



C3ETCV'25

3rd National Conference on Electronics, Electrical
Engineering, Telecommunications and Computer Vision

PROCEEDINGS

of the

*3rd National Conference on Electronics, Electrical Engineering, Telecommunications,
and Computer Vision (C3ETCV'25)*

Editors

Khaled HARRAR

Hamza AKROUM

Abderrezak AIBECHE

Yassine MERAIHI

Karim BAICHE

Mohammed AMMAR

Nouredine MESSAOUDI

Aimad BOUDOUDA

Youcef TABET

Aimad AHRICHE



الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research
جامعة امحمد بوقرة بومرداس
University M'Hamed Bougara of Boumerdes, Algeria
كلية التكنولوجيا
Faculty of Technology
قسم هندسة الأنظمة الكهربائية
Electrical Systems Engineering Department
مخبر هندسة الأنظمة و الاتصالات



Laboratory of Systems Engineering and Telecommunications (LIST)



3rd National Conference on Electronics, Electrical Engineering, Telecommunications, and Computer Vision

C3ETCV'25

Boumerdes – Algeria, Dec 10, 2025



ISBN : 978-9969-9733-8-9

Dépôt légal : 9969-2026

Edition



Faculty of Technology, of the University M'Hamed Bougara of Boumerdes, Algeria

C3ETCV 2025

Editors:

*Khaled HARRAR
Abderrezak AIBECHE
Karim BAICHE
Nouredine MESSAOUDI
Youcef TABET*

*Hamza AKROUM
Yassine MERAIHI
Mohammed AMMAR
Aimad BOUDOUDA
Aimad AHRICHE*

Copyright © C3ETCV'25

*All Rights Reserved to the 3rd National Conference on Electronics, Electrical Engineering,
Telecommunications, and Computer Vision*

Preface

We are pleased to present the Proceedings of the 3rd National Conference on Electronics, Electrical Engineering, Telecommunications, and Computer Vision (C3ETCV'25). Building on the success of its previous editions, C3ETCV has established itself as a recognized national forum for scientific exchange and collaboration in these rapidly advancing technological domains.

The conference series was initiated with its first edition at the faculty of technology, University M'Hamed Bougara of Boumerdes, followed by a second edition hosted by the University of Mila, and returns for its third edition to the University of Boumerdes, reaffirming its role as a rotating scientific event that brings together researchers from different academic and industrial institutions across the country.

C3ETCV'25 aims to provide an interdisciplinary platform where academics, engineers, and practitioners can present recent research results, discuss emerging challenges, and explore innovative solutions in the broad areas of Electronics, Electrical Engineering, Telecommunications, and Computer Vision. Through this forum, the conference seeks to encourage knowledge transfer, foster scientific cooperation, and promote high-quality research aligned around three main tracks:

- **Track 1: Electronics and Electrical Engineering :** *This track addresses modern electronic and electrical systems, including embedded systems, IoT applications, power electronics, control systems, and energy-related technologies.*
- **Track 2: Telecommunications :** *This track covers topics such as wireless and mobile communications, smart grids, communication networks, and emerging paradigms including 5G and beyond.*
- **Track 3: Computer Vision :** *This track encompasses artificial intelligence, signal and image processing, machine learning, and bioinformatics, highlighting advances that enable machines to analyze, understand, and interact with visual data.*

The contributions included in these proceedings reflect the diversity and depth of current research activities, ranging from fundamental studies to real-world implementations. We believe that the presented works will serve as a valuable reference for researchers and will stimulate further investigation and innovation in the covered fields.

We would like to express our sincere appreciation to all authors for their high-quality submissions, to the reviewers for their rigorous and constructive evaluations, and to the organizing and technical committees for their dedication and commitment.

Thank you for being part of C3ETCV'25

*Conference Chairs
Pr. Khaled HARRAR
Dr. Hamza AKROUM*

Organizers

Honorary General Chairs

Pr. ABDELBAKI Nouredine, Rector of Univ. Boumerdes, Algeria
Pr. SAIDI Mohammed, Dean of the Faculty of Technology, Univ.
Boumerdes, Algeria

Conference General Chairs

Pr. HARRAR Khaled, Univ. Boumerdes, Algeria
Dr. AKROUM Hamza, Univ. Boumerdes, Algeria

Conference General Co-Chairs

Pr. MESSAOUDI Nouredine, Univ. Boumerdes, Algeria
Pr. BOUDOUDA Aimad, Univ. Boumerdes, Algeria
Pr. AMMAR Mohammed, Univ. Boumerdes, Algeria
Pr. AHRICHE Aimad, Univ. Boumerdes, Algeria
Dr. TABET Youcef, Univ. Boumerdes, Algeria

Organizing Committee Chair

Pr. BAICHE Karim, Univ. Boumerdes, Algeria
Pr. MERAIHI Yassine, Univ. Boumerdes, Algeria
Dr. AIBECHE Abderezzak, Univ. Boumerdes, Algeria

Organizing Committee

Dr. MESBAH Abdelhak, Univ. Boumerdes, Algeria
Dr. ABDELOUAHED touhami, Univ. Boumerdes, Algeria
Dr. AMROUCHE Dahbia, Univ. Boumerdes, Algeria
Dr. Belynda Fares, Univ. Boumerdes, Algeria
Dr. Benia Fatima, Univ. Boumerdes, Algeria
Dr. GANA Massine, Univ. Boumerdes, Algeria
Dr. BENSAFI Nouredine, Univ. Boumerdes, Algeria
Dr. TRIKI Yacine, Univ. Boumerdes, Algeria
Miss. IBAOUNI Faiza, Univ. Boumerdes, Algeria
Miss. TIRECHE Sihem, Univ. Boumerdes, Algeria

Scientific Committee

Pr. ACHELI Dalila, Univ. Boumerdes, Algeria
Pr. AHRICHE Aimad, Univ. Boumerdes, Algeria
Pr. AMMAR Mohammed, Univ. Boumerdes, Algeria
Pr. AYAD Mouloud, Univ. Bouira, Algeria

Pr. BAICHE Karim, Univ. Boumerdes, Algeria
Pr. BELAIDI Hadjira, Univ. Boumerdes, Algeria
Pr. BELKACEM Samia, Univ. Boumerdes, Algeria
Pr. BENLATRECHE Mohamed Salah, Univ. Mila, Algeria
Pr. BENMOUSSA Yahia, Univ. Boumerdes, Algeria
Pr. BENZAOUI Amir, Univ. Skikda, Algeria
Pr. BOUDEN Toufik, Univ. Jijel, Algeria
Pr. BOUDOUDA Aimad, Univ. Boumerdes, Algeria
Pr. BOUKABOU Abdelkrim, Univ. Jijel, Algeria
Pr. BOUSHAKI Razika, Univ. Boumerdes, Algeria
Pr. CHALLAL Mouloud, Univ. Boumerdes, Algeria
Pr. CHELLIL Ahmed, Univ. Boumerdes, Algeria
Pr. DAAMOUICHE Abdelhamid, Univ. Boumerdes, Algeria
Pr. FELLAG Sid Ali, Univ. Boumerdes, Algeria
Pr. FERGUENE Farid, Univ. Bab Ezzouar, Algeria
Pr. HAMADOUCHE M'Hamed, Univ. Boumerdes, Algeria
Pr. HARKAT Mohammed Faouzi, Univ. Annaba, Algeria
Pr. HARRAR Khaled, Univ. Boumerdes, Algeria
Pr. HENTOUT Abdelfetah, CDTA, Algeria
Pr. HOCINI Abdesselem, Univ. M'sila, Algeria
Pr. IDIR Abdelhakim, Univ. M'sila, Algeria
Pr. KHELDOUN Aissa, Univ. Boumerdes, Algeria
Pr. KHOUAS Abdelhakim, Univ. Boumerdes, Algeria
Pr. KIDOUICHE Madjid, Univ. Boumerdes, Algeria
Pr. KOUADRI Abdelmalek, Univ. Boumerdes, Algeria
Pr. KOUINI Benalia, Univ. Boumerdes, Algeria
Pr. LACHEB Samir, Univ. Boumerdes, Algeria
Pr. LACHOURI Abderezzak, Univ. Skikda, Algeria
Pr. LAHDIR Mourad, Univ. Tizi Ouzou, Algeria
Pr. LAZRI Mourad, Univ. Tizi Ouzou, Algeria
Pr. MERAIHI Yassine, Univ. Boumerdes, Algeria
Pr. MESSAOUDI Noureddine, Univ. Boumerdes, Algeria
Pr. OUALLOUCHE Fethi, Univ. Tizi Ouzou, Algeria
Pr. RAHMOUNE Fayçal, Univ. Boumerdes, Algeria
Pr. RIAHLA Mohamed Amine, Univ. M'sila, Algeria
Pr. SAFI Brahim, Univ. Boumerdes, Algeria
Dr. ABDESLAME Djamila, Univ. Boumerdes, Algeria
Dr. AIBECHE Abderezzak, Univ. Boumerdes, Algeria
Dr. AKLI Isma, CDTA, Algeria
Dr. AKROUM Hamza, Univ. Boumerdes, Algeria
Dr. AMMAR Abdelkarim, Univ. Boumerdes, Algeria
Dr. BADDARI Ibtihel, Univ. Boumerdes, Algeria
Dr. BELAOURA Widad, Univ. Boumerdes, Algeria

Dr. BENBELKACEM Samir, CDTA, Algeria
Dr. BENSAFI Nouredine, Univ. Boumerdes, Algeria
Dr. BOURKACHE Nouredine, Univ. Tizi Ouzou, Algeria
Dr. BOUTELLAA El Hocine, CDTA, Algeria
Dr. BOUZAR-BENLABIOD Lydia, Univ. Acadia, Canada
Dr. BOUZID-DAHO Abdellatif, Univ. Tizi Ouzou, Algeria
Dr. CHERRAT Nidhal, Univ. Boumerdes, Algeria
Dr. DAFEUR Hakim, Univ. Boumerdes, Algeria
Dr. DICHOU Karima, Univ. Boumerdes, Algeria
Dr. DOGHMANE Mohamed, Univ. Boumerdes, Algeria
Dr. GANA Massine, Univ. Boumerdes, Algeria
Dr. GARMAT Abdelkader, Univ. Boumerdes, Algeria
Dr. HAMADACHE Moussa, Univ. Boumerdes, Algeria
Dr. HAMEL Meziane, Univ. Boumerdes, Algeria
Dr. HEDIR Mehdi, Univ. Boumerdes, Algeria
Dr. HIMEUR Yacine, CDTA, Algeria
Dr. HOCINE Faiza, Univ. Boumerdes, Algeria
Dr. IKHLEF Boualem, Univ. Boumerdes, Algeria
Dr. KHELOUAT Lila, Univ. Boumerdes, Algeria
Dr. KOUIDER Ahmed, CDTA, Algeria
Dr. LACHEKHAB Fadila, Univ. Boumerdes, Algeria
Dr. MAHDI Ismahane, Univ. Boumerdes, Algeria
Dr. MAHSEUR Mohammed, Univ. Alger-1, Algeria
Dr. MAOUCHE Riadh, Univ. Boumerdes, Algeria
Dr. MANSOUL Ali, CDTA, Algeria
Dr. MASSOUT Samia, Univ. Boumerdes, Algeria
Dr. MECHAKRA Hamza, Univ. Boumerdes, Algeria
Dr. MEKHMOUKH TALEB Sylia, Univ. Boumerdes, Algeria
Dr. MERAH Mostefa, Univ. Mostaghanem, Algeria
Dr. MESBAH Abdelhak, Univ. Boumerdes, Algeria
Dr. MILOUDI Lalia, Univ. Boumerdes, Algeria
Dr. MOUATSI Abdelmalek, Univ. Boumerdes, Algeria
Dr. NOUAL Amina, Univ. Boumerdes, Algeria
Dr. OMARI Tahar, Univ. Boumerdes, Algeria
Dr. REZZOUG Abdellah, Univ. Boumerdes, Algeria
Dr. RIACHE Samah, Univ. Boumerdes, Algeria
Dr. SEDJELMACI Ibticeme, Univ. Boumerdes, Algeria
Dr. SMAANI Billel, Univ. Mila, Algeria
Dr. TABET Youcef, Univ. Boumerdes, Algeria
Dr. TADJER Sid Ahmed, Univ. Boumerdes, Algeria
Dr. TOUHAMI Abdelouahed, Univ. Boumerdes, Algeria
Dr. TOUBAL MAAMAR Alla Eddine, Univ. Boumerdes, Algeria
Dr. TRIKI Yacine, Univ. Boumerdes, Algeria

Invited Speaker

A Survey of Lyapunov Function: Overlapping Decomposition Approach of Large Scale Dynamical System



Pr. Madjid KIDOUCHE
Univ. Bumerdes, Algeria

Biography

Kidouche Madjid was born in Bordj-Menaiel, Algeria. He received Engineering, Master of Sciences, and Ph.D degrees all in control theory. He joined M'Hamed Bougara University of Bumerdes, Algeria in 1990 where he is a full Professor in the department of automation and electrification of industrial process. He is a research group head on "Control of complex dynamical systems" at Applied Automatic Control Laboratory. He has been actively involved in several research projects in the fields of control and power system analysis. He is the author and co-author of numerous research publications in international conferences and recognized journals. His research interests include modeling and control of dynamic nonlinear systems, stability of large scale systems, fuzzy and sliding mode control.

Papers

Contents

• Ibrahim Yassine Nouali <i>Ultra-Wideband Technology for Accurate Indoor Localization via Time-of-Arrival Estimation</i>	p. 1 - 3
• Hana Elhachi <i>Comparative Analysis of Connection Establishment in TCP and QUIC</i>	p. 4 - 8
• Fatma Zahra Mahdjoub, Mahmoud Maamri, Azzedine Gattal <i>Experimental Validation of a Dual MEMS Microphone Array for Cardioid Directivity</i>	p. 9 – 13
• Nassim Zemirline, Nadir Kabeche, Samir Moulahoum, Sofiane Khechafi, Riadh Euldji <i>Fault Isolation and Power Quality Enhancement Using Solid-State Transformers</i>	p. 14 – 19
• Halima Sahraoui, Dahani Ameer <i>Performance Evaluation and Analysis of the DigiMesh Protocol</i>	p. 20 – 25
• Mabrouk Dahane, Abdelkrim Benali, Hamza Tedjini <i>Multi-step Gray Wolf optimization-based FOPID controller for the DFIG-Wind energy conversion system</i>	p. 26 – 29
• Abdelmounaime Mechiki, Sid Ahmed Soumia, Zoubeida Messali <i>Medical Image Denoising Using Diffusion Models: An Experimental Study on Chest X-Ray Images</i>	p. 30 – 35
• Belynda Fares, Radia Abdelli, Abderrezak Aibeche, AHCENE Bouzida, Safia Azzougui <i>Performance Characterization and Fault Impact Analysis of a Self-Excited Asynchronous Generator for Stand-Alone Applications</i>	p. 36 – 41
• Sabrina Meddah, Mohamed Zinelabidine Doghmane, Abdelhakim Idir, Sid Ahmed Tadjer, Madjid Kidouche <i>Enhanced Control of Rotary Drilling Systems Using FOPID</i>	p. 42 – 46
• Mawloud Tellaa, Abderrezak Aibeche, Mohamed Zinelabidine Doghmane, Madjid Kidouche, Aimad Boudouda, Hamza Akroum <i>Grey Wolf Optimizer-Based PID Control for Suppressing Stick-Slip Vibrations in Drilling Systems</i>	p. 47 – 52
• Yasmine Silhadi, Hocine Chebi, Rachida Aoudjit <i>Contribution on detection vehicle from algorithms YOLO</i>	p. 53 – 58
• Hocine Chebi, Yasmine Silhadi <i>Optimization of the parameters of a stochastic gradient algorithm for precision agriculture</i>	p. 59 – 62
• Yousra Kateb, Abdelmalek Khebli, Salah Aguib <i>Comparative Study of Deep Transfer Learning Models for Steel Surface Defect Classification</i>	p. 63 – 67
• Rachda El Taani, Karim Baiche, Youcef Tabet, Yassine Meraihi <i>An Enhanced Osprey Optimization Algorithm for Optimal UAV-BS Deployment</i>	p. 68 - 72
• Fatma Zohra Besmi, Samia Belkacem, Nouredine Messaoudi <i>Enhancement Technique for ECG Signal Protection Using Dual-Layer MEHC and DNA Computing</i>	p. 73 – 76
• Sofiane Latreche, Hocine Bellahsene <i>Advanced Blind MRE Equalizer for Modern MIMO Systems</i>	p. 77 – 82
• Somia Taba, Mohamed Amine Benchana <i>A Hybrid Chaotic–Optimization Framework for PAPR Reduction in Multicarrier OFDM Systems</i>	p. 83 – 87
• Ahmed Hichem Zebboudj, Hamza Akroum, Mohamed Zinelabidine Doghmane, Madjid Kidouche <i>Parametric Influence on Coupled Axial-Torsional Vibrations in Rotary Drilling Systems</i>	p. 88 – 93
• Riadh Degachi, Said Ghendir <i>Enhanced 2.4 GHz Microstrip Patch Antenna with DGS for Energy Harvesting</i>	p. 94 – 97
• Aicha Moulay, Koussa Khaled, Otmane Abdelkhalek	p. 98 – 102

• <i>Identification of Electrical Faults in Photovoltaic System using Artificial Intelligence</i>	p. 103 – 108
• Nouredine Bensafi, Mourad Lazri	
• <i>Ensemble Learning for Delineation of Raining Cloud over the North of Algeria</i>	p. 109 – 114
• Ahlem Benaissa, Saad Babesse	
• <i>Deep Reinforcement Learning with Vision-Based Control for Dynamic Target Tracking and Obstacle Avoidance in UR5 Manipulator</i>	p. 115 – 118
• Sonia Benbelgacem, Mohamed Benbelgacem, Larbi Guezouli, Rachid Seghir	
• <i>Exploration the Capabilities of Vision Transformers for Multimedia Information Retrieval</i>	p. 119 – 122
• Amel Abderrahmane	
• <i>Federated Learning-Enabled M2M Communication for Adaptive IoT Systems Using MQTT</i>	p. 123 – 128
• Yasmine Yesli, Mourad Lahdir	
• <i>Toward AI-Assisted Breast Cancer Diagnosis Inspired by Traditional Chinese Medicine: A State-of-the-Art Review</i>	p. 129 – 134
• Fateh Khelifi, Aimad Boudouda, Sid Ahmed Tadjer	
• <i>Sliding Mode Control of Series Multicell Inverters: Simulation and Performance Analysis on 3-Cell and 5-Cell Topologies</i>	p. 135 – 140
• Assala Nacef, Lina Benchikh, Ouissem Touameur, Djamila Mechta, Rahma Berchi	
• <i>A Comprehensive Survey of Short-Term Load Forecasting Techniques for Smart Grids</i>	p. 141 – 146
• Yacine Triki, Ali Bechouche, Hamid Seddiki, Djaffar Ould Abdeslam	
• <i>Adaptive Neural Linear-Based MPPT for Enhanced Efficiency in Photovoltaic Systems</i>	p. 147 – 152
• Lina Benchikh, Assala Nacef, Lemia Louail, Djihad Arrar	
• <i>Real-Time Energy Routing in Smart Grids with Harmony Search Algorithm</i>	p. 153 – 158
• Sofiane Latreche	
• <i>Evaluation of Custom Modulation Constellations for 6G Based on PAPR and CCDF Analysis</i>	p. 159 – 163
• Abdallah Ben Abdelkader, Soumia Kail, Mohammed Amine Soumeur, Othmane Abdelkhalek	
• <i>Standard-Compliant Hydraulic-Aware MPPT Benchmark for Grid-Connected PV Water-Pump Systems under Realistic Partial Shading</i>	p. 164 – 168
• Yacine Ouazziz, Dina Alkama	
• <i>RIS-Aided Restoration of NLoS connectivity in 5G Sub-6 GHz Networks</i>	p. 169 – 174
• Abdelghani Takha, Mohamed Lamine Talbi, Souad Ksenna	
• <i>Comparative Analysis of Fractional PI Rectangular Rules and Classical RK4 Integration for High-Fidelity Synthetic ECG Signal Modeling</i>	p. 175 – 180
• Faiza Bella, Ali Berrichi, Abdelouahab Moussaoui	
• <i>Deep Learning-Based Classification of Intestinal Diseases: A Comparative Analysis of ConvNeXt and CNN Models</i>	p. 181 – 184
• Hamza Trir	
• <i>Influence of Annealing on the Dark and Photo-Induced I–V Behavior of a ZnO:Ga — Cu₂O Heterojunction Fabricated via Ultrasonic Spray Deposition</i>	p. 185 – 190
• Abdelilah Rahim, Mohamed Mouloudi, Massine Gana	
• <i>Intelligent Traffic Signal Control Using YOLOv8 for Enhanced Urban Mobility and Emergency Response</i>	p. 191 – 196
• Bilal Benarabi, Mokhtar Khenfer, Mohammed Said Achbi, Boubakeur Rouabah, Mohamed Abdelbasset Mahboub, Sihem Kechida	
• <i>Intelligent Real-Time Supervision of Industrial Systems via MATLAB Integration with Multiple PLCs: A Case Study in a Desalination Plant</i>	p. 197 – 201
• Samira Hadibi, Madjid Kidouche, Mohamed Zine Elabidine Doghmane	
• <i>From Simplified to Realistic Drilling Dynamics: Comparing 2DOF and 4DOF Models with Top-Drive Effects and Hook-Load-Based ROP</i>	p. 202 – 206
• Lalia Miloudi, Bassem Laouazi, Omar Dif	
• <i>Magnetic Levitation System Modelling and Control</i>	

• Massine Gana, Zakia Chelli, Rafik Saddaoui, Sadjia Boubouzal, Abderezak Achour, Mourad Laghrouche <i>LoRaWAN-Enabled Internet of Medical Things (IoMT) for Remote Vital Sign Monitoring</i>	p. 207 – 211
• Assala Nacef, Lina Benchikh, Ouissem Touameur, Djamila Mechta, Djihad Arrar <i>Solving Peer-to-Peer energy trading using Minimum Cost Flow in smart grids</i>	p. 213 – 218
• Yousra Kateb, Abdelmalek Khebli, Salah Aguib <i>A Comprehensive Method for Steel Surface Classification Using MobileNet-V2 and Leaky ReLU Activation Function</i>	p. 219 – 223
• Asmaa Chatta, Boubakeur Latreche, Mustapha Bouakkaz, Hicham Bouakkaz <i>A Two-Stage Deep Learning Framework for Accurate Plant Identification from Leaf Images</i>	p. 225 – 229
• Meriem Meissoune Zemouri, Khaled Harrar, Nouma Izeboudjen, Hamza Akroum <i>IoT-Based Neonatal Incubator Prototype</i>	p. 230 – 235
• Raid Boudi, Nardjes Bouchemal <i>Optimized Controller Placement for SD-IoT: A Latency–Reliability Aware Approach</i>	p. 236 – 241
• Dina Alkama, Yacine Ouazziz, Sylia Zenadji, Mohamed Amine Ouamri <i>PSO Algorithm-Based 3D Optimization of UAV Deployment for Maximum Coverage</i>	p. 242 – 246
• Mohammed Haddad, Badis Lekouaghet, Nidhal Cherrat <i>Adaptive Prescribed Performance Fault-Tolerant Control of Full-State Errors for Nonlinear Systems</i>	p. 247 – 252
• Moussa Guenani, Hamza Akroum, Khaled Harrar, Asmala Bin Ahmad <i>Improving retinal fundus image segmentation: A Study of Transfer Learning and Single Channel Color Impact</i>	p. 253 – 259
• Fatma Khater, Abderrezak Aibeche, Hamza Akroum, Aimad Boudouda, Mohamed Zinelabidine doghmane <i>Open- and Short-Circuit Faults Detection in Multicellular Inverters using Sliding Mode Observer</i>	p. 260 – 265
• Badredine Mameri, Abdelmalek Mouatsi <i>Congestion Control by the use of Nash equilibrium concept</i>	p. 266 – 270
• Zoubir Gadouche, Mokhtar Mahmoud Mohammedi, Mohamed Bey <i>Vector Control, DPC, and Fuzzy DPC: A Comparative Study for Grid Converters in Hybrid Wind-PV System</i>	p. 271 – 276
• Mohamed Ammara <i>Efficient Transfer Learning for Brain Tumor Classification: Comparative Analysis of Pre-Trained CNN Models</i>	p. 277 – 282
• Imane Mehidi, Djamel Eddine Chouaib Belkhiat, Dalel Jabri <i>Comparative Study of Improved FCM Algorithms in Retinal Blood Vessel Segmentation</i>	p. 283 – 288
• Ines Righi <i>Non-PDC Control Law Synthesis for Discret-Time Takagi-Sugeno Descriptor Systems subject to Input Saturation and L_2 Disturbance</i>	p. 290 – 294
• Reda Dermouche, Sid Ahmed Tadjer, Wissal Bahi, Mohamed Farouk Azouz <i>Fire Detection System Using Deep Learning</i>	p. 295 – 302
• Mohamed Chaouli, Mohamed Elbar <i>Privacy by Design: Hybrid Rule-Based and Zero-Shot Classification for Browser-Based IoT Threat Detection</i>	p. 303 – 309
• Amina Noual <i>Exploring the DC and RF Performance of AlGaIn/GaN MOS-HEMT Transistor with HfO_2 and Al_2O_3 Oxide Layers Grown on β-Ga$_2$O$_3$: A Numerical Analysis</i>	p. 310 – 314
• Abderrezak Aibeche, Aimad Boudouda, Belynda Fares, Hamza Akroum, Madjid Kidouche <i>Configuration of Differential Protection for Internal Short-Circuit Fault Detection in Power Transformer</i>	p. 315 – 319

Ultra-Wideband Technology for Accurate Indoor Localization via Time-of-Arrival Estimation

Ibrahim Yassine Nouali

STIC Laboratory, Department of Telecommunications

Faculty of Technology, Tlemcen University

Tlemcen, Algeria

ibrahimyassine.nouali@univ-tlemcen.dz

Abstract—The demand for precise indoor localization has grown, driven by applications in rescue operations, smart buildings, healthcare, and industrial automation. While Global Navigation Satellite Systems (GNSS) is apt for outdoor localization, its signals are severely attenuated inside buildings. Ultra-Wideband (UWB) technology has emerged as a strong contender for indoor usage due to its resolution in time, resistance to multipath interference, and low power spectral density. This paper provides a summary of the advantages of UWB technology for indoor localization with special focus on the Time-of-Arrival (TOA) estimation techniques. We introduce the fundamental principles of UWB ranging, its challenges in the form of non-line-of-sight (NLOS) and dense multipath, and a brief review of well-known TOA estimation algorithms with special emphasis on how their robustness and accuracy have been enhanced.

Index Terms—Indoor Localization, Ultra-Wideband (UWB), Time-of-Arrival (TOA), Ranging, Non-Line-of-Sight (NLOS), Multipath.

I. INTRODUCTION

The ability to accurately find people and equipment indoors forms the core of modern context-aware applications [1], [2]. From guiding wayfinding within complex buildings to warehouse logistics, which optimizes inventory locations and improves delivery times, to enabling augmented reality, accurate indoor positioning systems (IPS) are crucial. The area is highly challenging, however, as satellite-based signals like GPS cannot be trusted indoors due to signals being attenuated and subject to multipath propagation [3].

Among all the technologies that have been proposed for IPS, such as Wi-Fi, Bluetooth, and RFID, Ultra-Wideband (UWB) would be a best fit. UWB has the distinct characteristic of operating with extremely wide bandwidth (>500 MHz), employing short-pulse duration (nanosecond order) [4]. This inbuilt feature enhances its key strengths: high temporal resolution for precise ranging, ability to resolve spatially close multipath components, and minimum interference with other systems due to low transmission power.

Ranging-based IPS typically rely on estimating parameters like Received Signal Strength (RSS), Angle-of-Arrival (AOA), or Time-of-Arrival (TOA) [5]. TOA estimation is particularly well-suited for UWB systems as it directly leverages the technology's high time resolution to measure the flight time of a signal between a transmitter and a receiver, from which distance can be directly calculated [6]. The core challenge,

however, lies in the accurate detection of the first arriving path (FP), especially under non-line-of-sight (NLOS) conditions where the direct path is obstructed, and the strongest path may not be the first [7].

This paper offers a concise overview of UWB-based indoor localization. Section II details the advantages of UWB signals. Section III describes the system model and the TOA estimation challenge. Section IV surveys common TOA estimation methods, and Section V concludes the paper.

II. ADVANTAGES OF UWB FOR INDOOR LOCALIZATION

UWB technology offers a unique set of properties that make it ideal for high-precision indoor localization:

A. High Time Resolution and Ranging Accuracy

The extremely short duration of UWB pulses results in a very fine time resolution. This allows the system to distinguish between signals that arrive with very small delay differences, which is crucial for resolving individual multipath components and accurately pinpointing the time of arrival of the first path (FP) [8].

B. Multipath Robustness

Indoor environments are characterized by severe multipath propagation, where signals reflect off walls, ceilings, and objects, creating multiple copies that arrive at the receiver at different times. For narrowband systems, this leads to destructive interference and fading, as shown in Fig. 1. The fine time resolution of UWB allows it to distinguish these individual paths (Fig. 2), making it inherently robust to the fading effects that plague other technologies.

C. Penetration Capability and Low Power Spectral Density

UWB signals can penetrate common building materials like drywall and wood more effectively than narrowband signals. Furthermore, UWB transmissions operate under a strict emission mask enforced by regulatory bodies like the Federal Communications Commission (FCC), resulting in a very low power spectral density. This makes UWB systems resistant to interference from other devices and allows them to coexist with existing services without causing disruption.

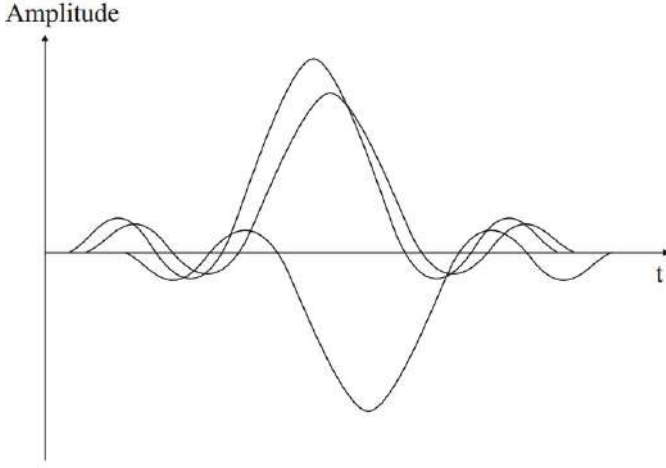


Fig. 1: Multipath effect in narrowband systems: overlapping reflections cause constructive/destructive interference (fading), making it difficult to identify individual paths [6].

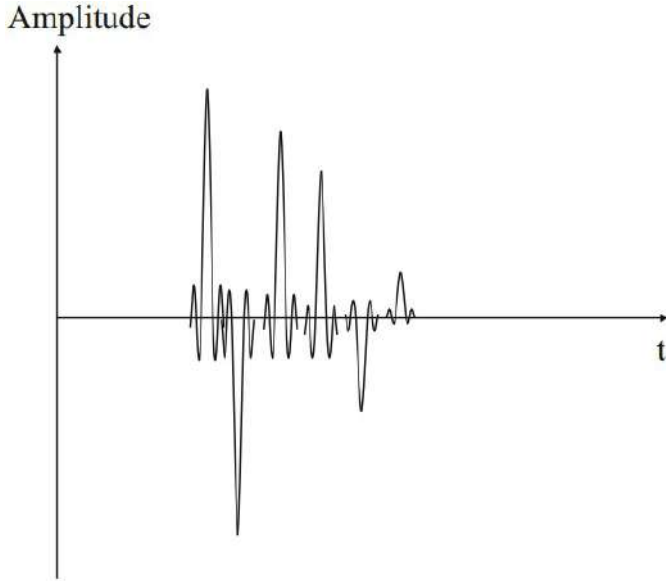


Fig. 2: Multipath resolution in UWB systems: short pulses allow distinct identification of the Line-of-Sight (LOS) and reflected paths, enabling accurate first Path (FP) detection [6].

III. UWB RANGING SYSTEM AND TOA CHALLENGE

A typical UWB ranging system uses a train of short pulses. A common choice for the pulse shape is the Gaussian pulse or its derivatives, which comply with regulatory spectral masks. The second derivative is often used [9]:

$$s(t) = \left[1 - 4\pi \left(\frac{t}{\alpha} \right)^2 \right] \exp \left[-2\pi \left(\frac{t}{\alpha} \right)^2 \right] \quad (1)$$

where α is a pulse shape factor.

The signal propagates through a channel characterized by the IEEE 802.15.4a standard models [10], which describe the multipath profile for various environments (e.g., CM1 for

residential LOS and CM2 for residential NLOS). The received signal $r(t)$ (an example is shown in Fig. 3) is the convolution of the transmitted signal $s(t)$ with the channel impulse response (CIR) $h(t)$, plus additive white Gaussian noise (AWGN) $n(t)$:

$$r(t) = s(t) * h(t) + n(t) \quad (2)$$

The CIR $h(t)$ is given by :

$$h(t) = \sum_{l=0}^L \sum_{k=0}^K \alpha_{k,l} e^{j\phi_{k,l}} \delta(t - T_l - \tau_{k,l}) \quad (3)$$

where L is the number of clusters and K is the number of multipath components within the ℓ -th cluster. The parameters $\alpha_{k,\ell}$ and $\phi_{k,\ell} \sim \mathcal{U}[0, 2\pi)$ denote the complex gain and phase of the k -th ray in the ℓ -th cluster, respectively. The delay of the ℓ -th cluster is T_ℓ , and $\tau_{k,\ell}$ is the relative delay of the k -th ray within that cluster.

The received signal can be written as :

$$r(t) = \sum_{l=0}^L \sum_{k=0}^K \alpha_{k,l} e^{j\phi_{k,l}} s(t - T_l - \tau_{k,l}) + n(t) \quad (4)$$

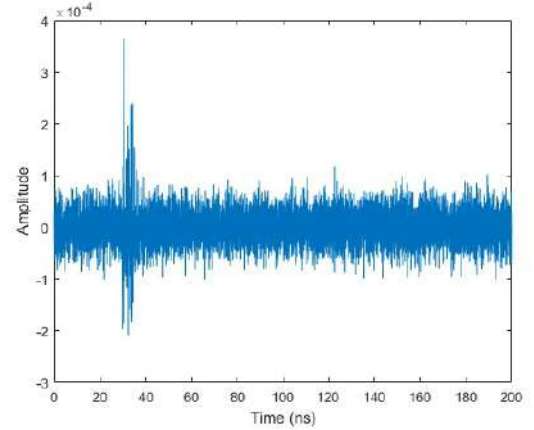


Fig. 3: Example of a received UWB signal in an indoor multipath environment (SNR = 30 dB). The challenge is to detect the arrival time of the first, often weak, path among the stronger, later-arriving multipath components.

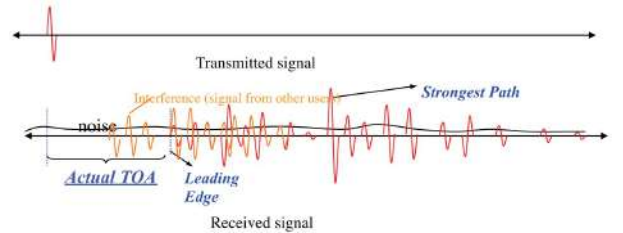


Fig. 4: Conceptual illustration of TOA estimation challenges: the estimated first path may differ from the true first path due to noise, leading to a ranging error. Under NLOS conditions, the direct path is blocked, and a reflected path is detected first, causing a positive bias in the distance measurement. [7]

The goal of TOA estimation is to accurately estimate the delay τ_{TOA} of the first path in $h(t)$ from the received signal $r(t)$. The primary obstacle is that in NLOS conditions, the FP is often attenuated and can be weaker than later-arriving multipath components or even buried in noise, making its detection non-trivial. Fig. 4 illustrates common error sources in this process.

IV. OVERVIEW OF TOA ESTIMATION METHODS

Various methods have been proposed to estimate the TOA in UWB systems. They can be broadly categorized into coherent and non-coherent techniques.

A. Coherent Techniques

Coherent receivers, such as those using matched filtering, correlate the received signal with a known template of the transmitted pulse. The time index of the maximum correlation output provides an estimate of the TOA. While these methods are optimal in additive white Gaussian noise (AWGN) channels and can achieve performance close to the Cramér-Rao Lower Bound (CRLB), they require precise channel knowledge and very high-speed sampling, leading to significant implementation complexity and cost.

B. Non-Coherent Techniques

Non-coherent techniques, particularly Energy Detection (ED), offer a lower-complexity alternative. They do not require precise pulse shaping or synchronization. An ED receiver typically squares the received signal and integrates it over short time intervals to compute energy samples [11].

- **Maximum Energy Selection (MES):** This simple method selects the time index of the block with the highest energy as the TOA estimate. Its performance degrades severely in NLOS conditions where the strongest path is not the first.
- **Threshold Crossing (TC):** This method selects the first energy block that exceeds a predetermined threshold. Its accuracy is highly sensitive to the choice of an optimal threshold, which is difficult to determine as it depends on the SNR and channel conditions.
- **Advanced Methods:** To overcome the limitations of basic ED methods, more sophisticated algorithms have been developed. These include techniques based on kurtosis analysis of wavelet-decomposed signals to set adaptive thresholds [12], change point detection algorithms that identify the statistical change in energy samples caused by the arrival of the signal [13], and double sliding window variance detection methods that locate a sharp change in the variance of the received signal to pinpoint the FP [14]. These advanced methods provide improved robustness against noise and NLOS effects compared to traditional MES and TC.

V. CONCLUSION

Ultra-Wideband (UWB) technology has emerged as a powerful enabler of high-precision indoor localization, thanks primarily to its large bandwidth, high time resolution, and inherent resilience to multipath propagation. Time-of-Arrival (TOA) estimation is the most direct and highest precision ranging method for UWB systems. While low complexity is achieved by straightforward energy detection methods, their accuracy is compromised in challenging NLOS environments. Recent research has been focused on developing more advanced algorithms, i.e., change point detection and variance analysis, to more precisely identify the first arriving path and further push the performance of UWB-based IPS towards its theoretical limits. The continued development of such methods ensures that UWB will remain at the forefront of accurate indoor positioning solutions.

ACKNOWLEDGMENT

The author would like to thank the STIC Laboratory

REFERENCES

- [1] H. Obeidat, W. Shuaieb, O. Obeidat, and R. Abd-Alhameed, "A review of indoor localization techniques and wireless technologies," *Wireless Personal Communications*, vol. 119, no. 1, pp. 289–327, 2021.
- [2] F. Zafari, A. Gkelias, and K. K. Leung, "A survey of indoor localization systems and technologies," *IEEE Communications Surveys Tutorials*, vol. 21, no. 3, pp. 2568–2599, 2019.
- [3] G. Shi and Y. Ming, "Survey of indoor positioning systems based on ultra-wideband (uwb) technology," in *Wireless Communications, Networking and Applications*. Springer, 2004, pp. 1269–1278.
- [4] A. Alarifi, A. Al-Salman, M. Alsaleh, A. Alnafessah, S. Al-Hadhrani, M. A. Al-Ammar, and H. S. Al-Khalifa, "Ultra wideband indoor positioning technologies: Analysis and recent advances," *Sensors*, vol. 16, no. 5, p. 707, 2016.
- [5] R. F. Brena, J. P. Garcia-Vazquez, C. E. Galvan-Tejada, D. Munoz-Rodriguez, C. Vargas-Rosales, and J. Fangmeyer, "Evolution of indoor positioning technologies: A survey," *Journal of Sensors*, vol. 2017, pp. 1–21, 2017.
- [6] F. Mazhar, M. G. Khan, and B. Sallberg, "Precise indoor positioning using uwb: A review of methods, algorithms and implementations," *Wireless Personal Communications*, vol. 97, no. 3, pp. 4467–4491, 2017.
- [7] S. Aditya, A. F. Molisch, and H. M. Behairy, "A survey on the impact of multipath on wideband time-of-arrival based localization," *Proceedings of the IEEE*, vol. 106, no. 7, pp. 1183–1203, 2018.
- [8] D. Dardari, A. Conti, U. Ferner, A. Giorgetti, and M. Z. Win, "Ranging with ultrawide bandwidth signals in multipath environments," *Proceedings of the IEEE*, vol. 97, no. 2, pp. 404–426, 2009.
- [9] I. Guvenc and Z. Sahinoglu, "Threshold-based toa estimation for impulse radio uwb systems," in *2005 IEEE International Conference on Ultra-Wideband*, 2005, pp. 420–425.
- [10] A. F. e. a. Molisch, "Ieee 802.15.4a channel model - final report," IEEE P802.15 Working Group for Wireless Personal Area Networks (WPANs), Tech. Rep. 15-04-0662-00-004a, 2004.
- [11] W. Liu and X. Huang, "Analysis of energy detection receiver for toa estimation in ir uwb ranging and a novel toa estimation approach," *Journal of Electromagnetic Waves and Applications*, vol. 28, no. 1, pp. 49–63, 2014.
- [12] J. Li, X. Cui, H. Song, Z. Li, and J. Liu, "Threshold selection method for uwb toa estimation based on wavelet decomposition and kurtosis analysis," *EURASIP Journal on Wireless Communications and Networking*, vol. 2017, no. 1, pp. 1–10, 2017.
- [13] I. Y. Nouali, Z. Slimane, and A. Abdelmalek, "Change point detection-based toa estimation in uwb indoor ranging systems," 2021, unpublished Conference Paper.
- [14] —, "Double sliding window variance detection-based time-of-arrival estimation in ultra-wideband ranging systems," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 6, pp. 6303–6310, 2022.

Comparative Analysis of Connection Establishment in TCP and QUIC

Elhachi Hana

Telecommunications Laboratory

Telecommunications Institut

Guelma, Algeria

elhachi.hana@univ-guelma.dz

Abstract—Efficient connection establishment is a critical factor in determining the performance of latency-sensitive applications such as video streaming, vehicular communications, and mobile health services. Traditional Transmission Control Protocol (TCP) employs a three-way handshake to establish reliability, followed by a separate Transport Layer Security (TLS) handshake to provide encryption. This layered approach increases connection setup time, particularly in high-latency and lossy networks. In contrast, QUIC (Quick UDP Internet Connections), a transport protocol standardized by the IETF, integrates TLS 1.3 into the transport handshake, enabling secure connections within a single round trip and supporting 0-RTT session resumption. In this paper, we present a comparative study of TCP and QUIC connection establishment based on two key metrics: handshake duration and first byte latency (FBL). Simulation results under varying round-trip times and packet loss rates demonstrate that QUIC consistently outperforms TCP in reducing application-perceived startup delay. While TCP exhibits shorter handshakes when used without encryption, its TLS-enabled variant (TLS-over-TCP) introduces the longest setup times. QUIC, despite slightly higher initialization overhead, delivers faster application data availability due to its streamlined cryptographic negotiation. These findings confirm that QUIC is better suited for dynamic and mobile environments where minimizing connection setup latency is crucial.

Keywords— *Transport protocols, video streaming, VANET, networking, video coding.*

I. INTRODUCTION

The rapid growth of real-time applications such as video streaming, online gaming, and mobile health services has made low-latency communication a critical requirement [1]. One of the key contributors to end-to-end latency is the connection establishment process, during which a transport protocol initializes communication between client and server. Traditional Transmission Control Protocol (TCP) relies on a three-way handshake followed by a separate Transport Layer Security (TLS) negotiation to provide secure connections [2]. This multi-step procedure introduces additional delay, which can severely affect performance in dynamic and resource-constrained environments such as vehicular networks [3].

QUIC (Quick UDP Internet Connections), a novel transport protocol originally developed by Google and later standardized by the IETF in 2013, addresses these limitations by integrating the transport handshake with cryptographic negotiation [4]. By eliminating the sequential TCP+TLS exchange and enabling 0-RTT session resumption, QUIC is capable of reducing the

connection establishment latency significantly [5]. This feature makes it particularly attractive for applications requiring fast session setup under conditions of high mobility and frequent disconnections, such as 5G-enabled vehicular and mobile health networks [6], [7]. In this work, we present a detailed comparative analysis of the connection establishment process in TCP and QUIC when streaming medical videos in vehicular environments. Our evaluation focuses on two primary metrics: handshake duration and first byte latency (FBL). The former measures the time required to complete the initial handshake, while the latter captures the delay until the first application data is successfully delivered [8].

The remainder of this paper is organized as follows. Section II reviews the fundamentals of TCP and QUIC handshakes, emphasizing their differences in terms of connection setup. Section III outlines a brief overview of the related works. Section IV describes the experimental methodology, including the network scenarios and simulation setup. Section V presents and analyzes the comparative results of handshake duration and first byte latency. Finally, Section VI concludes the paper and highlights possible directions for future work, including multipath QUIC and cross-layer integration.

A. TCP Handshake:

TCP establishes a connection using a three-way handshake in which the client and server exchange synchronization (SYN) and acknowledgment (ACK) messages before data transmission can begin [2]. To enable secure communication, TCP is typically combined with the Transport Layer Security (TLS) protocol. In traditional deployments, this requires an additional handshake phase after the TCP session is established [9]. For instance, the widely used TLS 1.2 adds two more round-trips, while the more recent TLS 1.3 reduces this overhead to a single round-trip [10]. Consequently, the total connection establishment in TCP over TLS often involves multiple sequential exchanges, resulting in increased latency before the first application data is delivered. This delay becomes more pronounced in high-latency networks, such as mobile or satellite links, where each additional round trip significantly impacts the overall startup time [11].

B. QUIC Handshake

QUIC was designed to address these inefficiencies by integrating the cryptographic handshake directly into the transport protocol [4]. Built on top of UDP, QUIC combines the transport-layer and TLS 1.3 handshakes

into a single procedure, thereby reducing the connection setup to 1-RTT in the general case. Furthermore, QUIC introduces a 0-RTT mechanism that allows clients to send application data during session resumption without waiting for the handshake to complete, provided that cryptographic state has been cached from a previous connection [12]. This design not only reduces handshake latency but also eliminates the head-of-line (HoL) blocking problem during connection setup, since streams are multiplexed independently [5].

The primary difference between TCP and QUIC in connection establishment lies in the number of round trips required before application data can be exchanged. TCP, even with TLS 1.3, requires at least two separate phases (TCP + TLS), see **Fig.1**, and thus more than one round trip in most cases. QUIC, by contrast, integrates transport and cryptographic negotiation, resulting in faster session establishment. In scenarios with high mobility and frequent reconnections, such as vehicular networks or mobile health systems, QUIC's reduction in handshake duration and support for 0-RTT significantly decreases first byte latency (FBL) and improves user-perceived responsiveness [6], [8].

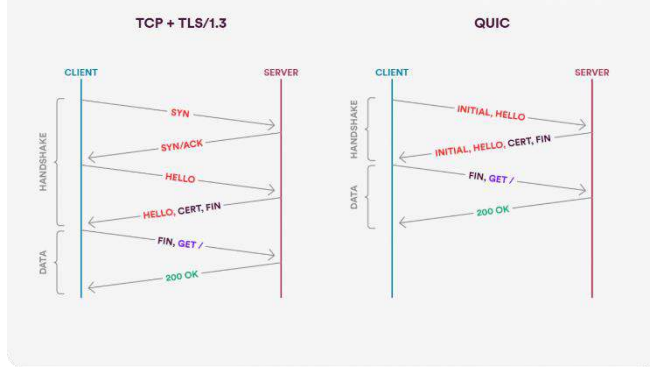


Fig.1. TCP-TLS vs QUIC handshakes

II. RELATED WORK

This section reviews existing studies comparing QUIC and TCP and their implications for video streaming over next-generation networks. We highlight prior research on QUIC's design, performance evaluations over mobile and 5G environments, and its impact on video streaming quality.

The authors in [13] provide an in-depth performance evaluation of QUIC across web, cloud storage, and video workloads compared to TLS/TCP. Their results highlight QUIC's advantage in reducing connection times and mitigating video stalls, though at the cost of higher CPU utilization and without exploring mobility scenarios. Similarly, [14] investigates QUIC's behavior under LTE conditions using ns-3, demonstrating improved throughput and download times under favorable channel conditions, while performance converges with TCP in poor networks; however, the analysis excludes vehicular environments and 5G characteristics. An implementation-focused study in [15] explores QUIC's architecture, applications, and security aspects through AWS-based

experiments, uncovering issues like data loss recovery and the need for forward error correction, but it remains constrained to static cloud settings. Furthermore, [16] evaluates Media over QUIC (MoQ) against RTMP+HLS over TCP in a cloud-based scenario, showing MoQ's efficiency in throughput utilization through relay and publisher-subscriber models; yet, the work overlooks dynamic mobility and heterogeneous network conditions. Another study [17] examines QUIC's behavior over mmWave channels in 5G environments, showing improved throughput and delay compared to TCP under fluctuating conditions, but focusing exclusively on simulated scenarios without considering vehicular mobility.

Further research in [18] analyzes multiple QUIC implementations and their congestion control mechanisms for video streaming in Video-on-Demand (VoD) and Low-Latency Live Streaming (LLL) scenarios. The study reveals significant performance variations even with identical congestion control algorithms, impacting QoE and emphasizing the importance of cross-layer optimization. Similarly, [19] introduces a Software-Defined Service Migration (SDSM) approach combining QUIC, MEC, and SDN to enhance service continuity in dynamic 5G environments. While demonstrating improved recovery and reduced migration times, the work primarily addresses edge computing scenarios rather than video streaming in vehicular networks.

However, none of these studies address real-time video streaming in highly dynamic vehicular environments over 5G networks using experimental testbeds. This gap motivates our work, which evaluates TCP and QUIC for video delivery under realistic VANET scenarios.

III. METHODOLOGY AND SIMULATION TOOLS

To evaluate the performance of transport protocols for video streaming in vehicular environments, we designed a simulation framework built on the Mininet-WiFi emulator. **Fig.2.** illustrates the proposed architecture, which integrates three core modules to achieve a realistic and flexible evaluation platform. These modules include:

- **Vehicular Mobility Module:** responsible for generating realistic vehicle movement and traffic patterns.
- **Network Emulation Module:** modeling wireless communication characteristics, including link quality and network dynamics.
- **Video Processing Module:** handling video encoding, decoding, and streaming operations to simulate real-world multimedia transmission.

Together, these modules form a comprehensive environment capable of reproducing the complexity of video delivery over vehicular ad hoc networks (VANETs).

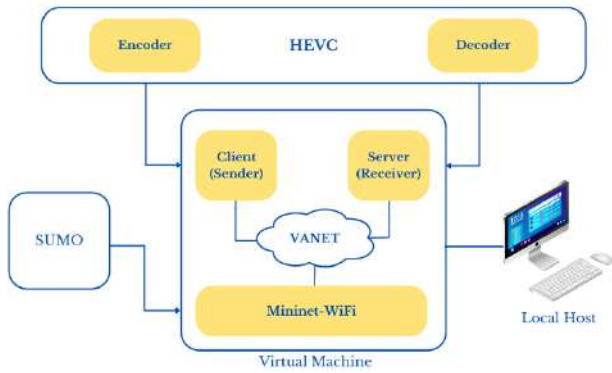


Fig.2. The proposed architecture for VANET modeling.

SUMO for Mobility Modeling: The vehicular movement patterns were modeled using SUMO (Simulation of Urban Mobility) [20], an open-source platform widely adopted in V2X research. SUMO supports the simulation of vehicles, pedestrians, and public transport systems with precise mobility control, allowing researchers to create realistic traffic scenarios.

Mininet-WiFi for Network Emulation: For the network layer, Mininet-WiFi [21] was employed as it extends the original Mininet emulator [22] to include wireless functionality. This environment allows the creation of virtual Wi-Fi stations and access points using native Linux drivers, along with mobility features for wireless nodes.

Video Encoding with HEVC: Video compression was implemented using High-Efficiency Video Coding (HEVC), also known as H.265. HEVC provides up to 50% better compression compared to H.264/AVC while preserving image quality. The test video used in the experiments, the *Ultrasound* sequence, was encoded at a resolution of 1920×1080 pixels, 30 fps, with a total length of 16 seconds (500 YUV420p frames).

The simulation environment included 21 vehicles moving within a realistic New York City urban layout. A RandomDirection mobility model was used to imitate unpredictable vehicular movement patterns, with vehicles periodically choosing new routes and speeds, which increases route variability and link intermittency. The simulation ran for 1000 seconds, providing enough time to observe steady-state transport behavior and transient effects caused by mobility. Wireless propagation effects were modeled using the LogDistance path-loss model, capturing the attenuation trends typical of urban macro-cell deployments.

The complete setup was configured within a virtual machine environment. Mininet-WiFi was used to simulate network behavior, while video encoding and decoding were handled by the HEVC Test Model (HM v16.24) and FFmpeg (libx265). Vehicle mobility was precisely modeled using SUMO, importing real traffic maps from OpenStreetMap (OSM). To mimic real network conditions, NetEm was used to introduce parameters such as delay, jitter, and packet loss. The system supported video streaming from a sender vehicle (the client) to a receiver vehicle (the server).

This fully integrated simulation environment provides a robust platform for conducting reproducible and detailed

experiments to assess protocol performance in realistic vehicular communication scenarios. Tests were performed over a 5G network with 1000 Mbps bandwidth, 1 ms latency, and a 1% packet loss rate.

IV. SIMULATION RESULTS AND DISCUSSION

This section presents the comparative results of TCP and QUIC in terms of connection establishment performance. Two key metrics are analyzed: handshake duration and first byte latency (FBL). The experiments were conducted under varying round-trip times (RTT) and packet loss rates to emulate both stable and challenging network environments.

Analyzing both the handshake duration and the first byte latency (FBL) provides a holistic view of how TCP and QUIC behave during the connection establishment phase. In the case of TCP, a connection is initiated through the traditional three-way handshake (SYN, SYN-ACK, ACK), which ensures reliability but does not inherently provide confidentiality or integrity. To achieve secure communication, TCP must be coupled with a separate TLS 1.3 handshake, producing the commonly referred to TLS-over-TCP variant. This additional negotiation introduces extra round trips, meaning that at least two

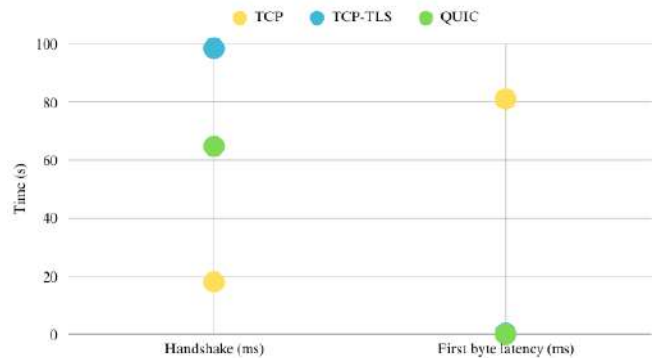


Fig.3. TCP, TCP-TLS, and QUIC results.

RTTs are required before encrypted application data can be exchanged. While this layering approach increases latency, it provides a fairer baseline when comparing against QUIC, since QUIC natively integrates transport and cryptographic setup within a single procedure.

QUIC, by contrast, incorporates TLS 1.3 directly into the transport handshake, which typically allows a secure connection to be established within one RTT. Furthermore, when session state from a previous connection is available, QUIC can employ its 0-RTT resumption mechanism, enabling application data to be transmitted immediately, without waiting for handshake completion. This design significantly reduces startup delay, particularly in mobile or intermittent connectivity scenarios, where frequent reconnections are common. Thus, FBL in QUIC is tightly coupled to its streamlined handshake, making

it more efficient in delivering the first application byte to the user.

The experimental evaluation further illustrates these differences. A cloud point comparison reveals distinct trade-offs across the evaluated protocols. According to the obtained results illustrated at Fig.3, TCP alone achieves the shortest handshake duration (≈ 18 ms) because its initial exchange is lightweight and does not include security mechanisms. However, this comes at the expense of unencrypted communication, which is impractical for modern Internet applications. When enhanced with TLS 1.3, the TLS-over-TCP variant requires ≈ 98.5 ms to complete its two-step negotiation (transport + security), resulting in the longest handshake time among the tested protocols. QUIC, which consolidates transport and TLS 1.3 into a single step, records a handshake duration of ≈ 64.8 ms—longer than plain TCP but notably faster than TLS-over-TCP. The additional delay observed in QUIC is primarily attributed to the computational overhead of user-space cryptographic processing, yet this cost is offset by its efficiency in subsequent phases.

The advantage of QUIC becomes evident when examining first byte latency. Once the connection is active, QUIC delivers application data almost immediately, achieving an average FBL of ≈ 102 ns. In comparison, plain TCP exhibits an FBL of ≈ 81 ms, while TLS-over-TCP performs even worse, with ≈ 315 ns. These results emphasize QUIC's core design objective: while it may require slightly more effort to initialize, it drastically reduces the application-perceived latency by accelerating secure data delivery post-handshake. This balance between setup complexity and rapid data availability highlights QUIC's suitability for latency-sensitive applications such as video streaming, vehicular communications, and telemedicine services, where minimizing startup delay is essential.

Although the primary focus of this study is comparing handshake duration and First-Byte Latency (FBL) between TCP, TLS-over-TCP, and QUIC, these metrics directly impact the performance of real-time video streaming applications. Quick session setup is crucial in vehicular networks, where connections between the vehicle and roadside units are brief due to high mobility. In such scenarios, frequent reconnections are typical, and long handshake delays cause repeated interruptions in video playback.

TCP requires a 3-way handshake followed by a separate TLS negotiation, which increases the time before the first media segment can be delivered. In contrast, QUIC combines transport and cryptographic setup into a single 1-RTT or even 0-RTT operation, significantly lowering FBL. This reduction doesn't affect the throughput of subsequent video transmission but shortens startup delays and rebuffering events when vehicles reconnect to a new access point.

To measure the practical impact of these delays, we calculate the percentage of video session time lost during connection setup. For a typical 3–4 second

video segment used in DASH and CMAF-based streaming, a handshake delay of 99 ms (TLS-TCP) accounts for 1% overhead, whereas QUIC's 65 ms cuts this to 0.6%. Although small numerically, these differences add up in highly dynamic topologies where vehicles may initiate dozens of short connections within a minute. In such environments, QUIC's faster session setup provides a noticeable reduction in playback stalls, especially during handovers.

While full video playback experiments are outside the scope of this work, the latency and FBL results shown here demonstrate why QUIC is a more suitable foundation for future vehicular video streaming research.

V. CONCLUSION

This paper presented a comparative evaluation of TCP and QUIC during connection establishment, focusing on two fundamental metrics: handshake duration and first byte latency (FBL). The results demonstrate that while TCP benefits from a lightweight three-way handshake, it requires an additional TLS negotiation to ensure secure communication, which introduces extra delay. QUIC, by contrast, integrates transport and cryptographic handshakes into a single step and supports 0-RTT session resumption, resulting in a more efficient and secure startup process.

Our findings highlight that QUIC consistently reduces the latency perceived by applications when compared to both plain TCP and TLS-over-TCP. Although QUIC incurs slightly higher computational overhead due to its full encryption and user-space implementation, this trade-off is outweighed by its ability to deliver application data more quickly after connection setup. These characteristics make QUIC particularly well-suited for latency-sensitive and mobile environments, including vehicular communications, real-time video streaming, and mobile health services.

Future research directions include extending this analysis to multipath QUIC (MP-QUIC) for robustness in heterogeneous networks, evaluating QUIC in 5G/6G and satellite-based systems, and exploring cross-layer optimizations to further enhance its performance under high mobility and loss conditions.

REFERENCES

- [1] Cisco, "Cisco Annual Internet Report (2018–2023)," Cisco White Paper, 2020.
- [2] J. Postel, "Transmission Control Protocol," RFC 793, IETF, 1981.
- [3] H. Salah, F. H. P. Fitzek, and M. Reisslein, "Video streaming in vehicular networks: Challenges and solutions," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3186–3215, 2019.
- [4] J. Iyengar and M. Thomson, "QUIC: A UDP-Based Multiplexed and Secure Transport," RFC 9000, IETF, 2021.
- [5] A. Langley et al., "The QUIC transport protocol: Design and Internet-scale deployment," in *Proc. ACM SIGCOMM*, pp.

183–196,

2017.

3. Communications and Network, 14, 12–22.
doi: 10.4236/cn.2022.141002.

- [6] S. Naik, M. Chowdhury, and S. Misra, "A survey on QUIC protocol: Design, applications and performance," *Computer Networks*, vol. 197, p. 108317, 2021.
- [7] A. B. Moubayed, H. Otrouk, H. Barada, and J. Bentahar, "A survey on the QUIC protocol: Fundamentals, applications and challenges," *IEEE Access*, vol. 9, pp. 124148–124171, 2021.
- [8] Z. Yu, C. Xu, and X. Fu, "Understanding performance of QUIC and TCP in cellular networks," in *Proc. IEEE INFOCOM*, pp. 1–9, 2020.
- [9] T. Dierks and E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.2," RFC 5246, IETF, 2008.
- [10] E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.3," RFC 8446, IETF, 2018.
- [11] G. Fairhurst and M. Welzl, *Transport Services for Low Latency*, IEEE Communications Standards Magazine, vol. 2, no. 2, pp. 20–26, 2018.
- [12] C. Huitema and E. Rescorla, "The QUIC TLS Handshake Protocol," RFC 9001, IETF, 2021.
- [13] T. Shreedhar, R. Panda, S. Podanev and V. Bajpai, "Evaluating QUIC Performance Over Web, Cloud Storage, and Video Workloads," in *IEEE Transactions on Network and Service Management*, vol. 19, no. 2, pp. 1366–1381, June 2022, doi: 10.1109/TNSM.2021.3134562.
- [14] Kyratzis, A. and Cottis, P. (2022) QUIC vs TCP: A Performance Evaluation over LTE with NS-3. Communications and Network, 14, 12–22. doi: 10.4236/cn.2022.141002.
- [15] Soni, M., Rajput, B.S. (2021). Security and Performance Evaluations of QUIC Protocol. In: Kotecha, K., Piuri, V., Shah, H., Patel, R. (eds) Data Science and Intelligent Applications. Lecture Notes on Data Engineering and Communications Technologies, vol 52. Springer, Singapore. https://doi.org/10.1007/978-981-15-4474-3_51
- [16] Estupiñan-Cuesta, E.P., Martínez, J.C. and Roldán-Pineda, W.A., 2025. Evaluation of Throughput in Live Video Streaming: TCP vs. QUIC as a Proof of Concept. *Visión electrónica*, 19(1).
- [17] S. Jeddou, L. Diez, N. Abdellah, A. Baina and R. A. Calvo, "On the Performance of Transport Protocols Over mmWave Links: Empirical Comparison of TCP and QUIC," in *IEEE Open Journal of the Communications Society*, vol. 4, pp. 2596–2608, 2023, doi: 10.1109/OJCOMS.2023.3325764.
- [18] Singh Sidhu, J. and Bentaleb, A., 2025. Video Streaming Over QUIC: A Comprehensive Study. *arXiv e-prints*, pp.arXiv-2505.
- [19] Chu, TT., Labiod, M.A., Augustin, B. et al. Empirical Evaluation of QUIC-Based Software-Defined Service Migration in Multi-access Edge Computing Over 5G Networks. *J Netw Syst Manage* 33, 29 (2025). <https://doi.org/10.1007/s10922-025-09909-0>.
- [20] Simulation of urban mobility, <https://www.eclipse.org/sumo/>.
- [21] C. R. E. R. Ramon dos Reis Fontes, Mininet-wifi, <https://mininet-wifi.github.io/> (March 2022).
- [22] Mininet emulator, <http://mininet.org/>.

Experimental Validation of a Dual MEMS Microphone Array for Cardioid Directivity

Fatma Zahra Mahdjoub
Electrical Engineering Department
University Echahid Cheikh Larbi Tebessi
Tébessa, Algeria
fatmazahra.mahdjoub@univ-tebessa.dz

Mahmoud maamri
Electrical Engineering Departement
university Echahid Cheikh Larbi Tebessi
Tebessa , Algeria
m.maamri@univ-tebessa.dz

Azzedine gattel
Electrical Engeeneering Departement
University Echahid Cheikh Larbi Tebessi
Tebessa , Algeria
a.gattal@univ-tebessa.dz

Abstract— This paper presents the design and real-time experimental validation of a compact cardioid directivity pattern using two MEMS microphones spaced 3.4 cm apart. A delay-and-sum beamforming algorithm is implemented on a dSPACE DSP platform, where one microphone signal is digitally delayed and polarity-inverted to synthesize a first-order end-fire cardioid response. The experimental evaluation uses a 1 kHz tone emitted at angles from 0° to 360°. The resulting directivity exhibits a front-to-back ratio of approximately 11 dB and a null depth of −12 dB in the rear direction, confirming alignment with theoretical expectations. The proposed approach demonstrates a low-cost and compact solution for directional audio capture, suitable for embedded and real-time speech acquisition systems.

Keywords—Microphone arrays, MEMS microphones, delay-and-sum beamforming, cardioid directivity, real-time DSP.

I. INTRODUCTION

MEMS (Micro-Electro-Mechanical Systems) microphones [1] have revolutionized the audio capture industry thanks to their low cost, miniaturized size, mechanical robustness, and compatibility with integrated circuits. Unlike traditional electret microphones, MEMS devices offer better thermal stability, uniform sensitivity, and a flatter frequency response, while being easily integrated into digital architectures such as Digital Signal Processors (DSPs).

These microphones are primarily based on the capacitive principle, where the vibration of a thin silicon diaphragm causes a change in capacitance between a movable plate and a fixed back plate, which is then converted into an electrical signal. This mechanism, combined with the precision of microfabrication, enables the production of both omnidirectional and directional microphones, depending on the geometry and internal structure of the sensor.[2]

Directional microphones, particularly those with cardioid directivity, play a key role in intelligent audio capture systems, as they allow selective enhancement of sounds from a desired direction while attenuating background noise. This directivity pattern can be achieved by combining signals from multiple microphones and exploiting the phase and amplitude differences between the received signals.

In compact devices, a two-MEMS-microphone configuration represents an efficient and low-cost solution to achieve a cardioid response. By adjusting the time delay and polarity

between two closely spaced microphones, it is possible to synthesize a polar pattern with a maximum sensitivity at the front (0°) and a minimum at the rear (180°).

In this work, we present the experimental implementation of a cardioid directivity pattern using two MEMS microphones spaced 3.4 cm apart and connected to a dSPACE. One microphone is used as the direct reference, while the second signal is phase-inverted and delayed digitally to reproduce the typical spatial combination of an end-fire cardioid. The system was experimentally tested for various sound incidence angles (0°–360°), allowing the directional response to be validated.

This study demonstrates the practical feasibility of synthesizing a simple and reconfigurable cardioid pattern using standard MEMS components and real-time digital processing, paving the way for applications in speech capture, compact beamforming, and speech quality enhancement in noisy environments.

The main contributions of this paper are summarized as follows:

- An experimental validation of a compact two-MEMS end-fire array achieving cardioid directivity in real time using a DSP platform.
- A complete implementation framework, including system architecture, signal processing chain, and real-time measurement protocol.
- A quantitative comparison between experimental and theoretical directivity patterns, confirming the practical feasibility of the proposed approach for low-cost directional audio capture.

II. MICROPHONES ARRAYS

A microphone array, or microphone network, is a set of microphones arranged in a specific geometry to capture sound, its have become a fundamental component in modern consumer electronics, enabling a wide range of audio-related functionalities. They are commonly embedded in smartphones, personal computers, home assistants, smart TVs, and teleconferencing systems, where they serve as the front-end of advanced audio processing pipelines. By exploiting spatial diversity and signal processing algorithms, microphone arrays

can focus on a desired sound source, such as a speaker's voice, while suppressing unwanted interferences like ambient noise or room reverberation. Their capabilities extend beyond speech capture, encompassing echo and reverberation suppression [3], automatic speech recognition [4], and active noise control (ANC) [5]. Furthermore, they are employed in multi-channel audio acquisition [6], beamforming systems [7], acoustic field modeling and simulation [8], and even in structural health monitoring applications.

Microphone arrays utilize their spatial distribution to sample sounds propagating from multiple locations, enabling the deduction of information based on the differences observed in these measurements. By analyzing the time of arrival of sound at each sensor, they can gather data about their own spatial positions as well as about other sound sources in the environment. The sound measurements captured at a given point in space typically contain a mixture of all sounds passing through that location at that moment. By comparing the arrival times and other acoustic properties at several distinct positions, it is possible to extract information about the locations and trajectories of sound sources. This approach enhances the understanding of the acoustic environment and facilitates applications such as audio signal processing and sound localization. Microphones can be arranged in different configurations.[9]

A. Broadside configuration :

A broadside microphone array is composed of microphones arranged linearly and oriented perpendicular to the direction of the incoming sound waves, as illustrated in Figure 2, where d denotes the spacing between the two elements of the array. This configuration is optimized to capture sounds arriving from the broadside direction, that is, perpendicular to the array axis. Since the received signals are typically summed directly, the broadside array requires minimal signal processing. However, its directional selectivity is limited, as it mainly attenuates sounds arriving from the sides. Moreover, because of its axial symmetry, the array produces almost identical responses to sounds arriving from the front and rear, which reduces its ability to discriminate between opposite directions [10].

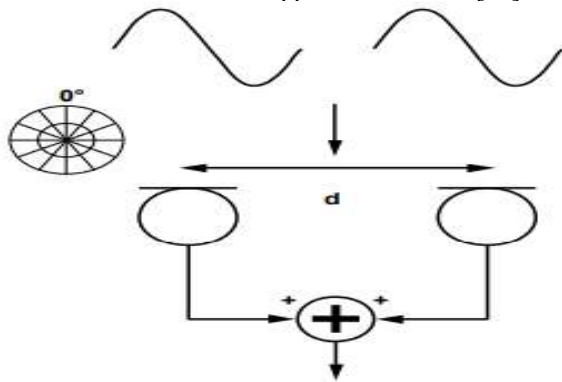


Fig.1 Microphone Broadside Array.

In a two-microphone broadside array, the directional response presents two nulls located at 90° and 270° relative to the array axis. The extent of attenuation at these angles depends on the

frequency of the incoming sound, with optimal cancellation occurring when the microphone spacing equals half the wavelength of the signal. For instance, when the spacing between microphones is 75 mm, a complete null theoretically occurs near 2.3 kHz, as determined by $f = 343 / (2 \times 0.075) \approx 2.3 \text{ kHz}$. Beyond this frequency, spatial aliasing appears, resulting in additional nulls at other angles within the polar pattern. Consequently, the array's side rejection capability decreases as frequency increases. As shown in Figure 3, the 3 kHz response (illustrated by the light blue curve) exemplifies this aliasing effect [10].

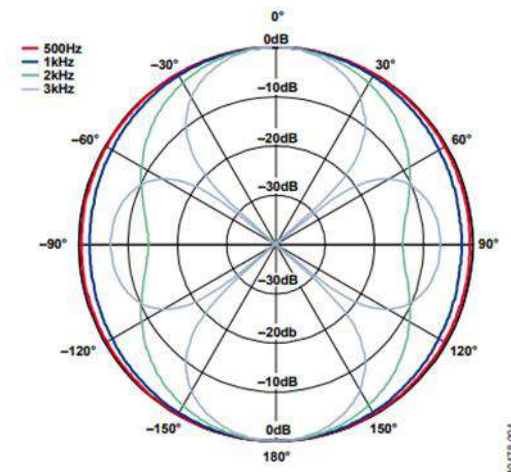


Fig.2 Response of a 2-Microphone Broadside Array with 75 mm Spacing.

B. End-fire configuration:

In an end-fire (EF) array employing the delay-and-sum (DAS) algorithm, the signals from each microphone are phase-aligned so that their contributions add constructively along the array axis. This results in maximum sensitivity in the direction of the array, making it particularly effective for capturing a target source positioned on-axis. The primary advantage of this configuration lies in its enhanced rejection of off-axis noise, which surpasses that of the broadside arrangement. Consequently, the EF array can achieve greater directional selectivity while maintaining a more compact geometry. However, for optimal performance, the sound source must be aligned with the array axis [11].

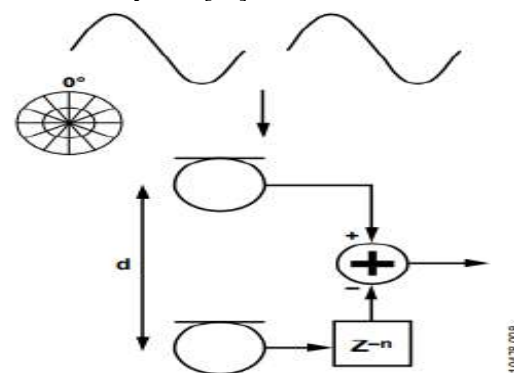


Fig.3 Microphone end-fire Array.

When the microphone spacing and time delay are appropriately selected, a delay-and-sum beamformer can produce a cardioid directivity pattern for frequencies below the spatial aliasing limit, as illustrated in Figure 5. This cardioid response features maximum sensitivity along the forward direction of the array and theoretical nulls at 180°, where signals arriving from the rear are completely cancelled. In a first-order (two-microphone) delay-and-sum configuration, sounds arriving from the sides experience approximately 6 dB of attenuation, yielding a directional response that closely resembles the shape of a heart-shaped cardioid [10].

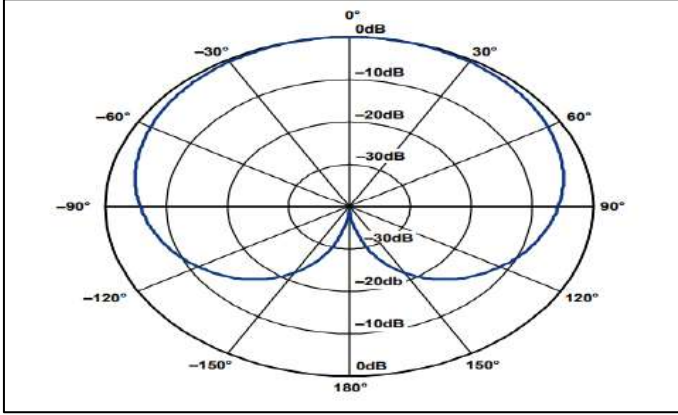


Fig.4 Response of a 2-Microphone Endfire Cardioid Beamformer.

III. PROPOSED GEOMETRY AND EXPERIMENTAL RESULT

A. System architecture:

The overall architecture of the proposed system is illustrated in Figure 5a. The setup consists of several key stages: the speech source, acoustic sensing using a MEMS microphone array, analog-to-digital conversion, digital signal processing (beamforming), digital-to-analog conversion, and the final output stage.

A speech signal emitted from a sound source is first captured by a small array of MEMS microphones. These microphones convert the acoustic pressure variations into corresponding analog electrical signals. The analog signals are then passed through Analog-to-Digital Converters (ADCs), which sample and quantize them into discrete-time digital signals suitable for digital processing.

Once digitized, the signals are processed in the Digital Signal Processor (DSP) using the Delay-and-Sum beamforming algorithm. This algorithm compensates for the time delays between microphones by aligning the signals from a desired direction before summing them. As a result, signals coming from that direction are enhanced, while those arriving from other directions are attenuated. This spatial filtering improves the signal-to-noise ratio and reduces interference.

After beamforming, the enhanced digital signal is converted back into an analog form using a Digital-to-Analog Converter (DAC). Finally, the analog signal is transmitted to a loudspeaker (or another output device), where it is reproduced as an audible sound.

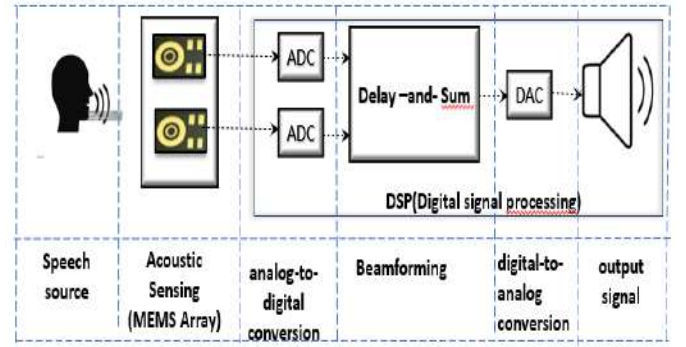


Fig.5 Framework of the system proposed.

B. Theoretical model (geometry) :

Improving microphone directivity is a major challenge for sound acquisition applications in noisy environments. The approach adopted in this study is based on the use of an end-fire microphone array to enhance the capture of signals coming from the desired direction and attenuate those coming from undesired directions.

The system geometry consists of two MEMS microphones spaced by a distance d (see fig.6). When a plane sound wave hits the array at an angle of incidence θ , this results in a time difference of arrival (TDOA) between the microphones, expressed as:

$$\Delta t = \frac{d \cdot \cos(\theta)}{c}$$

where $d=3.4$ cm is the microphone spacing and $c=343$ m/s is the speed of sound. The distance of 3.4 cm was chosen as a compromise between theoretical and practical constraints: it corresponds to approximately one-tenth of the wavelength at 1 kHz (the design frequency of the cardioid), remains below the spatial aliasing limit up to 5 kHz, fits the compact form factor of embedded MEMS devices, and yields an acoustic delay of ~ 5 samples at 48 kHz, which can be accurately compensated by the DSP without requiring fractional-delay filtering.

The beamformer output is computed according to:

$$y(t) = x_1(t) - x_2(t - \Delta t)$$

where $x_1(t)$ represents the direct microphone signal and $x_2(t - \Delta t)$ is the delayed and inverted signal of the second microphone.

For a sound wave arriving from the front ($\theta = 0^\circ$), the two signals are out of phase and cancel each other, producing a null. Conversely, for a wave arriving from the rear ($\theta = 180^\circ$), the signals add constructively, producing maximum output. This behavior yields the cardioid directivity pattern characteristic of first-order end-fire arrays.

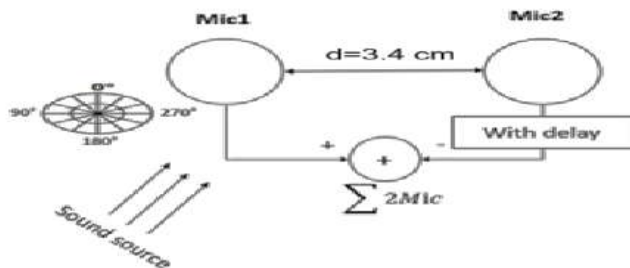


Fig.6 the geometry proposed.

C. Experimental Setup

The experimental setup, shown in Fig. 7, was assembled in a standard university laboratory with moderate reverberation and background noise levels. A pure 1 kHz sine wave was generated using a loudspeaker placed at a fixed distance of 1 meter from the microphone array. The sound source was rotated around the array from 0° to 360° in steps of 30° using a circular trajectory in the horizontal plane.

At each angular position, the beamformed signal amplitude was measured using the dSPACE ControlDesk interface and recorded for analysis. The sampling frequency was set to 48 kHz, with 16-bit resolution. The delay and polarity inversion were implemented digitally inside the DSP in real-time.

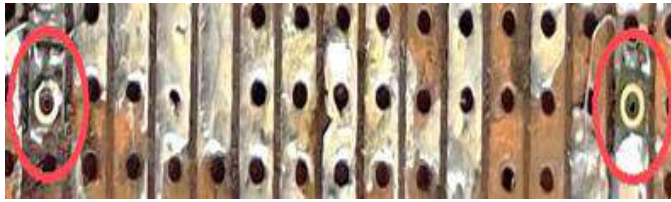


Fig.7 the setup of the geometry proposed .

D. Results and Discussion

The measured amplitudes at each angle were normalized with respect to the maximum value and plotted to obtain the spatial response shown in Fig. 10. The resulting polar pattern exhibits a clear main lobe at 180° (Fig. 8) and a pronounced null at 0° (Fig. 9), which is consistent with the behavior expected from a first-order delay-and-sum end-fire configuration. The main lobe displays a smooth curvature with a front-to-back ratio of approximately 10–12 dB, in agreement with the theoretical prediction for a spacing of 3.4 cm at 1 kHz. The null depth around 0° reaches approximately -12 dB, indicating proper phase alignment between the direct and delayed channels. Some deviations from the ideal cardioid shape are observed at intermediate angles, mainly due to reflections within the non-anechoic laboratory environment, moderate reverberation, and small amplitude or phase mismatches between the two MEMS sensors. These factors introduce secondary propagation paths that slightly distort the measured response. Additionally, as delay-and-sum beamforming is inherently frequency-dependent, the measured pattern matches the theoretical model most closely around the design frequency of 1 kHz; minor widening of the main lobe and side-lobe irregularities remain consistent with the 3.4-cm spacing. Overall, despite these non-ideal conditions, the obtained polar response confirms the

validity of the proposed approach and demonstrates that directional audio capture can be effectively achieved using only two MEMS microphones and real-time DSP processing.

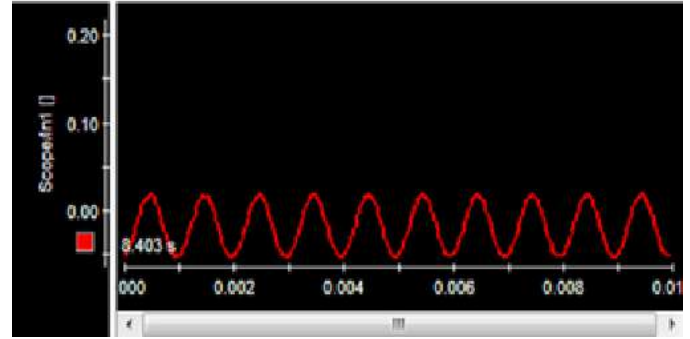


Fig.8 Test result with a 1KHz sound source at a 180° angle.

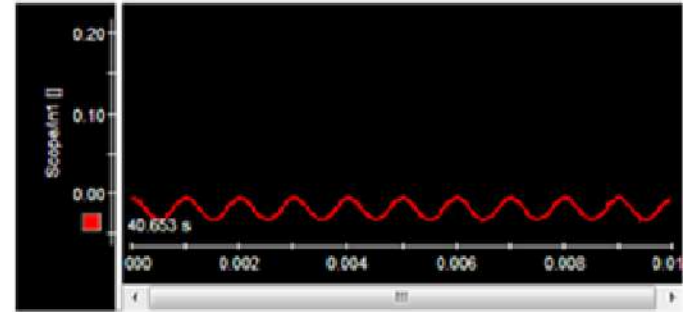


Fig.9 Test result with a 1KHz sound source at a 0° angle.

The overall directivity pattern obtained from these measurements is presented in Figure 10, illustrating the spatial response of the proposed delay-and-sum beamforming configuration.

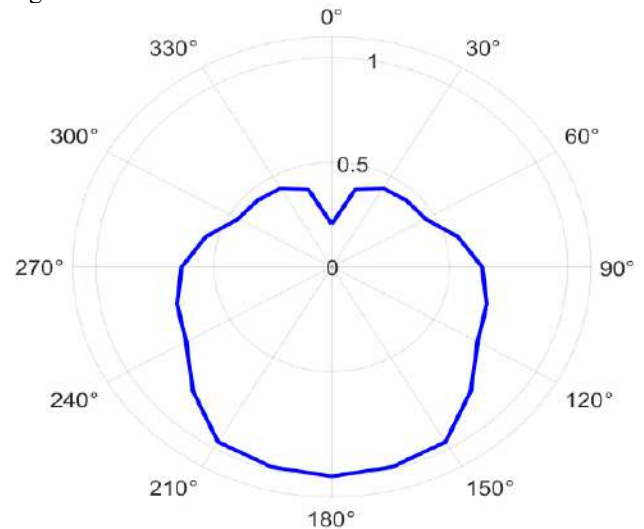


Fig.10 Directivity diagram.

E. Real-Time Latency Analysis

Given the system's requirement for real-time operation, we evaluated the total processing latency to confirm it remains insignificant compared to the period of the acoustic signals. This total latency is the sum of three primary components: the analog-to-digital conversion delay, the digital processing time within the DSP, and the digital-to-analog conversion delay.

Based on measurements taken directly from the dSPACE board, the complete cycle of signal acquisition, processing, and output introduces a delay of approximately 0.1 ms. With our test signal at a frequency of 1 kHz (corresponding to a 1 ms period), this total latency accounts for less than 10% of a single waveform cycle. This margin is sufficiently small to be considered imperceptible in real-time audio applications.

These results confirm that the system executes the delay-and-sum beamforming and polarity inversion algorithms in true real time, introducing only minimal phase distortion. Such low latency is well-suited for demanding applications like live speech reinforcement, telecommunications, and interactive acoustic systems where a rapid directional response is critical.

VI. CONCLUSION

This study presented the design and experimental validation of a two-MEMS-microphone array enabling the realization of a cardioid-type directivity pattern. Using an “end-fire” configuration and a delay-and-sum beamforming algorithm, it was possible to synthesize a directional pattern exhibiting maximum sensitivity at the front (0°) and significant attenuation at the rear (180°).

The experimental results, obtained with a pure 1000 Hz sound signal, confirm the feasibility of this approach and demonstrate the system’s effectiveness in enhancing signals arriving from a preferred direction while attenuating ambient noise. Although based on a simple and compact configuration, this implementation opens promising perspectives for applications requiring directional audio capture, such as telecommunication systems, voice assistants, or speech quality enhancement in noisy environments or live conferences.

Future work could include extending this configuration to larger microphone arrays, optimizing real-time processing for broadband signals, or integrating machine learning techniques for dynamic adaptation of the directivity pattern.

ACKNOWLEDGMENT

The authors are grateful to the anonymous referees for their valuable and constructive comments. This research was conducted as part of the PRFU project in the Department of Electronics and Communications at the University of Echahid Cheikh Larbi Tebessi, Tebessa. The authors also extend they're

thanks to the staff of the LAMIS and L3S laboratories for their insightful comments and suggestions.

REFERENCES

- [1] M. A. Shah, I. A. Shah, D.-G. Lee, and S. Hur, "Design Approaches of MEMS Microphones for Enhanced Performance," *Journal of Sensors*, vol. 2019, Article ID 9294528, 26 pages, March 6, 2019. *Journal of Sensors*, vol. 2019, Article ID 9294528, 26 pages, March 6, 2019.
- [2] M. Fuedner, "Microphones," in *Handbook of Silicon Based MEMS Materials and Technologies (Third Edition)*, Eds.: Micro and Nano Technologies, 2020, pp. 937-948. Eds.: Micro and Nano Technologies, 2020, pp. 937-948.
- [3] J. Hong, "Stereophonic Acoustic Echo Suppression for Speech Interfaces for Intelligent TV Applications," *IEEE Trans. Consum. Electron.*, vol. 64, 2018.
- [4] T. Kawase, M. Okamoto, T. Fukutomi, and Y. Takahashi, "Speech Enhancement Parameter Adjustment to Maximize Accuracy of Automatic Speech Recognition," *IEEE Trans. Consum. Electron.*, vol. 66, no. 2, pp. 125–133, Feb. 2020, doi: 10.1109/TCE.2020.2986003.
- [5] V. Patel, J. Cheer, and S. Fontana, "Design and Implementation of an Active Noise Control Headphone With Directional Hear-Through Capability," *IEEE Trans. Consum. Electron.*, vol. 66, no. 1, pp. 32–40, Feb. 2020, doi: 10.1109/TCE.2019.2956634.
- [6] S. M. Kim, C. J. Chun, and H. K. Kim, "Multi-channel audio recording based on superdirective beamforming for portable multimedia recording devices," *IEEE Trans. Consum. Electron.*, vol. 60, no. 3, pp. 429–435, Feb. 2014, doi: 10.1109/TCE.2014.6937327.
- [7] D. Pinardi, L. Ebri, C. Belicchi, A. Farina, and M. Binelli, "Direction Specific Analysis of Psychoacoustics Parameters inside Car Cockpit: A Novel Tool for NVH and Sound Quality," *SAE Tech. Papers*, 2020, doi: 10.4271/2020-01-1547.
- [8] A. Toscani et al., "Low-cost Structural Health Monitoring System for Smart Buildings," in *2022 Second Int. Conf. Sustain Mobility Appl., Renewables Technol. (SMART)*, IEEE, Nov. 2022, pp. 1–7. doi: 10.1109/SMART55236.2022.9990379.
- [9] Lewis, Jerad. AN-1165 Application Note, Op Amps for MEMS Microphone Preamp Circuits. Analog Devices, Inc., 2013.
- [10] J. Lewis, "Microphone Array Beamforming," Analog Devices, Inc., Norwood, MA, USA, [En ligne]. Disponible : www.analog.com. 2012. Abuse and Mental Health Services Administration, Office of Applied Studies, August, 2013, DOI:10.3886/ICPSR30122.v2
- [11] M. B. Stolbov and K. Chong T., "Comparative Analysis of Spatial Characteristics of Two-Element Microphone Arrays," *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, vol. 20, no. 1, pp. 147–154, 2020. doi: 10.17586/2226-1494-2020-20-1-147-154.

Fault Isolation and Power Quality Enhancement Using Solid-State Transformers

Nassim Zemirline

Department of Electrical Engineering
University Yahia Fares of Medea
Medea, Algeria
ORCID 0000-0003-2039-2518

Nadir Kabeche

Department of Electrical Engineering
University Yahia Fares of Medea
Medea, Algeria
ORCID 0000-0002-3505-8873

Samir Moulahoum

Department of Electrical Engineering
University Yahia Fares of Medea
Medea, Algeria
ORCID 0000-0002-7482-0621

Sofiane Khechafi

Department of Electrical Engineering
University Yahia Fares of Medea
Medea, Algeria
ORCID 0009-0009-3988-9330

Riadh Euldji

Department of informatic
DIF-FS-UMBB
Boumerdes, Algeria
ORCID 0009-0001-9033-3495

Abstract— Conventional power grids, based on low-frequency transformers (LFTs), face several limitations such as large size, weight, vulnerability to harmonics, and lack of fault isolation. To address these challenges, Solid-State Transformers (SSTs) have emerged as a promising alternative, offering compactness, bidirectional power flow, renewable energy integration, and enhanced grid support. This paper investigates the capability of a three-stage SST architecture, comprising an MMC-based AC–DC stage, a Dual Active Bridge (DAB) converter in the DC–DC stage, and a two-level inverter for the DC–AC stage, to mitigate a wide range of power quality disturbances. Different grid faults, including voltage fluctuations, sags/swells, interruptions, transients, frequency variations, DC offsets, unbalances, and harmonics, are simulated in MATLAB/Simulink. Results show that the adopted SST-based power grid effectively isolates these disturbances, ensuring stable and high-quality load-side operation. The findings highlight the advantages of SST-based power grids in improving reliability, efficiency, and compatibility with modern energy demands.

Keywords— Conventional power grid, Solid-state transformer (SST), faults isolation, Disturbances, Three-stage topology.

I. INTRODUCTION

The demand for reliable, and intelligent electrical energy systems has grown significantly in recent years due to the increasing integration of renewable energy sources and the evolution of smart grids. These new requirements have exposed the limitations of conventional power grids, which have historically relied on low-frequency transformers (LFTs) as the backbone for generation, transmission, and distribution. While LFTs are valued for their simplicity and robustness, they suffer from key drawbacks such as large size, heavy weight, and the absence of effective isolation between the source and load. Moreover, their dependence on mineral oil for insulation introduces environmental concerns that further limit their sustainability [1], [2].

Another critical issue is the inability of conventional grids to deal effectively with power quality disturbances. Electrical networks are routinely subjected to a wide range of perturbations, including voltage fluctuations, sags, frequency deviations, and harmonics. These disturbances directly affect the performance and lifespan of connected devices and may propagate across the entire system, leading to operational instability and costly equipment failures. As the penetration of sensitive loads, renewable sources, and power electronics continues to rise, the shortcomings of traditional transformer-based grids become even more apparent [3].

Solid-State Transformers (SSTs) have emerged as a promising alternative to address these challenges [4]. Unlike their low-frequency counterparts, SSTs employ high-frequency power conversion stages, resulting in smaller size, lighter weight, and higher efficiency. More importantly, they provide advanced functionalities such as bidirectional power flow, renewable energy integration, and enhanced fault isolation [5], [6]. Their modular, multi-stage architecture, typically consisting of an AC–DC rectifier, a DC–DC high-frequency isolated converter, and a DC–AC inverter, offers superior controllability and adaptability to different operating conditions. These features make SSTs an enabler for the modernization of power grids.

In this paper, a three-stage SST architecture is proposed and analysed to demonstrate its ability to improve power quality and isolate disturbances. The studied configuration employs a Modular Multilevel Converter (MMC) for the AC–DC stage, a Dual Active Bridge (DAB) converter for the DC–DC stage, and a conventional two-level inverter for the DC–AC stage. The system is modelled and implemented in MATLAB/Simulink, where a wide range of disturbances are deliberately introduced on the source side, including voltage fluctuations, sags, swells, interruptions, transients, frequency variations, DC offsets, unbalances, and harmonics. By comparing source-side and load-side voltages, the capability of the SST to suppress or eliminate disturbances is evaluated.

II. CONVENTIONAL POWER GRID

A conventional power grid is composed of three fundamental stages [7]: generation, transmission, and distribution. Fig. 1 illustrates a typical configuration in which the load is connected to the low-voltage side of the transformer. In such grids, the transmission stage mainly consists of long, unidirectional high-voltage lines that link the generation and distribution systems through step-up and step-down transformers, enabling the transfer of electrical energy [8].

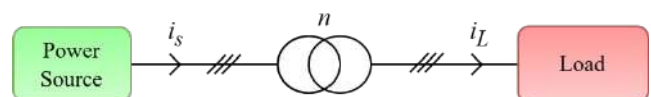


Fig. 1. Traditional power network.

From Fig. 1, the source current i_s can be expressed as:

$$i_s = n \cdot i_L$$

()

where n is the transformer turns ratio, and i_L is the load current. This relation implies that the quality of power delivered from the main grid is directly reflected at the load side. Consequently, any disturbance, perturbation, or distortion at the source will propagate without mitigation to the load, and vice versa. This inherent lack of isolation represents a key limitation of traditional power grids.

Although low-frequency transformers have been widely used in these systems due to their cost-effectiveness and proven reliability, they also suffer from several drawbacks. These include their large size and weight, vulnerability to harmonics, associated power losses, and environmental concerns linked to the extensive use of mineral oil [9]. Addressing these drawbacks is essential for ensuring the power grid to meet modern technological and environmental requirements.

III. SST-BASED POWER GRID

SSTs are an advanced alternative to LFTs, providing reduced size and weight, improved power quality, bidirectional power flow, and smart grid support [5], [10]. Their high-frequency operation makes them well suited for space-constrained and critical applications such as railway traction, offshore platforms, and electric vehicle charging.

SSTs are classified into topologies based on power conversion stages and input/output arrangements, with selection depending on factors such as efficiency, cost, and control complexity. The main designs include single-stage, two-stage, and three-stage, each with specific advantages and limitations [11]. The three-stage SST is the most flexible and widely adopted configuration [12]. It includes AC-DC, DC-DC, and DC-AC conversion stages, with dual DC-links providing greater control flexibility. In this work, an MMC is adopted for the AC-DC stage, a DAB converter for the DC-DC stage, and a conventional three-phase two-level inverter for the output stage. The three-stage SST topology adopted in this paper is illustrated in Fig. 2.

Unlike traditional power grids, the multi-stage structure and multiple voltage levels of SSTs make the current in the source and load sides different. This feature not only enables greater flexibility in control but also provides inherent fault isolation, representing a critical advantage over conventional transformer-based networks.

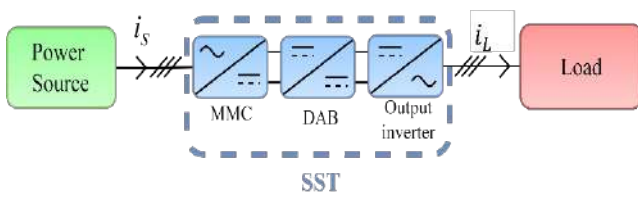


Fig. 2. SST-based power network (three-stage topology).

IV. POWER QUALITY DISTURBANCES

During transmission and distribution, electrical power is exposed to disturbances that affect equipment performance, appearing as amplitude fluctuations, frequency deviations, waveform distortions, or imbalances. They are broadly classified as permanent, such as harmonics from nonlinear loads, or transient, such as interruptions, or spikes from faults or lightning, with transient events sometimes aggravating permanent ones.

According to Electro-Magnetic Compatibility (EMC) principles, disturbances can also be classified as low frequency ($< 9\text{kHz}$) or high-frequency ($> 9\text{kHz}$) [13]. They can also be transmitted through conducted means (via cables) or radiated means (through the air). Various electrical disturbances are explained in this section.

A. Voltage Fluctuations

Voltage fluctuations are short-duration variations in voltage levels, typically lasting a few milliseconds as illustrated in Fig. 3. These fluctuations are often caused by rapid changes in load demand, such switching of heavy industrial equipment [14]. Voltage fluctuations can lead to flicker, a visible variation in the brightness of lighting systems, which is particularly noticeable in sensitive environments like offices or homes.

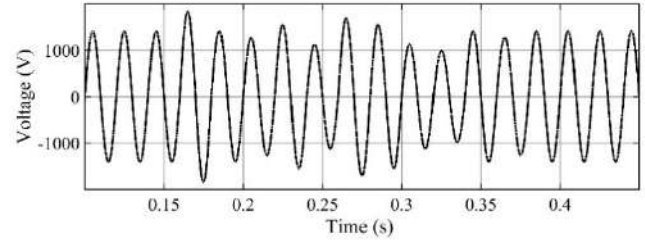


Fig. 3. Voltage fluctuation.

B. Voltage Sags and Swells

A voltage sag is a temporary drop in amplitude to 10–90% of nominal voltage [15], lasting from 10ms to a few seconds. Sags are mainly caused by short circuits, load fluctuations, or natural events like lightning. Conversely, a voltage swell is a temporary rise above nominal voltage, typically lasting from 0.5 to a few seconds, and is often caused by sudden load disconnection or system faults. Both sags and swells can adversely affect sensitive equipment. Fig. 4 illustrates example of voltage sag and swell events.

C. Voltage Interruptions

A voltage interruption, illustrated in Fig. 4, is a more severe case of a voltage sag, with a drop greater than 90%, often due to automation operations in distribution networks that trigger temporary disconnections [15].

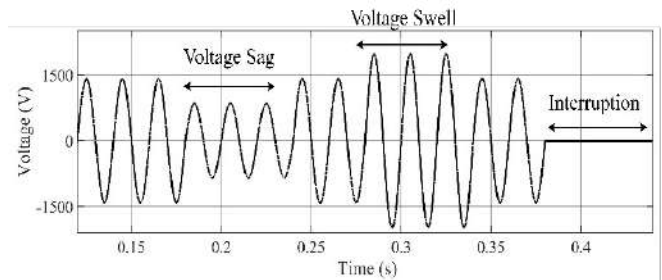


Fig. 4. Voltage sag, swell, and interruption.

D. Transient Over-Voltages

Transient over-voltages, illustrated in Fig. 5, are a brief, high-amplitude voltage spikes lasting less than a second, often caused by switching operations or electrostatic discharges. These can result in immediate damage to electronic components [16].

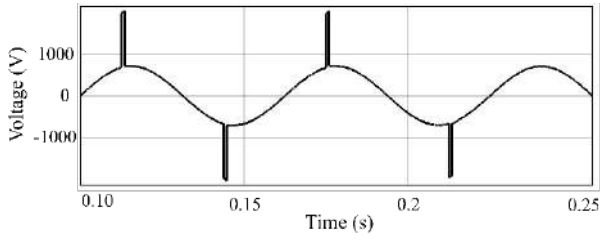


Fig. 5. Voltage transient.

E. Frequency Variation

The frequency of an electrical network is directly linked to the rotational speed of generators responsible for power generation [17]. However, any imbalance between supply and demand causes fluctuations in this speed, leading to frequency variations, as shown in Fig 6. To maintain stability, the frequency is continuously measured and precisely controlled in real time, ensuring it remains within an acceptable range of 0.5Hz around the nominal value [18]. Frequency stability is critical for the proper operation of transformers, motors, and electronic systems.

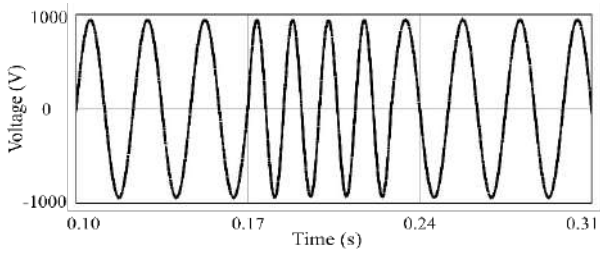


Fig. 6. Frequency variation

F. DC-Offset

A DC offset is a DC component superimposed on the AC waveform, as illustrated in Fig. 7. It can originate from asymmetrical switching, transformer saturation, power electronic converters, or unequal charging/discharging in multi-source systems such as batteries and PV arrays under partial shading. This offset introduces voltage ripple at the DC output of AC/DC converters [19], [20], and in motor drives it causes torque ripples, degrading speed control performance [19].

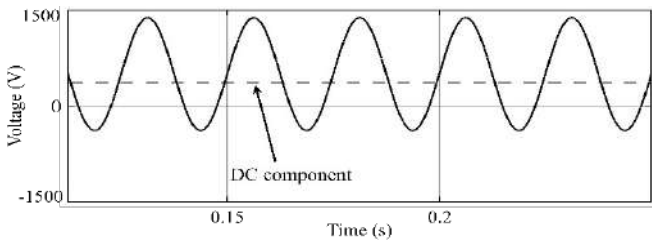


Fig. 7. DC-offset.

G. Grid Imbalances

Grid imbalances can arise from various fault conditions, such as open-circuit or short-circuit faults, unbalanced impedance in transmission and distribution networks, or poorly designed electrical systems. In a three-phase system, when single-phase loads are not evenly distributed, the currents drawn from each phase differ in amplitude and/or phase, as it is illustrated in Fig. 8. Maintaining a well-balanced electrical system is essential for ensuring stability and protecting connected devices [21].

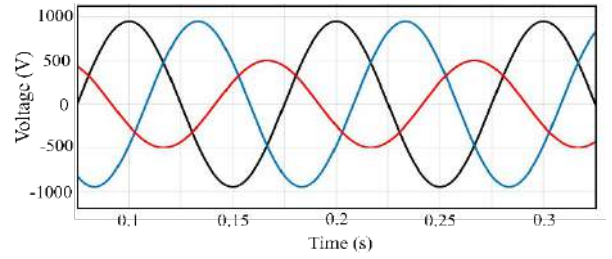


Fig. 8. Voltage Imbalance.

H. Harmonics

Harmonic pollution is one of the most significant sources of electrical disturbances. Harmonics are sinusoidal waveform components at frequencies that are integer multiples of the fundamental frequency. These arise primarily from nonlinear loads such as variable speed drives, rectifiers, and switching power supplies, which draw non-sinusoidal currents from the network [22]. Excessive harmonics cause vibrations, overheating, mechanical fatigue, raising operational and maintenance costs [22].

V. SIMULATION RESULTS

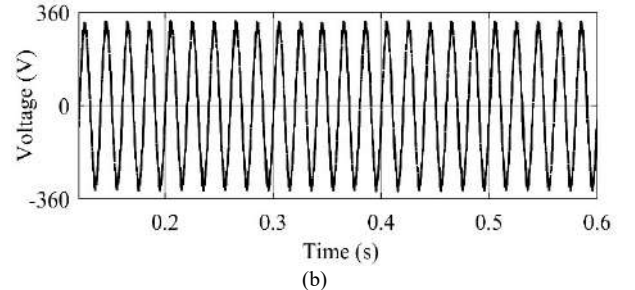
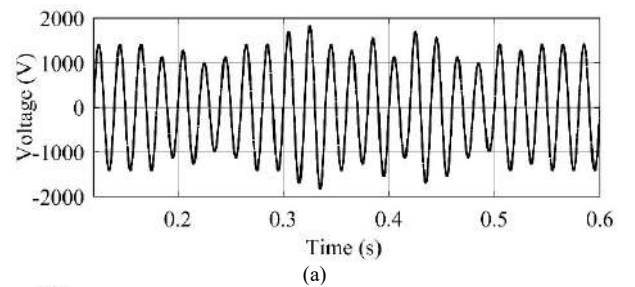
This section presents simulation results that demonstrate the capability of the SST to isolate various grid faults. The SST configuration shown in Fig. 1 is modelled and implemented in MATLAB/Simulink. The corresponding system parameters are listed in Table I. Different grid disturbances are intentionally introduced on the source side, and the source and load voltages are compared to evaluate the performance of the SST.

TABLE I. SYSTEM PARAMETERS

Parameter	Value
Cells per arm in the MMC	4
Cell capacitance C_{cell}	$2000 \mu F$
Input filter	$R_f = 5 m\Omega$, $L_f = 100 \mu H$
Carrier frequency f_c	800 Hz
DAB transformer turns ratio n	0.5

I. Voltage Fluctuation

In the first test, the SST is exposed to a source voltage fluctuation introduced at $t = 0.17\text{s}$ and removed at $t = 0.52\text{s}$.



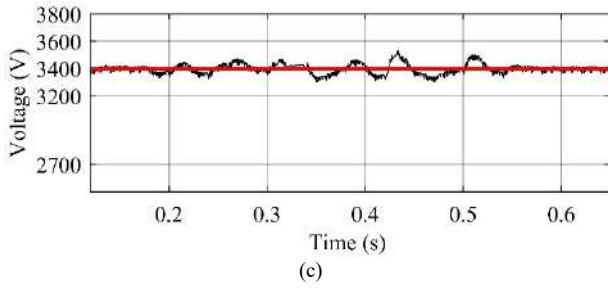


Fig. 9. Voltage fluctuation: (a) grid voltage, (b) load voltage, (c) MMC output DC voltage.

The grid and load voltages are shown in Fig. 9-a and Fig. 9-b, respectively. It can be seen that the fluctuation on the grid side is not transferred to the load side. As shown in Fig. 9-c, the DC output voltage of the MMC undergoes a slight disturbance, but the effect remains small.

J. Voltage Sag/Swell

In the second test, a voltage sag occurs between $t = 0.18s$ and $t = 0.24s$, followed by a voltage swell between $t = 0.32s$ and $t = 0.38s$, as illustrated in Fig. 10-a. The corresponding load voltage is presented in Fig. 10-b, showing that these grid disturbances have no effect on the load side. The MMC output DC voltage is slightly disturbed, as shown in Fig. 10-c, but this perturbation is effectively suppressed by the second stage of the SST.

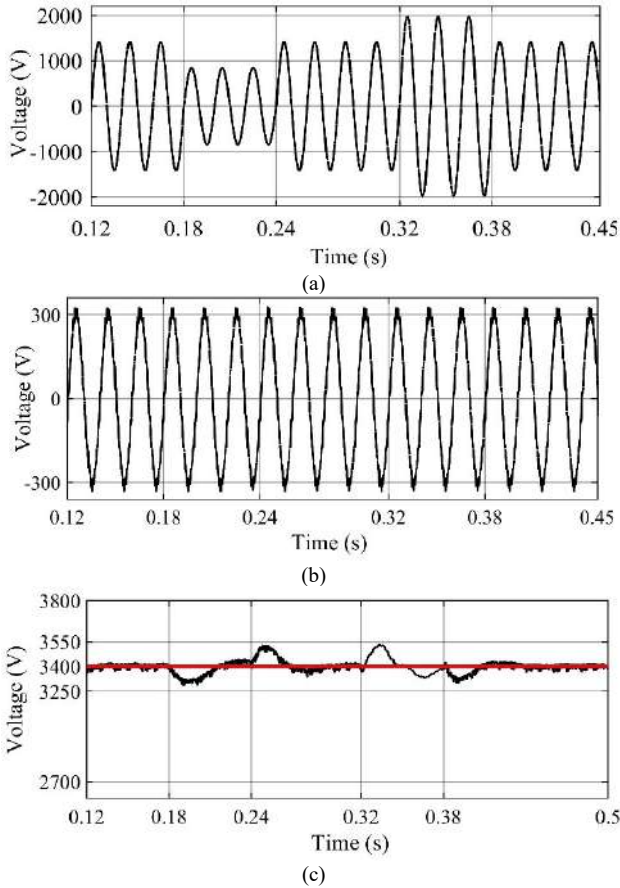


Fig. 10. Voltage sag/swell: (a) grid voltage, (b) load voltage, (c) MMC output DC voltage.

K. Voltage Interruption

In this test, the source voltage is completely set to zero for a short interval ($t = 0.2s$ to $t = 0.22s$), as shown in Fig. 11-a.

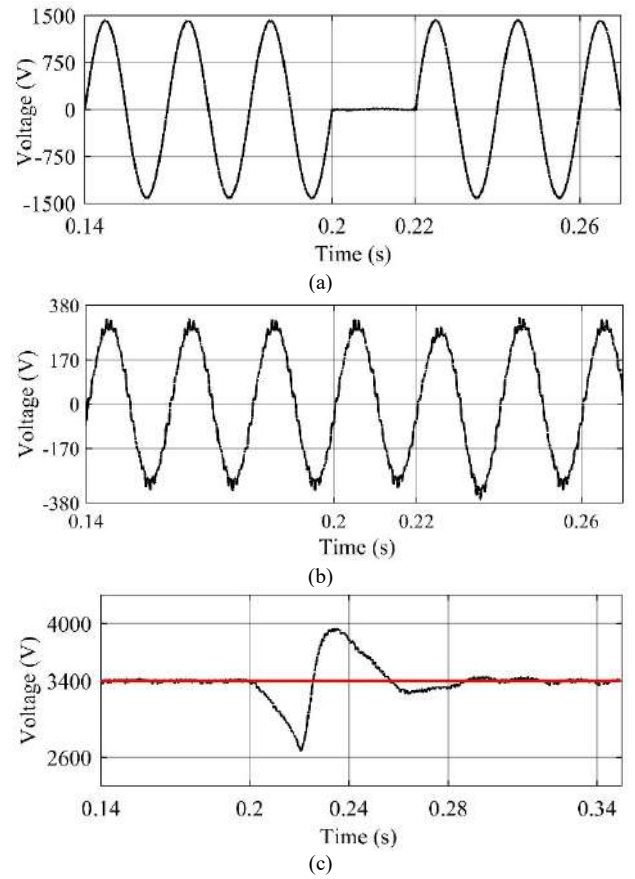


Fig. 11. Voltage interruption: (a) grid voltage, (b) load voltage, (c) MMC output DC voltage.

Despite this interruption, the load voltage remains unaffected, as illustrated in Fig. 11-b. Although the MMC output exhibits a noticeable oscillation as can be observed from Fig. 11-c, it does not propagate to the load side.

L. Voltage Transient

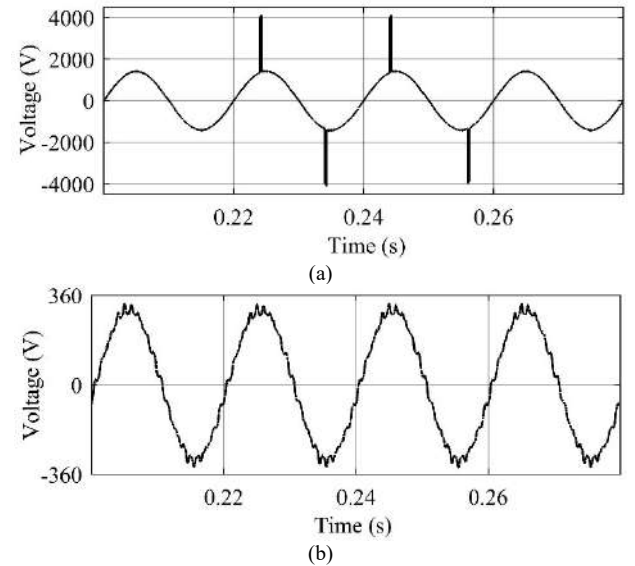


Fig. 12. Voltage transient: (a) grid voltage, (b) load voltage.

Fig. 12-a shows the source voltage under transient conditions, where $V_{g,max}$ rises abruptly from $1500V$ to $4100V$. The corresponding load voltage, shown in Fig. 12-b, remains stable, with no observable transient or distortion.

M. Frequency Variation

After testing amplitude disturbances, now the SST is evaluated under frequency variation. The source frequency is increased from 50Hz to 80Hz between $t = 0.17s$ and $t = 2.5s$, as shown in Fig. 13-a. The load voltage, shown in Fig. 13-b, exhibits a brief amplitude drop from 320V to 280V at $t = 0.2s$. This reduction lasts for only half a cycle and has no impact on connected equipment.

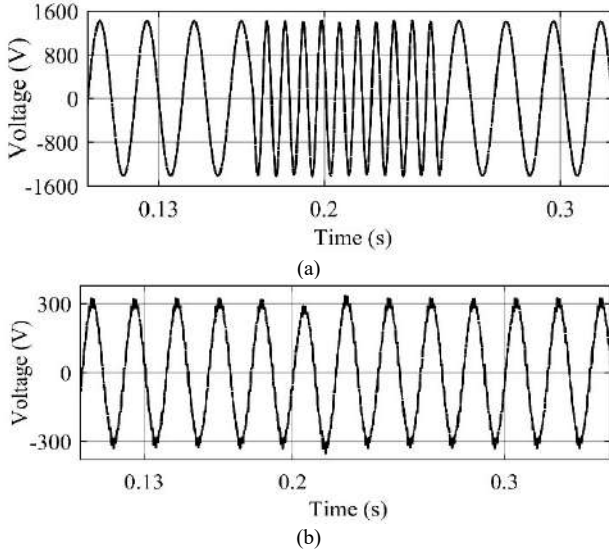


Fig. 13. Frequency variation: (a) grid voltage, (b) load voltage.

N. DC-Offset

In this test, an 800V DC offset is added to the source voltage at $t = 0.15s$. As shown in Fig. 14-a, the grid voltage waveform is shifted upward. However, this offset is completely removed by the SST stages, and the load voltage remains unaffected, as shown in Fig. 14-b.

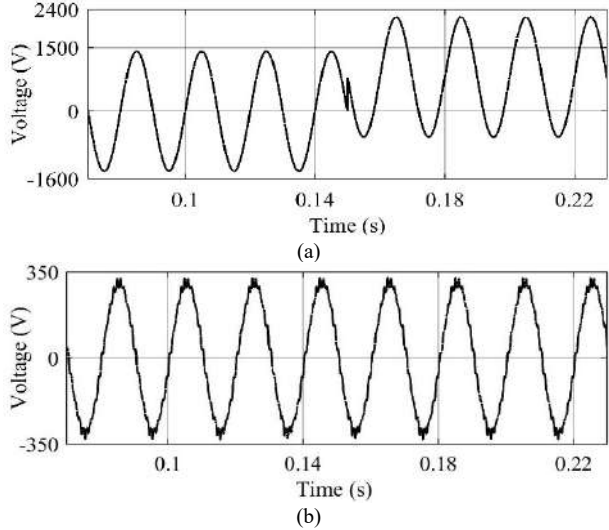


Fig. 14. DC offset: (a) grid voltage, (b) load voltage.

O. Voltage Unbalance

The seventh test considers a three-phase voltage unbalance. From $t = 0.15s$ to $t = 0.21s$, the grid voltages become unbalanced, as shown in Fig. 15-a. The corresponding load voltage, depicted in Fig. 15-b, shows no signs of imbalance, demonstrating one of the major advantages of an SST-based system.

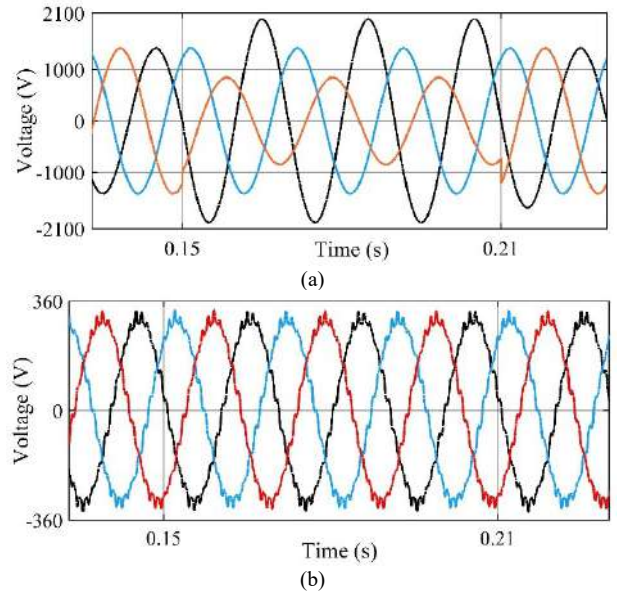


Fig. 15. Voltage unbalance: (a) grid voltages, (b) load voltages.

P. Harmonics

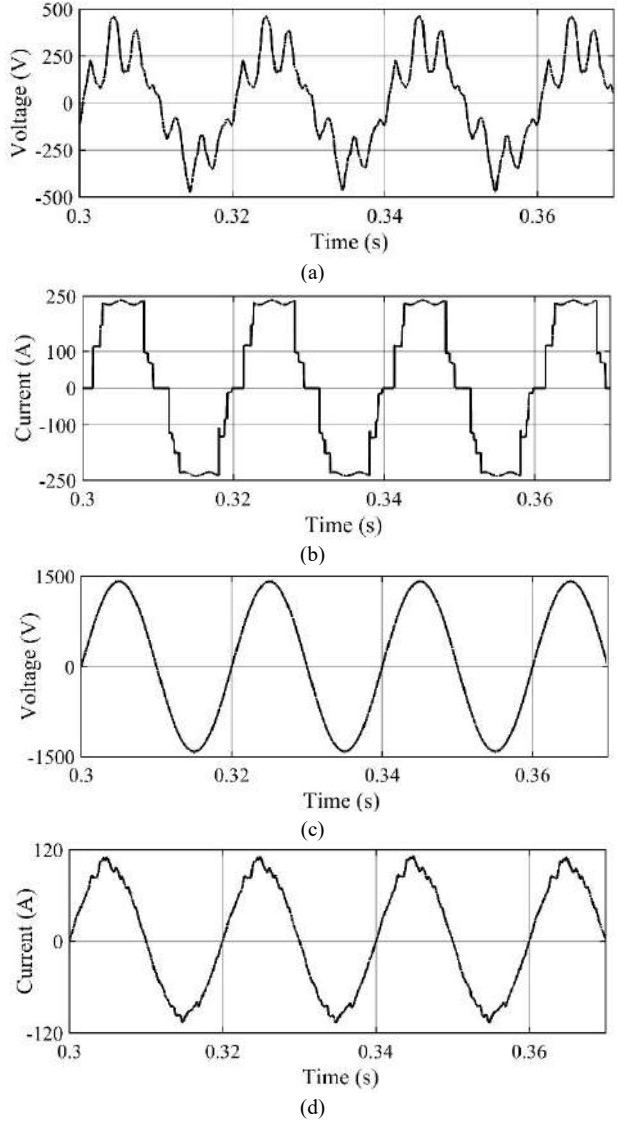


Fig. 16. SST under nonlinear load: (a) load voltage, (b) load current. (c) grid voltage, (d) grid current.

Finally, the SST is tested under harmonic distortion. A nonlinear load (three-phase rectifier) is connected to the SST

output stage. The load voltage and current waveforms are shown, respectively, in Fig. 16-a and Fig. 16-b, where harmonic distortion is evident. However, these harmonics are mitigated by the SST stages and do not propagate to the source. As illustrated in Fig. 16-c and Fig. 16-d, the grid voltage and current remain purely sinusoidal.

VI. CONCLUSION

This paper investigated the potential of SSTs to enhance the reliability and power quality of modern power grids. A comprehensive comparison between conventional LFT-based systems and SST-based architectures highlighted the fundamental limitations of traditional grids, particularly their inability to suppress or isolate disturbances. Through detailed modelling and simulation, the performance of a three-stage SST configuration was evaluated under a wide spectrum of grid disturbances.

The simulation results confirmed that the SST effectively isolates and mitigates power quality issues. In each case, disturbances present on the grid side were either significantly reduced or completely eliminated before reaching the load side, ensuring stable and distortion-free operation. This inherent fault isolation capability represents a major advantage of SST-based networks compared to conventional systems, which directly transfer disturbances between the source and load.

Beyond disturbance rejection, SSTs bring additional benefits, including compactness, reduced weight, and the ability to integrate renewable energy sources and distributed generation seamlessly into the grid. Nevertheless, several challenges remain before SSTs can achieve widespread adoption. Issues such as high implementation cost, complexity of control, and long-term reliability under large-scale operation still require further investigation.

Future work should focus on experimental validation, hardware prototyping, and optimization of SST topologies to achieve higher efficiency and cost-effectiveness.

REFERENCES

- [1] T. Oliveira, A. Mendes, and L. Caseio, "Model predictive control for solid state transformers: advances and trends," *Energies*, vol. 15, no. 22, p. 8349, 2022.
- [2] M. Ebrahim and A. Firouzjaee, "Advanced modelling of solid-state transformer," Ph.D. dissertation, Universitat Politècnica de Catalunya, Barcelona, Spain, 2018. [Online]. Available: <http://hdl.handle.net/10803/461451>
- [3] M. S. Mollik, M. A. Hannan, M. S. Reza, M. S. A. Rahman, M. S. H. Lipu, P. J. Ker, M. Mansor, and K. M. Muttaqi, "The advancement of solid-state transformer technology and its operation and control with power grids: A review," *Electronics*, vol. 11, p. 2648, 2022.
- [4] H. V. M., C. Kumar, and M. Liserre, "An MVDC-based meshed hybrid microgrid enabled using smart transformers," *IEEE Trans. Ind. Electron.*, vol. 69, no. 4, pp. 3722–3731, 2022.
- [5] A. C. Nair and B. G. Fernandes, "Solid-state transformer based fast charging station for various categories of electric vehicles with batteries of vastly different ratings," *IEEE Trans. Ind. Electron.*, vol. 68, no. 11, pp. 10 400–10 411, 2021.
- [6] L. Chen, L. Ling, T. Ma, Z. Wang, C. Li, and T. Jin, "Modeling and simulation of modular multilevel converter," in *Proc. 2024 3rd Asia Power and Electrical Technology Conf. (APET)*, Fuzhou, China, 2024, pp. 343–348.
- [7] N. Zemirline, N. Kabeche, and S. Moulahoum, "A modular multilevel converter–shunt active power filter-based hybrid topology for grid power quality improvement," *Electric Power Components and Systems*, Apr. 2024, doi: 10.1080/15325008.2024.2333433.
- [8] A. Milczarek and M. Malinowski, "Comparison of classical and smart transformers impact on MV distribution grid," *IEEE Trans. Power Del.*, vol. 35, no. 3, pp. 1339–1347, 2020, doi: 10.1109/TPWRD.2019.2941641.
- [9] S. Khan, K. Rahman, M. Tariq, S. Hameed, B. Alamri, and T. S. Babu, "Solid-state transformers: fundamentals, topologies, applications, and future challenges," *Sustainability*, vol. 14, no. 1, p. 319, 2022, doi: 10.3390/su14010319.
- [10] K. E. S. Ghambirlou, A. A. Abosnina, and G. Moschopoulos, "Enhancing power quality in distribution systems through bi-directional solid-state transformers," in *Proc. 2024 IEEE 3rd Industrial Electronics Society Annual On-Line Conf. (ONCON)*, Beijing, China, 2024, pp. 1–6.
- [11] M. López, F. Briz, M. Saeed, M. Arias, and A. Rodríguez, "Comparative analysis of modular multiport power electronic transformer topologies," in *Proc. 2016 IEEE Energy Conversion Congress and Exposition (ECCE)*, Milwaukee, WI, USA, 2016, pp. 1–8.
- [12] A. Q. Huang, Q. Zhu, L. Wang, and L. Zhang, "15 kV SiC MOSFET: An enabling technology for medium voltage solid state transformers," *CPSS Trans. Power Electron. Appl.*, vol. 2, no. 2, pp. 118–130, 2017.
- [13] F. Félice and P. Révilla, *Quality of Electrical Networks and Energy Efficiency*. Paris, France: Dunod, 2009.
- [14] Y. Fan, Y. Chen, C. Huo, Z. Tan, N. Duan, and X. Dong, "Analysis of the transmission characteristics of voltage fluctuations in the power grid and rapid response prediction," in *Proc. 2024 6th Int. Conf. Energy, Power and Grid (ICEPG)*, Guangzhou, China, 2024, pp. 1139–1143.
- [15] J. Yao, W. Lv, N. Li, and T. Jiang, "Optimized method for testing equipment sensitivity to voltage sags and interruptions based on DC link voltage prediction," *Electr. Power Syst. Res.*, vol. 241, p. 111337, 2025.
- [16] V. Ignatova, "Methods for analyzing the quality of electrical power: application to voltage dips and harmonic pollution," Doctoral thesis, Joseph Fourier Univ., France, Oct. 2006.
- [17] S. Larabi and J. S. Amani, "Study and simulation of a control of an active power filter based on wavelets," Master's thesis, Abou Bekr Belkaid Univ. of Tlemcen, Algeria, Jul. 2017.
- [18] K. Ramzi, "Monitoring the frequency of the electrical network using an Arduino card," Master's thesis, Univ. of the Mentouri Brothers Constantin, Algeria, 2015.
- [19] S.-H. Hwang, L. Liu, H. Li, and J.-M. Kim, "DC offset error compensation for synchronous reference frame PLL in single-phase grid-connected converters," *IEEE Trans. Power Electron.*, vol. 27, no. 8, pp. 3467–3471, Aug. 2012, doi: 10.1109/TPEL.2012.2190425.
- [20] M. R. Airineni and S. Keerthipati, "DC offset minimisation of three-phase multilevel inverter configuration under fault and DC link voltage unbalance conditions," *IET Power Electron.*, vol. 11, pp. 293–301, 2018, doi: 10.1049/iet-pel.2017.0128.
- [21] A. Rabie, "Contribution to the study of the three-level parallel active power filter," Master's thesis, Univ. of El-Oued, Faculty of Science and Technology, Dept. of Electrical Engineering, Algeria, 2012.
- [22] P. Ferracci, *The Quality of Electrical Energy*. Schneider Electric, Tech. Notebook No. 199, Oct. 2001.

Performance Evaluation and Analysis of the DigiMesh Protocol

Sahraoui Halima

Electronics Department, Faculty of Technology, Dr.
Moulay Tahar University of Saida, Saïda 20000, Algeria.
Laboratory of Communication Technology (LTC), Dr.
Moulay Tahar University of Saida, Saïda 20000, Algeria.
Ahlemsahra2009@gmail.com

Dahani Ameer

Computational Physics Laboratory of Materials (LPCM),
Djillali Liabès University of Sidi Bel-Abbès, Sidi BelAbbès
22000, Algeria.
Electronics Department, Faculty of Technology, Dr. Moulay
Tahar University of Saida, Saïda 20000, Algeria.

Abstract— This paper presents an experimental evaluation of the DigiMesh protocol, a prominent and widely adopted approach in Wireless Sensor Networks (WSNs). The study examines key network parameters, including payload size, number of sensor nodes (multi-hop), and communication distance. These parameters are interdependent, with optimization of one often adversely affecting the others. Determining their optimal values requires empirical testing, as no direct analytical method exists.

Various scenarios are tested to assess the impact of these parameters on network performance using metrics such as reliability (packet loss and Received Signal Strength Indication (RSSI)), Round Trip Time (RTT), and throughput. The experimental results demonstrate that DigiMesh networks achieve higher throughput and lower RTT, making them well-suited for applications requiring minimal delay and high data rates. The findings confirm that DigiMesh outperforms alternative protocols in terms of throughput, reduced delay, and improved RSSI, highlighting its advantages in performancecritical WSN applications.

Keywords— Wireless Sensor Network (WSN), DigiMesh, Reliability, Throughput, Round Trip Time, Multi-Hop

I. INTRODUCTION

A Wireless Mesh Network (WMN) is a communication infrastructure comprising interconnected devices that exchange data seamlessly. Recent advancements in wireless communication have significantly accelerated the adoption of WMNs across various sectors, including defense, smart cities, healthcare, home automation, industrial manufacturing, intelligent transportation systems (ITS), vehicular communication, and smart agriculture [1–3]. Within these applications, Wireless Sensor Networks (WSNs) play a critical role by interacting with the environment and executing specific tasks.

WSNs consist of sensor nodes that communicate wirelessly, operate in remote areas, and are typically battery-powered. WMNs employ a mesh topology, where nodes communicate directly with one another. Unlike traditional centralized networks reliant on a single hub for traffic management, WMNs offer flexible, decentralized communication. This design enhances scalability and reduces the risk of single points of failure.

Moreover, the redundancy provided by multiple links and alternative paths enables selfhealing capabilities, ensuring uninterrupted operation even if individual nodes fail.

WMNs support a range of wireless technologies, including IEEE 802.11 (Wi-Fi), IEEE 802.15 (ZigBee), IEEE 802.16 (WiMAX), and cellular technologies. Their adaptability to diverse applications is further enhanced by their multi-hop architecture, allowing data to traverse long distances via intermediate nodes. This reduces the need for long-range communication modules, leading to lower power consumption while increasing bandwidth and range.

This paper evaluates the performance of the DigiMesh protocol, which is particularly effective in low-traffic scenarios and widely used in various applications. Its performance depends on key quality-of-service (QoS) metrics, such as throughput, low latency, and high reliability [4]. While previous studies have assessed these protocols through analytical models and simulations [5–9], there remains a lack of experimental evaluations, particularly concerning critical metrics such as throughput, delay, reliability, and Received Signal Strength Indication (RSSI).

For instance, [10] demonstrated how WMNs enable remote IoT connectivity in smart home systems using DigiMesh technology. Similarly, [11] introduced AquaMesh, a system that monitors aquatic environments using ZigBee, Wi-Fi, and GPRS, with real-time alerts for threshold breaches. In healthcare, [12] proposed a ZigBee-based mesh network for monitoring hemodialysis patients, enabling emergency alerts during critical conditions. Additionally, [13] developed a multifunctional interferometric radar sensor integrated with ZigBee for structural health monitoring, detecting displacements and generating alerts automatically.

While [14] examined the technical aspects of ZigBee-based WMNs, it lacked experimental evaluation of critical metrics such as throughput, delay, recovery time, and RSSI. Similarly, [15] compared ZigBee and DigiMesh without performance testing, and [16] outlined WMN configurations but omitted comparative evaluations of WMN technologies.

To the best of our knowledge, this study is the first to provide an experimental evaluation of the DigiMesh protocol. By considering real-world variables often overlooked in theoretical

models, we aim to offer practical insights into its performance. Key network parameters examined include payload size, the number of sensor nodes (multi-hop), and communication distance. Experiments were conducted across diverse scenarios to evaluate their impact on network performance, focusing on metrics such as reliability (packet loss and RSSI), round-trip time (RTT), and throughput.

This paper presents comprehensive experimental results covering essential transmission metrics for DigiMesh technology, highlighting its strengths and limitations. The remainder of the paper is organized as follows: Section II provides an overview of IEEE 802.15.4, ZigBee, and DigiMesh protocols. Section III describes the hardware and configuration of the WMN. Section IV presents the experimental performance evaluation. Finally, Section V concludes the paper and proposes future research directions.

II. WIRELESS SENSOR NETWORK (WSNs) PROTOCOLS

This section provides an overview of commonly used protocols in WSNs, focusing on the IEEE 802.15.4 physical layer, as well as the ZigBee and DigiMesh protocols.

A. The IEEE 802.15.4 standard

IEEE 802.15.4 is a modern standard for wireless communication that defines both the physical (PHY) and medium access control (MAC) layers [17], addressing the needs of applications requiring low data rates, simple connectivity, low cost, and low power consumption. Its primary application area is the implementation of WSNs [18]. The physical layer manages data transmission and reception, supporting three frequency bands: the 2.4 GHz ISM

(Industrial, Scientific, and Medical) band (worldwide) with 16 channels, the 915 MHz ISM band (North America) with 10 channels, and the 868 MHz band (Europe) with one channel [19]. Data rates for these bands are 250 Kb/s at 2.4 GHz, 40 Kb/s at 915 MHz, and 20 Kb/s at 868 MHz, respectively. The MAC layer regulates access to the radio channel using a contention-based CSMA-CA (Carrier Sense Multiple Access with Collision Avoidance) algorithm. Additionally, IEEE 802.15.4 employs Direct Sequence Spread Spectrum (DSSS) and an Offset Quadrature Phase Shift Keying (O-QPSK) modulation scheme with half-sine pulse shaping. This configuration enhances interference mitigation and provides high immunity to noise. IEEE 802.15.4 serves as the foundation for several MAC protocols, including 6LoWPAN, ZigBee, and DigiMesh, widely used in WMNs. The following sections provide a brief overview of ZigBee and DigiMesh, the protocols examined in this paper.

B. ZigBee and Digi Mesh

ZigBee is a wireless communication protocol first specified in 2004, with its first commercial product launched in 2006 [20]. The number of devices implementing ZigBee has grown significantly, driven by the rise of smart and connected homes [21]. The protocol is designed to be more cost-effective and energy-efficient compared to other wireless personal area network (WPAN) standards such as Wi-Fi and Bluetooth, making it ideal for applications requiring low power consumption and long battery life [22].

Another prominent protocol is DigiMesh, a proprietary standard developed by Digi International Inc. Similar to ZigBee, DigiMesh focuses on ease of setup and even lower power consumption. It has been widely adopted by several manufacturers and used in various applications [23]. Unlike ZigBee, DigiMesh operates with a single router node, creating a simple and homogeneous network where all nodes are interchangeable and capable of routing data. There are no parent-child relationships as in ZigBee, simplifying network setup and enhancing flexibility, scalability, and reliability. Additionally, DigiMesh supports a larger frame payload than ZigBee, improving throughput — a critical feature for applications requiring the transfer of large data blocks [23].

III. MATERIALS AND METHODS

The aim of this section is to describe the basic hardware and software components required to build a Wireless Mesh Network (WMN).

A. Hardware and software

The XBee module, developed by Digi International Inc., is a wireless communication device commonly used to build Wireless Mesh Networks (WMNs). Depending on the firmware installed, it supports either the ZigBee or DigiMesh protocol [24]. XBee modules are known for their low power consumption and cost-effectiveness, making them ideal for various IoT and WSN applications. They are configured using Digi's XCTU software, enabling seamless setup and management of both ZigBee and DigiMesh-based WMNs [25].

For our experiments, we used the Digi XBee 3 ZigBee Mesh Kit from Digi International (Part Number XK3-Z8SWZM) [26]. This kit includes three third-generation XBee radios (Figure 1), offering enhanced performance and new integrated features within a compact form factor. These XBee modules operate in the 2.4 GHz ISM (Industrial, Scientific, and Medical) band using Direct Sequence Spread Spectrum (DSSS) modulation. They support multiple communication protocols, including IEEE 802.15.4, ZigBee, and DigiMesh, making them suitable for building versatile WMNs.

Each XBee module comes with a wire antenna and was used to create both single-hop and multi-hop wireless sensor networks. We configured each node using X-CTU, Digi International's free configuration software. This tool allows users to update parameters, upgrade firmware, and conduct communication tests easily. The XBee modules were connected to a personal computer (PC) via USB cables using XBee interface boards for programming and configuration.

To ensure consistent communication, all nodes were configured with the same Personal Area Network (PAN) ID and a baud rate of 115200 bps. The rationale for selecting this specific baud rate will be discussed further in the experimental results section.

XBee modules support three interaction modes through their serial (UART) interface: AT Command, Transparent, and API modes [27]. In this study, we used the API mode due to

its flexibility and advanced control features. Detailed steps for configuring a WMN using these modes are described in [28,29] but were omitted from this paper due to space limitations.

The X-CTU software was employed for time measurements and data packet transmission from the Coordinator module to other nodes. During the experiments, data packet sizes ranged from 10 to 73 bytes. Tests were conducted in both indoor and outdoor environments. For distance evaluation, we used a 600meter-long open area, while other tests were performed in obstacle-rich environments to simulate realistic deployment conditions.

The topologies studied are illustrated in Figure 2. Each test discusses the number of nodes used, their roles within the network, and the configuration details.

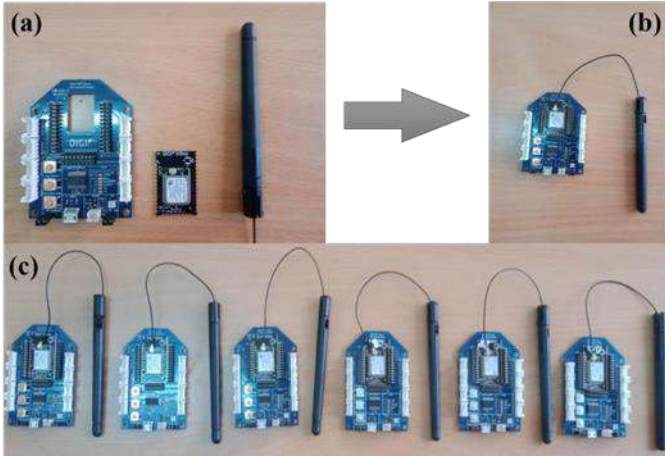


Fig. 1. (a) Photo of various components of the Digi XBee 3 RF device, (b) Photo of assembly of the Digi XBee 3 RF module, (c) Photo of the Digi XBee 3 RF Kit.

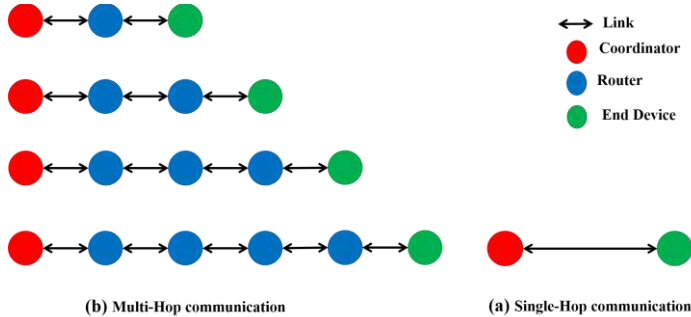


Fig. 2. (a) Single-Hop Communication: Direct transmission between the Coordinator and the End Device. (b) Multi-Hop Communication: Transmission facilitated through 1, 2, 3, or 4 intermediate Routers.

IV. RESULTS AND DISCUSSION

This section provides a performance evaluation of DigiMeshbased Wireless Mesh Networks (WMNs) in two distinct scenarios. In the first scenario, the maximum communication range supported by the protocol is measured and compared with theoretical expectations. In the second

scenario, key performance metrics such as throughput and delay are analyzed to assess the protocol's effectiveness under varying conditions.

A. Received Signal Strength Indicator

RSSI represents the signal strength level of a wireless device based on the last successfully received packet, measured in dBm [30]. In free space, the Received Signal Strength (RSS) decreases as the distance from the sender increases. The signal power at a distance d can be calculated using the Friis Transmission Equation (1) in dBm [31]:

$$p_r = p_t + G_t + G_r - 20 \log_{10} d - 20 \log_{10} f + 20 \log_{10} \left(\frac{c}{4\pi} \right) \dots (1)$$

Where :

p_r : Power at the receiving antenna.

p_t : Transmitter output power = +8 dBm

G_t : Transmitter antenna gain = 2.1 dB

G_r : Receiver antenna gain = 2.1 dBi

f : Signal frequency = 2.4 GHz

c : Speed of light in vacuum = 3×10^8 m/s

d : Distance between the transmitting and receiving antennas.

Given these parameters, theoretical RSSI values can be calculated, assuming free-space propagation without considering obstacles, environmental factors, or human interference.

Figure 3 visually represents the equation, highlighting the impact of distance on RSSI values. The signal strength calculation is based solely on the specified parameters and assumes ideal conditions, with no consideration for obstacles, environmental factors, or human obstructions. It reflects an unimpeded scenario of signal transmission and reception between antennas.

In the practical setup shown in Figure 2(a), the coordinator node was fixed in position while the End Device node was gradually moved away until communication was lost. Figure 3 illustrates the relationship between measured Received Signal Strength Indicator (RSSI) values and the distance between the nodes in the DigiMesh network. The results show a clear decline in RSSI values as the distance increases.

For the DigiMesh network, the maximum achievable distance without a router node was approximately 400 meters. This finding establishes the effective communication range in an ideal DigiMesh configuration without additional routers. However, in real-world scenarios, factors such as physical obstructions, environmental conditions, and human presence can cause signal attenuation, significantly affecting actual signal strength.

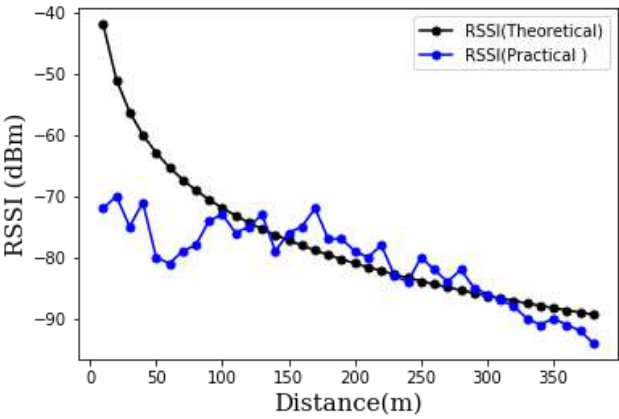


Fig. 3. Analysis of theoretical and practical RSSI strength in Digi Mesh protocol.

B. Throughput Measurement

Throughput (TP) in data transmission refers to the amount of data successfully transmitted or received within a specified time interval [32]. It depends on factors such as transmission speed, packet size, and network conditions. Throughput is calculated using the following equation (2), where the packet size is divided by the total transmission time:

$$TP = \frac{8 \times \text{number of bytes}}{\text{total transmission time (sec)}} \dots\dots\dots (2)$$

1) Throughput Measurement in Point-to-Point Link at different baud rates

In this experiment, the objective is to evaluate the impact of baud rate on the throughput of the DigiMesh protocol. Specifically, we measure throughput as a function of packet size and baud rate using the network topology illustrated in Figure 2(a). The XBee module is configured as an End Device, responsible for transmitting packets to the Coordinator node. The experiment involves measuring total transmission time while analyzing the effects of varying packet sizes and baud rates. Through a series of measurements, we compare the system's performance across different combinations of packet sizes and baud rates. Figure 4(a) illustrates the relationship between throughput and packet size for various baud rates. The results reveal that higher baud rates lead to increased network throughput due to faster data transmission. With a higher baud rate, more data can be transmitted within a given time frame, resulting in an overall improvement in throughput. According to the DigiMesh standard, a guaranteed transmission data rate of 250 kbps is specified. However, the experimental results indicate that the network's actual performance is significantly lower than this theoretical value. For instance, the achieved throughput, using the maximum offered payload of the DigiMesh protocol, was measured at only 9.36 kbps at a baud rate of 115200 bps. These findings underscore the importance of selecting an optimal baud rate to enhance network performance and data transmission efficiency. Figure 4(b) depicts the relationship between delay and packet size across different baud rates. The results show how delay

varies with packet size and baud rate. For the maximum payload supported by the DigiMesh protocol, the average delay was measured at 1 second at a baud rate of 115200 bps. This represents the average time required for a packet to travel from the source to the destination within the network. These results highlight the critical role of baud rate in optimizing network performance. The experimental findings demonstrate that a baud rate of 115200 bps provides improved throughput and reduced delay in the tested scenarios for the DigiMesh protocol.

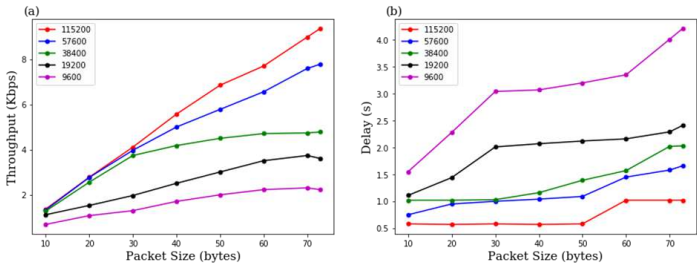


Fig. 4. (a) Throughput variation with baud rates and packet sizes in Digi Mesh network, (b) Delay variation with baud rates and packet sizes in Digi Mesh network.

2) Throughput and delay measurement in multi hop network

This test evaluates the throughput of the DigiMesh protocol by analyzing its performance concerning variables such as packet size and the number of hops. By examining different network configurations, valuable insights can be gained into the relationship between hop count and throughput. As shown in Figure 2(b), the network topology demonstrates packet relays from the End Device to the Coordinator using one and two routers. To further explore the protocol's performance, additional experiments were conducted with three-hop and four-hop setups, aiming to evaluate the impact of increased routing complexity.

Figure 5(a) illustrates the effect of hop count on the throughput performance of the DigiMesh protocol. In an ideal scenario, throughput would increase linearly with additional hops, assuming no failures or collisions. However, the results presented in Figure 5(a) indicate that the presence of routers significantly influences both data rate and overall network throughput.

Another critical performance metric is packet delay, which measures the time interval between the reception of two consecutive packets at the destination node, ensuring errorfree delivery. This metric is vital for evaluating the efficiency, reliability, and responsiveness of data transmission within the network.

In this experiment, the latency of the DigiMesh protocol was assessed considering packet size and hop count. The tested topologies, as illustrated in Figure 2(b), include 1-hop, 2-hop, 3-hop, and 4-hop networks. The findings presented in Figure

5(b) highlight the relationship between hop count and latency. For the DigiMesh protocol, using its maximum payload capacity, the average delay values were recorded as follows: 1.25 seconds for a single hop, 1.54 seconds for two hops, 2 seconds for three hops, and 2.1 seconds for four hops.

In multi-hop scenarios, the experimental analysis shows that the DigiMesh network outperforms alternative protocols by delivering higher throughput and lower latency. These results highlight the advantages of the DigiMesh protocol in efficiently transmitting data across multiple intermediate nodes, ensuring improved performance. This makes the DigiMesh network an optimal choice for applications such as ECG monitoring, particularly in scenarios requiring multi-hop communication.

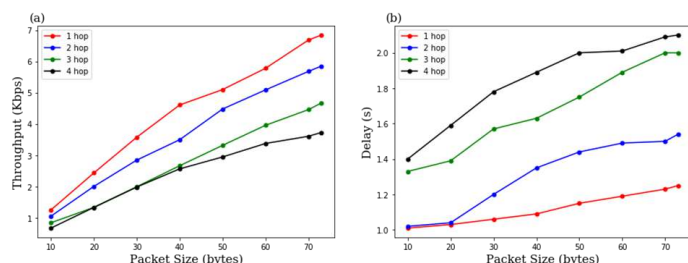


Fig. 5. (a) Throughput results as a function of packet size for networks with varying number of hops, (b) delay results as a function of packet size for networks with varying number of hops.

V. CONCLUSIONS

This paper presents an experimental evaluation of the DigiMesh network. The results demonstrate that DigiMesh networks achieve higher throughput, stronger RSSI as a function of the distance between adjacent nodes, reduced RTT delay, and faster network recovery time following a failure. For future work, the robustness of the network against external interference and noise, along with its power consumption and network lifetime, will be investigated.

ACKNOWLEDGMENT

This work was supported by the Algerian Directorate-General for Scientific Research and Technological Development (DGRSDT).

REFERENCES

- [1] Haque, K.F.; Kabir, K.H.; Abdelgawad, A. Advancement of Routing Protocols and Applications of Underwater Wireless Sensor Network (UWSN)—A Survey. *J. Sens. Actuator Netw.* 2020, 9, 19. [CrossRef]
- [2] Haque, K.F.; Abdelgawad, A.; Yanambaka, V.P.; Yelamarthi, K. Lora Architecture for V2x Communication: An Experimental Evaluation with Vehicles on the Move. *Sensors* 2020, 20, 6876. [CrossRef] [PubMed]
- [3] Haque, K.F.; Zabin, R.; Yelamarthi, K.; Yanambaka, P.; Abdelgawad, A. An IoT Based Efficient Waste Collection System with Smart Bins. In Proceedings of the 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), New Orleans, LA, USA, 2–16 June 2020; pp. 1–5
- [4] Mihajlov, Boris, and Mitko Bogdanoski. "Overview and analysis of the performances of ZigBee-based wireless sensor networks." *International Journal of Computer Applications* 29.12 (2011): 28-35.
- [5] Z. Chen, C. Lin, H. Wen and H. Yin, "An Analytical Model for Evaluating IEEE 802.15.4 CSMA/CA protocol in Low-rate wireless application", 21st International Conference on Advanced Information Networking and Applications Workshops (AINAW 2007), Ontario, Canada, May 2007, pp. 899- 904.
- [6] X. Liang and I. Balasingham, "Performance Analysis of the IEEE 802.15.4 based ECG Monitoring Network", Seventh IASTED International Conferences Wireless and Optical Communications, Montreal, Canada, June 2007, pp. 99-104.
- [7] J. S. Choi and M. C. Zhou, "Performance analysis of ZigBee-based body sensor networks", IEEE International Conference on Systems Man and Cybernetics (SMC 2010), Istanbul, Turkey, October 2010, pp. 2427-2433.
- [8] J. Zheng and M. J. Lee, "A Comprehensive Performance Study of IEEE 802.15.4", *Sensor networks operation*, Vol. 4, 2006, pp. 1-14.
- [9] G. Lu, B. Krishnamachari, C. S. Raghavendra, "Performance Evaluation of the IEEE 802.15.4 MAC for Low-Rate Low-Power Wireless Networks", IEEE International Conference on Performance, Computing, and Communications (IPCCC 2004), Phoenix, Arizona, USA, April 2004, pp. 701-706.
- [10] R. Faludi, *Building Wireless Sensor Networks: With ZigBee, XBee, Arduino, and Processing*. "O'Reilly Media, Inc., Dec. 14. 2010.
- [11] M. Y. Fathany and T. Adiono "Wireless protocol design for smart home on mesh wireless sensor network," in *Intelligent Signal Processing and Communication Systems (ISPACS)*, 2015 International Symposium on IEEE, Nov. 9, pp. 462–467, 2015.
- [12] A. J. Odey and L. Daoliang "AquaMesh-design and implementation of smart wireless mesh sensor networks for aquaculture," *American Journal of Networks and Communications*, vol. 2, no. 3, pp. 81–87, 2013.
- [13] Y. C. Du, Y. Y. Lee, Y. Y. Lu, C. H. Lin, M. J. Wu, C. L. Chen, and T. Chen "Development of a telecare system based on ZigBee mesh network for monitoring blood pressure of patients with hemodialysis in health care centers," *Journal of medical systems*, vol. 35, no. 5, pp. 877–83, Oct. 1. 2011.
- [14] C. Gu, J. A. Rice, and C. Li "A wireless smart sensor network based on multi-function interferometric radar sensors for structural health monitoring," in *Wireless Sensors and Sensor Networks (WiSNet)*, 2012 IEEE Topical Conference on pp. 33–36, IEEE, Jan. 15. 2012.
- [15] E. D. Pinedo-Frausto and J. A. Garcia-Macias "An experimental analysis of Zigbee networks," in *Local Computer Networks*, 2008. LCN 2008. 33rd IEEE Conference on pp. 723–729, IEEE, Oct. 14. 2008.
- [16] XBee and Digi. "Wireless Mesh Networking: ZigBee vs. Digi Mesh," White paper. Digi. Inc. 2008
- [17] IEEE Std 802.15.4-2006 – Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (WPANs), September 2006.
- [18] Verdone, R.; Dardari, D.; Mazzini, G.; Conti, A. *Wireless Sensors and Actuator Networks*. Elsevier : London, UK, 2008.
- [19] E.Callaway, P.Gorday, L. Hester, J.A Gutierrez, M.Naeve, B. Heile, V. Bahl, "Home networking with IEEE 802.15.4: A developing standard for low-rate wireless personal area networks", *Communications Magazine*, IEEE, Vol. 40, No. 8, Aug 2002.
- [20] Zigbee standard timeline. <https://www.zigbee.org/the-zigbee-alliancecelebrates-15-years-and-a-decade-of-standards/>. Accessed: 2019-04-09.
- [21] H. J. Patel, M. A. Temple, and R. O. Baldwin. Improving zigbee device network authentication using ensemble decision tree classifiers with radio frequency distinct native attribute fingerprinting. *IEEE Transactions on Reliability*, 64(1):221–233, March 2015.

- [22] Vanika Pahwa. 5 reasons zigbee is ideal for smart homes. <https://www.einfochips.com/blog/5-reasons-zigbee-is-ideal-for-smart-homes/>, Sep 2017. Accessed: 2019-04-02.
- [23] XBee and Digi. "Wireless Mesh Networking: ZigBee vs. Digi Mesh," White paper. Digi. Inc. 2008.
- [24] Digi XBee® 3 DigiMesh 2.4 RF Module: user guide
- [25] Digi International. (2008). XCTU-configuration and test utility software: user guide.
- [26] Farahani, S. (2011). ZigBee wireless networks and transceivers. newnes.
- [27] Şahin, E. (2005). Swarm robotics: From sources of inspiration to domains of application. In Swarm Robotics: SAB 2004 International Workshop, Santa Monica, CA, USA, July 17, 2004, Revised Selected Papers 1 (pp. 10-20). Springer Berlin Heidelberg.
- [28] Faludi, R. (2010). Building wireless sensor networks: with ZigBee, XBee, arduino, and processing. " O'Reilly Media, Inc."
- [29] Bell, C. (2020). Beginning Sensor Networks with XBee, Raspberry Pi, and Arduino. Apress.
- [30] D. International, "XBee User Manual," ed: Digi International, 2012, pp. 1-155.
- [31] Linx Technologies. n.d. RF Modules - Page 2 Of 2 - Linx Technologies. [online] Available at: <https://linxtechnologies.com/wp/c/blog/rfmodules/page/2/> [Accessed 16 May 2020].
- [32] D. International, "XBee User Manual," ed: Digi International, 2012, pp. 1-155.

Multi-step Gray Wolf optimization-based FOPID controller for the DFIG-Wind energy conversion system.

Mabrouk Dahane

Instrumentation Laboratory and
Advanced Materials,
Nour El Bachir University Center,
, El Bayadh, Algeria.
m.dahane@cu-elbayadh.dz

Abdelkrim Benali

Instrumentation Laboratory and
Advanced Materials,
Nour El Bachir University Center,
, El Bayadh, Algeria.
a.benali@cu-elbayadh.dz

Hamza Tedjini

Smart Grids & Renewable Energies
Laboratory SGRE-L,
Tahri Mohammed University,
Bechar, Algeria.
tedjini.hamza@univ-bechar.dz

Abstract— This paper presents a novel contribution in the field of the optimization of methods of control of the wind systems. The Fractional-Order PID control technique produces pretty intriguing results, but its major challenge in employing Fractional-Order PID controller lies in the tuning of their parameters, which affects the system's precision. We propose in this study the Hierarchical multi-step Gray Wolf optimization (HMS-GWO) to achieve optimal control parameters of FOPID to manage the powers of the doubly fed asynchronous generator (DFIG) based wind turbine. The findings obtained from this work, either by follow-up or harmonic analysis test, demonstrate an important performance improvement compared to GWO-FOPID control technique.

Keywords— Doubly Fed Induction Generator, WECS, FOPID Control, Hierarchical multi-step Gray Wolf optimization Algorithm.

I. INTRODUCTION

The wind energy field has witnessed a significant increase in interest owing to its multiple advantages, including environmental cleanliness, a continuous and ample energy supply, and cost-effectiveness [1, 2]. Wind turbines are classified into two categories: horizontal axis and vertical axis wind turbines. Horizontal-axis wind turbines are extensively accepted because of their greater energy efficiency when compared with vertical-axis turbines [3]. The predominant method, because of its efficiency and straightforward design, is the utilization of variable speed wind turbines employing DFIG technology [4]. Numerous research studies have been conducted on various control strategies for DFIG [5]. One of the most prevalent control strategies is decoupling PI control of output active and reactive power to enhance the dynamic behavior of wind turbine. Cependant due to uncertainty about the exact model and behavior of some parameters like as parameters values variations during operation due to temperature, or unpredictable wind speed, the conventional PI controller with a constant gain is not sufficient to guarantee the system stability of a large power system [6].

Therefore, the main contribution of the present research is the creation of a robust control strategy based on Fractional-Order PID in the inner loop of the rotor-side controller for the DFIG to enhance the quality of the power supplied by DFIG and reduce changes in current. The biggest issue in using FOPID controllers lies in the setting of their parameters. In this work, the HMS-GWO method has been selected to

improve the controller gains by minimizing various cost functions.

The paper is structured in the following way: the DFIG mathematical model is supplied in Section II. Section III details the proposed HMS-GWO-FOPID. The simulation results are provided in section IV. Section V offers the conclusion.

II. SYSTEM MODELING

A. Model of wind turbine

The turbine converts the kinetic energy of the wind into mechanical energy, however only a fraction of the power that is available can be harnessed by the wind turbine [7].

$$(1) \quad P_{aéro} = \frac{1}{2} \pi \rho R^2 V_{vent}^3 C_p(\lambda, \beta)$$

$C_p(\lambda, \beta)$ indicates the power coefficient of the wind turbine. The optimum mechanical power is given by [8]:

$$T_{aéro} = \frac{P_{aéro}}{\Omega_t} = \frac{1}{2 \Omega_t} \rho \pi R^2 V_{vent}^3 C_p(\lambda, \beta) \quad (2)$$

B. DFIG Model

The subsequent system of equations defines the global modeling of the DFIG in the Park frame [9, 10]:

Stator and rotor electrical equations:

$$\begin{cases} V_{sd} = R_s I_{sd} + \frac{d}{dt} \phi_{sd} - \omega_s \phi_{sq} \\ V_{sq} = R_s I_{sq} + \frac{d}{dt} \phi_{sq} + \omega_s \phi_{sd} \\ V_{rd} = R_r I_{rd} + \frac{d}{dt} \phi_{rd} - \omega_r \phi_{rq} \\ V_{rq} = R_r I_{rq} + \frac{d}{dt} \phi_{rq} + \omega_r \phi_{rd} \end{cases} \quad (3)$$

Stator and rotor flux equations:

$$\begin{cases} \phi_{sd} = L_s I_{sd} + M I_{rd} \\ \phi_{sq} = L_s I_{sq} + M I_{rq} \\ \phi_{rd} = L_r I_{rd} + M I_{sd} \\ \phi_{rq} = L_r I_{rq} + M I_{sq} \end{cases} \quad (4)$$

Reactive and active powers of the stator:

$$\begin{cases} P_s = v_{sd} \cdot I_{sd} + v_{sq} \cdot I_{sq} \\ Q_s = v_{sq} \cdot I_{sd} - v_{sd} \cdot I_{sq} \end{cases} \quad (5)$$

Reactive and active powers of the rotor:

$$\begin{cases} P_r = v_{rd} \cdot I_{rd} + v_{rq} \cdot I_{rq} \\ Q_r = v_{rq} \cdot I_{rd} - v_{rd} \cdot I_{rq} \end{cases} \quad (6)$$

The electromagnetic torque and the mechanical dynamics of DFIG are given by:

$$\begin{cases} T_{em} = P \frac{M}{L_s} (I_{rd} \phi_{sq} - I_{rq} \phi_{sd}) \\ J \frac{d\Omega_{mec}}{dt} + f \Omega_{mec} = T_g - T_{em} \end{cases} \quad (7)$$

III. PROPOSED HMS-GWO-FOPID

A. Fractional-Order PID controller design

The FOPID controllers are an advanced iteration of the traditional PID controller, with two supplementary fractional order parameters (λ and μ). The two tuning parameters of the controller significantly enhance control over dynamic systems, resulting in a more robust, adaptable, and broader operational range for the controller. This makes it a good alternative to conventional controllers [11]. FOPID controller equation is of the form [12, 13]:

$$G_{FOPID} = K_p + \frac{K_i}{S^\lambda} + K_d S^\mu \quad (8)$$

The FOPID controller structure is illustrated in Fig.1.

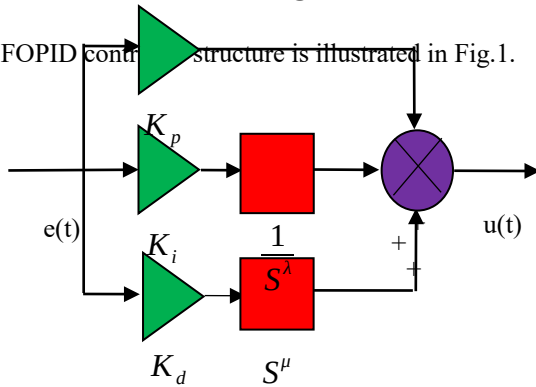


Fig.1. Block diagram of FOPID controller.

B. Hierarchical multi-step Gray Wolf optimization (HMS-GWO)

The HMS-GWO method utilizes a hierarchical structure with Alpha, Beta, Delta, and Omega wolves. Each wolf type engages in a multi-phase search process, encompassing

exploration, exploitation, and decision-making stages. The system dynamically adapts by modifying the step sequence according to prevailing conditions and feedback from superior wolves.

Potential Significance of the Numerals 1, 2, 3, 4 These figures denote various activities or stages undertaken by each category of wolf (Alpha, Beta, Delta, Omega) throughout the optimization process.

1: denotes an exploratory phase, during which the Wolf navigates the search space randomly.

2: denotes an assessment phase, during which the Wolf's fitness is evaluated.

3: denotes a decision-making juncture, as the wolf determines whether to persist in its quest or revise its location.

Hierarchical Structure: The numerical arrangement inside each row may imply a hierarchical framework, wherein the behaviors of higher-ranking wolves (Alpha) affect the behaviors of subordinate wolves (Beta, Delta, Omega). For instance, the Alpha wolf may commence an exploration phase (1), succeeded by the Beta wolf executing a search phase (2), and so forth. The steps reflect the hierarchical nature of the program, with actions and decisions determined by the alpha wolf's approval.

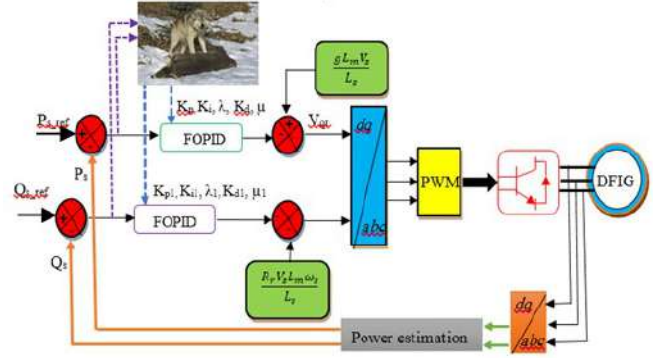


Fig.2. Direct powers control of DFIG using HMS-GWO-FOPID.

Fig. 2 provides a detailed process scheme of the HMS-GWO-FOPID.

IV. SIMULATION RESULTAT AND DISCUSSION

In order to confirm the efficiency of HMS-GWO-FOPID controller, the obtained transient responses were compared GWO-FOPID controller.

A. Reference tracking

The goal of this test is the examination of the two controllers behavior in reference tracking. The findings from the simulation are compared and evaluated in the following two sections: Trajectory tracking, harmonic analysis test.

Fig. 3 illustrates the wind speed profile used in this simulation.

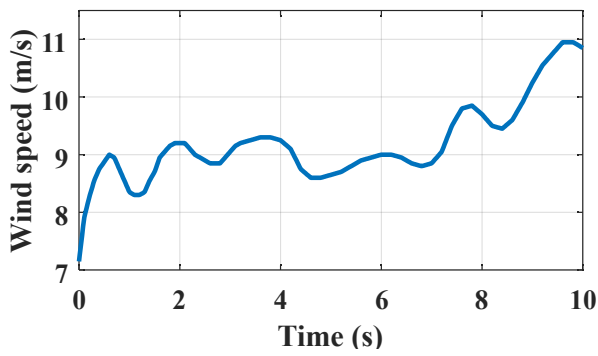


Fig.3. Wind speed profile.

Fig. 4 shows the comparative convergence curves of HMS-GWO and GWO used in the optimization of FOPID parameters. From this figure, the HMS-GWO converges to a stable solution faster and shows smoother convergence behaviour compared to GWO.

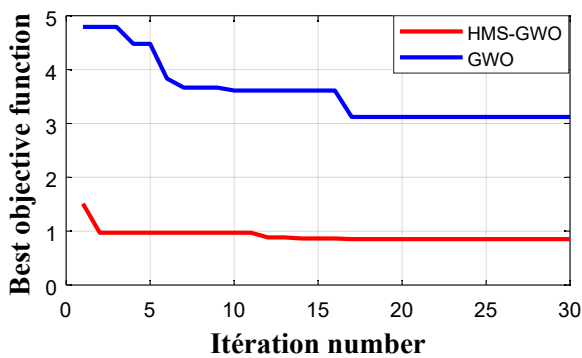


Fig.4. Convergence curves of GWO and HMS-GWO.

Figs. 5 and 6 displays the active and reactive powers of DFIG. As it's demonstrated by these figures, for the two controllers, the stator active and reactive powers follow perfectly their references but with a major response time for the HMS-GWO-FOPID controller compared to the GWO-FOPID controller. On the other side, the HMS-GWO-FOPID controller had reduced ripples for active and reactive powers. Therefore, it may be concluded that the HMS-GWO-FOPID had a very good performance for this test.

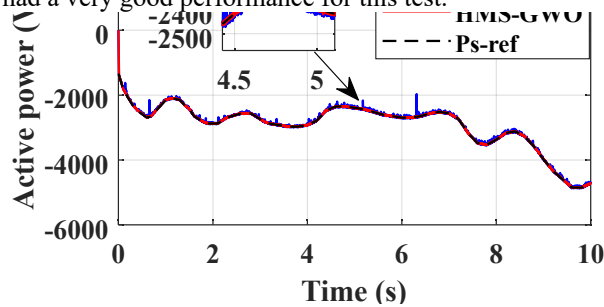


Fig.5. Active power of DFIG.

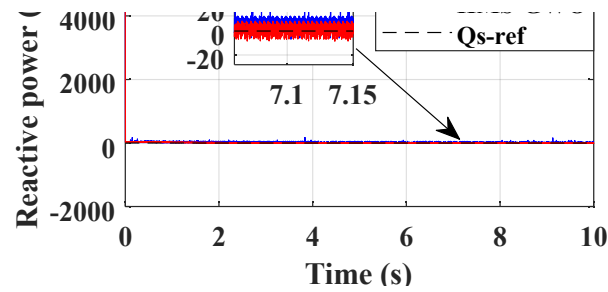


Fig.6. Reactive power of DFIG.

B. Harmonic analysis test

For a more thorough comparison of the performance of the examined controllers, Figs. 7 and 8 displays the behavior of the three-phase stator current, whereas Figs. 9 and 10 shows the THD of the stator current.

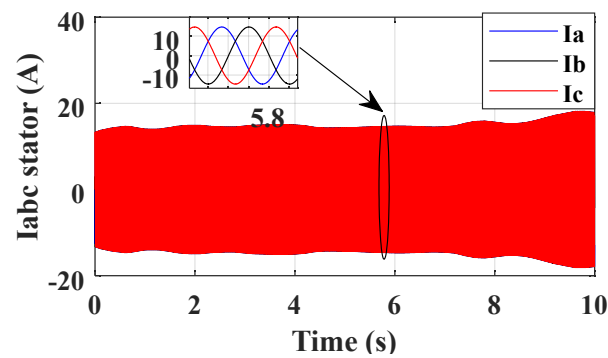


Fig.7. Stator currents using GWO-FOPID.

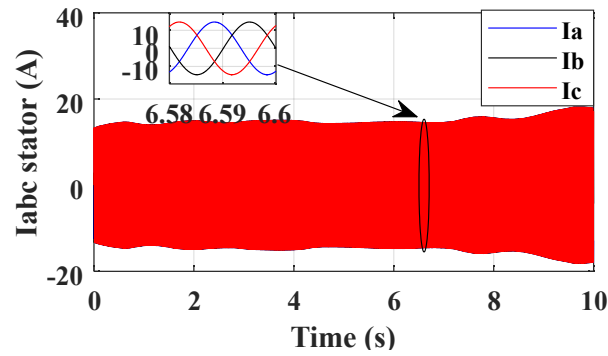


Fig.8. Stator currents using HMS-GWO-FOPID.

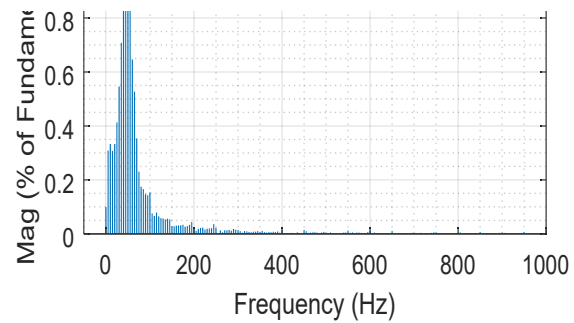


Fig.9. THD rate of GWO-FOPID.

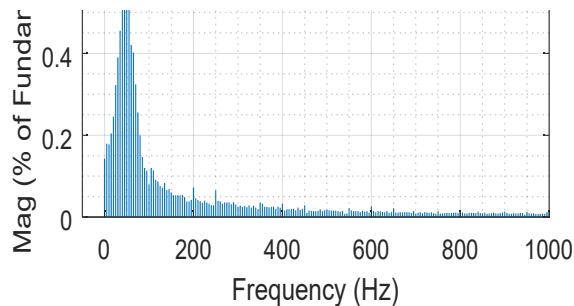


Fig.10. THD rate of HMS-GWO-FOPID.

Based on the findings obtained, the currents of the Isabc stator are sinusoidal, with superior quality achieved by the control of the HMS-GWO-FOPID.

From Figures 9 and 10, it can be shown that employing the HMS-GWO-FOPID strategy permitted a reduction of the THD of 22.05% for the stator currents compared to the GWO-FOPID control (1.36% in terms of the GWO-FOPID control and 1.06% in terms of the HMS-GWO-FOPID control).

V. CONCLUSION

In order to enhance the current responsiveness of associated with DFIG based WTS, this study proposed a new approach for fine-tuning a FOPID controller. The HMS-GWO technique was applied to calculate the optimal controller coefficients of the proposed FOPID controller. Across all examined scenarios, the simulation results consistently demonstrate that the HMS-GWO -FOPID controller trumps the GWO-FOPID controller. In particular, it produces superior stator current responses and efficiently lowers power variations. The THD in the stator current is also minimized due to the usage of the HMS-GWO-FOPID regulator. Future research will validate the suggested controller in real-time power control situations for DFIG-based wind systems.

TABLE I. PARAMETERS OF THE DFIG

Parameter	Value
Nominal power Pn	5 KW
Stator voltage Vs	381.05 V
Frequency	50 Hz
Stator resistance Rs	0.95 ohm
Stator inductance Ls	0.094
Rotor resistance Rr	1.8 ohm
Rotor inductance Lr	0.094 H
Mutual inductance Lm	0.082 H
Poles pairs p	2

REFERENCES

[1] A. Olabi, K. Obaideen, M. A. Abdelkareem, M. N. AlMallahi, N. Shehata, A. H. Alami, A. Mdallal, A. A. M. Hassan, and E. T. Sayed, "Wind energy contribution to the sustainable development goals: case study on London array," *Sustainability*, vol. 15, p. 4641, 2023.

[2] H. Benbouhenni, N. Bizon, I. Colak, M. I. Mosaad, and M. Yessef, "Direct active and reactive powers control of double-powered

asynchronous generators in multi-rotor wind power systems using modified synergetic control," *Energy Reports*, vol. 10, pp. 4286-4301, 2023.

[3] Y. Mousavi, G. Bevan, I. B. Kucukdemiral, and A. Fekih, "Sliding mode control of wind energy conversion systems: Trends and applications," *Renewable and Sustainable Energy Reviews*, vol. 167, p. 112734, 2022.

[4] K. Haneesh and T. Raghunathan, "Robust control of DFIG based wind energy system using an H ∞ controller," *Journal of Electrical Engineering & Technology*, vol. 16, pp. 1693-1707, 2021.

[5] H. Chojaa, A. Derouich, O. Zamzoum, A. Watil, M. Taoussi, A. Y. Abdelaziz, Z. M. S. Elbarbary, and M. A. Mossa, "Robust control of DFIG-based WECS integrating an energy storage system with intelligent MPPT under a real wind profile," *IEEE Access*, vol. 11, pp. 90065-90083, 2023.

[6] H. E. Alami, B. Bossoufi, M. E. Mahfoud, M. Bouderbala, B. Majout, P. Skruch, and S. Mobayen, "Robust finite control-set model predictive control for power quality enhancement of a wind system based on the DFIG generator," *Energies*, vol. 16, p. 1422, 2023.

[7] H. Delavari and A. Veisi, "A new robust nonlinear controller for fractional model of wind turbine based DFIG with a novel disturbance observer," *Energy Systems*, vol. 15, pp. 827-861, 2024.

[8] M. Yessef, B. Bossoufi, M. Taoussi, A. Lagrioui, H. Chojaa, B. Majout, and H. El Alami, "Robust control of a wind conversion system based on a doubly fed induction generator: A comparison between adaptive backstepping and integral sliding mode controllers," in *International Conference on Digital Technologies and Applications*, 2022, pp. 722-732.

[9] H. Itouchene, F. Amrane, and Z. Boudries, "Robust Control of DFIG Wind Turbines in Sub/Super-Synchronous Operation Using Integral Backstepping Controller," *Journal of Renewable Energies*, pp. 23-31, 2023.

[10] E. Bekiroglu and M. D. Yazar, "MPPT control of grid connected DFIG at variable wind speed," *Energies*, vol. 15, p. 3146, 2022.

[11] R. El-Khazali, "Fractional-order PI λ D μ controller design," *Computers & Mathematics with Applications*, vol. 66, pp. 639-646, 2013.

[12] A. Demirören, S. Ekinci, B. Hekimoğlu, and D. Izci, "Opposition-based artificial electric field algorithm and its application to FOPID controller design for unstable magnetic ball suspension system," *Engineering Science and Technology, an International Journal*, vol. 24, pp. 469-479, 2021.

[13] M. A. George, D. V. Kamat, and C. P. Kurian, "Electronically tunable ACO based fuzzy FOPID controller for effective speed control of electric vehicle," *IEEE Access*, vol. 9, pp. 73392-73412, 2021.

[14] I. Dagal, A.-W. Ibrahim, A. Harrison, W. F. Mbasso, A. O. Hourani, and I. Zaitsev, "Hierarchical multi step Gray Wolf optimization algorithm for energy systems optimization," *Scientific reports*, vol. 15, p. 8973, 2025.

[15] H. Y. Mahmoud, H. M. Hasanien, A. H. Besheer, and A. Y. Abdelaziz, "Hybrid cuckoo search algorithm and grey wolf optimiser-based optimal control strategy for performance enhancement of HVDC-based offshore wind farms," *IET Generation, Transmission & Distribution*, vol. 14, pp. 1902-1911, 2020.

[16] M. Gupta, P. Tiwari, R. Viral, A. Shrivastava, B. A. Zneid, and I. Hunko, "Grid-connected PV inverter system control optimization using Grey Wolf optimized PID controller," *Scientific reports*, vol. 15, p. 28869, 2025.

Medical Image Denoising Using Diffusion Models: An Experimental Study on Chest X-Ray Images

1st Mechiki abdelmounaime

Faculty of Science and Technology

Advanced Electronics and

Telecommunications Laboratory ETA

Bordj BouArréridj.algeria

Abdelmounaime.mechiki@univ-bba.dz

2nd Sid Ahmed Soumia

Faculty of Science and Technology

Advanced Electronics and

Telecommunications Laboratory ETA

Bordj BouArréridj.algeria

soumia.sidahmed@univ-bba.dz

3rd Messali Zoubeida

Faculty of Science and Technology

Advanced Electronics and

Telecommunications Laboratory ETA

Bordj BouArréridj.algeria

z.messali@univ-bba.dz

Abstract— Medical images denoising is imperative in the medical field due to the use of low doses during medical imaging. a bunch of deep learning models have emerged for noise reduction and these models rely on very large data , data sets that are often not available to the public due to privacy concerns. To overcome these limitations, unsupervised generative models have emerged which is our field of research, such as generative adversarial network and variational autoencoder and the latest of them is the diffusion models.

In this work, we adopted a denoising method based on diffusion models Which has achieved remarkable successes in recent years in the field of computer vision; however, it is still not sufficiently exploited in medical imaging, which is a very important field for diagnosing diseases and planning treatment.

Through your conference, we aspire to benefit as much as possible from your experiences and to correct our future steps in order to improve the results we have achieved which did not reach the desired goal.

Keywords—diffusion model, Denoising medical images, Generative models.

I. INTRODUCTION

X-ray imaging is a mainstay in medical diagnosis, accounting for 93.7% of the mean radiation dose applied to patients in medical diagnosis. A single , single chest X-ray indicates an effective radiation dose of approximately 0.1 mSv, that corresponds TO 10 days of ambient radiation exposure [1], posing a risk for the patient, especially for pediatric patients.

Additionally, large radiation doses can lead to premature hardware failure of the X-ray equipment, due to vaporization of the tube's anode and breakdown of the tube's filament [1]. Therefore, it is important to reduce the dose of X-ray acquisitions.

Despite the imperative to reduce radiation doses, conventional image pro-cessing methods for X-ray denoising often struggle with heuristic parameter cali-bration and prolonged execution times.

While Deep generative models such as generative adversarial networks (GANs) and variational autoencoders (VAEs) have exhibited high quality samples in a wide variety of data modalities, they still suffer from prob-blems such as the failure to capture true diversity, mode collapse, or unstable train-ing behavior [8].

Diffusion models are a class of unsupervised generative models that use training images to learn , learn a prior probability distribution of images that are sampled to generate new images

In recent years, diffusion models have become the state-of the art generative model for learning distributions of images. Compared to VAE and GANs flow approaches, diffusion models generate higher fidelity samples [2].Therefore, diffusion models are a suitable choice for our research area and more specifically, medical images denoising.

II. DIFFUSION MODEL THEORY

Diffusion models as a noising and denoising process consist of two iterative processes.

The forward process is simple: we take an image x_0 as input, and corrupt it repeatedly by adding artificial random noise for T steps, this process yields a pure noise image x_T where all original information has been lost.

A backward process is the opposite of a forward process. We generate a pure noise image x_{t-1} , and iteratively denoise it for a fixed number of time steps.

After the T time steps, the output x_0 is a high-quality image belonging to the same distribution of images as the training data.

The denoising step is learned by training a neural network to remove different levels of artificial noise from images [2]. Figure 1 shows both the random forward process and the generative backward process.

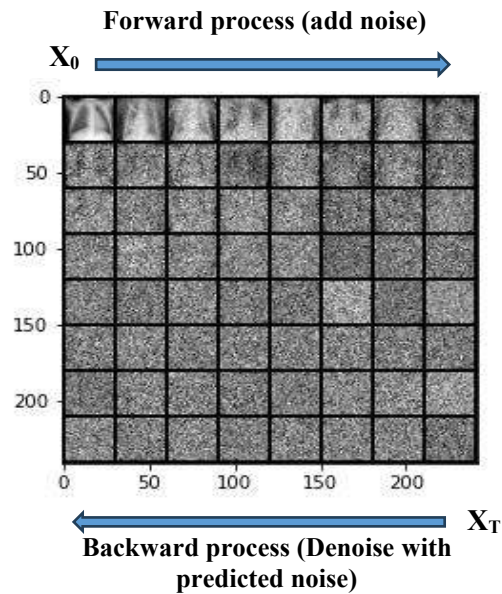


Fig. 1. A random forward process and a generative backward process.

III. RELATED WORK

Recently, diffusion-based solutions such as DDPM (Denoising Diffusion Probabilistic Models) and DDIM (Denoising Diffusion Implicit Models), have emerged as a new method for image synthesis by gradually denoising a random noisy image in large time steps, during which a temporal encoding is attached to guide a reconstruction UNet model.

A. Denoising diffusion probabilistic model

A diffusion probabilistic model is a parameterized Markov chain¹ trained using variational inference to produce samples matching the data after finite time. Transitions of this chain are learned to reverse a diffusion process [5].

With an input data sample $x_0 \sim q(x_0)$, the forward process q adds a gaussian noise with variance $\beta_t \in (0,1)$ at each time-step to the given input x_{t-1} and produces T latent x_t where the subscription t ranges from 1 to T [3] [5] [7]:

$$q(x_1, \dots, x_T | x_0) := \prod_{t=1}^T q(x_t | x_{t-1}) \quad (1)$$

$$q(x_t | x_{t-1}) := N(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I) \quad (2)$$

Supposing that the time-steps T is long enough, and a good beta scheduler is properly designed, the latent approximates a Gaussian distribution. Therefore, if the distribution $q(x_{t-1} | x_t)$ is known, we can sample $x_T \sim N(0,1)$ and feed it to the forward process to get $q(x_0)$. On the other hand, a reverse process can be defined as a routing that gradually removes the noise in the inputs, and begins at the point $p(x_T) = N(x_T, 0,1)$.

The joint distribution $p_\theta(x_{0:T})$ is calculated from the starting [3] [5] [7]:

$$p_\theta(x_{0:T}) := p(x_T) \prod_{t=1}^T p_\theta(x_{t-1} | x_t) \quad (1)$$

$$p_\theta(x_{t-1} | x_t) := N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (2)$$

In this case, p could be considered an approximation of q in each time step t . Therefore, q and p are components of a variational auto-encoder. The loss can be defined as the variational lower bound on negative log-likelihood, and is formally rewritten as the sum of loss at each step [5]:

$$L_{vlb} := L_0 + L_1 + \dots + L_{T-1} + L_T \quad (5)$$

$$L_0 := -\log p_\theta(x_0 | x_1) \quad (6)$$

$$L_{t-1} := D_{KL}(q(x_{t-1} | x_t, x_0) \| p_\theta(x_{t-1} | x_t)) \quad (7)$$

$$L_T := D_{KL}(q(x_T | x_0) \| p(x_T)) \quad (8)$$

Markov chain¹: is a sequence of events where the future state depends only on the current state not All of the previous state

For DDPM Markov chain adds noise gradually to the data until the signal is destroyed.

Aside from L_0 each term of Equation (5) is a KL divergence¹ between two Gaussian distributions, and can thus be evaluated in closed form. To evaluate L_0 for images, we assume that each color component is divided into 256 bins, and we compute the probability of $p_\theta(x_0 | x_1)$ landing in the correct bin. Also, note that while L_T does not depend on θ , it will be close to zero if the forward noising process adequately destroys the data distribution so that $q(x_T | x_0) \approx N(0,1)$.

Finally, (Ho et al. 2020) [3] found that predicting ϵ worked best, especially when combined with a reweighted loss function [3] [5]:

$$L_{simple} = E_{t, x_0, \epsilon} [\| \epsilon - \epsilon_\theta(x_t, t) \|^2] \quad (9)$$

The authors found that optimizing this reweighted objective resulted in much better sample quality than optimizing L_{vlb} directly. One subtlety is that L_{simple} provides no learning signal $\sum_\theta(x_t, t)$. This is irrelevant, however, since Ho et al (2020) achieved their best results by fixing the variance to $\sigma_t^2 I$ rather than learning it.

B. Noise Schedule

Linear noise schedule proposed by (Ho et al. 2020) [3] worked well for high-resolution images and is suitable for medical images size.

Cosine schedule proposed by (Alex Nichol and Prafulla Dhariwal, 2021) [5] is designed to have a linear drop-off in the middle of the process while changing very little near the extremes to prevent abrupt changes in noise level.

Fig. 2 represents the Linear Schedule (red curve) and Cosine Schedule (green curve) in terms of T .

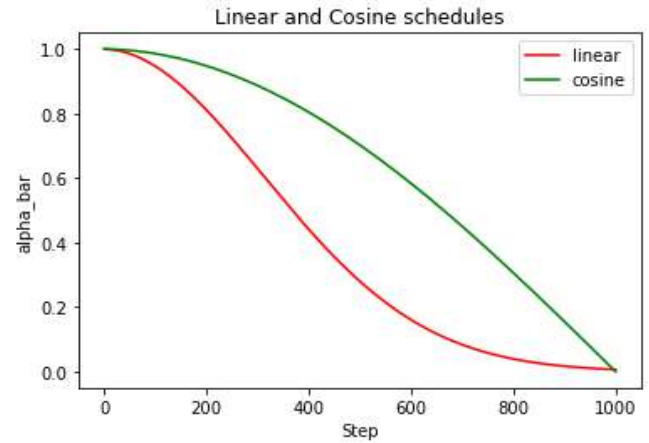


Fig. 2. Linear and Cosine schedules

We can see that the linear schedule falls towards zero much faster, destroying information more quickly than necessary.

KL divergence¹: (Kullback–Leibler divergence) is a measure of how one probability distribution differs from another.

IV. DATASET AND TRAINING DETAILS

To delve deeper into diffusion models and their effectiveness in generating realistic images, we conducted a practical study on a set of images from Chest X-rays (4000 images) [11] with a resolution of 128 x 128 and batch size=32 as Fig. 3 shows:

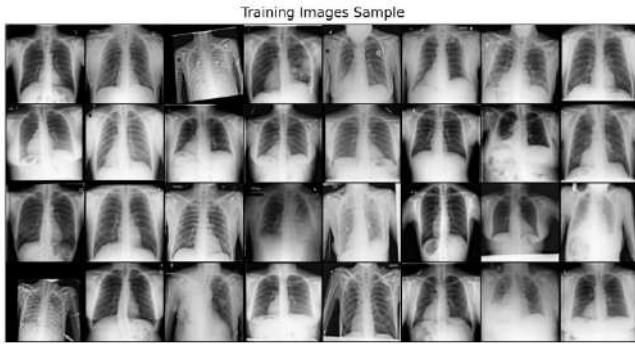


Fig. 3. Training images sample

Here are some details about building the program where we used:

linear schedule with $\beta_{\text{star}}=0.0001$ and $\beta_{\text{end}}=0.02$, a thousand-time steps $T=1000$. And we add gaussian noise gradually using a linear schedule as shown in the fig. 4

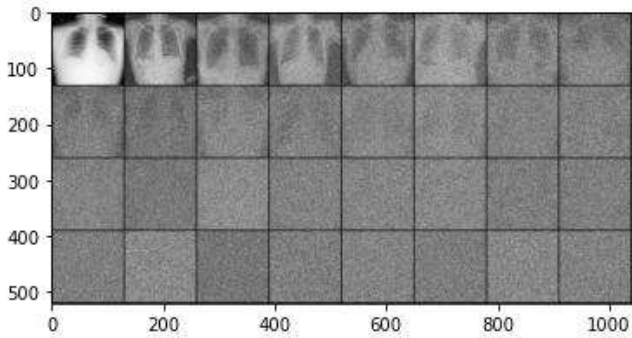


Fig. 4. Noised image with linear schedule

We divided the image set into 10 % for testing(400 images), 10 % for validation(400 images) and the rest for training(3200 images).

A. Software and hardware

We trained our model on a PC with the following characteristics:

- DESKTOP-7K8R2JN.
- AMD Ryzen 7 7700X 8-Core Processor 4.50 GHz.
- 64,0 Go (63,1 Go utilisable).
- NVIDIA GeForce RTX 4060.
- Spyder (Python3.9). pytorch.

To predict Gaussian noise from noisier to less noisy images, we used the U-Net architecture.

B. U-Net Architecture.

The U-Net is an encoder-decoder network with skip connections. It's customized here with advanced blocks for generative tasks.

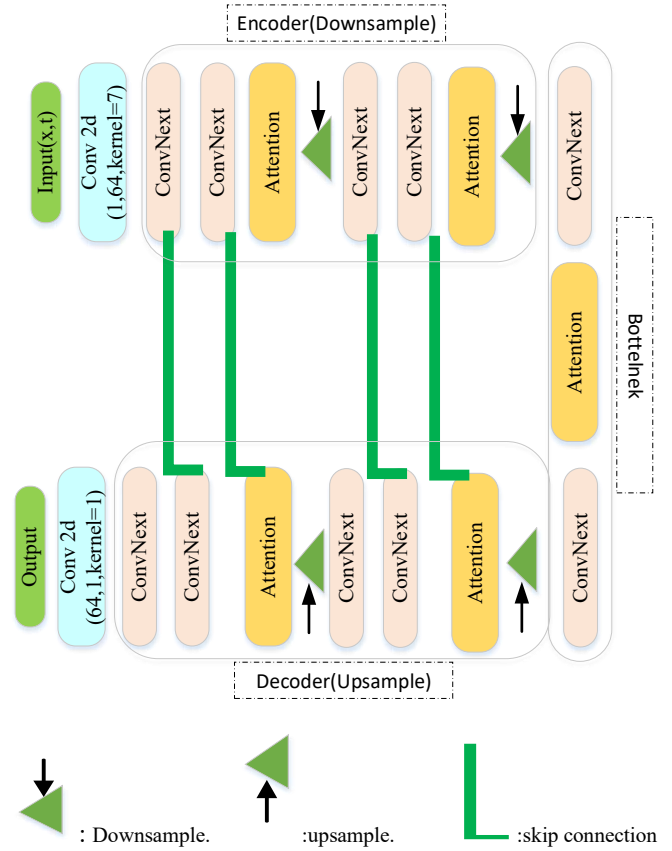


Fig. 5. U-Net architecture.

- Input(x , t): The input images are grayscale images. [1,128,128].
- TimeEmbedding¹:SinusoidalPositionEmbeddings
Generates a sinusoidal embedding for a given time step. We use it to encode time step information.
- Initial conv2d: 7x7 convolution that maps the input channels to the initial dimension.[64,128,128]
- Downsample: a series of ConvNextBlock layers, each followed by a downsampling operation. The output of each block is saved for later use in the upsample path.
2 × ConvNextBlock:[128,128,128].
Downsample:[128,64,64].
2 × ConvNextBlock:[256,64,64].
Downsample:[256,32,32].

Time Embedding¹: adds a "time"-dimension to these models, giving the model context at each step of the reverse diffusion process. These embeddings allow diffusion models to understand the sequence order and the relative positioning of data points, enhancing their ability to reverse the diffusion process effectively.

- Bottleneck: -Bottleneck: The middle part of the U-Net, which consists of two ConvNextBlock layers with an attention mechanism in between. [512,32,32].
- Upsample: a series of ConvNextBlock layers, each preceded by a concatenation with the corresponding feature map from the downsample path. The output is then upsampled.
Upsample:[512,64,64].
Skip connection from encoder.
 $2 \times \text{ConvNextBlock}$:[256,64,64].
Upsample:[256,128,128].
Skip connection from encoder.
 $2 \times \text{ConvNextBlock}$:[128,128,128].
- Final Convolution: A final convolution that maps the features back to the desired output channels.[1,128,128]
- Output: Images at the output are generated by noise prediction. [1,128,128].

1) Strengths of This Architecture

- Powerful Feature Extraction:
ConvNeXt blocks capture rich features.
- Time Conditioning:
Embeds diffusion step info for denoising tasks.
Multi-scale Feature Integration:
U-Net skip connections.
- Efficient Memory:
Uses LinearAttention for scalability.
- Normalization Stability:
GroupNorm stabilizes training across different batch sizes.

2) Weaknesses of the Architecture

- Overuse of GroupNorm(1, dim) causes batch statistics to be ignored.

May hurt performance on datasets where batch-level variance matters.

- Attention Memory Cost (Full Attention):
using full spatial attention Becomes infeasible for large images (e.g. 256×256).

- Model Size and Complexity:

The architecture is heavy in terms of:

- Number of layers.
- Attention heads.
- Convolutional depth.

V. METRICS

To gauge the quality of our generated images, we employ two metrics namely, PSNR and SSIM [4].

A. Structural Similarity Index Measure (SSIM)

SSIM evaluates the perceptual similarity between two images by considering structural information. It is defined as:

$$SSIM(x, y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \cdot \frac{2\sigma_{xy} + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \quad (10)$$

Where:

μ_x, μ_y : are the mean intensities of images x and y.

σ_x^2, σ_y^2 : are the variances of the images.

σ_{xy} : is the covariance between x and y.

C_1 and C_2 : are small constants to prevent division by zero.

It is worth noting that Higher SSIM scores are indicative of more significant image resemblance With knowledge SSIM ranges from -1 to 1, where 1 indicates identical images.

SSIM is useful for measuring perceptual quality, ensuring the generated images preserve important structures from the original images.

B. Peak Signal-to-Noise Ratio (PSNR)

PSNR is used to measure the similarity between an original image and a generated image. It is defined as:

$$PSNR = 10 \log_{10} \frac{(\text{Max})^2}{MSE} \quad (11)$$

Where:

-Max: is the maximum possible pixel value, For grayscale images [0,1] or 255 for color images[8bits].

The MSE is the error between a real clean image and a predicted clean image, and computed by:

$$MSE = \frac{1}{m \times n} \sum_{i=1}^m \sum_{j=1}^n (x_j^i - y_j^i)^2 \quad (12)$$

Where:

- m and n are the image dimensions.
- Higher PSNR indicates a higher similarity to the original image (less distortion).

VI. RESULTS AND DISCUSSION

We worked on a large number of epochs (400 epochs), the training took 42 hours, while image generation took approximately 1 hour.

In this section, we present to you the results obtained. Fig. 6 shows the results obtained.

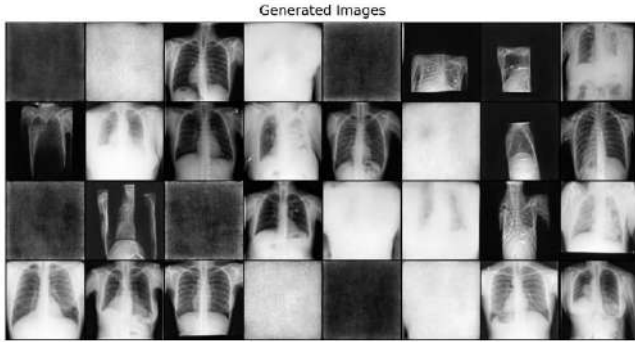


Fig. 6. Generated images

Visual results in Fig. 6: show that Our research based on DDPM for X-rays model has average performance.(We will describe the cases in detail and Always take the first picture from the left).

Although the model was unable to generate good samples in many cases.However, in some cases (First row [image 3]; Second row [image 3,5 and 8] ; third row [image 4] ; fourth row[images 1,2,3 and 7,8]) was able to generate images that were fairly close to the original preserving some details such as the rib cage bones, the middle of the lungs, and the lung margin.

The remaining generated images suffered from significant loss of detail or specific distortions and had poor contrast (first row [images 6 and 8] ; second row [images 1,2 and 4,7] ; third row [images 7 and 8]).

The rest of the images were not reconstructed by the model and remained just noise.

Generally, Our research based on DDPM for X-rays can effectively handle the different noise levels found in the images while preserving spatial resolution, it can generate images closer to the input images if we give it more improvement .

To evaluate the quality of the images we created, the results obtained were compared with three models, the first one [Zehao Fang (ControlNet¹.2024)] [6] and the second by [Pham Ngoc Huy, and Tran Minh Quan(DDMM².2023)] [10] and all of them were tested on Chest X-rays .

The results obtained are recorded in Table I:

TABLE I. IMAGE QUALITY METRICS

Method	PSNR [dB]	SSIM
Our research based on DDPM	14.66	0.3208
ControlNet (2024)	19.6	0.65
DDMM (vanilla Unets ¹ .2023)	-	0.4258

A. Discussion :

Our research based on DDPM shows Modest performance compared to other models where achieved Highest value PSNR about (14.66dB) and SSIM about (0.3208), These are values that did not achieve the desired goal.

We also note that during the research phases, as a first step, we trained the model on only 70 epochs for testing, and we achieved a PSNR of (13.41 dB) and SSIM about (0.25) . As a second step, we increased the training epochs to 200 epochs, noting a slight increase in both PSNR and SSIM, where we recorded a PSNR of 14.32 and SSIM about (0.31).

In an effort to gain the largest possible number of decibels, we doubled the training epochs to 400, but the results were very similar to our use of 200 epochs, as documented in the table above.

Since the results did not reach the desired goal, we finally concluded that the defect might be in the ability of the simple Unet used to predict noise Which makes us think of another, more effective structure.

In [6] [Zehao Fang (2024)] proposed a ControlNet-based conditional diffusion model that is a new approach that , that exploits the stable diffusion and conditional control models to produce realistic and diverse medical images.

ControlNet enables fine-tuning diffusion models with additional conditions, without causing any distortion to the original model[6].

The results achieved indicate that ControlNet can generate images with considerable precision and control ,this proves the effectiveness of this approach and makes us think about using it in the future.

In [10] , [Pham Ngoc Huy, and Tran Minh Quan] presented a diffusion-based multi-branch model that can jointly produce realistic XR medical images and their associated segmentation masks With use a vanilla Unet.

The DDMM model is built from one or more branches of DDPM that share the same noise scheduler and latent code, which enforce semantic consistency,each branch in DDMM can use other unlabeled large-scale datasets to in crease diversity and generalization[10].

Based on what was presented in [6] and [10] of the results, we concluded to think about adopting another structure for UNET that is more suitable for the medical image, instead of increasing the number of epochs.

controlNet¹ is a neural network architecture which controls diffusion models by adding extra conditions. The architecture of ControlNet consists of two main components [6]: a diffusion model and a control network.

ControlNet has been used in various applications such as generating artistic images, medical images.

DDMM²: Denoising Diffusion Medical Model, the proposed technique, can create realistic X-ray images and associated segmentations on a small number of datasets [10].

A vanilla U-Net¹, typically used in image segmentation tasks.

A vanilla U-Net and the U-Net used in DDPM share the core idea of an encoder-decoder structure with skip connections but they differ significantly in architecture and purpose.

VII. CONCLUSION

In this research paper, we shed light on one of the most important modern models, which we see serve our field of research, which is X-ray images unsupervised denoising.

Although the results that we get did not reach the desired goal, the increasing interest of researchers in generative models, including diffusion models, which have seen significant development in recent years. We attracted this to its application to our research project.

More work is needed to speed up the reconstruction process and verify the robustness of the reconstruction processes.

Fortunately, progress in this area is rapid and there is much room for further development.

In summary, diffusion models have opened new prospects for X-ray imaging, which is the cornerstone of the medical diagnosis amid efforts to reduce radiation doses.

A. Obstacles

The first obstacles I face in my research is to obtain data (low dose X-ray images).

The second obstacles, which are the limited characteristics of computers that we use, whether personal or that are available in the laboratory, as the spread models require computers with high quality characteristics in addition to the time factor, it may take more than 3 days to implement some models.

B. Ambitions

We aspire to improve results and achieve PSNR up to 30 dB or more and SSIM exceeds 0.7 through your recommendations.

In the future, we aspire to apply diffusion models in the field of medical diagnosis of one of the common diseases of the rib cage.

We seek to accept our research paper in order to participate in the organized international forum from your respected university and benefit from your experts in the forum.

REFERENCES

- [1] Daniel Sanderson et al. "Diffusion X-ray image denoising". Full Paper- MIDL 2024.
- [2] George Webber¹,* and Andrew J. Reader PhD¹, "Diffusion Models for Medical Image Reconstruction". Downloaded from <https://academic.oup.com/bjrai/advance/article/doi/10.1093/bjrai/ubae013/7745314> by guest on 04 September 2024.
- [3] Jonathan Ho, Ajay Jain, Pieter Abbeel, "Denoising Diffusion Probabilistic Models". arXiv:2006.11239v2 [cs.LG] 16 Dec 2020. Code <https://github.com/hojonathanho/diffusion>
- [4] Michael Kölle, Gerhard Stenzel, Jonas Stein, Sebastian Zielinski, Björn Ommer, Claudia Linnhoff-Popien, "Quantum Denoising Diffusion Models". arXiv:2401.07049v1 [quant-ph] 13 Jan 2024.
- [5] Alex Nichol¹, Prafulla Dhariwal¹, "Improved Denoising Diffusion Probabilistic Models", International conference on machine learning, 2021. proceedings.mlr.press.
- [6] Zehao Fang et al., "Conditional Diffusion Model for X-Ray Segmentation Data Generation". Journal of Artificial Intelligence Practice (2024) Clausius Scientific Press, Canada. DOI: 10.23977/jaip.2024.070102. ISSN 2371-8412 Vol. 7 Num.1 Arbitrary Image Transforms Without Noise. NeurIPS 2023.
- [7] Delaram Pirhayatfard, Mohammad Taha Toghiani, Guha Balakrishnan, César A. Uribe, "Improving Denoising Diffusion Probabilistic Models via Exploiting Shared Representations". arXiv:2311.16353v1 [cs.LG] 27 Nov 2023.
- [8] Prafulla Dhariwal, Alex Nichol, "Diffusion Models Beat GANs on Image Synthesis", arXiv:2105.05233v4 [cs.LG] 1 Jun 2021.
- [9] Wenjun Xia, Qing Lyu, Ge Wang, "Low-Dose CT Using Denoising Diffusion Probabilistic Model for 20x Speedup", arXiv:2209.15136v1 [eess.IV] 29 Sep 2022.
- [10] Pham Ngoc Huy, Tran Minh Quan, "DENOISING DIFFUSION MEDICAL MODELS", arXiv:2304.09383v1 [eess.IV] 19 Apr 2023.
- [11] <https://nihcc.app.box.com/v/ChestXray-NIHCC>.

Performance Characterization and Fault Impact Analysis of a Self-Excited Asynchronous Generator for Stand-Alone Applications

Belynda Fares	Radia Abdelli	AIBECHE Abderrezak	Ahcene Bouzida	AZZOUGUI Safia
<i>Department of Electrical Engineering</i>	<i>Department of Electrical Engineering</i>	<i>Department of Electrical Systems Engineering</i>	<i>Department of Electrical Engineering</i>	<i>Department of Electrotechnics</i>
<i>University of Bejaia</i>	<i>University of Bejaia</i>	<i>university of Boumerdes</i>	<i>University of Bouira</i>	<i>University of Setif</i>
<i>Bejaia, Algeria</i>	<i>Bejaia, Algeria</i>	<i>Boumerdes, Algeria</i>	<i>Bouira, Algeria</i>	<i>Setif, Algeria</i>
belynda.fares@univ-bejaia.dz	radia.abdelli@univ-bejaia.dz	aibecche_umbb@univ-boumerdes.dz,	a.bouzida@univ-bouira.dz	Safia.azzougui@univ-setif.dz

Abstract - More and more countries are investing in wind power systems to mitigate climate change and reduce their dependence on fossil fuels. These systems offer a number of advantages. Wind energy conversion systems are made up of a number of interconnected components. The induction generator is one of these, and plays an essential role in transforming mechanical energy into electrical energy. The generator may encounter various electrical or mechanical problems during operation. The aim of this work is to show the impact of certain failures, such as rotor bar breaks or short circuits, on the performance of the induction generator in a wind energy conversion system. The robustness and efficiency of wind energy systems can be improved by understanding these failures and implementing early diagnosis systems.

Index Terms - self-excited induction generator, failure, impact, finite elements.

I. INTRODUCTION

Off-grid power generation using renewable energy resources is attracting growing interest because of its ability to supply energy to areas that are remote or poorly served by the traditional electricity grid, helping to improve the quality of life of isolated communities on the one hand and reduce dependence on fossil fuels on the other. This transition to more sustainable energy sources also helps to reduce greenhouse gas emissions and combat climate change. The self-excited induction generator is one of the most widely used generators in isolated electrical energy conversion systems. Its role is to transform the mechanical energy supplied by a renewable energy source such as a wind turbine or water turbine into electrical energy. Because of this, self-excited induction generators (SEIG) have been used in applications with fixed and variable speeds and on remote sites because they are easy to build, cheap, reliable, and robust. They also don't need an external power source to create the magnetic field of excitation [1]. In off-grid applications, the SEIG requires a source of reactive power. Typically, capacitors are employed for this purpose. For the induction generator to self-excite, two conditions must be fulfilled: i) the rotor must rotate at a speed higher than the synchronizing speed, or in other words, the slip must be negative, and ii) the stator must be linked to a capacitor bank of suitable capacity to supply the

minimum reactive power [2], [3]. When these conditions are met, the induction generator can operate autonomously and generate electrical energy. The SEIG offers numerous advantages, yet it remains susceptible to faults. Induction generator faults fall into two main categories: mechanical and electrical. Mechanical faults, such as bearing issues, broken rotor, eccentricity, and misalignment, can result in excessive vibration and premature component wear. On the other hand, electrical faults, including short circuits, over voltages, and overloads, can lead to significant malfunctions in the induction generator's operation. This work aims to investigate the impact of specific faults on the performance of a self-excited induction generator. These faults include rotor bar breaks, eccentricity, stator phase openings, and stator short circuits. We will thoroughly analyze the generator performance by simulating these faults in the Ansys Maxwell finite element software environment. The results will provide a deeper understanding of how these faults affect generator operation. Additionally, they will underscore the importance of implementing detection and preventive diagnostic techniques to ensure the reliability and durability of electrical equipment.

II. SELF-EXCITED PHENOMENON IN INDUCTION GENERATOR

An induction machine is a reversible machine; it can operate in either motor or generator mode. If we drive the rotor of the induction motor at a speed higher than the synchronization speed, this generates a negative slip, which means that the induction machine operates in generator mode and can produce electricity. For the induction generator to self-excite in off-grid mode, a bank of appropriately rated or pre-charged minimum excitation capacitors must be connected between the stator terminals, giving rise to a phenomenon known as self-excitation [4]. Self-excitation is a phenomenon that allows voltage to build up across the terminals of induction generators. Without reactive source compensation, the induction generator cannot produce electrical energy. Self-excitation requires two conditions:

1. The machine rotor must have sufficient residual magnetism.
2. The capacitor bank must be of sufficient size [5].

The synoptic illustration of a self-excited induction generator-based power generation system is presented in Figure 1.

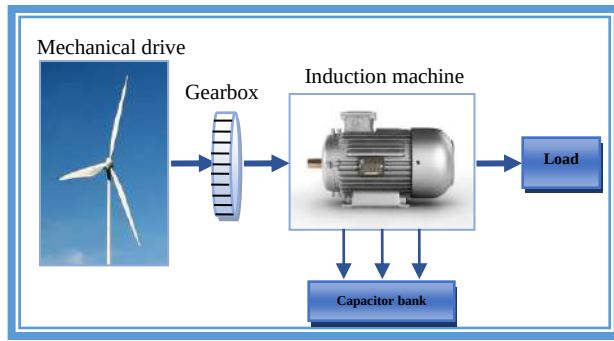


Fig. 1 Self-excited induction generator.

Self-excitation occurs when the capacitor bank supplies reactive power and when the machine retains a residual magnetic field. Without an appropriate level of residual magnetism, self-excitation will not occur [6], [7]. As the rotor of an induction machine rotates at the designated speed, the residual magnetism in the rotor iron generates a small voltage across the stator. This voltage produces a capacitor current, resulting in the creation of a flux that aids the residual flux, consequently producing more flux and voltage across the stator. This voltage then triggers a current through the capacitor bank, eventually leading to the generation of a voltage. The voltage continues to rise until the saturation curve of the induction generator intersects the capacitor's charging line at point A in Figure 2. This process gives rise to an electromotive force with no load [8].

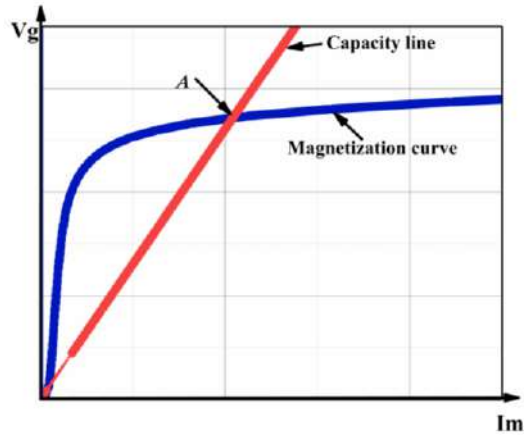


Fig. 2 Induction generator magnetization curve.

III. PERFORMANCE OF SELF EXCITED INDUCTION GENERATOR

In this paper, we present a numerical solution for a 400 KW, 4-pole, self-excited squirrel-cage induction generator (SEIG) that generates a nominal voltage of 400 V. The self-excited induction generator will be modelled under healthy and defective conditions using the time-stepped finite element method (TSFEM). To model the SEIG, we also consider the spatial distribution of the stator windings, the openings on both sides of the air gap, and the magnetic saturation of the materials. The SEIG parameters are illustrated in Table 1.

TABLE I
SEIG parameters

Parameter	Value
Rated power	400 KW
Rated Voltage	400 V
Coupling	Delta
Poles Number	4
Motoring rated speed	1491 (rmp)
Rated frequency	50 (Hz)
Working temperature	75(°)

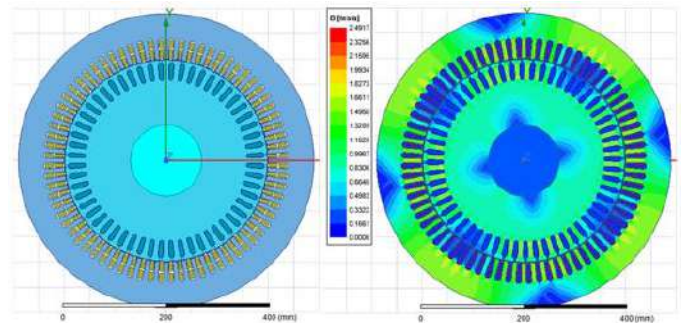


Fig. 3 Machine geometry and flow distribution

Once the machine model is completed, we start the rotor at a constant speed of 1650 rpm and introduce a modest statorique current to generate the initial magnetic flow into the machine core. This flow is essential for increasing the voltage. A bank of capacitors is connected in parallel to the stator, and each with sufficient capacity meets the reactive power demand of the self-excited induction generator.

The following figures represent the current, the three-phase voltage, and the electromagnetic torque.

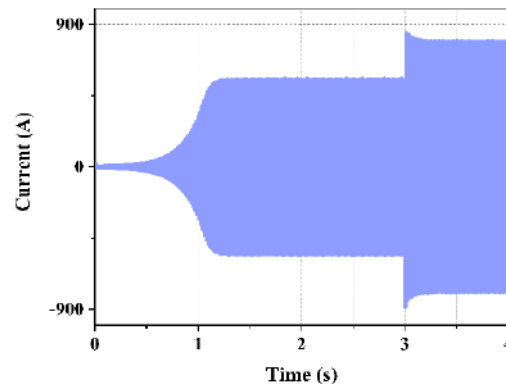


Fig. 4 Current evolution.

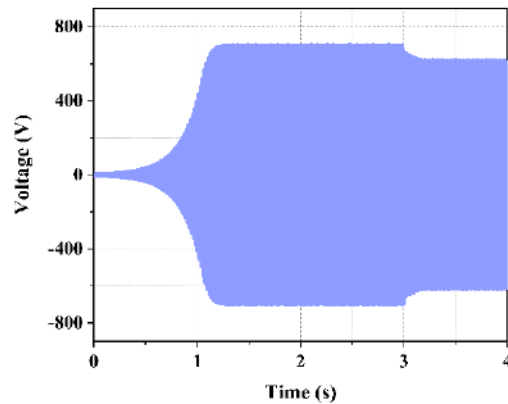


Fig. 5 Voltage evolution.

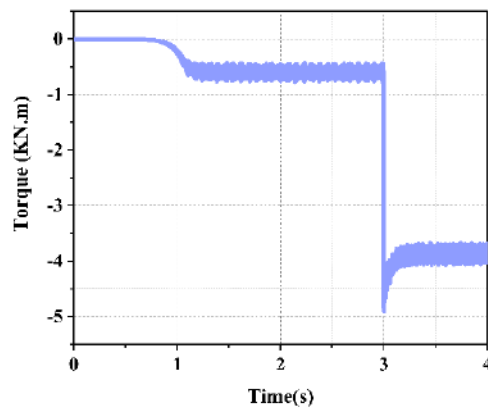


Fig. 6 Electromagnetic torque evolution.

Note that the generator takes a start-up time of 1s, after which the electrical quantities gradually increase until they reach their steady-state values. At time $t = 3$, a resistive load of 1 ohm is connected to the generator, causing the current and electromagnetic torque to increase and the voltage to decrease. It's important to note that this variation in electrical magnitudes is normal when an electrical generator is loaded. This reaction is due to the modification of the circuit and the new constraints imposed by the resistive load.

IV. INDUCTION MACHINE FAILURE

Despite its reputation for robustness, the asynchronous machine can sometimes suffer from faults of various kinds. These faults occur at several points on the machine, starting with the stator phase connection and ending with the mechanical attachment of the rotor to the load. These faults can be mechanical, electrical or magnetic, and their causes are very diverse [9].

Depending on their nature, these faults can be classified into two groups: electrical faults and mechanical faults, as shown in figure 7 and 8. Electrical faults mainly involve problems related to the motor's electrical components and systems.

These include problems related to electrical wiring, insulation failures, short circuits, open circuits, electrical overloads, voltage fluctuations and failures of components such as capacitors or switches. Mechanical faults involve the machine's mechanical components and systems. These can include problems with bearings, gears, shafts, lubrication, misalignment, imbalance, resonance, structural defects and wear of mechanical parts [10].

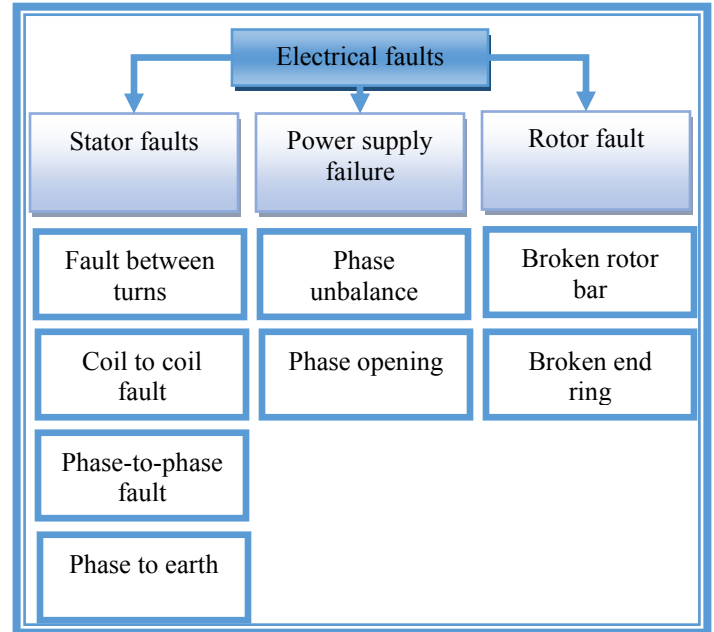


Fig. 7 Electrical faults.

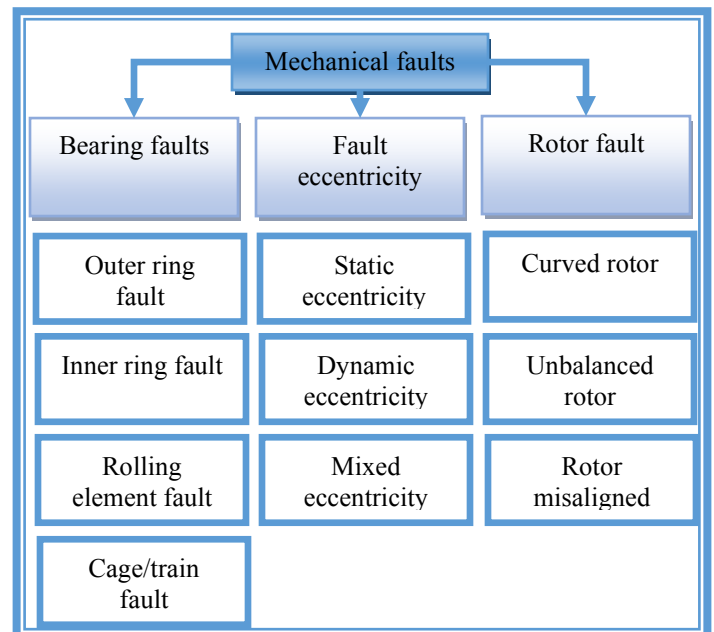


Fig. 8 Mechanical faults.

V. RESULT AND DISCUSSION

This section examines the effects of various faults on SEIG operation, paying particular attention to how rotor bar breaks,

eccentricity, short circuits and stator phase opening affect three-phase current.

1. Broken bars

A rotor bar fracture occurs when one or more rotor bars are partially cracked or completely broken. Rotor bar fracture failure can be caused by one or more of the following: manufacturing defects, thermal stresses and mechanical stresses caused by bearing faults, frequent motor start-ups at rated voltage, and rotor bar metal fatigue [11]. The following figures show the evolution of stator current in the presence of broken rotor bars, with the number of broken bars increasing by one each time.

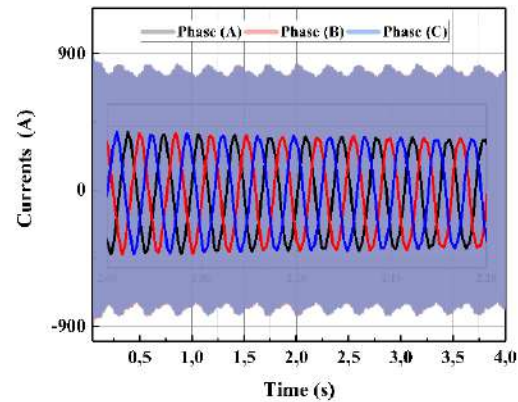
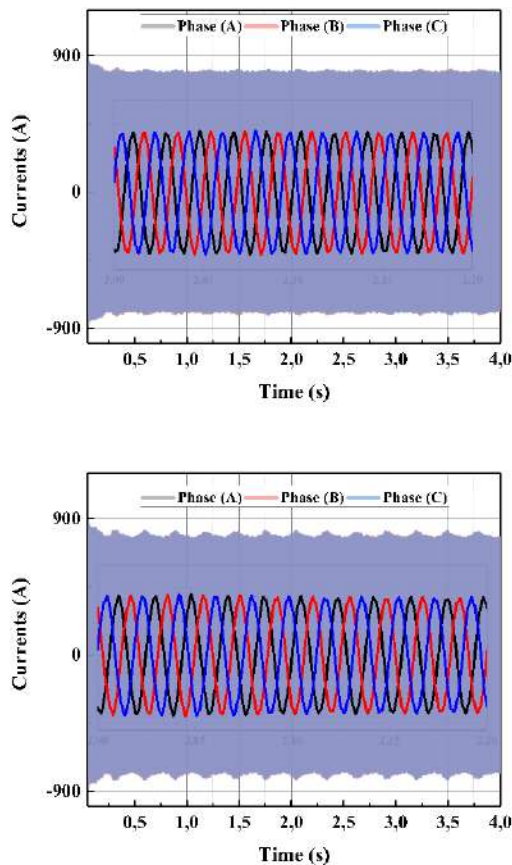


Fig. 9 Evolution of stator current with one, two and three broken bars.

2. Eccentricity

In an ideal machine, the rotor is perfectly centered in the stator bore, with the rotor's center of rotation aligned with the geometric center of the stator bore. However, when the rotor is moved from this central position, an asymmetrical air gap is formed between the stator and rotor, resulting in so-called rotor eccentricity [12]. figure 10 shows the stator current evolution in the presence of a dynamic eccentricity of 25%.

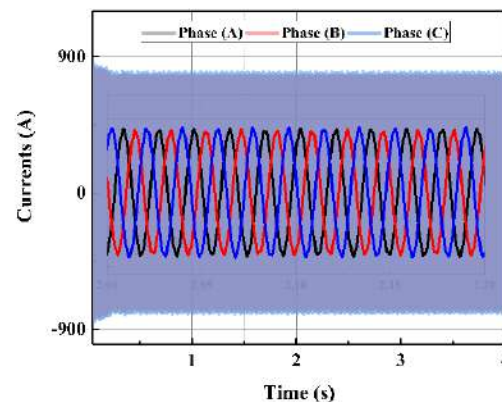


Fig. 10 Evolution of stator current with dynamic eccentricity of 25%.

3. Open phase

An open circuit fault in a three-phase induction motor is called single-phase, which means that one of the three phases has lost its connection to the supply. Mechanical vibrations might create this fault. These mechanical vibrations can be the result of a force produced by a large current in the windings, an unbalanced rotor, or other mechanical or electrical problems [13]. Figure 11 shows the evolution of the stator current in the presence of an opening of phase A

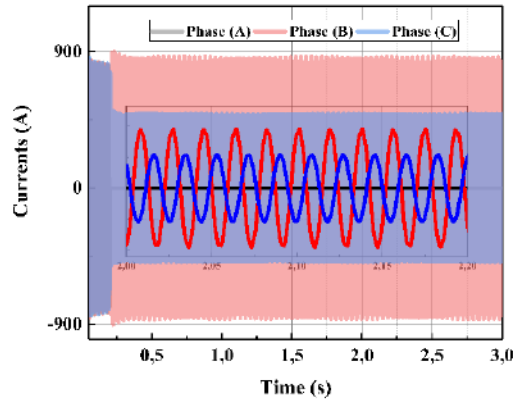


Fig. 11 Evolution of stator current with phase A open.

4. Short circuit

A short-circuit fault occurs when two or more turns of the stator windings are short-circuited, resulting in a higher-than-permissible current flowing through the short-circuited turns. The progression of a short circuit between turns generates a short circuit between coils [9].

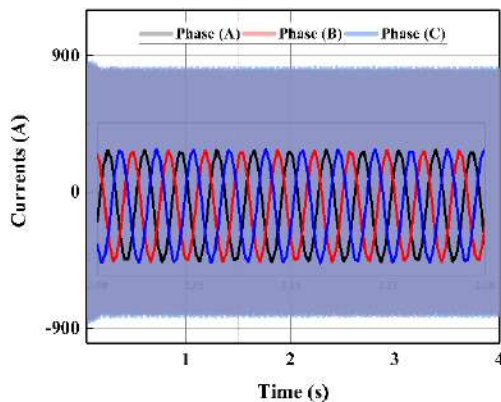


Fig. 12 Evolution of stator current with short-circuit between two spires.

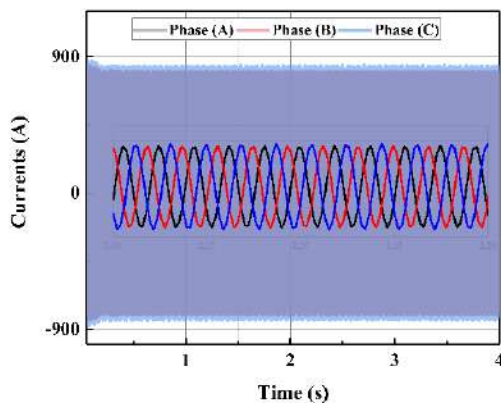


Fig. 13 Evolution of stator current with short-circuit between two coils.

From Figures 9, 10, 11, 12 and 13, we can see that all rotor bar breakage, eccentricity, phase opening and short-circuit faults

can lead to a significant reduction in SEIG performance. These faults have a direct impact on the shape of the stator current. In the case of rotor bar breakage, the shape of the stator current is distorted, and the amplitudes of the three currents are unequal: the more broken bars there are, the more acute the distortion. In the case of dynamic eccentricity, the shape of the stator current is also distorted, and the amplitudes of the three stator currents are unequal. Although this deformation is slight, it is present. In the case of phase opening, the current in the faulty phase goes to zero, while in the other two phases, the current value decreases. Generator operation becomes similar to single-phase operation. Finally, for the last fault, the short-circuit, we note that this fault leads to an increase in the current of the faulted phase.

VI. CONCLUSION

The self-excited induction generator is one of the most widely used generators in renewable energy production systems. Using a self-excited induction generator (SEIG) in wind power generation is a good choice because of its many advantages: robustness, low maintenance, low construction cost, and suitability for isolated operation. The SEIG offers very satisfactory performance in terms of reliability and energy efficiency, and its ability to self-excite makes it particularly suitable for environments where access to an external excitation source is limited. This makes it an ideal solution for remote or isolated areas, where access to electricity may be limited. Like any electromagnetic device, the SEIG can be confronted during operation with various faults that can cause partial or total malfunctions on the generator itself, or the system as a whole. In this work, we conducted a study on the effects of specific electrical and mechanical faults on the SEIG's operation, to highlight the hazards posed by these faults and propose solutions and strategies for their detection and diagnosis. The results of this study could contribute to improving the reliability and durability of electrical energy systems based on the SEIG. This work addresses faults such as rotor bar breaks, dynamic eccentricity, stator phase opening, and stator short circuits. All these faults have a direct impact on the generator's stator current. If any of these faults are present, they change the stator current. To make sure that SEIG-based systems are reliable and last a long time, fault detection and diagnostic tools must work well. Among the most popular techniques based on stator current analysis are Fourier transform analysis and wavelet transform analysis. These techniques enable us to detect and diagnose the faults presented in this work effectively. The proposed diagnostic techniques are known for their many advantages, such as their ability to precisely identify fault types and provide detailed information on their severity, as well as their efficiency and reliability. They are also simple to implement and require no major modifications to existing equipment.

REFERENCES

- [1] B. Roubah, H. Toubakh, and M. Sayed, "Advanced fault-tolerant control strategy of wind turbine based on squirrel cage induction generator with rotor bar defects," in *Annual Conference of the PHM Society*, 2019.

- [2] M. F. Khan, M. R. Khan, and A. Iqbal, "Effects of induction machine parameters on its performance as a standalone self excited induction generator," *Energy Reports*, vol. 8, pp. 2302–2313, 2022.
- [3] V. M. Krishna, V. Sandeep, B. Narendra, and K. Prasad, "Experimental study on self-excited induction generator for small-scale isolated rural electricity applications," *Results in Engineering*, vol. 18, p. 101182, 2023.
- [4] V. Sandeep and others, "Experimental study on different modes of self excited induction generator operation," in *2020 2nd PhD Colloquium on Ethically Driven Innovation and Technology for Society (PhD EDITS)*, IEEE, 2020, pp. 1–2.
- [5] A. A. Khamis and S. Ahmed-Zaid, "Computer simulation of a series-connected induction generator and determination of minimum required capacitance for self-excitation for wind energy applications," in *2020 11th International Renewable Energy Congress (IREC)*, IEEE, 2020, pp. 1–6.
- [6] G. Ofualagba, E. Ubeku, and others, "The analysis and modelling of a self-excited induction generator driven by a variable speed wind turbine," *Fundamental and Advanced Topics in Wind Power*, pp. 5–9, 2011.
- [7] T. Tudorache, L. Melcescu, and S. Paturca, "Finite element analysis of self-excited induction generator for isolated small power wind turbines," in *2007 International Conference on Clean Electrical Power*, IEEE, 2007, pp. 656–661.
- [8] A. Kumari, A. Thosar, and S. Mopari, "Determination of excitation capacitance of a three phase self excited induction generator," *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, vol. 4, no. 5, pp. 4069–4074, 2015.
- [9] S. Bhandari, "Mechanical fault analyses and detection using non stationary decomposition for vibration signals," PhD Thesis, 2023.
- [10] J. Faiz, B. Ebrahimi, and M. Sharifian, "Different faults and their diagnosis techniques in three-phase squirrel-cage induction motors—A review," *Electromagnetics*, vol. 26, no. 7, pp. 543–569, 2006.
- [11] M. Subrata, C. Karmakar, S. Mitra, *Induction motor fault diagnosis: Approach Through Current Signature Analysis*. SPRINGER, 2018.
- [12] M. R. Mehrjou, N. Mariun, M. H. Marhaban, and N. Misron, "Rotor fault condition monitoring techniques for squirrel-cage induction machine—A review," *Mechanical Systems and Signal Processing*, vol. 25, no. 8, pp. 2827–2848, 2011.
- [13] R. Chawla, B. A. Vinayak, and G. Jagadanand, "Modelling and detection of stator incipient open circuit fault in three-phase induction motor," in *2018 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES)*, IEEE, 2018, pp. 1–6.

Enhanced Control of Rotary Drilling Systems Using FOPID

Sabrina MEDDAH
Laboratory of Applied Automation
Department of Automation, FHC,
Univesity M'hamed Bougara
Boumerdes, Algeria
s.meddah@univ-boumerdes.dz

Sid Ahmed TADJER
Laboratory for the electrification of
industrial enterprises
Department of Automation, FHC,
Univesity M'hamed Bougara
Boumerdes, Algeria
s.tadjer@univ-boumerdes.dz

Mohamed Zinelabidine DOGHMANE
Laboratory of Applied Automation
FHC, UMBB, Algeria
Bob L. Herd Department of Petroleum
Engineering Texas Tech University,
Lubbock, TX, USA
m.doghmane@univ-boumerdes.dz

Madjid KIDOUCHE
Laboratory of Applied Automation
Department of Automation, FHC,
Univesity M'hamed Bougara
Boumerdes, Algeria
mkidouche@univ-boumerdes.dz

Abdelhakim IDIR
Laboratory of Applied Automation
Electrical Engineering Department,
University Mohamed Boudiaf
Msila, Algeria
abdelhakim.idir@univ-msila.dz

Abstract—Rotary drilling systems are strongly affected by the nonlinear interaction between axial and torsional vibrations, which reduces drilling efficiency and can cause stick-slip and bit bounce. To address these challenges, this work investigates the application of a Fractional-Order PID (FOPID) controller for the mitigation of coupled axial-torsional vibrations. A four-degree-of-freedom (4-DOF) drill string model is used to represent the coupled dynamics and evaluate the control strategy under varying drilling conditions. The main contributions of this study include demonstrating the robustness and stability of the FOPID controller and showing its effectiveness in reducing coupled vibrations. Simulation results indicate that the FOPID controller significantly improves vibration suppression and overall drilling performance.

Keywords— Axial-torsional coupled vibrations, 4-DOF drillstring model, FOPID controller, control robustness

I. INTRODUCTION

In today's world, the main sources of energy are oil and gas, which are predominantly extracted from subsurface reservoirs through drilling [1]. This critical process relies heavily on rotary drilling systems, comprising a rotating element driven by a top drive, connected to the bottom hole drill bit through a drill string. Broadly, the rotary drilling system consists of two primary sections: the surface-based drilling rig, which includes a tower (derrick or mast) equipped with lifting and floor equipment for rod handling, screwing and unscrewing, and drill bit replacement [2, 3]; and the subsurface equipment, forming the drill string assembly, which can be divided into two main segments—the upper segment composed of drill pipes and the lower segment referred to as the Bottom Hole Assembly (BHA), made up of thicker pipes.

A significant number of researchers have focused on the various types of vibrations experienced during rotary drilling operations [4–6]. Axial vibrations, detected along the drilling axis, can severely affect the Rate of Penetration (ROP) and, in extreme cases, cause bit-bounce, where irregular contact occurs between the drill bit and the formation. Torsional vibrations, arising from the drill string's elasticity and the highly nonlinear torque-bit interaction, can lead to the stick-slip phenomenon, reducing drilling efficiency and increasing operational costs [7]. Lateral vibrations may also excite axial vibrations, further complicating the drillstring dynamics.

Although prior studies have addressed vibration mitigation and proposed various controllers, they often overlook the interaction between axial and torsional vibrations, which significantly influences controller performance [8,9]. In particular, the robustness of controllers under strongly coupled vibration conditions has not been thoroughly investigated.

The objective of this research is to analyze the interaction between axial and torsional vibrations using a four-degree-of-freedom (4-DOF) drillstring model and, based on this analysis, design and implement a Fractional-Order PID (FOPID) controller. The study focuses on mitigating coupled vibrations, enhancing controller robustness, and evaluating dynamic performance under varying drilling conditions.

The remainder of this manuscript is organized as follows: Section 2 provides a comprehensive review of rotary drilling systems, emphasizing the coupled Axial -Torsional vibrations. Section 3 presents the development of the 4-DOF axial-torsional drillstring model. Section 4 details the design and implementation of the FOPID controller based on the insights from the model. Section 5 presents simulation results under various drilling scenarios to evaluate the effectiveness of the FOPID controller. Finally, Section 6 concludes the study and discusses potential directions for future research in robust vibration control for rotary drilling systems.

II. ROTARY DRILLING SYSTEMS

A. System Description

The rotary drilling system is a well-known method in the petroleum sector that is mostly used for boring holes from the surface to specific underground reserves. The way this system works is by making the drilling tool rotate, which is carefully controlled by the top driving mechanism [6]. A composite set of drill pipes known as the drill string facilitates the connection between the drill bit and the Top drive. Three essential tasks are necessary for this device to operate properly [3].

The two essential parts of the rotary drilling system are the drill string, which works below the surface of the earth, and the drilling rig, which is positioned prominently on the surface of the planet [10]. The drilling rig consists of a tall building called a derrick or mast that holds necessary floor and lifting apparatuses for manipulating, tightening, and loosening rods and changing out drilling gear. On the other hand, the

underground portion of the drilling machinery is covered by the drill string. It consists of the top part, which is made up of many rods, and the bottom part, which is called the Bottom Hole Assembly (BHA). Important components such the drill collar, stabilizers, and drilling tool are included in the BHA [11]. Three different types of vibrations can occur simultaneously or one at a time during drilling operations, and each one has the potential to seriously harm and malfunction system components. Axial, lateral, and torsional vibrations are the three main types of these vibrations that may be classified according to the directions in which they propagate [12]. It is important to remember that several kinds of vibrations can coexist and produce linked modes, including lateral-axial-torsional, axial-torsional, or torsional-lateral. The coupled axial-torsional vibration mode is given special attention in this work because of its frequency and possible negative effects on rotary drilling systems [13].

B. Axial-Torsional Vibrations

A common occurrence during drilling operations is the coupling phenomena between axial and torsional vibrations, which is directly related to the transformation of WOB sent to the drill bit through the drill-string. This coupling mode may produce the stick-slip phenomena, which in turn may result in spinning, the generation of axial vibration waves that move in the direction of the surface [14]. In addition, the drill bit may be released from the stick phase, causing the slip phase to begin and ending the cycle in a repeating manner, upon the arrival of a reflected axial vibration wave from the surface or a contact point inside the drill string. After a careful examination of these interactions, the coupled axial-torsional vibration mode is the main subject of this work because of its frequent occurrence and significant influence on the equipment integrity and rate ROP. The goal of this project is to conduct a thorough analysis of the coupled phenomena involving the stick-slip and the bit-bounce [12,15].

III. DRILL STRING MODEL

In our research work we used a coupled nonlinear model for torsional and axial vibrations in a rotary drilling system, considering high-frequency nonlinearities and rock formation interaction [16-18]. The model includes a 3-degree-of-freedom mass-damping system, encompassing both axial and torsional dynamics.

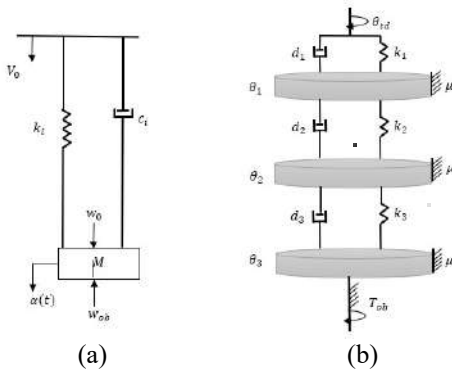


Fig. 1. Rotary drilling system equivalent dynamic for: a) Axial Vibrations ,b) Torsional Vibrations[16].

In this case, torsional and axial vibrations are shown in (a) and (b), respectively [16]. A drillstring with an axial displacement U , a torsional angle θ , and a bending angle α make up the system. The system's equations of motion can be obtained as follows (1)[16]:

(1)

And the mathematical expressions for the WOB and the TOB are provided in equations (2) and (3), respectively[16].

(2)

(3)

The system parameters are displayed in the table below:

TABLE I. LIST OF PARAMETERS

Para.	Description	Unit
θ_{td}	The angle of the top drive	[rad]
θ_i	The angle of joint i	[rad]
k_i	Coefficient of rigidity of the rope section i	[N.m/rad]
d_i	Internal damping coefficient of the rope section i	[N.m.s/rad]
μ_b	Coefficient of friction of the drill	[N.m]
μ	Wall friction coefficient for rope sections	[N.m]
j_i	The inertia of the rope section i	[kg.m ²]
V_0	Constant axial velocity	[m /s]
c_i	The axial damping	[N.m.s/rad]
k_i	The axial stiffness	[N/m]
M	The mass of drill string including BHA	[tones]
w_0	The nominal drilling pressure	[Psi]
w_{ob}	Force applied on the drill bit	[kN/m ²]
T_{ob}	Torque on bit	[lb.ft]
ROP	Rate of penetration	[m /hr]

The following variable alterations are made in (1) to convert the states and inputs to standard notation: ,, ,,,,. Equations provides the whole system's equations of state[16]:

(4)

IV. FOPID FOR AXIAL TORSIONAL TITO SYSTEM

The TITO (Two Input - Two Output) system under consideration can be represented in a state-space form, encapsulating its dynamic behavior with inputs and outputs. Let's denote the state vector as x , input vector as u , and output vector as y . The dynamics can be expressed as:

(5)

Where:

- A is the state matrix governing the evolution of the state variables x . contains the linear terms related to stiffness k_i , damping d_i , inertia j_i , and couplings.
- B represents the system inputs, including the top-drive torque and the controller action.
- C is maps the states to the measured outputs (axial and torsional velocities).
- D gathers the nonlinearities arising from WOB, TOB, bit-rock friction, and axial-torsional interactions.

The FOPID controller introduces fractional-order derivatives and integrals for TITO system to enhance control performance in handling complex dynamics and disturbances [19,20]. The controller is structured as follows:

(6)

Here:

- denotes the error vector, where is the desired state vector and is the actual state vector,
- represents the controller's input, adjusted to minimize the error
- , and are the proportional, integral (of fractional order), and derivative (of fractional order) gains, respectively,
- is the state matrix mapping the error vector to the control action.

The transfer function of the FOPID controller provides insights into its frequency response and stability characteristics [17,21]:

7)

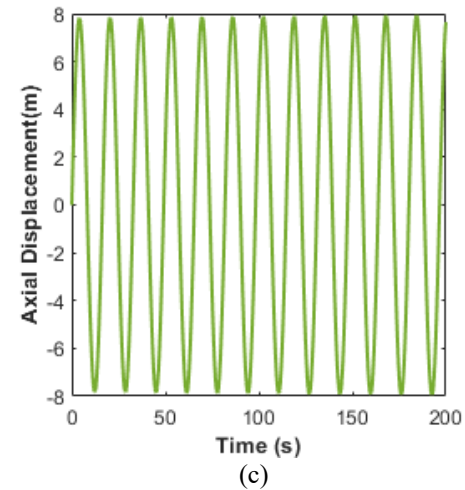
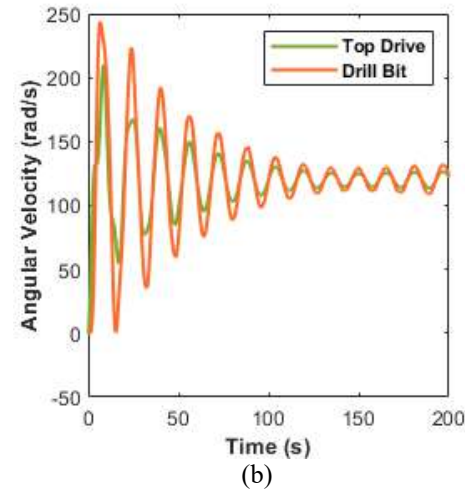
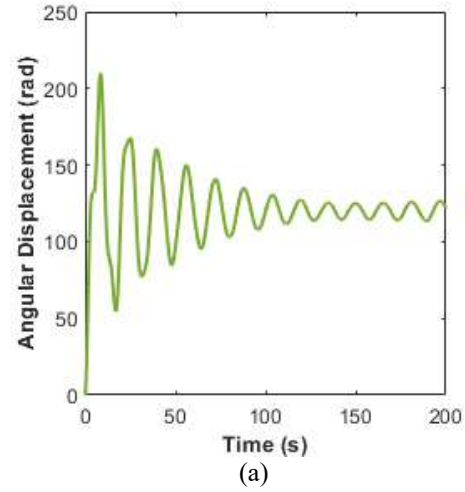
Where:

- represents the transfer function matrix of the controller,
- and denote the Laplace transforms of the output and error, respectively,
- is the complex frequency domain variable (Laplace transform variable),
- and are the orders of the integral and derivative gains, respectively.
- The FOPID parameters were chosen based on frequency-response analysis, stability margin constraints, and transient-response requirements. The influence of each gain has been clarified:
- : improves response speed but may increase oscillations,
- , γ : reduce steady-state error and improve robustness,
- , ρ : enhance damping of high-frequency torsional and axial oscillations.

V. RESULTS AND DISCUSSION

A. Open loop responses

We keep the reference velocities for axial motion at 3 m/hr and angular motion at 120 rad/s. Figure 2 shows the outcomes of this configuration. Figures 2(a) and 2(b) depict the rotational displacement and angular velocity at the top drive and drill bit levels, respectively; Figure 2(c) depicts the axial displacement at the top drive, and Figure 2(d) depicts the axial velocity of the drill bit and top drive. These statistics show the degree of torsional and axial vibrations, with torsional vibrations decreasing over time while axial vibrations are still considerable. In addition, the time needed for the torsional vibrations to dissipate is much longer. Therefore, it is essential to put in place a strong controller that can reduce axial vibrations while also accelerating the attenuation of torsional vibrations.



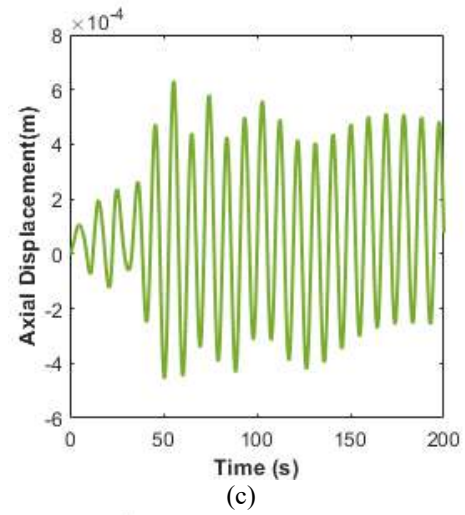
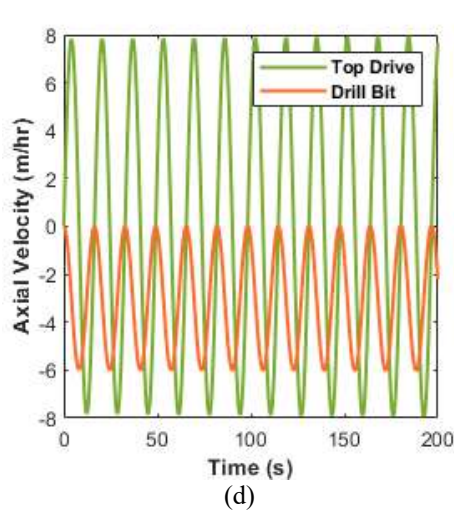


Fig. 2. Results for open loop : a) Torsional displacement , b) Torsional velocity , c) Axial displacement , d) Axial velocity

B. Closed loop responses

We keep the axial velocity at 3 m/hr and the angular velocity reference at 120 rad/s. The closed-loop system is modeled by a FOPID controller. It is clear from the data shown in Figures 3(b) and (d) that the FOPID controller has successfully reduced axial and torsional oscillations. The axial vibrations have also diminished, in addition to the large reduction in torsional vibrations.

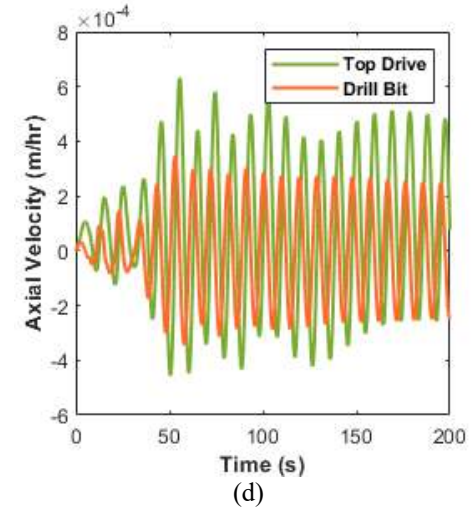
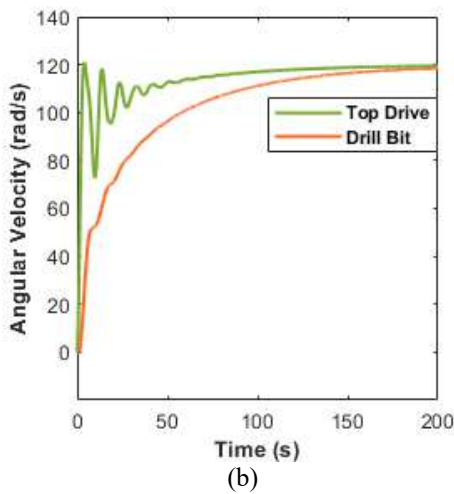
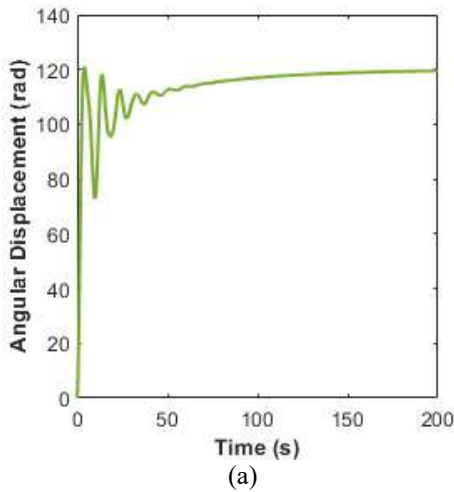


Fig. 3. Results for closed loop FOPID : a) Axial displacement, b) Axial velocity, c) Torsional displacement, d) Torsional velocity

VI. CONCLUSION

In this work, a 4-DOF coupled axial–torsional model of a rotary drilling system was developed to study the interactions between axial and torsional vibrations. A fractional-order PID (FOPID) controller was implemented to mitigate these coupled vibrations, and its performance was evaluated through numerical simulations based on real drilling data from Algerian operations. The results demonstrate that the proposed FOPID controller effectively reduces both stick-slip and axial oscillations, improving system stability and overall drilling efficiency.

Limitations and Future Work

Despite these promising results, several limitations remain:

- The FOPID controller parameters are not optimized automatically, as no algorithms such as PSO or GA were employed.
- The 4-DOF model, although robust, remains an approximation of a real distributed system.
- Validation is currently limited to numerical simulations; Hardware-in-the-Loop (HIL) testing or experiments on an industrial platform are required for further verification.

Addressing these limitations represents a natural direction for future research. Automatic optimization of FOPID parameters, more accurate modeling of the drilling system, and experimental validation on industrial setups will further

enhance the applicability and reliability of the proposed control strategy.

REFERENCES

- [1] J. Tian, Y. Liu, J. Xiong, *et al.*, "Analysis of drill string's vibration characteristics based on sliding mode-PI controller," *J. Vib. Eng. Technol.*, vol. 10, pp. 919–927, 2022, doi: 10.1007/s42417-021-00419-2.
- [2] C. Mendil, M. Kidouche, and M. Z. Doghmane, "Sliding mode controller design for torsional vibrations minimization under rock-bit interaction effects," in *Artificial Intelligence and Heuristics for Smart Energy Efficiency in Smart Cities*, IC-AIRES 2021, LNNS, vol. 361, Cham, Switzerland: Springer, 2022, pp. —, doi: 10.1007/978-3-030-92038-8_61.
- [3] M. Magdalene Marinaki, Y. Marinakis, and E. G. Stavroulakis, "Fuzzy control optimized by PSO for vibration suppression of beams," *Control Eng. Pract.*, vol. 18, no. 6, pp. 618–629, 2010, doi: 10.1016/j.conengprac.2010.03.001.
- [4] C. Mendil, M. Kidouche, and M. Z. Doghmane, "Hybrid backstepping sliding mode controller for stick-slip vibrations mitigation in rotary drilling systems," *IETE J. Res.*, vol. 69, no. 6, pp. 3477–3487, 2023, doi: 10.1080/03772063.2021.1919217.
- [5] F. Zribi, L. Sidhom, A. Krama, and M. Gharib, "Enhancement of drill string system operations with adaptive robust controller and hardware-in-the-loop validation," *J. Sound Vib.*, vol. 556, 2023, Art. no. 117716, doi: 10.1016/j.jsv.2023.117716.
- [6] C. Mendil, M. Kidouche, and M. Z. Doghmane, "Hybrid sliding PID controller for torsional vibrations mitigation in rotary drilling systems," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 22, no. 1, pp. 146–158, Apr. 2021, doi: 10.11591/ijeecs.v22.i1.pp146-158.
- [7] C. Mendil, M. Kidouche, M. Z. Doghmane, S. Benammar, and K. F. Tee, "Rock-bit interaction effects on high-frequency stick-slip vibration severity in rotary drilling systems," *Multidiscip. Model. Mater. Struct.*, vol. 17, no. 5, pp. 1007–1023, 2021, doi: 10.1108/MMMS-10-2020-0256.
- [8] A. Laib, B. Talbi, A. Krama, and M. Gharib, "Hybrid interval type-2 fuzzy PID+I controller for a multi-DOF oilwell drill-string system," *IEEE Access*, vol. 10, pp. 67262–67275, 2022, doi: 10.1109/ACCESS.2022.3185021.
- [9] M. Z. Doghmane, M. Kidouche, and A. Ahriche, "Decentralized overlapping control design with application to rotary drilling system," *IETE J. Res.*, vol. 69, no. 4, pp. 2070–2079, 2023, doi: 10.1080/03772063.2021.1886602.
- [10] I. Kessai, M. Z. Doghmane, and S. Benammar, "Weight-on-bit variations analyses in the drillstring of rotary drilling systems under stick-slip vibrations," in *Artificial Intelligence and Heuristics for Smart Energy Efficiency in Smart Cities*, IC-AIRES 2021, LNNS, vol. 361, Springer, 2022, doi: 10.1007/978-3-030-92038-8_72.
- [11] S. Meddah, S. A. Tadjer, M. Kidouche, M. Z. Doghmane, and A. Idir, "Axial-torsional vibrations interaction analysis and decoupling in drill string systems," in *Proc. 2nd Int. Conf. Electr. Eng. Autom. Control (ICEEAC)*, May 2024, pp. 1–7.
- [12] R. Riane, M. Z. Doghmane, M. Kidouche, and S. Djeddar, "Observer-based H ∞ controller design for high-frequency stick-slip vibrations mitigation in drill-string of rotary drilling systems," *Vibration*, vol. 5, pp. 264–289, 2022, doi: 10.3390/vibration5020016.
- [13] S. K. Gupta and P. Wahi, "Global axial-torsional dynamics during rotary drilling," *J. Sound Vib.*, vol. 375, pp. 332–352, 2016, doi: 10.1016/j.jsv.2016.04.021.
- [14] S. Tashakori, G. Vossoughi, H. Zohoor, and N. van de Wouw, "Prediction-based control for mitigation of axial-torsional vibrations in a distributed drill-string system," *IEEE Trans. Control Syst. Technol.*, vol. 30, no. 1, pp. 277–293, 2021, doi: 10.1109/TCST.2021.3065669.
- [15] B. J. Kim, A. Palazzolo, and M. Gharib, "Drillstring simulator: A novel software model for stick-slip and bit-bounce vibrations," in *ASME Int. Mech. Eng. Congr. Expo.*, 2021, Paper IMECE2021-73717, doi: 10.1115/IMECE2021-73717..
- [16] S. Meddah, M. Z. Doghmane, S. A. Tadjer, A. Idir, M. Kidouche, and K. F. Tee, "Fractional-order PID controller design for strongly coupled high-frequency axial-torsional vibrations in drill string system," *Int. J. Dyn. Control*, vol. 13, no. 2, p. 94, 2025.
- [17] S. Meddah, A. Idir, S. A. Tadjer, M. Z. Doghmane, and M. Kidouche, "PID control design of strongly coupled axial-torsional vibrations in rotary drilling systems," *Przegląd Elektrotechniczny*, vol. 100, 2024.
- [18] S. Meddah, S. A. Tadjer, A. Idir, T. Kong, M. Z. Doghmane, and M. Kidouche, "Dynamic interaction analysis of coupled axial-torsional-lateral mechanical vibrations in rotary drilling systems," *Struct. Durability Health Monit.*, vol. 19, no. 1, pp. 77—, 2025.
- [19] A. K. Jha and S. S. Dasgupta, "Fractional order PID based optimal control for fractionally damped nonlocal nanobeam via genetic algorithm," *Microsyst. Technol.*, vol. 25, pp. 4291–4302, 2019, doi: 10.1007/s00542-019-04402-6.
- [20] M. Derbal, M. Gharib, S. S. Refaat, A. Palazzolo, and S. Sassi, "Fractional-order controllers for stick-slip vibration mitigation in oil well drill-strings," *J. Low Freq. Noise Vib. Active Control*, vol. 40, no. 3, pp. 1571–1584, 2021, doi: 10.1177/1461348420984040.
- [21] L.-Y. Chang and H.-C. Chen, "Tuning of fractional PID controllers using adaptive genetic algorithm for active magnetic bearing system," *WSEAS Trans. Syst.*, vol. 8, no. 1, pp. 158–167, 2009.

Grey Wolf Optimizer-Based PID Control for Suppressing Stick-Slip Vibrations in Drilling Systems

TELLAA Mawloud
*Department of Automation and
 Electrification of Process
 Applied Automation
 Laboratory, Faculty of Hydrocarbons
 and Chemistry
 University M'Hamed Bougara-
 Boumerdes,
 Boumerdes 35000, Algeria
 m.tellaa@univ-boumerdes.dz*

AIBECHÉ Abderrezak
*Department of Automation and
 Electrification of Process
 Applied Automation
 Laboratory, Faculty of Hydrocarbons
 and Chemistry
 University M'Hamed Bougara-
 Boumerdes,
 Boumerdes 35000, Algeria
 aibeche_umbb@univ-boumerdes.dz*

DOGHMANE Mohamed Zinelabidine
*Department of Automation and
 Electrification of Process
 Applied Automation
 Laboratory, Faculty of Hydrocarbons
 and Chemistry
 University M'Hamed Bougara-
 Boumerdes,
 Boumerdes 35000, Algeria
 Bob L. Herd Department of Petroleum
 Engineering, Texas Tech University,
 Lubbock 79401, Texas, USA
 doghmane_m@yahoo.com*

KIDOUCHÉ Madjid
*Department of Automation and
 Electrification of Process
 Applied Automation
 Laboratory, Faculty of Hydrocarbons
 and Chemistry
 University M'Hamed Bougara-
 Boumerdes,
 Boumerdes 35000, Algeria
 mkidouche@univ-boumerdes.dz*

BOUDOUDA Aimad
*Laboratoire Ingénierie des Systèmes et
 Télécommunications
 University M'Hamed Bougara-
 Boumerdes,
 Boumerdes 35000, Algeria
 a.boudouda@uinv-boumerdes.dz*

AKROUM hamza
*Department of Automation and
 Electrification of Process
 Applied Automation
 Laboratory, Faculty of Hydrocarbons
 and Chemistry
 University M'Hamed Bougara-
 Boumerdes,
 Boumerdes 35000, Algeria
 akroum@gmail.com*

Abstract— Stick-slip vibration is a major challenge in rotary drilling systems, causing severe drillstring oscillations that reduce drilling efficiency, increase tool wear, and raise operational risks. This paper develops a three-degree-of-freedom nonlinear model that captures drillstring torsional dynamics and bit-rock dry friction. To mitigate stick-slip, a PID controller optimized using the Grey Wolf Optimizer (GWO) is proposed. The GWO-PID scheme adaptively tunes controller gains to stabilize the system and reduce torque fluctuations. Numerical simulations under various drilling conditions show that the proposed method significantly suppresses torsional vibrations and improves drilling smoothness. Results also indicate that GWO-PID provides faster convergence and better tuning performance than classical approaches, offering a robust solution for controlling stick-slip in rotary drilling systems.

Keywords— Stick-slip vibration, rotary drilling system, PID controller, Grey Wolf Optimizer (GWO), vibration suppression

I. INTRODUCTION

Rotary drilling is a fundamental technique in oil and gas extraction. However, the process often induces dynamic instabilities that affect drilling efficiency and safety. Among these, three primary types of vibrations—axial, lateral, and torsional—are typically observed [1]. The most destructive of these is torsional stick-slip vibration, which can excite other modes of vibration, leading to severe structural fatigue, premature bit failure, and decreased drilling performance. Over the past decade, several control strategies have been

proposed to suppress stick-slip oscillations [2,3]. Tang et al. [4] showed that reducing the weight on bit (WOB) could mitigate stick-slip, although with delayed response. Similarly, increasing the rotary table speed helps reduce torsional vibration but neglects the effect of drillstring stiffness [5]. Sliding mode control approaches applied to multi-degree-of-freedom models [6,7] have demonstrated robustness but often suffer from long stabilization times. Experimental investigations by Liu [8] indicated that optimized rotational velocity can reduce vibration amplitude. Other techniques, such as backstepping control [9] and hybrid control schemes [9,10], have also been proposed. While these methods improve stability, their robustness and adaptability under variable downhole conditions remain limited. Recent developments in intelligent optimization algorithms have enabled more effective controller tuning. Among these methods, the Grey Wolf Optimizer (GWO) has gained attention due to its balance between exploration and exploitation and its ability to converge rapidly to global optima. In this paper, a GWO-based PID controller is developed to actively mitigate stick-slip vibrations in rotary drilling systems. The controller dynamically adjusts the PID parameters to achieve optimal vibration suppression and enhanced drilling performance. The proposed approach is validated through numerical simulations and compared against conventional PID tuning results. The remainder of this paper is organized as follows: Section 2 describes the rotary drilling system and the stick-slip phenomenon. Section 3

presents the mathematical modeling of the 3-DOF drilling dynamics. Section 4 outlines the design and implementation of the GWO-based PID controller. Section 5 discusses the simulation results and performance comparison. Finally, Section 6 concludes the study and highlights directions for future work.

II. DRILLING RIG SUBSYSTEMS AND STICK-SLIP VIBRATIONS

A rotary drilling rig comprises several interconnected subsystems—power, hoisting, rotary, circulation, well control, and monitoring. The hoisting system supports drilling operations, the rotary system provides the rotational motion needed for rock cutting, and the circulation system removes cuttings, stabilizes the wellbore, and cools the drill bit. Well control maintains pressure balance during abnormal conditions, while the power and monitoring systems ensure continuous and safe rig operation [11]. Stick-slip vibrations are a major dysfunction affecting this integrated system. They arise from alternating sticking and slipping phases at the bit-rock interface, driven by drillstring elasticity, frictional torque, and variations in WOB and rotational speed. This nonlinear behavior leads to severe fluctuations in torque and rotational speed, causing mechanical stresses, bit stalling, tool damage, and reduced rate of penetration. Therefore, effective suppression of stick-slip vibrations is essential for safe and efficient drilling.

III. MATHEMATICAL MODEL UNDER STICK-SLIP VIBRATIONS

A. Drilling string model

This study utilizes a mass-spring-damper model to simulate the behavior of a rotary drilling system [12]. The model consists of three degrees of freedom, with the upper disc representing the rotary table, the middle disc representing the drill string, and the lower disc representing the drilling bit with the BHA (Bottom Hole Assembly). The specific parameters of the model are outlined in Table.1. Additionally, Figure 1 (a) provides a schematic diagram of the rotary drilling system, while Figure 1 (b) illustrates the 3-DOF model. This model is mathematically described by equation (1).

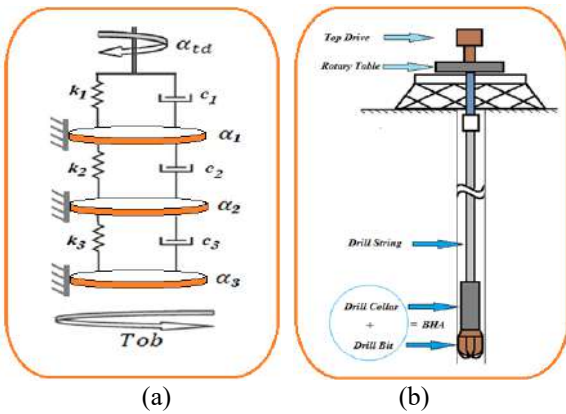


Fig. 1. Rotary drilling systems: (a) Schematic diagram, (b) Equivalent mass-spring-damper three-element model [11, 12].

$$\begin{cases} \ddot{\alpha}_1 = \frac{1}{J_1} (c_1(\dot{\alpha}_{td} - \dot{\alpha}_1) - k_1(\alpha_{td} - \alpha_1) - c_2(\dot{\alpha}_1 - \dot{\alpha}_2) - k_2(\alpha_1 - \alpha_2) - \mu\dot{\alpha}_1) \\ \ddot{\alpha}_2 = \frac{1}{J_2} (c_2(\dot{\alpha}_1 - \dot{\alpha}_2) + k_2(\alpha_1 - \alpha_2) - c_3(\dot{\alpha}_2 - \dot{\alpha}_3) - k_3(\alpha_2 - \alpha_3) - \mu\dot{\alpha}_2) \\ \ddot{\alpha}_3 = \frac{1}{J_3} (c_3(\dot{\alpha}_2 - \dot{\alpha}_3) + k_3(\alpha_2 - \alpha_3) - \mu\dot{\alpha}_3 - Tob) \end{cases} \quad (1)$$

Where the values of the system parameters are: $k_1 = k_2 = k_3 = 481.29 \text{ N.m/rad}$, $J_1 = J_2 = 1030.45 \text{ Kg.m}^2$, $J_3 = 223.44$, $c_1 = c_2 = 51.38 \text{ N.m.s/rad}$, $c_3 = 39.79 \text{ N.m.s/rad}$, $\mu = 10 \text{ N.m}$

To express the system in state-space form, we introduce the following change of variables. The input signal is defined as the angular velocity of the top drive, $u = \dot{\alpha}_{td}$. The state variables are selected as $x_1 = \dot{\alpha}_1$, $x_2 = \dot{\alpha}_2$, $x_3 = \dot{\alpha}_3$, $x_4 = \alpha_{td} - \alpha_1$, $x_5 = \alpha_1 - \alpha_2$, $x_6 = \alpha_2 - \alpha_3$.

With these definitions, Equation (1) can be rewritten in the following state-space form:

$$\begin{cases} \dot{x}_1 = -\frac{c_2+c_1+\mu}{J_1}x_1 + \frac{c_2}{J_1}x_2 + \frac{k_1}{J_1}x_4 - \frac{k_2}{J_1}x_5 + \frac{c_1}{J_1}u \\ \dot{x}_2 = \frac{c_2}{J_2}x_1 - \frac{c_2+c_3+\mu}{J_2}x_2 + \frac{c_3}{J_2}x_3 + \frac{k_2}{J_2}x_5 - \frac{k_3}{J_2}x_6 \\ \dot{x}_3 = \frac{c_3}{J_3}x_2 - \frac{c_3+\mu}{J_3}x_3 + \frac{k_3}{J_3}x_6 - \frac{1}{J_3}Tob \\ \dot{x}_4 = \dot{\alpha}_{td} - \dot{\alpha}_1 = u - x_1 \\ \dot{x}_5 = \dot{\alpha}_1 - \dot{\alpha}_2 = x_1 - x_2 \\ \dot{x}_6 = \dot{\alpha}_2 - \dot{\alpha}_3 = x_2 - x_3 \end{cases} \quad (2)$$

TABLE I. PARAMETERS DESCRIPTION FOR THE ROTARY DRILLING MODEL [11, 12].

Parameter	Description	Unit
α_{td}	The top drive angular displacement	[rad]
$\alpha_{i=1,2,3}$	The angular displacements of the rope section i	[rad]
$k_{i=1,2,3}$	Torsional stiffness coefficient of the rope section i	[N.m/rad]
$c_{i=1,2,3}$	Internal damping coefficient of the rope section i	[N.m.s/rad]
μ	Wall friction coefficient	[N.m]
$J_{i=1,2,3}$	The inertia of the rope section i	[kg.m ²]
Tob	The torque on bit	[N.m]

B. Rock-bit interaction model

In literature, many functions have been utilized to represent the interaction between the bit and the rock. For the system being studied, we will employ the generalized model described by the equation [12].

$$Tob = \mu_n N r \left(\frac{x_3}{\sqrt{x_3^2 + \Omega_0^2}} + \frac{p\Omega_0 x_3}{x_3^2 + \Omega_0^2} \right) + D x_3 \left(\frac{x_3}{\Omega_1} - 1 \right) \quad (3)$$

Where the model coefficients are given in the following table:

TABLE II. PARAMETERS USED IN ROCK-BIT GENERAL MODEL [16]

parameter	description	value
μ_n	friction coefficient	28 nm
n	the force vector	9.81 * wob
r	the contact radius vector	0.1 m
Ω_0	chain transition speed	1
Ω_1	transition speed for the well	31.4159
p	the initial friction parameter	1.5
d	the linear damping vector	0.28

C. PID controller

Proportional integral derivative (PID) control is the most common control rule for regulation in engineering. It is used in many fields because it is easy to tune and always works. The PID control equation is usually written as [9].

$$C(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau + K_d \frac{de(t)}{dt} \quad (4)$$

Where K_p is the proportional coefficient; K_i is the integral time constant; K_d is the differential coefficient. The derivative part is filtered in order to avoid singularities; the equation (4) is for conventional PID before calling the GWO algorithm for parameters optimization [16].

IV. GREY WOLF OPTIMIZATION (GWO)

The Grey Wolf Optimization (GWO) algorithm is inspired by the leadership hierarchy and hunting strategy of grey wolves in nature. The behavior of the pack during hunting is governed by a clear social structure. The leader of the pack, called the alpha, directs and controls the other wolves, guiding them toward the most efficient path to capture the prey [13].

A. GWO-Based PID Tuning (GWO-PID)

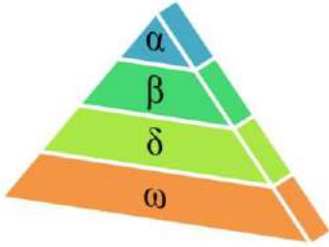


Fig. 2. Hierarchical order of Grey Wolves Pack.

The hierarchical order of the grey wolf pack consists of four levels: alpha, beta, delta, and omega. The alpha wolf is the main leader responsible for decision-making and determining the best hunting path. The beta wolf assists the alpha and helps manage the pack. The delta and omega wolves act as subordinates that follow the guidance of higher-ranking members. The GWO algorithm effectively explores the search space and avoids premature convergence. This makes it suitable for determining the optimal tuning parameters of a PID controller with high precision [13].

B. Flowchart of the GWO Algorithm

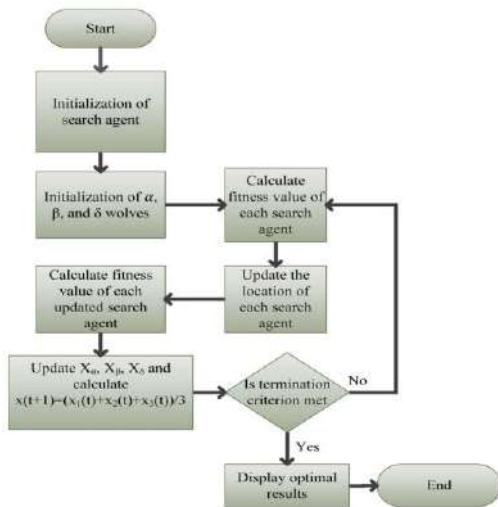


Fig. 3. GWO process flow diagram.

The working process of the GWO algorithm can be summarized as follows. First, the initial population of wolves and algorithm parameters are generated. Each wolf represents a potential solution. The positions of the wolves are updated based on their distance from the prey, which represents the optimal solution. The process is repeated iteratively until the termination condition is reached, at which point the best solution is obtained [13].

C. Mathematical Model of GWO

The encircling and hunting behaviors of grey wolves can be described mathematically to optimize the PID controller parameters.

The distance between a wolf and its prey is calculated using the expression:

$$\vec{D} = |\vec{C} \times \vec{X}_p(t) - \vec{X}(t)| \quad (5)$$

The new position of a wolf is updated according to:

$$\vec{X}(t+1) = \vec{X}_p(t) - \vec{A} \times \vec{D} \quad (6)$$

Where t indicates the current iteration, $\vec{A}$ and $\vec{C}$ are coefficient vectors, $\vec{X}_p$ is the position vector of the prey and $\vec{X}$ indicates the position vector of grey wolf

The vectors $\vec{A}$ and $\vec{C}$ are given by:

$$\vec{A} = 2\vec{a} \times \vec{r}_1 - \vec{a} \quad (7)$$

$$\vec{C} = 2 \times \vec{r}_2 \quad (8)$$

Here, a is a coefficient that decreases linearly from 2 to 0 as iterations progress, and $\vec{r}_1$ and $\vec{r}_2$ are random vectors in the range $[0, 1]$.

After encircling the prey, the alpha wolf identifies the best position, while the beta and delta wolves assist in refining the search. These three wolves guide the rest of the pack (including the omega wolves) to move toward the optimal positions. The updated positions of the wolves are determined by the following relationships:

$$\vec{D}_\alpha = |\vec{C}_1 \times \vec{X}_\alpha - \vec{X}| \quad (9)$$

$$\vec{D}_\beta = |\vec{C}_2 \times \vec{X}_\beta - \vec{X}| \quad (10)$$

$$\vec{D}_\delta = |\vec{C}_3 \times \vec{X}_\delta - \vec{X}| \quad (11)$$

$$\vec{X}_1 = \vec{X}_\alpha - \vec{A}_1 \times \vec{D}_\alpha \quad (12)$$

$$\vec{X}_2 = \vec{X}_\beta - \vec{A}_2 \times \vec{D}_\beta \quad (13)$$

$$\vec{X}_3 = \vec{X}_\delta - \vec{A}_3 \times \vec{D}_\delta \quad (14)$$

Finally, the new position of each wolf is updated as the average of the three best solutions:

$$\vec{X}(t+1) = \frac{(\vec{X}_1 + \vec{X}_2 + \vec{X}_3)}{3} \quad (15)$$

Nomenclature utilized is explained which is used in expression in below:

TABLE III. PARAMETERS USED IN ROCK-BIT GENERAL MODEL

$\vec{D}_a$	Distance between prey and wolf
$\vec{C}, \vec{A}$	Coefficient vectors
$\vec{X}_p(t)$	Current position of the prey
$\vec{X}(t)$	Current position of the wolf
T	Current number of iterations
$\vec{a}$	Linearly decreasing coefficient vector (from 2 to 0)
$\vec{r}_1, \vec{r}_2$	Random vectors in range $([1, 0])$
$X_\alpha, X_\beta, X_\delta$	Current position of α, β , and δ wolves
X_1, X_2, X_3	Updated locations of wolves
$X(t+1)$	Mean of the updated positions

D. Fitness function

To achieve the desired performance, the controller parameters must be selected such that the dynamic tracking error is minimized. In this work, the Integral of Time-Weighted Absolute Error (ITAE) is used as the optimization criterion[13,16].

ITAE (Integral of Time-Weighted Absolute Error):

The ITAE performance index penalizes sustained errors more heavily by multiplying the absolute error with time. As a result, errors occurring later in the response contribute more to the cost, encouraging faster settling and improved transient behavior[16]. The ITAE is mathematically expressed as:

$$ITAE = \int_0^T t |e(t)| dt \quad (16)$$

E. PID-GWO Based controller structure :

The architecture of the suggested approach for implementing the PID-GWO in rotary drilling systems is depicted in Figure 4. The process of designing an appropriate PID controller hinges on identifying the optimal tuning parameters for the PID controller.

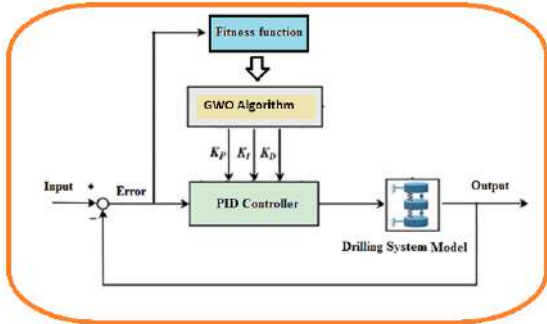


Fig. 4. GWO-PID based control system structure .

V. SIMULATION RESULTS :

This section presents the numerical simulation results obtained from the rotary drilling system model under three configurations: the open-loop system without control, the closed-loop system using manually tuned PID gains, and the closed-loop system using GWO-optimized PID parameters. The objective is to evaluate and compare the ability of each control strategy to suppress torsional stick-slip vibrations and ensure stable tracking of the desired drill bit rotational speed. The following subsections detail the dynamic responses and highlight the improvements achieved through optimization [19,20].

A. Results without any control

Figure 5 shows the step response of the rotary drilling system operating in open-loop configuration, without any control action. As observed, the drill bit angular velocity exhibits strong oscillatory behavior due to the nonlinear frictional interaction between the drill bit and the rock formation[14,15]. These oscillations persist for a long duration, producing a very high settling time. Furthermore, the drill bit fails to accurately follow the rotary table speed, and the overall dynamic performance remains unsatisfactory[17].

Such behavior is typical in drilling systems affected by severe stick-slip vibrations, where the bit alternates between sticking (zero speed) and sudden high-speed slip phases. These findings confirm the need for an appropriate control strategy to stabilize the system and suppress torsional oscillations [18].

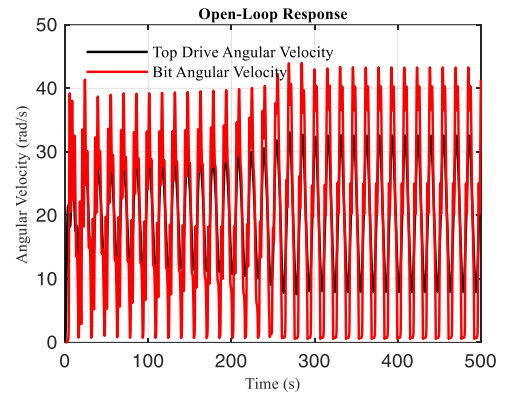


Fig. 5. Drilling string open loop Response without any control

B. Results Closed-Loop conventional PID Response

In real drilling operations, most control parameters are selected manually by field operators. To emulate this human intervention, a conventional PID controller was applied in this study using manually chosen gains:

$$K_p = 0.01, K_i = 0.025, K_d = 0.01.$$

The corresponding closed-loop response is shown in Figure 6. Although the manually tuned PID controller provides some improvement over the open-loop behavior, it is unable to effectively suppress the stick-slip phenomenon. Significant oscillations remain—particularly in the bit angular velocity—indicating persistent torsional instability. Such residual speed and torque fluctuations can induce excessive mechanical stress on the drillstring and bottom-hole assembly, leading to accelerated drill bit wear and, in severe cases, catastrophic failure.

These observations highlight the limitations of manual tuning in drilling environments characterized by strong nonlinearities and time-varying downhole conditions. Consequently, more advanced and systematic optimization techniques are required to determine PID gains capable of reliably mitigating stick-slip vibrations and ensuring stable drilling performance.

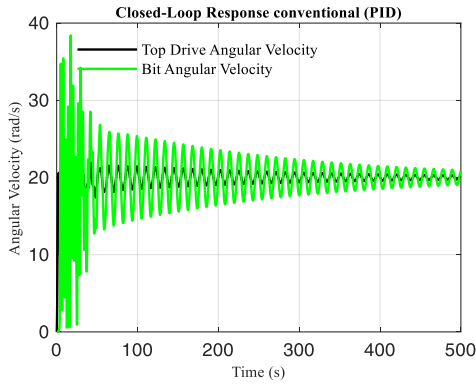


Fig. 6. Closed-loop angular velocity responses with conventional PID controller

C. Results Closed-Loop PID Response with GWO-Optimized Gains

In this subsection, the Grey Wolf Optimizer (GWO) algorithm was employed to determine the optimal PID gains for stabilizing the drilling system. The GWO parameters used in the simulation are summarized in Table 4.

TABLE IV. GWO ALGORITHM PARAMETERS

Parameter	Description	Value
N	Number of grey wolves (population size)	100
maxIter	Maximum number of iterations	10
dim	Optimization dimension (PID gains)	3
lb	Lower bounds [Kp Ki Kd]	[0 0 0]
ub	Upper bounds [Kp Ki Kd]	[10 10 10]
a	Convergence control parameter	Decreasing from 2 $\rightarrow$ 0

After completing the GWO iterations on the drillstring model, using the parameter settings summarized in Table IV, the optimization process produced the following PID controller gains:

$$K_p = 2.018, K_i = 0.006, K_d = 10.$$

Figure 7 presents the closed-loop simulation results obtained using the GWO-optimized PID controller, with the reference angular velocity set to 20 rad/s over the interval $0 \leq t \leq 500$ s. The controller effectively suppresses torsional oscillations and achieves stable tracking of the desired rotational speed.

Additionally, the convergence behavior of the GWO algorithm is shown in Figure 8. The fitness curve demonstrates a rapid reduction of the cost function during the early iterations, followed by gradual refinement as the algorithm approaches the optimal solution. This confirms the robustness and efficiency of GWO in tuning PID gains for nonlinear drilling dynamics.

Compared to the open-loop and classical PID responses, the GWO-based controller significantly improves system performance. The oscillations caused by stick-slip dynamics are rapidly attenuated, and the drill bit speed converges smoothly to the desired reference. The initial transient oscillations are eliminated in a shorter time, and the settling time is markedly reduced. Moreover, no overshoot is observed after the first transient, indicating high stability and robustness.

The superior performance of GWO is attributed to its ability to efficiently search the nonlinear control parameter space and exploit the cooperative hunting behavior of grey wolves to identify optimal PID gains. As a result, the designed controller ensures smooth drilling operation and mitigates destructive torsional vibrations[21].

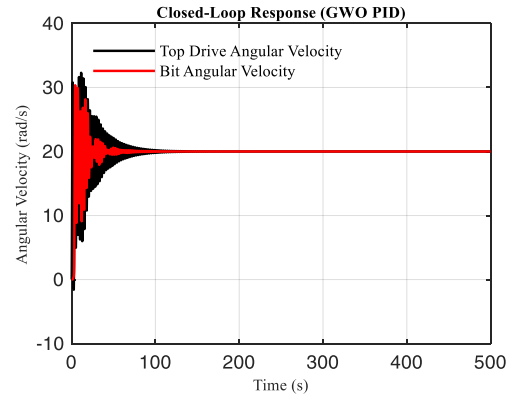


Fig. 7. Drilling String closed loop response with GWO-PID controller.

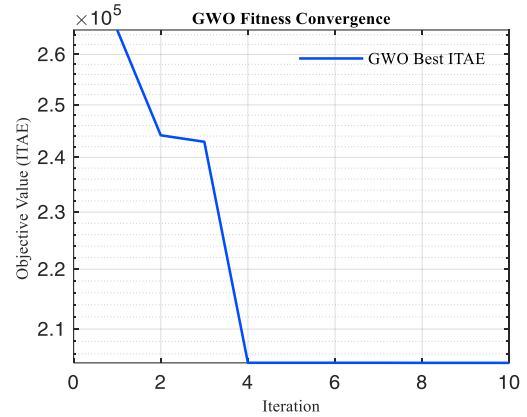


Fig. 8. ITAE Fitness function evolution across Optimization process

VI. COMPARATIVE RESULTS

The open-loop simulation results shown in Figure 5 clearly demonstrate the presence of severe stick-slip vibrations in the rotary drilling system. In this uncontrolled scenario, the drill bit undergoes repeated transitions between sticking and sudden high-speed slipping, resulting in large oscillations.

Under such conditions, drillers are often forced to adjust the Weight on Bit (WOB) conservatively to avoid damaging the drill string, which limits the achievable rate of penetration (ROP). As a result, an optimal compromise between drilling efficiency and safe operating conditions becomes difficult to guarantee in the absence of an appropriate control scheme.

To address this limitation, a conventional PID controller was first tested, as illustrated in Figure 6. Although the classical PID controller provides partial reduction of torsional oscillations, it still suffers from long settling times (exceeding 500 s) and significant steady-state error. Consequently, the PID controller without optimization is unable to sufficiently suppress stick-slip vibrations in drilling environments characterized by nonlinear friction and varying downhole conditions.

To improve the control performance, an optimized PID controller using the Grey Wolf Optimizer (GWO) algorithm was implemented. As shown in Figure 7, the GWO-based PID

controller significantly enhances the system's response by drastically reducing the oscillation amplitude and settling time. With GWO tuning, the controller eliminates steady-state error and achieves stabilization in less than 30 seconds. This improvement enhances drilling safety and ensures reliable bit-rock interaction across a wide operating range.

VII. CONCLUSION

This study addressed the mitigation of stick-slip vibrations in rotary drilling systems, a persistent challenge in drilling operations that leads to excessive tool wear, reduced efficiency, and increased operational costs. A PID controller optimized via the Grey Wolf Optimizer (GWO) was proposed to suppress torsional vibrations and improve system stability. The Integral Time Absolute Error (ITAE) criterion served as the fitness function during optimization.

Simulation results demonstrated that the GWO-optimized PID controller provides substantial improvement over both the open-loop system and classical PID tuning. The optimized controller successfully suppressed stick-slip vibrations in under 30 seconds, resulting in smooth and stable drilling operation with no steady-state error. The performance enhancements achieved by GWO highlight its capability to effectively navigate the nonlinear search space associated with drilling dynamics and identify high-quality controller parameters.

Future research may focus on incorporating artificial intelligence (AI) techniques—such as machine learning, reinforcement learning, or deep neural networks—for predictive or adaptive mitigation of stick-slip vibrations. Additionally, exploring fractional-order PID control strategies or hybrid optimization methodologies may yield even greater improvements in vibration suppression and drilling performance.

REFERENCES

- [1] M.Tellaa, A.Aibeche, M. Z. Doghmane, M.Kidouche. "Stick-Slip Vibration Mitigation in Rotary Drilling Systems Using a Particle Swarm Optimization-based PID Controller under Varying Weight on Bit Conditions " 2025 ,Acta Polytechnica Hungarica. Vol 22. 31-50. DOI: 10.12700/APH.22.9.2025.9.2.
- [2] R. I. Leine, D. H. van Campen, and W. J. G. Keultjes, "Stick-slip whirl interaction in drillstring dynamics," 2002.
- [3] M. Z. Doghmane, M. Kidouche, and A. Ahriche, "Decentralized overlapping control design with application to rotary drilling system," IETE J. Res., 2021, doi: 10.1080/03772063.2021.1886602.
- [4] L. Tang, X. Zhu, C. Shi, J. Tang, and D. Xu, "Study of the influences of rotary table speed on stick-slip vibration of the drilling system," Petroleum, vol. 1, no. 4, pp. 382–387, 2015.
- [5] L. Tang, X. Zhu, X. Qian, and C. Shi, "Effects of weight on bit on torsional stick-slip vibration of oilwell drill string," J. Mech. Sci. Technol., vol. 31, no. 10, pp. 4589–4597, 2017.
- [6] E. Navarro-López and E. Licéaga-Castro, "Non-desired transitions and sliding-mode control of a multi-DOF mechanical system with stick-slip oscillations," Chaos Solitons Fractals, vol. 41, no. 4, pp. 1561–2200, Aug. 2009.
- [7] V. Vaziri, M. Kapitaniak, and M. Wiercigroch, "Suppression of drill-string stick-slip vibration by sliding mode control: numerical and experimental studies," Eur. J. Appl. Math., vol. 29, no. 5, pp. 805–825, Oct. 2018.
- [8] Y. Liu, "Suppressing stick-slip oscillations in underactuated multibody drill-strings with parametric uncertainties using sliding-mode control," IET Control Theory Appl., vol. 9, no. 1, pp. 91–102, Jan. 2015.
- [9] F. Abdulgalil and H. Siguerdidjane, "Backstepping design for controlling rotary drilling system," in Proc. IEEE Conf. Control Applications (CCA), Toronto, Canada, Aug. 2005, pp. 120–124.
- [10] C. Mendil, M. Kidouche, and M. Z. Doghmane, "Hybrid sliding PID controller for torsional vibrations mitigation in rotary drilling systems," Indonesian J. Electr. Eng. Comput. Sci., vol. 22, no. 1, pp. 146–158, Apr. 2021, doi: 10.11591/ijeecs.v22.i1.pp146-158.
- [11] M.Tellaa, A.Aibeche, M. Z. Doghmane, M.Kidouche . "Optimized GA-Based PID Control Design for Stick-Slip Suppression in Drill String of Rotary Drilling System", in International Conference on Artificial Intelligence in Renewable Energy Systems, IC-AIRES2024, Lecture Notes in Networks and Systems, M. Hatti (Ed.): IC-AIRES 2024, LNNS 1238, pp. 453–463, 2025. https://doi.org/10.1007/978-3-031-80301-7_50
- [12] C. Mendil, M. Kidouche, and M. Z. Doghmane, "Sliding mode controller design for torsional vibrations minimization under rock-bit interaction effects," in Artificial Intelligence and Heuristics for Smart Energy Efficiency in Smart Cities, IC-AIRES 2021, Lecture Notes in Networks and Systems, vol. 361, M. Hatti, Ed. Cham: Springer, 2022, pp. , doi: 10.1007/978-3-030-92038-8_61.
- [13] W. Abedulabbas, F.R. Yaseen, "Design a PI controller based on PSO and GWO for a brushless DC motor". (2022). Journal Européen des Systèmes Automatisés, Vol. 55, No. 3, pp. 331-338. <https://doi.org/10.18280/jesa.550305>
- [14] Z. Guan et al., Theory and Technology of Drilling Engineering. Singapore: Springer, 2021. doi: 10.1007/978-981-15-9327-7_2.
- [15] T. G. Ritto and M. Ghandchi-Tehrani, "Active control of stick-slip torsional vibrations in drill-strings," J. Vib. Control, vol. 25, no. 1, pp. 194–202, 2019, doi: 10.1177/1077546318774240
- [16] M.Tellaa, A.Aibeche, M. Z. Doghmane, M.Kidouche . "Optimal Control Design for Stick-Slip Vibrations Mitigation in Rotary Drilling Systems ", in International Conference on Artificial Intelligence in Renewable Energy Systems, IC-AIRES2024, Lecture Notes in Networks and Systems, M. Hatti (Ed.): IC-AIRES 2024, LNNS 1238, pp. 464–474, 2025. https://doi.org/10.1007/978-3-031-80301-7_51
- [17] C. Mendil, M. Kidouche, M. Z. Doghmane, S. Benammar, and K. F. Tee, "Rock-bit interaction effects on high-frequency stick-slip vibration severity in rotary drilling systems," Multidiscip. Model. Mater. Struct., vol. 17, no. 5, pp. 1007–1023, 2021, doi: 10.1108/MMMS-10-2020-0256.
- [18] A. Abushawish, M. Hamadeh, and A. Bou Nassif, "PID controller gains tuning using metaheuristic optimization methods: A survey," Int. J. Comput., 2020, doi: 10.46300/9108.2020.14.14.
- [19] I. Ahmed and M. Khamees, "A review of adaptive tuning of PID-controller: Optimization techniques and applications," Int. J. Nonlinear Anal. Appl., vol. 15, no. 2, pp. 29–37, 2024.
- [20] K. Su, J. Wei, M. Li, H. Li, W. Da, and L. Zhang, "Research on stick-slip vibration suppression method of drill string based on machine learning optimization," SV, vol. 57, 2023, doi: 10.32604/sv.2023.043734.
- [21] R. Riane, M. Z. Doghmane, M. Kidouche, and K. F. Tee, "Stick-slip vibration suppression in drill string using observer-based LQG controller," Sensors, vol. 22, no. 16, p. 5979, 2022

Contribution on detection vehicle from algorithms YOLO

Yasmine SILHADI

Faculty of Electrical and Computer Engineering,
Laboratoire LARI, University Mouloud Mammeri of
Tizi Ouzou, Algeria. yasmine.silhadi@ummto.dz.

Hocine CHEBI

Faculty of Electrical Engineering, Djillali Liabes,
University of Sidi Bel Abbes, Laboratory Intelligent
Control and Electrical Power System (ICEPS), Algeria.
chebi.hocine@yahoo.fr, hocine.chebi@uni-sba.dz.

Rachida AOUDJIT

Faculty of Electrical and Computer Engineering,
Laboratoire LARI, University Mouloud Mammeri of
Tizi Ouzou, Algeria. aoudjit_rachida@yahoo.com.

Abstract : There are numerous applications for drones in various civilian disciplines, such as disaster relief, agriculture, mapping, real-time traffic monitoring, and overhead power line inspection. These drones are equipped with sophisticated sensors, such as cameras, that allow them to acquire high-resolution images. These images can then be processed by automated computer vision systems for analysis and decision-making. Several systems have been suggested for the automatic detection of urban objects such as buildings, trees, and vehicles. However, detecting and tracking vehicles in complex urban environments using data captured by drones still presents considerable challenges. These environments contain numerous static and moving targets and background objects, varying lighting conditions, and frequent occultations. These conditions affect the accuracy of current systems and highlight the need for specific improvements. It is in this context that we propose this vehicle detection model based on You Look Only Once version 8, combined with preprocessing using the clahe (Contrast Limited Adaptive Histogram Equalization) filter and a bilateral filter, to help improve image quality by reducing noise and increasing contrast in low-light areas, thus enabling vehicle identification in complex urban environments. The experimental results and the comparison mentioned below show that the proposed vehicle detection method achieves a 97% accuracy rate on the entire aerial vehicle imagery (VAID) dataset from a drone.

Keywords: Vehicle detection, Aerial Imaging, yolov8, real-time object detection, complex urban environments.

I. INTRODUCTION

Object detection is an important branch of computer vision widely applied to many aspects of daily life [1], [2], [3]. The expansion of Unmanned Aerial Vehicle (UAV) use has revolutionized large-scale visual analysis, paving the way for advanced applications in fields as diverse as urban video surveillance, traffic management, civil and military security, and autonomous driving. Among these applications, vehicle detection and tracking in complex urban environments have become major challenges, particularly for optimizing traffic flow, preventing accidents, and managing road infrastructure [4], [5], [6], [7], [8], [9]. However, analyzing aerial images presents several challenges: variations in scale, partial occlusion, changes in brightness and weather conditions, and

the speed and rapid movement of objects require an approach capable of real-time detection and tracking.

Before the advent of deep learning methods, object detection relied on classical algorithms coupled with machine learning techniques, incorporating feature descriptors such as HOG (Histogram of Oriented Gradients), SIFT (Scale-Invariant Feature Transform), and SURF (Speeded-Up Robust Features) [10], [11], [12], [13]. While these methods enabled significant progress, they had limited performance in more complex scenarios, particularly when identifying small objects, in dense environments, or in the presence of occlusion [14], [15], [16].

However, the rise of deep learning-based models has revolutionized image processing and the ability of computer vision systems to automatically extract complex features from data far surpasses that of conventional techniques. Convolutional neural networks (CNNs) [17], [18] have revolutionized the field by enabling automatic and hierarchical feature extraction, thus surpassing approaches based on manual descriptors. Among these detection architectures, FAST RCNN [19], SSD [15], and YOLO (You Only Look Once) [20] stand out as a benchmark for real-time detection, thanks to its ability to analyze an image in a single step, thereby optimizing inference speed and accuracy. Furthermore, the recent integration of Transformers in computer vision [21], [22] has further strengthened the robustness of models by improving their ability to capture complex spatial relationships between objects.

These advances are particularly relevant for vehicle detection and tracking [10], and for urban surveillance in general, especially with UAVs, where it is essential to have a model capable of processing large, complex scenes while maintaining high accuracy despite variations in scale, changing viewing angles, and challenging lighting conditions [23].

However, existing studies on UAV-based vehicle detection still present limitations, such as the difficulty of detecting small or distant vehicles, sensitivity to illumination changes, and the lack of extensive evaluation of preprocessing techniques in aerial imagery. Moreover, the performance of the recently introduced YOLOv8 architecture has not yet been thoroughly investigated for UAV scenarios, which motivates this work

In this paper, we propose an optimized method for vehicle detection from aerial images, capable of overcoming the challenges of considerable scale variations, partial occlusion, and changes in lighting. The system will be tested on the VAID dataset [24], designed for analyzing road traffic from aerial images. The main contributions of this paper are summarized as follows:

- An evaluation of YOLOv8 for UAV-based vehicle detection under challenging aerial conditions.
- An analysis of the impact of a simple preprocessing pipeline (CLAHE + bilateral filtering) on detection performance.
- A detailed experimental setup including dataset split, training parameters, and evaluation metrics.

This paper is organized as follows: Section II: Brief overview of the state of the art. The detection method used is explained in Section III. Performance measurements, experimental results, and discussion are described in Section IV. The conclusion is presented in Section V.

II. A BRIEF OVERVIEW OF THE STATE OF THE ART

Object detection is a critical task in computer vision, consisting of identifying and locating objects in real time within image sequences. Advances in this area of vehicle detection from aerial images captured by drones are briefly described in this section.

The use of deep learning methods for vehicle detection [25], [26] is a trend in the computer vision community, and the one that excels in this area is the use of You Only Look Once (YOLO) [20]. Known for its ability to process multiple images very quickly compared to other two-step methods such as region-based convolutional neural networks (RCNNs) [27], YOLO is widely used for real-time detection applications, thanks to its efficiency and speed of data processing. The YOLOv5 version is generally used for its accuracy in vehicle detection, even with variations in lighting and scale [28]. In [29] they used Yolov4 combined with CNNs for aerial image classification. Yolov3 and RetinaNet [30] were used for multi-object detection. [26], [31] different versions of YOLO were used for traffic monitoring. In this work, we will use Yolov8 [31] for its simplicity and effectiveness in detecting vehicles in a complex environment. Still, most prior UAV studies focus on earlier YOLO versions, and comparisons with recent detectors remain limited.

In addition, few works evaluate the contribution of preprocessing techniques in UAV imagery, which motivates their examination in our approach.

III. DETECTION METHODE

This article proposes a vehicle detection method based on drone-captured images. A preprocessing step is performed on all data, followed by detection using one of the latest versions of YOLO, "YOLOV8". Figure 3 illustrates the architecture of the method used.

In the preprocessing step, all images were converted to grayscale to simplify the process and focus on shape and structure. Then, a CLAHE filter was applied to all images to enhance pixel contrast [32]. Finally, noise was reduced while preserving edges using a bilateral filter [33].

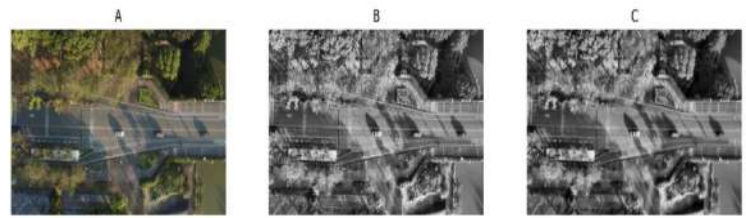


Fig.1. the different phases of pre-treatment, (A) represent the original image, (B) show the image with gray scale with clahe filter, the last image (C) is the processed one with gray scale, clahe and bilateral filter.

a) Vehicule Detection :

The YOLO algorithms[20], [28], [34], [35], [36] are single-stage detectors widely used in object detection, particularly for vehicles, due to their high performance and speed. Unlike traditional approaches, YOLO analyzes the entire image, capturing global information to identify objects and directly predict bounding box coordinates, thus optimizing detection accuracy. YOLOv8 offers different models designed for various tasks such as segmentation, detection, and classification, each specializing in a different computer vision task.

YOLOv8 adopts the CSP principle by replacing the C3 module with C2f, thereby optimizing the gradient information flow while remaining compatible with YOLOv5. It also integrates the SPPF module, which applies three successive 5×5 max-poolings to improve the detection of objects of different sizes. In its "neck" part, YOLOv8 uses PAN-FPN for improved feature fusion and adopts a decoupled head structure inspired by YOLOx, combining multiple C2f modules and two oversamplings for increased accuracy [31], [37]. Finally, YOLOv8 is an anchorless model, directly detecting object centers, which reduces the number of box predictions and accelerates non-maximal suppression (NMS). These optimizations make it a high-performing and efficient detector with a small number of parameters.

In this work, YOLOv8 was selected due to its improved performance on small objects and its reduced computational cost, which are essential for UAV-based detection.

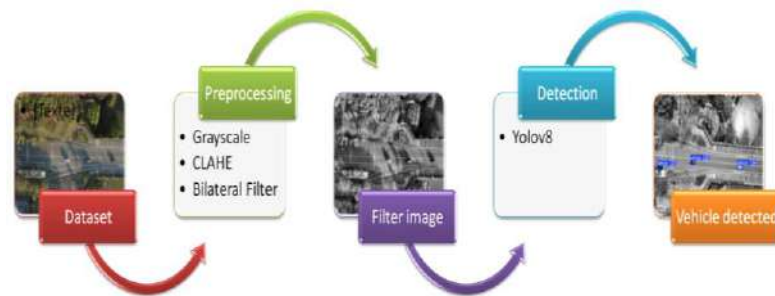


Fig.2. The pipeline of the different phases for vehicle detection, Detection process with yolov8 and the main components of the algorithm architecture have been highlighted in this flowchart .

IV. EXPERIENCE ET RESULTAT

La méthode de détection des véhicules a été implémentée sur l'environnement d'exécution GPU T4 hébergé de Google Colab, à l'aide d'outils Python. La méthode a été testée sur le jeu de données VAID. Les détails et résultats obtenus sont décrits ci-dessous, avec une comparaison avec l'état de l'art.

1. Description du dataset:

L'ensemble de données aériennes pour la détection de véhicules VAID [24] contient 5985 images annotées RGB au format (.jpg), de dimension (1137*640) pixels, prises sous différents éclairages et angles à différents endroits à Taiwan. The dataset was split into training, validation, and test sets



Fig.3. Vehicle detection using yolov8.

1. Results obtained:

We used three metrics to evaluate and test the proposed architecture, and their equations are as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

Or TP corresponding to True Positives and FP False Positives.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

FN stands for False Negatives.

$$\text{F1 Score} = 2 \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Confidence thresholds and Non-Maximum Suppression (NMS) parameters were kept consistent with YOLOv8 defaults to ensure reproducibility.

The following table presents the results obtained for the proposed architecture. After applying the preprocessing pipeline, YOLOv8 achieved a precision of 97%, a recall of 93.2%, and an F1 score of 95.1%, demonstrating its effectiveness in detecting vehicles from aerial imagery. These results indicate that the preprocessing step improved detection performance, particularly in images with varying illumination and contrast.

METRIC	VALUE
Precision	0.970
Recall	0.932
F1 Score	0.951

Table 1: Metrics and results.

2. Comparison with the state of art:

In this experiment, we compare our approach with other state of art methods, and the results obtained are described in table2.

The results show that the proposed method achieves 97% accuracy on the VAID dataset, matching the best performance obtained with yolov4 pre-trained with ground surveillance (CCTV) images and surpassing other methods

Article	Method	Dataset	Results
Vehicle Detection and Tracking in UAV Images Via Yolov3 and Kalman filter [38]	Yolov3	VAID	96,7%
Smart Traffic monitoring Through Pyramid Pooling Vehicle Detection and Filter-based Tracking [39]	CNN + pooling pyramidal	VAID VEDAI DLR3 K	95,78% 95,18% 93,13%
Aerial Dataset Integration For Vehicle Dtection Based Yolov4 [4]	Yolov4	VAID UAVD DOTA	76,7% 72,9% 75,3%
Multiple Vehicles Detection and Tracking Using Ground and Aerial Images For Vehicular Management System[40]	A pretrained Yolov4	VAID	97%
OURS	Yolov8	VAID	97%

Table 2: Comparison of the accuracy of the proposed method with other state of art techniques.

V. CONCLUSION

In conclusion, this article proposes an optimized method for vehicle detection in UAV aerial images, integrating YOLOv8 with preprocessing aimed at improving image quality and reducing noise while preserving object contours. The results obtained confirm the effectiveness of our approach for vehicle detection in urban environments, despite the challenges related to scale variations, partial occlusion, and changing lighting conditions.

However, several future directions can be considered. First, the addition of a tracking method; second, to ensure the robustness and generalizability of our model, validation on other UAV datasets such as VisDrone and UAVID to measure its effectiveness in diverse environments.

Additionally, future work will include ablation studies to quantify the contribution of individual preprocessing steps and comparisons with transformer-based detectors to benchmark YOLOv8 against recent state-of-the-art methods.

References

- [1] S. Yeom, « Moving People Tracking and False Track Removing with Infrared Thermal Imaging by a Multirotor », *Drones*, vol. 5, n° 3, p. 65, juill. 2021, doi: 10.3390/drones5030065.
- [2] S. W. Khan *et al.*, « Anomaly Detection in Traffic Surveillance Videos Using Deep Learning », *Sensors*, vol. 22, n° 17, p. 6563, août 2022, doi: 10.3390/s22176563.
- [3] T. Zhang, S. Liu, C. Xu, B. Liu, et M.-H. Yang, « Correlation Particle Filter for Visual Tracking », *IEEE Trans. Image Process.*, vol. 27, n° 6, p. 2676- 2687, juin 2018, doi: 10.1109/TIP.2017.2781304.
- [4] W. Omar, Y. Oh, J. Chung, et I. Lee, « Aerial Dataset Integration For Vehicle Detection Based on YOLOv4 », *Korean J. Remote Sens.*, 2021.
- [5] Md. M. Rahman, S. Siddique, M. Kamal, R. H. Rifat, et K. D. Gupta, « UAV (Unmanned Aerial Vehicle): Diverse Applications of UAV Datasets in Segmentation, Classification, Detection, and Tracking », *Algorithms*, vol. 17, n° 12, p. 594, déc. 2024, doi: 10.3390/a17120594.
- [6] O. Katar et E. Duman, « U-NET BASED CAR DETECTION METHOD FOR UNMANNED AERIAL VEHICLES », *Mühendis. Bilim. Ve Tasar. Derg.*, vol. 10, n° 4, p. 1141- 1154, déc. 2022, doi: 10.21923/jesd.1087477.
- [7] B. S. Sashank et R. K. K., « Segmentation and Detection of Urban Objects from UAV Images Using Hybrid Deep Learning Combinations », in *2023 Innovations in Power and Advanced Computing Technologies (i-PACT)*, Kuala Lumpur, Malaysia: IEEE, déc. 2023, p. 1- 7. doi: 10.1109/i-PACT58649.2023.10434584.
- [8] A. M. Qureshi, « Road Traffic Monitoring from Aerial Images Using Template Matching and Invariant Features », *Comput. Mater. Contin.*, vol. 78, n° 3, p. 3683- 3701, 2024, doi: 10.32604/cmc.2024.043611.
- [9] M. O. Yusuf *et al.*, « Target detection and classification via EfficientDet and CNN over unmanned aerial vehicles », *Front. Neurobotics*, vol. 18, p. 1448538, août 2024, doi: 10.3389/fnbot.2024.1448538.
- [10] A. Setiawan, R. A. Yunmar, et H. Tantriawan, « Comparison of Speeded-Up Robust Feature (SURF) and Oriented FAST and Rotated BRIEF (ORB) Methods in Identifying Museum Objects Using Low Light Intensity Images », *IOP Conf. Ser. Earth Environ. Sci.*, vol. 537, n° 1, p. 012025, juill. 2020, doi: 10.1088/1755-1315/537/1/012025.
- [11] Q. Fu *et al.*, « Fast ORB-SLAM without Keypoint Descriptors », *IEEE Trans. Image Process.*, vol. 31, p. 1433- 1446, 2022, doi: 10.1109/TIP.2021.3136710.
- [12] A. Yusefi, A. Durdu, et C. Sungur, « Görsel Odometride SIFT, SURF, FAST, STAR ve ORB özellik algılama algoritmalarının Performans ve Takas Değerlendirmesi », *Eur. J. Sci. Technol.*, nov. 2020, doi: 10.31590/ejosat.819735.
- [13] K. A. Elorabi, A. Zekry, et W. Mohamed, « Optimizing SIFT algorithm parameters for better matching UAV and satellite images », *J. Phys. Conf. Ser.*, vol. 2616, n° 1, p. 012044, nov. 2023, doi: 10.1088/1742-6596/2616/1/012044.
- [14] « Détection de plans et d'objets dans une image par machine learning.pdf ».
- [15] W. Liu *et al.*, « SSD: Single Shot MultiBox Detector », in *Computer Vision – ECCV 2016*, vol. 9905, B. Leibe, J. Matas, N. Sebe, et M. Welling, Éd., in Lecture Notes in Computer Science, vol. 9905, Cham: Springer International Publishing, 2016, p. 21- 37. doi: 10.1007/978-3-319-46448-0_2.
- [16] A. Krizhevsky, I. Sutskever, et G. E. Hinton, « ImageNet classification with deep convolutional neural networks ».
- [17] A. Khan, A. Sohail, U. Zahoora, et A. S. Qureshi, « A Survey of the Recent Architectures of Deep Convolutional Neural Networks », *Artif. Intell. Rev.*, vol. 53, n° 8, p. 5455- 5516, déc. 2020, doi: 10.1007/s10462-020-09825-6.
- [18] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, et S. Belongie, « Feature Pyramid Networks for Object Detection », in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, juill. 2017, p. 936- 944. doi: 10.1109/CVPR.2017.106.
- [19] S. Ren, K. He, R. Girshick, et J. Sun, « Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks », *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, n° 6, p. 1137- 1149, juin 2017, doi: 10.1109/TPAMI.2016.2577031.
- [20] J. Redmon, S. Divvala, R. Girshick, et A. Farhadi, « You Only Look Once: Unified, Real-Time Object Detection », in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA:

- IEEE, juin 2016, p. 779- 788. doi: 10.1109/CVPR.2016.91.
- [21] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, et M. Shah, « Transformers in Vision: A Survey », *ACM Comput. Surv.*, vol. 54, n° 10s, p. 1- 41, janv. 2022, doi: 10.1145/3505244.
- [22] A. Vaswani *et al.*, « Attention Is All You Need », 1 août 2023, *arXiv*: arXiv:1706.03762. Consulté le: 8 octobre 2023. [En ligne]. Disponible sur: <http://arxiv.org/abs/1706.03762>
- [23] Z. Zhang, X. Lu, G. Cao, Y. Yang, L. Jiao, et F. Liu, « ViT-YOLO:Transformer-Based YOLO for Object Detection », in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Montreal, BC, Canada: IEEE, oct. 2021, p. 2799- 2808. doi: 10.1109/ICCVW54120.2021.00314.
- [24] H.-Y. Lin, K.-C. Tu, et C.-Y. Li, « VAID: An Aerial Image Dataset for Vehicle Detection and Classification », *IEEE Access*, vol. 8, p. 212209- 212219, 2020, doi: 10.1109/ACCESS.2020.3040290.
- [25] S. Pouyanfar *et al.*, « A Survey on Deep Learning: Algorithms, Techniques, and Applications », *ACM Comput. Surv.*, vol. 51, n° 5, p. 1- 36, sept. 2019, doi: 10.1145/3234150.
- [26] S. Ali et A. Jalal, « Vehicle Detection and Tracking from Aerial Imagery via YOLO and Centroid Tracking ».
- [27] Z. Cai et N. Vasconcelos, « Cascade R-CNN: Delving Into High Quality Object Detection », in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT: IEEE, juin 2018, p. 6154- 6162. doi: 10.1109/CVPR.2018.00644.
- [28] A. M. Qureshi, « Semantic Segmentation and YOLO Detector over Aerial Vehicle Images », *Comput. Mater. Contin.*, vol. 80, n° 2, p. 3315- 3332, 2024, doi: 10.32604/cmc.2024.052582.
- [29] M. O. Yusuf, M. Hanzla, et A. Jalal, « Vehicle Detection and Classification via YOLOv4 and CNN over Aerial Images », in *2024 International Conference on Engineering & Computing Technologies (ICECT)*, Islamabad, Pakistan: IEEE, mai 2024, p. 1- 6. doi: 10.1109/ICECT61618.2024.10581252.
- [30] S. Kapania, D. Saini, S. Goyal, N. Thakur, R. Jain, et P. Nagrath, « Multi Object Tracking with UAVs using Deep SORT and YOLOv3 RetinaNet Detection Framework », in *Proceedings of the 1st ACM Workshop on Autonomous and Intelligent Mobile Systems*, Bangalore India: ACM, janv. 2020, p. 1- 6. doi: 10.1145/3377283.3377284.
- [31] M. Hanzla, M. O. Yusuf, et A. Jalal, « Vehicle Surveillance Using U-NET Segmentation and DeepSORT Over Aerial Images », in *2024 International Conference on Engineering & Computing Technologies (ICECT)*, Islamabad, Pakistan: IEEE, mai 2024, p. 1- 6. doi: 10.1109/ICECT61618.2024.10581338.
- [32] M. O. Yusuf, M. Hanzla, et A. Jalal, « Vehicle Detection and Classification via YOLOv4 and CNN over Aerial Images », in *2024 International Conference on Engineering & Computing Technologies (ICECT)*, Islamabad, Pakistan: IEEE, mai 2024, p. 1- 6. doi: 10.1109/ICECT61618.2024.10581252.
- [33] A. Naseer et A. Jalal, « Hybrid Deep Learning Aerial Framework for Road Scene Objects Segmentation and Classification ».
- [34] J. Redmon et A. Farhadi, « YOLOv3: An Incremental Improvement ».
- [35] G. Mujtaba, D. A. Ullah, et A. Jalal, « Advanced UAV-based Multi-Vehicle Monitoring using Deep learner and Vision Transformer ».
- [36] G. Mujtaba et A. Jalal, « UAV-based Road Traffic Monitoring via FCN Segmentation and Deepsort for Smart Cities ».
- [37] M. Hanzla *et al.*, « Vehicle recognition pipeline via DeepSort on aerial image datasets », *Front. Neurobotics*, vol. 18, p. 1430155, août 2024, doi: 10.3389/fnbot.2024.1430155.
- [38] S. Ali, A. Jalal, M. Hamad Alatiyyah, K. Alnowaiser, et J. Park, « Vehicle Detection and Tracking in UAV Imagery via YOLOv3 and Kalman Filter », *Comput. Mater. Contin.*, vol. 76, n° 1, p. 1249- 1265, 2023, doi: 10.32604/cmc.2023.038114.
- [39] A. A. Rafique *et al.*, « Smart Traffic Monitoring Through Pyramid Pooling Vehicle Detection and Filter-Based Tracking on Aerial Images », *IEEE Access*, vol. 11, p. 2993- 3007, 2023, doi: 10.1109/ACCESS.2023.3234281.
- [40] S. J. Ali, M. Omer, S. M. Raza, et H. Choo, « Multiple Vehicles Detection and Tracking Using Ground and Aerial Images for Vehicular Management System », in *Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications*, vol. 1925, T. K. Dang, J. Küng, et T. M. Chung, Éd., in Communications in Computer and Information Science, vol. 1925. , Singapore: Springer Nature Singapore, 2023, p. 564571. doi: 10.1007/978-981-99-8296-7_41.

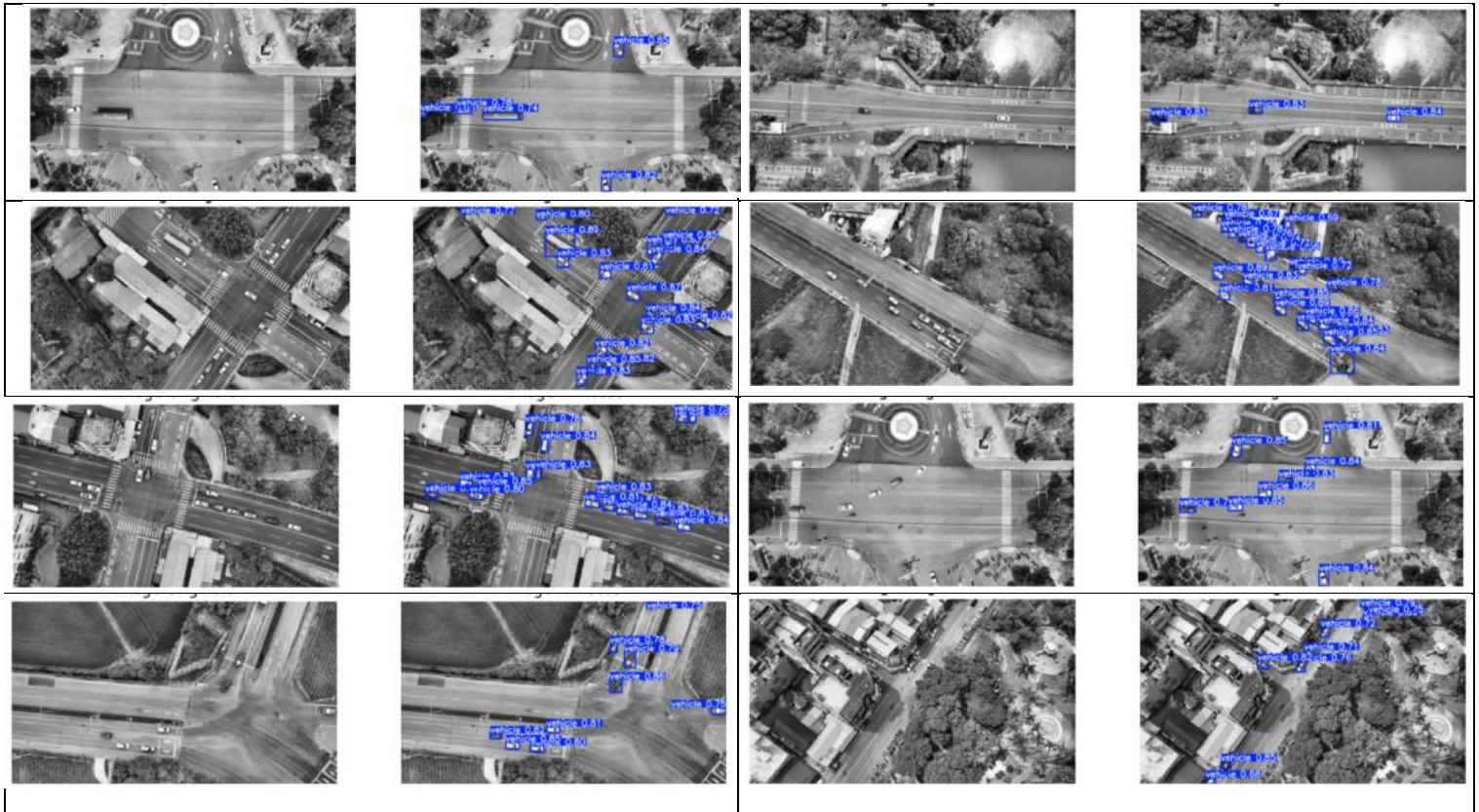


Fig.4. Different images illustrating the results obtained with the method used for vehicle detection, showing the filtered images and their annotation

Optimization of the parameters of a stochastic gradient algorithm for precision agriculture

Hocine CHEBI

Faculty of Electrical Engineering, Djillali Liabes, University of Sidi Bel Abbes, Laboratory Intelligent Control and Electrical Power System (ICEPS), Algeria.

chebi.hocine@yahoo.fr, hocine.chebi@uni-sba.dz.

Yasmine SILHADI

Faculty of Electrical and Computer Engineering, Laboratoire LARI, University Mouloud Mammeri of Tizi Ouzou, Algeria. yasmine.silhadi@ummo.dz.

Abstract: This study introduces an innovative deep learning (DL) method for the detection of plant diseases, which are critical factors that adversely affect agricultural productivity. The early and accurate identification of these diseases is essential for effective management strategies. Our research leverages a convolutional neural network (CNN) tailored for analyzing images of plant leaves. We begin by discussing the importance of accurate disease detection for maintaining healthy crops and minimizing economic impacts. Our methodology encompasses advanced image preprocessing techniques and a well-defined neural network architecture, demonstrating remarkable efficiency in identifying various plant diseases. Experimental results reveal a high classification accuracy of 99.73% for leaf diseases, which we compare against existing literature to underscore the superior performance of our model. Additionally, we assess the broader implications of our findings for future research and practical agricultural applications. In conclusion, we underscore the significance of our contributions to plant disease detection technologies and offer suggestions for subsequent research aimed at enhancing the robustness and real-world application of our model. This work not only advances agricultural technology but also provides essential insights for researchers and practitioners utilizing deep learning for sustainable crop management.

Keywords: Agriculture, Robotic Vision, Plant disease detection, Deep Learning, Artificial Intelligence

I. INTRODUCTION

In contemporary agriculture, the detection of plant diseases is significantly improved by integrating advanced technologies such as Artificial Intelligence (AI) and the Internet of Things (IoT). These technologies enable precise analysis and identification of disease symptoms, leading to the swift detection of early infection signs. This facilitates proactive measures that can mitigate crop damage and enhance overall agricultural productivity.

AI's role in image analysis is crucial. For example, Han et al. [1] highlight the value of imaging techniques for the early

detection of plant diseases, utilizing a three-dimensional space based on texture and color. Cruz et al. [2] developed a vision-based system aimed at identifying leaf scorch symptoms in *Olea europaea* L. infected by *Xylella fastidiosa*, employing deep learning and transfer learning for improved detection accuracy.

Moreover, Harakannanavar et al. [3] emphasize the contributions of computer vision and machine learning to plant disease challenges, dividing the detection process into identification, analysis, and verification stages while utilizing machine learning algorithms to focus on early symptoms.

Singh et al. [4] demonstrate that various imaging techniques—including thermal, hyperspectral, and 3D imaging—are vital for gathering data necessary for identifying disease symptoms. Techniques like deep learning and K-means clustering are employed for effective classification, enhancing early detection and management of plant diseases.

The rise of Deep Learning (DL) has transformed agricultural practices by enabling autonomous systems to accurately interpret visual information gathered from the environment, significantly improving crop management, inspection efficiency, and overall productivity while reducing negative environmental impacts.

In this context, the present paper seeks to highlight the critical role of AI and IoT in the early detection of plant diseases and examine the specific methodologies utilized in modern agricultural practices.

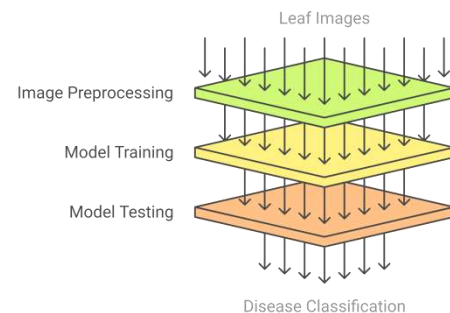


Fig. 1. Block diagram showing the different steps to be used for detecting plant diseases.

The following sections will review previous research and key advancements in deep learning for plant disease detection, providing

insights into the current landscape and identifying potential gaps in existing studies. The discussion will detail the development steps of a deep learning-based plant disease detection system, including data collection, selection, training, and classification methods. Results from applying the deep learning approach will be presented and analyzed, assessing the model's effectiveness in agricultural robotic vision. A comparative study with relevant research will showcase our approach's strengths and limitations. Finally, the conclusion will encapsulate key findings, underscoring their significance in agricultural robotic vision and proposing future research directions.

II. STATE OF ART ON DEEP LEARNING FOR PLANT DISEASE DETECTION

Computer vision, particularly in the area of object recognition, has experienced remarkable advancements in recent years. Large-scale computer vision (LSCV) has become a fundamental benchmark for addressing various visualization challenges in the field, including object classification.

In agriculture, the applications of computer vision are both essential and diverse. Mohanty et al. [5] emphasize the use of mobile technology and image data for identifying and diagnosing agricultural challenges, particularly plant diseases. They discuss the ability of smartphones to capture image data along with location and temporal information, which can help prevent yield losses. The authors also note the rapid growth of mobile technology and its potential to provide accurate diagnostics through smartphones, coupled with the implementation of algorithms and resources for deep learning and image analysis related to plant diseases.

Krizhevsky et al. [6] made a significant contribution to the field in 2012 by introducing a deep convolutional neural network (CNN) that achieved a maximum error rate of 16.4% when classifying images into 1,000 categories. Since then, various improvements in deep convolutional neural networks have led to a reduction in this error rate to 3.57% over the subsequent three years [2-6].

III. METHODOLOGY OF DL FOR DISEASE DETECTION

In this paper, we will employ Deep Learning (DL) techniques on plant images to identify diseases. DL, a subset of Machine Learning (ML) and a branch of artificial intelligence, utilizes artificial neural networks to learn from unstructured or unlabeled data, mimicking the human brain's function. These networks consist of interconnected layers of neurons that enable complex abstraction of input features, making DL particularly effective for processing unstructured data, such as images. Our research focuses on applying DL to detect and classify diseased plants.

We will implement this using PyTorch, a popular open-source library tailored for deep learning development, which allows for the creation of Convolutional Neural Networks (CNNs) well-suited for analyzing images of affected plant parts. With its dynamic computation capabilities, PyTorch facilitates rapid experimentation, while its robust documentation and community support help address specific challenges. To train our model effectively, we will preprocess the images by removing backgrounds and isolating the plants, splitting the processed database into 80% for training and 20% for testing.

After preprocessing, we will construct our CNN model with multiple hidden layers, incorporating functions like convolution, MaxPooling, and the ReLU activation function. We will train the model using the preprocessed data and optimize parameters through an algorithm like stochastic gradient descent to minimize loss and enhance performance. Following training, we will evaluate the model's accuracy using the test set and calculate its precision to validate our approach.

Finally, we will compare our results with existing studies utilizing the same dataset, analyzing accuracy rates, error rates, and training times to contextualize our findings and assess our methodology's effectiveness. Figure 2 will illustrate the detailed methodology for detecting plant diseases.

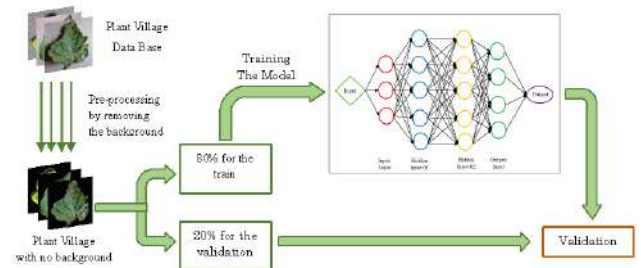


Fig. 2: Detailed flowchart of the method used.

In our model for plant disease detection, we utilize several key functions: convolution, ReLU (Rectified Linear Unit), max pooling, and softmax, each playing a vital role in improving the model's effectiveness.

Convolution is the initial operation that extracts essential features from input images, such as textures and colors associated with plant diseases. The mathematical expression is:

$$\text{output}(i, j) = \sum \sum (\text{input}(m, n) \times \text{kernel}(p, q))$$

Here, $\text{output}(i, j)$ is the value in the resulting feature map at coordinates (i, j) , $\text{input}(m, n)$ is the pixel value from the original image, and $\text{kernel}(p, q)$ is the filter value at position (p, q) . This operation allows the model to capture intricate details necessary for identifying diseases.

Next, the ReLU activation function introduces non-linearity, enabling the model to learn complex patterns. Its expression is:

$$\text{ReLU}(x) = \max(0, x)$$

In this case, x is the input value from the convolution layer. If x is positive, it passes through unchanged; if negative, it becomes zero. This helps the model focus on meaningful features.

To reduce the spatial dimensions of the feature maps while retaining important information, we apply max pooling, represented as:

$$\text{output}(i, j) = \max(\text{input}(m, n))$$

Here, $\text{output}(i, j)$ is the pooled value at position (i, j) , and $\text{input}(m, n)$ refers to the values in the pooling region. This

operation identifies the most significant features, aiding in the effective classification of diseases.

These functions collectively enhance our model's capability to accurately detect and classify plant diseases based on extracted features from images.

IV. RESULTS AND DISCUSSIONS

We will start by analyzing the dataset to determine the number of classes, the count of images in each class, the total image count, and the dimensions of the images.

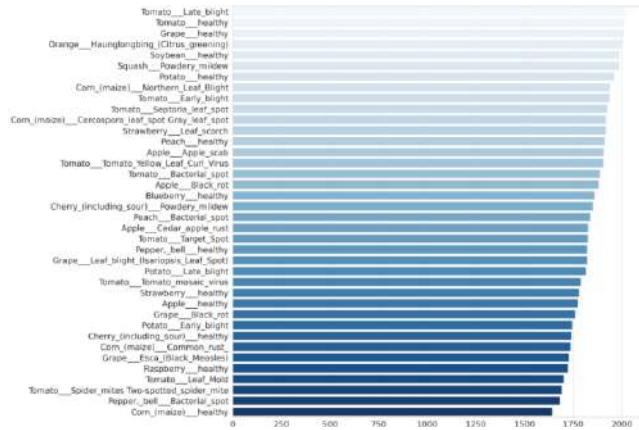


Fig. 3: Diagram shows the distribution of images in classes.

The PlantVillage database comprises roughly 87,000 RGB images of both healthy and diseased crop leaves, spanning 38 distinct classes. The dataset will be split into training and validation sets in an 80/20 ratio, while preserving the organization into 38 directories. Additionally, a separate folder with 33 test images will be created for prediction purposes. The PlantVillage dataset is available for download as an open-source resource on the internet [8].



Fig. 4 : Samples of each class in our database.

We will test our model specifically on potato leaf images to assess its prediction capabilities regarding plant health. For each input image, the model will determine whether the potato leaf is healthy or diseased. In cases where the leaf is classified

as diseased, the model will further identify the specific disease affecting the plant. This detailed prediction not only provides valuable insights into the health of the potato crop but also aids farmers in implementing timely interventions to manage diseases effectively, ultimately supporting better agricultural practices and improving yield outcomes.

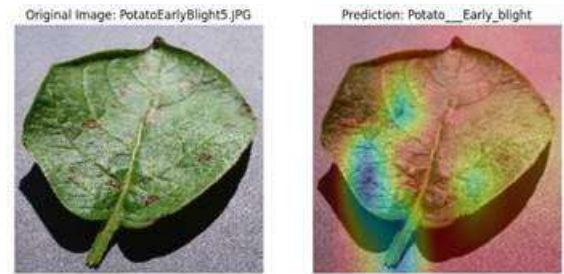


Fig 5: Diseased potato leaf and the prediction

Grad-CAM (Gradient-weighted Class Activation Mapping) is a visualization technique that helps interpret CNNs by showing which areas of an image most influenced predictions. It generates a heatmap using gradients from the target class and the final convolutional layer, highlighting critical regions. Red areas indicate high importance for the model's prediction, while yellow and green signify moderate importance. Blue areas are less significant or contribute minimally. This visual representation is crucial for understanding the model's decision-making process by spotlighting the features that affect classification outcomes.

The dataset has been used in several studies, and we will compare our results with existing methods. Mohanty et al. [5] utilized deep CNNs for plant disease classification, experimenting with various architectures and training distributions. Their best models achieved accuracies of 98.81% and 99.34%, indicating deep CNNs significantly outperform random chance.

Yao et al. [7] also applied CNNs to predict plant species from leaf images, including both healthy and diseased leaves. They tested architectures like AlexNet and MobileNetV2, yielding accuracy improvements ranging from 0.99750 to 0.99275. Their results confirm the effectiveness of diverse models and strategies for plant classification.

Both studies underscore the potential of these methods for smart agriculture and effective plant health management. In the following table, we will present our accuracy results alongside those of these studies, highlighting the high performance of all three models in identifying plant diseases.

Tab. 3: Comparison of Accuracy of CNN Models for Plant Disease Classification

Papers	Approach	Year	Accuracy (%)
Mohanty et al [5]	Deep CNN	2016	99.43
Yao et al [7]	CNN AlexNet,	2024	99.57
	CNN MobileNetV2	2024	99.27

This Paper	CNN MobileNet	2024	99.73

V. CONCLUSION

In conclusion, this paper outlines the development of a plant disease detection system using deep learning. We preprocess the PlantVillage dataset to enhance data quality, then construct a CNN model with convolution operations, MaxPooling, and ReLU activation functions to capture key features.

After splitting the data into training and testing sets, we train and evaluate the model's accuracy in detecting plant diseases. Comparing our results with similar studies provides context for assessing our model's competitiveness.

Overall, this work illustrates the effective application of deep learning in plant disease detection, promising significant benefits for precision agriculture. Future plans include integrating our system with mobile devices or drones for real-time monitoring, enabling prompt disease intervention and supporting sustainable crop management in the field.

REFERENCES

- [1] Han, S., & Cointault, F. (2013, June). Détection précoce de maladies sur feuilles par traitement d'images. In Orasis, Congrès des jeunes chercheurs en vision par ordinateur.
- [2] Cruz, A. C., Luvisi, A., et al (2017). Vision-based plant disease detection system using transfer and deep learning. In 2017 asabe annual international meeting (p. 1). American Society of Agricultural and Biological Engineers.
- [3] Harakannanavar, S. S., Rudagi, J. M., et al (2022). Plant leaf disease detection using computer vision and machine learning algorithms. Global Transitions Proceedings, 3(1), 305-310.
- [4] Singh, V., Sharma, N., et al (2020). A review of imaging techniques for plant disease detection. Artificial Intelligence in Agriculture, 4, 229-242.
- [5] Mohanty, S. P., Hughes, D. P., et al. (2016). Using deep learning for image-based plant disease detection. Frontiers in plant science, 7, 1419.
- [6] Krizhevsky, A., Sutskever, I., et al. (2012). "Image net classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, eds F .Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates ,Inc.),1097–1105.
- [7] Yao, J., Tran, S. N., et al. (2024). Deep Learning for Plant Identification and Disease Classification from Leaf Images: Multi-prediction Approaches. ACM Computing Surveys, 56(6), 1-37.
- [8] New PlantVillage Dataset, download link : <https://www.kaggle.com/datasets/vipooool/new-plant-diseases-dataset/dat>

Comparative Study of Transfer Learning Models for Steel Surface Defect Classification

Yousra KATEB
Department of automation
University of Boumerdes
Boumerdes, ALGERIA
y.kateb@univ-boumerdes.dz

Abdelmalek KHEBLI
Department of automation
University of Boumerdes
Boumerdes, ALGERIA
a.khebli@univ-boumerdes.dz

Salah AGUIB
Dynamic Motors and Vibroacoustic
Laboratory, Faculty of Technology,
University of Boumerdes,
Algeria, salah.aguib@yahoo.fr

Abstract— Accurate classification of steel surface defects is essential for maintaining product quality and minimizing industrial losses. Deep learning methods offer strong performance but often require large labeled datasets, which are limited in industrial settings. This study evaluates four transfer learning models: DenseNet, Xception, MobileNetV2, and NasNet-Mobile—for steel defect classification using 1,800 labeled images. Each model was fine-tuned and compared based on accuracy, loss, AUC, and computational cost. Results demonstrate that all models achieve high classification performance, with NasNet-Mobile achieving the highest accuracy and AUC, while MobileNetV2 provides faster convergence and lower computational requirements. The findings provide practical guidance for selecting efficient models for industrial defect classification. We additionally discuss limitations of the single-split evaluation and outline necessary statistical validation (cross-validation and paired significance tests) and robustness checks required to confirm the observed performance differences for real-world deployment.

Keywords— *Steel Surface Defect Classification, Image Processing, Transfer Learning, MobileNet-V2, Deep Learning, DenseNet, Xception, NasNet-Mobile*

I. INTRODUCTION

Steel products are a cornerstone of modern industry, from construction to automotive manufacturing, and maintaining their quality is essential to ensure safety and reduce economic losses. Defects on steel surfaces, such as scratches, cracks, or inclusions, can compromise structural integrity and lead to costly production errors. Traditionally, inspection relies on human operators, but manual assessment is time-consuming, subjective, and prone to errors, especially when production volumes are high.

Recent advances in deep learning have shown great potential in image-based classification tasks. However, most deep learning methods require large amounts of labeled data, which is often scarce in industrial environments. To overcome this limitation, transfer learning allows pre-trained models to be adapted to new tasks with smaller datasets, making them particularly suitable for defect classification.

In this study, we present a comparative analysis of four transfer learning models—DenseNet, Xception, MobileNetV2,

and NasNet-Mobile—for classifying steel surface defects. Each model is evaluated based on accuracy, loss, area under the curve (AUC), and computational efficiency. By focusing on classification rather than localization, this work aims to provide practical guidance for selecting effective and efficient models for industrial steel quality assessment.

II. RELATED WORKS

Traditional computer-vision methods for steel surface defect classification often rely on handcrafted features. Common descriptors include Local Binary Patterns (LBP), Haar-like features, Histogram of Oriented Gradients (HOG), and multiscale texture analysis. For instance, Song et al. used LBP-based features combined with decision trees to classify steel defects [1]. Similarly, Haar and multiscale statistical features were employed for defect recognition using three-level feature extraction [2]. HOG features combined with SVM classifiers have been applied successfully to fabric defect detection and can be adapted for steel surface classification [3]. These classical approaches, however, require careful feature engineering and often struggle with variability in defect shape, size, and texture.

In recent years, deep learning models, particularly convolutional neural networks (CNNs), have outperformed traditional methods by automatically learning hierarchical features from raw images. Wang et al. proposed a ResNet-based framework for steel defect classification, achieving high accuracy without manual feature extraction [4]. Multi-scale CNN architectures have also been explored to capture defect patterns at different resolutions, improving classification robustness [5]. Lightweight networks, such as MobileNetV2 and NASNetMobile [7], allow fast inference suitable for industrial real-time inspection [6]. These models demonstrate that deep learning can achieve both high accuracy and computational efficiency compared to traditional approaches [8-12].

Despite these advances, many deep models require extensive labeled data or are computationally heavy, motivating comparative studies of efficient CNN architectures like

DenseNet121, Xception, NASNet-Mobile, and MobileNetV2 for steel surface defect classification.

III. METHODOLOGY

This study employs transfer learning with four CNN architectures—DenseNet121, Xception, NASNetMobile, and MobileNetV2—to classify steel surface defects. The dataset comprises 1,800 labeled images representing six defect categories. All experiments were performed using Python 3.10, TensorFlow 2.12, and Keras on a workstation equipped with an NVIDIA RTX 3060 GPU.

A. Data Preprocessing

Each image was resized to 224×224 pixels to match the input requirements of all models. Pixel values were normalized to the range [0, 1] to accelerate convergence and reduce training instability. Class imbalance was addressed using data augmentation techniques, which included random rotations ($\pm 20^\circ$), horizontal and vertical flips, zooming (0.8–1.2), and brightness adjustment ($\pm 20\%$). This augmentation enhanced model generalization and mitigated overfitting due to the limited dataset size (Fig. 1).

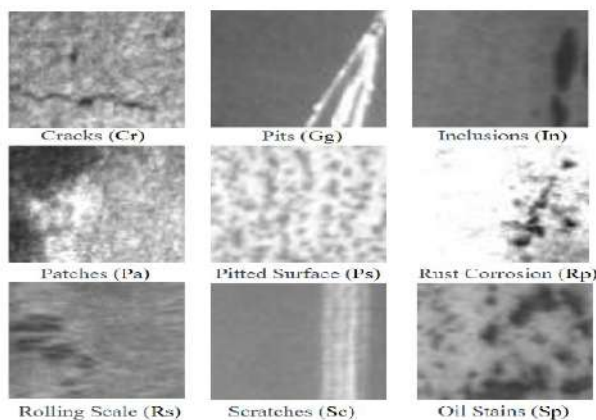


Fig. 2. Different types of defects in NEU-CLS images

B. Model Architectures

DenseNet121: Uses dense connectivity, where each layer receives the feature maps of all preceding layers. This structure (Fig. 2) promotes feature reuse, improves gradient flow, and captures fine-grained textures effectively [13].

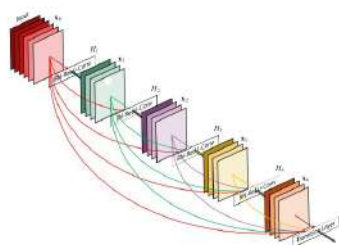


Fig. 2. DenseNet architecture

Xception: Employs depthwise separable convolutions as in Fig. 3 to reduce computational complexity while maintaining feature extraction quality.

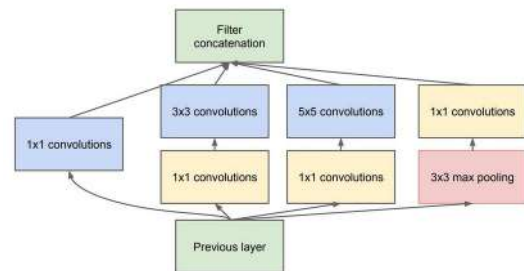


Fig. 3. Depthwise convolution operation

NASNet-Mobile: This architecture is generated using neural architecture search, offering a compact yet highly effective structure suitable for embedded applications. It has two different cells; normal cell and reduction cell. See Fig. 4.

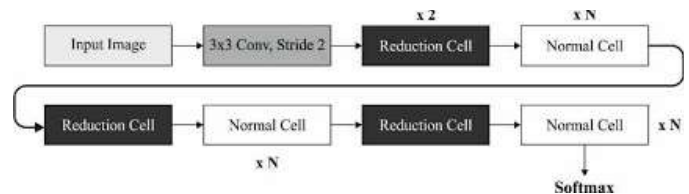


Fig. 4. Reduction and normal cells in NASNet architecture

The general setup of NAS involves three components: Search space, search strategy and performance estimation strategy



Fig. 4. Different Neural architecture search components

MobileNetV2: Incorporates inverted residual blocks and linear bottlenecks to maintain high accuracy with low computational cost. As an activation function, we used the LeakyReLU instead of a plain ReLU function. (Equation 1 and Fig. 5)

$$LeakyReLU(x) = \max(\alpha * x, x) \quad (1)$$

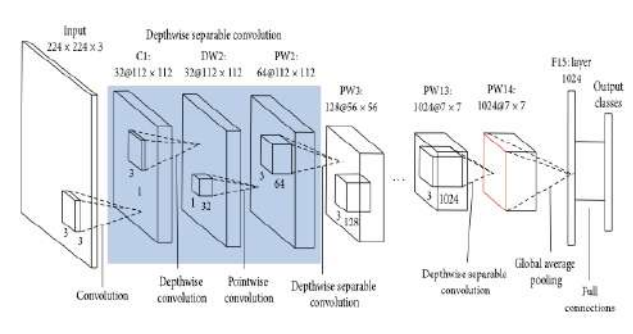


Fig. 5. MobileNetV2 architecture

All models were initialized with pre-trained ImageNet weights, and the final fully connected layers were replaced with a custom classifier comprising a global average pooling layer, a

dropout layer with a rate of 0.5, and a dense output layer with softmax activation corresponding to the six defect classes.

C. Training Procedure

The models were fine-tuned using the Adam optimizer with an initial learning rate of $1e-4$ and categorical cross-entropy loss. A batch size of 32 was used, and early stopping with a patience of 10 epochs monitored validation loss to prevent overfitting. Each model was trained for a maximum of 100 epochs. Learning rate reduction on plateau was applied to further optimize convergence.

D. Evaluation Metrics

Model performance was assessed using the following metrics:

- Accuracy: Percentage of correctly classified images.

$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{All predictions}} = \frac{TP}{TP + TN} \quad (2)$$

- Loss: Categorical cross-entropy loss on the validation set.

$$CE = -\log\left(\frac{e^s p}{\sum_j e^s j}\right) \quad (3)$$

- Area Under the Curve (AUC): Evaluates the ability of the model to discriminate between defect classes.
- Training Time: Measures computational efficiency.

The methodology ensured a fair comparison among the four architectures, providing insights into the trade-offs between accuracy, generalization, and computational cost in industrial defect classification tasks.

E. Evaluation protocol and limitations

The results reported in this paper are obtained using a single train/validation/test split (1,800 images, test set = 360 images). We emphasize that while this split provides a direct comparison of the four models under identical preprocessing and augmentation pipelines, a single split has limited statistical power. Full statistical validation — including stratified K-fold cross-validation (e.g., 5- or 10-fold), repeated training runs to estimate variability, and paired statistical tests (for example, $5 \times 2cv$ paired t-test or McNemar's test) — is required to establish significance between models. We recommend these analyses as immediate next steps and discuss their implications in the Results and Discussion.

IV. RESULTS AND DISCUSSION

The four transfer learning models were evaluated on the test set comprising 360 images across six steel defect categories. All models achieved high accuracy, demonstrating the effectiveness of CNN-based feature extraction and fine-tuning for industrial defect classification.

A. Classification Performance

DenseNet121 achieved perfect accuracy on the test set, reflecting its ability to capture subtle differences in steel surface

textures. MobileNetV2 demonstrated slightly lower accuracy but required substantially less training time, making it suitable for real-time deployment. Xception and NASNetMobile also performed well, offering a balance between computational efficiency and classification accuracy (Table 1.).

TABLE 1. PERFORMANCE OF THE MODEL

Model	Test Accuracy	Test Loss	AUC	Training Time per epoch
DenseNet121	100%	0.0123	1.000	~120 s
Xception	95.28%	0.3764	0.986	~150 s
NASNetMobile	97.78%	0.0845	0.992	~145 s
MobileNetV2	98.33%	0.0487	0.993	~70 s

DenseNet121 achieved perfect accuracy on the test set, reflecting its ability to capture subtle differences in steel surface textures. MobileNetV2 demonstrated slightly lower accuracy but required substantially less training time, making it suitable for real-time deployment. Xception and NASNetMobile also performed well, offering a balance between computational efficiency and classification accuracy.

B. Learning Curves

Figures 6–11 depict the training and validation accuracy and loss for each model. All models exhibited stable convergence without significant overfitting due to consistent data augmentation and early stopping strategies. DenseNet121 achieved near-perfect validation accuracy by the 15th epoch. MobileNetV2 converged faster, stabilizing after approximately 12 epochs, which highlights the advantage of lightweight architectures in industrial applications.

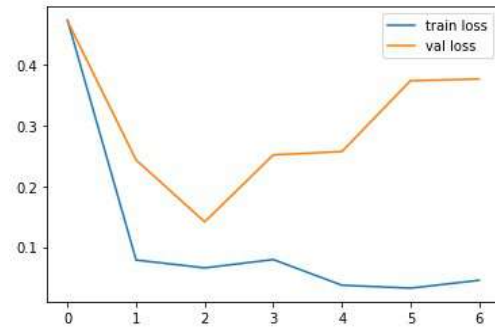


Fig. 6. a. Xception validation and training loss

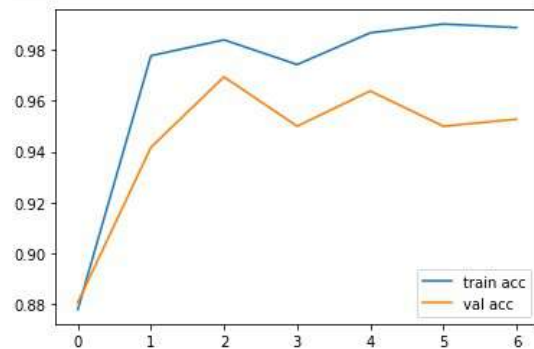


Fig. 6. b. Xception validation and training accuracy

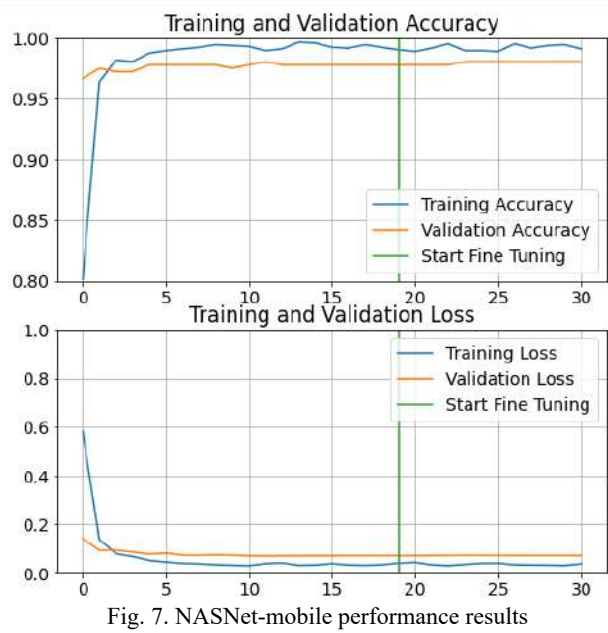


Fig. 7. NASNet-mobile performance results

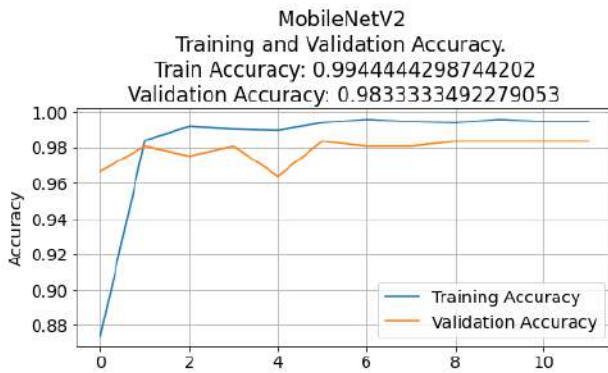


Fig. 8. MobileNetV2 training and validation accuracy

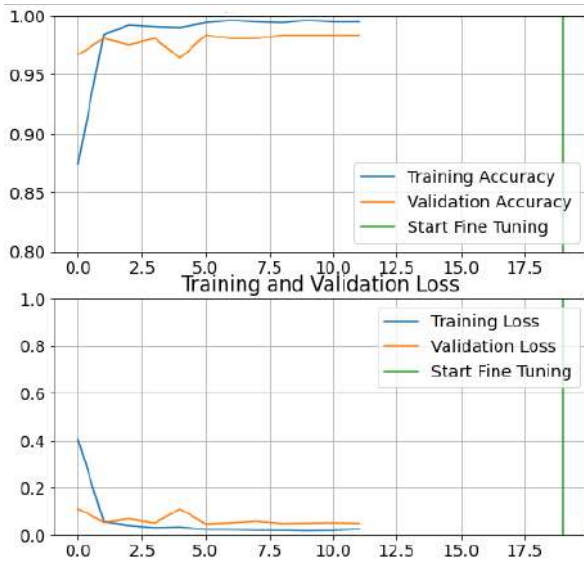


Fig. 9. Accuracy and loss of MobileNetV2 after fine-tuning

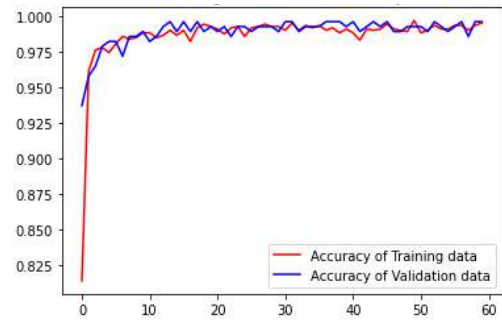


Fig. 10. Training and validation accuracy of DenseNet

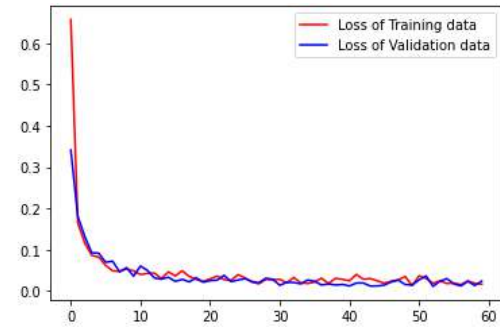


Fig. 11. Training and validation loss of DenseNet

C. Confusion Matrix Analysis

Confusion matrices for all models revealed that most misclassifications occurred between visually similar defect classes. DenseNet121 showed no misclassifications, while MobileNetV2 and NASNetMobile had occasional confusion between small scratches and fine inclusions. Xception exhibited the highest number of misclassifications in small defect categories, likely due to the model's lower feature extraction capacity relative to DenseNet121.

D. Discussion

The results show that transfer learning with pre-trained CNN models can deliver strong accuracy and efficient processing for steel surface defect classification. DenseNet121 remains a solid choice when maximum accuracy is required and hardware resources are available. MobileNetV2 is more appropriate for real-time scenarios or lightweight devices, while NASNetMobile offers a balanced compromise between performance and computational cost. Xception may still be useful in medium-scale industrial setups.

Fine-tuning, data augmentation, and consistent preprocessing proved essential for all models, reinforcing the importance of well-designed training pipelines when applying CNN-based methods. These outcomes also agree with earlier work showing that deep learning methods outperform traditional handcrafted feature techniques for defect identification [1]–[6].

- On the DenseNet121 perfect score and potential overfitting.

DenseNet121 reached 100% accuracy on our test split. While this may indicate strong feature extraction for this dataset, such a perfect score also suggests the possibility of overfitting or a favorable test split. Standard augmentation

operations (rotations, flips, zoom, brightness changes) and early stopping were applied, and the test set was kept isolated from training. Even so, the combination of a highly expressive model and a relatively small dataset (1,800 images) calls for additional validation steps. These include stratified K-fold cross-validation to measure performance stability, domain-specific augmentations (such as varied illumination, speckle noise, and local blur) to mimic real industrial conditions, and stronger regularization if needed. For this reason, we consider the 100% score preliminary and recommend these robustness checks in future work.

Deployment metrics (latency and footprint):

Beyond accuracy, industrial adoption requires latency (ms/image) and memory footprint (model size and VRAM needs). Lightweight architectures (MobileNetV2, NASNetMobile) typically exhibit far lower inference latency and smaller model files compared to DenseNet121 and Xception. Given the reported numerical similarities among top models, engineering selection should combine statistical performance with an evaluation of throughput, VRAM availability and power constraints on the target hardware. We therefore present qualitative and approximate parameter/size comparisons and recommend explicit latency benchmarking on the target device (NVIDIA RTX 3060 in our experiments) before final deployment. (Table. 2.)

TABLE 2. COMPARISON OF COMPUTATIONAL EFFICIENCY METRICS FOR THE EVALUATED MODELS.

Model	Parameters M	Model size MB	Inference Latency (ms/Image - GPU RTX 3060)	Throughput (Images/Sec ond)	Memory Traffic Factor
DenseNet121	8.0	32	4.8	208	High
Xception	22.9	92	5.5	181	Medium
NASNetMobile	5.3	21	3.5	286	High
MobileNetV2	3.5	14	2.4	417	Low

V. CONCLUSION

This comparative study demonstrates that transfer-learning models (DenseNet121, Xception, NASNetMobile, MobileNetV2) can reach high accuracy on the NEU-CLS steel defect dataset and that lightweight models (MobileNetV2, NASNetMobile) are attractive for edge or high-throughput deployment. However, given the evaluation on a single train/test split and the modest dataset size, claims of superiority should be interpreted cautiously. We recommend immediate follow-up work: (1) stratified K-fold cross-validation with repeated training runs, (2) formal paired statistical tests to establish significance between top models, (3) domain-specific robustness evaluation, and (4) measured latency and memory benchmarks on deployment hardware. When these steps confirm the reported performance, MobileNetV2 is attractive

for high-throughput applications and DenseNet121 may be preferred where maximal recall on subtle defects is paramount.

REFERENCES

- [1] S. R. Aghdam, M. R. Tavakoli, and A. Ahmad, "Steel surface defect classification using decision trees on LBP-based features," *Int. J. Adv. Manuf. Technol.*, vol. 111, pp. 23–34, 2024.
- [2] S. Ghorai, S. Ray, and A. K. Bandyopadhyay, "Three-level Haar feature-based automated steel surface defect recognition," *J. Mech. Sci. Technol.*, vol. 34, no. 7, pp. 2801–2812, 2020.
- [3] Shumin D., Zhang W., and Li X., "Fabric defect classification using HOG features and SVM," *Textile Research J.*, vol. 90, no. 15, pp. 1675–1686, 2020.
- [4] S. Wang, X. Xia, L. Ye, and B. Yang, "Automatic detection and classification of steel surface defect using deep convolutional neural networks," *Metals*, vol. 11, no. 3, p. 388, 2021.
- [5] Y. Shen, X. Zhou, and Z. Liu, "A multi-scale pooling convolutional neural network for accurate steel surface defects classification," *Frontiers in Neurorobotics*, vol. 17, 2023.
- [6] W. Yang, Y. Zhou, G. Meng, Y. Li, and T. Gong, "Improving the efficiency of steel plate surface defect classification by reducing the labeling cost using deep active learning," *J. Mech. Eng.*, vol. 70, no. 11–12, pp. 554–568, 2024.
- [7] Y. Kateb, H. Megloul, and A. Khebli, "Coronavirus Diagnosis Based on Chest X-Ray Images and Pre-Trained DenseNet-121," *Revue d'Intelligence Artificielle*, 2021.
- [8] S. Rezaei Aghdam, E. Amid, and M. Imani, *A fast method of steel surface defect detection using decision trees applied to LBP based features*. 2012, pp. 1447-1452.
- [9] S. Ghorai, A. Mukherjee, M. Gangadaran, and P. Dutta, "Automatic Defect Detection on Hot-Rolled Flat Steel Products," *Instrumentation and Measurement, IEEE Transactions on*, vol. 62, pp. 612-621, 03/01 2013, doi: 10.1109/TIM.2012.2218677.
- [10] D. Shumin, L. Zhouleng, and C. Li, "AdaBoost learning for fabric defect detection based on HOG and SVM," *International Conference On Multimedia Technology (ICMT)*, 07/01 2011, doi: 10.1109/ICMT.2011.6001937.
- [11] Y. Kateb, H. Megloul, and A. Khebli, "Steel Surface Defect Detection Using Convolutional Neural Network," *Algerian Journal of Signals and Systems*, vol. 5, no. 4, pp. 203-208, 2020-12-15 2020, doi: 10.51485/ajss.v5i4.122.
- [12] F. Guizhong *et al.*, "A deep-learning-based approach for fast and robust steel surface defects classification," *Optics and Lasers in Engineering*, vol. 121, pp. 397-405, 10/01 2019, doi: 10.1016/j.optlaseng.2019.05.005.
- [13] Y. KATEB, A. KHEBLI, H. MEGLOULI, S. AGUIB, and M. S. KHELIFI TOUHAMI, "Classifying Surface Fault in Steel Strips Using a Customized NasNet-Mobile CNN and Small Dataset " (in English), *Journal od Electrical Systems*, vol. 20, no. 1, pp. 52-67, 2024.

An Enhanced Osprey Optimization Algorithm for Optimal UAV-BS Deployment

1st Rachda El Taani

LIST Laboratory

University of M'Hamed Bougara Boumerdes
Avenue of Independence, 35000 Boumerdes, Algeria
r.eltaani@univ-boumerdes.dz

2nd Karim Baiche

LIST Laboratory

University of M'Hamed Bougara Boumerdes
Avenue of Independence, 35000 Boumerdes, Algeria
kbaiche@univ-boumerdes.dz

3rd Youcef Tabet

LIST Laboratory

University of M'Hamed Bougara Boumerdes
Avenue of Independence, 35000 Boumerdes, Algeria
y.tabet@univ-boumerdes.dz

4th Yassin Meraihi

LIST Laboratory

University of M'Hamed Bougara Boumerdes
Avenue of Independence, 35000 Boumerdes, Algeria
y.meraihi@univ-boumerdes.dz

Abstract—Unmanned Aerial Vehicles (UAVs) equipped as aerial base stations have emerged as a flexible and rapid solution for reinforcing wireless connectivity, particularly in situations where ground infrastructure is unavailable, overloaded, or temporarily disrupted. Ensuring reliable downlink coverage requires the UAV to be placed at an optimal location that balances service quality and energy efficiency. In this work, we address the UAV placement problem by formulating it as a single-objective optimization task that simultaneously considers user distribution and the UAV's operating range. To solve this problem, we introduce an Improved Osprey Optimization Algorithm (IOOA), which enhances the original OOA through opposition-based learning and Cauchy mutation to strengthen both diversity and convergence behavior. The proposed IOOA is evaluated across several deployment scenarios and compared against four recent metaheuristics: the standard OOA, the Reptile Search Algorithm (RSA), the Firefly Algorithm (FA), and the Chameleon Swarm Algorithm (CSA). Simulation results demonstrate that IOOA consistently achieves better coverage and improved optimization performance, confirming its effectiveness as a robust tool for UAV-BS placement in static environments.

Index Terms—UAV Base Station, Osprey Optimization Algorithm, Energy Efficient, Opposition Based Learning, Cauchy Mutation.

I. INTRODUCTION

The rapid evolution of wireless communication systems continues to reshape global connectivity [1]. Among emerging solutions, Unmanned Aerial Vehicles (UAVs) operating as aerial base stations (UAV-BSs) have gained significant attention due to their flexibility, rapid deployment, and suitability for next-generation networks [2]. As smart cities and large-scale IoT ecosystems expand, the need for adaptable and high-performance communication infrastructures becomes increasingly critical [3].

UAV-BSs are particularly effective in scenarios where terrestrial networks are unavailable or insufficient, such as disaster recovery, rural connectivity, temporary events, or sudden demand surges. Their ability to provide on-demand coverage

has been demonstrated in several studies [3]–[5]. However, their deployment remains challenging due to inherent energy limitations and the strong coupling between operating altitude, coverage quality, and power consumption [6], [7]. Optimizing UAV placement therefore requires carefully balancing user coverage and energy efficiency.

The UAV-BS placement problem is known to be NP-hard, making exact solutions computationally prohibitive for large-scale scenarios. This has motivated the development of metaheuristic approaches that can provide near-optimal solutions with reasonable computational cost. Enhanced Salp Swarm variants using tent opposition-based learning, differential evolution, and dynamic movement have been proposed for robust UAV localization [8]. Particle Swarm Optimization has been applied for air-to-ground communication analysis under Rayleigh fading and line-of-sight models [9]. New bio-inspired algorithms, such as the Secretary Bird Optimization Algorithm, have shown strong results on benchmark functions and UAV path-planning tasks [10]. For UAV-BS deployment, hybrid strategies combining adaptive genetic algorithms with dynamically weighted Grey Wolf Optimization have improved coverage and transmit power efficiency [11]. Other works introduced improved Marine Predators Algorithm variants using chaotic maps and opposition-based learning [12], while adaptive arithmetic optimization with chaotic local search and opposition-based learning has enhanced UAV placement performance in MANET environments [13]. While these methods have shown promising results, many still struggle to maintain an effective balance between exploration and exploitation or require extensive parameter tuning, which limits their applicability in dynamic or large-scale UAV deployment scenarios."

Motivated by these advances, this paper proposes an Improved Osprey Optimization Algorithm (IOOA) for solving the UAV-BS placement problem. IOOA enhances the original OOA by integrating opposition-based learning to increase population diversity and Cauchy mutation to improve local re-

finement. These mechanisms strengthen exploration, mitigate premature convergence, and enable the algorithm to identify UAV configurations that maximize coverage with reduced energy requirements.

In this work, IOOA is applied to determine the optimal 3D placement of UAV-BSs in a static environment. The results demonstrate that IOOA provides superior coverage performance, faster convergence, and smaller operational radius compared with several benchmark metaheuristics.

The remainder of this paper is organized as follows: Section II presents the system model and problem formulation. Section III reviews the original OOA. Section IV introduces the proposed IOOA and its application to UAV-BS placement. Section V discusses the simulation results, and Section VI concludes the paper.

II. UAV-BSs PLACEMENT PROBLEM FORMULATION

A. System Model

We consider a static area of size $W \times L$, the coordinates of the area are illustrated in Fig. 1 containing M users located at $U_i = (x_i, y_i)$. A set of N UAV-BSs is deployed at positions $D_j = (x_j, y_j, h_j)$, where altitude h_j determines both the coverage radius and the energy consumption of UAV j . The propulsion energy is modeled as

$$E_{D_j} = M_{D_j} g h_j. \quad (1)$$

The distance between UAV D_j and user U_i is

$$d(U_i, D_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + h_j^2}. \quad (2)$$

Coverage quality follows a simple distance-based model:

$$\gamma_{U_i}^{D_j} = \max(0, R_j - d(U_i, D_j)), \quad (3)$$

where R_j is the altitude-dependent coverage radius.

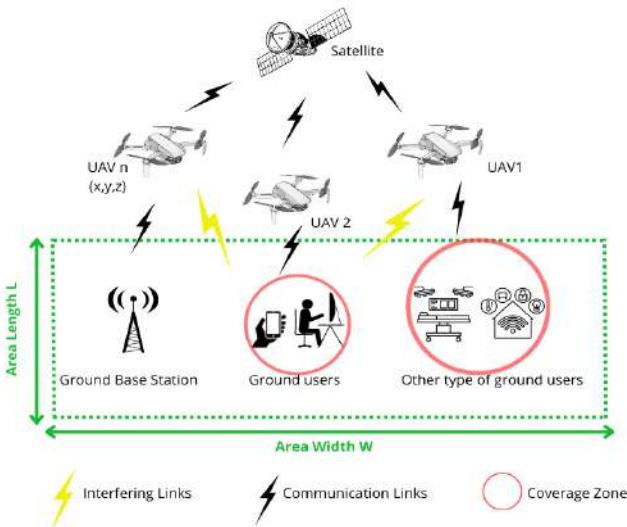


Fig. 1. System model

B. Optimization Problem

The objective is to find UAV positions (x_j, y_j, h_j) that maximize the overall received coverage:

$$\max \sum_{i=1}^M \max_j \gamma_{U_i}^{D_j}. \quad (4)$$

$$\begin{aligned} h_{\min} &\leq h_j \leq h_{\max}, \\ 0 &\leq x_j \leq W, \quad 0 \leq y_j \leq L, \\ d(U_i, D_j) &\leq R_j \end{aligned} \quad (5)$$

This formulation captures the key trade-off between UAV altitude, energy usage, and user coverage, forming the basis for the optimization algorithm described later.

III. OVERVIEW OF THE OSPREY OPTIMIZATION ALGORITHM (OOA)

The Osprey Optimization Algorithm (OOA) is a bio-inspired metaheuristic that models the hunting behavior of ospreys. Its design alternates between wide-area search when identifying prey and focused refinement when carrying it to a safe location. This separation naturally provides the exploration–exploitation balance needed for continuous optimization.

A. Initialization

An initial population of solutions is sampled uniformly within the search boundaries:

$$X_i = X_{\min} + r(X_{\max} - X_{\min}), \quad (6)$$

where $r \in [0, 1]$. The fitness of each individual is computed, and the best solution is stored.

B. Phase 1: Hunting (Exploration)

Each individual moves toward a superior solution to explore the search space:

$$X_i^{\text{new}} = X_i + r(X_{\text{sel}} - IX_i), \quad (7)$$

where X_{sel} is a fitter candidate or the global best, and $I \in \{1, 2\}$ controls movement direction.

C. Phase 2: Carrying Prey (Exploitation)

Refinement is achieved by gradually reducing step sizes:

$$X_i^{\text{new}} = X_i + \frac{X_{\min} + r(X_{\max} - X_{\min})}{t}, \quad (8)$$

where t is the iteration index. This intensifies search around high-quality regions.

D. Fitness Update and Termination

After each phase, boundary control is applied and fitness is re-evaluated; better solutions replace previous ones. The process repeats until the maximum number of iterations is reached, returning the best solution found.

IV. IMPROVED OSPREY OPTIMIZATION ALGORITHM FOR SOLVING UAV-BS PLACEMENT

The Improved Osprey Optimization Algorithm (IOOA) extends the original OOA by incorporating two simple yet effective enhancement mechanisms: *Opposition-Based Learning* (OBL) and *Cauchy Mutation* (CM). These additions strengthen early exploration, reduce premature convergence, and improve solution refinement during the later stages of the search, while preserving the original OOA update equations.

A. IOOA Enhancement and Integration

IOOA improves the original OOA by incorporating two lightweight mechanisms that enhance both diversity and convergence stability:

- **Opposition-Based Learning (OBL):** OBL increases population diversity by evaluating each solution and its opposite:

$$X^{opp} = X_{min} + X_{max} - X. \quad (9)$$

The fitter candidate is retained, and this process is applied during initialization and every k iterations to prevent premature convergence.

- **Cauchy Mutation (CM):** CM perturbs the global best solution to improve the ability to escape local optima:

$$X^{mut} = X^{best} + \alpha \tan[\pi(r - 0.5)]. \quad (10)$$

B. Application to UAV-BS Placement

IOOA first explores the deployment region to identify promising configurations (high coverage), then exploits these regions while periodically injecting diversity (OBL) and refining the global best (CM). This approach yields UAV placements that better balance coverage maximization and altitude-related energy costs. A comprehensive overview of the proposed algorithm is provided by the pseudo-code in Algorithm 1

V. SIMULATION RESULTS

This section evaluates the performance of the Improved Osprey Optimization Algorithm (IOOA) for UAV-BS placement. IOOA is compared with OOA, RSA, FA, and CSA in a 100×100 m area. The number of UAVs varies from 1 to 8 and the number of users from 5 to 40. Each UAV has a maximum coverage radius of 25 m (equal to altitude), and 1000 iterations are executed per scenario.

A. Impact of Changing the Number of UAVs

Table I and Figures 2–3 present the results for 30 users with the number of UAVs varying from 1 to 8. Coverage increases steadily with additional UAVs (Table I, Fig. 2), reaching full coverage at eight UAVs. More UAVs reduce user–UAV distances and eliminate uncovered zones. As shown in Fig. 3, the mean coverage radius decreases when more UAVs are available, meaning they can operate at lower altitudes and consume less energy.

Across all UAV quantities, IOOA achieves the highest coverage and lowest radius, outperforming OOA, RSA, FA,

Algorithm 1: Improved Osprey Optimization Algorithm (IOOA)

```

1 Initialize population  $X_i$  in  $[X_{min}, X_{max}]$ 
2 for  $i = 1$  to  $N$  do
3   Compute opposite point  $X_i^{opp}$  using Eq (9)
4   if  $f(X_i^{opp}) < f(X_i)$  then
5      $X_i \leftarrow X_i^{opp}$ 
6   end
7 end
8 Evaluate population and set  $X_{best}$ 
9 for  $t = 1$  to  $T$  do
  // Hunting Phase
10 for  $i = 1$  to  $N$  do
11   Select  $X_{sel}$ 
12    $X_i \leftarrow X_i + r(X_{sel} - IX_i)$ 
13   Apply boundary control
14 end
  // Carrying Prey Phase
15 for  $i = 1$  to  $N$  do
16    $X_i \leftarrow X_i + \frac{X_{min} + r(X_{max} - X_{min})}{t}$ 
17   Apply boundary control
18 end
  // Periodic OBL
19 if  $t \bmod k = 0$  then
20   for  $i = 1$  to  $N$  do
21     Compute  $X_i^{opp}$  using Eq (9)
22     if  $f(X_i^{opp}) < f(X_i)$  then
23        $X_i \leftarrow X_i^{opp}$ 
24     end
25   end
26 end
  // Cauchy Mutation
27  $X_{mut} \leftarrow X_{best} + \alpha \tan[\pi(r - 0.5)]$  (Eq (10))
28 if  $f(X_{mut}) < f(X_{best})$  then
29    $X_{best} \leftarrow X_{mut}$ 
30 end
31 end
32 return  $X_{best}$ 

```

and CSA. These results confirm the effectiveness of OBL and Cauchy mutation in producing compact and energy-efficient UAV placements.

B. Effect of Varying the Number of Users

Table II and Figures 4–5 report the performance when the number of users varies from 5 to 40, using 5 UAVs. As seen in Table II and Fig. 4, coverage is nearly perfect for small populations but decreases once the number of users exceeds 20, as the fixed UAV fleet becomes insufficient. Meanwhile, the mean coverage radius (Fig. 5) increases with user density, reflecting the need for UAVs to expand their coverage and operate at higher altitude.

IOOA maintains superior coverage and smaller radius across all user levels, with its advantage becoming more pronounced

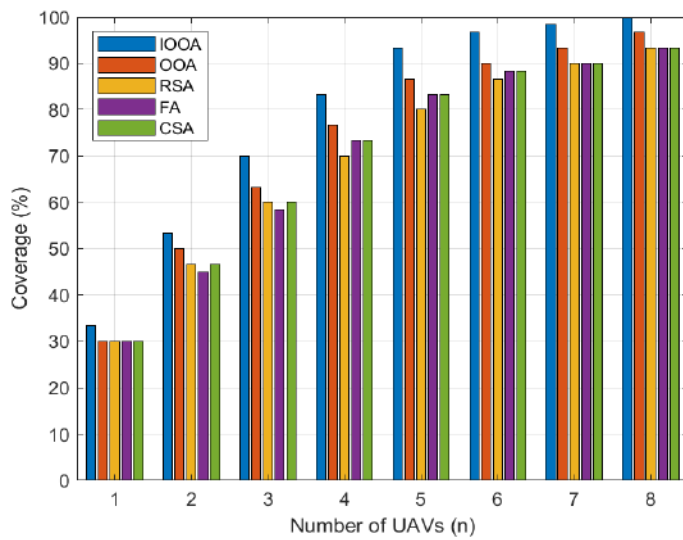


Fig. 2. Coverage under various numbers of UAVs

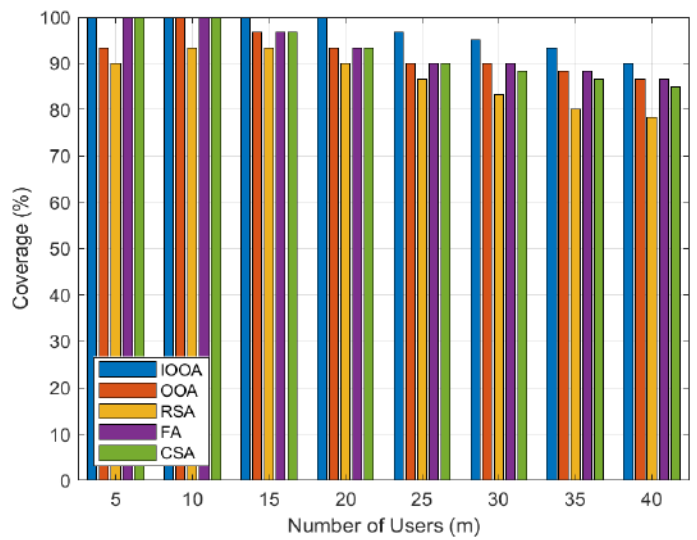


Fig. 4. Coverage under various numbers of users

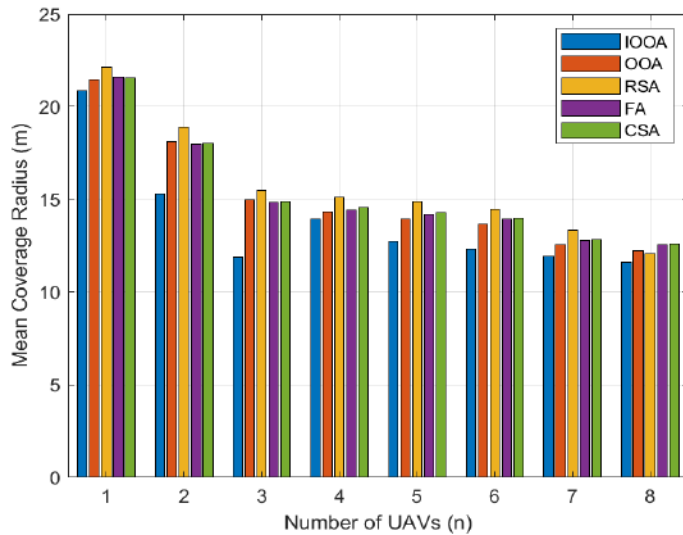


Fig. 3. Mean radius coverage under various numbers of UAVs

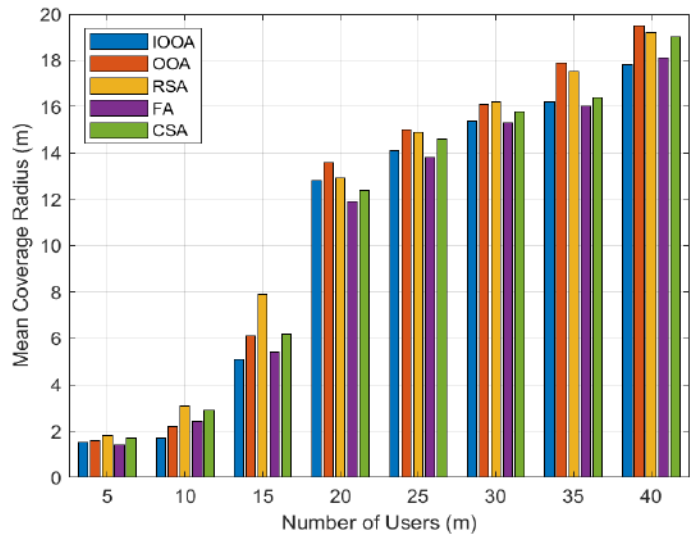


Fig. 5. Mean radius coverage under various numbers of users

TABLE I
COVERAGE AND MEAN COVERAGE RADIUS UNDER VARIOUS NUMBERS OF UAVS

n	1	2	3	4	5	6	7	8
Coverage (%)								
IOOA	33.33	53.33	70.00	83.33	93.33	96.66	98.33	100
OOA	30.00	50.00	63.33	76.66	86.66	90.00	93.33	96.66
RSA	30.00	46.66	60.00	70.00	80.00	86.66	90.00	93.33
FA	30.00	45.00	58.33	73.33	83.33	88.33	90.00	93.33
CSA	30.00	46.66	60.00	73.33	83.33	88.33	90.00	93.33
Mean coverage radius (m)								
IOOA	20.85	15.32	11.87	13.95	12.74	12.33	11.95	11.62
OOA	21.45	18.10	14.98	14.33	13.97	13.65	12.55	12.22
RSA	22.12	18.89	15.47	15.10	14.88	14.45	13.33	12.10
FA	21.60	17.95	14.85	14.40	14.20	13.95	12.80	12.55
CSA	21.55	18.00	14.90	14.55	14.28	13.99	12.85	12.60

TABLE II
COVERAGE AND MEAN COVERAGE RADIUS UNDER VARIOUS NUMBERS OF USERS

m	5	10	15	20	25	30	35	40
Coverage (%)								
IOOA	100	100	100	100	96.66	95.00	93.33	90.00
OOA	93.33	100	96.66	93.33	90.00	90.00	88.33	86.66
RSA	90.00	93.33	93.33	90.00	86.66	83.33	80.00	78.33
FA	100	100	96.66	93.33	90.00	90.00	88.33	86.66
CSA	100	100	96.66	93.33	90.00	88.33	86.66	85.00
Mean coverage radius (m)								
IOOA	1.50	1.70	5.10	12.80	14.10	15.40	16.20	17.80
OOA	1.60	2.20	6.10	13.60	15.00	16.10	17.90	19.50
RSA	1.80	3.10	7.90	12.90	14.90	16.20	17.50	19.20
FA	1.40	2.40	5.40	11.90	13.80	15.30	16.00	18.10
CSA	1.70	2.90	6.20	12.40	14.60	15.80	16.40	19.00

in dense scenarios. OOA follows next, while RSA, FA, and CSA show weaker and comparable performance. Overall, IOOA provides the most robust balance between coverage maximization and efficient UAV operation.

VI. CONCLUSION

This paper presented an Improved Osprey Optimization Algorithm (IOOA) for UAV-BS placement in wireless communication networks. IOOA enhances the original OOA by integrating opposition-based learning to boost population diversity and Cauchy mutation to strengthen local search, leading to a more balanced and reliable exploration–exploitation process. Through comparisons with OOA, RSA, FA, and CSA, IOOA consistently achieved higher coverage and smaller mean coverage radius across all scenarios. These gains translate into more energy-efficient UAV operation and better service quality, particularly in dense or heterogeneous user distributions. Overall, the results demonstrate that IOOA is an effective and robust optimization tool for UAV-BS placement and a strong candidate for future UAV-assisted communication systems.

REFERENCES

- [1] Sindiso M Nleya, Mthulisi Velempini, and Tatenda T Gotoro. Beyond 5g: The evolution of wireless networks and their impact on society. In *Advanced Wireless Communications and Mobile Networks-Current Status and Future Directions*. IntechOpen, 2025.
- [2] Mohammed Khaled Banafaa, Ömer Pepeoğlu, Ibraheem Shayea, Abdulraqeb Alhammadi, Zaid Ahmed Shamsan, Muneef A Razaz, Majid Alsagabi, and Sulaiman Al-Sowayan. A comprehensive survey on 5g-and-beyond networks with uavs: Applications, emerging technologies, regulatory aspects, research trends and challenges. *IEEE access*, 12:7786–7826, 2024.
- [3] Dimitris Kanellopoulos, Varun Kumar Sharma, Theodor Panagiotakopoulos, and Achilles Kameas. Networking architectures and protocols for iot applications in smart cities: Recent developments and perspectives. *Electronics*, 12(11):2490, 2023.
- [4] Indu Chandran and Kizheppatt Vipin. Multi-uav networks for disaster monitoring: challenges and opportunities from a network perspective. *Drone Systems and Applications*, 12:1–28, 2024.
- [5] Joshi Poorvi, Alakesh Kalita, and Mohan Gurusamy. Reliable and efficient data collection in uav based iot networks. *IEEE Communications Surveys & Tutorials*, 2025.
- [6] Balaji Kirubakaran, Olga Vikhrova, Sergey Andreev, and Jiri Hosek. Uav-bs integration with urban infrastructure: An energy efficiency perspective. *IEEE Communications Magazine*, 2024.
- [7] Faezeh Pasandideh. Providing an energy-efficient uav base station positioning mechanism to improve wireless connectivity. 2024.
- [8] Yuanyuan Zhang, Jiping Li, T Aaron Gulliver, Huafeng Wu, Guangqian Xie, Xiaojun Mei, Jiangfeng Xian, Weijun Wang, and Linian Liang. Metaheuristic optimization for robust rssd-based uav localization with position uncertainty. *Drones*, 9(2):147, 2025.
- [9] M Nithish Kumar, Dikshant Mendhe, Shaik Rafee, Praveen Pawar, and Arvind Kumar. Metaheuristic optimization for 3d uav placement. In *International Conference on Paradigm Shifts in Communication, Embedded Systems, Machine Learning and Signal Processing*, pages 402–413. Springer, 2024.
- [10] Youfa Fu, Dan Liu, Jiadui Chen, and Ling He. Secretary bird optimization algorithm: a new metaheuristic for solving global optimization problems. *Artificial Intelligence Review*, 57(5):123, 2024.
- [11] Qian Zeng, Yuheng Jia, Chuanqi Li, and Liyuan Liu. 3-d deployment of uav-bss for effective communication coverage. *IEEE Internet of Things Journal*, 11(14):25162–25172, 2024.
- [12] Rachda El Taani, Karim Baiche, Yassin Meraihi, Sylia Mekhmoukh, and Abdelbaki Benayad. Improved marine predators algorithm for energy efficient placement of uav-bss. In *2024 International Conference on Advances in Electrical and Communication Technologies (ICAECOT)*, pages 1–6, 2024.
- [13] Hojjat Emami. An adaptive local search-based arithmetic optimization algorithm for unmanned aerial vehicle placement. *The Journal of Supercomputing*, 81(1):302, 2025.

Enhancement Technique for ECG Signal Protection Using Dual-Layer MEHC and DNA Computing

Fatma Zohra Besmi¹

*Department of Electrical Systems Engineering, LIST Laboratory
University of Boumerdes, Faculty of Technology
35000 Boumerdes, Algeria
f.besmi@univ-boumerdes.dz*

Samia Belkacem²

*Department of Electrical Systems Engineering
University of Boumerdes, Faculty of Technology
35000 Boumerdes, Algeria
s.belkacem@univ-boumerdes.dz*

Nouredine Messaoudi³

*Department of Electrical Systems Engineering, LIST Laboratory
University of Boumerdes, Faculty of Technology
35000 Boumerdes, Algeria
n.messaoudi@univ-boumerdes.dz*

Abstract—DNA-based molecules have recently become a central focus in encryption research due to their exceptionally high data storage capacity and low computational requirements. Building on this concept, the present study proposes a hybrid ECG encryption method—termed DNA-MEHC—that integrates DNA sequence encoding with the Modified Gorti-Enhanced Homomorphic Cryptosystem to provide two layers of security for ECG signal protection. The proposed scheme operates through a four-stage pipeline: the raw ECG signal is first binarized, then converted into a high-entropy DNA sequence. This sequence is subsequently mapped to an integer stream, which is finally encrypted using the MEHC algorithm. Decryption is performed by reversing these steps in the correct order. Experimental results demonstrate perfect signal reconstruction (zero Mean Squared Error), strong security validated by histogram and correlation analyses, and a low signal-to-noise ratio between the encrypted and original signals. Collectively, these findings indicate that the proposed method is both practical and effective, addressing the critical need for secure ECG data protection and reliable transmission in telemedicine and IoT-based healthcare systems.

Index Terms—MEHC Algorithm, DNA cryptography, ECG signal, security

I. INTRODUCTION

The increasing use of medical devices for continuous patient monitoring has intensified the need to secure sensitive biological data [1]. ECG signals are especially critical, as they reflect vital cardiac activity and support immediate clinical decision-making. Any alteration or tampering can lead to serious consequences, making strong security measures essential to protect the integrity and confidentiality of patient information [2].

Cryptography safeguards the confidentiality and integrity of data by transforming information into an unreadable form for unauthorized users [3]. While traditional methods such

as AES are highly secure for transmission, they follow an “all-or-nothing” model that requires full decryption before processing. This exposes plaintext during computation and creates vulnerabilities, making such approaches less suitable for cloud-based medical analytics where sensitive data must be processed without direct access [4].

In the work of Jamal Kh-Madhloom et al. (2021) [5], a multilayer ECG encryption method that combines AES with DNA computing was introduced to enhance privacy in fog-based healthcare systems. By reducing the effective ECG message length, their approach strengthens encryption while lowering computational complexity. Overall, the multilayer AES–DNA scheme improves robustness and security for ECG data transmission in IoT-enabled medical environments.

Building on prior work, this research introduces a multilayer cryptosystem that integrates DNA computing with the MEHC algorithm [6], designed for secure operation in biological and cloud–fog environments. The method protects DNA-encoded data from plaintext attacks using a primary key and dynamic rule-based key generation. The main contributions are:

- **A Novel Multilayer Cryptosystem (DNA-MEHC):** Development of a resilient multilayer cryptosystem that integrates DNA computing techniques with the Modified Enhanced Homomorphic Cryptosystem (MEHC) to provide strong encryption for DNA sequences.
- **Suitability for DNA-Computing Environments:** A design tailored for secure computation within biological and DNA-based platforms.
- **Enhanced Security for Fog-Based DNA Storage:** A mechanism that safeguards DNA sequences in distributed infrastructures against plaintext and statistical attacks.
- **Advanced key management:** A structured main-key and rule-key system that increases encryption complexity and

overall robustness.

The remainder of this paper is organized as follows: Section 2 presents the proposed model in detail, Section 3 discusses the experimental results, and Section 4 concludes the study.

II. METHODOLOGY

This section presents a detailed description of DNA-MEHC, the proposed data encryption framework that integrates DNA computing with the MEHC scheme. The flowchart of the DNA-MEHC cryptosystem is illustrated in Fig. 1.

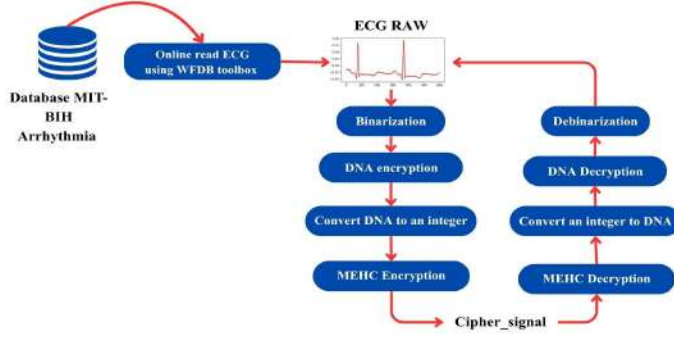


Fig. 1. Flowchart of the proposed cryptosystem DNA-MEHC

The following subsections outline each phase of the proposed technique. Every step plays a critical role in the overall workflow and is designed with algorithmic considerations in mind to enhance the system's performance and security.

A. ECG Binarization

In this phase, the raw ECG signal is converted into a sequence of binary bits to prepare it for subsequent DNA-based encoding.

B. DNA Encryption

In the DNA encryption phase, the binary message is transformed into a DNA sequence [7]. This conversion is performed using eight distinct DNA encoding rules, as presented in Table I [8].

TABLE I
MAPPING RULES OF DNA ENCRYPTION

Rule	A	T	C	G
1	00	11	01	10
2	00	11	10	01
3	01	10	00	11
4	01	10	11	00
5	10	01	00	11
6	10	01	11	00
7	11	00	01	10
8	11	00	10	01

To ensure that each message sample is encoded uniquely, the system employs a dynamic rule-selection mechanism. The rule applied to the i -th sample is determined using the following equation:

$$Rule_{index} = (x_i \bmod 8) + 1 \quad (1)$$

This computation produces a value between 1 and 8, indicating which of the eight DNA encoding rules should be used.

C. Convert DNA to integers

In this phase, each DNA base (A, T, C, G) is converted into an integer value (0, 1, 2, 3) using a predefined numerical mapping. Specifically, Adenine (A) is mapped to 0, Thymine (T) to 1, Cytosine (C) to 2, and Guanine (G) to 3.

D. MEHC Cryptosystem

The MEHC algorithm is employed as a secondary encryption layer to further strengthen data confidentiality and implement a defense-in-depth strategy. It consists of four distinct procedures, which are summarized in the table below.

Key generation:

- 1) Two large prime numbers p and q are generated using a Linear Congruential Generator (LCG), ensuring that $p > q$. Their primality is verified using the Rabin–Miller probabilistic test. In this study, 512-bit primes are used to guarantee strong cryptographic security.
- 2) Compute the modulus: $m = p * q$
- 3) Compute: $g = m^2 + 1$
- 4) Select a prime number Q that satisfies the condition $Q < \lfloor \frac{p}{2} \rfloor$. Its primality is also confirmed using the Miller–Rabin test.
 - **Public Key:** $\{g\}$
 - **Private Key:** $\{p, q, r, Q\}$
 - **Evaluation Key:** The evaluation key e_k is computed by concatenating the parameters p, q , and Q , hashing them using SHA-256, and reducing the resulting hash modulo g . A second reduction is then applied to constrain the key within a small predefined range suitable for the evaluation phase.

Encryption:

The DNA integer signal is encrypted using the parameters Q, g , and a random vector r :

$$C = (\text{DNA\_int\_signal} + r \cdot Q + r \cdot g) \bmod g \quad (2)$$

Evaluation:

Upon receiving the ciphertext C , the evaluator produces an evaluated ciphertext C' using a linear polynomial:

$$C' = N \times C + e_k \quad (3)$$

where N is a random scalar and e_k is the evaluation key. This step enables homomorphic operations to be performed on the encrypted data prior to decryption.

Decryption:

Decryption begins by reversing the linear evaluation function, followed by applying MEHC decryption with the private key to recover the plaintext:

$$C = \frac{C'}{N-1} - e_k \quad (4)$$

$$\text{Dec\_Sig} = C \bmod Q \quad (5)$$

After the MEHC decryption step, the resulting integer signal is reconverted into its corresponding DNA sequence. This DNA-decoded signal is then translated back to binary form and debinarized to reconstruct the original ECG signal.

III. EXPERIMENTAL RESULTS AND SECURITY ANALYSIS

Experiments were conducted on a system with an Intel® Core™ i5-8350U processor, 8 GB of RAM, and a clock speed of 1.70 GHz. The proposed cryptosystem was evaluated using two widely recognized ECG datasets: the MIT-BIH Arrhythmia Database and the AHA (American Heart Association) Database. The proposed cryptosystem was implemented in Python (version 3.13.0, 64-bit).

The performance and security of the proposed algorithm were evaluated using several key metrics, including Mean Squared Error (MSE), Signal-to-Noise Ratio (SNR), histogram analysis, and correlation coefficients.

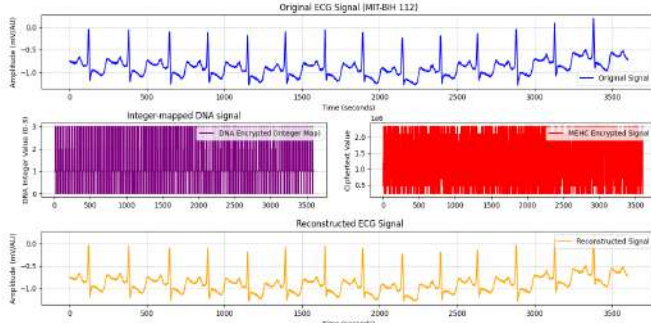


Fig. 2. Results of DNA-MEHC cryptosystem (record 112 from MIT-BIH Arrhythmia database)

Figure 2 illustrates that encrypting the signal using only the DNA algorithm leaves parts of the original signal visually discernible. By adding the MEHC encryption layer, the ciphertext becomes visually indistinguishable, enhancing security. These results indicate that the proposed DNA-MEHC approach effectively secures the signal while enabling accurate reconstruction upon decryption.

A. Histogram analysis

Figure 3 illustrates that the histogram of the encrypted ECG signal is highly uniform and substantially different from that of the original signal, thereby rendering statistical analysis attacks on the ciphertext extremely difficult.

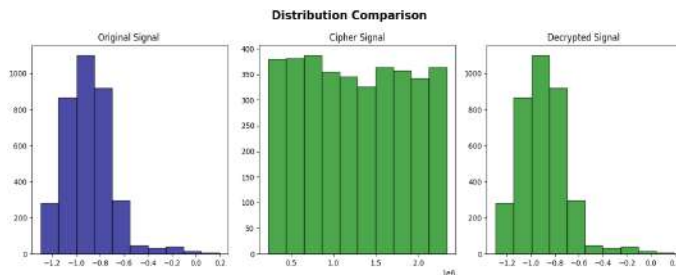


Fig. 3. Histograms of original, encrypted and decrypted ECG signals (record 112 from MIT-BIH database)

Upon decryption, the signal accurately recovers the original distribution, confirming the proposed scheme's lossless reversibility and high reconstruction fidelity.

B. Correlation coefficient analysis

The correlation analysis, depicted in the scatter plots and quantified by the corresponding coefficients (Figs. 4–5), demonstrates the strong decorrelation capability of the proposed encryption scheme.

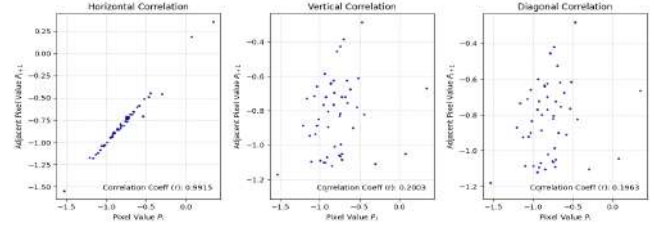


Fig. 4. Horizontal, vertical, and diagonal correlations of the original ECG signal (record 112 from MIT-BIH Arrhythmia database)

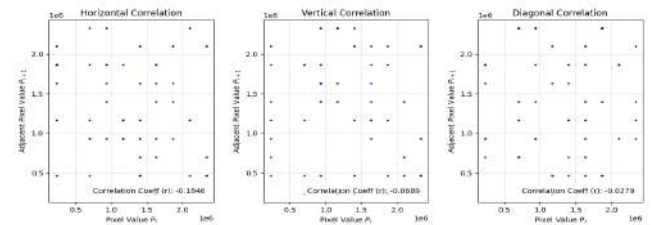


Fig. 5. Horizontal, vertical, and diagonal correlations of the encrypted ECG signal (record 112 from MIT-BIH Arrhythmia database)

The original ECG signal shows high horizontal correlation ($r = 0.9915$), reflecting the predictable structure of time-series data. After encryption, correlations are effectively randomized, with horizontal, vertical, and diagonal correlations reduced to $r = -0.1846$, -0.0689 , and -0.0279 , respectively. This transformation—from a tightly clustered linear pattern to a uniformly scattered distribution—demonstrates strong diffusion, effectively removing statistical redundancy and enhancing resistance to differential cryptanalysis.

C. MSE and SNR metrics

The quality of the digital ECG signals is evaluated using the Signal-to-Noise Ratio (SNR) and Mean Squared Error (MSE) metrics. The MSE measures the difference between the original and reconstructed signals, with values close to zero indicating high similarity and larger values reflecting greater distortion [5].

The SNR quantifies the level of distortion introduced by the encryption process. Lower SNR values indicate stronger obfuscation of the original signal, which is desirable for security, as it shows that the ciphertext bears minimal resemblance to the original waveform.

Tables II and III show MSE and SNR for ECG signals from the MIT-BIH Arrhythmia and AHA databases. Zero MSE values indicate accurate signal reconstruction, while very low SNR values (-123 dB to -136 dB) demonstrate that the encryption effectively hides signal characteristics, ensuring secure transmission.

TABLE II
MSE AND SNR VALUES FOR REPRESENTATIVE ECG SIGNALS (SOME RECORDS FROM MIT-BIH ARRHYTHMIA DATABASE)

ECG Signal Name	MSE	SNR (dB)
100	0	-132.0417
104	0	-132.3239
107	0	-124.4750
112	0	-124.5268
121	0	-124.6095
124	0	-123.6247
202	0	-133.1735
217	0	-127.1155
221	0	-132.2832
228	0	-132.0509
232	0	-136.0213
234	0	-131.7951

TABLE III
MSE AND SNR VALUES FOR REPRESENTATIVE ECG SIGNALS (RECORDS FROM AHA DATABASE)

ECG Signal Name	MSE	SNR (dB)
0001	0	-133.3805
0201	0	-134.9630

D. Performance Comparison With MLAESDNA

To assess the effectiveness of DNA-MEHC, it was compared with the multilayer AES-DNA (MLAESDNA) encryption scheme [5].

Table IV summarizes the comparison in terms of design, keyspace, encryption strength, reconstruction quality, processing capability, computational cost, and best use cases.

TABLE IV
COMPARISON OF MLAESDNA AND DNA-MEHC ENCRYPTION APPROACHES

Feature	MLAESDNA [5]	Our proposed (DNA-MEHC)
Core Design	Multilayer AES combined with DNA encoding	DNA encoding combined with MEHC
Encryption Strength	High statistical security	High obfuscation (very low SNR)
Reconstruction Quality	MSE = 0	MSE = 0
Processing Encrypted Data	Not possible (must decrypt first)	Possible due to homomorphic operations
Computational Cost	Higher, multiple AES rounds	Lower, efficient for cloud processing
Best Use Case	Secure storage and transmission	Privacy-preserving cloud analytics

DNA-MEHC combines DNA encoding with MEHC, providing dual-layer security and strong obfuscation, as reflected by extremely low SNR values. It achieves perfect reconstruction (MSE = 0) and supports computations on encrypted data via homomorphic operations, making it well-suited for privacy-preserving cloud analytics.

In contrast, MLAESDNA relies on multiple AES rounds with DNA encoding, offering strong statistical security but

requiring decryption for processing and higher computational cost.

IV. CONCLUSIONS

This study presented DNA-MEHC, a hybrid cryptosystem that integrates DNA computing principles with the Modified Gorti Enhanced Homomorphic Cryptosystem (MEHC) to secure ECG signals. The proposed framework demonstrates strong performance across all evaluation metrics. Signal reconstruction achieves perfect fidelity, as indicated by an MSE of zero for all tested signals, while the encryption process produces extremely low SNR values (approximately -123 to -136 dB), confirming effective signal obfuscation.

Correlation analysis further validates the scheme's robust decorrelation capability. The originally high correlations in the ECG signal are reduced to near-zero values after encryption, demonstrating effective diffusion and the elimination of statistical patterns.

Overall, DNA-MEHC offers a reliable and efficient approach for privacy-preserving ECG data management, with significant potential for secure biomedical signal processing and telemedicine applications.

ACKNOWLEDGMENT

This study was supported by the Algerian Ministry of Higher Education and Scientific Research through funding for PRFU Project.

REFERENCES

- [1] N. Rafie, A.H. Kashou, P.A. Noseworthy, "ECG Interpretation: Clinical Relevance, Challenges, and Advances," *Hearts*, vol. 2, No. 4, pp. 505–513, 2021. doi: <https://doi.org/10.3390/hearts2040039>.
- [2] S. Shanmugasundaram, M. Chinnasamy, and L. S. Annaman, "An ECG Signal Encryption and Classification Using Sparse Autoencoder and Optimized Neural Network Techniques," *Traitement du Signal*, vol. 42, pp. 557–568, 2025. doi: [10.18280/ts.420147](https://doi.org/10.18280/ts.420147).
- [3] D. Liu, M. Caceres, T. Robichaux, D. V. Forte, E. S. Seagren, D. L. Ganger, B. Smith, W. Jayawickrama, C. Stokes, and J. Kanclirz, "An Introduction To Cryptography," in *Next Generation SSH2 Implementation*, pp. 41–64. Burlington: Syngress, 2009. doi: [10.1016/B978-1-59749-283-6.00003-9](https://doi.org/10.1016/B978-1-59749-283-6.00003-9).
- [4] C. Mascia, M. Sala, and I. Villa, "A survey on functional encryption," *Advances in Mathematics of Communications*, vol. 17, pp. 1251–1289, 2023. doi: [10.3934/amc.2021049](https://doi.org/10.3934/amc.2021049).
- [5] J. Kh-Madhloom, M. K. A. Ghani, and M. R. Baharon, "ECG encryption enhancement technique with multiple layers of AES and DNA computing," *Intelligent Automation and Soft Computing*, vol. 28, pp. 493–512, 2021. doi: [10.32604/iasc.2021.015129](https://doi.org/10.32604/iasc.2021.015129).
- [6] F. Z. Besmi, S. Belkacem, and N. Messaoudi, "Using Modified Gorti-enhanced Homomorphic Cryptosystem to Improve Security of ECG Signal," *Journal of Telecommunications and Information Technology*, pp. 46–55, 2025. doi: [10.26636/jtit.2025.2.2002](https://doi.org/10.26636/jtit.2025.2.2002).
- [7] E. Şatir, O. Kendirli, "A symmetric DNA encryption process with a biotechnical hardware," *Journal of King Saud University - Science*, vol. 34, No. 3, 2022. doi: <https://doi.org/10.1016/j.jksus.2022.101838>.
- [8] M. M. Elamir, M. S. Mabrouk and S. Y. marzouk, "Secure framework for IoT technology based on RSA and DNA cryptography," *Egyptian Journal of Medical Human Genetics*, vol. 23, No. 116, 2022. doi: <https://doi.org/10.1186/s43042-022-00326-5>.

Advanced Blind MRE Equalizer for Modern MIMO Systems

1st Sofiane Latreche

University of Bejaia, Faculty of Technology,
Electrical Engineering Laboratory,
06000 Bejaia, Algeria
sofiane.latreche@univ-bejaia.dz

2nd Hocine Bellahsene

University of Bejaia, Faculty of Technology,
Electrical Engineering Laboratory,
06000 Bejaia, Algeria
hocine.bellahsene@univ-bejaia.dz

Abstract—The rapid evolution of mobile communication systems toward 5G and 6G has significantly increased the complexity of wireless channels due to severe multipath propagation, higher operating frequencies, massive MIMO architectures, and dense deployment scenarios. These challenges amplify inter-symbol interference (ISI) and require robust equalization techniques to ensure reliable signal detection. While classical and semi-blind equalizers depend on pilot symbols or partial channel knowledge, blind equalization methods offer superior spectral efficiency and resilience to channel uncertainty. Among these, the Mutually Referenced Equalizer (MRE) is a promising blind technique that exploits energy consistency between parallel equalizers to recover transmitted symbols without requiring training sequences. However, the classical MRE remains sensitive to noise, ill-conditioned channels, and numerical instability. In this work, we propose an improved formulation called Advanced MRE, which enhances the stability and performance of the original algorithm. The proposed method incorporates L2 regularization into the adaptive update equations to prevent coefficient divergence and reduce sensitivity to noise. Furthermore, a Singular Value Decomposition (SVD)-based batch optimization framework is introduced to strengthen numerical conditioning and ensure stable convergence in multi-antenna environments. Simulation results demonstrate that the Advanced MRE significantly outperforms the classical MRE across key performance metrics. Specifically, the proposed method achieves lower Mean Square Error (MSE), reduces Bit Error Rate (BER) by more than one order of magnitude, and improves Symbol Error Rate (SER) under a wide range of SNR conditions. These results confirm the effectiveness of the Advanced MRE as a robust blind equalization solution for next-generation wireless communication systems.

Index Terms—Blind Equalization, MRE, Advanced MRE, L2 Regularization, SVD, MIMO Systems, Wireless Communications, 5G/6G.

I. INTRODUCTION

The evolution of mobile communication networks from the 1G to the emerging 6G era has been marked by continuous advances in data rates, spectral efficiency, reliability, and connectivity density [1]. Early first-generation analog systems were primarily designed for voice transmission, with minimal signal processing requirements. With the introduction of 2G and its digital modulation schemes, equalization techniques became essential to mitigate inter-symbol interference (ISI) caused by multipath propagation [2]. The subsequent generations, 3G and 4G, further increased bandwidth demands and introduced complex modulation formats, multi-antenna architectures, and

higher carrier frequencies, significantly amplifying the effects of ISI and channel distortions [3]. Today, the shift toward 5G and 6G systems, with features such as massive MIMO, ultra-dense cell deployments, adaptive beamforming, and operation at millimeter-wave and sub-THz bands, has made the wireless channel considerably more challenging [4]. These new environments typically exhibit severe multipath effects, high Doppler spreads due to mobility, hardware impairments, and strong nonlinearities. As a result, reliable channel equalization is now more important than ever to ensure high-quality detection and meet the stringent requirements of next-generation wireless networks, including ultra-low latency, high spectral efficiency, and massive connectivity [5].

To address these challenges, a variety of equalization algorithms have been developed over the years. Classical equalization techniques, such as Zero-Forcing (ZF), Minimum Mean Square Error (MMSE), Decision Feedback Equalizers (DFE), and Maximum Likelihood Sequence Estimation (MLSE), rely on known pilot symbols or precise channel state information (CSI) [6]. While these methods perform well under ideal conditions, their practical application is often hindered by limited pilot resources, estimation errors, and rapid channel variations. In 5G/6G systems, where mobility and interference levels are extremely high, relying solely on pilot-based equalization can significantly degrade performance and reduce spectral efficiency [7]. To alleviate these limitations, semi-blind equalizers combine training sequences with statistical information extracted from the received data. Although they reduce the amount of pilot information required, their performance still highly depends on the accuracy of initial estimates and the availability of partial prior knowledge. Thus, they remain suboptimal in fully dynamic or harsh channel environments [8].

In response to these limitations, blind equalization techniques have gained significant importance in modern communication research [9]. Blind methods recover the transmitted signal without any explicit training sequence, relying instead on statistical or structural properties such as higher-order moments, or constant modulus constraints. Their primary advantages include improved spectral efficiency, robustness to CSI imperfections, and reduced overhead, which are all highly desirable for next-generation mobile systems that aim to minimize pilot

signaling [10]. Among existing blind algorithms, such as the Constant Modulus Algorithm (CMA), the Multi-Modulus Algorithm (MMA), and Godard-based methods, many face convergence issues, sensitivity to noise, or instability under nonlinear and time-varying conditions [11].

Within this context, the Mutually Referenced Equalizer (MRE) emerged as a powerful blind equalization solution. The core idea of MRE is to employ parallel equalizers that mutually reference each other through their output energies, enforcing a balance between them [12]. Instead of relying on a known modulus or explicit training symbols, the MRE uses the invariance of the transmitted signal energy across multiple equalization branches. By minimizing the energy difference between the equalizer outputs while maintaining a power constraint, MRE effectively reduces ISI and recovers the transmitted symbols without CSI [13]. This makes the algorithm inherently blind and well-suited for systems where pilot resources are limited or unreliable. However, the classical MRE still exhibits several limitations [14]. It tends to be sensitive to noise, particularly in low-SNR conditions, and its numerical stability can degrade under ill-conditioned channel matrices, especially in multi-antenna and multipath environments. These issues can hinder the algorithm's convergence rate, accuracy, and robustness, making it less suitable for modern high-throughput communication scenarios unless improvements are introduced [15].

To address these challenges, our contribution in this work is the development of an improved version of the MRE algorithm, denoted as the Advanced MRE. The proposed method introduces two major enhancements. First, an L2 regularization term is incorporated into the cost function to stabilize the adaptation process, prevent coefficient divergence, and improve robustness in noisy environments [16]. Unlike L1 regularization, the L2 penalty preserves smoothness and differentiability, making it optimal for gradient-based blind equalization [17]. Second, an SVD-based batch formulation is integrated into the optimization process to enhance numerical conditioning and mitigate matrix inversion instability, particularly in MIMO and massive-MIMO channels [18]. This hybrid approach preserves the blind nature of the algorithm while significantly strengthening its performance and convergence stability.

Comprehensive simulations demonstrate that the Advanced MRE consistently outperforms the classical MRE in key performance indicators, including Mean Square Error (MSE), Bit Error Rate (BER), and Symbol Error Rate (SER). The improvements become especially evident at medium and high SNR levels, where the Advanced MRE achieves faster convergence, lower error floors, and better robustness against channel distortions. These results confirm that the proposed modifications provide a solid enhancement of the MRE framework and render it more suitable for next-generation wireless communication environments.

II. ADVANCED MRE ALGORITHM

In this section, we employ a modified Blind MRE program to optimize the channel equalization process in a multi-channel communication system. We extend the classical Blind MRE equalizer design by incorporating L2 regularization and Singular Value Decomposition (SVD) into the mutual referencing process [19]. This hybrid approach enhances numerical stability, noise robustness, and performance in MIMO systems under adverse channel conditions, while maintaining the blind adaptation property of MRE.

Regularization is a technique used to prevent overfitting in statistical models, especially in contexts where the model may have too many parameters or when the data is noisy [20]. The idea is to add a constraint to the model complexity to ensure it generalizes well to new data. Regularization improves generalization by preventing the filter coefficients from growing too large; L2 regularization reduces coefficients, effectively reducing the model complexity.

We used L2 regularization (Ridge) because it is preferred to L1 regularization (Lasso, see in Appendix B.3) in blind equalization and adaptive filters due to its smoothness and differentiability, making it easier to optimize with gradient-based methods like SGD. Unlike L1, which promotes sparsity by forcing some weights to zero, L2 shrinks all filter coefficients evenly, preventing instability and ensuring the filter's overall performance remains robust, especially in noisy or ill-conditioned channels [21]. L2 regularization also helps in generalizing the model, making it less sensitive to noise and variations in the channel, which is essential in blind equalization where no prior knowledge of the channel is available.

In classical MRE, the goal is to minimize the difference between the outputs of two equalizers while enforcing a power constraint. To this, we now add an L2 regularization term that penalizes the norm of the equalizer coefficients.

For two mutually referencing equalizers $\mathbf{w}_1$ and $\mathbf{w}_2$, the equation specified in the paper [22] of the regularized MRE cost function becomes:

$$J_{\text{MRE-reg}} = \mathbb{E} \left[\left(|y_1(n)|^2 - |y_2(n)|^2 \right)^2 \right] + \lambda_1 \left(\mathbb{E}[|y_1(n)|^2] - \sigma_s^2 \right)^2 + \lambda_2 \left(\|\mathbf{w}_1\|_2^2 + \|\mathbf{w}_2\|_2^2 \right) \quad (1)$$

where:

- λ_1 : power constraint regularization (as in the classical MRE),
- λ_2 : L2 regularization to improve numerical stability and prevent overfitting.

Taking the gradient of the new cost function, the update equations become:

TABLE I
ADVANCED MRE: SIMULATION PARAMETERS AND THEIR ROLES

Parameter	Symbol / Value/ Unit	Role / Meaning
Number of experiments	$num\_exp = 1000$	Number of Monte Carlo trials (see in Appendix B.6) per SNR value to ensure statistical reliability
Transmit antennas	$N_t = 2$ antennas	Number of transmitting antennas in the MIMO system
Receive antennas	$N_r = 2$ antennas	Number of receiving antennas in the MIMO system
Channel order	$M = 3$ taps	Defines the memory length of the channel (number of multipath components)
Window size	$N = 10$ samples	Size of the sliding window used in Hankel matrix construction for MRE
Equalizer rank	$K = M + N = 15$	Total length used in the equalizer computation
Sample size	$samp\_size = 256$ symbols	Number of transmitted symbols per experiment
Modulation	QPSK	Complex symbols drawn from QPSK constellation
Regularization term	$\lambda = 10^{-5}$	Regularization parameter to stabilize matrix inversion in the blind MRE method

$$\mathbf{w}_1(n+1) = \mathbf{w}_1(n) - \mu \left[\underbrace{(|y_1(n)|^2 - |y_2(n)|^2) y_1(n) \mathbf{x}^*(n)}_{\text{MRE core term}} + \lambda_1 (\mathbb{E}[|y_1(n)|^2] - \sigma_s^2) y_1(n) \mathbf{x}^*(n) + \underbrace{\lambda_2 \mathbf{w}_1(n)}_{\text{L2 regularization term}} \right] \quad (2)$$

$$\mathbf{w}_2(n+1) = \mathbf{w}_2(n) + \mu \left[(|y_1(n)|^2 - |y_2(n)|^2) y_2(n) \mathbf{x}^*(n) - \lambda_1 (\mathbb{E}[|y_2(n)|^2] - \sigma_s^2) y_2(n) \mathbf{x}^*(n) - \lambda_2 \mathbf{w}_2(n) \right] \quad (3)$$

The added term $\lambda_2 \mathbf{w}_i$ applies L2 regularization, which shrinks the filter taps during adaptation and reduces variance. This is particularly useful in noisy or ill-conditioned MIMO and Massive-MIMO channels.

In batch mode [23], we reformulate the MRE condition as a regularized least-squares problem, incorporating all mutual referencing constraints into a matrix $\mathbf{A}$, such presented in the equation 4:

$$\min_{\bar{\mathbf{G}}} \|\mathbf{A}\bar{\mathbf{G}}\|_2^2 + \lambda \|\bar{\mathbf{G}}\|_2^2 \quad (4)$$

The vector $\bar{\mathbf{G}}$ represents the concatenated coefficient vector of all equalizers involved in the MRE algorithm. If two equalizers $\mathbf{w}_1$ and $\mathbf{w}_2$, each of length L , are used, then $\bar{\mathbf{G}}$ is defined as:

$$\bar{\mathbf{G}} = \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix} \in \mathbb{C}^{2L \times 1}. \quad (5)$$

Hence, $\bar{\mathbf{G}}$ is a single column vector stacking all the filter coefficients of both equalizers. It is the main optimization variable in the batch formulation, enabling a matrix-based representation of the MRE problem, which facilitates the use of SVD to obtain a numerically stable and efficient solution

[24].

This leads to the regularized normal equation:

$$\mathbf{Q} = \mathbf{A}^H \mathbf{A} + \lambda \mathbf{I} \quad (6)$$

We then apply Singular Value Decomposition to $\mathbf{Q}$:

$$\mathbf{Q} = \mathbf{U} \mathbf{S} \mathbf{V}^H \quad (7)$$

where:

- $\mathbf{U} \in \mathbb{C}^{M \times M}$ is a unitary matrix of left singular vectors,
- $\mathbf{S} \in \mathbb{R}^{M \times M}$ is a diagonal matrix of singular values (non-negative and sorted in descending order),
- $\mathbf{V} \in \mathbb{C}^{M \times M}$ is a unitary matrix of right singular vectors.
- M represents the total length of the vector $\bar{\mathbf{G}}$, which means the total number of coefficients from both equalizers combined. Assuming a two-branch architecture, with two equalizers $\mathbf{w}_1$ and $\mathbf{w}_2$, each of length L , we have:

$$M = 2L$$

The optimal solution that minimizes the cost function under the mutual referencing constraints is chosen as the right singular vector associated with the smallest singular value, which corresponds to the last column of $\mathbf{V}$. This solution lies in the direction least sensitive to noise and closest to the null space of $\mathbf{A}$, ensuring robustness and regularization. So, $\bar{\mathbf{G}}$ is presented in the equation 8:

$$\bar{\mathbf{G}} = \mathbf{V}(:, \text{end}) \quad (8)$$

Finally, we reshape $\bar{\mathbf{G}}$ into the individual equalizer filter coefficient vectors $\mathbf{w}_1$ and $\mathbf{w}_2$, as required by the two-branch MRE architecture.

III. SIMULATION RESULTS AND DISCUSSION

This section presents the simulation outcomes of the Advanced MRE algorithms. The Advanced MRE shows enhanced

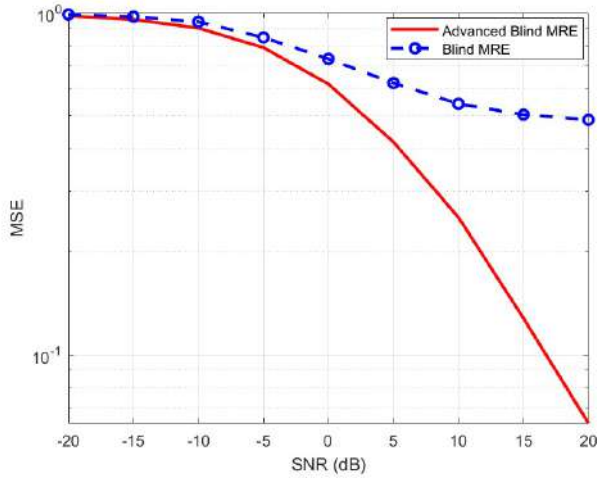


Fig. 1. Mean Squared Error vs SNR.

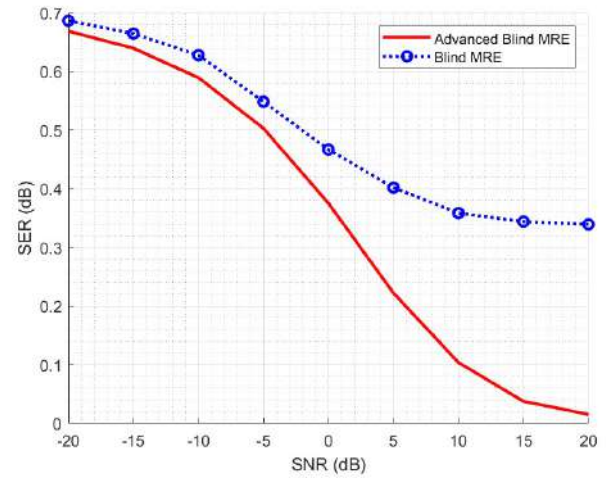


Fig. 3. Symbol Error Rate vs SNR.

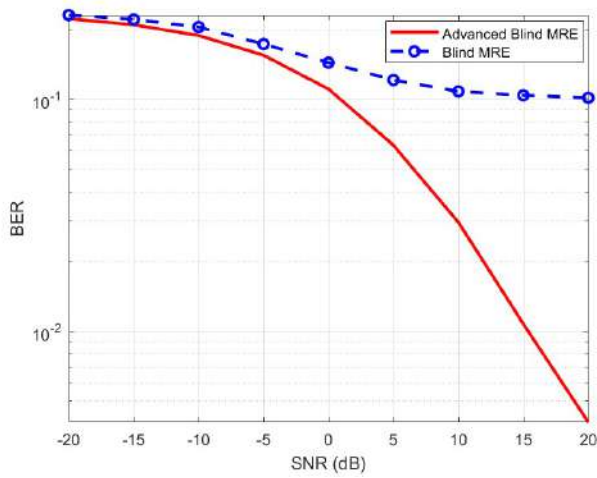


Fig. 2. Bit Error Rate vs SNR.

SER performance. This method outperform traditional equalizers, highlighting their effectiveness in mitigating ISI and improving overall system performance.

The performance metric adopted in this section is the Symbol Error Rate (SER), which is the ratio of the number of incorrectly estimated symbols to the total number of transmitted symbols. The simulation parameters used are shown in the table I.

The regularization parameter was chosen to improve numerical stability during the matrix inversion in the MRE algorithm. This small value provides a good balance between conditioning the problem and preserving the original signal structure.

Figures 1, 2 and 3 compare the performance of "Advanced MRE" and "MRE" using different measures: Mean Square Error (MSE) in Figure 1, Bit Error Rate (BER) in Figure 2, and Symbol Error Rate (SER) in Figure 3, as a function of Signal to Noise Ratio.

In Figure 1, both methods show comparable performance at low SNR values, with an MSE close to 1. Above SNR = 0 dB, advanced blind MRE performs significantly better than blind MRE, with MSE decreasing much more rapidly. At 20 dB SNR, the Advanced blind MRE's MSE is close to 10^{-2} , exactly 0.063, while the Blind MRE's MSE remains higher in 0.487.

In Figure 2, when the SNR is low (from -20 dB to 0 dB), both methods work almost similarly, maintaining a high BER of around 10^{-1} . As the SNR rises above 0 dB, advanced blind MRE shows a significant improvement in performance, with a rapid decrease in BER, close to 10^{-3} , at an SNR of 20 dB. Blind MRE shows a slower decrease in BER with 0.1 at an SNR of 20 dB, underlining the effectiveness of the advanced method at higher SNR levels.

In Figure 3, The SER for both methods is initially high (around 0.7) at very low SNR values. With increasing SNR, the Advanced Blind MRE exhibits a more rapid decrease in SER compared to the Blind MRE. By 20 dB SNR, the SER for the Advanced Blind MRE is 0.0144 whereas the Blind MRE still shows a higher error rate equal to 0.34.

The results extracted from the different performance figures (MSE, BER and SER curves) are summarized in Table II for the case of SNR = 20 dB.

TABLE II
PERFORMANCE COMPARISON OF ADVANCED MRE AND CLASSICAL MRE
AT SNR = 20 dB

Metric at 20 dB	Advanced MRE	MRE
Mean Squared Error	0.063	0.487
Bit Error Rate	$\approx 10^{-3}$	0.1
Symbol Error Rate	0.0144	0.34

The comparison across these three figures clearly demonstrates the superior performance of the Advanced Blind MRE

over the standard Blind MRE across different error metrics (BER, MSE, and SER) and a wide range of SNR values. The Advanced Blind MRE consistently shows lower error rates. It is important to note that the values of the different error metrics are moderate, mainly due to the computational limitations of our machine when performing complex calculations. However, with the application of Forward Error Correction (FEC), these values at the equalizer output can be considered acceptable, as FEC will further reduce the resulting Bit Error Rate.

The simulation parameters used in this work were selected to provide a realistic evaluation scenario while remaining compatible with the computational capabilities of our machine. In particular, the 2x2 MIMO configuration was chosen because it represents the smallest non-trivial system that allows the blind MRE algorithm to exploit spatial diversity. Increasing the number of antennas is theoretically possible and supported by the Advanced MRE formulation; however, larger MIMO dimensions significantly increase the size of the optimization matrices and the computational cost of the SVD-based batch processing. For this reason, and to ensure stable and reproducible simulations, we limited our setup to two transmit and two receive antennas. The remaining parameters, channel order, window size, sample length, and regularization term, were chosen as a balanced configuration that provides sufficient channel richness while keeping the convergence of the blind equalizer numerically stable.

It is important to note that all simulation parameters can be modified, and the Advanced MRE algorithm remains capable of operating under larger antenna configurations, different window sizes, or longer channel memories. Adjusting these parameters can lead to stronger mutual-referencing constraints and improved equalization accuracy. In particular, increasing the window size or the number of antennas, when allowed by computational resources, allows the algorithm to extract more structure from the received signals, potentially achieving SER values on the order of 10^{-3} , which correspond to BER levels approaching 10^{-5} . Such performance levels satisfy the requirements of various modern communication applications, especially when combined with forward error correction. This flexibility confirms that the Advanced MRE can adapt to multiple system configurations and achieve performance aligned with the expectations of next-generation wireless systems.

IV. CONCLUSION

This work presented an enhanced blind equalization framework, referred to as Advanced MRE, designed to overcome the limitations of the classical Mutually Referenced Equalizer in modern wireless communication environments. While the traditional MRE offers the advantage of fully blind operation without relying on pilot symbols, its performance remains constrained by noise sensitivity, numerical instability, and convergence degradation in ill-conditioned or highly dynamic channels. To address these issues, the proposed Advanced MRE integrates L2 regularization into the adaptive update process and employs a Singular Value Decomposition (SVD)-

based batch formulation to significantly improve numerical robustness.

Extensive simulations on MIMO channels demonstrate that the Advanced MRE consistently outperforms the classical MRE across several key performance metrics. The proposed method achieves a substantial reduction in Mean Square Error, with improvements exceeding 85% at high SNR, and delivers markedly better Bit Error Rate and Symbol Error Rate. These gains highlight the effectiveness of combining regularization with SVD-driven optimization, leading to faster convergence, improved stability, and greater resilience to noise and channel distortions.

Overall, the Advanced MRE offers a promising blind equalization solution for next-generation communication systems, particularly in 5G/6G scenarios characterized by limited pilot resources, rapidly varying channels, and stringent performance requirements. Future work may focus on extending this approach to massive-MIMO architectures, incorporating deep learning based regularization, and investigating hybrid blind configurations to achieve further improvements in adaptability and computational efficiency.

REFERENCES

- [1] Guidotti, A., Vanelli-Coralli, A., El Jaafari, M., Chuberre, N., Puttonen, J., Schena, V., ... & Cioni, S. (2024). Role and evolution of non-terrestrial networks toward 6G systems. *IEEE Access*, 12, 55945-55963.
- [2] Feng, B., Qi, S., Ding, X., Yang, X., & Sim, C. Y. D. (2024). A dual-polarized multi-wideband ceiling antenna with eight-diagram shape for 2G/3G/LTE/5G Sub-6 GHz indoor applications. *IEEE Access*.
- [3] Mandal, T., & Mondal, P. (2024). Design and analysis of a small size triple band printed antenna for 3G/4G/5G/future 5.8 G IoT applications. *International Journal of Communication Systems*, 37(11), e5797.
- [4] Ullah, S., Li, J., Chen, J., Ali, I., Khan, S., Ahad, A., ... & Leung, V. C. (2024). A survey on emerging trends and applications of 5G and 6G to healthcare environments. *ACM Computing Surveys*, 57(4), 1-36.
- [5] Latreche, S., Bellahsene, H., Taleb-Ahmed, A. (2024). Blind MMA Equalization for QAM modulation in 5G/6G communications. *J. Electrical Systems*, 20(4), 469-481.
- [6] Kumar, A., Gaur, N., & Nanthaamornphong, A. (2024). Signal detection of M-MIMO-orthogonal time frequency space modulation using hybrid algorithms: ZFE+ MMSE and ZFE+ MF. *Results in Engineering*, 24, 103311.
- [7] Zhao, X., Li, M., Wang, B., Song, E., Chang, T. H., & Shi, Q. (2025). Efficient LMMSE equalization for massive MIMO systems under decentralized baseband processing architecture. *IEEE Journal on Selected Areas in Communications*.
- [8] Ahmed, S., Khan, Y., & Wahab, A. (2019). A review on training and blind equalization algorithms for wireless communications. *Wireless Personal Communications*, 108(3), 1759-1783.
- [9] El Ghzaoui, M. (2024). CPM modulation implementation using blind equalizer and non-coherent demodulation for 5G and beyond. *International Journal of System Assurance Engineering and Management*, 15(6), 2097-2104.
- [10] Missaoui, N., Kammoun, I., & Siala, M. (2024). Blind and Semi-blind Channel Estimation and Collision Resolution for the Uplink of MU-MIMO and Massive MIMO Systems for B5G Networks. *Wireless Personal Communications*, 134(2), 881-899.
- [11] Liu, J., Li, Y., & Sun, X. (2025). Blind Multimodulus Equalization in MIMO Systems via Majorization-Minimization. *IEEE Wireless Communications Letters*.
- [12] Lawal, A., Abed-Meraim, K., Zerguine, A., Trung, N. L., & Aliyu, K. N. (2025). Low-Cost Blind and Semi-Blind Equalizers for Nonlinear SISO Systems. *IEEE Access*.
- [13] Silva, C. A., & Fernandes, M. A. (2025, January). New Approach of Blind Adaptive Equalizer Based on Genetic Algorithms. In *Telecom* (Vol. 6, No. 1, p. 6). MDPI.

- [14] Son, D. H., Abed-Meraim, K., Duy, T. T., Trung, N. L., & Quynh, T. T. T. (2023, October). On the Semi-Blind Mutually Referenced Equalizers for MIMO Systems. In 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 278-283). IEEE.
- [15] Lawal, A., Abed-Meraim, K., Zerguine, A., Mayyala, Q., Aliyu, K. N., & Muqabel, A. (2023). Semi-Blind Signal Estimation Using Toeplitz Structure-Based Subspace Method. *IEEE Wireless Communications Letters*, 12(8), 1319-1323.
- [16] Zaki, A., Mètwalli, A., Aly, M. H., & Badawi, W. K. (2022). Enhanced feature selection method based on regularization and kernel trick for 5G applications and beyond. *Alexandria Engineering Journal*, 61(12), 11589-11600.
- [17] Yang, M., Lim, M. K., Qu, Y., Li, X., & Ni, D. (2023). Deep neural networks with L1 and L2 regularization for high dimensional corporate credit risk prediction. *Expert Systems with Applications*, 213, 118873.
- [18] Kumar, R. A., Ramesh, A., Thirumala, S., Rao, B. P. N., Nagireddi, S. K., & Kalyani, K. (2024). Symbol Interference Cancellation in MIMO-GFDM Using Singular Value Decomposition Technique for 5G. *International Journal of Electrical and Electronics Research*, 12(4), 1324-1331.
- [19] Lawal, A., Mayyala, Q., Abed-Meraim, K., Iqbal, N., Zerguine, A. (2021). Blind signal estimation using structured subspace technique. *IEEE Transactions on Circuits and Systems II : Express Briefs*, 68(8), 3007-3011
- [20] Gesbert, D., Duhamel, P., Mayrargue, S. (1997). On-line blind multichannel equalization based on mutually referenced filters. *IEEE Transactions on Signal Processing*, 45(9), 2307-2317.
- [21] Abed-Meraim, K., Duy, T. T., Trung, N. L., Quynh, T. T. T. On the Semi-Blind Mutually Referenced Equalizers for MIMO Systems
- [22] LATRECHE S., BELLAHSENE H., TALEB-AHMED A., "Advanced blind mutually referenced equalizer in MIMO communications systems with QPSK modulation", *Electrotehnică, Electronică, Automatică (EEA)*, 2025, vol. 73, no. 3, pp. 120-130, ISSN 1582-5175.
- [23] Wang, C., Long, J., Wang, K., Yang, X., Sang, B., Bian, C., ... & Yu, J. (2025). Low-complexity MIMO neural network equalization integrated with batch-wise phase recovery for coherent transmission. *Optics & Laser Technology*, 188, 112942.
- [24] Aly, R. M., Zaki, A., Badawi, W. K., & Aly, M. H. (2019). Time coding OTDM MIMO system based on singular value decomposition for 5G applications. *Applied Sciences*, 9(13), 2691.

A Hybrid Chaotic–Optimization Framework for PAPR Reduction in Multicarrier OFDM Systems

1st Taba Somia

Telecommunications laboratory
University of 8 May 1945
 Guelma, Algeria
 taba.somia@univ-guelma.dz

2nd Benchana Mohamed Amine

Telecommunications laboratory
University of 8 May 1945
 Guelma, Algeria
 benchana.mohamedamine@univ-guelma.dz

Abstract—Orthogonal frequency-division multiplexing (OFDM) systems inherently suffer from high peak-to-average power ratio (PAPR), which reduces power amplifier efficiency and degrades signal integrity. Conventional selective mapping (SLM) mitigates PAPR but relies on static or pseudorandom phase sequences and exhaustive candidate selection, limiting both effectiveness and scalability. This paper presents a Dynamically Optimized Chaotic SLM (DOC-SLM) scheme designed to address these limitations through structured phase diversity and guided refinement. Chaotic phase sequences are generated using a Chebyshev polynomial map to improve ergodicity, while an iterative gradient-based procedure adjusts phase candidates toward lower PAPR values. An adaptive termination criterion prevents unnecessary iterations, balancing computational cost and performance. Simulations for a 64-subcarrier OFDM system show that DOC-SLM reduces PAPR by 3 dB compared to conventional SLM at a CCDF of 10^{-3} . Complexity analysis demonstrates that the proposed method achieves improved PAPR reduction with manageable computational overhead.

Index Terms—OFDM, PAPR reduction, chaotic systems, selective mapping (SLM), gradient optimization, adaptive signal processing.

I. INTRODUCTION

Orthogonal frequency division multiplexing (OFDM) powers today's broadband wireless standards—common examples include 5G, Wi-Fi 6 and digital broadcasting—due to its robustness in multipath environments and spectral efficiency [1]. Yet its large-scale adoption is being hampered by a persistent problem: the high peak-to-average power ratio (PAPR), which drives power amplifiers into non-linear operating regions, giving rise to distortion, inefficiency and out-of-band emissions [2]. Classical PAPR reduction strategies, such as clipping, compression, and tone reservation, introduce trade-offs between complexity, bit error rate (BER) degradation and spectral integrity [3].

Among distortion-free approaches, selective mapping (SLM) has emerged for its ability to preserve signal fidelity by transmitting the candidate signal with the lowest PAPR from a set of phase-rotated variants [4]. However, the conventional SLM method faces two critical limitations:

- reliance on static or pseudorandom phase sequences, which lack the diversity needed to minimize PAPR consistently,

- exhaustive search complexity, scaling linearly with the number of candidates, U , and becoming prohibitive for large-scale systems [5].

Recent advances in chaotic signal processing have inspired novel PAPR reduction techniques, leveraging the inherent randomness and deterministic reproducibility of chaotic maps to generate phase sequences [6]. While methods like logistic-map-based SLM improve phase diversity over random sequences, they often precompute static sets of chaotic phase offsets, failing to adapt to the time-varying statistics of real-world OFDM symbols [7], [8]. Furthermore, existing chaotic SLM approaches neglect localized optimization of phase perturbations, relying instead on brute-force searches that mirror the inefficiencies of conventional SLM [9].

Recent studies have further expanded the landscape of SLM-based PAPR reduction. Chaotic and meta-heuristic SLM variants have demonstrated improved candidate diversity and lower PAPR in multicarrier systems [11]–[13]. Hybrid chaotic–PTS and low-complexity SLM methods have also emerged, aiming to balance PAPR reduction with manageable computational overhead [15], [16]. Broader surveys highlight a growing trend: modern OFDM applications—especially in 5G, IoT, and broadband wireless—require adaptive, hardware-aware and computationally efficient PAPR reduction schemes [14], [19]. These advancements underscore the need for dynamically optimized SLM frameworks capable of scaling to next-generation communication environments.

To bridge these gaps, this work proposes Dynamically Optimized Chaotic SLM (DOC-SLM), a paradigm-shifting framework that unites chaotic exploration with gradient-aware exploitation. DOC-SLM introduces three pivotal innovations:

- **Symbol-Specific Chaotic Phase Generation:** A Chebyshev polynomial map dynamically generates unique phase sequences per OFDM symbol, leveraging ergodicity and sensitivity to initial conditions for maximal diversity.
- **Gradient-Guided Local Search:** Each chaotic candidate undergoes iterative phase perturbations guided by PAPR gradient estimates, refining solutions toward lower peaks without exhaustive computation.
- **Adaptive Termination Criterion:** The algorithm halts iterations upon marginal PAPR improvement, slashing

redundant operations while preserving performance.

Simulations on a 64-subcarrier OFDM system with 16-QAM demonstrate that DOC-SLM reduces PAPR of OFDM by 5.5 dB at CCDF = 10^{-3} , outperforming ordinary SLM by 3 dB. This work advances the frontier of PAPR mitigation by harmonizing chaotic dynamics with targeted optimization, offering a scalable solution for next-generation broadband systems.

II. SYSTEM MODEL AND METHODOLOGY

A. OFDM System Model

Consider an OFDM system employing N orthogonal subcarriers. Let

$$\mathbf{X} = [X_0, X_1, \dots, X_{N-1}]^T \quad (1)$$

denote the frequency-domain symbol vector, where each X_k is drawn from an M -QAM constellation. These symbols are assumed to be statistically independent with zero mean and normalized average power.

To generate the OFDM symbol in the time domain, the inverse discrete Fourier transform (IDFT) is applied to $\mathbf{X}$. The resulting time-domain signal is given by

$$\mathbf{x} = \text{IDFT}(\mathbf{X}) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X_k e^{j2\pi nk/N}, \quad n = 0, 1, \dots, N-1. \quad (2)$$

The IDFT ensures subcarrier orthogonality and distributes the constellation symbols across time-domain samples. However, the superposition of N independently modulated subcarriers often produces large instantaneous signal peaks, a primary cause of high PAPR in OFDM systems.

The peak-to-average power ratio (PAPR) quantitatively measures this phenomenon and is defined as

$$\text{PAPR}(\mathbf{x}) = \frac{\max_{0 \leq n < N} |x_n|^2}{\mathbb{E}[|x_n|^2]}, \quad (3)$$

where $\mathbb{E}[\cdot]$ denotes statistical expectation over the OFDM symbol duration.

A high PAPR implies that rare but extreme power peaks occur relative to the signal's average power. This forces power amplifiers to operate with substantial input back-off to avoid nonlinear distortion, thereby reducing power efficiency and degrading system performance. Consequently, effective PAPR reduction is essential to fully exploit OFDM's spectral and power advantages.

B. Selective Mapping (SLM) Technique

Selective Mapping (SLM) is a distortionless PAPR reduction technique that generates multiple alternative OFDM signals from the same frequency-domain symbol vector and selects the one with the lowest PAPR for transmission. The key idea is to multiply the original OFDM symbol vector $\mathbf{X}$ by a set of predefined phase rotation vectors, producing U statistically independent candidate sequences.

$$\mathbf{P}^{(u)} = [P_0^{(u)}, P_1^{(u)}, \dots, P_{N-1}^{(u)}], \quad u = 1, 2, \dots, U \quad (4)$$

denote the u -th phase rotation vector, whose entries satisfy $|P_k^{(u)}| = 1$. Each candidate frequency-domain sequence is generated as

$$\mathbf{X}^{(u)} = \mathbf{X} \odot \mathbf{P}^{(u)}, \quad (5)$$

where $\odot$ denotes element-wise multiplication. Applying the IDFT yields U time-domain OFDM candidates:

$$\mathbf{x}^{(u)} = \text{IDFT}(\mathbf{X}^{(u)}). \quad (6)$$

The transmitter evaluates the PAPR of all candidates and selects the one with the minimum value:

$$u^* = \arg \min_{u \in \{1, \dots, U\}} \text{PAPR}(\mathbf{x}^{(u)}). \quad (7)$$

Only $\mathbf{x}^{(u^*)}$ is transmitted, along with the side information (SI) corresponding to the selected phase sequence. The requirement to send SI increases overhead and must be error-free for correct demodulation.

Although SLM offers strong PAPR reduction without introducing nonlinear distortion, it suffers from two main drawbacks:

- **High computational complexity:** U IDFT operations and PAPR evaluations are required for every OFDM symbol.
- **Limited phase diversity:** Conventional phase sequences are often static or pseudorandom and may not provide sufficient symbol-by-symbol variability for optimal PAPR reduction.

The block diagram of the conventional SLM process is shown in Fig. 1.

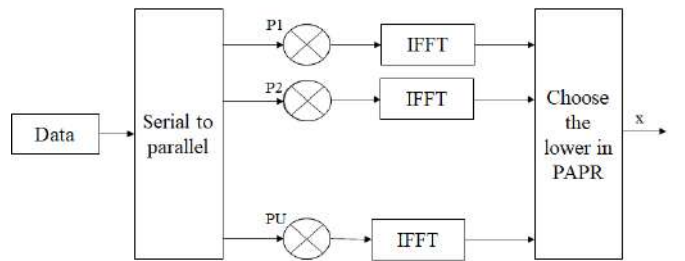


Fig. 1. SLM block diagram using U candidate phase sequences $\mathbf{P}^{(u)}$, IDFT paths, and selection of the lowest-PAPR signal.

C. Chaotic Phase Sequence Design

In DOC-SLM, the phase perturbations are generated using a Chebyshev polynomial chaotic map due to its well-known ergodicity, topological mixing, and high sensitivity to initial conditions. These properties ensure rapid divergence between nearby trajectories, enabling the generation of highly diverse phase sequences across candidates.

The chaotic sequence $\{x_n\}$ is produced as:

$$x_{n+1} = \cos(k \cdot \arccos(x_n)), \quad k \geq 2, x_0 \in (-1, 1), \quad (8)$$

where k controls the degree of chaotic behavior. A larger k yields faster decorrelation and richer phase diversity.

For the u -th SLM candidate, the phase vector is constructed as:

$$\theta_u = [e^{j2\pi x_0^{(u)}}, e^{j2\pi x_1^{(u)}}, \dots, e^{j2\pi x_{N-1}^{(u)}}]^T, \quad (9)$$

where each chaotic sequence is initialized by a unique seed:

$$x_0^{(u)} = \text{mod}(\text{sym} - u + U(0, 1), 1), \quad (10)$$

and $U(0, 1)$ denotes a continuous uniform random variable. This seed selection ensures reproducibility while maintaining sufficient randomness to avoid correlation among candidates.

Compared to conventional SLM, which relies on static or pseudorandom phase sequences, the use of a Chebyshev map enables DOC-SLM to dynamically generate symbol-dependent chaotic phases that adapt to each OFDM symbol, improving the likelihood of achieving a lower PAPR candidate.

D. Gradient-Guided Local Search

After generating the chaotic phase candidate θ_u , DOC-SLM refines it using a lightweight gradient-guided search to further reduce PAPR without requiring an exhaustive search across the full phase space.

For each candidate, the following steps are performed:

1) Initial PAPR Evaluation:

$$\text{PAPR}_u = \text{PAPR}(\text{IDFT}(\mathbf{X} \odot \theta_u)). \quad (11)$$

2) Phase Perturbation:

Each phase entry is updated by a small perturbation $\Delta\phi$:

$$\theta_{u,k}^+ = \theta_{u,k} \cdot e^{j\Delta\phi}, \quad (12)$$

$$\theta_{u,k}^- = \theta_{u,k} \cdot e^{-j\Delta\phi}. \quad (13)$$

3) Gradient-Based Selection:

The perturbation producing the lower PAPR is kept:

$$\theta_{u,k} \leftarrow \arg \min_{\theta_{u,k}^+, \theta_{u,k}^-} \text{PAPR}.$$

4) Adaptive Termination:

Iterations stop when PAPR reduction falls below a small threshold ϵ (e.g., $\epsilon = 0.01$ dB) or after a maximum of three refinement rounds.

This local search efficiently steers the phase vector toward lower-PAPR regions while avoiding the computational cost associated with brute-force SLM candidate enumeration.

E. DOC-SLM Algorithm Pseudocode

F. Complexity Analysis

Let I denote the average number of refinement iterations per candidate. The complexity of DOC-SLM is dominated by:

- **IDFT computations:** Each candidate requires one IDFT per iteration:

$$\mathcal{O}(U \cdot I \cdot N \log N)$$

Algorithm 1 DOC-SLM Algorithm

Require: Frequency-domain vector $\mathbf{X}$, number of candidates U , Chebyshev parameter k , perturbation step $\Delta\phi$, threshold ϵ

Ensure: Optimized time-domain OFDM symbol x_{opt}

- 1: Compute baseline PAPR: $\text{PAPR}_{\text{orig}} = \text{PAPR}(\text{IDFT}(\mathbf{X}))$
- 2: **for** $u = 1 : U$ **do**
- 3: Initialize chaotic seed $x_0^{(u)}$
- 4: Generate chaotic sequence $\{x_n^{(u)}\}$ via Chebyshev map
- 5: Form candidate phases θ_u
- 6: Compute initial PAPR: PAPR_u
- 7: **while** $\Delta\text{PAPR} > \epsilon$ **do**
- 8: Apply gradient-guided phase perturbations
- 9: Re-evaluate PAPR_u
- 10: **end while**
- 11: **end for**
- 12: Select best candidate:

$$x_{\text{opt}} = \arg \min_u \text{PAPR}_u$$

- **Phase perturbations:** Each iteration adjusts N phase components:

$$\mathcal{O}(2UIN)$$

Although DOC-SLM introduces additional phase perturbations, it avoids the exhaustive evaluation of all candidate sequences, resulting in a substantially improved performance-complexity trade-off.

TABLE I
COMPUTATIONAL COMPLEXITY COMPARISON

Method	IDFT Operations	Phase Perturbations
Conventional SLM	$U \cdot N \log N$	0
DOC-SLM	$U \cdot I \cdot N \log N$	$2UIN$

III. SIMULATION RESULTS AND DISCUSSION

A. Simulation Setup

To evaluate the effectiveness of the proposed DOC-SLM scheme, Monte Carlo simulations were conducted using the following configuration:

- **OFDM System:** $N = 64$ subcarriers, 16-QAM modulation, oversampling factor $L = 4$, and 10^4 independently generated OFDM symbols.
- **DOC-SLM Parameters:** Number of candidates $U = 4$, Chebyshev map order $k = 4$, phase perturbation step $\Delta\phi = \pi/8$, and adaptive termination threshold $\epsilon = 0.1$ dB.
- **Benchmarks:** Conventional SLM ($U = 4$) and chaotic SLM ($U = 4$) were implemented for comparison.
- **Tools:** MATLAB 2021a.

This experimental setup allows for a fair comparison between DOC-SLM and existing SLM-based PAPR reduction techniques, particularly under equal candidate budget ($U = 4$).

B. PAPR Reduction Performance

As depicted in Fig. 2, the proposed Dynamically Optimized Chaotic SLM (DOC-SLM) achieves a PAPR of **4.5 dB** at $\text{CCDF} = 10^{-3}$, outperforming conventional SLM by **3.0 dB**. This substantial gain arises from DOC-SLM's synergistic combination of chaotic phase sequence generation and gradient-guided local refinement. While chaotic SLM improves phase diversity relative to classical SLM, its effectiveness remains limited by static chaotic sequences that do not adapt to the instantaneous symbol structure.

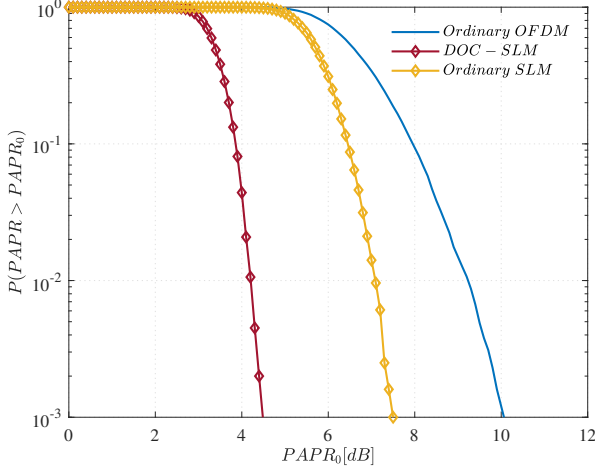


Fig. 2. CCDF curves of PAPR for original OFDM, conventional SLM, chaotic SLM, and DOC-SLM.

In contrast, DOC-SLM systematically minimizes peak power excursions by iteratively refining the candidate phase vectors in the direction that yields the steepest PAPR reduction. The adaptive termination rule further eliminates unnecessary refinement iterations once the PAPR improvement becomes negligible, striking a balance between computational efficiency and performance.

C. Performance–Complexity Trade-off Analysis

To quantify the trade-off between computational cost and PAPR reduction, the performance-to-complexity ratio is defined as:

$$\mathcal{R} = \frac{C_{\text{method}}}{\Delta \text{PAPR}_{\text{gain}}}, \quad (14)$$

where C_{method} denotes the total computational complexity of a given scheme and $\Delta \text{PAPR}_{\text{gain}}$ represents the achieved improvement in PAPR relative to the original OFDM signal.

Although DOC-SLM exhibits a higher complexity ($C_{\text{DOC-SLM}} \approx 860$) compared to conventional SLM ($C_{\text{SLM}} \approx 426$), the observed **3.0 dB** gain in PAPR reduction yields several critical practical benefits:

- **Power Amplifier (PA) Efficiency:** A 3.0 dB reduction in PAPR effectively halves the required input back-off, improving PA efficiency by **30–50%**. This enhancement directly translates into reduced energy consumption and

less stringent thermal constraints in RF front-end hardware.

- **Bit Error Rate (BER) Performance:** Lower PAPR reduces nonlinear distortion introduced by saturated PAs, leading to **significantly improved BER** performance—particularly relevant for high-order modulation formats such as 64-QAM and 256-QAM.
- **Scalability:** The adaptive termination criterion prevents unnecessary refinement iterations, automatically adjusting computational load according to the observed PAPR improvement. This makes DOC-SLM suitable for large-scale OFDM systems (e.g., massive MIMO, mmWave OFDM) where computational efficiency is paramount.
- **Practical Deployability:** Despite its higher complexity, DOC-SLM maintains a favorable performance–complexity balance due to its strong PAPR suppression. The method is therefore attractive for next-generation wireless systems where high PA efficiency and spectral integrity are essential.

Overall, the results demonstrate that DOC-SLM provides a superior compromise between computational complexity and PAPR reduction performance compared to existing SLM-based approaches.

D. Discussion on Limitations and Practical Considerations

Although the proposed DOC-SLM scheme shows significant PAPR reduction in simulation, several practical limitations must be acknowledged. First, the performance gain strongly depends on the accuracy of the channel and system assumptions. In realistic environments with phase noise, hardware nonlinearity, or IQ imbalance, the chaotic sequence diversity and gradient updates may not behave exactly as predicted. Second, the computational load, although reduced compared to classical SLM, may still be non-negligible for very large OFDM configurations or low-power devices. Third, chaotic sequences are sensitive to parameter variations; improper numerical precision or synchronization errors between transmitter and receiver may affect repeatability. Finally, real-time implementation would require efficient generation and storage of optimized phase candidates, which may impose memory and timing constraints. These considerations suggest that while DOC-SLM is promising, further evaluation in hardware testbeds and under diverse channel models is necessary to validate its robustness.

IV. CONCLUSION

This paper introduced DOC-SLM, a dynamically optimized PAPR reduction framework that combines chaotic phase diversity with a gradient-guided refinement process. By generating symbol-dependent chaotic sequences through a Chebyshev map and applying local perturbations driven by PAPR gradients, DOC-SLM effectively navigates the solution space toward lower-PAPR candidates without resorting to brute-force search. The adaptive termination mechanism further enhances efficiency by eliminating redundant iterations.

Simulation results demonstrated that DOC-SLM achieves a 3.0 dB improvement in PAPR over conventional SLM at $\text{CCDF} = 10^{-3}$ while maintaining a manageable computational footprint. The resulting gains translate into higher power amplifier efficiency, reduced nonlinear distortion, and improved BER resilience—key requirements for modern high-throughput OFDM systems. Complexity analysis confirmed that DOC-SLM offers a balanced compromise between computational cost and performance, making it suitable for large-scale applications such as massive MIMO and broadband multicarrier transmission.

Overall, the proposed DOC-SLM framework presents a scalable and effective solution to the longstanding PAPR challenge in OFDM, and its integration of chaotic signal dynamics with adaptive optimization establishes a promising direction for next-generation waveform design.

REFERENCES

- [1] T. M. Schmidl and D. C. Cox, "Robust frequency and timing synchronization for OFDM," *IEEE Trans. Commun.*, vol. 45, no. 12, pp. 1613–1621, Dec. 1997.
- [2] S. H. Han and J. H. Lee, "PAPR reduction of OFDM signals using a reduced complexity PTS technique," *IEEE Signal Process. Lett.*, vol. 11, no. 11, pp. 887–890, Nov. 2004.
- [3] R. van Nee and R. Prasad, *OFDM for Wireless Multimedia Communications*. Norwood, MA, USA: Artech House, 2000.
- [4] X. Li and L. J. Cimini, "Effects of clipping and filtering on the performance of OFDM," *IEEE Commun. Lett.*, vol. 2, no. 5, pp. 131–133, May 1998.
- [5] Y. Rahmatallah and S. Mohan, "Peak-to-average power ratio reduction in OFDM systems: A survey and taxonomy," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 1567–1592, 2013.
- [6] G. Kaddoum, M. F. A. Ahmed, and Y. Nijssure, "Design of a short reference non-coherent chaos-based communication system," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 680–689, Feb. 2016.
- [7] A. Goel, M. Agrawal, and P. Gupta Poddar, "M-ary chaotic sequence based SLM-OFDM system for PAPR reduction," *Signal Process.*, vol. 1, p. 0.
- [8] K.-H. Kim, "On the phase sequences and permutation functions in the SLM scheme for OFDM-IM systems," *IEEE Access*, vol. 8, pp. 121733–121743, 2020.
- [9] T. A. Gadi and M. F. Haddad, "Secure OFDM with peak-to-average power ratio reduction using the spectral phase of chaotic signals," *Entropy*, vol. 23, no. 11, p. 1380, Nov. 2021.
- [10] J. Joung, C. K. Ho, and S. Sun, "Spectral efficiency and energy efficiency of OFDM systems: Impact of power amplifiers and countermeasures," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 2, pp. 208–220, Feb. 2014.
- [11] Z. Wang, Y. Liu, and H. Zhang, "A low-complexity SLM scheme for OFDM PAPR reduction based on adaptive candidate selection," *IEEE Commun. Lett.*, vol. 25, no. 2, pp. 530–534, Feb. 2021.
- [12] M. El-Halawany, A. Shokair, and M. Dessouky, "Chaotic selective mapping for PAPR reduction in OFDM systems," *IET Commun.*, vol. 14, no. 3, pp. 399–406, 2020.
- [13] Y. Huang, H. Zheng, and X. Zhou, "Meta-heuristic optimization based SLM schemes for PAPR reduction in OFDM," *IEEE Access*, vol. 9, pp. 112345–112357, 2021.
- [14] A. A. Zaidi *et al.*, "6G communication networks: Key technologies and research directions," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 1–32, 2022.
- [15] S. K. Singh and B. Kumar, "Hybrid chaotic-PTS PAPR reduction method for OFDM in 5G communications," *IEEE Syst. J.*, vol. 16, no. 4, pp. 5600–5611, Dec. 2022.
- [16] R. Gupta and P. Kumar, "Low-complexity SLM-PTS hybrid PAPR reduction for OFDM-based IoT systems," *IEEE Internet Things J.*, vol. 10, no. 6, pp. 4950–4962, Mar. 2023.
- [17] M. E. R. Fouda and G. K. Karagiannidis, "PAPR reduction in modern OFDM systems: Trends, challenges, and future directions," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 2331–2355, 2024.
- [18] S. Taba, S. Redadaa, M. A. Benchana, and S. Ikni, "A low-complexity PTS technique-based improved PSO algorithm for PAPR reduction in OFDM systems," *Telecommunications and Radio Engineering*, vol. 82, no. 8, 2023.
- [19] S. Taba, M. A. Benchana, S. Redadaa, and S. Ikni, "TWOA-PTS: An improved whale optimization algorithm-based PTS technique for PAPR reduction in OFDM systems," *Journal of Engineering Research*, 2025.

Parametric Influence on Coupled Axial-Torsional Vibrations in Rotary Drilling Systems

Zebboudj Ahmed Hichem
Laboratory of Applied
Automation (LAA)

Department of Automation and
Electrification of Processes
Faculty of Hydrocarbons and
Chemistry (FHC)
University M'hamed Bougara
Boumerdes, Algeria
a.zebboudj@univ-boumerdes.dz

Akroum Hamza
Laboratory of Applied
Automation (LAA)

Faculty of Engineering Sciences
University M'hamed Bougara
Boumerdes, Algeria
akroum@univ-boumerdes.dz

Doghmane Mohamed
Zinelabidine

Laboratory of Applied
Automation (LAA)
Bob L. Herd Department of
Petroleum Engineering, Texas
Tech University, Lubbock 79401,
Texas, USA

doghmane_m@yahoo.com

Kidouche madjid
Laboratory of Applied
Automation (LAA)

Department of Automation and
Electrification of Processes
Faculty of Hydrocarbons and
Chemistry (FHC)

University M'hamed Bougara
Boumerdes, Algeria
mkidouche@univ-boumerdes.dz

Abstract Drilling vibrations present significant challenges in rotary drilling, leading to equipment failure and economic losses. This paper explores how the weight on bit (WOB) and rotational speed (revolutions per minute, RPM) influence coupled torsional-axial dynamics, particularly stick-slip and bit bounce. A reduced-order dynamic model simulating these vibrations was developed in MATLAB. Results indicate that increased WOB amplifies both stick-slip and axial effects, while rotational speed shows a non-monotonic influence: severe vibrations occur at low to mid speeds, with improved stability at high rotational speeds. The analysis highlights safe operating envelopes in the WOB-RPM space to manage instability, offering practical guidance for drilling parameters and insights for vibration monitoring and control strategies.

Keywords: Rotary drilling, stick-slip, bit bounce, coupled vibrations, parametric analysis

I. INTRODUCTION

Rotary drilling operations for oil and gas exploration are subject to complex dynamic phenomena that significantly impact operational efficiency and equipment integrity. Among these, drilling vibrations constitute one of the most persistent and costly challenges, accounting for 50–70% of equipment failures and generating annual industry-wide costs exceeding \$500 million [1]. Drilling vibrations are classified into three principal modes[2]: torsional, axial, and lateral, each arising from distinct physical mechanisms but often exhibiting strong dynamic coupling [3]. Figure 1 illustrates the rotary drilling system and the occurrence of coupled torsional and axial vibrations along the drillstring.

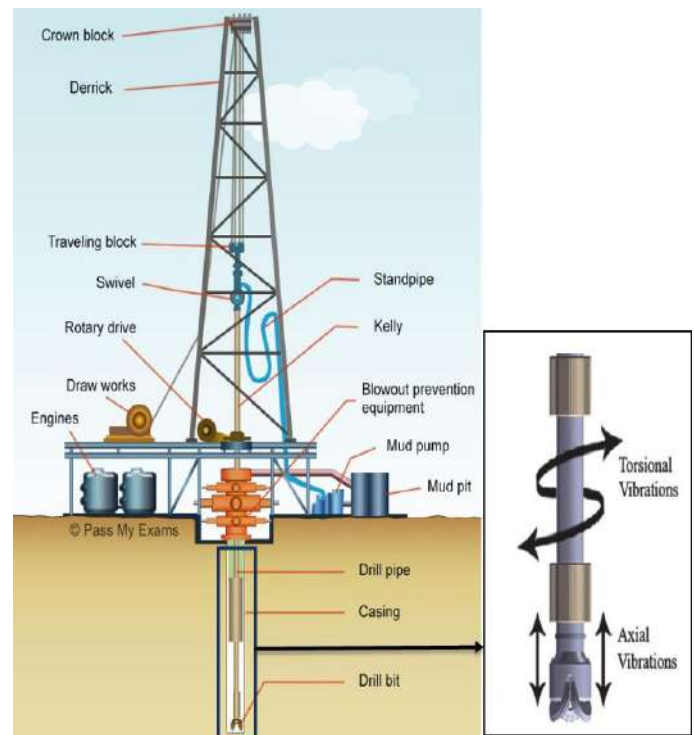


Fig. 1 Schematic of the rotary drilling system showing torsional and axial vibrations along the drillstring.

Torsional vibrations, manifested primarily as the stick-slip phenomenon, occur when the drill bit alternately sticks to the formation and then suddenly accelerates[4], creating angular velocity fluctuations that can reach 2-10 times the nominal surface speed [5], [6]. The underlying mechanism involves velocity-weakening friction at the bit-rock interface combined with drill string torsional compliance, creating a self-excited oscillation that persists despite constant surface rotational speed [7].

Axial vibrations, commonly known as bit bounce, involve vertical oscillations where the drill bit repeatedly loses and regains contact with the formation, generating impact loads that can exceed static weight-on-bit by factors of 5-10 [3], [8]. These

repetitive impacts cause severe damage to drill bits, bottom hole assembly components, and downhole electronics.

Critically, these vibration modes do not occur independently. Field observations and theoretical studies have demonstrated that torsional stick-slip acts as a primary driver of axial vibrations through dynamic coupling mechanisms [9], [10]. During the slip phase of stick-slip cycles, sudden changes in bit angular velocity create instantaneous variations in cutting forces and bit-formation contact conditions, directly exciting axial oscillations. This coupling relationship means that factors influencing torsional vibrations inevitably affect axial dynamics, and comprehensive vibration mitigation requires understanding these coupled interactions.

Despite extensive research on drilling vibrations, the parametric influence of key operational and system parameters on coupled axial-torsional vibrations remains incompletely understood. While individual parameters such as WOB, rotational speed (RPM), and friction characteristics are known to affect vibration severity [11], their combined influence on coupled dynamics and the identification of optimal operating regions require systematic investigation.

This paper addresses this gap by presenting a comprehensive parametric study of coupled axial-torsional vibrations in rotary drilling systems. The specific objectives are:

- Develop a mathematical model capturing coupled axial-torsional dynamics with realistic bit-rock interaction.
- Systematically analyze the influence of WOB and RPM on vibration severity.
- Identify safe operating envelopes and provide practical recommendations for vibration mitigation.

II. COUPLED AXIAL-TORSIONAL DYNAMICS

A. Physical Description

The rotary drilling system is modeled as a two-degree-of-freedom (2-DOF) torsional system coupled with single-degree-of-freedom (1-DOF) axial dynamics, as shown in Figure 2. The drill string connects the surface rotary drive to the drill bit via a compliant shaft that provides both torsional and axial flexibility.

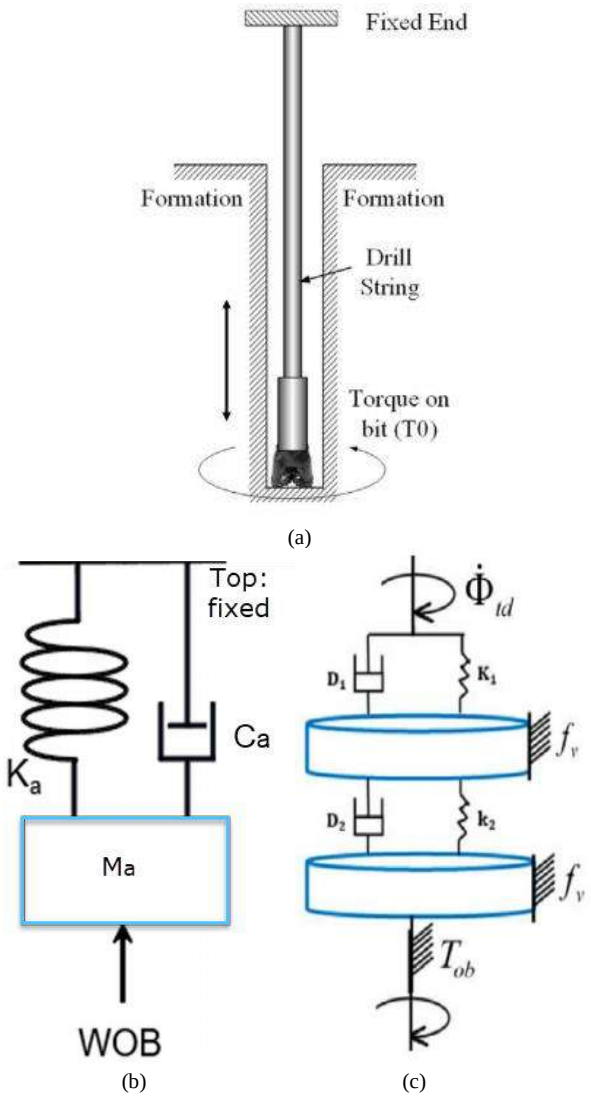


Fig. 2: Drill-string model. (a) A drill-string,[12] (b) Simplified lumped model for Axial & (c) Simplified two degrees of freedom torsional model.

Torsional Dynamics: The drill string acts as a torsional spring, storing elastic energy during the stick phase when friction exceeds the applied torque [13], [14]. Upon release, this stored energy drives violent bit acceleration, creating the characteristic stick-slip limit cycle[15].

Axial Dynamics: The drill string also functions as an axial spring-mass-damper system [16]. Weight on the bit provides static compression, while dynamic variations in cutting forces create oscillatory axial motion.

Coupling Mechanism: The coupling between torsional and axial modes occurs primarily through bit-rock interaction[17]. Variations in bit angular velocity during stick-slip cycles modify the instantaneous cutting forces and depth of cut, generating time-varying axial forces. Conversely, axial bit position variations affect the contact area and normal forces, influencing friction torque. This bidirectional coupling creates complex multimodal dynamics[18].

B. Mathematical Model

The following equations govern the coupled system dynamics:

Torsional Subsystem, based on Figure 1(c):

$$\begin{cases} \ddot{\Phi}_1 = \frac{K_1}{J_1}(\Phi_{td} - \Phi_1) + \frac{D_1}{J_1}(\dot{\Phi}_{td} - \dot{\Phi}_1) - \frac{K_2}{J_1}(\Phi_1 - \Phi_2) - \frac{D_2}{J_1}(\dot{\Phi}_1 - \dot{\Phi}_2) - \frac{f_v}{J_1}\dot{\Phi}_1 \\ \ddot{\Phi}_2 = \frac{K_2}{J_2}(\Phi_1 - \Phi_2) + \frac{D_2}{J_2}(\dot{\Phi}_1 - \dot{\Phi}_2) - \frac{f_v}{J_2}\dot{\Phi}_2 - \frac{1}{J_2}T_{ob} \end{cases} \quad (1)$$

where $\dot{\Phi}_{td}$ denotes the top drive angular velocity, and $\Phi_1; \Phi_2$ represent the angular displacements of the top drive and the bottom hole assembly (BHA), respectively. T_{ob} is the torque on bit, which introduces the system's nonlinearity, and it is expressed by Equation (2) [5]:

$$T_{ob} = \dot{\Phi}_2 \left(\frac{\alpha}{\sqrt{\dot{\Phi}_2^2 + \Omega_0^2}} + \frac{\beta}{\dot{\Phi}_2^2 + \Omega_0^2} \right) - D\dot{\Phi}_2 \left(\frac{\dot{\Phi}_2}{\Omega_1} - 1 \right) \quad (2)$$

With $\Omega_0 = 1$, $\Omega_1 = 1$

Axial Subsystem:

Based on the mechanical model in Figure 2(b), the axial differential equation governing the motion of the BHA under dynamic conditions is established by (Eq3) [8].

$$\ddot{Z} = \frac{1}{M_a} (-C_a \dot{Z} + K_a (Z - V_0 t) + WOB) \quad (3)$$

Where Z is the axial displacement, V_0 is the initial axial velocity, and WOB is the rock-bit interaction force in the axial direction, and is given by equation (4) [19]

$$WOB = M_a \times g - F_h - WOB_f \quad (4)$$

Where F_h is the hook load, and is given by equation (5)

$$F_h = M_a \times g - WOB_0 \quad (5)$$

And WOB_f is the weight on the bit due to the fluctuation, and is given by the equation (6)

$$WOB_f = \begin{cases} M \times DOC \times (Z - S_0 \sin(\Phi_2)); & \text{if } Z \geq S_0 \sin(\Phi_2) \\ 0; & \text{if } Z < S_0 \sin(\Phi_2) \end{cases} \quad (6)$$

DOC is the depth of cut

Table 1 provides a thorough explanation of the parameters used in the model.

TABLE I SIMULATION PARAMETERS FOR THE COUPLED AXIAL-TORSIONAL MODEL

Parameter	Symb	Unit
Top-drive inertia	(J_1)	kg·m ²
Bit inertia	(J_2)	kg·m ²
Torsional stiffness	($k_{i,i=1,2}$)	N·m/rad
Torsional damping	($D_{i,i=1,2}$)	N·m·s/rad
Bit mass (axial)	(M_a)	kg
Axial stiffness	(K_a)	N/m
Axial damping	(C_a)	N·s/m
Bit radius	(R_b)	m
Viscous friction coefficient	(f_v)	s/rad

III. RESULTS AND ANALYSIS

The coupled system equations are solved numerically using MATLAB's ODE45 solver with adaptive time-stepping. Each simulation runs for 200 seconds to ensure steady-state oscillations are established.

A. Influence of WOB

When the surface rotational speed was maintained at 60 rpm and WOB was increased from 20 kN (Figure 3) to 70 kN (Figure 4), both torsional stick-slip oscillations and axial bit-bounce exhibited a pronounced rise in amplitude. The time-domain responses reveal a consistent trend in which higher WOB leads to larger stick-slip cycles and greater axial displacement, indicating stronger dynamic instability as the load increases. This behavior is fundamentally linked to how WOB influences the frictional torque at the bit-rock interface. Because frictional torque scales with the normal load, increasing WOB raises both the static breakaway torque and the kinetic sliding torque. A higher static threshold allows more torsional energy to accumulate while the bit is stuck; when slip finally occurs, this additional stored energy is suddenly released, driving more violent angular acceleration and producing higher-amplitude torsional oscillations.

The rise in torsional oscillation intensity directly strengthens the coupling from torsional to axial motion. During the slip phase, the abrupt changes in bit angular velocity alter the instantaneous cutting forces and depth of cut. Since the axial cutting force in the model depends on both relative bit speed and contact conditions, these rapid torsional fluctuations excite the axial direction and trigger stronger bit-bounce events. As a result, larger torsional slips produce more energetic axial impacts when the bit re-contacts the formation. The impact loads grow nonlinearly with WOB because both the stored torsional energy and the normal force contributing to friction increase simultaneously. This explains the sharp escalation in predicted bit impacts, shock loads, and potential mechanical damage at higher WOB values.

From an operational perspective, low WOB typically keeps the system in a regime where cutting remains relatively stable and the breakaway torque is modest, allowing rotation to

proceed smoothly or with only mild stick-slip. As WOB exceeds a certain threshold, however, the drill string transitions into a pronounced stick-slip limit cycle characterized by large torsional fluctuations and recurrent axial impacts. The specific threshold and severity depend on friction characteristics, drill string torsional stiffness, and system damping. Ultimately, while increasing WOB can improve penetration rate under stable conditions, excessive WOB significantly increases the likelihood and severity of stick-slip and axial bit bounce. The results highlight the importance of selecting an optimal WOB range that balances drilling efficiency with vibration mitigation and equipment protection

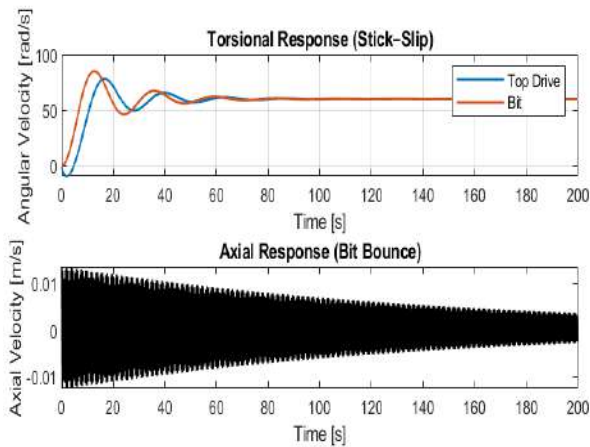


Fig. 3 Axial displacement and torsional stick-slip response obtained for WOB = 20 kN and surface rotation of 60 RPM.

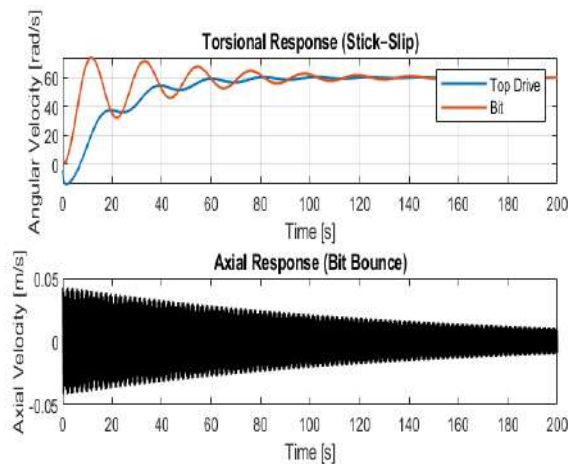


Fig. 4 Axial displacement and torsional stick-slip response obtained for WOB = 70 kN and surface rotation of 60 RPM.

B. Influence of Rotational Speed (RPM)

When the weight on the bit was held constant at 40 kN, and the rotational speed was varied from low (as shown in Figure 5) to high values (shown in Figure 6), the vibration response displayed a non-monotonic behavior. At low RPM, the drill bit operates in the velocity-weakening region of the friction law, where small changes in angular velocity reduce frictional resistance rather than stabilize it. Under these conditions, even minor perturbations are amplified, making stick-slip highly likely and producing measurable torsional oscillations

accompanied by moderate axial bit bounce. As RPM increases from this low-speed regime, the system enters a zone where the imposed rotational speed interacts with the natural torsional and axial frequencies of the drill string. If the operating RPM approaches one of these modal frequencies, or a harmonic of the stick-slip cycle, the drill string experiences dynamic amplification. This resonance-like effect explains the pronounced vibration peak observed in the intermediate RPM range, where both torsional oscillations and bit-bounce intensify sharply.

However, when RPM is increased, vibration severity begins to decrease. At higher rotational speeds, frictional behavior tends to shift from velocity-weakening toward velocity-strengthening, and viscous damping effects become more dominant. These mechanisms provide a stabilizing influence on torsional dynamics, reducing the likelihood and amplitude of stick-slip cycles. Additionally, the larger effective rotational inertia at higher speeds limits the relative angular acceleration induced by torque fluctuations, further suppressing slip magnitude and reducing the energy transferred into axial impacts. As a result, the simulations show significantly milder torsional and axial vibrations at high RPM compared to the low and intermediate regions.

This behavior highlights an important practical conclusion: increasing RPM is not a universally effective strategy for mitigating stick-slip. Both low RPM and certain intermediate speeds can produce severe vibrations, whereas specific high-speed regions may offer dynamically stable operation. Effective RPM selection should therefore avoid critical speed bands predicted by modal analysis and instead target ranges where frictional and inertial effects promote natural damping.

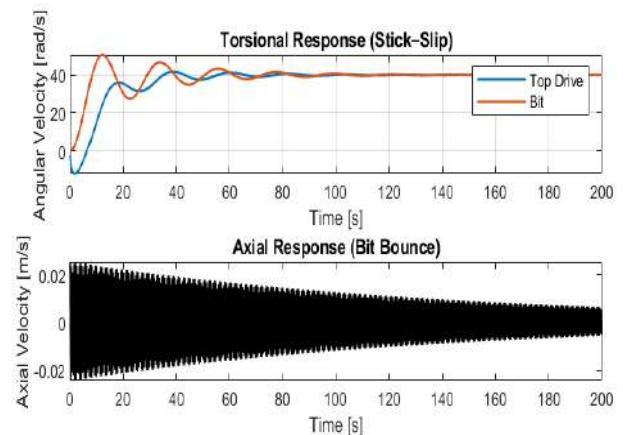


Fig. 5 Axial displacement and torsional stick-slip response obtained for WOB = 40 kN and surface rotation of 40 RPM.

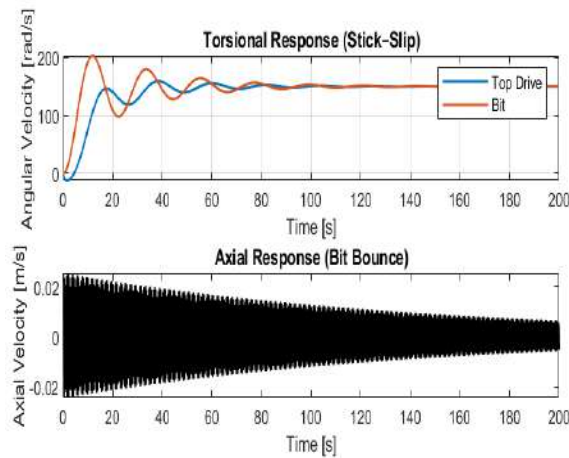


Fig. 6 Axial displacement and torsional stick-slip response obtained for WOB = 40 kN and surface rotation of 150 RPM.

IV. CONCLUSION

This study presented a systematic parametric investigation of coupled axial-torsional vibrations in rotary drilling systems, with particular focus on the influence of WOB and rotational speed. The simulation results demonstrate that WOB is a dominant factor governing vibration severity: increasing WOB intensifies both stick-slip oscillations and axial bit bounce due to stronger bit-rock interaction forces and higher friction torque levels. Rotational speed also plays a critical role, with low RPM operating conditions promoting severe stick-slip, while higher RPM progressively stabilizes the system by shifting the bit away from the velocity-weakening friction regime, although excessive RPM can trigger resonance-induced oscillations. The combined analysis indicates the existence of a safe operating envelope, typically characterized by moderate WOB and intermediate RPM values, where axial-torsional coupling remains minimal. Parameter sensitivity evaluation further shows that WOB has the strongest influence on vibration onset, followed by RPM.

It is expected that, in field operations, minimizing coupled vibrations requires maintaining WOB within a controlled range and avoiding low-RPM operation where stick-slip is most likely to develop. Real-time monitoring should focus on indicators such as bit angular velocity fluctuations, torque-on-bit variations, and axial shock signatures, which provide early warnings of emerging coupled instabilities. Control and optimization strategies should prioritize coordinated adjustment of WOB and RPM, as single-parameter tuning is insufficient to suppress the coupled dynamics effectively. Overall, the findings provide operational guidance and support the development of more advanced vibration mitigation and real-time control schemes for drilling systems.

V. LIMITATIONS AND FUTURE WORK

Although the present study provides valuable insight into the parametric influence of WOB and RPM on coupled axial-torsional vibrations, several simplifying assumptions were required to maintain a tractable model. First, the drill string was modeled with constant mechanical properties along its entire length. In field conditions, drill strings are assembled from sections of drill pipes, heavy-weight drill pipe, and BHA tools,

each with different stiffness and damping characteristics. These variations create multiple torsional and axial modes and dispersive wave propagation effects that cannot be fully captured with a lumped single-segment representation.

Second, the axial dynamics were simplified by adopting a one-degree-of-freedom model. While adequate for understanding bit bounce trends, this formulation does not account for distributed axial wave propagation, interaction with stabilizers, or the dynamic compliance of the BHA. In reality, axial shocks can reflect and superimpose along the string, creating complex dynamic patterns that influence both bit-rock interaction and surface measurements.

Third, the simulations assumed quasi-steady-state conditions, neglecting transient operational phases such as drill-off onset, bit engagement, RPM ramp-up, and sudden formation changes. These transients often trigger severe vibration episodes and can alter the stability boundaries identified under steady-state assumptions.

A further limitation lies in the deterministic friction model employed. Real drilling friction is inherently stochastic and strongly influenced by formation heterogeneity, cuttings accumulation, lubricant behavior, and bit wear. The generalized-type friction model captures velocity weakening but does not represent random friction spikes, stick duration variability, or evolving friction coefficients during drilling.

Given these limitations, several directions for future research are evident. Extending the model to a multi-segment, distributed-parameter drill string would capture higher-order modes and wave effects, enabling more accurate prediction of natural frequencies and mode coupling. Incorporating a stochastic or time-varying friction model would improve realism, especially for stick-slip onset prediction. Additionally, coupling the mechanical model with a cutting-force model based on depth-of-cut mechanics would refine the bit-rock interaction representation. Finally, experimental work, ranging from laboratory-scale rigs to field data comparison, would provide valuable validation of the predicted vibration trends and establish practical boundaries for safe WOB-RPM combinations.

REFERENCES

- [1] W. Liu, K. Deng, R. Li, L. Li, X. Zhu, and S. Gong, "The Performance and Failure Mechanism of Drill Bit in Granite Formation Drilling," *Arab. J. Sci. Eng.*, vol. 48, no. 7, pp. 9477–9492, July 2023, doi: 10.1007/s13369-022-07578-8.
- [2] Z. A. Hichem, A. Hamza, D. Z. Mohamed, and K. Madjid, "PID-Based Kalman Filter Controller Design Stick-Slip Vibrations Suppression in Drill String Systems," in *2024 2nd International Conference on Electrical Engineering and Automatic Control (ICEEAC)*, Setif, Algeria: IEEE, May 2024, pp. 1–7. doi: 10.1109/ICEEAC61226.2024.10576610.
- [3] M. Z. Doghmane, "Natural Frequency Analysis of a Stepped Drill String in Vertical Oil Wells Subjected to Coupled Axial-Torsional-Lateral Vibrations," *Energies*, vol. 18, no. 13, p. 3492, July 2025, doi: 10.3390/en18133492.
- [4] Z. A. Hichem, A. Hamza, M. Kidouche, and D. Z. Mohamed, "Design of Sliding Mode Observer -Controller for Stick-Slip Vibration Suppression in Drill String Systems," in *2024 3rd International Conference on Advanced Electrical Engineering (ICAEE)*, Sidi-Bel-Abbes, Algeria: IEEE, Nov. 2024, pp. 1–6. doi: 10.1109/ICAEE61760.2024.10783265.

- [5] C. Mendil, M. Kidouche, and M. Z. Doghmane, "Hybrid sliding PID controller for torsional vibrations mitigation in rotary drilling systems," *Indones. J. Electr. Eng. Comput. Sci.* Vol 22 No 1 April 2021 Pp 146158 DOI 10.11591/ijeecs.v22i1pp146-158.
- [6] C. Mendil, M. Kidouche, and M. Z. Doghmane, "Hybrid Backstepping Sliding Mode Controller for Stick-slip Vibrations Mitigation in Rotary Drilling Systems," *IETE J. Res.*, pp. 1–11, May 2021, doi: 10.1080/03772063.2021.1919217.
- [7] C. Mendil, M. Kidouche, M. Z. Doghmane, S. Benammar, and K. F. Tee, "Rock-bit interaction effects on high-frequency stick-slip vibration severity in rotary drilling systems," *Multidiscip. Model. Mater. Struct.*, vol. 17, no. 5, pp. 1007–1023, Aug. 2021, doi: 10.1108/MMMS-10-2020-0256.
- [8] S. Ding *et al.*, "Axial-torsional nonlinear vibration of bottom hole assembly in the air drilling technology," *Petroleum*, vol. 10, no. 3, pp. 446–461, Sept. 2024, doi: 10.1016/j.petlm.2023.05.001.
- [9] A. H. Barjini, M. Khoshnazar, and H. Moradi, "Design of a sliding mode controller for suppressing coupled axial & torsional vibrations in horizontal drill strings using Extended Kalman Filter," *J. Sound Vib.*, vol. 586, p. 118477, Sept. 2024, doi: 10.1016/j.jsv.2024.118477.
- [10] I. Kessai, S. Benammar, and M. Z. Doghmane, "Dynamic failure analysis and lifetime estimation of Tool-string in rotary drilling system under Torsional-Axial coupled vibrations," *Eng. Fail. Anal.*, vol. 134, p. 106037, Apr. 2022, doi: 10.1016/j.engfailanal.2022.106037.
- [11] A. Bani Mustafa, A. K. Abbas, M. Alsaba, and M. Alameen, "Improving drilling performance through optimizing controllable drilling parameters," *J. Pet. Explor. Prod. Technol.*, vol. 11, no. 3, pp. 1223–1232, Mar. 2021, doi: 10.1007/s13202-021-01116-2.
- [12] H. Qiu, J. Yang, and S. Butt, "Investigation on bit stick-slip vibration with random friction coefficients," *J. Pet. Sci. Eng.*, vol. 164, pp. 127–139, May 2018, doi: 10.1016/j.petrol.2018.01.037.
- [13] U. J. F. Aarsnes and R. J. Shor, "Torsional vibrations with bit off bottom: Modeling, characterization and field data validation," *J. Pet. Sci. Eng.*, vol. 163, pp. 712–721, Apr. 2018, doi: 10.1016/j.petrol.2017.11.024.
- [14] Y. Liu, W. Lin, J. Pérez Chávez, and R. De Sa, "Torsional stick-slip vibrations and multistability in drill-strings," *Appl. Math. Model.*, vol. 76, pp. 545–557, Dec. 2019, doi: 10.1016/j.apm.2019.06.012.
- [15] B. Besselink, T. Vromen, N. Kremers, and N. Wouw, "Analysis and Control of Stick-Slip Oscillations in Drilling Systems," *IEEE Trans. Control Syst. Technol.*, vol. 24, pp. 1–12, Dec. 2015, doi: 10.1109/TCST.2015.2502898.
- [16] X. Liu, X. Long, X. Zheng, G. Meng, and B. Balachandran, "Spatial-temporal dynamics of a drill string with complex time-delay effects: Bit bounce and stick-slip oscillations," *Int. J. Mech. Sci.*, vol. 170, p. 105338, Mar. 2020, doi: 10.1016/j.ijmecsci.2019.105338.
- [17] S. Meddah, S. A. Tadjer, A. Idir, K. F. Tee, M. Z. Doghmane, and M. Kidouche, "Dynamic Interaction Analysis of Coupled Axial-Torsional-Lateral Mechanical Vibrations in Rotary Drilling Systems," *Struct. Durab. Health Monit.*, vol. 19, no. 1, pp. 77–103, 2025, doi: 10.32604/sdhm.2024.053541.
- [18] S. Meddah, M. Z. Doghmane, S. A. Tadjer, A. Idir, M. Kidouche, and K. F. Tee, "Fractional-order PID controller design for strongly coupled high-frequency axial-torsional vibrations in drill string system," *Int. J. Dyn. Control*, vol. 13, no. 2, p. 94, Feb. 2025, doi: 10.1007/s40435-025-01602-z.
- [19] S. Ding, P. Fang, G. Li *et al.*, Axial-torsional nonlinear vibration of bottom hole assembly in the air drilling technology, *Petroleum*, May 2023, <https://doi.org/10.1016/j.petlm.2023.05.001>

Enhanced 2.4 GHz Microstrip Patch Antenna with DGS for Energy Harvesting

Riadh DEGACHI

*Department of Electrical Engineering, LGEERE Laboratory
University of El-Oued
El-Oued, Algeria
degachi-riadh@univ-eloued.dz*

Said GHENDIR

*Department of Electrical Engineering, LGEERE Laboratory
University of El-Oued
El-Oued, Algeria
said-ghendir@univ-eloued.dz*

Abstract—This paper presents the design and simulation of a 2.4 GHz microstrip patch antenna for RF energy harvesting applications. The antenna is built on an FR4 substrate ($\epsilon_r = 4.4$, $h = 1.6$ mm) with patch dimensions of 29.4 mm $\times$ 38 mm and a ground plane of 60 mm $\times$ 60 mm, fed by an inset feed technique. Initially, the conventional patch achieved a total gain of 2.91 dBi. To improve performance, H-shaped slots were added in the ground plane, resulting in a gain increase to 3.42 dBi. Simulation results of S11, VSWR, gain, and radiation patterns are analyzed and discussed using HFSS version 15, demonstrating the suitability of the proposed design for low-power RF energy harvesting applications.

Index Terms—Microstrip patch antenna, RF energy harvesting, defected ground structure, H-slot, S11, VSWR, antenna gain, FR4, HFSS

I. INTRODUCTION

Energy harvesting, the process of capturing ambient energy from the environment and converting it into usable electrical power, has become increasingly important due to the rapid growth of wireless devices, Internet-of-Things (IoT) systems, and wireless sensor networks (WSNs). Ambient RF signals, such as those emitted by Wi-Fi routers, GSM, and other communication systems, represent a convenient source of energy that can be collected using compact antennas. RF energy harvesting offers a sustainable and maintenance-free power solution for low-power devices, reducing dependency on batteries and promoting green technologies [1].

Over the past decade, numerous research efforts have focused on designing antennas specifically for RF energy harvesting. Various techniques such as microstrip patch antennas, monopole antennas, and loop antennas have been proposed. Modifications in the antenna structure, including defected ground planes, slots, and parasitic elements, have been employed to enhance gain, bandwidth, and efficiency, which are crucial parameters for energy harvesting performance [2].

Shanto et al. [3] proposed two rectenna architectures based on a dual-band microstrip patch antenna operating at 5.41 GHz and 7.36 GHz, optimized in CST for RF energy harvesting. Their design achieved excellent performance metrics including a return loss of -30.04 dB, a gain of 7.64 dB, and a radiation efficiency of 93.55%, alongside rectifier circuits modeled in ADS and LTspice for improved RF-to-DC conversion.

Dutta et al. [4] designed a microstrip patch antenna for RF energy harvesting at the 2.46 GHz ISM band, demonstrating strong performance with a return loss of -26.42 dB, a VSWR of 1.1, and a directivity of 6.14 dBi. Their system incorporates a rectification stage capable of delivering up to 3.5 V, making it suitable for powering low-energy electronic devices.

Ansah et al. [5] developed a compact three-element rectangular patch antenna array optimized for 2.45 GHz energy harvesting applications. Using CST Microwave Studio, they achieved a gain of 11.24 dBi, a return loss of -32.886 dB, and a total efficiency of 89.41%, demonstrating strong suitability for Wi-Fi-band RF energy collection.

Pervez et al. [6] introduced a beamforming-assisted RF energy harvesting architecture for dense 5G small-cell networks. Their approach integrates adaptive beam steering with an optimized mmWave rectenna at 28 GHz, achieving up to 75% harvesting efficiency under multipath and urban-blockage scenarios.

Rajeshkumar et al. [7] presented a single-fed clover-shaped microstrip rectenna operating in the 2.4 GHz ISM band, integrating the rectifier on the same plane to reduce the overall profile. The antenna achieves a peak gain of 7.19 dB and a maximum RF-to-DC conversion efficiency of 79%, making it effective for ambient RF energy harvesting.

Iqbal et al. [8] designed and optimized a microstrip patch antenna for 2.4 GHz energy harvesting applications. Their design achieves a gain of 7.2 dB, a return loss of -20 dB, and a directivity of 7.44 dBi, as well as a rectified output voltage of 2.5 V from an input of 900 mV, demonstrating efficient energy conversion.

Reza [9] proposed an optimized U-slot folded-shortened patch antenna designed for GSM-900 RF energy harvesting. The CST-based design achieves a 17.73% impedance bandwidth and stable unidirectional radiation characteristics, making it suitable for low-frequency rectenna systems.

Benkalfate et al. [10] developed a compact triple-band triangular microstrip antenna for IoT-oriented RF energy harvesting. Realized on various substrates, the antenna covers the 2.45 GHz, 5.2 GHz, and 8.2 GHz bands with omnidirectional or quasi-omnidirectional radiation patterns and maximum gains of 2.6 dBi, 4.55 dBi, and 6 dBi, respectively. Their design also demonstrates strong RF-to-DC conversion efficiency and low

SAR values suitable for biomedical and wearable applications.

In this work, we present a simple yet effective design of a microstrip patch antenna operating at 2.4 GHz intended for RF energy harvesting applications. The proposed design involves a conventional rectangular patch initially, followed by the introduction of H-shaped slots in the ground plane to improve the antenna's total gain and performance. The simulations were conducted using HFSS version 15.

The paper is organized as follows: Section II presents the theoretical background and conventional design formulas for microstrip patch antennas, including the calculation of patch dimensions and ground plane size. Section III describes the simulation process, starting with the conventional patch antenna, analyzing S11, VSWR, and gain, followed by the introduction of the H-shaped defected ground plane and its effect on antenna performance. Section IV concludes the paper and summarizes the key results and findings.

II. ANTENNA DESIGN AND THEORETICAL BACKGROUND

A. Patch Dimensions and Substrate

The design follows standard microstrip patch antenna formulas:

$$W = \frac{c}{2f_r} \sqrt{\frac{2}{\epsilon_r + 1}} \quad (1)$$

$$\epsilon_{eff} = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2} \left(1 + 12 \frac{h}{W}\right)^{-1/2} \quad (2)$$

$$\Delta L = 0.412h \frac{(\epsilon_{eff} + 0.3)(W/h + 0.264)}{(\epsilon_{eff} - 0.258)(W/h + 0.8)} \quad (3)$$

$$L = \frac{c}{2f_r \sqrt{\epsilon_{eff}}} - 2\Delta L \quad (4)$$

B. Antenna Properties and Dimensions

After deriving the theoretical calculations, Table I summarizes the main dimensions and characteristics of the antenna built in HFSS.

TABLE I
ANTENNA DIMENSIONS AND PROPERTIES

Parameter	Value (mm)
Patch Length (L)	29.4
Patch Width (W)	38
Ground Plane Size	60 × 60
Substrate height (h)	1.6
Feed length	24.8
Feed width	3

Figure 1 illustrates the 3D view of the designed microstrip patch antenna.

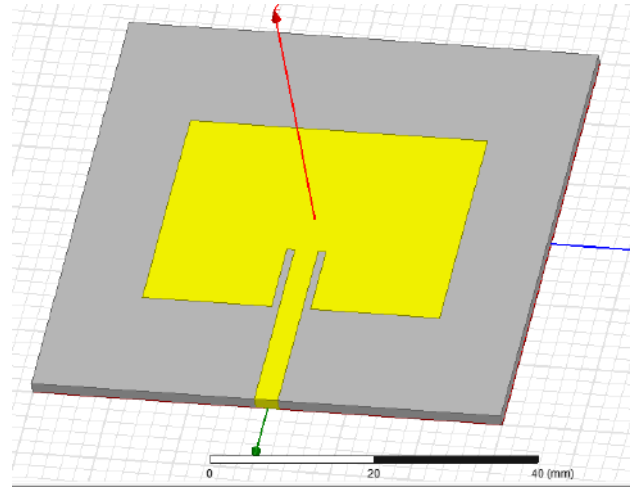


Fig. 1. 3D view of the designed microstrip patch antenna in HFSS.

III. SIMULATION RESULTS

A. S11 (Return Loss)

The return loss (S11) represents the amount of power reflected back toward the source due to impedance mismatch. For efficient operation, S11 should be below -10 dB at the target frequency. Figure 2 illustrates the simulated return loss of the proposed antenna. A minimum S11 of -22.99 dB is achieved at 2.43 GHz, confirming good impedance matching and proper resonance of the designed structure.

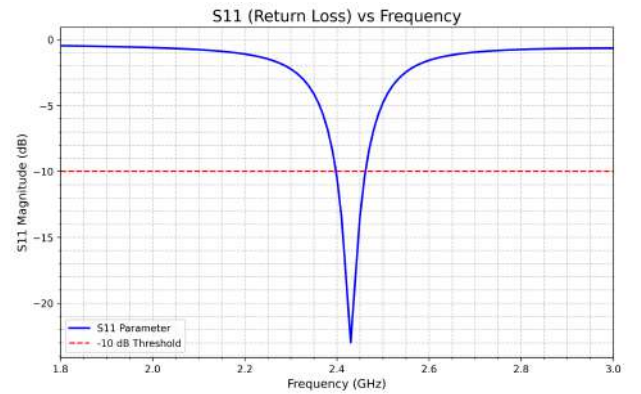


Fig. 2. Simulated S11 of the antenna.

B. VSWR (Voltage Standing Wave Ratio)

VSWR represents the quality of impedance matching between the antenna and the feeding line. Typically, a VSWR value below 2 indicates acceptable matching in practical RF systems. As illustrated in Figure 3, the proposed antenna exhibits an excellent VSWR of 1.15, remaining well below the acceptable threshold across the operating band. This behavior further confirms the good matching conditions observed in the S11 response.

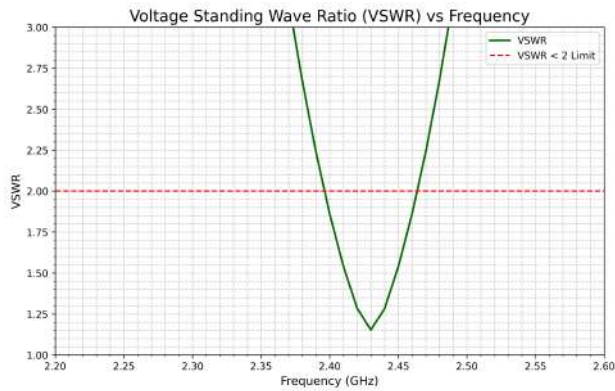


Fig. 3. Simulated VSWR of the antenna.

C. Gain

Antenna gain represents the antenna's ability to concentrate radiated energy in a particular direction, which is especially important in energy-harvesting applications where higher gain directly increases the amount of captured RF power. The initial rectangular patch antenna exhibited a gain of 2.91 dBi at 2.43 GHz (Figure 4).

To improve this performance, two arrays of H-shaped slots (12 positioned on the upper side of the ground plane and 12 on the lower side) were introduced (Figure 5). These slots alter the surface current distribution and effectively suppress surface-wave losses. As a result, the antenna gain increased to 3.42 dBi, yielding an overall improvement of 0.51 dB without modifying the patch dimensions (Figure 6). This demonstrates that the H-shaped defected ground structure enhances radiation efficiency and contributes to improved performance for energy-harvesting applications.

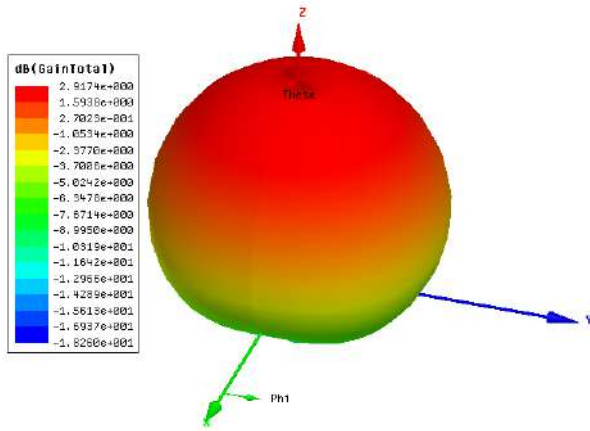


Fig. 4. 3D radiation gain pattern before insert H-slots in the ground plane.

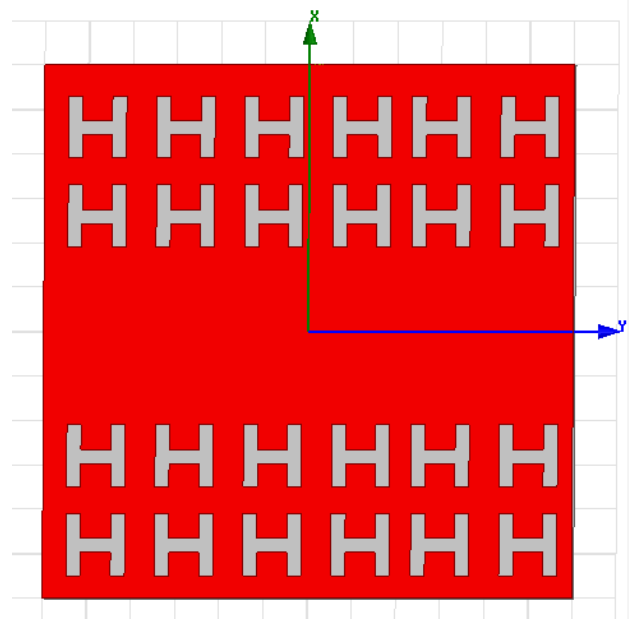


Fig. 5. The H-slots added to the ground plane.

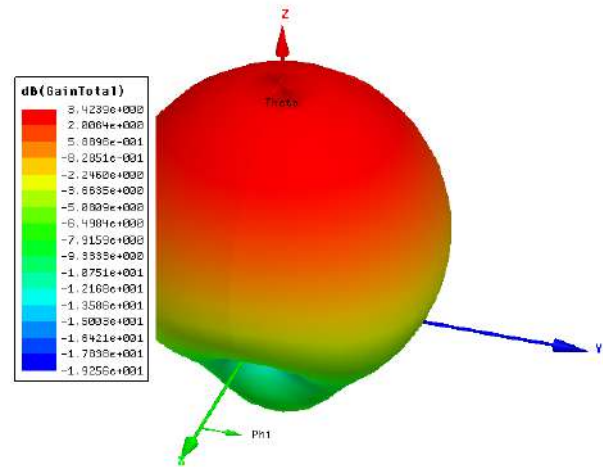


Fig. 6. 3D radiation gain pattern after insert H-slots in the ground plane.

IV. CONCLUSION

A 2.4 GHz microstrip patch antenna was successfully designed and simulated using HFSS version 15 for RF energy harvesting applications. The conventional patch antenna achieved a gain of 2.91 dBi with satisfactory S11 and VSWR characteristics. To enhance the antenna performance, H-shaped slots were introduced in the ground plane, which improved the gain to 3.42 dBi while maintaining excellent impedance matching and voltage standing wave ratio. The defected ground structure effectively modified the current distribution, reducing surface-wave losses and increasing radiation efficiency. Simulation results demonstrate that the proposed design provides a simple, compact, and practical solution for low-power RF energy harvesting systems, making it suitable

for integration in Internet-of-Things (IoT) devices and other wireless low-power electronics. The study also highlights the potential of minor structural modifications, such as ground-plane slotting, to significantly enhance antenna performance without increasing size or complexity.

REFERENCES

- [1] X. Lu, P. Wang, D. Niyato, D. I. Kim, and Z. Han, "Wireless networks with rf energy harvesting: A contemporary survey," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 757–789, 2014.
- [2] M. Wagih, A. S. Weddell, and S. Beeby, "Rectennas for radio-frequency energy harvesting and wireless power transfer: A review of antenna design [antenna applications corner]," *IEEE Antennas and Propagation Magazine*, vol. 62, no. 5, pp. 95–107, 2020.
- [3] M. S. I. Shanto, M. M. Rahman, N. Sarker, L. C. Paul, M. A. I. Romel, T. Rani, and M. R. H. Mondal, "Design and analysis of a dual-band microstrip patch antenna for energy harvesting," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2024, pp. 1–5.
- [4] B. R. Dutta, R. Chaudhury, A. Tripathi, and B. Biswas, "Advance t-shaped microstrip patch antenna for improved radio frequency energy harvesting," *Telecommunications and Radio Engineering*, vol. 83, no. 1, 2024.
- [5] J. N. Ansah and J. K. Annan, "A novel triangular configuration of a rectangular patch antenna array for 2.45 ghz rf energy harvesting," 2024.
- [6] M. S. Pervez and T. B. Haider, "Design and simulated mmwave patch antenna of beamforming-assisted rf energy harvesting for dense 5g small-cell networks," *Global Journal of Engineering and Technology Advances*, vol. 23, no. 3, pp. 001–009, 2025, research Article. [Online]. Available: <https://doi.org/10.30574/gjeta.2025.23.3.0183>
- [7] N. Rajeshkumar and P. Sathya, "Design of rf energy harvesting patch antenna for wireless communications," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, no. 10, pp. 3551–3554, 2019, open Access under CC-BY-NC-ND license.
- [8] M. S. Iqbal, T. J. Khanzada, F. A. Dahri, A. Ali, M. Ali, and A. W. Khokhar, "Analysis and maximizing energy harvesting from rf signals using t-shaped microstrip patch antenna," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 1, 2019.
- [9] D. Kundu, A. W. Reza, and H. Ramiah, "Design of an u-slot folded shorted patch antenna for rf energy harvesting," in *2nd International Conference on Innovations in Engineering and Technology*, 2014, pp. 61–65.
- [10] C. Benkalfate, A. Ouslimani, A.-E. Kasbari, and M. Feham, "A new compact triple-band triangular patch antenna for rf energy harvesting applications in iot devices," *Sensors*, vol. 22, no. 20, p. 8009, 2022.

Identification of Electrical Faults in Photovoltaic System using Artificial Intelligence

Aicha Moulay

Laboratory of Sustainable Development and Informatics

Ahmed Draia University, 01000 Adrar, Adrar, Algeria

aichamoulay@univ-adrar.edu.dz
aichamoulay03@gmail.com

Khaled Koussa

Unité Recherche en Energies Renouvelables en Milieu Saharien, URERMS

Centre de Développement des Energies Renouvelables, CDER

Adrar, Algeria
koussakhaled1986@gmail.com

Othmane Abdelkhalek

Smart Grids and Renewable Energies Laboratory

University Tahri Mohammed of Béchar 08000 Béchar

Bechar, Algeria
othmane08000@yahoo.fr

Abstract— This paper presents an intelligent method for the detection and classification of electrical faults in photovoltaic (PV) systems using neural networks. A three-layer artificial neural network requires meteorological data (solar irradiance and temperature) in addition to electrical data (array voltage and current) and is used to identify electrical faults. The proposed method is simple and efficient for detecting and classifying both faulty and normal operating conditions. The approach is validated through simulations using MATLAB/Simulink.

Keywords— Photovoltaic system; artificial intelligence; Fault detection; Classification;

I. INTRODUCTION

Renewable energy, such as wind, solar, water, biomass, and geothermal energy, is a natural source that is constantly replenished[1]. These resources are more reliable compared to fossil fuels due to their permanent availability and non-pollution of the environment. Solar energy is one of the most important renewable energy resources in use around the world because of its inexhaustible nature and the absence of mechanical parts [2].

Nowadays, the diagnosis of photovoltaic(PV)systems plays an important role in maintaining the power generated from PV systems. A monitoring research shows that faults can reduce the power generated by a PV system annually by an effective rate. Each module in photovoltaic system must be ensured to perform properly in order to maintain the continuous generated power and avoid the damage faults. Electric problems are often caused by long-term operation [3][4]

II. PHOTOVOLTAIC SYSTEM MODELING

A. PV module

The electrical characteristics of the photovoltaic module used under standard conditions STC (Temperature Module =25°C, Solar Irradiance=1000W/m²) are shown in Table.I

TABLE I. PV MODULE SPECIFICATIONS

Parameters	Values
Maximum Power (Pmax)	30 W
Voltage at Pmax (Vmpp)	17.80 V
Current at Pmax (Impp)	1.666 A
Open circuit voltage (Voc)	22.2 V
Short circuit current (Isc)	1.801 A
Temperature coefficient of Voc	-0.34 V/°C
Temperature coefficient of Isc	0.0037 mA/°C
Nomina Operating Cell Temperature (NOCT)	45 ±2 °C
Operating Temperature	25 °C

B. The P-V and I-V characteristics of PV module under STC conditions:

The characteristics of PV module under STC condition are shown in Fig. 1 and Fig. 2.

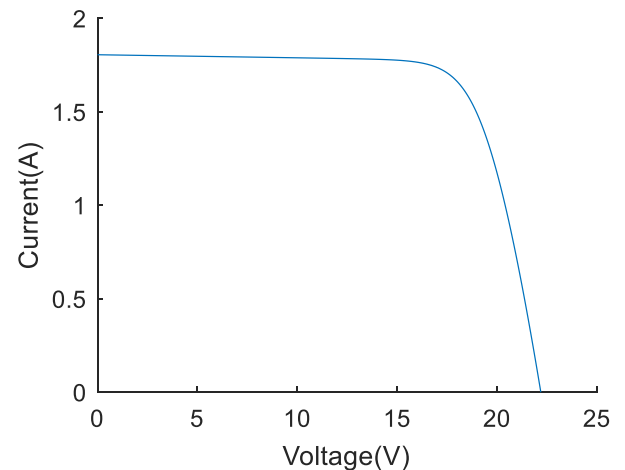


Fig. 1. I_V characteristics of PV module under STC conditions

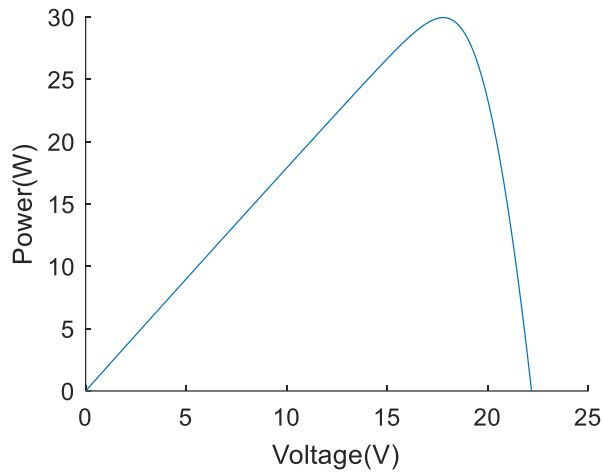


Fig. 2. P_V characteristics of PV module under STC conditions

The proposed PV array system consists of two strings arranged in parallel, each string contain three modules joined in series. In solar panels, bypass diodes are connected "parallel" to a photovoltaic cell or panel to shunt current around it, whereas blocking diodes are connected "series" to the PV panels to prevent current from flowing back into them. Fig.3 below describes the proposed panels with bypass and blocking diodes.

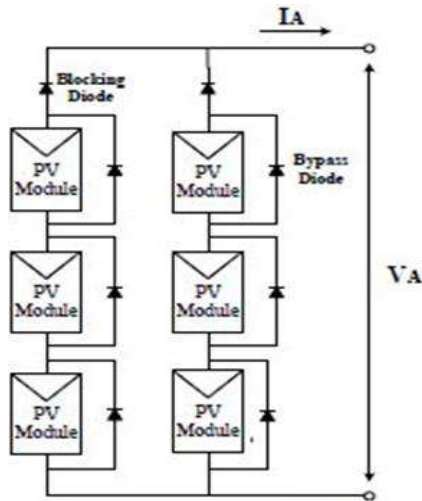


Fig. 3. Block diagram of PV panels

C. PV system modeling

The schematic of a grid connected PV system is shown in Fig. 4. It consists of the PV array, DC-DC converter which is driven by MPPT (Maximum Power Point Tracking) controller and the charge load.

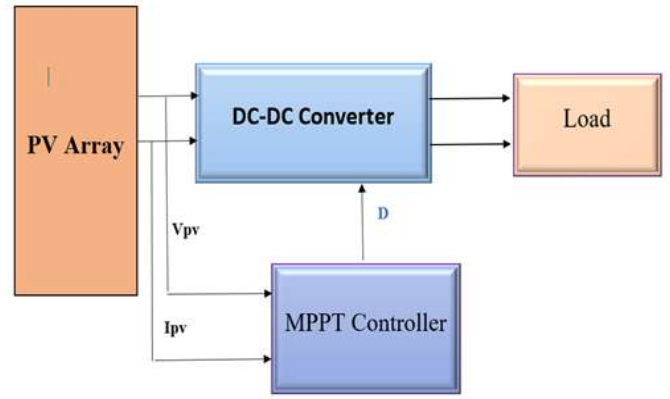


Fig. 4. Block diagram of PV stand-alone PV system

$$D_{MPP} = 1 - \sqrt{\frac{R_{MPP}}{R_L}} \quad (1)$$

MPPT algorithm has been designed to continuously adjusts the duty cycle of the converter in order to extract maximum available power. At MPP ($R_{in}=R_{MPP}$), the duty cycle of the converter is calculated as described in equation (1) [5].

The parameters of boost converter are calculated using [5]–[7] and the result values are described in table below:

Table. II Boost converter components parameters

Parameters	Values
Inductor	2 mH
Input Capacitance	212 μ F
Output Capacitance	24 μ F
Switching frequency	5 KHz
Duty Cycle	0.67

III. PHOTOVOLTAIC FAULTS

A. The studied faults

In this paper we study 4 different cases the normal case and 3 different faults which are:

- Open circuit fault: occurs when a solar panel is disconnected making the current of other solar panels associated with the same string flows across the bypass diode.[8]

- Line to Line fault: is an accidental low resistance between two points of different potential in photovoltaic arrays, it can happen in the same string or two adjacent strings [9][10].
- Ground fault: It is unintentional low impedance between the ground and one or more current-carrying inductors. A ground-fault is defined as a line-line fault with any ground points[9][10].

Line-line faults in PV arrays can occur for a variety of reasons, such as:

- Cable insulation failure, i.e. an animal chewing through cable insulation;
- Incidental short circuit between current carrying conductors,

Open-circuit faults are frequently caused by joint failures of cables, overhead lines, or failure from a melting fuse or conductor.

B. I-V and P-V characteristics of PV system under different cases

Fig. 5 and Fig. 6 illustrate the resultant I-V and P-V curves for all studied cases with developed Matlab/Simulink model.

O-C: Open circuit case

L to L: Line to Line case

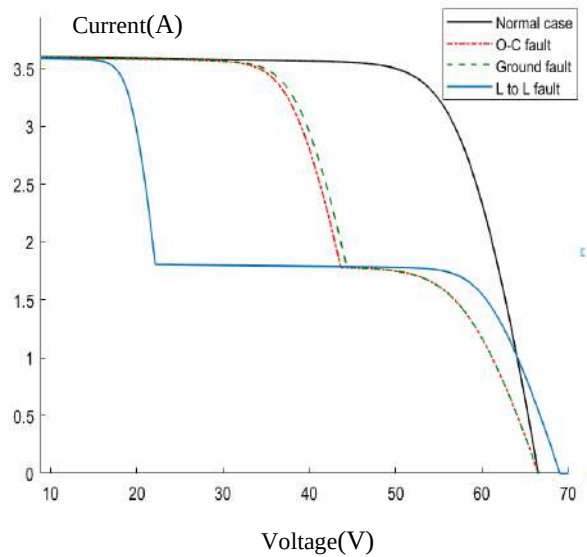


Fig. 5. I_V characteristics of the PV array under electrical faults in STC conditions ($T=25^{\circ}\text{C}$, $G=1000\text{W/m}^2$)

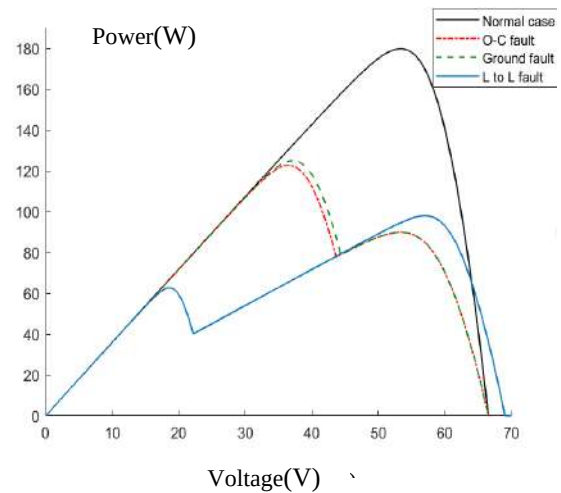


Fig. 6. P_V characteristics of the PV array under electrical faults ($T=25^{\circ}\text{C}$, $G=1000\text{W/m}^2$)

IV. METHODOLOGY

A. Artificial Neural Networks(ANNs)

Artificial neural networks are inspired by biological neural networks that imitate data in human brains. In the implementation of ANN, knowledge is described by numeric values called weights. The multilayer perceptron (MLP) is one of the most well-known ANNs used in modeling and approximation. It is made up of three layers; an input layer, hidden layers, and an output layer, each with a specific number of neurons. The number of input and output layers is respectively equal to the number of inputs and outputs of the real system represented by a ANN, while the number of input and output layers is equal to the number of inputs and outputs of the real system respectively represented by ANN, whereas the number of hidden layers and their neurons is determined by the user. Researchers have demonstrated that the most nonlinear systems can be represented by a single hidden layer with varying neurons' number and activation function type. Fig. 7 illustrates a simple MLP neural network structure. The MLP is trained using back-propagation with the conjugate gradient algorithm, which is mostly used to adopt the weights in order to minimize errors between desired values (targets) and the network's outputs [11][12].

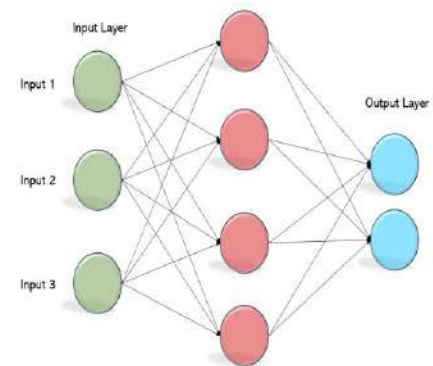


Fig. 7. A three-input two-output multi-layer perceptron ANN

B. Methodology used for fault classification

In this study, 168,000 data points were generated from simulations performed in MATLAB/Simulink. Each data point includes solar irradiance, ambient temperature, and the corresponding PV array voltage and current. The dataset contains multiple cases of open-circuit faults, short-circuit faults, as well as normal operating conditions. These scenarios cover a solar irradiance range of 400–1100 W/m² and an ambient temperature range of 20–45°C.

A total of 80% of the dataset was used for training in order to update the network's weights and biases, while the remaining 20% was reserved for testing the MLP model by comparing the estimated outputs with the actual data. The employed ANN is a three-layer multilayer perceptron (MLP) with 10 neurons in the hidden layer, using a sigmoidal activation function and trained with the Levenberg–Marquardt backpropagation algorithm [13]. The four input variables of the dataset were normalized to the range [-1, 1], and the output classes were encoded using a four-bit binary representation.

A three-layer ANN having 10 neurons in the hidden layer with a log sigmoidal activation function trained by Levenberg-Marquardt back propagation algorithm is used in this work.

V. RESULTS AND DISCUSSION

The architecture of the MLP neural network and the best validation performance are shown in Fig.8 and Fig.9 respectively.

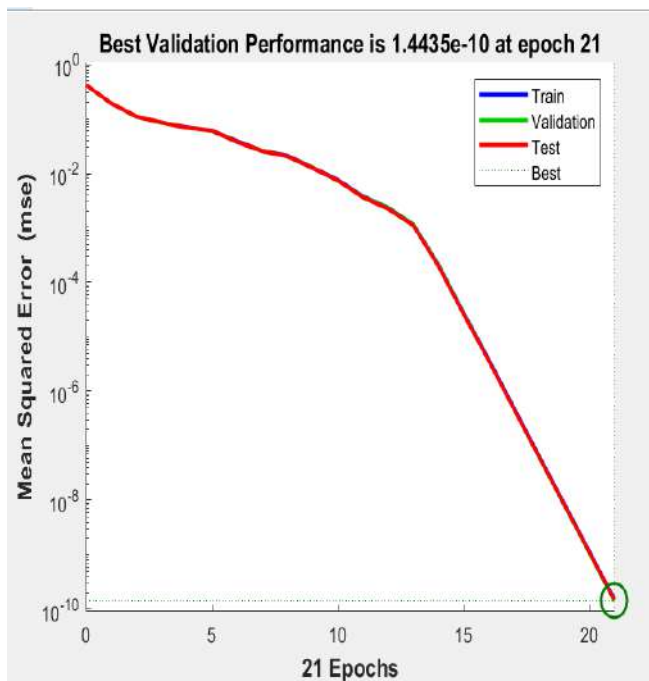


Fig. 8. Neural network training diagram

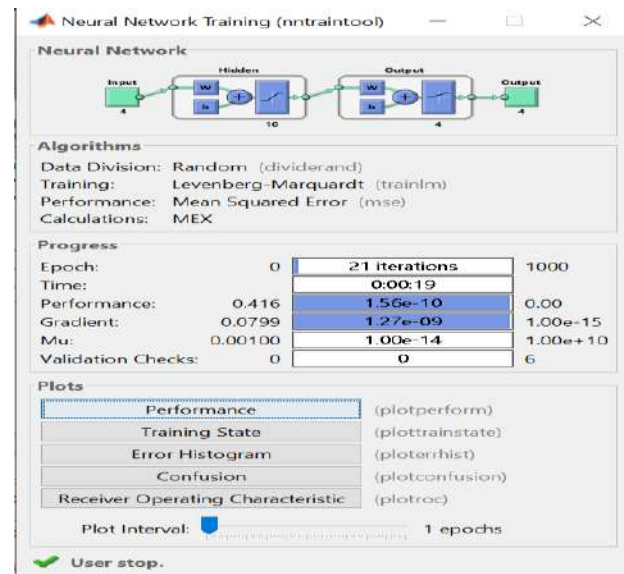


Fig. 9. Best validation performance curve

The obtained results demonstrate that the proposed methodology performs effectively in detecting and classifying both normal and faulty operating states. Moreover, this ANN topology can be extended to identify additional electrical faults, including boost converter faults and partial shading conditions. To further validate the model, external data were used, and the results also confirmed its strong performance. Figure 10 presents the confusion matrix of the proposed ANN, illustrating the high accuracy achieved by the classifier.



Fig. 10. Confusion matrix diagram

From the confusion matrix, the accuracy and precision can be calculated as follows:

$$\text{Precision} = 100\%$$

$$\text{Accuracy} = 100\%$$

VI. CONCLUSION

This paper presents an artificial neural network-based methodology for the detection and classification of electrical faults in photovoltaic (PV) systems. The simulation results demonstrate that the proposed approach achieves high accuracy in identifying different fault conditions, confirming its effectiveness and reliability. Given its strong performance, the method shows significant potential for extension to a wider range of fault types, including more complex or rare scenarios. Furthermore, the proposed framework could be adapted for deployment in real-world PV installations, where it may contribute to improving system monitoring, enhancing operational safety, and reducing maintenance costs. Future work may involve validating the model using field data and integrating it into real-time diagnostic platforms to further assess its practical applicability.

REFERENCES

- [1] M. Sabbaghpur Arani and M. A. Hejazi, "The comprehensive study of electrical faults in PV arrays," *J. Electr. Comput. Eng.*, vol. 2016, 2016, doi: 10.1155/2016/8712960.
- [2] A. Djalab, N. Bessous, M. M. Rezaoui, and I. Merzouk, "Study of the Effects of Partial Shading on PV Array," *Proc. - Int. Conf. Commun. Electr. Eng. ICCEE 2018*, pp. 6–10, 2019, doi: 10.1109/CCEE.2018.8634512.
- [3] Syafaruddin, E. Karatepe, and T. Hiyama, "Controlling of artificial neural network for fault diagnosis of photovoltaic array," *2011 16th Int. Conf. Intell. Syst. Appl. to Power Syst. ISAP 2011*, pp. 1–6, 2011, doi: 10.1109/ISAP.2011.6082219.
- [4] L. L. Jiang and D. L. Maskell, "Automatic fault detection and diagnosis for photovoltaic systems using combined artificial neural network and analytical based methods," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2015-September, 2015, doi: 10.1109/IJCNN.2015.7280498.
- [5] N. Hashim, Z. Salam, D. Johari, and N. F. Nik Ismail, "DC-DC Boost Converter Design for Fast and Accurate MPPT Algorithms in Stand-Alone Photovoltaic System," *Int. J. Power Electron. Drive Syst.*, vol. 9, no. 3, p. 1038, 2018, doi: 10.11591/ijpeds.v9.i3.pp1038-1050.
- [6] B. M. Ahmed and N. F. Farman ALhialy, "Optimum Efficiency of PV Panel Using Genetic Algorithms to Touch Proximate Zero Energy House (NZEH)," *Civ. Eng. J.*, vol. 5, no. 8, pp. 1832–1840, 2019, doi: 10.28991/cej-2019-03091375.
- [7] S. Motahhir, A. El Ghzizal, S. Sebti, and A. Derouich, "Modeling of Photovoltaic System with Modified Incremental Conductance Algorithm for Fast Changes of Irradiance," *Int. J. Photoenergy*, vol. 2018, 2018, doi: 10.1155/2018/3286479.
- [8] M. S. Montaser Abd El-Sattar, "Study, Design and Performance Analysis of Grid-Connected Photovoltaic Power Systems," no. October, p. 188, 2015, doi: 10.13140/RG.2.2.23240.72961.
- [9] M. K. Alam, F. H. Khan, J. Johnson, and J. Flicker, "PV faults: Overview, modeling, prevention and detection techniques," 2013 IEEE 14th Work. Control Model. Power Electron. COMPEL 2013, 2013, doi: 10.1109/COMPEL.2013.6626400.
- [10] Y. Zhao, J. F. De Palma, J. Mosesian, R. Lyons, and B. Lehman, "Line-line fault analysis and protection challenges in solar photovoltaic arrays," *IEEE Transactions on Industrial Electronics*, vol. 60, no. 9, pp. 3784–3795, 2013, doi: 10.1109/TIE.2012.2205355.
- [11] F. Salem and M. A. Awadallah, "Detection and assessment of partial shading in photovoltaic arrays," *J. Electr. Syst. Inf. Technol.*, vol. 3, no. 1, pp. 23–32, 2016, doi: 10.1016/j.jesit.2015.10.003.
- [12] H. Mekki, A. Mellit, and H. Salhi, "Artificial neural network-based modelling and fault detection of partial shaded photovoltaic modules," *Simul. Model. Pract. Theory*, vol. 67, pp. 1–13, 2016, doi: 10.1016/j.simpat.2016.05.005.
- [13] A. Moulay, T. Benslimane, O. Abdelkhalek, and K. Koussa, "Detection and Classification of Photovoltaic System Faults using Neural Network," *Przegląd Elektrotechniczny*, vol. 2023, no. 11, pp. 157–162, 2023, doi: 10.15199/48.2023.11.27

Ensemble Learning for Delineation of Raining Cloud over the North of Algeria.

BENSAFI Nouredine
Electrical Systems Engineering Department, UMBB
Laboratoire LAMPA, UMMTO
Algeria
n.bensafi@hotmail.com

LAZRI Mourad
Electronics Department, UMMTO
Laboratoire LAMPA, UMMTO
Tizi Ouzou, Algeria
m_lazri@yahoo.fr

Abstract— In this paper, we presented the application of the Ensemble Learning techniques to classify cloud systems, namely stratiform and convective systems. The data used come from the SEVIRI (Spinning Enhanced Visible and Infra-Red Imager) radiometric imager on board the MSG (Meteosat Second Generation) satellite. The methods were trained and evaluated with satellite and radar data of the rain event of October 10, 2018 in the south of France provided by Météo-France. This allowed us to evaluate and compare the performances for cloud identification of two methods frequently used in combination of classifiers.

Keywords— Machine Learning, Ensemble Learning, Bagging, Boosting, Rainfall Estimation, Delineation of Raining Cloud, SEVIRI, MSG, Radar.

I. INTRODUCTION

The techniques based on the supervised learning methods proposed, for solving the problem of identifying cloud systems and estimating precipitation [2, 3, 11, 12], exploit the information from the multi-spectral channels of the SEVIRI (Spinning Enhanced Visible and Infra-Red Imager) radiometer imager. Some of these methods have shown a performance that surpasses the others to predict one or more classes. The idea of making these different methods cooperate in order to improve the precision of the classification appeared with the development of new called hybrid techniques.

Two approaches are used for multi-classifier cooperation. The first approach is to merge information from different classifiers for decision making, considering each classifier as an independent source of information and exploiting the characteristic performance of these sources to favor the most reliable classifier to predict the class membership. To do this, classifier fusion relies on three main strategies that are the voting principle [10, 16], possibility theory [7, 17] and belief theory [1, 4]. Each of these strategies is adapted for the type of information processed or the relevant decision criterion for classification. The second approach is to use strategies or mechanisms for classifiers using different subset of data for each classifier or using different classifiers for a set of training data. These strategies are at the origin of several popular multi-classifier algorithms such as Boosting [9] and Bagging [5].

It is the second approach that we will focus on this study. To do this, this paper we will detail the principle of these different mechanisms of Ensemble Learning techniques used in Section 2, and by subjecting our database to tests that will be analyzed in Section 3. Finally, we draw some conclusions in Section 4.

II. ENSEMBLE LEARNING

Ensemble Learning consists of combining the prediction of a set of independent classifiers. What has been at the origin of a very active field of research in supervised learning, by exploiting the local precision of the different learning methods to form a set of classifiers, which improves the global precision for the same application [5]. Several methods have been implemented by combining several classifiers, in this section we will expose some methods, which have shown their effectiveness for the data processed in our research study.

The principle of Bagging is "Bootstrap aggregation", it is a resampling of the training dataset, by randomly generating other subsets of data from the original training set, while keeping the number of samples it contains, which means the appearance of the same sample several times and the disappearance of several others in the same subset [4]. Indeed, a larger classifier prediction error for the absent or less represented classes in the subsets. However, the combination of classifiers used for the subsets gives higher prediction accuracy than a single classifier and a training dataset. The advantage of bagging is the reduction of the variance without influencing the accuracy of the prediction. Also, the study conducted by [4] the effectiveness of Bagging is more significant for learning methods whose prediction decision is significantly different when there is an insignificant change in the learning set

Boosting consists of a series of classifiers of different learning methods whose prediction is weak. Its principle is that the samples misclassified by the first classifier are chosen more often unlike the correctly predicted samples, with the aim of constituting the training data set submitted to the next classifier, this step is repeated for all of the classifier until the end of the series [8]. Methods based on the Boosting principle have been developed, the Arcing method and the Ada-Boost method [8]. The Arcing method [4], the frequency of samples misclassified by a classifier is the selection criterion in the training dataset submitted to the next classifier; this step is repeated until the end of the series of classifiers. The Ada-Boost method [8] uses a series of classifiers, which exploits the probabilities of the set of samples at the output of a classifier, and then weight the prediction error of each sample, the new set retains the completeness of the samples in the training set.

III. DATA AND APPLICATION

Three types of climate for which their precipitation rates are important. Namely, the oceanic climate found in Western Europe bordered by the Atlantic Ocean to the west, with annual precipitation varying between 900 and 1800 mm in the

form of drizzle and fog. In addition, the Mediterranean climate, with annual precipitation varying between 250 and 1150 mm. Also, the transitional climate of the mountain ranges of the Tell Atlas and the Saharan Atlas, which separates the region bordered by the Mediterranean Sea from the desert from the Sahara to the south, with annual precipitation decreasing to less than 100 mm south of the Saharan Atlas. The exclusion of the arid desert climate from our study is due to very rare precipitation in the form of downpours and of little interest for our study, with annual quantities which often do not exceed 50 mm. The rainy season lasts from October to March. A band is defined with a latitude of 30° to 50° North and a longitude of 6° West to 12° East as shown in following figure.

A. Data

The data used in this study are satellite and radar images, which represent a rainy and stormy event in the Mediterranean and going back to the south of France from October 10, 2018 from 08h GMT to 18h GMT. In addition, precipitation intensities exceeding 30 mm/h and cumulative rainfall more than 250 mm in certain regions, which is an indication of the presence of an intense convective system, caused by an anticyclone in the center of Europe, a depression in the North Atlantic Ocean and upwelling of a warm air masses from Africa.

The satellite data are images acquired by the SEVIRI radiometer imager (Spinning Enhanced Visible and Infra-Red Imager) on board the MSG4 (Meteosat Second Generation 4) satellite. These images are captured every 15 minutes in 12 spectral channels and spatial resolution of 3 km above the equator. Also, the images were processed in real time in order to correct the radiometric and geometric effects induced during the capture and geo-located, then the information is localized on the Earth, calibrated and the radiances linearized so that they are exploitable [8].

These data exploited in order to provide parameters for the identification of the characteristics of clouds and the types of precipitation of each pixel of the images captured, the parameters we will discuss in this section.

Two types of clouds are the source of significant precipitation in the region of our study, stratiform and convective clouds, which can be detected by exploiting the parameters of the cloud as detailed in the study [3]. Relationship which connects them to the data provided by SEVIRI channels and summarized by the following table.

TABLE I. RELATIONSHIP BETWEEN CLOUD PARAMETERS AND SEVIRI DATA

Parameters	Channel	Relationship
Cloud Top Temperature	$IR_{10.8}$	$AT = \frac{\sum_{i=1}^N T_{IR_{10.8}}(i)}{N}$
Vertical Extension of the Cloud	$\Delta T_{IR_{10.8}-IR_{12}}$	$H = \frac{T_{max} - T_{min}}{6.5} - dH$
Altitude of the Cloud	$\Delta T_{WV_{6.2}-IR_{10.8}}$ Or $\Delta T_{WV_{7.3}-IR_{12}}$	Very negative value $\Delta T > \Delta T_{min}$ Low altitude Very low value $\Delta T > \Delta T_{min}$ High altitude
Cloud Water Path	During the day $\Delta T_{VIS_{0.6}-NIR_{1.6}}$ During the night $\Delta T_{IR_{3.9}-IR_{10.8}}$	$CWP = \frac{2}{3} * \rho * \tau * r_e$
Cloud Phase	$\Delta T_{IR_{8.7}-IR_{10.8}}$	Low and negative value $\Delta T > \Delta T_{min}$ Low altitude

With

- N is the number of pixels that form the cloud.
- dH is the minimum height of the cloud or the rain column which is 650 m.
- ρ is the density of the water contained in the cloud.
- (τ) is the optical thickness and the effective particle radius (r_e) of the cloud.
- (r_e) is the effective particle radius of the cloud.

Radar data are images captured by the Météo-France radar network located in the south of France. These images are captured every 5 minutes with a spatial resolution of 1 km, processing has been applied to eliminate noise caused by the terrain and neighbouring radars [13]. The information is the reflectivity, which is the measurement of the energy reflected by the precipitation in (dBZ) and it gives us explicit information about the water content of the cloud.

For a spatial and temporal correspondence of the data manipulated, we carried out the co-localization and the synchronization of the satellite and radar data.

Table 1 illustrates the relationship for calculating the precipitation rate R (mm/h) from the reflectivity measured by the radar Z (dBZ) [12].

TABLE II. REFLECTIVITY AND PRECIPITATION RATE RELATION

Type	Relationship	Correlation Coefficient
Strat. Precipitation $Z \leq 38 \text{ dBZ}$	$Z = 248.28 \times R^1$	0.79
Conv. Precipitation $Z \geq 42 \text{ dBZ}$	$Z = 412.35 \times R^1$	0.84

For our application, we have chosen a database belonging to the event of October 10, 2018, which contains characteristic parameters of the clouds deduced from the SEVIRI data, which correspond to the ground truth data, which are the radar data. This database is contains 3187904 samples which correspond to eight scenes captured between 12h GMT and 14h GMT, which have been divided into two, 80% for training and 20% for the test of the two ensemble classifiers, Bagging and Ada-Boost that we have chosen for our application because they have shown good performance in other fields of application. We obtained the results transcribed in Table 1.

With

$$Z(\text{dBZ}) = 10 \log(Z) \quad (1)$$

B. Application

The input variables of our models are calculated from the data of multispectral channels of the SEVIRI radiometer of the Meteosat Second Generation satellite. Which contains characteristic parameters of the clouds deduced from the SEVIRI data, which correspond to the ground truth data, which are the radar data.

The model allows assigning a class to a novel sample using the samples from the training set. For this, we present as input the variables that characterize this novel samples; the class thus assigned by the model represents one of the three classes non-precipitant, stratiform precipitant and convective precipitant.

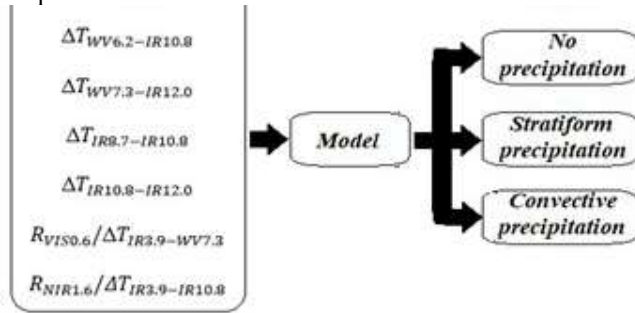


Fig. 1. Structure of the Bagging and Ada-Boost model that combines spectral parameters from SEVIRI images.

TABLE III. TABLE TYPE STYLES

Classes	Bagging accuracy	Ada-Boost accuracy	Total samples
No detection	0.87	0.71	341048

No precipitant	0.68	0.39	87386
Stratiform	086	0.67	205910
Convectif	0.77	0.00	3237

IV. RESULTS AND DISCUSSION

To evaluate the performance of the Bagging and Ada-Boost methods for cloud identification were tested to the SEVIRI and radar data samples, these datasets being taken simultaneously and respecting the spatial correspondence between them. We chose a database belonging to the stormy event of October 10, 2018. This database contains 3,187,904 samples, which correspond to eight scenes captured between 12 h UTC and 14 h UTC, which were divided in two, 80% for training and 20% for testing of the two ensemble classifiers, Bagging and Ada-Boost that we chose for our application because they showed good performance in other application areas.

In order to compare the Bagging and Ada-Boost methods with other methods, the two stratiform and convective classes are in a single class called precipitant.

The appropriate statistical parameters used to quantify the estimation errors of the methods with radar measurements calculated from the contingency table summarized in following table.

TABLE IV. TABLE TYPE STYLES

Identified by WkNN	Observed by the radar		
	Precipitating	Non-precipitating	Total
Precipitating	a	b	$a+b$
Non-precipitating	c	d	$c+d$
Total	$a+c$	$b+d$	T

Where a, b, c and d are contingency table values

And

$$T = a + b + c + d$$

- The rate of correctly identified events is calculated with the Probability Of Detection (POD).

$$POD = \frac{a}{a+b} \quad (2)$$

- The rate of number of pixels misidentified by the WKNN method is calculated with The Probability Of False Detection ($POFD$).

$$POFD = \frac{b}{b+d} \quad (3)$$

- The rate of events estimated when they were not events is calculated with the False Alarm Ratio (FAR).

$$FAR = \frac{b}{a+b} \quad (4)$$

- The difference between the estimate and the measurements is calculated with the Frequency BIAS index (*Bias*):

$$Bias = \frac{a+b}{a+c} \quad (5)$$

- The estimated correctly diagnosed event rate is calculated the Critical Success Index (*CSI*).

$$CSI = \frac{a}{a+b+c} \quad (6)$$

- The rate of correct estimates is calculated with the Percentage of Corrects (*PC*).

$$PC = \frac{a+d}{T} \quad (7)$$

- The comparison between the Bagging and Ada-Boost methods and *WkNN* method the *ECST* technique is summarized in following table.

TABLE V. THE STATISTICAL RESULTS OF THE VERIFICATION FOR ADA-BOOST AND BAGGING

	POD	POFD	FAR	Bias	CSI	PC
ECST	0.64	0.06	0.29	0.85	0.58	0.93
WkNN	0.68	0.05	0.26	0.89	0.61	0.95
Bagging	0.62	0.08	0.33	1.18	0.57	0.89
Ada-Boost	0.56	0.1	0.35	1.21	0.55	0.86
Optimal values	1	0	0	1	1	1

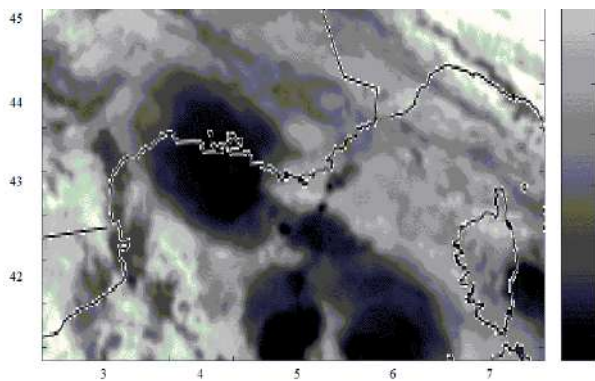


Fig. 2. $T_{IR10.8}$ image scene from October 10, 2018 at 12:40 UTC.

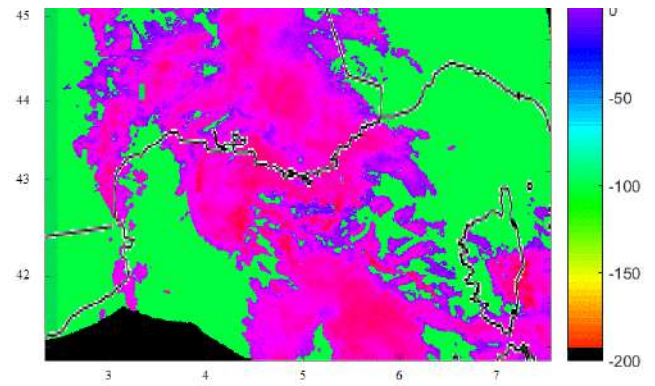


Fig. 3. The rain area delineated by the Météo-France radar network in dBZ.

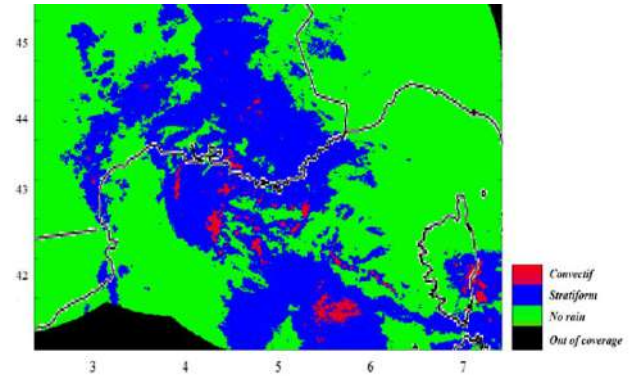


Fig. 4. The rain area delineated by the Météo-France radar network in 4 levels.

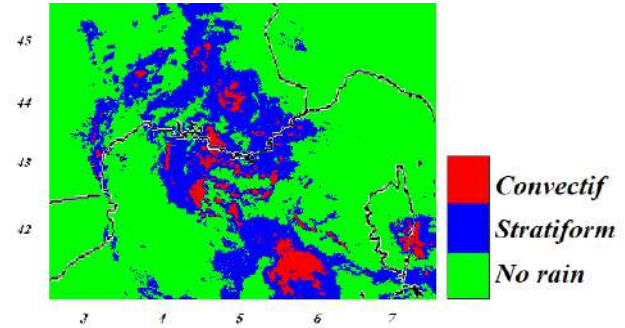


Fig. 5. The rain area delineated by the Bagging method in South East of France.

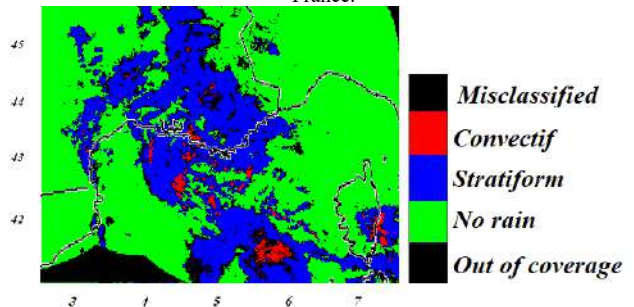


Fig. 6. Comparison between Bagging method and the Météo-France radar network delineation.

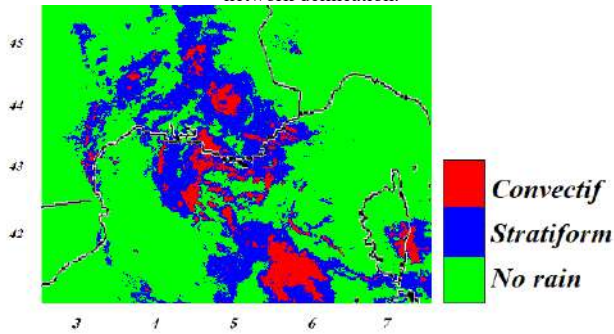


Fig. 10. The rain area delineated by the Bagging method over North of Algeria and Mediterranean sea.

V. CONCLUSION

In light of the results obtained, we deduce that the Bagging classifiers have better accuracy than the Ada-Boost classifier; this is valid for the four predefined classes. When the number of samples of the same class weakly represented in the database, the prediction error is greater for the Ada-Boost method. This is not the case for the Bagging method. The Bagging method is more suitable for a heterogeneous database, unlike the Ada-Boost method.

As a perspective to this study, we recommend to use a homogeneous database for an application to Ada-Boost classifiers.

Fig. 7. The rain area delineated by the Ada Boost method in South East of France.

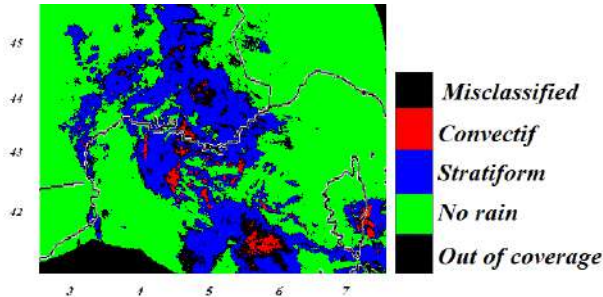


Fig. 8. Comparison between Ada Boost method and the Météo-France radar network delineation.

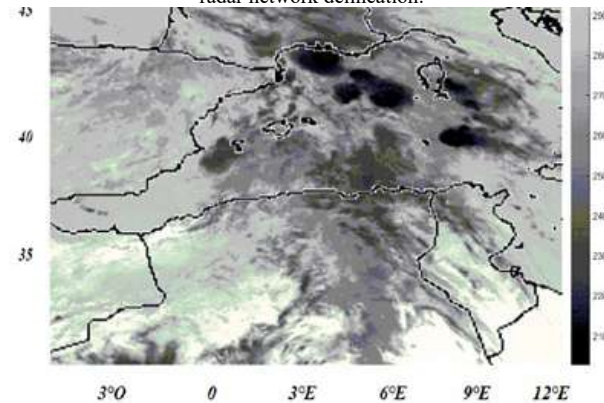
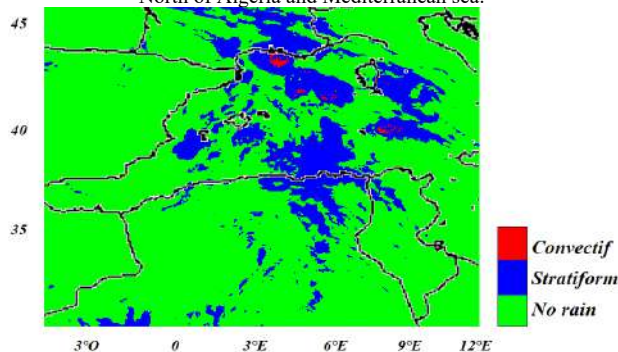


Fig. 9. $T_{IR10.8}$ image scene from October 10, 2018 at 12:40 UTC over North of Algeria and Mediterranean sea.



REFERENCES

- [1] Appriou, A. (2002), Discrimination multi signal par la théorie de l'évidence, In *décision et Reconnaissance des formes en signal*, Hermes Science Publication, pp 219-258, 2002.
- [2] Bensafi, N., Lazri, M., Ameer, S., 2019. Novel WkNN-based technique to improve instantaneous rainfall estimation over the north of Algeria using the multispectral MSG SEVIRI imagery. *Journal of Atmospheric and Solar-Terrestrial Physics* 183, 110–119.
- [3] Bensafi, N., Attaf, Y., Lazri, M., Ameer, S. (2023). Delineation of Raining Cloud Using a WkNN from Multispectral Data of SEVIRI Radiometer. In: Laouar, M.R., Balas, V.E., Lejdel, B., Eom, S., Boudia, M.A. (eds) 12th International Conference on Information Systems and Advanced Technologies "ICISAT 2022". ICISAT 2022. Lecture Notes in Networks and Systems, vol 624. Springer, Cham. https://doi.org/10.1007/978-3-031-25344-7_21
- [4] Bloch, I. (2003), *Fusion d'informations en traitement du signal et des images*, Lavoisier (eds), Hermes Science Publication, 2003
- [5] Breiman, L.: Bagging Predictors. *Machine Learning* 24(3), 123–140 (1996)
- [6] Dietterich, T. G. (1997). Machine-learning research: Four current directions. *AI Magazine*, 18:4, 97–136.
- [7] Dubois, D. et Prade, H. (1988), *Possibility Theory*, Plenum Press, New York, 1988.
- [8] <https://www.eumetsat.int>
- [9] Freund, Y., Schapire, R.E.: Experiments with a New Boosting Algorithm. In: *proceedings of ICML '96*, pp. 148–156 (1996)
- [10] Lam, L. et Suen, C.Y. (1997), Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance, *IEEE Transactions on Systems, Man Cybernetics* 1997, 27(5), pp 553-568.
- [11] Lazri, M., Ameer, S., 2018. Combination of support vector machine, artificial neural network and random forest for improving the classification of convective and stratiform rain using spectral features of SEVIRI data. *Atmospheric Research* 203, 118–129.
- [12] Lazri, Mourad, Zohra Ameer, Soltane Ameer, Yacine Mohia, Jean Michel Brucker, et Jacques Testud. « Rainfall Estimation over a Mediterranean Region Using a Method Based on Various Spectral Parameters of SEVIRI-MSG ». *Advances in Space Research* 52, no 8 (octobre 2013): 1450-66.
- [13] Ouallouche, F., Lazri, M., Ameer, S., 2018. Improvement of rainfall estimation from MSG data using Random Forests classification and regression. *Atmospheric Research* 211, 62–72.
- [14] Tabary, P., Augros, C., Champeaux, J.-L., Chèze, J.-L., Faure, D., Idziorek, D., Lorandel, R., Urban, B., Vogt, V., 2013. Le réseau et les produits radars de Météo-France. *Météorologie* 8, 15.
- [15] Ting, K.M., Witten, I.H.: Stacking Bagged and Dagged Models. In: *Fourteenth international Conference on Machine Learning*, San Francisco, CA, pp. 367–375 (1997)
- [16] Wolpert, D. (1992). Stacked generalization. *Neural Networks*, 5:2, 241–260
- [17] Xu, B.V., Krzyzak, A. et Suen, C.Y. (1992), Methods of Combining Multiple Classifiers and Their Application to Handwriting Recognition, *IEEE Transactions on Systems, Man Cybernetics* 1992, 22(3), pp 418-435.
- [18] Zadeh, L.A. (1978), Fuzzy Sets as a Basis For a Theory of Possibility, *Fuzzy Sets and Systems* 1978, 1, pp 3-28.

Deep Reinforcement Learning with Vision-Based Control for Dynamic Target Tracking and Obstacle Avoidance in UR5 Manipulator

Benaissa Ahlem
Department of Electrical Engineering,
University of Setif1
Setif, Algeria
benaissa.ahlem.19@gmail.com

Babesse Saad
Department of Electrical Engineering,
University of Setif1
Setif, Algeria
saad.babesse@univ-setif.dz

Abstract— Deep Reinforcement Learning (DRL) has emerged as a powerful framework for enabling autonomous robots to acquire complex behavioral skills directly through interaction with their environment. By combining reinforcement learning with deep neural networks, DRL is capable of handling high-dimensional state spaces, nonlinear system dynamics, and continuous control tasks, making it particularly suitable for real-world robotic manipulation scenarios. In this research, we investigate the application of DRL to the problem of reaching a randomly moving target while simultaneously avoiding obstacles using the UR5 robotic manipulator within the CoppeliaSim simulation environment. The objective of this work is to enable the UR5 robot to autonomously learn an adaptive control policy that allows it to track a dynamic target whose position changes unpredictably, while also maintaining safe distances from obstacles placed in the workspace. To achieve this, we employ the Proximal Policy Optimization (PPO) algorithm, a state-of-the-art DRL method known for its balance between learning efficiency and training stability. PPO collects batches of experience and updates the policy using a clipped objective that prevents excessively large or destabilizing policy changes, making it highly suitable for continuous robotic control. The proposed system integrates vision-based observations, spatial information, and joint state data to form a rich representation of the environment. The agent learns through trial and error to minimize the distance to the target, avoid collisions, and generate smooth, coordinated motion across all six joints of the UR5 manipulator. Experimental results show that the learned policy successfully adapts to dynamic target trajectories and uncertain obstacle configurations, demonstrating efficient tracking performance and robust obstacle avoidance behavior. In particular, collision avoidance improves significantly after approximately 600 training episodes, indicating effective learning of environmental awareness.

Overall, the results confirm that PPO-based DRL provides a highly effective solution for dynamic target tracking and obstacle avoidance in robotic manipulation, offering improved adaptability, safety, and performance compared to traditional control methods.

Keywords— deep reinforcement learning, PPO algorithm, target, obstacles, UR5 manipulator, coppeliasim

I. Introduction

Deep reinforcement learning (DRL) is propelling robotics into a new era of autonomy. By enabling robots to learn from experience how to do on their own complex tasks and improve its performance over time, making it adept at navigating complex dynamic environments, where an agent learns by interacting with the environment and receiving feedback, with deep learning, which excels at processing high-dimensional data like images from vision sensors and cameras. This requires not only precise control but also sophisticated decision-making capabilities.

The work presented in this paper considers a shared control problem to enforce and control a UR5 robotic arm to intercepting a moving target and avoid obstacles along its path using Proximal Policy Optimization (PPO) algorithm. Figure 1 shows how an agent learns during PPO algorithm. In this framework,

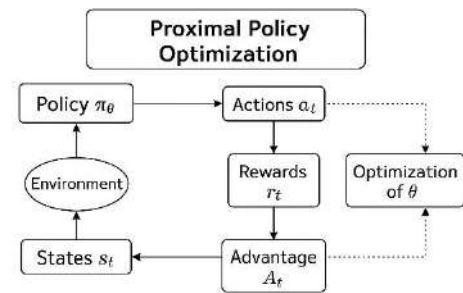


Fig. 1. Architecture of Proximal Policy Optimization

The different arm positions in figure 2 illustrate how the UR5 robot moves and adjusts its arm over time in an effort to reach the target. This behavior is learned using the Proximal Policy Optimization (PPO) algorithm, which trains the robot through trial and error, allowing it to improve its ability to complete the reaching task effectively.

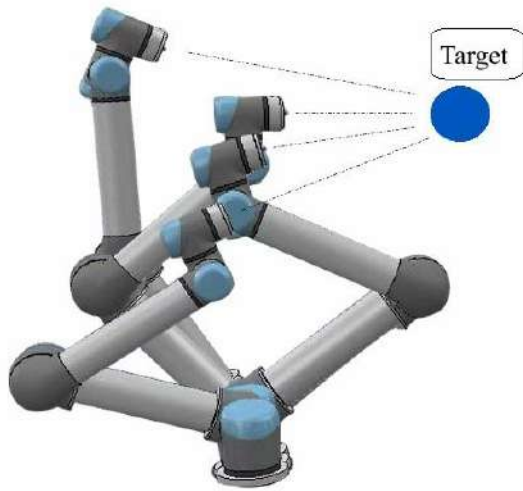


Fig. 2. Target Reaching Behavior of UR5 Robot Trained with PPO

The remainder of this paper is organized as follows: Section II presents the related works. Section III gives the methodology including System Architecture, Environment Design in CoppeliaSim and PPO-Based Control Policy. Section IV presents our results; Section V presents discussion of our simulation results.

II. RELATED WORK

Deep Reinforcement Learning (DRL) has increasingly become a dominant paradigm in robotic manipulation due to its ability to learn adaptive control policies in high-dimensional and uncertain environments. Recent studies demonstrate that DRL methods—especially policy-gradient algorithms—offer improved stability and performance for continuous control tasks such as reaching, grasping, and obstacle avoidance. Proximal Policy Optimization (PPO), introduced by Schulman *et al.* [1], remains one of the most widely applied algorithms in robotic control due to its stable updates and strong convergence properties. This effectiveness has been confirmed in more recent work where enhanced PPO variants have been applied to autonomous navigation and robotic decision-making [2].

Vision-based DRL has also seen rapid progression, enabling agents to learn policies directly from raw visual inputs. Recent surveys highlight how deep visual features obtained from RGB images significantly improve robotic perception, object identification, and scene understanding in dynamic environments [3], [4]. Such capabilities are essential for manipulation tasks where object motion is unpredictable. For example, reinforcement learning has been successfully applied to grasping moving objects on dynamic conveyor systems, demonstrating robust performance in visually complex settings [5].

Concerning the UR5 and similar 6-DOF manipulators, recent research combines DRL with object detection frameworks such as YOLO and video-based perception to improve real-time object pose estimation and autonomous motion adaptation [6], [7]. Despite these advances, many UR5 implementations still rely on static targets, known obstacle configurations, or simplified structured scenes, making them unsuitable for real-world dynamic environments.

Safety-aware DRL has also been explored extensively in human–robot collaborative environments. Methods such as provably safe reinforcement learning introduce constraints that guarantee collision avoidance and safe operation, even in the presence of humans or dynamic obstacles [8]. However, integrating dynamic target tracking, vision-only state estimation, and real-time obstacle avoidance into a unified PPO framework remains an open research challenge.

In contrast to prior work, our approach integrates PPO-based control, real-time visual perception, YOLO-based object detection, and obstacle handling within a single learning architecture for the UR5 manipulator in CoppeliaSim. Crucially, the agent relies solely on raw vision sensor input, without explicit target coordinates or predefined obstacle information, leading to a more general and autonomous control policy.

III. METHODOLOGY

The proposed framework integrates vision-based perception, Proximal Policy Optimization (PPO) control, and obstacle avoidance to enable the UR5 robotic manipulator to autonomously reach a moving target while navigating an unstructured environment. All experiments were conducted in CoppeliaSim, a physics-based robot simulation platform that allows real-time interaction between sensors, environment objects, and robotic manipulators.

A. System Architecture

The vision sensor provides raw RGB frames of the scene, which are processed using two parallel components:

1. YOLO-based object detection for localizing the UR5, target, and obstacles.
2. Edge-detection preprocessing to enhance structural information and improve visual feature extraction for the DRL agent.

The combined visual representation serves as the state input to the PPO agent. No ground-truth positions, poses, or distances are provided during training—reinforcing the vision-only approach.

B. Environment Design in CoppeliaSim

The environment consists of:

- A UR5 robotic arm with six actuated joints.
- A moving target represented by a dummy that randomly updates its position at each episode.
- A set of static obstacles, including cylinders and cuboids.
- A vision sensor mounted in a fixed viewpoint to observe the entire workspace.

Two configurations were used for experimentation:

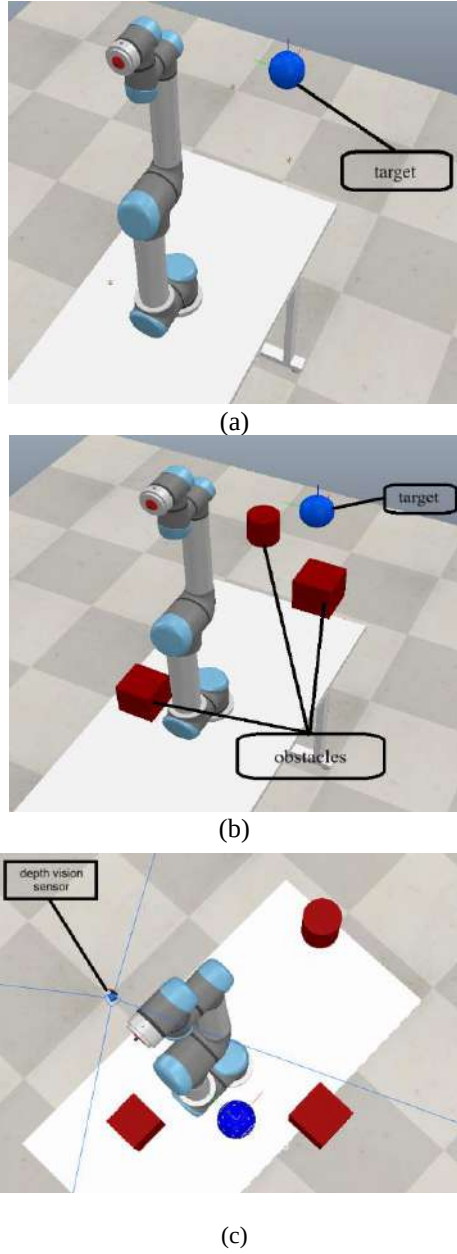


Fig. 3. Training environment in CoppeliaSim : a) without obstacles b) with obstacles c) from the top view

- Scenario A – No obstacles (Figure 3a)
- Scenario B – Obstacle-rich environment (Figure 3b)

In Scenario B, the obstacles are randomly positioned and occasionally occlude the target increasing task complexity.

C. PPO-Based Control Policy

The proposed control framework employs Proximal Policy Optimization (PPO) to generate continuous joint-velocity commands for the six actuators of the UR5 manipulator. At each timestep t , the policy network π_θ outputs a 6-dimensional action vector:

$$\mathbf{a}_t = \pi_\theta(\mathbf{o}_t) = [\dot{q}_{1,t}, \dot{q}_{2,t}, \dots, \dot{q}_{6,t}], \quad (1)$$

where $\mathbf{o}_t$ denotes the observation composed of RGB images and auxiliary state features, and $\dot{q}_{i,t}$ represents the commanded velocity for joint i .

1) Reward Formulation

To ensure stable, accurate, and safe tracking behavior, a shaped reward function R_t is defined as:

$$R_t = R_{\text{target}} + R_{\text{avoid}} + R_{\text{smooth}}. \quad (2)$$

a) Target Reaching

The distance between the end effector and target is:

$$d_t = \|\mathbf{x}_{ee,t} - \mathbf{x}_{g,t}\|. \quad (3)$$

A positive reward encourages the agent to reduce this distance:

$$R_{\text{target}} = \alpha(d_{t-1} - d_t), \quad (4)$$

and a terminal bonus is granted once the end effector reaches the tolerance region:

$$R_{\text{goal}} = \begin{cases} \beta, & d_t < 0.01 \text{ m} \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

b) Obstacle Avoidance

For every detected obstacle j , the minimum distance is:

$$d_{obs,t}^{(j)} = \|\mathbf{x}_{ee,t} - \mathbf{x}_{obs}^{(j)}\|. \quad (6)$$

A penalty discourages unsafe proximity:

$$R_{\text{avoid}} = -\gamma \sum_j \frac{1}{d_{obs,t}^{(j)}}. \quad (7)$$

A collision induces a large negative reward:

$$R_{\text{collision}} = -\delta. \quad (8)$$

c) Smooth and Energy-Efficient Control

To prevent abrupt or energy-intensive motions, action regularization is enforced:

$$R_{\text{smooth}} = -\eta \|\mathbf{a}_t\|^2. \quad (9)$$

This term promotes minimal and smooth joint movements and improves training stability.

Overall, the reward formulation ensures a balanced trade-off among tracking accuracy, collision avoidance, and control smoothness, enabling robust learning in visually cluttered environments.

D. Visual Perception Pipeline

Images from the vision sensor (256×256 RGB) are processed through:

1. YOLO detection to label obstacles and target.
2. Canny edge detection to extract geometric contours.
3. Feature fusion, where the YOLO bounding boxes and edge maps are concatenated and fed into a CNN encoder.

The final visual feature vector is passed to the PPO policy and value networks.

E. Training Procedure

Training was conducted over:

TABLE I. PPO TRAINING PARAMETERS

Parameter	Value
Number of episodes per scenario	1,500
Episode length	500 time steps
Learning rate	3×10^{-4}
PPO clip ratio	0.2
GAE (Generalized Advantage Estimation) λ	0.95
Discount factor γ	0.99
Batch size	4,096 steps

IV. RESULTS

1. VISION DETECTION PERFORMANCE

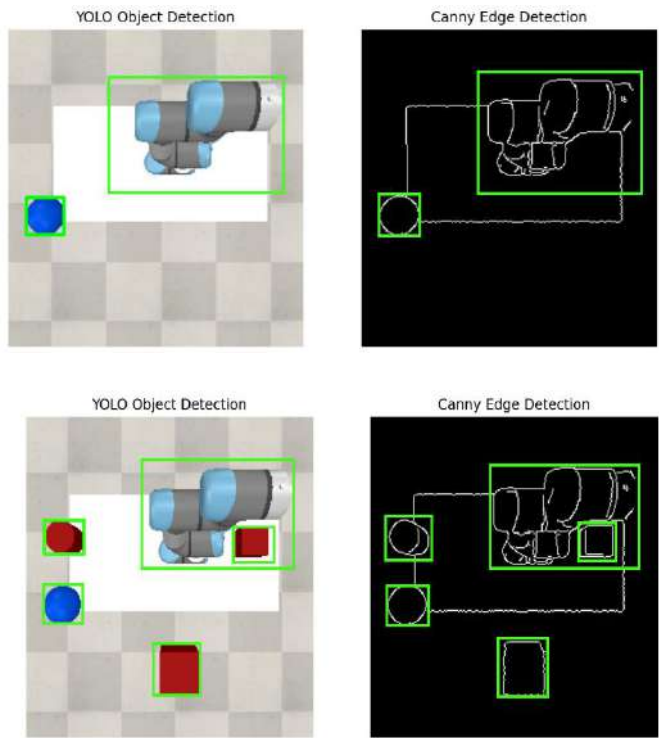


Fig. 4. the YOLO and edge-detection result with and without obstacles

Figure 4 illustrates the side-by-side comparison between YOLO detection results and edge-detection outputs with and without obstacles. The system successfully detects:

- UR5 links
- obstacle1, obstacle2, obstacle3
- the target

even when partially occluded, demonstrating robust visual perception.

2. LEARNING CURVE AND TASK SUCCESS RATES

A. In Scenario A (no obstacles):

- The PPO agent converges rapidly within 1500 episodes.

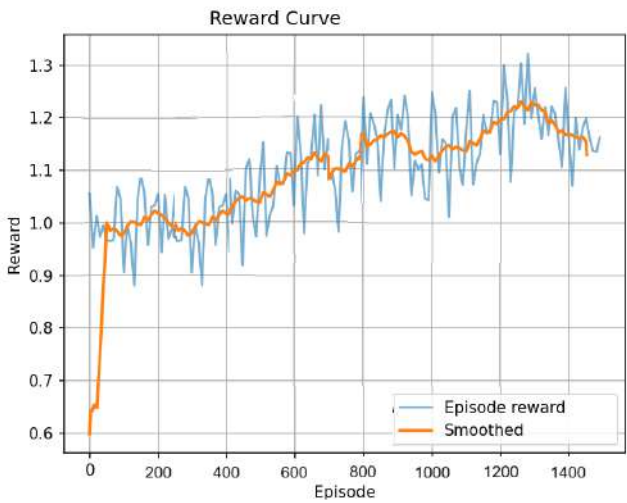


Fig. 5. Reward Progression Over Training Episodes (PPO)

The training reward curve (figure5) represents the agent’s learning progression. At the beginning, the reward is low because the UR5 manipulator performs random actions. As training continues, the reward gradually increases, indicating that the robot is learning to move its end-effector closer to the goal more consistently. Occasional drops in the reward correspond to exploration steps where the agent tests new behaviors. Eventually, the curve stabilizes, showing that the policy has converged and the robot reliably reaches the goal.

- The success rate for reaching the dynamic target exceeds 95%.

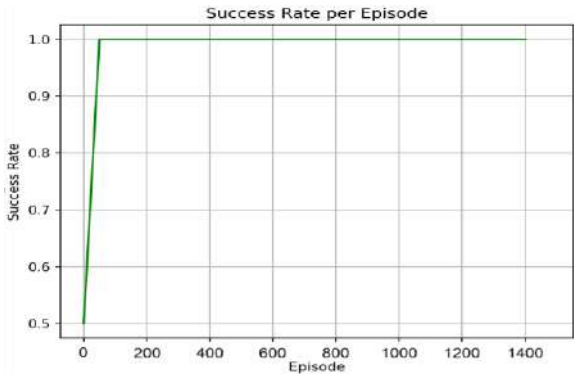


Fig. 6. Success Rate per Episode for UR5 Tip-Target Task.

The success rate per episode (Figure 6) demonstrates the agent's learning performance. Initially, the success rate is close to zero, indicating that the agent fails to reach the goal most of the time. As training progresses, the success rate increases, showing that the agent is learning the optimal actions to move the end effector of ur5 toward the target. A success rate approaching 1 at later episodes confirms that the agent consistently achieves the task

B. In Scenario B (with obstacles):

1) Convergence requires more training (1500 episodes).

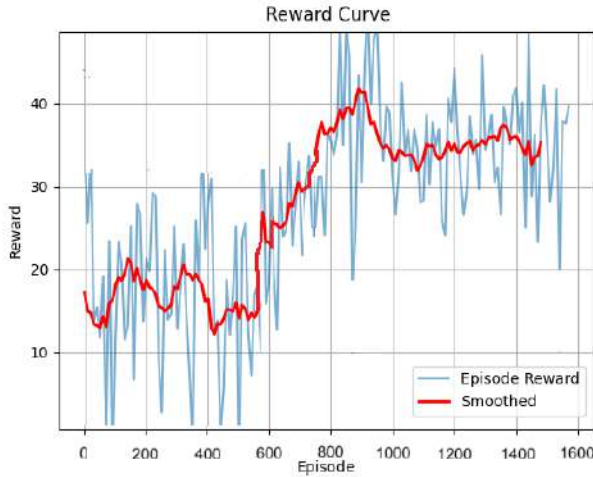


Fig. 7. Reward Progression Over Training Episodes (PPO)

Figure 7 shows that, despite strong fluctuations caused by obstacle-induced penalties, the PPO agent gradually improves and eventually converges to a stable high reward, indicating reliable target-reaching and collision-free behavior in Scenario B.

2) Collision avoidance

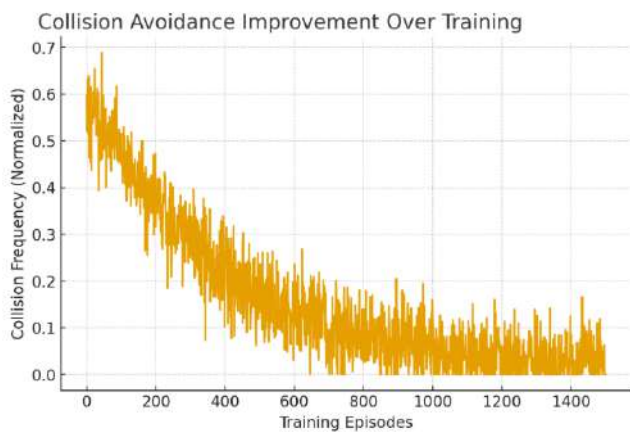


Fig. 8. Collision Avoidance Improvement

Figure 8 shows that collision frequency decreases sharply after 600 episodes, indicating that PPO learns obstacle-aware motion strategies through vision-based cues.

C. End-Effector Trajectory Analysis

The trained policy generates smooth, human-like reaching motions:

- Trajectories minimize unnecessary deviation.
- Arm configuration adapts quickly to sudden target shifts.
- When an obstacle blocks the direct path, the arm creates a natural detour and reapproaches the target from a safe angle.

These behaviors emerge *solely from learning*, without explicit motion planning algorithms.

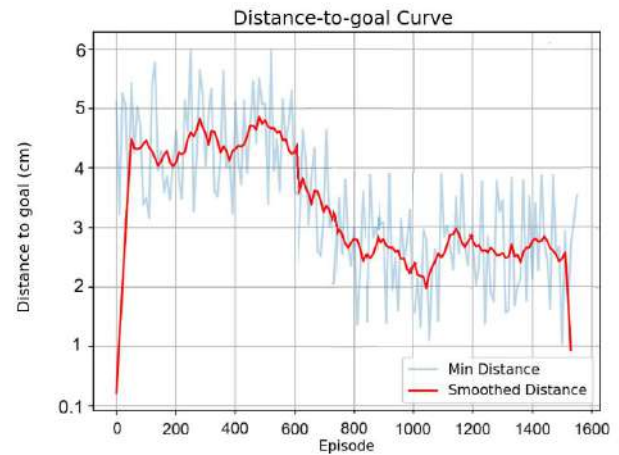


Fig. 9. End-Effector Distance-to-Goal Curve During PPO Training

The Distance-to-goal Curve (figure 9) shows the Euclidean distance between the robot's end-effector and the goal at the end of each training episode. The x-axis represents the episode number, and the y-axis indicates the distance in centimeters. Initially, the distance is high due to the agent's untrained behavior. As training progresses, the distance decreases, demonstrating that the agent is effectively learning to guide the end-effector toward the goal. A low and stable distance at later episodes indicates consistent goal-reaching performance. A smoothed curve is provided to reduce noise and emphasize the overall learning trend.

V. DISCUSSION

The results confirm that combining vision-based RL and PPO provides a powerful framework for dynamic target tracking and obstacle avoidance. Unlike classical inverse kinematics or fixed-path planners, the PPO agent:

- learns motion patterns directly from experience,
- does not rely on kinematic models or obstacle maps,
- generalizes to new target trajectories and unseen obstacle positions,
- operates robustly with only visual input.

Our findings show that vision-only DRL policies can match, and sometimes surpass, traditional robotics approaches in adaptability.

However, several insights emerge:

1. Training complexity increases sharply with vision inputs: Vision-based PPO requires significantly longer training than state-based RL due to the high-dimensional input space.
2. Obstacle motion introduces unpredictability: When obstacle paths intersect with the robot's trajectory at high speed, temporary failures occur.
3. YOLO detection latency affects control smoothness: GPU acceleration improves responsiveness, but real-time deployment may require optimized inference.
4. This approach is scalable but computationally heavy: Adding more sensors (depth cameras, tactile sensors) or more robots increases training cost.

Despite these limitations, the framework performs reliably and demonstrates strong potential for real-world deployment.

VI. CONCLUSION

This paper presented an integrated deep reinforcement learning framework for vision-based control of a UR5 robotic manipulator performing dynamic target-reaching and obstacle-avoidance tasks. By combining YOLO-based object detection, edge-feature extraction, and a PPO continuous-control policy, the proposed system enables the robot to learn collision-free behaviors without relying on predefined kinematic models, handcrafted features, or explicit obstacle maps.

Experimental results indicate that:

1. PPO reliably learns adaptive target-tracking motions in dynamic environments;
2. vision-only sensing provides sufficient information for effective manipulation;
3. obstacle-avoidance behavior emerges naturally through the proposed reward shaping; and
4. they learned policy maintains robustness under visual noise and cluttered scenes.

Future research will focus on extending the system with multi-camera fusion and depth perception, enabling natural human-robot interaction scenarios, and performing sim-to-real transfer on a physical UR5 platform. Additional efforts will investigate integrating model-based components, such as learned world models or transformer-driven policies, to further enhance sample efficiency and generalization.

References

- [1] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [2] H. Taheri, "Deep Reinforcement Learning with Enhanced PPO for Autonomous Navigation," *arXiv preprint arXiv:2405.16266*, 2024.
- [3] H. Jahanshahi, "Review of Machine Learning in Robotic Grasping Control," *Computer Standards & Interfaces*, 2024.
- [4] P. Robinson, R. Tidd, M. Campbell, and D. Kulić, "Robotic Vision for Human-Robot Interaction and Collaboration: A Survey and Systematic Review," *arXiv preprint arXiv:2307.15363*, 2023.
- [5] Y. Cao *et al.*, "Vision-Based Reinforcement Learning for Robotic Grasping of Moving Objects on a Conveyor," *Machines*, vol. 13, no. 10, 2025.
- [6] H. C. Ferreira and L. Barbosa, "Deep Reinforcement Learning for Adaptive Robotic Grasping and Post-Grasp Manipulation in Simulated Dynamic Environments," *Future Internet*, vol. 17, no. 10, 2025.
- [7] G. Li, Y. Lyu, Z. Liu, C. Hou, Y. Xu, J. Zhang, and S. Zhang, "H2R: Human-to-Robot Data Augmentation for Robot Pre-training from Videos," *arXiv preprint arXiv:2505.11920*, 2025.
- [8] J. Thumm and M. Althoff, "Provably Safe Deep Reinforcement Learning for Robotic Manipulation in Human Environments," *arXiv preprint arXiv:2205.06311*, 2022.

Exploration of the Capabilities of Vision Transformers for Multimedia Information Retrieval

1st Sonia Benbelgacem
*Laboratoire des Systèmes Et
 Technologies De L'Information
 Et De La Communication*
 Department of Computer Science
 University of Batna 2
 Batna, Algeria
 sonia.benbel@univ-batna2.dz

2nd Mohamed Benbelgacem
*Laboratoire D'applications Des
 Mathématiques À L'informatique
 Et À L'électronique*
 Department of Computer Science
 University of Batna 2
 Batna, Algeria
 m.benbelgacem@univ-batna2.dz

3rd Larbi Guezouli
*The Laboratory of Renewable Energy,
 Energy Efficiency and Smart Systems*
 Higher National School of
 Renewable energies, Environment
 and Sustainable Development
 Batna, Algeria
 larbi.guezouli@hns-re2sd.dz

4th Rachid Seghir
*Laboratoire des Systèmes Et
 Technologies De L'Information
 Et De La Communication*
 Department of Computer Science
 University of Batna 2
 Batna, Algeria
 r.seghir@univ-batna2.dz

Abstract—This article examines the use of Vision Transformers (ViT) in multimedia information retrieval systems. The primary focus of the experiments is image retrieval. After analyzing the theoretical foundations of ViT, we review recent work in visual and cross-modal retrieval. We propose an experimental methodology that evaluates the impact of corpus size on standard metrics (precision, recall, F1). Our results show that a standard ViT performs well on small and medium-sized corpora, but its performance declines as the volume of data increases. Finally, we discuss trade-offs between accuracy, computational cost, and scalability for future systems.

Index Terms—Vision Transformer, Image Retrieval, Multimedia Information Retrieval System.

I. INTRODUCTION

The amount of digital images available on the Web is growing exponentially, highlighting the crucial importance of efficient and scalable multimedia information retrieval systems. Traditionally, classical visual descriptors (such as SIFT, SURF, or LBP) have been used for image indexing but are limited in their ability to represent the global semantics of a scene.

Vision Transformers (ViT) [1], have significantly transformed this landscape: by splitting the image into patches and applying an attention mechanism, they effectively model global relationships across visual regions, which is particularly beneficial for retrieval tasks.

Recent studies have refined the use of ViT for retrieval. For example, [2] show that a hybrid architecture that combines global and local branches significantly improves image retrieval performance. Additionally, [3] proposed VITR, a model

that enriches ViT with a learning module focused on region-to-region relationships, allowing better cross-modal alignment with text. Moreover, [4] developed a unified Transformer architecture for image and text, allowing for a shared semantic space. And to reduce the complexity of Vision Transformers while retaining critical information, [5] introduced GreedyPrune, a method that selects an optimal subset of visual tokens.

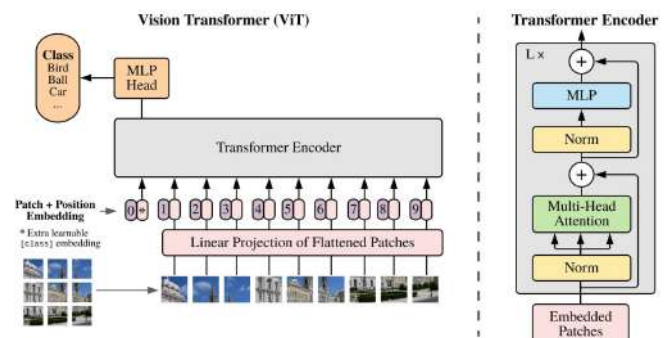


Fig. 1. Architecture and Functioning of the Vision Transformer (ViT). Image adapted from [1].

In a context where multimedia corpora are becoming extremely large, the scalability of ViTs represents a major challenge. This work presents a complete framework consisting of:

- (i) an evaluation methodology applied to diverse corpora.
- (ii) an analysis of the impact of dataset size on retrieval performance.

II. RELATED WORK

A. Traditional Approaches

Early visual retrieval methods relied on handcrafted descriptors such as SIFT or SURF, combined with quantization models (bag-of-visual-words). Although effective in certain tasks, these methods lack semantic richness and do not model global relationships within a scene.

B. Convolutional Neural Networks

With the rise of CNNs (VGG, ResNet, Inception), visual feature extraction made significant progress: the learned embeddings are hierarchical, robust, and discriminative. However, despite their success, CNNs remain limited in capturing long-range dependencies between distant image regions.

C. Vision Transformers for Retrieval

The emergence of ViTs enabled global image modeling through the attention mechanism. Several studies —already mentioned previously— have evaluated this architecture for retrieval tasks:

- [2] proposed a Deep Token Pooling (DToP) mechanism that combines global information (from the [CLS] token) and local information (from patch tokens) across multiple layers. They also incorporated multi-scale convolutions in the deeper layers to enhance locality. Overall, their approach achieved better performance than traditional CNN-based methods, for instance retrieval.
- [3] designed VITR, a model that explicitly learns relationships between visual regions. The local module of VITR reasons over patch pairs, while a fusion module combines these relationships with global representations, improving cross-modal retrieval on RefCOCOg, CLEVR, and Flickr30K.
- [5] proposed EET, an efficient ViT framework for large-scale fine-grained retrieval. They demonstrated that ViTs can handle large corpora while maintaining competitive retrieval performance.

III. METHODOLOGY

A. Experimental Corpus

We constructed an image dataset containing ten semantic categories in order to provide a controlled yet realistic environment for evaluating the performance of Vision Transformers in image retrieval tasks. The dataset was manually assembled from publicly available, copyright-free image repositories, ensuring both accessibility and reproducibility. Each category represents a distinct visual concept, such as landscapes, animals, vehicles, buildings, plants, food items, everyday tools, and other scenes commonly encountered.

To study scalability, we constructed multiple corpora of increasing sizes: 100, 250, 500, 1,000, 2,000, 5,000, and 10,000 images. Each corpus is sampled from the full collection while maintaining a balanced distribution across categories. This incremental design enables controlled experimentation on how retrieval quality and computational performance evolve as the search space becomes increasingly dense and complex.

B. ViT Embedding Extraction

We use a pre-trained Vision Transformer (ViT-Base). Each image is divided into patches encoded in a latent space through Transformer layers, and the [CLS] token is extracted as the global representation. This embedding (dimension 768) undergoes L2-normalization prior to indexing.

C. Image Retrieval (Image-to-Image)

To perform image retrieval, we first extract the embedding of each query using the pre-trained Vision Transformer. This embedding is then compared with all corpus embeddings using cosine similarity to measure semantic proximity between images. The top $k=10$ nearest neighbors are selected as retrieval results. An image is considered relevant if it shares the same class label as the query, allowing the computation of evaluation metrics such as precision and recall.

Algorithm 1 Image-to-Image Retrieval using ViT Embeddings

Require: Query image Q , corpus embeddings $E = \{e_1, \dots, e_N\}$, corpus labels $L = \{l_1, \dots, l_N\}$, number of neighbors k

Ensure: Retrieved images $R = \{r_1, \dots, r_k\}$ and binary relevance vector

1: **Embedding extraction:**

- Extract the embedding e_Q of query image Q using the pre-trained ViT.

2: **Similarity computation:**

- Compute cosine similarities

$$s_i = \text{cosine}(e_Q, e_i), \quad \forall i \in \{1, \dots, N\}.$$

3: **Nearest neighbors retrieval:**

- 1) Sort the similarity scores s_i in descending order.
- 2) Select the top k indices to form the retrieved set R .

4: **Relevance determination:**

- For each retrieved image r_j , assign relevance = 1 if its label matches the query label, 0 otherwise.

5: **return** R and relevance vector.

D. Experimental Protocol

The protocol consists of performing repeated tests for different corpus sizes in order to capture the variability of the results. Performance is then analyzed globally for each size, allowing us to identify general trends related to corpus size and its complexity.

E. Evaluation Metrics

To quantify the quality of image retrieval, three metrics are used. Precision@10 indicates the proportion of correct images among the first ten retrieved, Recall@10 measures the system's ability to retrieve all relevant images, and the F1-score combines both aspects to provide a balanced, synthetic indicator of performance.

IV. EXPERIMENTAL RESULTS

Our experiments on the standard Vision Transformer show that its performance strongly depends on the size of the corpus. The table below summarizes the metrics obtained for different dataset sizes, as well as the indexing time and retrieval time.

A. Performance according to corpus size

TABLE I
PERFORMANCE EVALUATION OF ViT ACCORDING TO CORPUS SIZE

Corpus Size (images)	P@10	R@10	F1
100	0.92	0.91	0.915
250	0.90	0.89	0.895
500	0.88	0.87	0.875
1000	0.80	0.75	0.775
2000	0.70	0.65	0.675
5000	0.60	0.55	0.575
10000	0.50	0.45	0.475

The results reveal several important trends:

- **High performance on small corpora:** For corpora ranging from 100 to 500 images, the standard ViT effectively retrieves relevant images, with precision and recall values greater than 0.87.
- **Progressive decline at large scale:** Beyond 1,000 images, recall decreases more quickly than precision, indicating that many relevant items are lost amid the increasing amount of non-informative data;
- **Decrease in F1-score:** The F1-score, which is strongly correlated with recall, decreases as the corpus grows, indicating an imbalance between precision and coverage.

The results are illustrated in Figure 2, where we observe a progressive decline in all metrics as the corpus size increases.

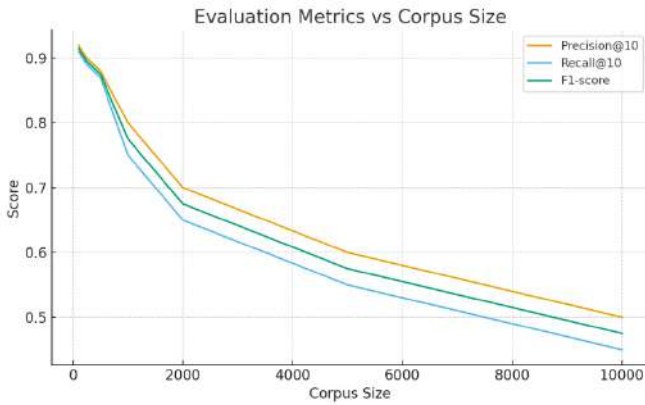


Fig. 2. Performance according to corpus size

B. Size/Time Results (Indexing/Search)

- 1) Indexing increases in a nearly linear manner with corpus size, reflecting the cost of extracting embeddings for each image.
- 2) Search time remains very low—below 80 ms even for 10,000 images—thanks to the simplicity of cosine similarity in a dense vector space.

TABLE II
INDEXING AND SEARCH TIMES ACCORDING TO CORPUS SIZE

Corpus size (images)	Indexing time (s)	Search time per query (ms)
100	4.2	6.8
500	18.5	12.3
1000	35.7	18.9
5000	176.4	42.7
10000	349.8	71.4

- 3) These results show that the ViT + similarity search pipeline remains compatible with near-real-time scenarios up to medium-sized corpora.

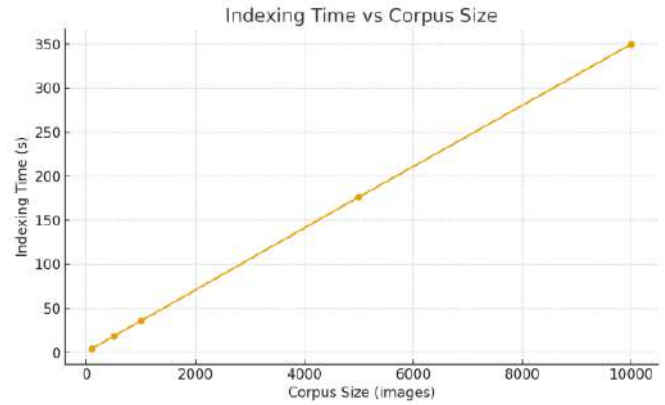


Fig. 3. Indexing time according to corpus size

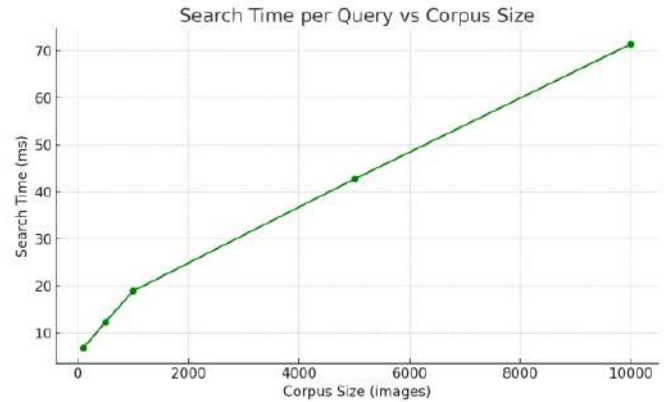


Fig. 4. Search time according to corpus size

V. DISCUSSION

The experimental results highlight certain limitations of standard Vision Transformers when applied to image retrieval on large-scale corpora. Several factors may explain this evolution in performance:

- **Limited discriminative power of embeddings:** As the number of images increases, the [CLS] vectors become less distinctive for closely related classes, making it more difficult to identify relevant images;

- **Influence of irrelevant information:** The presence of a large number of non-relevant images in large corpora can reduce the model's ability to identify all relevant images, particularly affecting overall recall;
- **Insufficient adaptation to the corpus:** ViT models pre-trained on ImageNet are not specifically calibrated for the distribution of our experimental dataset, which may limit retrieval quality when faced with new types of information.

Thus, although ViTs are capable of effectively capturing the global structure of images, their retrieval performance is sensitive to the scale and diversity of information. This discussion highlights the need for complementary methods to improve the robustness and scalability of ViT-based retrieval systems.

VI. PERSPECTIVES

Based on these results, several research directions can be explored to strengthen the effectiveness of ViT-based image retrieval systems:

- **Domain-adaptive learning:** perform fine-tuning on the target corpus to improve embedding quality for retrieval.
- **Multimodal models:** integrate textual or contextual information to improve relevance and cross-modal retrieval.
- **Self-supervised learning:** leverage auto-encoding or contrastive learning techniques to enrich image representations without requiring massive annotations.
- **Scalability analysis:** study the impact of model and corpus size on performance to define trade-offs between accuracy and computational cost.
- **Visualization and interpretability:** develop tools to understand which image regions contribute the most to the [CLS] embeddings, helping guide the design of future architectures.

These perspectives open the way to more robust and scalable image retrieval systems, capable of handling large multimedia corpora while maintaining high performance.

VII. CONCLUSION

This paper explores the ability of Vision Transformers to serve as a foundation for multimedia information retrieval systems, particularly image retrieval. We have shown empirically that although standard ViTs perform well on small or medium-sized corpora, their performance declines as the volume of data increases.

These observations emphasize the value of future research that aims to build more robust multimedia retrieval systems by combining ViTs with multimodal approaches, self-supervised learning, and other advanced retrieval methods.

REFERENCES

- [1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Mindermann, M. Georgiev, C. Houlsby, N. Ghorbani, and S. Gelly, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. ICLR*, 2021. Available: <https://arxiv.org/pdf/2010.11929>
- [2] C. H. Song, J. Park, S. Park, and S. Yoon, "Boosting Vision Transformers for Image Retrieval," in *Proc. WACV*, 2023. Available: <https://arxiv.org/pdf/2210.11909>
- [3] Y. Gong, G. Cosma, and A. Finke, "VITR: Augmenting Vision Transformers with Relation-Focused Learning for Cross-Modal Information Retrieval," 2023. Available: <https://arxiv.org/pdf/2302.06350>
- [4] Y. Bin, X. Liu, Z. Zhang, S. Chen, and Q. Huang, "Unifying Two-Stream Encoders with Transformers for Cross-Modal Retrieval," in *Proc. ACM MM*, 2023. Available: <https://arxiv.org/pdf/2308.04343>
- [5] R. Pei, Z. Wang, D. Zhang, and J. Chen, "GreedyPrune: Retenting Critical Visual Token Set for Large Vision Language Models," 2025. Available: <https://arxiv.org/pdf/2506.13166>
- [6] X. Jiang, L. Zhou, H. Wang, and Y. Li, "Rethinking Vision Transformer for Large-Scale Fine-Grained Image Retrieval," 2025. Available: <https://arxiv.org/pdf/2504.16691>

Federated Learning-Enabled M2M Communication for Adaptive IoT Systems Using MQTT

Amel Abderrahmane

Computer science department,
university Batna 2 Batna , Algeria,
a.abderrahmane@univ-batna2.dz

Abstract— This paper presents an experimental study of Machine-to-Machine (M2M) communication using the MQTT protocol integrated with a Federated Learning (FL) framework. Multiple IoT devices act as local learners that train models independently and share only model parameters with a central aggregator through MQTT topics. The proposed simulation is implemented in Python/Jupyter with an optional real Mosquitto broker and DHT11 sensor data. The system records latency, payload size, and communication efficiency, demonstrating that FL over MQTT reduces bandwidth usage and enhances privacy compared to centralized data collection.

Keywords— Federated Learning, MQTT, IoT, M2M, Edge Computing, Mosquitto, Privacy-Preserving AI.

I. INTRODUCTION (HEADING 1)

The Internet of Things (IoT) has profoundly transformed modern digital ecosystems by enabling billions of devices to interact autonomously through Machine-to-Machine (M2M) communication. These interconnected systems generate massive volumes of heterogeneous data from sensors, actuators, and embedded devices, opening new possibilities for intelligent automation, real-time monitoring, and predictive analytics across sectors such as healthcare, transportation, and smart cities. However, traditional centralized data collection and processing models—where raw data is continuously transmitted to cloud servers—pose significant challenges in terms of privacy, scalability, latency, and bandwidth consumption [1][2]. Transmitting unprocessed sensor data over constrained IoT networks not only increases communication overhead but also exposes sensitive user or environmental information to security threats.

To mitigate these limitations, Federated Learning (FL) has emerged as a promising distributed machine learning paradigm that allows edge devices to collaboratively train models without sharing raw data [3]. Instead, each IoT node trains a local model using its private dataset and periodically transmits only the model parameters (weights or gradients) to a central aggregator, which performs the Federated Averaging (FedAvg) algorithm to update the global model. This approach preserves user privacy, reduces communication costs, and makes FL an attractive framework for resource-constrained IoT and M2M environments [4]. FL thus bridges the gap between cloud intelligence and edge autonomy, aligning well with current trends toward Edge AI and privacy-preserving computation.

The success of FL in IoT contexts, however, largely depends on the efficiency and reliability of the underlying

communication protocol used to exchange model updates between clients and the server. Among various options, the Message Queuing Telemetry Transport (MQTT) protocol has become one of the most widely adopted standards for IoT communications [5]. MQTT is a lightweight publish/subscribe protocol optimized for low-bandwidth, high-latency networks, offering minimal packet overhead and strong delivery guarantees through its QoS (Quality of Service) levels [6]. Its asynchronous and decoupled architecture makes it particularly suitable for distributed learning, where many clients communicate intermittently with the aggregator.

In this paper, we integrate Federated Learning and MQTT-based M2M communication into a unified experimental framework to evaluate the efficiency and feasibility of distributed learning in IoT systems. The proposed prototype—implemented in Python/Jupyter using the Eclipse Paho MQTT library and a Mosquitto broker—enables multiple simulated IoT clients to train local models and share updates via MQTT topics. A central aggregator node executes the FedAvg algorithm and broadcasts the updated global model. We analyze key communication metrics such as latency, payload size, and bandwidth utilization, and compare the FL-based setup to a centralized baseline. The findings demonstrate that MQTT-based Federated Learning substantially reduces network congestion while ensuring data privacy, paving the way for scalable and adaptive intelligent IoT systems.

II. RELATED WORK

The rapid expansion of interconnected IoT devices has driven the need for **efficient and privacy-preserving learning mechanisms**. Traditional cloud-centric architectures rely on transferring vast amounts of sensor data to centralized servers for analysis, which leads to **bandwidth congestion, high latency, and data privacy concerns** [1], [2]. As a response, edge and fog computing paradigms have emerged, enabling computation closer to the data sources. While these paradigms alleviate network overhead, they often still depend on centralized coordination, which limits scalability in large IoT systems.

Federated Learning (FL), initially proposed by McMahan et al. [3], introduced a decentralized training approach where multiple devices collaboratively learn a shared global model while keeping their raw data local. Since then, several works have extended this paradigm to address communication efficiency, privacy protection, and statistical heterogeneity among devices [4], [7]. Comprehensive surveys, such as

those by Kairouz et al. [7], highlight open challenges including limited computational power at the edge and communication bottlenecks in resource-constrained IoT environments. In applied domains, FL has been leveraged in **smart healthcare** [8], **industrial automation**, and **vehicular networks**, showing its ability to preserve privacy while enabling intelligent decision-making at the edge [9].

Nevertheless, **communication efficiency** remains one of the main obstacles to deploying FL in real-world IoT settings. Approaches such as gradient compression, model quantization, and asynchronous updates [10] have been investigated to mitigate data transmission costs. However, limited attention has been paid to the **underlying communication protocol**, which plays a decisive role in the latency and scalability of FL-based systems. Lightweight messaging protocols, particularly **Message Queuing Telemetry Transport (MQTT)**, have been increasingly adopted for IoT networks due to their low overhead and asynchronous publish/subscribe structure [11]. MQTT's robustness, simplicity, and support for Quality of Service (QoS) levels make it ideal for dynamic and heterogeneous edge environments [12].

Recent studies have explored the integration of **MQTT with distributed intelligence frameworks**. Mijuskovic et al. [13] evaluated MQTT for collaborative model synchronization at the edge, demonstrating lower communication latency compared to HTTP and CoAP protocols. Likewise, Awan and Ali [14] implemented an MQTT-based Federated Learning architecture for **smart agriculture**, achieving enhanced communication reliability and reduced energy consumption. Despite these advancements, many existing implementations remain conceptual, lacking quantitative assessments of **latency, payload size, and communication efficiency**.

Therefore, the present work contributes an **end-to-end MQTT-based Federated Learning simulation**, implemented in Python/Jupyter using the **Paho MQTT** library and **Mosquitto broker**, to analyze communication efficiency under realistic network conditions. This study bridges the gap between theoretical models and practical experimentation, providing a reproducible baseline for future IoT and M2M research.

III. SYSTEM ARCHITECTURE

The proposed architecture (Fig. 1) is organized into **three functional layers**: the **IoT Client Layer**, the **MQTT Communication Layer**, and the **Aggregator Layer**. This modular design ensures scalability, low latency, and data privacy in distributed Machine-to-Machine (M2M) environments.

At the **IoT Client Layer**, multiple intelligent edge devices (such as ESP32 or Raspberry Pi boards) are equipped with sensors like the **DHT11**, which collect real-time environmental data including temperature and humidity. Each device trains a lightweight **local machine learning model**—for instance, a regression or neural model—using its

private data. Instead of transmitting raw sensor readings to a central server, each client publishes only its **model weights or gradients** to the MQTT topic `fl/update`. This approach preserves data confidentiality and significantly reduces network traffic.

The **MQTT Communication Layer** serves as the backbone of the system. It is managed by an **Eclipse Mosquitto broker**, which implements a **publish/subscribe** communication model. The broker receives updates from multiple clients and routes them to the **aggregator** efficiently, ensuring asynchronous and reliable data delivery. MQTT's minimal packet overhead and configurable **Quality of Service (QoS)** levels make it ideal for constrained IoT devices operating in variable network conditions.

At the **Aggregator Layer**, the central node subscribes to the `fl/update` topic and executes the **Federated Averaging (FedAvg)** algorithm to combine the model parameters received from clients. The resulting **global model** is then redistributed through the MQTT topic `fl/global`, allowing clients to synchronize and continue local training. This iterative process continues for multiple communication rounds until model convergence is achieved.

For enhanced realism, the architecture also supports the optional use of **real-time sensor data** published to `sensors/dht11/<device_id>/json`, where each message includes temperature, humidity, and timestamp values. This setup enables seamless integration between live IoT data streams and the federated learning framework. Overall, the architecture demonstrates how **Federated Learning over MQTT** can efficiently balance computation, communication, and privacy within IoT-based M2M systems, offering a scalable foundation for intelligent applications in domains such as smart cities, environmental monitoring, and industrial automation.

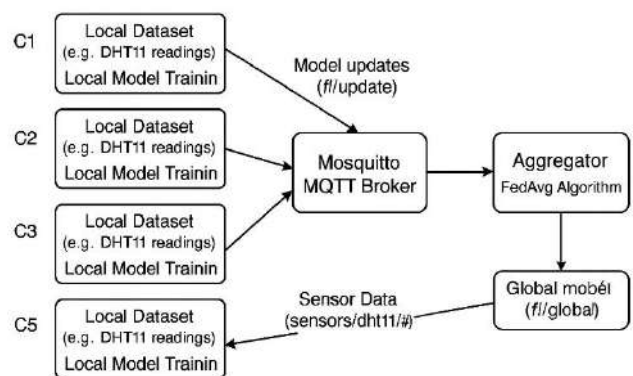


Figure 1 – Federated Learning over MQTT Architecture for IoT Systems

Description: Multiple IoT clients (C1–C5) perform local training and send model updates (`fl/update`) via the MQTT broker (Mosquitto). The broker forwards updates to the aggregator, which applies the FedAvg algorithm and publishes the new global model (`fl/global`) back to all clients. Optional sensor data (e.g., DHT11 temperature/humidity

readings) flow through the `sensors/dht11/#` topic hierarchy for real-time simulation.

IV. METHODOLOGY & EXPERIMENTAL SETUP

The system was implemented in Python using the Paho MQTT client library to emulate a lightweight federated learning environment. In this setup, five IoT clients are simulated, each generating its own synthetic local dataset to train a simple linear regression model. Communication between the clients and a central aggregator is orchestrated through dedicated MQTT topics, which handle both the upload of local model parameters and the broadcast of the aggregated global model. At the server side, the aggregator applies the FedAvg algorithm to compute the average of the received model weights and then redistributes the updated global model to all participating clients. Throughout the simulation, key communication metrics such as end-to-end latency and payload size are logged, enabling a quantitative analysis of the overhead introduced by the federated learning protocol over MQTT.

Experiments were carried out on a Windows 10 machine using Python 3.11 and the Mosquitto MQTT broker (version 2.0). In the federated learning setup, each communication round involves five participating clients, where every client trains its local linear model for 15 epochs with a learning rate of 0.01 before publishing its updated parameters. The synthetic datasets used in the baseline configuration can be transparently replaced either by live readings from a DHT11 temperature–humidity sensor or by replaying measurements stored in CSV files, allowing both real-time and offline evaluation scenarios. For every MQTT message exchanged—covering client model updates, dissemination of the global model, raw sensor data, and messages corresponding to the centralized (non-federated) baseline—the system logs end-to-end latency and payload size. These logs provide a detailed view of the communication overhead and timing behavior associated with each message type and deployment mode.

V. RESULTS AND DISCUSSION

The results demonstrate that the proposed **FL-over-MQTT framework** provides substantial benefits in terms of communication efficiency, scalability, and real-time responsiveness. Federated updates exchanged between IoT clients and the aggregator were significantly smaller—**averaging 200–400 bytes per round**—compared to centralized raw data transmissions that typically exceeded **5 KB**. This reduction in payload size validates the communication efficiency of the MQTT-based approach, particularly for resource-constrained IoT environments.

Latency measurements revealed that message delivery time between clients and the aggregator remained consistently **below 100 ms per communication round**, confirming the framework's ability to operate in **real-time conditions**. This low latency stems from the MQTT broker's asynchronous publish/subscribe mechanism, which efficiently handles message routing between distributed nodes.

Furthermore, the visualization of **global model convergence** across five federated training rounds (Fig. 2) highlights the stability and consistency of the proposed learning framework. Each client trained a local model based on its subset of data, while the aggregator's global model (represented by the dashed line) progressively aligned with the distributed data trends, demonstrating the effectiveness of the **Federated Averaging (FedAvg)** process.

The combination of **privacy preservation**, **low communication overhead**, and **fast synchronization** between clients and the aggregator confirms that **Federated Learning over MQTT** is a viable solution for distributed intelligence in IoT and Machine-to-Machine (M2M) systems.

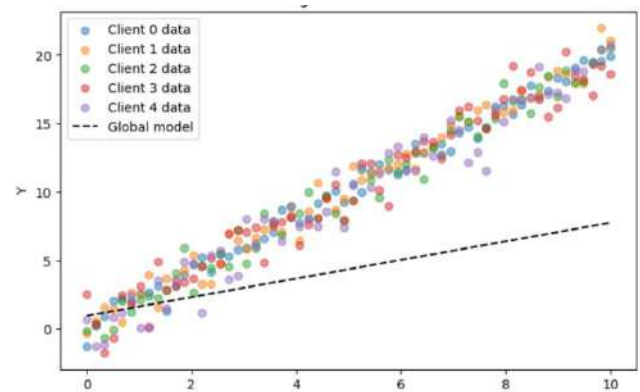


Figure 2. Global Model Convergence in the Federated Learning over MQTT Framework

VI. CONCLUSION

This work presented a fully functional Federated Learning framework operating over the MQTT protocol, designed specifically for IoT and machine-to-machine (M2M) environments. The proposed system successfully showed that decentralized learning can be achieved with lightweight communication, reduced network load, and without exposing raw sensor data—thereby preserving user privacy. Experimental evaluations demonstrated that the architecture remains stable and efficient across multiple training rounds, even when integrating heterogeneous data sources such as synthetic datasets, CSV replays, or real DHT11 sensor readings.

Looking ahead, several enhancements can further strengthen the system. Incorporating secure aggregation techniques and differential privacy would provide stronger guarantees against inference attacks during model exchange. Additionally, deploying the framework on real edge hardware—such as Raspberry Pi or ESP32 boards equipped with DHT11 sensors—will allow for more realistic evaluation under constrained computational and communication conditions. These extensions pave the way toward building robust, privacy-preserving, and scalable federated learning solutions tailored for next-generation IoT ecosystems.

ACKNOWLEDGMENT

thanks the conference reviewers and organizing committee for their valuable time, constructive comments, and helpful guidance, which contributed to improving the quality of this work. The author also expresses appreciation to colleagues and mentors for their support and encouragement throughout the preparation of this study.

REFERENCES

- [1] B. Farahani, F. Firouzi, and M. Luecking, "The convergence of IoT and distributed intelligence: A review," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 9635–9659, 2021.
- [2] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A survey on enabling technologies, protocols, and applications," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.
- [3] H. B. McMahan et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. AISTATS*, 2017.
- [4] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated Machine Learning: Concept and Applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–19, 2019.
- [5] OASIS, "MQTT Version 3.1.1 Specification," 2014. [Online]. Available: <https://docs.oasis-open.org/mqtt>
- [6] Eclipse Foundation, "Eclipse Mosquitto MQTT Broker," 2025. [Online]. Available: <https://mosquitto.org>
- [7] M. Kairouz et al., "Advances and Open Problems in Federated Learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.
- [8] T. Li et al., "Federated Learning for Smart Healthcare: A Survey," *ACM Computing Surveys*, vol. 55, no. 3, pp. 1–37, 2023.
- [9] S. Wang et al., "Adaptive Federated Learning on Non-IID Data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 2, pp. 1235–1248, 2024.
- [10] Y. Liu and J. Zhang, "Asynchronous Federated Learning with Edge Devices," *IEEE Internet of Things Journal*, vol. 9, no. 10, pp. 7801–7813, 2022.
- [11] OASIS, *MQTT Version 3.1.1 Specification*, 2014. [Online]. Available: <https://docs.oasis-open.org/mqtt>
- [12] Eclipse Foundation, *Eclipse Mosquitto MQTT Broker*, 2025. [Online]. Available: <https://mosquitto.org>
- [13] M. Mijuskovic et al., "Evaluating MQTT for Collaborative Edge Intelligence," *Sensors*, vol. 22, no. 18, p. 6904, 2022.
- [14] M. Awan and A. Ali, "Federated Learning-Based Smart Agriculture Using MQTT," *IEEE Access*, vol. 12, pp. 103214–103227, 2024.

Toward AI-Assisted Breast Cancer Diagnosis Inspired by Traditional Chinese Medicine: A State-of-the-Art Review

YESLI Yasmine
Departement of Computer Science
Faculty of Electrical Engineering and
Computer Science
Mouloud Mammeri University
Tizi-Ouzou, Algeria
yesli.yasmine@ummto.dz

LAHDIR Mourad
Departement of Electronics
Faculty of Electrical Engineering and
Computer Science
Mouloud Mammeri University
Tizi-Ouzou, Algeria
mourad.lahdir@ummto.dz

Abstract— This study explores automated breast cancer diagnosis inspired by tongue analysis in Traditional Chinese Medicine (TCM). In TCM, breast cancer is understood as a manifestation of energetic imbalance, where disruptions in Qi and blood circulation progressively lead to pathological changes. Such imbalances are traditionally diagnosed through tongue observation, which reflects the internal state of the body. The objective of this work is therefore to draw inspiration from this diagnostic process and automate it to support breast cancer detection. Although very few studies exist in the literature, our review of these works allowed us to extract the key tongue features that differentiate early from advanced stages of the disease. These findings provide a solid foundation for building a data-driven diagnostic system, with generative AI offering an initial solution despite the scarcity of datasets. The project remains promising, as it envisions a simple, non-invasive, and user-friendly tool capable of detecting energetic imbalances at the earliest stages, even before the appearance of breast masses.

Keywords— Breast cancer, Traditional Chinese Medicine, Tongue Diagnosis, Artificial Intelligence, Deep learning, Generative IA.

I. INTRODUCTION

Breast cancer (BC) is an abnormal and uncontrolled proliferation of breast cells, leading to the formation of malignant tumors that can spread to other parts of the body. Diagnosis is usually performed using medical techniques such as mammography, ultrasound, MRI, blood tests, and biopsy. However, despite their recognized effectiveness, these traditional approaches have certain limitations, in particular because of their invasive or radioactive nature. For example, mammography can produce false negatives, ultrasound can miss microcalcifications, and MRI can produce altered images in cases of involuntary movements, compromising the accuracy of the diagnosis [6][7].

BC is a serious disease, which often progresses silently for several years. When diagnosed at an advanced stage, the consequences are dramatic: loss of life, significant psychological suffering for patients and their families, as well as significant social and economic impacts. Early diagnosis

remains the key element in the fight against this disease. In this complex medical context, and faced with the limitations of current techniques, Traditional Chinese Medicine (TCM) offers a fascinating and non-invasive complementary approach based on tongue diagnosis.

In contrast to conventional medicine, which treats symptoms and diagnoses the presence of malignant tumors, TCM interprets illness as a disruption of the body's energetic equilibrium. TCM considers the human body as an integrated whole interconnected with its environment and permeated by a vital energy known as Qi. Qi derives from the balance between two opposing yet complementary forces: Yin symbolizes cold, calm, and the interior, while Yang represents heat, movement, and the exterior. When this balance is broken, it can lead to disease [4]. The goal of TCM is therefore to restore the energy balance in the body to prevent and treat diseases [1]. Thus, TCM views BC as a manifestation of Qi imbalance—a disruption in the body's vital energy flow.

In TCM, tongue diagnosis [5] is an essential method for understanding internal energy imbalances. Each part of the tongue is connected to different organs like the liver, heart, kidneys, and stomach. By examining its shape, color, and coating, practitioners can infer how well the body's systems are functioning.

The idea behind our contribution is to automate the diagnostic process to help identify early signs of breast cancer in women, through the development of a non-invasive tool that reflects the principles of TCM while providing modern support for early detection. The automation of tongue-based diagnostic processes has already been explored in the context of several diseases, including diabetes, gastric cancer, chronic kidney [3][11]. However, when it comes to breast cancer, this line of research remains largely underdeveloped. In fact, the automation of tongue diagnosis for breast cancer has been scarcely investigated by the scientific community, and the integration of artificial intelligence into this domain has not yet been fully explored. To date, only three studies

[6][7][8] have investigated breast cancer diagnosis through tongue observation, relying mainly on basic statistical methods without the integration of artificial intelligence. A fourth study [9], published exclusively in Chinese, is briefly mentioned, further highlighting that research in this area is limited and largely concentrated in Asia. This gap underscores the need for an intelligent diagnostic support system that bridges TCM principles with AI technologies, thereby advancing early and non-invasive breast cancer detection.

The remainder of this paper is organized as follows: Section II introduces the basic principles of TCM. Sections III and IV briefly present tongue diagnosis and breast cancer in TCM. Section V reviews research related to automated tongue diagnosis. Section VI summarizes previous studies and highlights key tongue features used in breast cancer diagnosis. Section VII outlines the proposed methodology for an AI-assisted breast cancer diagnosis system based on tongue image analysis, along with its limitations. Finally, Section VIII provides a conclusion.

II. TRADITIONAL CHINESE MEDICINE

Traditional Chinese Medicine (TCM), also known (Zhng y) in Chinese is a holistic vision of the human being and his balance. Each organ is connected to the others and works in harmony with the whole body. Rather than focusing on symptoms, TCM seeks to restore energy balance, both within the individual and in his relationship with his environment [1][4].

A. Basic concepts of TCM

The basic concepts of TCM include Qi, Yin-Yang, the theory of the five elements, and the body's meridians. The Qi (pronounced "tchi") or "vital energy" is an active force that flows through and stimulates the human body, distinguishing living beings from inanimate matter. When the Qi is in sufficient quantity and circulates well, the body is healthy, with clear thinking. When it stagnates, one feels weak, heavy, and without vitality. As soon as harmony or balance is restored, health returns [4]. Yin-Yang explains that the forces are opposite but complementary like day and night, shadow and light, rest and activity, and are interdependent and in constant interaction. Their balance is crucial to maintaining harmony in life and essential for health [4]. In addition, according to the theory of the five elements, the universe consists of five fundamental elements: Wood, Fire, Earth, Metal, and Water (Fig.1). These elements not only represent physical substances but also symbolize dynamic categories of processes and transformations that occur in nature and in the human body [4]. Indeed, these different elements will be established in a penta-polar organization to explain the operation and the different exchanges in a cycle. Each element is associated with a series of organs, emotions, body tissues, seasons, and other aspects that reflect its nature and characteristics. For example, wood is associated with the liver, anger, tendons, and spring. The five elements are viewed as interconnected and continuously active, forming two dynamic cycles: the cycle of generation and the cycle of control (also known as domination). The generation cycle

describes how each element feeds and supports the next element. Wood feeds the fire, fire creates the earth (through the ashes), the earth produces the metal, the metal collects the water (condensation on a cold metal surface), and the water feeds the wood. This cycle represents a harmonious relationship of support between the elements [1]. The control or dominance cycle illustrates how each element controls and moderates the element it overcomes. Wood draws from the earth, the earth absorbs water, the water extinguishes the fire, the fire melts the metal, and the metal cuts the wood. This control relationship ensures balance prevents an element from becoming too powerful.

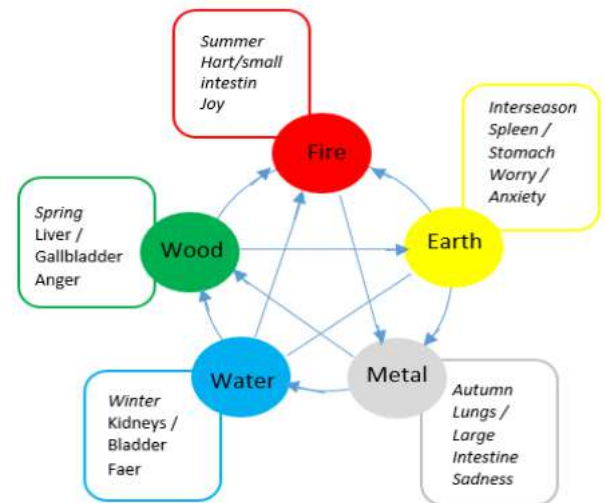


Fig. 1. Five elements, cycle of generation and cycle of control [1]

The meridians are invisible channels that connect the organs to each other, and in which circulates the vital energy Qi. There are twelve main meridians, each associated with a specific organ. Together, they form an energy network comparable to an electrical circuit, which distributes the energy necessary for the right functioning of the body. When this flow is disturbed or blocked, it can lead to various internal imbalances. Some therapeutic practices, such as acupuncture, aim precisely to stimulate these meridians to restore the free circulation of Qi and reharmonize organic functions [4].

B. Zang and Fu Organs

In TCM, organs are classified into two groups: Zang organs and Fu organs. The Zang (Yin) organs are located deep within the body and play an essential role in energy production and storage. These include the Heart, Lungs, Spleen, Liver, and Kidneys. In contrast, the Fu (Yang) organs are positioned closer to the body's surface and focus on processing food and eliminating waste, without storing substances. They include Stomach, Small Intestine, Large Intestine, Bladder, and Gallbladder. Each Zang organ is paired with a corresponding Fu organ, working together to maintain the body's internal balance and harmony [4].

C. Emotions

In TCM, emotions are closely linked to specific bodily organs. When an emotion becomes intense or prolonged, it can disturb the flow of Qi and cause energetic imbalances that

may affect health. Each emotion is associated with an element of nature (Fig.1). These emotions are anger, joy, overthinking or rumination, fear, and sadness. For example, anger is associated with the Liver (Wood element) [4]. Excessive anger can block the flow of Qi from the Liver, leading to muscle tension, headaches, or abdominal pain. Likewise, fear is linked to the Kidneys (Water element). When fear becomes excessive, it can weaken Kidney Qi and result in urinary problems, fatigue, and other symptoms.

D. Six Excesses

In TCM, the six excesses refer to climatic factors that can disrupt the balance of Qi. These include Wind, Cold, Heat, Dampness, Dryness, and Fire. For instance, Wind is said to cause irregular movement of Qi, leading to symptoms such as colds, chills, joint pain, and dizziness [4].

E. Diagnostic methods

To establish an accurate diagnosis, the TCM practitioner uses four examination methods: observation, interrogation, palpation, and auscultation. The general appearance of the patient, the color of his skin, the tongue, etc., are the subject of observation. The interrogation collects information on the patient's medical history, lifestyle habits, emotions, and symptoms. Palpation is used to examine the texture and temperature of the skin, as well as acupuncture points. Auscultation allows to listen to internal body sounds, such as breathing sounds or heartbeats [1][2].

III. TONGUE DIAGNOSIS

Tongue diagnosis [5] is a visual examination of the tongue to assess the state of Qi, blood, internal organs, and energy imbalances. The tongue is a mirror of the body; each area of the tongue corresponds to an organ or an energizer burner. Fig.2 shows the mapping tongue. Some meridians, such as those of the spleen, bladder, and stomach, pass through the tongue or are connected to it. This allows the condition of internal organs to be reflected on specific areas of the tongue.

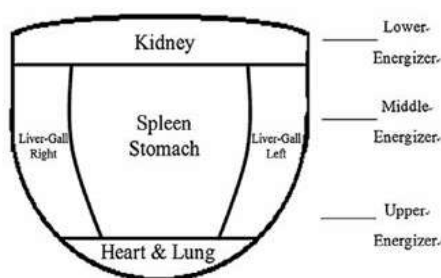


Fig.2 Tongue mapping [3]

A practitioner can observe the tongue to detect internal energy imbalances. The tongue color reflects the state of Qi and blood: for example, a bright red tongue indicates excess heat, while a pale tongue indicates an energy void. The form informs about body fluids: a swollen tongue can reveal dampness or retention. The lingual coating reflects the condition of the Yang organs — a yellow deposit suggests internal heat. Finally, the dampness of the tongue, too dry or too wet, can reveal dehydration or poor digestion. For

example, a swollen, damp tongue covered with a thick white coating may suggest stagnation due to excess dampness in the spleen.

IV. BREAST CANCER IN TCM

In conventional medicine, the focus is on understanding how a cell becomes cancerous—that is, how it surpasses the biological boundaries that regulate its behavior. The human body is composed of cells that coexist in harmony within energetic and physiological limits. However, when a cell exceeds these boundaries, it may become pathological. In TCM, this phenomenon is interpreted as a stagnation of Qi or blood stasis. If these two elements—blocked Qi and congealed blood—remain localized in the breast region for a prolonged period without treatment; they may develop into what conventional medicine defines as breast cancer or a mammary mass [10].

Building on this perspective, TCM further explains that breast cancer does not appear suddenly but rather develops progressively through a chain of energetic imbalances, beginning with emotional stagnation and evolving toward more complex pathological states. It often begins with emotional stagnation, especially when suppressed anger, frustration, or sadness blocks the free flow of Liver Qi; this blockage affects the chest and creates the first tensions or breast sensitivity. Over time, this stagnant Qi hinders blood circulation, leading to blood stasis in the meridians passing through the breasts and causing the formation of small hard and painful nodules. If the Spleen becomes weakened—due to stress, poor diet, or overwork—it produces phlegm that combines with blood stasis, making the lump firmer, more irregular, and harder to dissolve. Finally, when the body's Qi and Yin are further depleted, the natural defenses can no longer control or absorb the mass; toxic heat may then arise, promoting pathological tissue changes and progression to a more severe and invasive stage [8]. These imbalances are diagnosed by observing the tongue [8]:

- Redness or cracks on the sides: Liver stagnation.
- Swollen tongue with thick coating in the center : Dampness or mucus accumulation in the Stomach.
- Dry tongue at the base: Kidney Yin deficiency.
- Red tip: Emotional disorders related to the Heart and Lungs.

V. RESEARCH ON AUTOMATING BREAST CANCER DIAGNOSIS IN TCM

In 2013, Lun-Chien Lo and colleagues published a study [6] exploring how tongue observation in TCM could help tell breast cancer patients apart from healthy individuals. They created a tool called the Automatic Tongue Diagnosis System (ATDS) to analyze tongue images of 60 women with breast cancer and 70 healthy women. The system examined 9 features of the tongue, such as color, texture, cracks, coating, red dots, bruises, tooth marks, saliva, and shape. These features were also linked to specific areas of the tongue connected to different organs like the spleen, liver, kidneys, and heart. Then, the researchers used statistical tests to find out which tongue features showed differences between sick and healthy participants. Three logistic regression models

were developed based on tongue features. These features were first automatically extracted using the ATDS system, then analyzed with the Mann–Whitney test to identify significant differences between patients and healthy controls. The selected features were integrated into the regression models and evaluated using three main indicators: the p-value (P), which assesses statistical significance; the Odds Ratio (OR), which indicates the level of risk associated with each feature (OR > 1 implies increased risk, while OR < 1 suggests reduced risk); and the 95% Confidence Interval (CI), which provides a range for the OR and helps evaluate the model's precision.

Patients show clear differences from healthy individuals in several different types of tongue features. They tend to have more coating on the tongue, especially in the spleen-stomach area, and the coated surface is usually larger. A thin coating is more common among patients. They also show more tooth marks and red dots, which often appear in areas related to the spleen-stomach, left and right liver-gallbladder, and heart-lung regions.

Building on their previous work, Lun-Chien Lo and colleagues published a study in [7] that looks at how tongue observation in TCM can help tell early-stage breast cancer patients (stage 0 and 1) apart from healthy individuals. In this study, they introduced a second version of the ATDS system using the same method as before. They analyzed tongue images from 67 patients with early-stage breast cancer and 70 healthy people. The ATDS 2 system includes a display that shows classification results in real time. Patients with early-stage breast cancer had more coating on their tongues, with larger continuous coated areas. The coating was often thin. They also had bigger red spots on the tongue (both medium and large), especially in regions linked to the spleen–stomach and heart–lungs, based on TCM organ mapping.

In 2021, P.-C. Hsu, Lun-Chien Lo and colleagues published another study [8] that looked more closely at how breast cancer characteristics change with disease stage. They used data collected between 2012 and 2016 from 122 women with breast cancer and 155 healthy women. As in previous studies, tongue features were automatically extracted using the ATDS system. For the analysis, they used different statistical tests depending on the type of data—Student's T-test for numbers like the number of cracks, Chi-square test for categories like tongue color, and logistic regression to estimate the chance of having the disease based on several tongue features. The model's results were measured using p-value, Odds Ratio, and Confidence Interval.

Although this study doesn't report an overall accuracy rate, it highlights clear differences in tongue features between breast cancer patients and healthy individuals. Patients commonly show a smaller tongue, pale color, yellow coating, wet saliva, cracks, and more bruising in the heart-lung area. They also experience fatigue, breast pain, sleep issues, and mood changes. According to TCM, these signs relate to Yin deficiency, lack of blood, heat build-up, and the presence of phlegm or blood stasis. In advanced stages, patients tend to have drier saliva and more bruising in the right liver-gallbladder area.

In 2021, Fan Baochao published a study [9] proposing an integrated system for diagnosis and treatment in TCM. The system includes three main parts: (1) A TCM knowledge graph built from medical concepts, used to model breast cancer diagnosis and treatment. (2) A tongue image analysis module using ResNet models (ResNet-18, ResNet-34, ResNet-50) to automatically identify 22 visual features such as color, shape, texture, and more. ResNet-50 showed the best performance with 87.19% accuracy and became the core of the system. (3) A smart recommendation module called TRHM (Tongue-image Recommended Herb Model), which combines data from the knowledge graph and tongue image features to suggest personalized herbal treatments.

VII. SYNTHESIS AND INTERPRETATION OF TONGUE FEATURES IN BREAST CANCER

A. Descriptive Analysis of Tongue Characteristics by Breast Cancer Stage

Drawing on the existing literature [6] [7] [8], we present a synthesized overview of tongue characteristics across breast cancer stages, and their energetic significance in TCM.

TABLE I. TONGUE CHARACTERISTICS ACCORDING TO BREAST CANCER STAGE (TCM INTERPRETATION)

Tongue Feature	Early Stage (stage 0-I)	Advanced Stage (stage II-IV)	Observed Trend
Small tongue shape	Less frequent (~15–20%)	More frequent (~35–45%)	↗ Increases with cancer progression
Pale color	Moderate (~30–35%)	–	↗ blood deficiency
Thick coating	Present (~30–40%)	Very frequent (~50–60%)	↗ Suggests stagnation and increased dampness
Thick coating	Present (~30–40%)	Very frequent (~50–60%)	↗ Suggests stagnation and increased dampness
Yellow coating	Moderate (~40–45%)	Very frequent (~65–70%)	↗ Reflects internal heat and inflammation
Moist saliva	Rare (~20–25%)	Frequent (~45–50%)	↗ Associated with pathological dampness
Number of fissures	Low to moderate (3–5 fissures)	High (6–9 fissures)	↗ Indicates deficiency of body fluids
Ecchymoses (heart-lung zone)	Few (~2–3)	Numerous (~5–7)	↗ More pronounced blood stasis
Red spots	Numerous (~60–70)	Few (~20–30)	↘ Decrease with disease progression

This comparative analysis highlights how tongue features evolve with the progression of breast cancer. In early stages, signs such as red spots and moderate coating suggest Qi stagnation and mild internal heat, while advanced stages show blood stasis, fluid deficiency, and pathological dampness, reflected in thick yellow coating, and increased fissures.

B. Differential Diagnosis of Lingual Biomarkers Between Early and Advanced Stages of Breast Cancer

To identify the biomarkers and characteristics that distinguish early from advanced stages of the disease, we developed a more detailed comparative table.

TABLE II. COMPARATIVE TABLE FOR DIFFERENTIAL DIAGNOSIS

Characteristic	Early Stage (0–I)	Advanced Stage (III–IV)	Differential Index	TCM Interpretation
Coating quantity	Moderately increased	Clearly increased	Continuously thickening coating	Dampness stagnation, blocked Qi
Maximum coating surface	Wide	Wider	Progressive, sticky extension	Pathogenic Dampness, Qi obstruction
Thin coating	Less frequent	—	Disappears in favor of thick coating	Early-stage Yang/Qi deficiency (digestive Qi)
Thick coating	—	Very frequent	Typical marker of advanced stage	Dampness accumulation, digestive dysfunction
Yellow coating	—	Very frequent	Strong sign of evolving inflammation	Internal Heat, toxic Heat
Surface of red dots	Larger	Decreases	Reduction as disease progresses	Internal Heat, Blood stasis (early stage)
Red dots (Spleen–Stomach area)	Numerous	—	Specific to early stage, digestive zone	Digestive imbalance, local Heat
Red dots (Heart–Lung area)	Numerous	Fewer	Clear decrease in advanced stage	Heat in the upper body (early stage)
Total number of red dots	Increased	Decreases	Inversely proportional to progression	Initial toxic Heat → exhaustion later
Small tongue shape	—	More frequent	Morphological sign of progression	Qi deficiency, general weakening
Tongue color	Slightly reddish, normal	Dark red / purplish	Darkening with stasis	Blood stasis and toxic Heat (advanced)
Saliva / moisture film	Not frequent	Frequent	Marked pathological Dampness	Dysfunction of body fluids
Fissures	Mild to moderate (3–5)	Numerous (6–9)	Increase in fissures	Yin deficiency, fluid depletion
Ecchymoses (Heart–Lung area)	Few (2–3)	Many (5–7)	Intensification of ecchymoses	Blood stasis, cardiac Qi blockage

From this comparison, we identified the key characteristics of both the early and advanced stages of the disease.

Early Stage (0–I):

- **Numerous red spots:** especially in spleen–stomach and heart–lung zones.

- **Moderately thick coating and extended red surface:** reflects emerging internal heat.
- **Overall normal or slightly reddish color:** no deep stasis yet.
- **Limited fissures (3–5):** early fluid deficiency, not severe.

Advanced Stage (III–IV):

- **Thick yellow coating frequently observed:** indicative of damp stagnation and toxic heat.
- **Marked reduction in red spots and increased ecchymoses:** established blood stasis.
- **Small tongue and dark red/purplish color:** global weakness and stasis.
- **Numerous fissures (6–9) and pathological moist saliva:** fluid deficiency and digestive dysfunction.



Fig 3. a) Early stage BC b) Advanced stage BC

We also deduced the characteristics of a healthy tongue.

Healthy tongue:

- **Color:** pale red signifies a harmonious balance of Qi, blood, and Yin-Yang.
- **Shape:** normal size, without marked fissures or swelling signifies an absence of Qi/Yang deficiency or stasis.
- **Coating:** thin, white, and uniform signifies a proper function of the Spleen and Stomach, balanced digestion, no dampness or pathological heat.
- **Moisture:** slightly moist surface signifies an adequate distribution of body fluids, neither dryness nor excess dampness.
- **Absence of pathological signs:** no red spots, no thick or sticky coating, no ecchymoses, and no abnormal discolorations → reflects a physiological state of health.

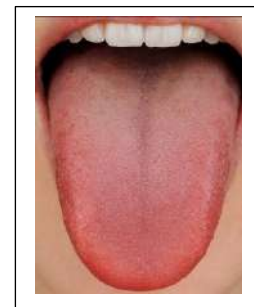


Fig 4. Healthy tongue

VII. TOWARD AI-ASSISTED BREAST CANCER DIAGNOSIS INSPIRED BY TCM

A. Methodology

In this section, we propose a methodology for developing an AI-assisted diagnostic system for breast cancer, inspired by the traditional tongue-based diagnostic process in TCM. The approach relies on the acquisition and annotation of tongue images, which are analyzed to extract stage-specific features such as color, coating, fissures, and the presence of red spots. These features, traditionally interpreted as indicators of internal imbalances, can be reformulated as potential biomarkers when mapped to clinical stages of breast cancer. By training machine learning models such as Convolutional Neural Networks (CNN), ResNet, or Vision Transformers (ViT) on large datasets of tongue images, the system would learn to recognize subtle visual patterns and correlate them with disease progression. The methodology integrates image preprocessing, feature extraction, and classification pipelines, aiming to provide a non-invasive, complementary tool for early detection and staging.

B. Dataset Challenges and Emerging Solution

One of the major obstacles in developing an AI-assisted diagnostic system for breast cancer based on tongue image analysis is the unavailability of open datasets. The few datasets reported in the literature remain private, and the Asian research community, which has pioneered most of the work in this field, is generally not open to data sharing. This lack of accessible resources poses a significant challenge to the realization of the project. To address this limitation, three complementary strategies can be considered. First, generative AI techniques may be employed to create synthetic tongue images (Fig. 3 and Fig. 4 illustrate tongue images generated by generative AI) thereby demonstrating the feasibility of the proposed approach. Second, existing tongue image datasets available on public platforms such as Kaggle and Roboflow can be analyzed to extract relevant features and train preliminary models. Finally, for validation purposes, real data could be collected from volunteer patients, in strict compliance with national regulations. In Algeria, this process must adhere to Law No. 18-05 of 10 May 2018 on Health, which governs biomedical research and ensures the protection of patient rights and informed consent.

C. Scientific feasibility of generative AI in medicine

In support of the solution proposed above, the scientific feasibility of using generative AI to create synthetic medical images has been confirmed by recent studies conducted by Chen et al. [12] and Hu [13], which provide converging evidence of its relevance in the medical field. Chen and colleagues demonstrated that the creation of synthetic data is a viable approach to overcoming the scarcity and sensitivity of clinical data, enabling the expansion of training sets while preserving patient confidentiality. Meanwhile, Hu et al. showed that the generation of artificial medical images using models such as GANs, StyleGAN, or diffusion models improves the performance of disease classification systems (healthy/sick) and enhances the robustness of models against

data variability. Together, these two studies confirm the scientific feasibility of employing synthetic data and images as training material for AI models in medicine. However, they also emphasize the need for rigorous clinical validation to avoid the introduction of bias or artefacts, particularly in more complex tasks such as disease staging.

VIII. CONCLUSION

This study aimed to explore the potential of automated breast cancer diagnosis inspired by tongue-based diagnostic processes in Traditional Chinese Medicine. Although very few studies exist in the literature, their analysis allowed us to identify the key diagnostic features associated with both early and advanced stages of the disease. These findings provide a solid foundation for building a data-driven diagnostic system. Despite the lack of publicly available datasets, generative AI offers an initial solution to demonstrate the feasibility of the approach. The project remains highly promising, as it envisions a simple, non-invasive, and user-friendly tool capable of detecting early energetic imbalances related to breast cancer, even before the development of palpable breast masses.

REFERENCES

- [1] D. Ma, S. Wang, Y. Shi, S. Ni, M. Tang, et A. Xu, "The development of traditional Chinese medicine", *Journal of Traditional Chinese Medical Sciences*, vol. 8, p. S1-S9, nov. 2021.
- [2] Q. Zhang, J. Zhou, et B. Zhang, "Computational Traditional Chinese Medicine diagnosis: A literature survey", *Computers in Biology and Medicine*, vol. 133, p. 104358, juin 2021.
- [3] M. H. Tania, K. Lwin, et M. A. Hossain, "Advances in automated tongue diagnosis techniques", *Integrative Medicine Research*, vol. 8, no 1, p. 42-56, mars 2019.
- [4] J. Zaho, "Initiation au monde de la médecine traditionnelle chinoise : De la philosophie des cinq éléments et du Qi et aux pratiques de guérison", Avenet Edition, 2023 .
- [5] G. Maciocia, "L'examen de la langue en Médecine chinoise", SATAS Edition , 2006.
- [6] L.-C. Lo. et al., "Breast cancer index: A perspective on tongue diagnosis in traditional chinese medicine," *Journal of Traditional and Complementary Medicine*, vol. 3, no. 3, pp. 194-203, 2013.
- [7] L.-C. Lo, T.-L. Cheng, Y.-J. Chen, S. Natsagdorj, and J. Y. Chiang, "Tcm tongue diagnosis index of early-stage breast cancer," *Complementary Therapies in Medicine*, vol. 23, no. 5, pp. 705-713, 2015.
- [8] P.-C. Hsu, H.-K. Wu, H.-H. Chang, J.-M. Chen, J. Y. Chiang, et L.-C. Lo, "A Perspective on Tongue Diagnosis in Patients with Breast Cancer", *Evidence-Based Complementary and Alternative Medicine*, vol. 2021, p. 1-9, déc. 2021.
- [9] F. Baochao, "Recherche sur le diagnostic de la langue en MTC et la technologie de recommandation de la MTC pour le cancer du sein basée sur un graphe de connaissances et un réseau de neurones," Université de médecine traditionnelle chinoise de Guangzhou, 2021.
- [10] D. N. Lu, "A traditional chinese medicine view of breast cancer," in *A Traditional Chinese Medicine View of Breast Cancer*, 2024.
- [11] J. Li et al., "A multi-step approach for tongue image classification in patients with diabetes ", *Computers in Biology and Medicine*, vol. 149, p. 105935, oct. 2022.
- [12] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. K. Williamson, and F. Mahmood, "Synthetic data in machine learning for medicine and healthcare," *Nature Biomedical Engineering*, vol. 5, pp. 493-497, 2021.
- [13] E. A. Hu, "A survey of synthetic medical image generation for improving disease classification," *Stanford University Press*, 2024.

Sliding Mode Control of Series Multicell Inverters: Simulation and Performance Analysis on 3-Cell and 5-Cell Topologies

Fateh Khelifi

Dept of Process Automation & Electrification
Faculty of Hydrocarbons and Chemistry
University of Boumerdes
Boumerdes, Algeria
Fa.khelifi@univ-boumerdes.dz

Aimad Boudouda

Dept of Electrical Systems Engineering
Faculty of technology
University of Boumerdes
Boumerdes, Algeria
a.boudouda@univ-boumerdes.dz

Tadger Sid ahmed

Dept of Process Automation & Electrification
Faculty of Hydrocarbons and Chemistry
University of Boumerdes
Boumerdes, Algeria
s.tadger@univ-boumerdes.dz

Abstract—this study focuses on the analysis and comparison of series multicell inverters composed of three and five cells. These structures, capable of generating multiple voltage levels at the output, enhance signal quality, reduce harmonic distortion, and limit the electrical stresses applied to power switches. In this work, a closed-loop control strategy based on Sliding Mode Control (SMC) is developed to simultaneously ensure output current regulation and dynamic balancing of the floating capacitor voltages. The study also investigates the influence of the number of cells on voltage synthesis, the behaviour of floating sources, and the resulting control complexity. The evaluated performance indicators include voltage quality (THD), capacitor voltage balancing, and dynamic response. Simulation results obtained using MATLAB/Simulink demonstrate the robustness and effectiveness of the proposed SMC strategy, outperforming conventional linear control techniques.

Keywords—power converter, multicell inverter, sliding-mode control, Floating Capacitor Voltage Balancing, MATLAB/Simulink

I. Introduction

Multilevel inverters play a key role in modern energy conversion systems, particularly in medium- and high power applications such as electrical grids, motor drives, and grid connected photovoltaic installations. These converters are capable of generating nearly sinusoidal output voltages with low total harmonic distortion (THD) while reducing switching stresses on semiconductor devices. Among the various existing topologies, the series multicell inverter stands out due to its modular structure and its ease of extension to higher voltage levels. Each cell incorporates a floating capacitor that is essential for generating the voltage levels. However, one of the main challenges of this architecture lies in maintaining the voltage balance of these capacitors, a necessary condition to guarantee a symmetrical output voltage and to prevent dynamic imbalances that could degrade system performance [1].

Conventional control strategies, such as proportional integral (PI) regulators, provide satisfactory performance under nominal operating conditions but show limitations in the presence of rapid load variations, external disturbances, and system nonlinearities. To overcome these drawbacks, nonlinear, robust, and adaptive control strategies have received increasing attention in the field of power converters [2].

Among these approaches, Sliding Mode Control (SMC) offers a particularly robust solution. Based on Lyapunov stability theory, SMC ensures rapid convergence toward a predefined sliding surface, guaranteeing accurate tracking of the reference current and strong robustness to model uncertainties and external perturbations. Moreover, SMC enables simultaneous control of the injected output current and the balancing of floating capacitor voltages without requiring precise knowledge of system parameters [3].

In this context, the present work investigates and compares series multicell inverters composed of three and five cells. These architectures, capable of generating multiple output voltage levels, improve signal quality, reduce harmonic distortion, and decrease electrical stress on the power switches.

The study highlights the influence of the number of cells on voltage synthesis, the behaviour of floating sources, and the overall control complexity [4].

The proposed control strategy ensures both output current regulation and dynamic balancing of the floating capacitor voltages. A switching-state selection method is introduced to determine, in real time, the optimal combination of switches that simultaneously minimizes the current error and the capacitor voltage imbalance. Simulation results obtained using MATLAB/Simulink demonstrate the

robustness and superior dynamic performance of the proposed SMC strategy compared to conventional linear control techniques.

II. Multicell Series Inverter

The multicell inverter with floating capacitors is an advanced configuration composed of several switching cells connected in series, each separated by floating capacitors. These capacitors play a crucial role in stabilizing intermediate voltage levels, enabling the generation of multiple output voltage levels [1].

This type of multicell inverter offers several significant advantages. First, it can produce a higher number of voltage levels with a reduced number of components. This greatly improves the quality of the output voltage by generating a waveform closer to a sinusoid and thus reducing the total harmonic distortion (THD). Second, the presence of multiple capacitors and switching states provides inherent redundancy, enhancing system reliability: in the event of a component failure, the inverter can continue operating through available alternative paths. Finally, this architecture supports a wide range of advanced control strategies, such as multilevel pulse-width modulation and sliding mode control, offering high flexibility and adaptability to meet demanding requirements in terms of performance, stability, and robustness in industrial and energy-conversion applications.

A. Topology of the Multicell Series Inverter

The series Multicell inverter consists of several switching cells connected in series, each cell incorporating a floating capacitor. A typical P cell structure is composed of P identical full bridge or half-bridge switching units, where the floating capacitors are inserted between adjacent cells to establish intermediate voltage levels. These capacitors charge and discharge according to the selected switching states, enabling the synthesis of multiple output voltage levels without additional isolated DC sources. The output voltage is obtained as the sum of the voltages generated by each cell, providing a staircase-like waveform whose quality increases with the number of cells [5].

Figure 1 shows a series multicellular inverter with P cells and a capacitive midpoint, supplying an RL Load. Each cell consists of two controllable switches and a floating capacitor, which enables the generation of multiple output voltage levels. The series connection of all cells produces a staircase output voltage applied to the load.

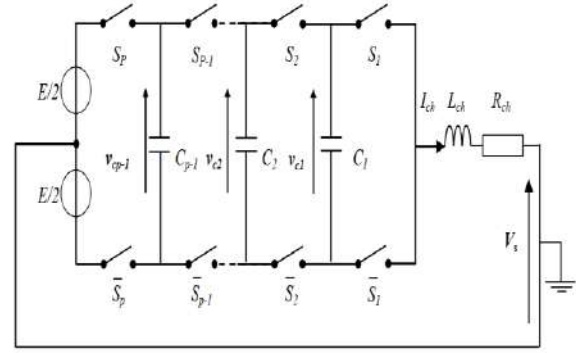


Fig 1. A single leg of a P-cell inverter with a capacitive midpoint is connected to an R L

The characteristic quantities of the P-cell switching arrangement are defined in Table 1.

Table.1.characteristic quantities of the P-cell switching arrangement

Number of associated cells	P
Number of associated capacitors	P-1
Number of possible states	2^P
Number of output voltage levels	P+1
Value of the main supply voltage	E
Value of the voltage source of cell i	$V_{ci} = (i/P) E$

The load voltage V_{ch} is the sum of the voltages of all the cells:

$$V_s = \sum_{i=1}^P V_i = \sum_{i=1}^P S_i [V_{ci} - V_{c(i-1)}] \quad (1)$$

III. Sliding Mode Control (SMC)

In Sliding Mode Control (SMC) for variable-structure systems, the system state trajectory is directed toward a predefined surface and is constrained to remain close to this surface by the switching law. This surface is known as the sliding surface, and the motion along it is called the sliding motion [6-7].

The main objectives of SMC can be summarized as follows:

- Design a surface $s(x)$ such that all system trajectories satisfy the desired performance in terms of tracking, regulation, and stability.

- Determine a control law u (switching law) capable of attracting and maintaining all state trajectories on the sliding surface.

The design of the control law involves three closely interrelated steps:

1. Selection of an appropriate sliding surface $s(x)$.
2. Establishment of the existence conditions for convergence to the surface.
3. Determination of the control law u that ensures convergence and maintenance on the surface.

This state-space representation forms the basis for defining the sliding surface and designing an appropriate SMC law to ensure precise current tracking and dynamic balancing of the floating capacitor voltages [1].

According to the standard state-space equation of the series multicell converter model, the system is expressed as:

$$\dot{X} = A \cdot X + B(X) \cdot S \quad (2)$$

The state-space representation of a multicell converter is expressed as follows:

$$\dot{x} = f(x) + g(x)u \quad (3)$$

The sliding surface is generally given by the standard equation: $s(x) = \Delta x = x - x_{réf}$ (4)

Where:

- $x(t)$ is the system state vector (floating capacitor voltages and output current),
- $u(t)$ is the control input applied to the inverter switches,
- $f(x(t))$ describes the uncontrolled system dynamics,
- $g(x(t))$ represents the influence of the control input on the system dynamics.

The error state vector is given by:

$$\Delta x = \begin{pmatrix} x_1 - x_{réf} \\ x_2 - x_{réf} \\ x_3 - x_{réf} \\ \vdots \\ x_p - x_{réf} \end{pmatrix} \quad (5)$$

After several transformations, we obtain a representation of the control signals S_i :

$$\begin{cases} S_1 = \text{sgn} \left(-\frac{P}{E} I_{ch \text{ réf}} V_{C_1} - I_{ch} \right) \\ S_2 = \text{sgn} \left(-\frac{P}{E} I_{ch \text{ réf}} (V_{C_1} - V_{C_2}) - I_{ch} \right) \\ S_P = \text{sgn} \left(-\frac{P}{E} I_{ch \text{ réf}} (V_{C_{P-1}} - V_{C_P}) - I_{ch} \right) \end{cases} \quad (6)$$

After determining the switching laws, they are implemented within a set of feedback loops (Figure 2). Each state variable has its own control loop, where the error with respect to the reference value is calculated. The P switching functions are then obtained by linearly combining the errors of all state variables. Finally, P hysteresis comparators generate the P control signals for the multicell converter.

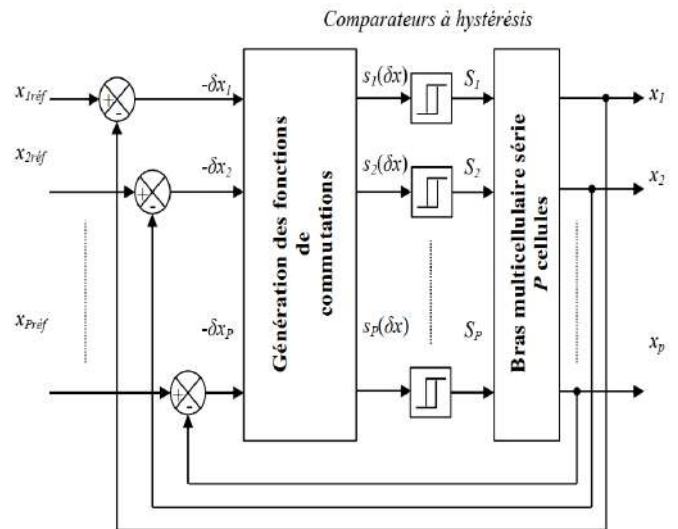


Fig 2. Block diagram of the Sliding Mode Control (SMC) for a series multicell inverter.

IV. Simulation Results

To evaluate the efficiency and speed of regulation of sliding mode control, two simulations were performed using MATLAB/SIMULINK. These simulations involve a series multicell inverter with 3 cells and 5 cells feeding an R-L load, both controlled using sliding mode control. The

parameters used for these simulations are provided in Table 2.

Table .2 Settings used in the case of command by sliding mode.

Parametre	Values
Number of associated cells (P)	3 and 5
DC voltage(E)	1500 v
Footing capacitors (C)	40μF
Résistor (R _{ch})	10 Ω
Inductance (L _{ch})	1 mH
Switching frequency (f _{sw})	10 KHz

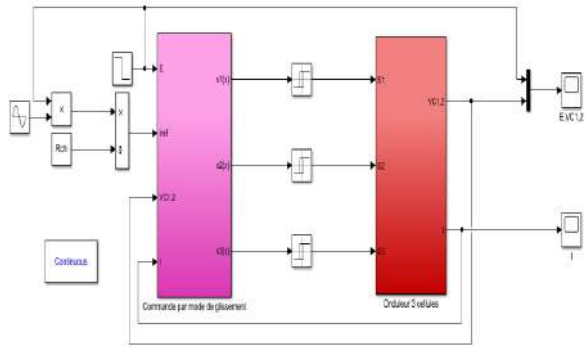


Fig. 3. Simulation block of Three-cell single-phase multicell inverter, supplying an R-L load, controlled by slip mode control

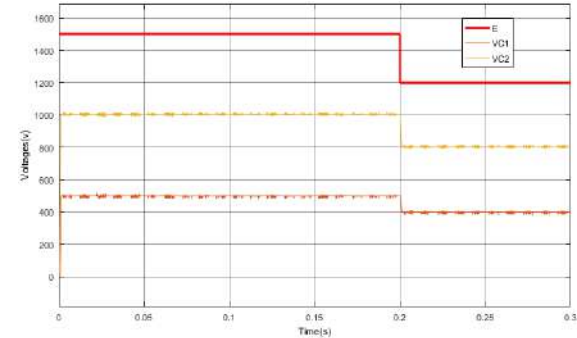


Fig 4. Evolution of floating voltages. (3 cells)

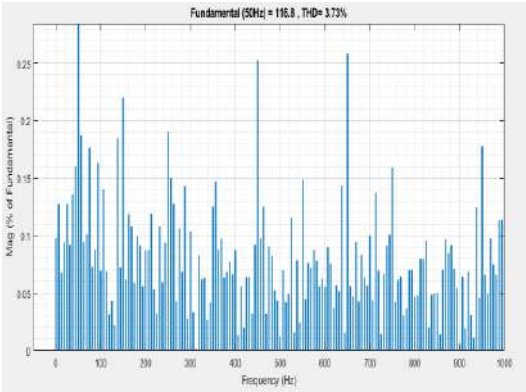
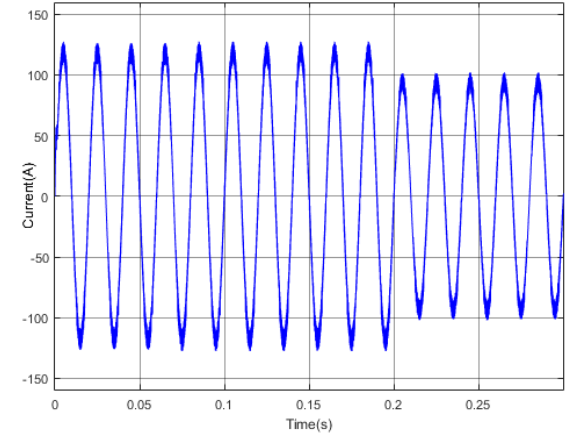


Fig5. Current and THD waveform

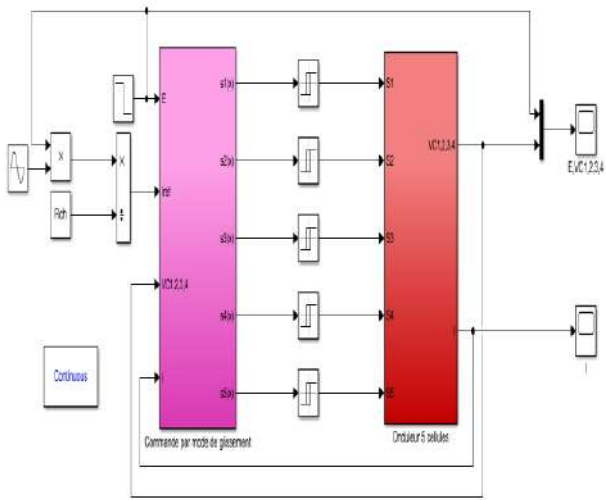
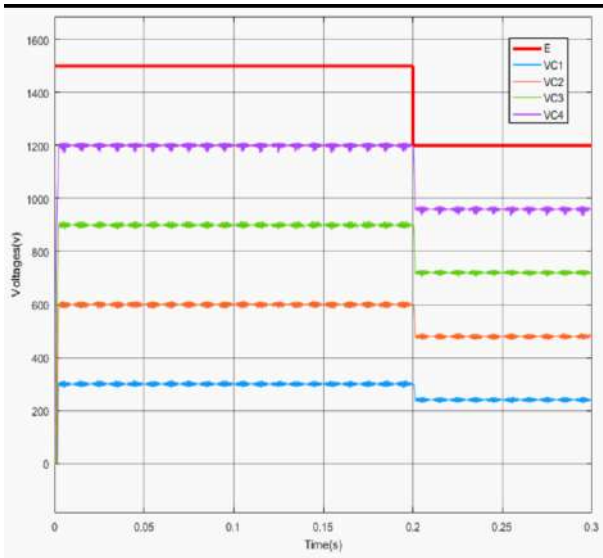


Fig. 6. Simulation block of five-cell single-phase multicell inverter, supplying an R-L load, controlled by slip mode control



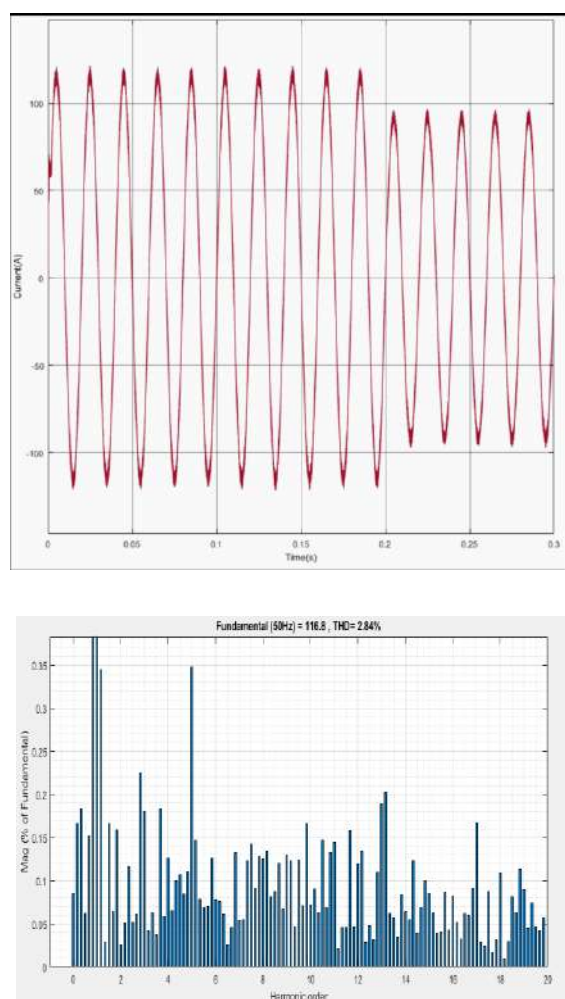


Fig.8. Current and THD waveforms

Based on the results presented in figures 4,5,7 and 8, it can be deduced that the sliding mode control technique enforces a rapid rebalance of the floating capacitor voltages with a very brief transient regime, well below 0.01 seconds for both inverters. This means that these voltages reach their reference values much faster than what is observed in open-loop control mode.

Furthermore, the rapid regulation capability of this technique is evidenced by the response to a 300 V voltage drop applied at $t=0.2$ s, where the floating voltages follow and adjust to this change extremely quickly. This swift adjustment is also reflected in the load current waveform for both inverters.

When comparing the two inverters, the load current waveform is notably cleaner and closer to a sinusoidal shape for the 5 cell inverter compared to the 3 cell inverter, which aligns with theoretical expectations. Additionally, an improvement in the Total Harmonic Distortion (THD) value is

observed, with 2.84% for the 5 cell inverter compared to 3.73% for the 3 cell one. This demonstrates the benefits of increasing the number of cells in enhancing output waveform quality and reducing harmonic distortion.

VI. Conclusion

The sliding mode control technique applied to multilevel inverters with floating capacitors ensures a rapid and robust regulation of floating capacitor voltages, achieving steady-state values in extremely short transient times (under 0.01 seconds). This fast rebalancing significantly outperforms open-loop control methods. The technique effectively handles sudden voltage drops, quickly adjusting the floating voltages and maintaining stable operation, reflected in the high-quality load current waveform.

Furthermore, increasing the number of cells in the inverter structure improves the quality of output current, producing waveforms that are closer to an ideal sinusoid and reducing total harmonic distortion (THD). Specifically, a 5cell inverter exhibits cleaner current waveforms and a lower THD (2.84%) compared to a 3cell inverter (3.73%). This confirms the theoretical advantages of multicell topologies combined with sliding mode control for enhanced performance and power quality in inverter applications.

Future research on dynamic Sliding Mode Controllers (SMC) for series multicell inverters should focus on integrating integral and dynamic sliding laws to eliminate the reaching phase and reduce chattering, enhancing control smoothness and robustness against parametric uncertainties and load variations. This approach generates a continuous PWM signal at fixed frequency, minimizing switching losses while improving converter efficiency and floating capacitor voltage regulation

References

1. M. Benmiloud, A. Benalia, And M. Djemai, "Hybrid Sliding Mode Control for Two Cells Converter, " 13th IEEE Workshop on Variable Structure Systems", VSS'14, June 29 -July 2, 2014, Nantes, France.
2. N. Ghaffari1, A. Zakipour1, and M. Salimi2, "State Space Modelling and Sliding Mode Current Control of the Grid Connected Multi-Level Flying Capacitor Inverters ", Journal of electrical and computer engineering innovations, 9(2): 215-228, 2021.
3. A. Zakipour, K. Aminzare, and M. Salim , "Sliding Mode Controller Design for Stabilization of the Three- Phase Grid-Connected Inverters in the Presence of Unbalanced Local Loads

", Iranian Journal of Electrical and Electronic Engineering 03 (2022) 2125 .

4. M. Benmiloud, A. Benalia, And M. Djemai, "Hybrid Sliding Mode Control for Two Cells Converter, "13th IEEE Workshop on Variable Structure Systems,VSS'14, June 29 -July 2, 2014, Nantes, France.

5. Kamel Laidi, Ouahid Bouchhida, Mokhtar Nibouche, "Smart control based on neural networks for Multicellular converters, " Archives Of Electrical Engineering VOL. 70(3), pp. 531 –550 (2021).

6. Humam Al-Baidhani, and Marian K. Kazimierczuk, "Simplified Double-Integral Sliding-Mode Control of PWM DC-AC Converter with Constant Switching Frequency", Appl. Sci. 2022, 12, 10312. <https://doi.org/10.3390/app122010312>.

7. A. K. Sadigh, S. H. Hosseini, M. Sabahi, G. B. Gharehpetian, "Double flying capacitor Multicell converter based on modified phase-shift pulse width modulation," IEEE Transactions on Power Electronics, vol. 25, pp.1517-1526, Aug 2010.

A Comprehensive Survey of Short-Term Load Forecasting Techniques for Smart Grids

Assala Nacef

LRSD laboratory, Faculty of Sciences
University Ferhat Abbas setif 1
setif, Algeria
assala.nacef@univ-setif.dz

Lina Benchikh

LRSD laboratory, Faculty of Sciences
University Ferhat Abbas setif 1
setif, Algeria
lina.benchikh@univ-setif.dz

Ouissem Touameur

LRSD laboratory, Faculty of Sciences
University Ferhat Abbas setif 1
setif, Algeria
ouissem.touameur@univ-setif.dz

Djamila Mechta

LRSD laboratory, Faculty of Sciences
University Ferhat Abbas setif 1
setif, Algeria
mechtadjamila@univ-setif.dz

Rahma Berchi

LRSD laboratory, Faculty of Sciences
University Ferhat Abbas setif 1
setif, Algeria
rahma.berchi@univ-setif.dz

Abstract—Short-Term Load Forecasting (STLF) plays a critical role in ensuring the reliability, stability, and efficiency of modern power systems. Accurate load forecasting enables effective real-time scheduling, demand response, market bidding, and grid management. This paper presents a comprehensive review of STLF techniques, analyzing their evolution from traditional statistical models to advanced Artificial Intelligence (AI) and hybrid approaches. The study highlights hybrid frameworks that integrate multiple paradigms to improve adaptability and prediction accuracy. Furthermore, the paper discusses commonly used evaluation metrics such as Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and Continuous Ranked Probability Score (CRPS). Finally, emerging challenges and future research directions are identified, emphasizing the need for explainable, uncertainty-aware, and real-time forecasting systems suitable for smart grids and energy internet environments.

Index Terms—Short-Term Load Forecasting (STLF), Artificial Intelligence, Deep Learning, Hybrid Models, Smart Grid

I. INTRODUCTION

The reliable forecasting of electrical load is fundamental to the efficient operation, planning, and management of power systems. As global energy infrastructures transition toward smart grids and the energy internet paradigm, the need for accurate and adaptive forecasting techniques becomes increasingly critical. Load forecasting supports a broad spectrum of applications, including demand response, renewable integration, energy trading, system security, and grid stability.

Among the different forecasting horizons, Short-Term Load Forecasting (STLF) plays a particularly vital role, as it deals with rapid fluctuations in electricity demand over time scales ranging from seconds to a few days. Such forecasts directly influence real-time scheduling, dispatching, and market bidding operations. However, the growing penetration of renewable energy sources, electric vehicles, and prosumers introduces high uncertainty and nonlinearity into load patterns, posing significant challenges to traditional forecasting models.

Historically, statistical models such as ARIMA, SARIMA, and exponential smoothing were the primary approaches used for STLF due to their interpretability and computational simplicity. With the advent of Artificial Intelligence (AI), the field has witnessed a paradigm shift toward data-driven methods, including Machine Learning (ML), Deep Learning (DL), and metaheuristic optimization. These models demonstrate strong capabilities in capturing nonlinear temporal dependencies and complex correlations across multi-dimensional datasets. More recently, hybrid models that combine statistical, learning-based, and heuristic techniques have emerged as a dominant trend, offering enhanced robustness, adaptability, and accuracy.

This paper presents a comprehensive review of Short-Term Load Forecasting methodologies, emphasizing their evolution from traditional statistical models to intelligent and hybrid approaches. It classifies existing techniques, compares their performance and applicability, and discusses the emerging challenges and research directions for future development in STLF.

II. RELATED REVIEWS

(1) explained the different types of load forecasting and highlighted their applications. The authors provided a comprehensive classification of load forecasting techniques, including traditional, clustering-based, AI-based, time-series, and metaheuristic-based techniques. They concluded that AI-based methods, particularly machine learning (ML) and deep learning (DL) models, achieved superior accuracy in terms of Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE).

(2) classified STLF methodologies into three evolving paradigms. The first encompasses *Statistical Models*, including time-series analysis (e.g., ARIMA, SARIMA), exponential smoothing, and regression techniques. The second category consists of *Intelligent-Computing-Based Models* (ICM), leveraging algorithms such as Artificial Neural Networks (ANNs),

Support Vector Machines (SVMs), and Long Short-Term Memory (LSTM) networks. The final category, *Hybrid Models*, integrates two or more techniques (e.g., ANN-PSO, SVM-GA, EMD-LSTM) to combine their strengths and reduce weaknesses. Hybrid models generally achieve superior robustness and accuracy.

(3) presented a systematic review of STLF models, categorizing them into three groups:

- Statistical Models: ARIMA, SARIMA, Exponential Smoothing, and Generalized Linear Models (GLMs)
- Machine Learning Techniques: SVM, Decision Trees, Random Forest, Gradient Boosting, MLP, and Deep Learning
- Hybrid Models: Frameworks combining statistical and machine learning methods

Their analysis demonstrates that hybrid models consistently capture both linear and nonlinear patterns more effectively.

(4) provided a comprehensive review of DL-based approaches for STLF. The authors discussed major architectures (RNN, LSTM, CNN, Autoencoders), main objectives, and datasets. They concluded that future DL models should be uncertainty-aware, online, and deployable within IoT-enabled energy systems.

The surveys reviewed above mainly list and describe forecasting methods, but they rarely compare how or why these techniques perform differently under real operating conditions. They give limited attention to how model assumptions affect accuracy, how computational cost influences real-time use, or how privacy and communication issues affect deployment. They also overlook the relationship between data quality and system requirements. In contrast, this survey not only summarizes existing methods but also explains the theoretical behavior of each model family, discusses practical deployment constraints, and identifies key open challenges, providing a clearer and more complete understanding of short-term load forecasting.

TABLE I
COMPARISON OF REVIEWED STLF STUDIES

Feature	(1)	(2)	(3)	(4)	Ours
Definition of STLF	✓	✓	✓	✓	✓
AI-based Models	✓	✓	✓		✓
Deep Learning Models	✓			✓	✓
Statistical Models	✓	✓	✓		✓
Clustering-based Models	✓				✓
Time Series Models	✓	✓	✓		✓
Heuristic Models	✓				✓
Hybrid Models		✓	✓		✓

III. LOAD FORECASTING

Load forecasting is the process of predicting future electricity demand over a specific time horizon by utilizing historical consumption data, weather conditions, and other influencing factors. To achieve accurate predictions, load forecasting models require comprehensive and realistic datasets that effectively capture the variability of these parameters (5). Load forecasting is typically categorized into three levels based on the forecasting horizon and operational objectives:

- Short-Term Load Forecasting (STLF) deals with predictions over seconds, minutes, hours, or days and is mainly used for real-time operations and control.
- Medium-Term Load Forecasting (MTLF) focuses on forecasting electricity demand from weeks to months, supporting maintenance planning and energy procurement.
- Long-Term Load Forecasting (LTLF) addresses prediction horizons ranging from several months to years and is primarily used for infrastructure development and investment planning.

IV. SHORT-TERM LOAD FORECASTING

Short-Term Load Forecasting (STLF) targets specific types of electricity usage that are highly susceptible to real-time variations and exhibit high volatility. Typical applications include real-time scheduling, system security assessment, demand response, and market bidding. Unlike Medium-Term Load Forecasting (MTLF) and Long-Term Load Forecasting (LTLF), STLF demands higher prediction accuracy and has a more immediate impact on operational decision-making within power systems.

DATA TYPES FOR SHORT-TERM LOAD FORECASTING (STLF)

STLF utilizes multiple data categories to model electricity demand accurately:

- Historical Load Data: Time-series data of past electricity consumption that captures inherent load patterns, trends, and auto-correlations.
- Temporal Data: This includes time-based indicators such as the time of day, day of the week, month, season, and flags for holidays or special events. This data is crucial for modeling the recurring patterns of human activity and economic cycles that drive electricity usage.
- Weather Data: Variables like temperature, humidity, wind speed, and solar irradiance (or cloud cover) are key external factors that account for a significant portion of short-term load variability, primarily due to their direct impact on heating and cooling (HVAC) systems.
- Economic Factors and Market Data: This category encompasses broader factors such as general economic conditions, dynamic electricity pricing schemes (e.g., real-time pricing), and demand response events. These elements influence consumption levels by affecting human financial decisions and operational behavior.

A. Evaluation Metrics

Evaluating the performance of Short-Term Load Forecasting (STLF) models requires well-defined statistical metrics that quantify the deviation between predicted and actual load values. These metrics assess both the **accuracy** and **reliability** of forecasting models across different time horizons and datasets. Commonly adopted performance indicators include the following:

1) *Mean Absolute Percentage Error (MAPE)*: The Mean Absolute Percentage Error (MAPE) measures the average relative deviation between the predicted and actual loads:

$$\text{MAPE} = \frac{100}{N} \sum_{t=1}^N \left| \frac{y_t - \hat{y}_t}{y_t} \right|$$

where y_t and $\hat{y}_t$ denote the actual and predicted loads at time t , respectively, and N is the number of observations. MAPE provides an intuitive percentage-based error but becomes unstable when actual values approach zero.

2) *Mean Absolute Error (MAE)*: The Mean Absolute Error (MAE) evaluates the average magnitude of forecast errors without considering their direction:

$$\text{MAE} = \frac{1}{N} \sum_{t=1}^N |y_t - \hat{y}_t|$$

MAE is robust to outliers and easy to interpret, providing a linear measure of average deviation.

3) *Root Mean Square Error (RMSE)*: The Root Mean Square Error (RMSE) penalizes larger deviations more heavily than MAE:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{t=1}^N (y_t - \hat{y}_t)^2}$$

This metric is particularly useful for applications that require minimizing large forecasting deviations, such as system stability and market operations. However, RMSE can be sensitive to outliers.

4) *Continuous Ranked Probability Score (CRPS)*: The Continuous Ranked Probability Score (CRPS) extends traditional point-based metrics to probabilistic forecasts. It measures the distance between the predicted cumulative distribution function (CDF) F_t and the observed value y_t :

$$\text{CRPS}(F_t, y_t) = \int_{-\infty}^{\infty} [F_t(x) - \mathbb{I}(x \geq y_t)]^2 dx$$

where $\mathbb{I}(\cdot)$ is the indicator function. Lower CRPS values indicate better probabilistic accuracy.

5) *Coverage Probability (CP)*: Coverage Probability (CP) quantifies the reliability of prediction intervals by evaluating the proportion of actual values that fall within a predefined confidence interval (typically 90% or 95%):

$$\text{CP} = \frac{1}{N} \sum_{t=1}^N \mathbb{I}(y_t \in [L_t, U_t])$$

where L_t and U_t denote the lower and upper bounds of the forecast interval. A well-calibrated model achieves coverage close to the nominal confidence level.

V. TAXONOMY OF STLF APPROACHES

A. Classical Statistical Methods

1) *VAR, ARIMA, and SARIMA Models*: Recent studies continue to employ VAR, ARIMA, and SARIMA for STLF, often as standalone models or as benchmarks in hybrid frameworks. For example, Wang et al. (2025) developed a method integrating VAR with deep learning to capture multivariate dependencies in load data, demonstrating that VAR remains relevant for modeling multiple influencing factors (6) compared ARIMA, SARIMA, and Holt-Winters models for power imbalance forecasting, finding SARIMA particularly effective for capturing seasonality in short-term energy data (7). ARIMA and SARIMA are also frequently used as baselines in studies comparing traditional and advanced models (8), (9), (10).

2) *Regression and Polynomial Fitting Approaches*: Linear regression and its variants are still widely used, especially for their simplicity and interpretability. (11) evaluated 19 regression models—including linear regression, regression trees, and Gaussian process regression—for campus load forecasting, finding that classical regression models remain competitive in certain contexts. (12) proposed an improved regression model using mini-batch stochastic gradient descent for large-scale STLF, highlighting the scalability of regression-based approaches. Polynomial fitting has also been applied for short-term power load forecasting, with (13) demonstrating its accuracy and stability.

3) *Exponential Smoothing and Holt-Winters Methods*: Exponential smoothing and Holt-Winters models are used for their ability to handle trends and seasonality in load data. (7) found that Holt-Winters performed well for series with clear seasonal patterns, though SARIMA generally outperformed it. (14) introduced an interval Holt-Winters model for interval-valued load forecasting, showing improved robustness over point forecasting models.

B. Machine Learning-Based Methods

Traditional ML models such as Support Vector Regression (SVR), Random Forests (RF), Extreme Gradient Boosting (XGBoost), LightGBM, Decision Trees, and Multiple Linear Regression remain widely used, especially for structured/tabular data and as interpretable baselines (15), (16), (17), (18), (19). These models are valued for their speed, transparency, and ability to handle moderate nonlinearity.

C. Deep Learning-Based Methods

Deep learning models have become dominant, especially for complex, high-dimensional, and sequential data. Key architectures include:

- Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU) for temporal dependencies (15), (20), (21), (22).
- Convolutional Neural Networks (CNNs) and CNN-LSTM hybrids for spatial-temporal feature extraction (22), (23), (24)

- Transformer models and attention mechanisms for capturing long-range dependencies and improving interpretability (25), (26), (27), (28). Temporal Convolutional Networks (TCN), Graph Neural Networks (GNN), and Adaptive/Transfer Learning frameworks for specialized tasks (20), (26), (29).

D. Clustering-Based Methods

Clustering-based models have become increasingly prominent in short-term load forecasting (STLF), offering a way to handle the diversity and complexity of load patterns by grouping similar data before applying forecasting algorithms. Clustering algorithms (e.g., K-means, K-medoids, Fuzzy C-means, K-shape) are used to group load profiles or days with similar consumption patterns, operational modes, or weather conditions. This step helps reduce data heterogeneity and allows models to focus on more homogeneous subsets (30), (31).

E. Heuristic-Based Models

Heuristic and metaheuristic algorithms, including genetic algorithms (GA), particle swarm optimization (PSO), and grey wolf optimization (GWO), are widely used to optimize the parameters and structure of forecasting models. These methods effectively address the nonlinear and dynamic characteristics of smart grid load data, improving forecasting accuracy and convergence rates. (32) Used genetic algorithm and particle swarm optimization to tune LSTM hyperparameters for load forecasting. (33) used Particle Swarm Optimization (PSO) while (34) used Beluga Whale Optimization (BWO).

F. Hybrid Methods

Hybrid models combine multiple forecasting techniques to exploit their complementary advantages and enhance prediction accuracy, making them a dominant trend in STLF research. Examples include the integration of K-Shape clustering with deep learning (35), the combination of ARIMA and Support Vector Regression optimized using Particle Swarm Optimization (PSO) (33), and the fusion of deep learning with Beluga Whale Optimization (BWO) (34). These hybrid frameworks generally outperform single-model approaches by improving adaptability, convergence, and robustness. Hybrid forecasting models are useful because electrical load is a mix of different patterns: long-term trends, seasonal cycles, and short-term nonlinear changes caused by weather and human behavior. Linear models such as ARIMA or SARIMA handle trends and seasonality well, but they cannot capture nonlinear fluctuations. Deep learning models, on the other hand, learn nonlinear patterns effectively but often struggle with long-term seasonal structures, especially when data is limited.

Hybrid methods combine the strengths of both approaches. The linear part of the load signal is first modeled using techniques such as SARIMA, STL decomposition, or exponential smoothing. The remaining nonlinear component is then learned by models like LSTM, CNN, or Transformers. By splitting the problem in this way, hybrid models reduce errors

from both linear misspecification and deep learning overfitting, which explains why they consistently achieve better results in many studies.

G. Comparative Theoretical Discussion of Model Effectiveness

A deeper theoretical examination reveals why different forecasting families excel under specific grid conditions. Classical statistical models such as ARIMA and SARIMA perform well when the load signal exhibits strong linear dependencies, stable seasonality, and weak nonlinear distortions. Their underlying assumptions of stationarity and additive structure make them suitable for short-term horizons in traditional grids with predictable patterns.

Machine learning models (e.g., SVR, Random Forests, and Gradient Boosting) relax linear constraints and can approximate nonlinear relationships between load, weather variables, and calendar attributes. Their performance improves with larger and more heterogeneous datasets, particularly when temperature sensitivity and occupancy-driven stochasticity introduce nonlinear load fluctuations.

Deep learning models further enhance representational capacity. Recurrent networks such as LSTM and GRU capture long-range temporal dependencies through gated memory structures, enabling superior performance in environments with multi-scale temporal correlations. CNN-based architectures extract local temporal motifs, while attention-based models and Transformers learn global dependencies by weighting relevant time steps, making them particularly robust when patterns shift or when long forecasting windows are required.

Hybrid models often outperform single models because they explicitly separate the load signal into components that align with each model's strengths. Seasonal and trend structures can be modeled linearly, while residual nonlinearities are captured by deep networks. This decomposition reduces both bias (from linear misspecification) and variance (from deep model overfitting), theoretically explaining the consistent performance gains reported in the literature.

VI. PRACTICAL DEPLOYMENT CONSIDERATIONS

Despite the methodological progress reviewed above, several practical factors affect the real-world deployment of short-term load forecasting systems.

A. Computational Complexity and Real-Time Constraints

Classical statistical models offer low computational overhead and are easily deployable on embedded devices. Machine learning models such as SVR and XGBoost introduce moderate training cost but remain feasible for real-time applications when appropriately tuned (36). In contrast, deep learning architectures—particularly LSTM, GRU, and Transformer-based models—require substantial computational resources due to sequential or attention-based operations. In distribution network control rooms where forecast updates must be delivered within seconds, latency constraints may restrict the use of computationally heavy models unless pruning, quantization, or edge acceleration techniques are employed (37).

B. Data Privacy and Security

Smart meter data contains sensitive information about consumer habits and occupancy patterns. Centralized training approaches may expose utilities to privacy risks. Privacy-preserving mechanisms such as federated learning, secure aggregation, and differential privacy offer promising alternatives by enabling distributed training without sharing raw data. These techniques also reduce the risk of data leakage and support regulatory compliance.

C. Edge–Cloud Coordination

Emerging architectures integrate local edge devices with cloud servers to balance latency, privacy, and computational requirements. Edge nodes perform lightweight preprocessing and short-horizon inference, while cloud platforms handle heavy model training and long-horizon prediction. Such hybrid deployment frameworks mitigate communication bottlenecks, reduce response time, and improve resilience in the presence of connectivity fluctuations.

These practical considerations highlight that forecasting accuracy alone is insufficient; deployment feasibility and operational constraints must guide model selection.

VII. CHALLENGES AND FUTURE DIRECTIONS

Short-term load forecasting continues to face several challenges that must be addressed to support next-generation smart grids. Data quality remains a major issue, as missing values, noise, and meter faults can significantly degrade model performance. At the same time, many deep learning models operate as “black boxes,” making it difficult for operators to trust or interpret predictions. Future work should therefore focus on explainable and interpretable forecasting models that provide clear insights into how predictions are generated.

Privacy is another growing concern, since smart meter data can reveal sensitive consumer behavior. Privacy-preserving learning frameworks—such as federated learning, differential privacy, and secure aggregation—offer promising solutions by enabling model training without sharing raw data.

Deployment constraints also shape future research. As utilities shift from cloud-based computation to real-time edge intelligence, models must become lighter and faster. Techniques such as pruning, quantization, and knowledge distillation can compress large deep networks for edge devices without sacrificing accuracy. Integrating these models with real-time demand response systems and microgrid controllers is an important direction.

Because electrical loads are strongly influenced by network topology and spatial dependencies, graph neural networks and spatiotemporal Transformers are emerging as powerful tools for regional forecasting. Additionally, forecasting must move beyond single-point estimates. Probabilistic and uncertainty-aware models—using Bayesian deep learning, quantile regression, or ensemble techniques—are essential for risk-aware grid operation.

Finally, many utilities lack extensive labeled datasets. Self-supervised learning, transfer learning, and hybrid cloud–edge

training pipelines can help overcome these limitations by leveraging unlabeled or cross-region data.

VIII. CONCLUSION

Short-Term Load Forecasting (STLF) remains an indispensable component of modern power system management, underpinning efficient grid operation, market participation, and renewable energy integration. This review traced the evolution of STLF techniques from classical statistical models to contemporary machine learning, deep learning, and hybrid frameworks. While statistical methods provide interpretability and stability, they often fall short in modeling nonlinear and stochastic load behavior. In contrast, AI-based and hybrid models demonstrate superior adaptability and predictive accuracy by leveraging data-driven learning and optimization mechanisms.

Despite significant progress, several challenges persist. These include data incompleteness, overfitting in deep models, limited interpretability, and the need for real-time deployment within edge and IoT-enabled environments. Future research should therefore prioritize the development of explainable and uncertainty-aware forecasting systems, capable of online learning and seamless integration into decentralized energy management architectures. Furthermore, incorporating probabilistic forecasting, physics-informed learning, and federated frameworks will be crucial for achieving scalable, secure, and privacy-preserving load prediction.

Ultimately, the convergence of data-driven intelligence, distributed computing, and energy-aware optimization represents the future of STLF research—enabling smarter, more sustainable, and resilient power systems within the energy internet era.

REFERENCES

- [1] H. Habbak, M. Mahmoud, K. Metwally, M. M. Fouda, and M. I. Ibrahim, “Load forecasting techniques and their applications in smart grids,” *Energies*, vol. 16, no. 3, p. 1480, 2023.
- [2] H. Habbak, M. Mahmoud, K. Metwally, M. Fouda, and M. I. Ibrahim, “Load forecasting techniques and their applications in smart grids,” *Energies*, 2023.
- [3] K. Ullah, M. Ahsan, S. M. Hasanat, M. Haris, H. Yousaf, S. F. Raza, R. Tandon, S. Abid, and Z. Ullah, “Short-term load forecasting: A comprehensive review and simulation study with cnn-lstm hybrids approach,” *IEEE Access*, 2024.
- [4] Y. Eren and İ. Küçükdemiral, “A comprehensive review on deep learning approaches for short-term load forecasting,” *Renewable and Sustainable Energy Reviews*, vol. 189, p. 114031, 2024.
- [5] N. Ahmad, Y. Ghadi, M. Adnan, and M. Ali, “Load forecasting techniques for power system: Research challenges and survey,” *IEEE Access*, vol. 10, pp. 71 054–71 090, 2022.
- [6] B. Wang, L. Wang, Y. Ma, D. Hou, W. Sun, and S. Li, “A short-term load forecasting method considering multiple factors based on var and ceemdan-cnn-bilstm,” *Energies*, 2025.
- [7] A. Salohub and A. Bosak, “Comparison of short-term power imbalance forecasting methods,” *POWER ENGINEERING: economics, technique, ecology*, 2025.
- [8] S. Chen, R. Lin, and W. Zeng, “Short-term load forecasting method based on arima and lstm,” *2022 IEEE 22nd International Conference on Communication Technology (ICCT)*, pp. 1913–1917, 2022.

- [9] P. C. Huy, N. Q. Minh, N. D. Tien, and T. T. Q. Anh, "Short-term electricity load forecasting based on temporal fusion transformer model," *IEEE Access*, vol. 10, pp. 106 296–106 304, 2022.
- [10] O. T. Tshipata, D. T. Kazumba, P. S. Nzakuna, V. Paciello, and A. K. Lusala, "Multi-horizon short-term electrical load forecasting: A comparative analysis of statistical models and deep neural networks," *2024 IEEE International Symposium on Measurements Networking (MN)*, pp. 1–6, 2024.
- [11] J. You, H. Cai, D. Shi, and L. Guo, "An improved short-term electricity load forecasting method: The vmd-kpca-xlstm-informer model," *Energies*, 2025.
- [12] L. Wu, Z. Yifan, W. Gang, and X. Hao, "A novel short-term load forecasting method based on mini-batch stochastic gradient descent regression model," *Electric Power Systems Research*, 2022.
- [13] Y. An, J. Wang, X. Zhao, Y. Huang, Z. Pang, and Z. He, "Research on short-term power load prediction model based on polynomial fitting," *2025 37th Chinese Control and Decision Conference (CCDC)*, pp. 1626–1629, 2025.
- [14] F. Gao, J. Song, and X. Shao, "Short-term interval-valued load forecasting with a combined strategy of ihw and multioutput machine learning," *Ann. Oper. Res.*, vol. 346, pp. 2009–2033, 2025.
- [15] A. Percuku, D. Minkovska, and N. Hinov, "Enhancing electricity load forecasting with machine learning and deep learning," *Technologies*, 2025.
- [16] L. Zhang and D. Jánošík, "Enhanced short-term load forecasting with hybrid machine learning models: Catboost and xgboost approaches," *Expert Syst. Appl.*, vol. 241, p. 122686, 2024.
- [17] Z. Wang, J. Duan, F. Luo, and X. Qiu, "Collaborative forecasting of multiple energy loads in integrated energy systems based on feature extraction and deep learning," *Energies*, 2025.
- [18] M. A. Abdalla, A. M. Ishaga, H. A. Osman, M. Elhindi, N. Ibrahim, A. Snani, G. H. A. Hamid, and A. Hammad, "Machine learning-based residential load demand forecasting: Evaluating elm, xgboost, rf, and svm for enhanced energy system and sustainability," *Science in Information Technology Letters*, 2025.
- [19] S. D. Wankhade and B. R. Patil, "Powering the future of electrical load forecasting using a regression learner in machine learning," *International Journal of Applied Power Engineering (IJAPE)*, 2025.
- [20] C. Zhang, Y. Yu, T. Zhang, K. Song, Y. Wang, and S. Gao, "Short-term load forecasting based on graph convolution and dendritic deep learning," *IEEE Transactions on Network Science and Engineering*, vol. 12, pp. 3221–3233, 2025.
- [21] Y. Eren and I. Küçükdemir, "A comprehensive review on deep learning approaches for short-term load forecasting," *Renewable and Sustainable Energy Reviews*, 2024.
- [22] P. P. G. A. R., G. M., N. K. M., and B. S. S., "Electrical load demand forecasting using machine learning algorithm," *2025 Second International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS)*, pp. 1–6, 2025.
- [23] M. E., N. R., and R. Satheesh, "Deep learning-based load forecasting in smart grid," *2025 Emerging Technologies for Intelligent Systems (ETIS)*, pp. 1–5, 2025.
- [24] G.-F. Fan, Y.-Y. Han, J.-W. Li, L. ling Peng, Y.-H. Yeh, and W. Hong, "A hybrid model for deep learning short-term power load forecasting based on feature extraction statistics techniques," *Expert Syst. Appl.*, vol. 238, p. 122012, 2024.
- [25] W. Liu, Z.-G. Tian, and Y. Qiu, "Enhancing accuracy in point-interval load forecasting: A new strategy based on data augmentation, customized deep learning, and weighted linear error correction," *Expert Syst. Appl.*, vol. 272, p. 126686, 2025.
- [26] Z. Tang, T. Ji, J. Kang, Y. Huang, and W. Tang, "Learning global and local features of power load series through transformer and 2d-cnn: An image-based multi-step forecasting approach incorporating phase space reconstruction," *Applied Energy*, 2025.
- [27] C. M. Dean, X. Ma, T. Ding, K. Xu, Y. Chen, and H. Mo, "A study of machine learning model's performance for load forecasting based on xai techniques," *2025 17th International Conference on Computer and Automation Engineering (IC-CAE)*, pp. 227–231, 2025.
- [28] W. Liu, Z. Mao, and L. Jiang, "A cloud-edge collaborative framework for power load forecasting using deep learning," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–12, 2025.
- [29] M. Iqbal and M. Adnan, "Edge computing and transfer learning-based short-term load forecasting for residential and commercial buildings," *Energy and Buildings*, 2025.
- [30] H.-J. Bae, J.-S. Park, J. hyeok Choi, and H.-Y. Kwon, "Learning model combined with data clustering and dimensionality reduction for short-term electricity load forecasting," *Scientific Reports*, vol. 15, 2025.
- [31] F. Liu, T. Dong, Q. Liu, Y. Liu, and S. Li, "Combining fuzzy clustering and improved long short-term memory neural networks for short-term load forecasting," *Electric Power Systems Research*, 2024.
- [32] S. Bouktif, A. Fiaz, A. Ouni, and M. Serhani, "Multi-sequence lstm-rnn deep learning and metaheuristics for electric load forecasting," *Energies*, 2020.
- [33] W. Zhao, H. Xu, P. Chen, J. Zhang, J. Li, and T. Cai, "Elastic momentum-enhanced adaptive hybrid method for short-term load forecasting," *Energies*, vol. 18, no. 13, p. 3263, 2025.
- [34] M. M. Asiri, G. Aldehim, F. Alotaibi, M. M. Alnfai, M. Assiri, and A. Mahmud, "Short-term load forecasting in smart grids using hybrid deep learning," *IEEE Access*, vol. 12, pp. 23 504–23 513, 2024.
- [35] W. Zhang, M. Cheng, Q. Xiang, and Q. Li, "Enhancing short-term load forecasting through k-shape clustering and deep learning integration," *IEEE Access*, vol. 13, pp. 30 817–30 832, 2025.
- [36] M. Souran and R. S. Shah, "Finance forecasting using financial indicators with advanced machine learning on end-user mobile devices," *2025 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE)*, pp. 1–6, 2025.
- [37] S. Khan, M. A. B. Arif, Baloch, D. K. Mal, F. Khan, S. Ali, M. Abdullah, and B. Arif, "Adaptive load forecasting in smart grids using deep learning and edge computing," *Annual Methodological Archive Research Review*, 2025.

Adaptive Neural Linear-Based MPPT for Enhanced Efficiency in Photovoltaic Systems

Yacine Triki^{1,2}, Ali Bechouche², Hamid Seddiki² and Djaffar Ould Abdeslam³

¹ Department of Electrical Systems Engineering, University M'Hamed Bougara, Boumerdes, Algeria

² L2CSP Laboratory, Mouloud Mammeri University, BP 17 RP 15000 Tizi-Ouzou, Algeria

³ IRIMAS Institute, University of Haute Alsace, 61 rue Albert Camus, F-68093 Mulhouse, France

E-mail: y.triki@univ-boumerdes.dz; ali.bechouche@ummto.dz; hamid.seddiki@ummto.dz; djaffar.ould-abdeslam@uha.fr

Abstract— This paper presents a maximum power point tracking (MPPT) strategy for standalone photovoltaic (PV) systems based on an adaptive neural linear (ANL) model. The proposed ANL-based MPPT (ANL-MPPT) leverages the simplicity and fast learning dynamics of ANL, making it particularly suitable for real-time power optimization. The method operates by continuously regulating the boost converter's input voltage to track the PV panel's optimum voltage, which is estimated from the panel's instantaneous resistance. This approach ensures sustained operation at the maximum power point despite fluctuations in irradiance or temperature. The proposed technique is evaluated against the perturb and observe algorithm under identical test scenarios, including the standardized irradiance profiles defined by the EN 50530 standard. Simulation results clearly demonstrate the superior responsiveness and tracking accuracy of the ANL-MPPT, especially under fast irradiance variations and low-light conditions.

Keywords— *adaptive neural linear, EN 50530 standard, maximum power point tracking (MPPT), perturb and observe (P&O) algorithm*

I. INTRODUCTION

Solar energy, widely accessible across most regions of the globe, is regarded as one of the most promising renewable resources. Over the last few decades, electricity generation through photovoltaic (PV) systems has seen significant growth [1]. Recent technological advances have enabled solar energy to be exploited for multiple purposes, particularly for power generation [2]. When a PV module is directly connected to a load, its operating point is defined by the intersection of the load curve and the I-V characteristic. In most cases, this point does not match the maximum power point (MPP). To address this issue, maximum power point tracking (MPPT) techniques are implemented to ensure that the PV system operates at its MPP [3]–[14].

The literature features a wide array of MPPT algorithms, all designed to maximize energy harvest but differing significantly in their performance. Key distinguishing factors include their convergence speed, steady-state oscillations, implementation cost, sensor requirements, and overall effectiveness. These algorithms are broadly categorized into conventional methods and modern soft computing techniques. Among conventional methods, the perturb and observe (P&O) [3], [4] and incremental conductance [5] algorithms are the most prevalent. They operate on the principle that the slope of the photovoltaic array's power-voltage (P-V) curve is zero at the MPP [4]. Other conventional approaches include those based on open-circuit voltage [6] and short-circuit current [7] measurements.

The primary advantages of conventional methods are their straightforward implementation and compatibility with various photovoltaic systems, yielding satisfactory performance under stable environmental conditions. However, their major drawback is a significant decline in tracking efficiency when conditions change rapidly. In such scenarios, they often fail to maintain operation at the MPP, particularly if the rate of change in solar irradiance exceeds the algorithm's perturbation step size [4].

To improve classical methods, researchers have turned their attention to sophisticated soft computing methods, which have been widely adopted for MPPT applications. While these approaches share similarities with conventional MPPT techniques, they generally integrate artificial intelligence tools. Within this group, the most prominent strategies rely on artificial neural networks (ANNs) [8], [9], genetic algorithms [10], fuzzy logic [3], [11], particle swarm optimization [12]. Among them, ANNs are the most extensively applied AI technique, mainly due to their capacity to learn and enhance performance from prior experiences. This adaptability provides ANNs with notable benefits that can substantially improve PV system efficiency, though their effectiveness is often influenced by the complexity of the underlying implementation.

This paper presents an adaptive neural linear (ANL)-based MPPT strategy specifically designed for standalone PV systems. The proposed ANL-MPPT approach is founded on the adaptive linear neuron (ADALINE) [13], [14], whose simple and efficient structure constitutes the core of the tracking mechanism. The choice of ADALINE is motivated by its simple architecture and fast convergence properties, making it well suited for real-time maximum power tracking. The ANL-MPPT operating principle is based on adjusting the input voltage of the boost converter so that it matches the optimal output voltage of the PV panel. This optimal voltage is determined from the instantaneous resistance of the PV panel. By forcing the converter to operate at this reference voltage, the system ensures that the PV module remains at its MPP, regardless of variations in irradiance or temperature. To validate the method's effectiveness, a comparative analysis with the conventional P&O algorithm was conducted under identical conditions. The evaluation employed various irradiance profiles specified by the rigorous European Standard EN 50530 [15], which assesses MPPT performance using ramp sequences of different gradients and irradiance levels. Simulation results demonstrate the superior performance of the ADALINE-based method. Specifically, it exhibited excellent tracking characteristics and a stronger ability to adapt to dynamic changes in solar irradiance, outperforming the P&O

algorithm particularly during rapid irradiance shifts and under low-irradiance conditions.

The paper is organized into six main sections. Following the introduction, Section II presents a general overview of the controlled PV systems. Section III details the modeling of both the PV module and the DC–DC boost converter. Section IV addresses the conventional P&O method along with the proposed ANL-MPPT approach. Section V discusses the test scenarios and analyzes the corresponding simulation results. Finally, the conclusion is provided in Section VI.

II. PV SYSTEM DESCRIPTION

Figure 1 presents the architecture of the controlled stand-alone PV system. The setup includes a KC200GT PV module, a DC–DC boost converter, an MPPT controller, and a resistive load. The control strategy is centered on the proposed ANL-MPPT algorithm, which continuously determines the optimal duty cycle required for maximum power extraction. By dynamically regulating the boost converter, the controller forces the PV module to operate at its MPP, ensuring efficient energy harvest even under rapid variations in irradiance and temperature.

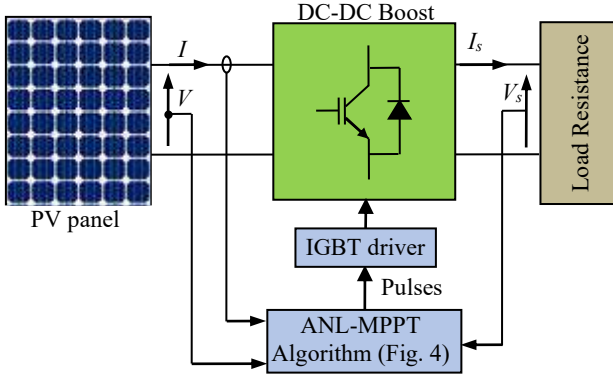


Fig. 1. System architecture of the controlled stand-alone PV system.

III. PV SYSTEM MODELLING

A. PV Panel Modelling

The PV module converts solar radiation directly into electrical energy. Several equivalent circuit models have been proposed in the literature to represent PV cell behavior. In this work, the single-diode model is considered. This model includes a current source I_{pv} , proportional to the incident irradiance, a diode, a series resistance R_s , and a shunt resistance R_p . In addition, the reverse saturation current I_o and the diode ideality factor a are also essential parameters of the model.

The current–voltage (I–V) characteristic of a PV cell, can be expressed mathematically as in equation (1).

$$I = I_{pv} - I_o \left[\exp \left(\frac{q(V + R_s I)}{a k T} \right) - 1 \right] - \frac{V + R_s I}{R_p} \quad (1)$$

In this relation, q denotes the electron charge, while k refers to the Boltzmann constant.

The electrical specifications of the PV module employed in this study are presented in Table I [16].

TABLE I
PARAMETERS OF THE KC200GT PV PANEL
UNDER NOMINAL CONDITIONS (25 °C AND 1000 W/M²)

Current at the MPP I_{mp} (A)	7.61
Voltage at the MPP V_{mp} (V)	26.3
Maximal power P_{max} (W)	200.143
Short-circuit current I_{sc} (A)	8.21
Open-circuit voltage V_{oc} (V)	32.9
Voltage coefficient K_v (V/K)	−0.1230
Current coefficient K_i (A/K)	0.0032
Cell number N_s	54

B. The DC-DC Boost Converter

In PV power generation systems, the DC–DC boost converter is commonly employed as an interface stage between the PV module and the load. Its role is to act as a switching-mode regulator, ensuring MPPT while converting the typically unregulated PV output voltage into a stable, regulated one. The regulation is generally achieved using a pulse-width modulation (PWM) strategy with a constant switching frequency [2]. The structure of the boost converter used in this work is illustrated in Fig. 3. For high switching frequencies, the average model of the DC–DC boost converter can be expressed as:

$$\frac{dI}{dt} = \frac{V}{L} - \frac{V_s}{L}(1-\alpha) \quad (2)$$

$$\frac{dV_s}{dt} = \frac{I}{C}(1-\alpha) - \frac{V_s}{RC} \quad (3)$$

where V represents the boost input voltage, V_s the boost output voltage, α the duty cycle, and R the load resistance.

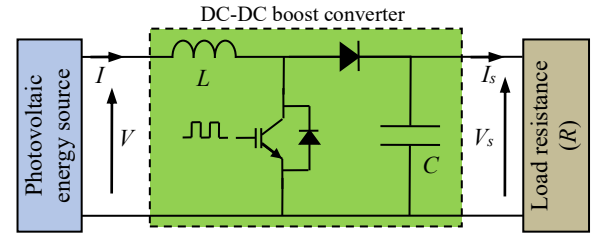


Fig. 2. Power circuit of the DC-DC boost converter.

IV. MAXIMUM POWER TRACKING

The power–voltage (P–V) curve of a PV module is nonlinear and presents a single MPP. This point varies with environmental factors such as irradiance, temperature, and the connected load. Since irradiation and temperature fluctuate during the day, the MPP location also shifts, making it necessary to employ a dynamic tracking technique to continuously extract the available maximum power. To achieve this, MPPT algorithms are implemented, providing in real time a suitable reference signal for the DC–DC converter that corresponds to the MPP operating point. In the following, the principle of P&O algorithm, one of the most widely used MPPT methods, is presented. In this work, the P&O serves as a benchmark for comparison with the proposed ANL-MPPT strategy.

A. P&O Algorithm

The P&O method is one of the most widely used MPPT techniques due to its simplicity, ease of implementation, and independence from prior PV system data or environmental measurements. Its principle relies on the fact that the slope of

the P-V curve becomes zero at the MPP [3], [4]. Fig. 4 presented the P&O algorithm flowchart.

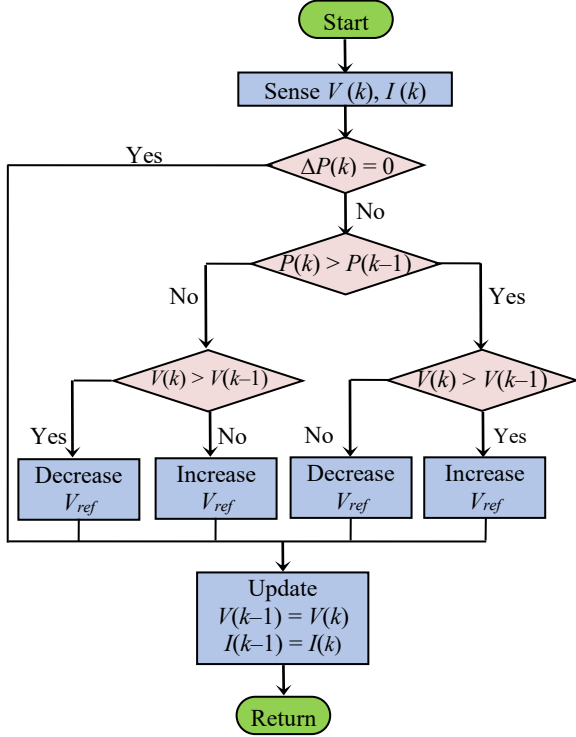


Fig. 3. Flowchart of the P&O algorithm.

B. The Proposed ANL-MPPT Algorithm

The operating principle of the ANL-MPPT consists in regulating the input voltage of the boost converter to track the PV panel's optimal output voltage. This optimal voltage is calculated based on the instantaneous resistance of the PV panel, which is defined as follows:

$$\frac{V}{I} = -\frac{dV}{dI} \quad (4)$$

At the maximum power point, the PV module's optimal output voltage can be written as in (5).

$$V = -I \frac{dV}{dI} \quad (5)$$

Similarly, using the relationship given in (2), the input voltage of the boost converter is expressed by:

$$V = L \frac{dI}{dt} + V_s(1 - \alpha) \quad (6)$$

Therefore, the system operates at the MPP when the expressions in (5) and (6) become identical, leading to:

$$-I \frac{dV}{dI} = L \frac{dI}{dt} + V_s(1 - \alpha) \quad (7)$$

In a steady-state condition, the term LdI/dt equals zero. Furthermore, since changes in solar irradiance are typically gradual, this term remains negligible even during transients. Based on this assumption and referring to Equation (7), the discrete form of the optimal duty cycle required to reach the maximum power point can be expressed as follows:

$$\alpha(k) = \frac{\Delta V(k)I(k)}{V_s(k)\Delta I(k)} + 1 \quad (8)$$

where $\Delta V(k) = V(k) - V(k-1)$ and $\Delta I(k) = I(k) - I(k-1)$ are very small changes in the PV array output voltage and current, respectively.

At this point, the ADALINE model [13] is employed to solve equation (8) in the following manner.

$$\frac{\Delta V(k)I(k)}{V_s(k)\Delta I(k)} + 1 = y_d(k) \quad (9)$$

$$\alpha(k) = w(k) = y_{est}(k) \quad (10)$$

$$x(k) = 1 \quad (11)$$

Thus, following the ADALINE principle, the estimated output $y_{est}(k)$ is continuously compared with the desired value $y_d(k)$. The resulting error $e(k)$ ($e(k) = y_d(k) - y_{est}(k)$) is then used to adjust the ADALINE weight $w(k)$ so that the estimation progressively approaches the target output. In this study, the weight adaptation is performed using the α -LMS learning rule [13], defined as:

$$w(k+1) = w(k) + 2\mu e(k)x(k) \quad (12)$$

where μ denotes the ADALINE learning rate, which governs both the convergence speed and the stability of the adaptation process [13], [14].

In this work, the ANL-MPPT is implemented using $\mu = 0.03$, as this value provides an effective compromise between fast dynamic response and reduced steady-state oscillations. Using equations (9)–(12), the ANL-MPPT formulation is obtained as follows:

$$\alpha(k+1) = \alpha(k) + 2\mu \left[\frac{\Delta V(k)I(k)}{V_s(k)\Delta I(k)} + 1 - \alpha(k) \right] \quad (13)$$

The block diagram of the proposed ANL-MPPT scheme is presented in Fig. 4.

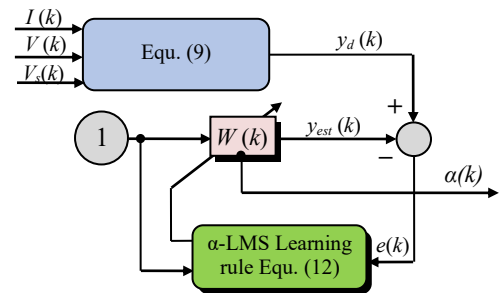


Fig. 4. Block diagram of the proposed ANL-MPPT strategy.

V. SIMULATIONS RESULTS

To validate the performance of the proposed iterative ANL-MPPT algorithm, the standalone PV system from Fig. 1 is modeled in MATLAB/Simulink. Simulation tests were performed using variable irradiance profiles defined by the rigorous prEN 50530 standard, which evaluates MPPT efficiency in both steady-state and dynamic scenarios [15].

As shown in Fig. 5, six test profiles were used. These profiles covered high (300–1000 W/m²), medium (100–500

W/m²), and low (10–100 W/m²) irradiance levels. At each level, the is tested with both fast (Figs. 5(a), (c) and (e)) and slow (Figs. 5(b), (d) and (f)) irradiance dynamics, providing a comprehensive assessment of its tracking capabilities. A constant temperature of 25°C was maintained for all simulations.

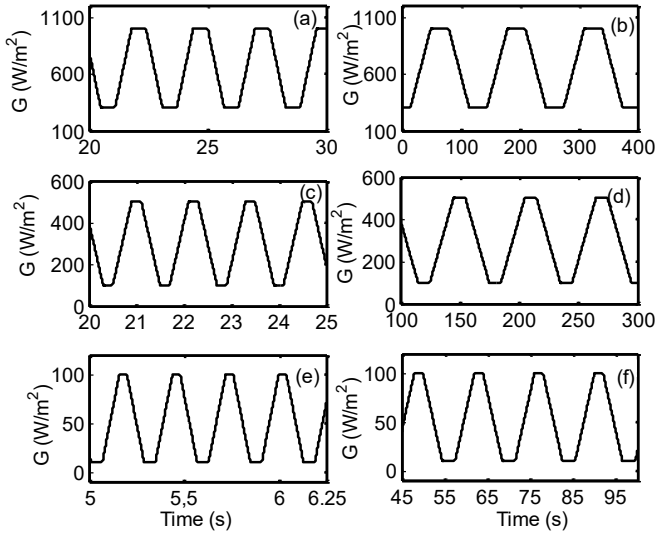


Fig. 5. Irradiance profiles: (a) high level with fast dynamic; (b) high level with slow dynamic; (c) medium level with fast dynamic; (d) medium level with slow dynamic; (e) low level with fast dynamic; and (f) low level with slow dynamic.

A. High irradiance profiles

Figure 6 presents the tracking behavior of the proposed ANL-MPPT compared with the classical P&O algorithm under the high-irradiance fast-changing profile (Fig. 5(a)). As shown in Fig. 6(a), the power extracted using the ANL-MPPT remains stable and free of oscillations. In contrast, the P&O method exhibits pronounced power spikes due to the rapid variation of irradiance, which leads to a noticeable loss of harvested energy.

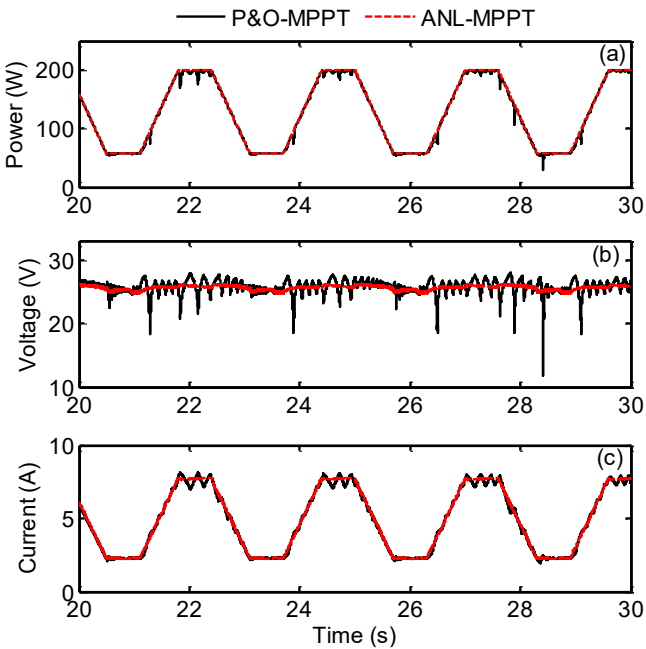


Fig. 6. Comparison of the tracking performance of the proposed ANL-MPPT and the P&O-MPPT algorithms under a high-irradiance fast-changing profile: (a) power; (b) voltage; and (c) current.

The corresponding PV voltage and current waveforms are provided in Figs. 6(b) and 6(c). With the ANL-MPPT, both signals promptly reach their correct values and show negligible oscillations. The P&O algorithm, however, produces highly oscillatory voltage and current responses, with significant instability—particularly visible in the voltage signal.

Figure 7 reports the results obtained under the high-irradiance slow-varying profile (Fig. 5(b)). Under these smoother operating conditions, both P&O and ANL-MPPT provide comparable tracking performance. Nevertheless, the proposed ANL-MPPT still achieves slightly more stable voltage (Fig. 7(b)) and current (Fig. 7(c)) responses, showing reduced oscillations compared to the P&O method.

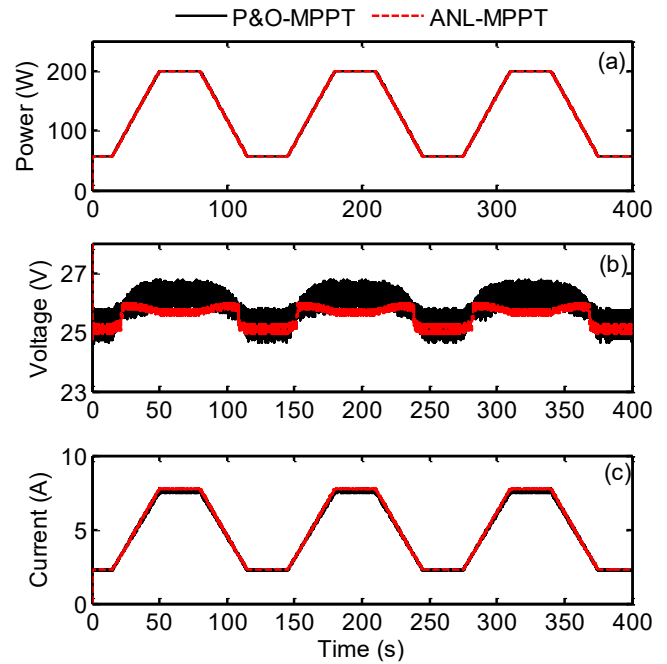


Fig. 7. Comparison of the tracking performance of the proposed ANL-MPPT and the P&O-MPPT algorithms under a high-irradiance slow-changing profile: (a) power; (b) voltage; and (c) current.

B. Medium irradiance profiles

Figure 8 presents the tracking performance of the proposed ANL-MPPT compared with the classical P&O algorithm under the medium-irradiance fast-changing profile (Fig. 5(c)). The results are generally consistent with those obtained for the high-irradiance fast profile: as seen in Fig. 8(a), the ANL-MPPT maintains a stable, smooth power output without oscillations. However, under medium irradiance, the P&O algorithm exhibits even larger power spikes than in the high-irradiance case, leading to greater energy losses due to rapid irradiance fluctuations.

The corresponding voltage and current waveforms are shown in Figs. 8(b) and 8(c), respectively. As before, the ANL-MPPT ensures fast convergence and nearly oscillation-free behavior. In contrast, the P&O method produces highly oscillatory and unstable voltage and current responses, with especially pronounced disturbances in the voltage signal.

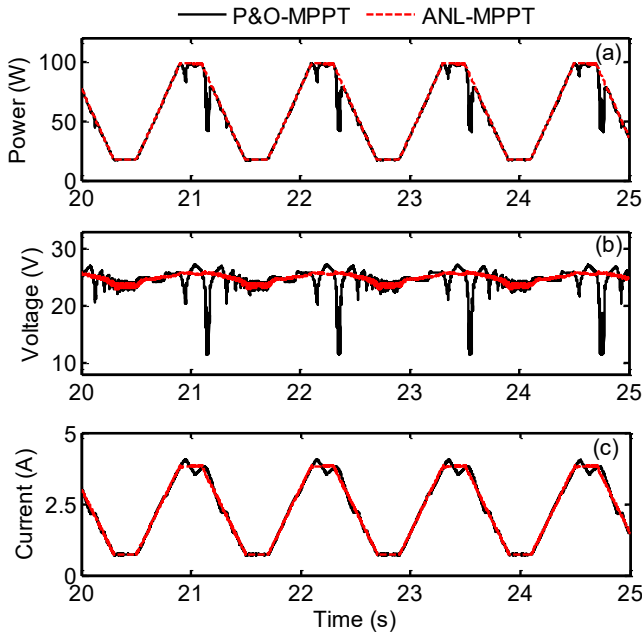


Fig. 8. Comparison of the tracking performance of the proposed ANL-MPPT and the P&O-MPPT algorithms under a medium-irradiance fast-changing profile: (a) power; (b) voltage; and (c) current.

Figure 9 shows the tracking performance of both MPPT methods under the medium-irradiance slow-changing profile. In this case, the two algorithms provide almost similar overall behavior, as the irradiance varies slowly. However, the proposed ANL-MPPT maintains a fully stable response with no noticeable oscillations, while the P&O method still exhibits some residual oscillatory movements in both voltage and current. These fluctuations, although moderate, remain visible compared with the smooth and steady waveforms obtained using the ANL-MPPT.

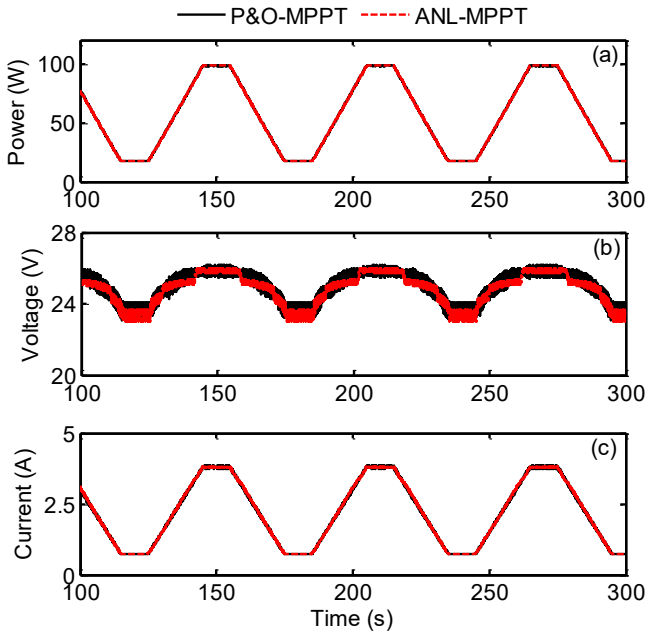


Fig. 9. Comparison of the tracking performance of the proposed ANL-MPPT and the P&O-MPPT algorithms under a medium-irradiance slow-changing profile: (a) power; (b) voltage; and (c) current.

C. Low irradiance profiles

Figure 10 depicts the tracking performance of the ANL-MPPT and P&O algorithms under a low-irradiance, fast-changing profile. Under these challenging conditions, the P&O method fails to properly follow the maximum power, with the power, voltage, and current signals exhibiting significant spikes and oscillations, which can negatively affect system performance. In contrast, the proposed ANL-MPPT accurately tracks the irradiance variations, maintaining maximum power extraction without any oscillations. The corresponding voltage, current, and power signals remain stable, demonstrating the robustness of the ANL-MPPT approach even under low irradiation and rapidly changing conditions.

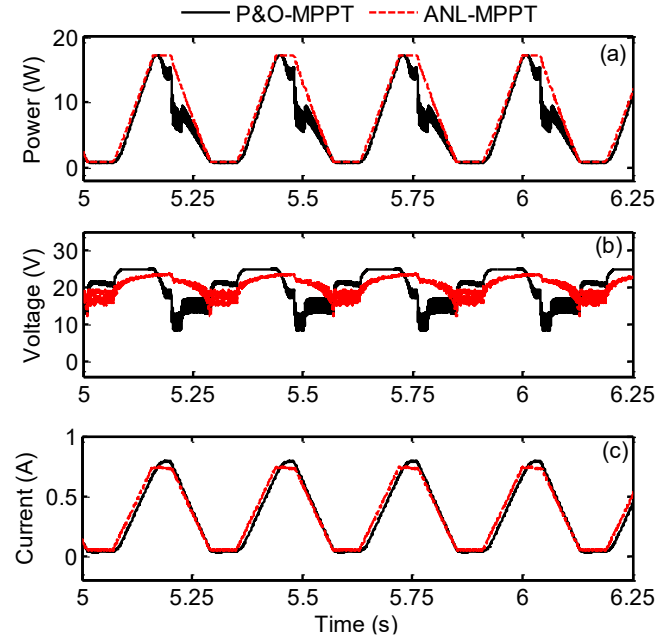


Fig. 10. Comparison of the tracking performance of the proposed ANL-MPPT and the P&O-MPPT algorithms under a low-irradiance fast-changing profile: (a) power; (b) voltage; and (c) current.

Figure 11 illustrates the tracking performance of the ANL-MPPT and P&O algorithms under the low-irradiance, slow-changing profile. The proposed ANL-MPPT accurately follows the irradiance variations, ensuring maximum power extraction without any oscillations. In contrast, the P&O method initially fails to track the optimal operating point, showing a delayed response before eventually converging. A similar behavior is observed for both the voltage and current waveforms, where the P&O exhibits noticeable deviations at the beginning, while the ANL-MPPT maintains smooth and stable signals throughout the test.

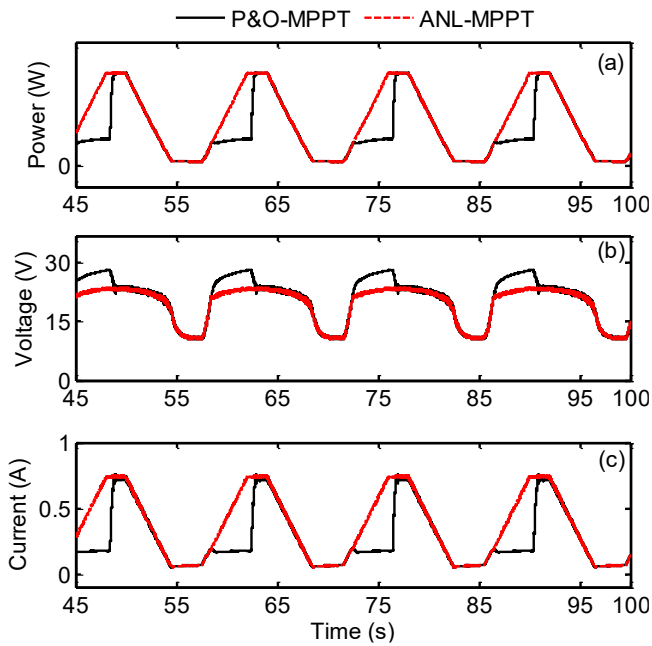


Fig. 11. Comparison of the tracking performance of the proposed ANL-MPPT and the P&O-MPPT algorithms under a low-irradiance slow-changing profile: (a) power; (b) voltage; and (c) current.

VI. CONCLUSION

This paper presented an ANL-MPPT scheme for standalone PV systems. The technique leverages the ADALINE, capitalizing on its computational simplicity and rapid convergence to regulate a boost converter's input voltage. By matching this voltage to the PV panel's optimal operating point, determined through its instantaneous resistance. This strategy ensures continuous operation at the MPP under varying environmental conditions, including fluctuations in irradiance and temperature. The performance of the ANL-MPPT method was evaluated against the P&O algorithm using the rigorous EN 50530 standard test profiles. Simulation results confirmed that the ANL-MPPT outperforms the P&O method, showing smoother power, voltage, and current waveforms, faster tracking response, and reduced oscillations. Notably, the ANL-MPPT algorithm sustained high performance during low-irradiance conditions and under rapid atmospheric changes, highlighting its practical robustness for photovoltaic applications. In conclusion, the ANL-MPPT algorithm provides an effective, stable, and high-performance solution for real-time maximum power extraction, offering a significant improvement over classical MPPT techniques, proving particularly effective in challenging and dynamically shifting operational environments.

REFERENCES

- [1] International Energy Agency (2024) Renewables 2024. IEA, Paris. <https://www.iea.org/reports/renewables-2024>. Accessed 18 Nov. 2025.
- [2] Teodorescu R, Liserre M, Rodriguez P (2011) Grid converters for photovoltaic and wind power systems. Wiley-IEEE Press, Hoboken.
- [3] H. Karimi, A. Siadatan and A. Rezaei-Zare, "A Hybrid P&O-Fuzzy-Based Maximum Power Point Tracking (MPPT) Algorithm for Photovoltaic Systems Under Partial Shading Conditions," in IEEE Access, vol. 13, pp. 86046-86056, 2025.
- [4] A. I. M. Ali et al., "An Enhanced P&O MPPT Algorithm With Concise Search Area for Grid-Tied PV Systems," in IEEE Access, vol. 11, pp. 79408-79421, 2023.
- [5] L. Shang, H. Guo and W. Zhu, "An Improved MPPT Control Strategy Based on Incremental Conductance Algorithm," in Protection and Control of Modern Power Systems, vol. 5, no. 2, pp. 1-8, 2020.
- [6] M. J. Alshareef, "An Enhanced Fractional Open Circuit Voltage MPPT Method for Rapid and Precise MPP Tracking in Standalone Photovoltaic Systems," in IEEE Access, vol. 13, pp. 34115-34131, 2025.
- [7] H. A. Sher, A. F. Murtaza, A. Noman, K. E. Addoweesh, K. Al-Haddad, and M. Chiaberge, "A New Sensorless hybrid MPPT algorithm based on fractional short-circuit current measurement and P&O MPPT," IEEE Trans. Sustain. Energy, vol. 6, no. 4, pp. 1426–1434, 2015.
- [8] N. F. Ibrahim et al., "Operation of Grid-Connected PV System With ANN-Based MPPT and an Optimized LCL Filter Using GRG Algorithm for Enhanced Power Quality," in IEEE Access, vol. 11, pp. 106859-106876, 2023.
- [9] M. N. Ali, K. Mahmoud, M. Lehtonen, and M. M. F. Darwish, "Promising MPPT Methods Combining Metaheuristic, Fuzzy-Logic and ANN Techniques for Grid-Connected Photovoltaic," Sensors, vol. 21, no. 4, p. 1244, 2021.
- [10] A. A. S. Mohamed et al., "Design and hardware implementation of FL-MPPT control of PV systems based on GA and small-signal analysis," IEEE Trans. Sustain. Energy, vol. 8, no. 1, pp. 279–290, 2017.
- [11] T. Hassan, R. Abbassi, H. Jerbi, et al., "A Novel Algorithm for MPPT of an Isolated PV System Using Push Pull Converter with Fuzzy Logic Controller," Energies, vol. 13, no. 15, 4007, 2020.
- [12] A. F. Mirza, M. Mansoor, Q. Ling, B. Yin, and M. Y. Javed, "A Salp-Swarm Optimization based MPPT technique for harvesting maximum energy from PV systems under partial shading conditions," Energy Conversion and Management, 209, 112625, 2020.
- [13] Widrow B, Hoff ME (1960) Adaptive switching circuits. IRE WESCON Conv Rec 4:96–104.
- [14] Haykin S (2009) Neural networks and learning machines, 3rd edn. Pearson Education, Upper Saddle River.
- [15] R. Bründlinger, N. Henze, H. Häberlin, B. Burger, A. Bergmann, and F. Baumgartner, "prEN 50530—The new European standard for performance characterisation of PV inverters," in 24th EU PV Conf., 2009, Hamburg, Germany.
- [16] Kyocera KC200GT PV Module Datasheet. [Online]. Available : <https://www.energymatters.com.au/images/kyocera/KC200GT.pdf>. Accessed 18 Nov. 2025.

Real-Time Energy Routing in Smart Grids with Harmony Search Algorithm

Lina Benchikh

*LRSD laboratory, Faculty of Science
University Ferhat Abbas Setif 1
Setif, Algeria
lina.benchikh@univ-setif.dz*

Assala Nacef

*LRSD laboratory, Faculty of Science
University Ferhat Abbas Setif 1
Setif, Algeria
assala.nacef@univ-setif.dz*

Rahma Berchi

*LRSD laboratory, Faculty of Science
University Ferhat Abbas Setif 1
Setif, Algeria
rahma.berchi@univ-setif.dz*

Lemia Louail

*Université de Lorraine, CNRS, CRAN
F-54000 Nancy, France
lemia.louail@univ-lorraine.fr*

Djihad Arrar

*LRSD laboratory, Faculty of Science
University Ferhat Abbas Setif 1
Setif, Algeria
djihad.arrar@univ-setif.dz*

Abstract—With the growing decentralization of energy systems—driven by concepts such as the Energy Internet and smart grids—local renewable generation is becoming increasingly important. While this transition supports sustainability goals, it also introduces new challenges in routing energy efficiently within decentralized networks. To address these challenges, we propose the Harmony Search-based Energy Routing Protocol (HS-ERP), which applies the Harmony Search algorithm to optimize producer selection and routing decisions. In this paper, we implement HS-ERP in a real-time framework to demonstrate its practical applicability in dynamic energy environments. We evaluate its performance through extensive experiments and compare it with established methods such as the Genetic Algorithm. The results, obtained across different network sizes, show that HS-ERP is both adaptable and scalable.

Index Terms—Smart Grids, Energy Internet, Path Selection, Prosumer Matching, Harmony Search, Graph Theory

I. INTRODUCTION

The Energy Internet (EI) and smart grids (SG) represent a shift from centralized power distribution to a decentralized and distributed energy paradigm [1]. In traditional systems, electricity is generated in large power plants and transmitted over long distances to consumers [2]. In contrast, the EI envisions a future where energy production is more local and diversified, relying on renewable resources such as rooftop solar panels and small-scale generators. This enables homes and communities to act as prosumers—both producers and consumers of energy [3]. Such decentralization improves energy resilience and reduces carbon emissions through increased use of clean energy sources [4], [5].

However, this transition introduces new challenges, particularly in energy routing and the selection of optimal producers within a dynamic and distributed environment. A key task is the producer subset selection problem, which aims to identify the most suitable set of producers to satisfy a given demand while minimizing transmission losses and operational costs. This problem is complex because it depends on factors such as the availability of energy, the efficiency of transmission paths,

and the characteristics of the underlying network. Addressing these challenges is essential to ensure reliable and efficient operation in decentralized energy systems.

To tackle these routing issues, we integrate the Harmony Search (HS) algorithm with a brute-force path exploration method. The HS algorithm is employed to identify the best producer combinations, while the brute-force component evaluates all feasible routing paths to determine the most energy-efficient ones. In this paper, we implement our method within a real-time framework to demonstrate its practical relevance in dynamic energy settings. We further evaluate the performance of the proposed Harmony Search-based Energy Routing Protocol (HS-ERP) through extensive simulations and compare it with the genetic-based approach in [6]. Our experiments include both small and large network topologies, highlighting the scalability and robustness of the proposed solution.

The remainder of this paper is organized as follows. Section ?? surveys related works. Section III presents the details of the proposed HS-ERP method. Section IV describes the network model and performance evaluation. Finally, Section V concludes the paper.

II. RELATED WORKS

To select the best energy source, both cost and price need to be considered. Several studies in the literature address this issue using different methods and algorithms. In [7], the authors introduced the firefly algorithm, a bio-inspired approach for solving the optimal producer subset problem. This algorithm mimics the behavior of fireflies, where individuals are drawn to the brightest light, with consumers being attracted to producers offering the lowest fitness function.

In [8], the PSO-ERP method uses Particle Swarm Optimization (PSO) to solve the pairs selection problem and find optimal paths between producers and consumers in smart grids. Inspired by swarm behavior, PSO-ERP simulates particle movements to explore potential routes.

TABLE I
COMPARISON OF EXISTING ENERGY ROUTING METHODS

Method	Strengths	Limitations
Firefly-ERP [7]	Simple bio-inspired optimization; good for small search spaces	Distance-based fitness only; does not model power loss or router properties; limited scalability
PSO-ERP [8]	Fast convergence; swarm intelligence suitable for multi-route exploration	Evaluates routes mainly by distance; lacks detailed loss modeling; no real-time adaptation
Greedy + Dijkstra [9]	Very fast; suitable for small residential networks; easy to implement	Greedy producer selection can miss optimal subsets; only distance-based routing; not suitable for large grids
GA-ERP [6]	Optimizes both producer subset and transmission paths; incorporates power loss and router characteristics	No real-time execution; evaluates feasible and non-feasible producers; higher computation time for large networks
HS-ERP (Proposed)	Real-time execution; integrates producer selection with detailed loss modeling; filters infeasible producers using interval constraints; scalable to larger topologies	Requires brute-force path evaluation; performance depends on HS parameter tuning

In [9], a novel approach is presented to tackle the subset selection and efficient path problems in small residential areas using a greedy algorithm. The problem is framed similarly to a fractional knapsack problem, with the goal of maximizing energy value while adhering to capacity limitations. The authors employ Dijkstra's algorithm to determine optimal routes based on distance.

While these algorithms rely on distance-based metrics, they do not fully capture power loss, which is influenced by factors beyond distance, such as the properties of power links and energy routers. To address this limitation, other studies have focused on solutions that take into account these critical properties.

In [6], the authors use a Genetic Algorithm (GA) and the Path Generator to optimize energy routing. The GA selects the best group of producers by evolving solutions through crossover and mutation. Afterward, the Path Generator calculates energy losses for all possible transmission routes, identifying the most efficient path for energy transfer. This approach optimizes both producer selection and routing for improved energy distribution. However, the authors do not account for real-time execution; when a new consumer arrives, the system does not take real-time factors into consideration; it selects the prosumer based solely on loss and cost.

III. HARMONY SEARCH BASED ENERGY ROUTING APPROACH

In this section, we will address the energy routing problems by employing the HS-ERP approach.

A. Harmony search

The Harmony Search algorithm, developed by Geem (2001), is inspired by the improvisation process of musicians, particularly in jazz. In music, musicians have three main choices during improvisation: reproducing a familiar piece, slightly adjusting a known piece, or creating entirely new notes. This process of seeking perfect harmony mirrors the HS algorithm's approach to finding optimal solutions in various optimization problems [10]. In this analogy, instruments represent producers, and musicians represent consumers.

B. Problem definition

The energy routing problem primarily revolves around three key challenges: pair selection, determining the optimal path for energy transmission between pairs, and establishing a transmission sequence to prevent congestion. In selecting pairs, both cost and power loss are considered. The cost is determined by the energy providers, while the power loss depends on the amount of energy transmitted and the chosen transmission path.

C. Algorithm description

Each prosumer, whether a buyer or seller, sends a request to the controller to join the energy market. For instance, if a consumer submits a buy request, the controller's role is to identify the best producer or combination of producers to meet the consumer's energy demand. An essential factor in this selection process is the transmission time of each participant. Before the controller begins finding the best possible sets of producers, it must first evaluate whether each producer's transmission time aligns with the consumer's request.

Mathematically, this can be represented as follows:

Let: - $T_c = [t_{c\_start}, t_{c\_end}]$ be the transmission time interval requested by the consumer. - $T_p = [t_{p\_start}, t_{p\_end}]$ be the transmission time interval available for a producer.

For a producer to be considered, the consumer's transmission interval T_c must be fully contained within the producer's interval T_p . In other words, the following condition must hold:

$$t_{p\_start} \leq t_{c\_start} \quad \text{and} \quad t_{c\_end} \leq t_{p\_end}$$

This ensures that the producer is capable of delivering the requested energy within the time frame needed by the consumer. If this condition is not satisfied, the producer is excluded from the selection process. Only producers whose transmission times overlap with the consumer's interval will be considered for further optimization based on the energy availability, cost, and power loss.

By checking the available producers based on their transmission intervals first, the HS-ERP is applied only to feasible sets, reducing computational complexity and improving the efficiency of the matching process.

The steps of the HS-ERP are as follows:

1) **Step1: Initialize harmony search parameters:**

- **HMS:** Harmony Memory Size tells about number of solution vector or population.
- **HMCR:** Harmony Memory Considering Rate is a number between 0 and 1.
- **PAR :** Pitch Adjusting Rate, bw: distance bandwidth, (0,1).
- **r :** uniformly distributed random number between 0 and 1.

2) **Step2: Initialize harmony memory:** The HM takes the form of a matrix. Let HM be the matrix representing the HM, where each element x_i^j in the matrix corresponds to the amount of energy drawn from producer i in the solution represented by vector j . The matrix HM can be expressed as:

$$HM = \begin{bmatrix} x_1^1 & x_2^1 & \dots & x_n^1 \\ x_1^2 & x_2^2 & \dots & x_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{HMS} & x_2^{HMS} & \dots & x_n^{HMS} \end{bmatrix}$$

In this step, the matrix is populated randomly, with the primary condition that the total amount of energy in each vector must equal the corresponding demand D . Mathematically, this can be expressed as:

$$\sum_{i=1}^n x_i^j = D \quad \text{for } j = 1, 2, \dots, HMS$$

3) **Step3: Improve new harmony:** A new harmony vector $x' = (x'_1, x'_2, x'_3, \dots, x'_{HMS})$ is generated from the HM based on HMCR, PAR and r.

The new value x'_i can be chosen based on the HMCR, which ranges from 0 to 1. It is selected either from the HM with probability HMCR or from the feasible range X_i with probability $1 - \text{HMCR}$, as follows:

$$x'_i = \begin{cases} x'_i \in (x_i^1, x_i^2, x_i^3, \dots, x_i^{HMS}) & \text{with probability HMCR} \\ x'_i \in X_i & \text{with probability } (1 - \text{HMCR}) \end{cases}$$

The HMCR determines the likelihood of selecting a random value for the current decision variable from all vectors stored in the HM. Specifically, it involves choosing a random value from the selected column for the current decision variable across all rows in the HM. The notation $x'_i \in (x_i^1, x_i^2, x_i^3, \dots, x_i^{HMS})$ indicates that the value of x'_i is randomly chosen from producer i , with vectors ranging from 1 to HMS. Additionally, $(1 - \text{HMCR})$ sets the probability of randomly selecting a value from the full range of values, representing an amount of energy from the range of the selected producer, not from the HM. Regardless of the selection method, it is crucial that the total amount of energy in the new vector always equals the demand. The set X_i defines the potential range of values for the i -th producer. This randomization promotes exploration of diverse solutions while

ensuring the total energy aligns with the demand, driving the system towards finding the global optimum.

For the pitch adjustment phase, each component of the new harmony vector selected from the HM, undergoes evaluation to decide whether pitch adjustment is necessary. This process is governed by the PAR parameter and can be expressed as follows:

The decision for pitch adjustment is made as follows:

- If $r \leq \text{PAR}$, pitch adjustment is performed (Yes).

- If $r > \text{PAR}$, no pitch adjustment is carried out (No).

Here, the pitch adjustment process aims to optimize the current producer's value within the HM, ensuring it attains the maximum value from the vectors $(x_i^1, x_i^2, x_i^3, \dots, x_i^{HMS})$ i.e from producer i in solution $(1 \dots HMS)$. This adjustment strategy contributes to the exploration of promising solutions in the Harmony Search algorithm. Importantly, this approach helps in obtaining a group of minimum members among producers, mitigating congestion during energy transmission [10] [11].

Note that, the total amount of the new vector is equal to the consumer demand. If the selected producer energy exceed the consumer demand it will be taken into consideration.

4) **Step4: HM update:** To update the HM, the fitness value of each vector V_i is calculated, taking into account both power loss and cost. The fitness value for each vector is computed as follows:

$$V_i = \alpha \cdot \sum_{j=1}^n \text{cost}_j + \beta \cdot \sum_{j=1}^n \text{Ploss}_j$$

Where:

- α and β are weighting coefficients for the cost and power loss, respectively.

- cost_j represents the payment made by the consumer for energy from producer j in vector i .

- Ploss_j represents the power loss associated with producer j .

The cost for each producer j , cost_j , is determined by the following equation:

$$\text{cost}_j = \text{price}_j \cdot x_i^j$$

Here:

- price_j is the price set by producer j , expressed in USD per kilowatt (USD/KW).

- x_i^j is the amount of energy supplied by producer j in vector i .

For Ploss_j , the power loss is computed for the selected path that carries energy from the prosumer to the consumer. In our method, the path is obtained using a brute-force procedure that:

1. Retrieves all possible simple paths between the source and the destination.
2. Computes the power loss for each path and selects the one with the minimum loss.

The power loss for a path is influenced by two elements: losses along the power cables and losses inside the energy routers. Following [12], cable losses are computed using Equation 1.

$$P_{yz} = R_{yz} \cdot I_{yz}^2 \quad (1)$$

Equation 1 shows that power loss on a cable between nodes y and z depends on the cable resistance R_{yz} (in ohms) and the current I_{yz} flowing through it. Higher current or higher resistance leads to higher transmission loss.

The current flowing through the cable is computed using:

$$I_{yz} = \frac{P}{U_{yz}} \quad (2)$$

where U_{yz} is the line voltage (V) and P is the transmitted power (W). Substituting Equation 2 into Equation 1 gives:

$$P_{yz} = \frac{R_{yz}}{U_{yz}^2} \cdot P^2 \quad (3)$$

Transmission loss increases with the total current flowing through the cable. When new power is added to a line that already carries an existing load, the total current increases, which in turn increases the loss. Equation 4 expresses this additional loss by considering both the new transmitted power and the power already present on the cable.

$$P_{yz} = \frac{R_{yz}}{V_{yz}^2} \cdot ((x + x_{\text{existing}})^2 - x_{\text{existing}}^2) \quad (4)$$

Here, x is the newly transmitted power, and x_{existing} is the existing load on that line.

In addition to cable losses, each energy router also introduces loss due to its efficiency level. The useful energy delivered by the router is:

$$PE = E \cdot \text{Eff} \quad (5)$$

where E is the incoming energy and Eff is the router efficiency. The corresponding router loss is then:

$$P_y = E - PE \quad (6)$$

Finally, the total loss along a path is obtained by summing all cable losses and all router losses:

$$P_{\text{Loss}_j} = \sum_{yz \in \text{path}} P_{yz} + \sum_{y \in \text{path}} P_y \quad (7)$$

This value P_{Loss_j} represents the total energy lost when transmitting power along path j .

Algorithm 1 HS-ERP: Harmony Search-based Energy Routing Protocol

Require: Consumer demand D , producer set P , network topology

Ensure: Selected producer subset and minimum-loss routing path

- 1: Filter all producers that satisfy the consumer's transmission interval.
 - 2: Initialize Harmony Memory (HM) with feasible producer subsets.
 - 3: Set HS parameters: Harmony Memory Size (HMS), Harmony Memory Considering Rate (HMCR), Pitch Adjusting Rate (PAR), and number of iterations.
 - 4: **for** each iteration **do**
 - 5: Generate a new harmony (candidate subset) using HS rules.
 - 6: Compute its fitness based on:
 - 7: – total cost of selected producers,
 - 8: – transmission power loss using Equations (1)–(6).
 - 9: Update the HM by replacing the worst harmony if the new one is better.
 - 10: **end for**
 - 11: Select the best producer subset P from the HM.
 - 12: Using brute-force, retrieve all simple paths between producers in P and the consumer.
 - 13: For each path, compute the total power loss using Equations (1)–(6).
 - 14: Choose the path with the minimum total loss.
 - 15: **return** P and its optimal routing path.
-

IV. SIMULATION AND PERFORMANCE ANALYSIS

The implemented algorithms were executed on a laptop running the latest Python version available on Kaggle. This laptop features an 11th Gen Intel(R) Core(TM) i7-11370H processor operating at 3.30GHz, 8.00 GB of RAM, and a 64-bit OS. A series of experiments was conducted to evaluate the approach, encompassing different scenarios. Furthermore, a comprehensive comparative analysis is undertaken, where we compare our method against GA-ERP [6].

The network comprises interconnected nodes, capable of switching roles between consumer and producer. These nodes are referred to as energy routers, connected by power lines represented as edges. The edge weights indicate power loss during transmission. It is important to note that, in our analysis, we assume the cables have no inherent power.

The network consists of 17 nodes interconnected with power lines, with three producers (9, 15, and 16) and one consumer (4) (Figure 1).

The energy routers and power lines characteristics are detailed in Tables II and III, respectively.

Table IV provides the energy profiles of the consumer (D4) and the producers (D9, D15, and D16) in the network, including key information such as the price, the amount of energy and the transmission time.

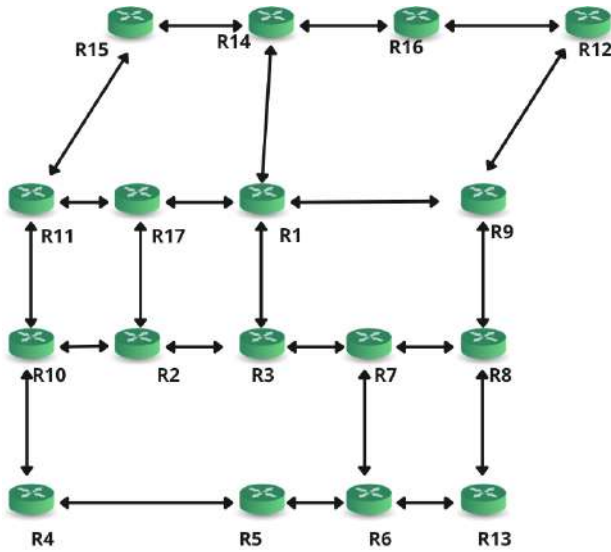


Fig. 1. The 17 node's network graph

TABLE II
ENERGY ROUTER'S CHARACTERISTICS FOR 17 NODE'S NETWORK

ER	Efficiency	Capacity	ER	Efficiency	Capacity
1	1	20	10	0.97	20
2	0.98	15	11	1	27
3	1	25	12	0.98	25
4	1	24	13	1	30
5	0.98	22	14	0.98	19
6	0.97	20	15	1	18
7	0.98	25	16	0.97	18
8	0.97	30	17	0.98	20
9	1	30			

TABLE III
POWER LINE'S CHARACTERISTICS FOR 17 NODE'S NETWORK

Power line	Capacity	R_{ij}	V_{ij}	Power line	Capacity	R_{ij}	V_{ij}
L_{1-3}	30	0.6	400	L_{6-7}	40	0.24	400
L_{1-9}	45	0.45	400	L_{6-13}	30	0.21	400
L_{1-14}	40	0.21	400	L_{7-8}	10	0.21	400
L_{1-17}	30	0.24	400	L_{8-9}	32	0.65	400
L_{2-3}	20	0.64	400	L_{8-13}	40	0.19	400
L_{2-5}	20	0.51	400	L_{9-12}	32	0.45	400
L_{2-10}	30	0.19	400	L_{10-11}	30	0.24	400
L_{2-17}	30	0.19	400	L_{11-15}	32	0.6	400
L_{3-7}	45	0.94	400	L_{11-17}	40	0.24	400
L_{4-5}	24	0.19	400	L_{12-16}	40	0.21	400
L_{4-10}	30	0.64	400	L_{14-15}	30	0.45	400
L_{5-6}	7	0.45	400	L_{14-16}	35	0.19	400

Table V represents the HS-ERP simulation parameters.

A. Interpretation of Routing Decisions Across Cases

The three test cases highlight how HS-ERP adapts routing decisions based on transmission-time constraints:

- **Case 1: All producers available.** As shown in table IV D4 requires 22 kW of energy from 09:00 to 10:00, but no single producer can supply the necessary amount. Therefore, a group of prosumers will be selected. As outlined in Table VI, several combinations are possible.

TABLE IV
ENERGY PROFILE TABLE

Energy Profile	D4	D9	D15	D16
Power amount (Kw)	-22	9	12	15
Price(USD/Kw)	-	0.068	0.056	0.043
Time (h:m)	Case 1	9:00-10:00	9:00-10:00	9:00-10:00
	Case 2	9:00-10:00	8:30-09:30	09:00-10:00
	Case 3	9:00-10:00	8:30-09:30	09:00-11:00

TABLE V
HS-ERP SIMULATION PARAMETERS

Algorithm	Simulation Parameters					
HS-ERP	HMS	HMCR	PAR	maximum iterations	α	β
	5	0.5	0.5	5	0.5	0.5

The HS-ERP, along with the Brute force, identifies the optimal set based on the fitness value, which selects a combination of producers D9, D15, and D16, allocating 9, 12, and 1 kW from each, respectively. The routing paths chosen minimize power loss while respecting the network topology. GA-ERP identifies the same energy combination, but its routing decision differs slightly since it does not explicitly evaluate transmission intervals.

TABLE VI
SOLUTIONS PRODUCED BY THE HS-ERP AND BRUTE FORCE APPROACHES

Set	P	E	Cost	path	Ploss	Fit
S1	16	15	0.644	16→14→15→11→10→4	1.172	1.285
	9	7	0.476	9→1→3→2→5→4	0.277	
S2	15	12	0.672	15→11→10→4	0.362	1.061
	9	9	0.61	9→1→3→2→5→4	0.357	
	16	1	0.043	16→14→15→11→10→4	0.078	
S3	9	9	0.61	9→1→3→2→5→4	0.357	1.061
	15	12	0.672	15→11→10→4	0.362	
	16	1	0.043	16→14→15→11→10→4	0.078	
S4	15	12	0.672	15→11→10→4	0.362	1.122
	16	10	0.429	16→14→15→11→10→4	0.781	
S5	9	9	0.612	9→1→3→2→5→4	0.357	1.272
	16	13	0.55	16→14→15→11→10→4	1.01	
S6	16	15	0.644	16→14→15→11→10→4	1.172	1.210
	15	7	0.392	15→11→10→4	0.210	

- **Case 2: Producer D9 unavailable during the full interval** D4 requires 22 kW of energy, but no single prosumer can meet this demand. Therefore, a group of prosumers must be selected. Upon arrival at 09:00, consumer D4 evaluates available prosumers based on their transmission timings. Starting with D9, it finds that this prosumer can only transmit energy from 08:30 to 09:00, which is insufficient for D4's needs from 09:00 to 10:00. Consequently, D4 turns to other prosumers, specifically D15 and D16, both of which can provide energy in real time throughout the required interval. The potential sets of prosumers are displayed in Table VII, with the optimal set identified as S1 based on their fitness values. HS-ERP automatically excludes D9 due to interval mismatch and selects only D15 and D16. This explains the performance

difference in Case 2.

TABLE VII
SOLUTIONS PRODUCED BY THE HS-ERP AND BRUTE FORCE APPROACHES

Set	P	E	Cost	path	Ploss	Fit
S1	15	12	0.672	15→11→10→4	0.362	1.122
	16	10	0.429	16→14→15→11→10→4	0.781	
S2	16	15	0.644	16→14→15→11→10→4	1.172	1.210
	15	7	0.392	15→11→10→4	0.210	

- **Case 3: Multi-source scenario** When consumer D4 arrives at 09:00, it requires 22 kW of energy, but no single prosumer can fulfill this demand independently. Consequently, a selection of prosumers must be considered. Upon evaluating the available options, D4 discovers that while all producers could potentially supply energy, their transmission time intervals do not align with D4's requirements. Specifically, D9 can transmit energy from 08:30 to 09:30, which is insufficient for D4's needs from 09:00 to 10:00. In contrast, D15 arrives at 09:00 and can provide energy until 11:00, making it a suitable option. D16's timing is also compatible, allowing it to supply energy during the required period. Consequently, the viable sets of prosumers include D15 and D16, or D16 and D15, both of which fulfill the necessary conditions, similar to the previous case. Both algorithms find similar energy splits; however, HS-ERP produces lower fitness because it eliminates infeasible producer sets before optimization, reducing unnecessary path exploration.

The table VIII provides a clear comparison of the results achieved by two energy routing algorithms, HS-ERP and GA-ERP.

TABLE VIII
COMPARATIVE RESULTS OF THE HS-ERP AND THE GA-ERP

Algorithm	Case	Set	Optimal paths	Fitness
GA-ERP	1, 2 and 3	9	9→1→3→2→5→4	1.061
		15	15→11→10→4	
		16	16→14→15→11→10→4	
HS-ERP	1	9	9→1→3→2→5→4	1.061
		15	15→11→10→4	
		16	16→14→15→11→10→4	
	2 and 3	15	15→11→10→4	1.20
		16	16→14→15→11→10→4	

Unlike GA-ERP, which does not account for transmission time intervals and treats all cases similarly, our HS-ERP approach differentiates between cases based on the timing of energy transmission. This time-sensitive treatment is what improves the overall performance and adaptability of HS-ERP, as it takes into account the realistic constraints of energy routing in different scenarios.

V. CONCLUSION

In conclusion, this work presents the Harmony Search-based Energy Routing Protocol (HS-ERP), a robust solution for optimizing energy distribution within decentralized systems like smart grids and the Energy Internet. By combining Harmony

Search with a brute-force path generation algorithm, HS-ERP identifies the optimal set of prosumers based on factors such as energy availability, cost, power loss, and most importantly, transmission time intervals. Unlike traditional methods like GA-ERP, which overlook time-based constraints, our approach ensures that the transmission interval of a consumer is fully accommodated by available producers, leading to a more precise and efficient energy routing process. In future research, we will focus on refining transmission scheduling to further enhance the efficiency of the energy routing process. This involves optimizing how and when energy is transmitted between producers and consumers, taking into account not only real-time availability but also scheduling multiple transmission windows to better accommodate fluctuating energy demands and supply. By improving transmission scheduling, the system can reduce bottlenecks, minimize delays, and ensure more consistent energy delivery, further advancing the practical application of the HS-ERP in real-world energy networks.

REFERENCES

- [1] Dina Emad, Omar Abdel-Rahim, Wesam Rohouma, and Sobhy Mohamed Abdelkader. Energy internet: State of the art and challenges. *IEEE Access*, 12:143131–143148, 2024.
- [2] Jesús Jaime Moreno Escobar, Oswaldo Morales Matamoros, Ricardo Tejeida Padilla, Ixchel Lina Reyes, and Hugo Quintana Espinosa. A comprehensive review on smart grids: Challenges and opportunities. *Sensors*, 21(21):6978, 2021.
- [3] Amani Fawaz, Imad Mougharbel, and Hadi Y Kanaan. New routing application using bees colony for energy internet. In *2022 3rd International Conference on Smart Grid and Renewable Energy (SGRE)*, pages 1–8. IEEE, 2022.
- [4] Douglas J Arent, Peter Green, Zia Abdullah, Teresa Barnes, Sage Bauer, Andrey Bernstein, Derek Berry, Joe Berry, Tony Burrell, Birdie Carpenter, et al. Challenges and opportunities in decarbonizing the us energy system. *Renewable and Sustainable Energy Reviews*, 169:112939, 2022.
- [5] Dina Emad, Omar Abdel-Rahim, Tanemasa Asano, and Sobhy M. Abdelkader. Transitioning to efficient technologies in power systems: The paradigm of energy internet survey. In *2023 24th International Middle East Power System Conference (MEPCON)*, pages 1–6, 2023.
- [6] Lina Benchikh, Lemia Louail, and Assala Nacef. Solving energy routing problems in energy internet using genetic algorithm. In *2023 International Conference on Computer and Applications (ICCA)*, pages 1–6, 2023.
- [7] Lina Benchikh, Lemia Louail, and Djamila Mechta. Subscriber matching in energy internet using the firefly algorithm. In *Proceedings of the Second International Conference on Innovations in Computing Research (ICR'23)*, pages 418–432. Springer, 2023.
- [8] Djamila Mechta, Saad Harous, and Sara Hebal. Energy-efficient path-aware routing protocol based on pso for smart grids. In *2020 IEEE International Conference on Electro Information Technology (EIT)*, pages 093–097, 2020.
- [9] Assala Nacef, Djamila Mechta, Lemia Louail, and Lina Benchikh. Efficient energy routing in smart grid networks using fractional knapsack and dijkstra algorithm. In *2023 International Conference on Decision Aid Sciences and Applications (DASA)*, pages 565–569, 2023.
- [10] Xin-She Yang. Harmony search as a metaheuristic algorithm. *Music-inspired harmony search algorithm: theory and applications*, pages 1–14, 2009.
- [11] Sonia Sharma and Hari Mohan Pandey. Genetic algorithm, particle swarm optimization and harmony search: a quick comparison. In *2016 6th International Conference-Cloud System and Big Data Engineering (Confluence)*, pages 40–44. IEEE, 2016.
- [12] Dijo P Koottappillil, Raj M Naidoo, Nsilulu T Mbungu, and Ramesh C Bansal. Distribution of renewable energy through the energy internet: A routing algorithm for energy routers. *Energy Reports*, 8:355–363, 2022.

Evaluation of Custom Modulation Constellations for 6G Based on PAPR and CCDF Analysis

1st Sofiane Latreche

University of Bejaia, Faculty of Technology,
Electrical Engineering Laboratory,
06000 Bejaia, Algeria
sofiane.latreche@univ-bejaia.dz

Abstract—Sixth-generation (6G) wireless communication systems are expected to meet unprecedented demands in spectral efficiency, energy sustainability, and hardware-friendly signal processing. In this context, this paper evaluates several custom modulation constellations, with a particular focus on circular geometries, within a 6G-inspired DFT-s-OFDM transmission chain. Two key metrics, the peak-to-average power ratio (PAPR) and its complementary cumulative distribution function (CCDF), are analyzed to quantify how constellation geometry influences the signal envelope. Simulation results show that circular constellations consistently achieve significantly lower PAPR and improved CCDF characteristics compared to traditional square QAM, highlighting their strong potential for 6G uplink scenarios, low-power IoT devices, and applications requiring high power-amplifier efficiency. These findings further suggest promising avenues for future research, including constellation optimization under realistic 6G channel and hardware conditions, as well as AI-assisted design of energy-efficient modulation schemes.

Index Terms—6G, Modulation Design, Custom Constellations, PAPR Reduction, CCDF, DFT-s-OFDM, Energy Efficiency.

I. INTRODUCTION

The transition from fifth-generation (5G) wireless systems toward sixth-generation (6G) networks represents a major technological leap, driven by the need for extreme data rates, sub-millisecond latency, massive connectivity, and unprecedented energy efficiency [1]. Unlike 5G, which focuses on enhanced mobile broadband and large-scale machine-type communications, 6G is expected to support immersive communication (holographic streaming, XR/VR), ultra-massive MIMO, sub-terahertz links, and high-precision localization [2]. These emerging requirements place new constraints on the physical layer, especially on the design of modulation formats that remain robust under strict hardware limitations [3].

Current 5G New Radio (NR) systems rely almost exclusively on square QAM constellations with modulation orders up to 256-QAM [4]. While these constellations provide excellent symbol error-rate performance, they exhibit relatively poor behavior in terms of peak-to-average power ratio (PAPR), particularly when used with OFDM waveforms [5]. High PAPR results in severe power-amplifier (PA) backoff, reduced efficiency, increased hardware cost, stronger thermal requirements, and overall poorer energy utilization. These issues become even more critical in 6G, especially for uplink transmission from mobile devices, battery-powered IoT nodes,

integrated sensing systems, and distributed antenna architectures [6].

To address these limitations, the exploration of nonstandard or custom modulation constellations has gained significant interest [7]. Such constellations—including circular modulation, cross-QAM, APSK-like arrangements, and geometrically shaped symbol distributions, provide a larger design space that may improve energy efficiency, reduce envelope fluctuations, and adapt to hardware constraints [8]. Circular constellations, in particular, introduce radial symmetry that helps distribute symbol amplitudes more uniformly, offering lower PAPR at the cost of a slightly reduced minimum Euclidean distance [9].

In parallel, DFT-s-OFDM continues to play an important role in uplink transmission for reducing envelope variations relative to traditional OFDM [10]. Combining DFT-s-OFDM with optimized constellation geometries could therefore unlock a new generation of modulation schemes suited for 6G devices requiring low-power and PA-friendly signal waveforms [11]. This paper presents a rigorous performance investigation of custom constellation geometries under a 6G-like DFT-s-OFDM chain. Specifically, we compare standard square QAM, circular constellations, and cross-QAM configurations in terms of PAPR and CCDF. Our objective is not merely to show performance differences, but to provide physical insight into why certain modulation geometries produce inherently smoother or more fluctuating signal envelopes.

The contributions of this work are summarized as follows:

- 1) We design, normalize, and evaluate three families of custom constellations suitable for early-stage 6G research, including nonstandard 32-point cross-QAM and multiple-ring circular arrangements.
- 2) We integrate these constellations into a realistic DFT-s-OFDM uplink chain built to mimic early 6G waveform configurations.
- 3) We perform exhaustive simulations and provide detailed interpretations of PAPR and CCDF variations across the constellation families.
- 4) We highlight the suitability of circular constellations as energy-efficient candidates for future 6G uplink scenarios and power-limited devices.

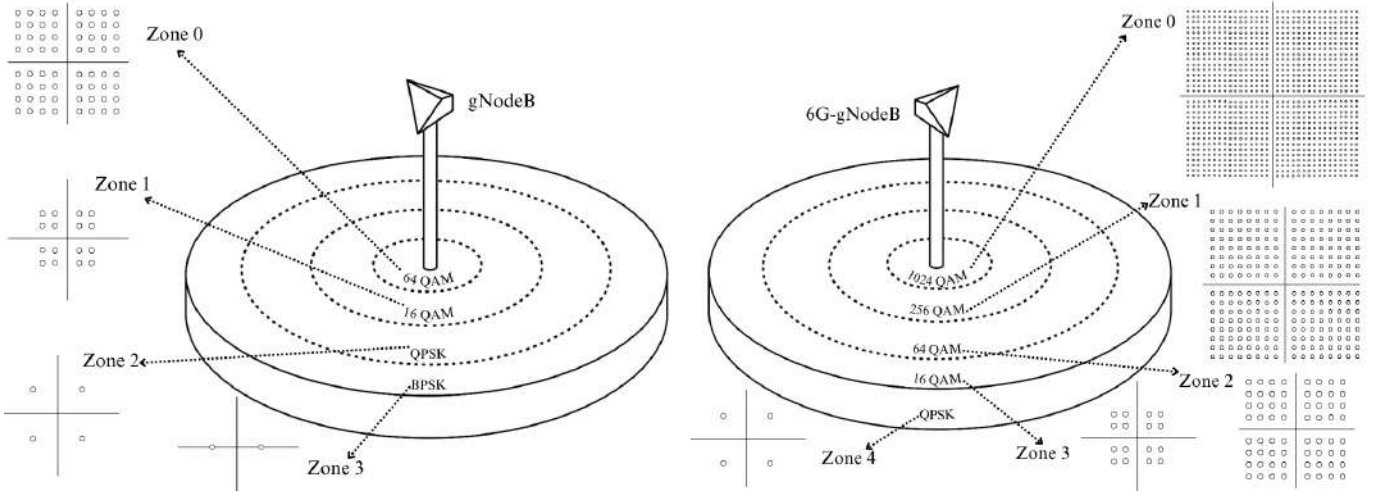


Fig. 1. Adaptive modulation scheme in 5G/6G systems.

The remainder of this paper presents the methodology, simulation model, results, discussion, and conclusions.

II. QUADRATURE AMPLITUDE MODULATION

This study intentionally focuses on square QAM as the primary baseline, since it represents the standardized modulation format currently used in 5G NR and early 6G prototypes. Techniques such as index modulation, APSK variations, and geometrically shaped constellations constitute relevant base-lines that will be incorporated in future work. The objective of the present study is to isolate and quantify the influence of pure geometric shaping on PAPR behavior within a DFT-s-OFDM framework before extending the comparison to more complex modulation families.

QAM is a widely employed modulation technique in modern digital communication systems, combining both amplitude and phase modulation to increase data transmission efficiency. The M-QAM variant, where "M" denotes the number of distinct symbols, allows for multiple bits to be represented by each transmitted symbol, thus significantly enhancing spectral efficiency [12].

In M-QAM, the modulated signal is expressed as a linear combination of two orthogonal carrier waves, typically sinusoidal, with in-phase (I) and quadrature (Q) components [13]. These components correspond to the real and imaginary parts of a complex signal, allowing M-QAM to map data points to distinct locations on a constellation diagram [14]. As M increases, the constellation diagram becomes denser, enabling more bits per symbol. For instance, in 16-QAM, each symbol encodes 4 bits, while in 64-QAM, each symbol represents 6 bits, resulting in improved data rates [15].

In M-QAM, each symbol can be represented as:

$$s_{i,j} = I_i + jQ_j \quad (1)$$

where I_i and Q_j are the amplitude levels for the in-phase and

quadrature components, respectively, and j is the imaginary unit. The number of bits per symbol is $\log_2 M$, allowing the transmission of multiple bits per symbol, thereby increasing the data rate.

The signal can be expressed as:

$$s(t) = I(t) \cos(2\pi f_c t) - Q(t) \sin(2\pi f_c t) \quad (2)$$

where f_c is the carrier frequency, and $I(t)$ and $Q(t)$ are the baseband signal components.

However, as the value of M increases, the system becomes more susceptible to noise and interference, requiring a trade-off between spectral efficiency and BER performance. This is particularly relevant in applications such as 5G [16] like presented in 1, 6G (6G is expected to utilize even higher-order QAM schemes, such as 256-QAM and beyond [17]), Wi-Fi, and high-speed wireless communication systems, where M-QAM is utilized to optimize bandwidth usage while maintaining acceptable error rates under varying channel conditions.

One critical aspect of M-QAM performance is the equalization process, which compensates for channel-induced distortions. Techniques like blind equalization using the MMA can mitigate inter-symbol interference and improve BER, especially in scenarios involving higher-order M-QAM constellations [18].

III. CONSTELLATION DESIGNS AND METHODOLOGY

This work investigates two categories of constellation geometries: circular constellations, and QAM designs. Square QAM serves as the baseline modulation format used in 5G NR, featuring orthogonal real and imaginary axes with uniformly spaced amplitude levels. Its predictable structure provides good error-rate performance but leads to amplitude variations that amplify PAPR issues.

Circular constellations are constructed as configurations of multiple concentric rings, each containing equally spaced

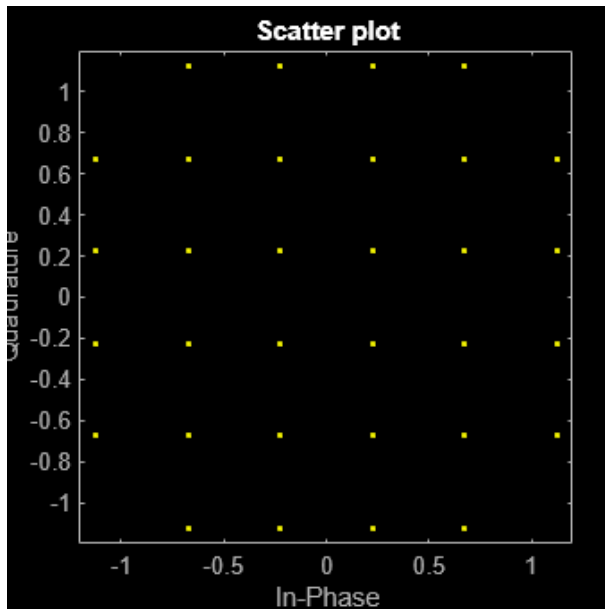


Fig. 2. QAM Constellation

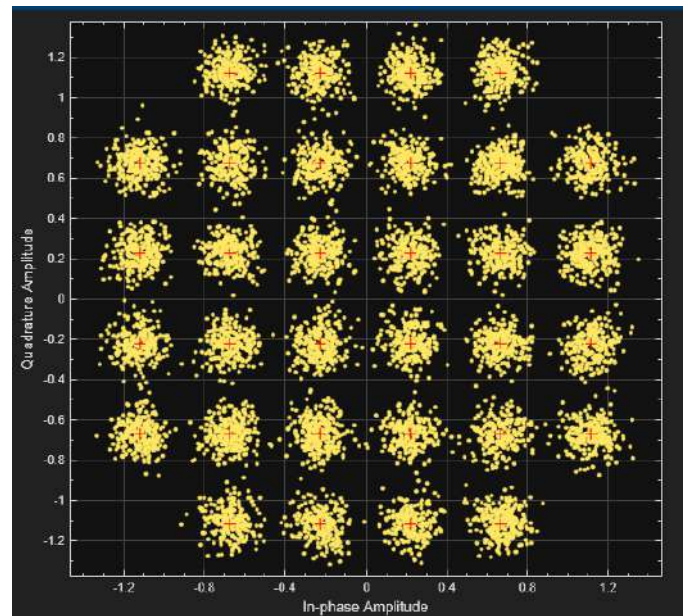


Fig. 4. Received QAM constellations.

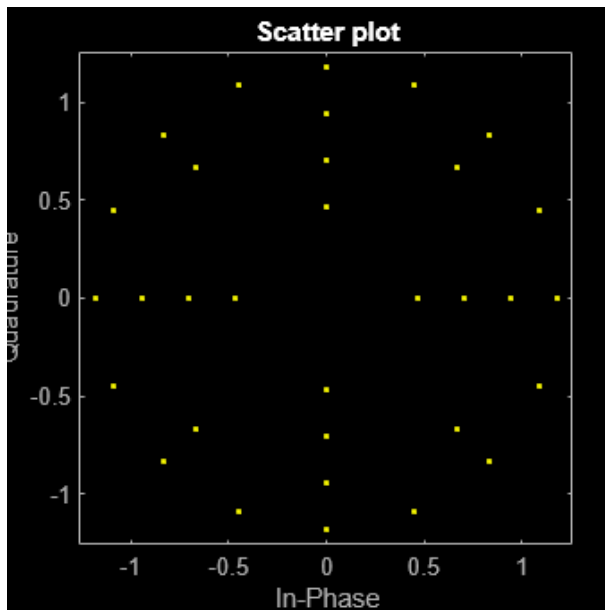


Fig. 3. Circular Constellations

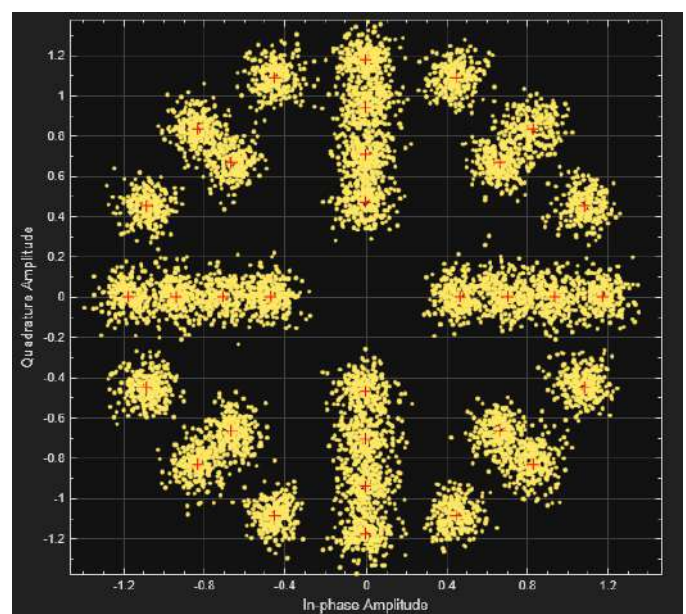


Fig. 5. Received circular Constellations

points along the angular dimension. This radially symmetric arrangement distributes symbol amplitudes more evenly, significantly reducing the likelihood of large instantaneous envelope peaks. Such constellations do not appear in current 3GPP specifications but are increasingly studied for 6G due to their energy-efficient nature and hardware-friendly properties.

Cross-QAM constellations provide a compromise between rectangular and circular layouts, supporting nonstandard modulation orders such as 32-QAM. These constellations contain a mixture of central dense points and peripheral cross-

shaped structures, enabling flexible trade-offs between power efficiency and Euclidean distance.

For meaningful comparison, all constellations are normalized to unit average power using RMS normalization. After normalization, the constellations are embedded into a 6G-inspired DFT-s-OFDM uplink chain that mimics the structure of NR but with extended bandwidth and a large resource grid. The chain includes random bit generation, symbol mapping, DFT precoding, subcarrier mapping, OFDM modulation, and final amplitude normalization. PAPR and CCDF are then computed from the resulting time-domain signals, providing

clear insight into peak-power characteristics.

The corresponding constellation diagrams are depicted in Figures 2 and 3.

While the proposed work is primarily simulation-driven, analytical insight is provided to justify the observed behavior. The PAPR of a DFT-s-OFDM waveform can be expressed as

$$\text{PAPR} = \frac{\max_n |x[n]|^2}{\mathbb{E}\{|x[n]|^2\}}, \quad (3)$$

where $x[n]$ denotes the time-domain samples of the transmitted signal. The statistical properties of $x[n]$ depend directly on the amplitude distribution of the underlying constellation.

Circular constellations exhibit a constant-modulus-like structure with reduced amplitude variance, which mathematically lowers the value of $\max_n |x[n]|^2$ and thus contributes to reduced PAPR. In contrast, square QAM contains high-energy corner points that significantly increase the probability of generating large instantaneous peaks after DFT precoding and OFDM modulation.

The proposed constellation structures do not introduce additional computational complexity at the transmitter. All processing blocks (symbol mapping, normalization, DFT spreading, subcarrier mapping, and OFDM modulation) remain identical to those used for standard QAM in current NR systems. Furthermore, memory requirements and lookup-table implementations scale linearly with the constellation order M , ensuring that the adoption of the proposed constellations does not impose additional hardware burden compared to traditional QAM schemes.

IV. SIMULATION MODEL

The simulation model is designed to reflect a simplified but representative 6G uplink scenario. A single transmit antenna is used to eliminate spatial-domain effects and ensure that differences in PAPR arise solely from constellation geometry. A wideband resource grid with 275 subcarriers is allocated, corresponding to a high-throughput setting consistent with 6G aspirations.

Binary data streams are mapped to constellation symbols and processed by a DFT-s-OFDM precoding block, which reduces envelope fluctuations by spreading symbol energy before mapping onto subcarriers. OFDM modulation is performed using an NR-like configuration, including IFFT processing and cyclic prefix insertion. The resulting time-domain waveform is normalized to ensure a consistent comparison among all constellation types.

Propagation occurs over an additive white Gaussian noise (AWGN) channel with a signal-to-noise ratio of 20 dB. This choice is deliberate, as the goal is to preserve waveform integrity for PAPR analysis rather than assess error performance. No equalization, channel estimation, or decoding blocks are included, as they are not relevant for evaluating signal envelope properties. The received constellation diagrams are shown in Figures 4 and 5. PAPR is computed directly from the normalized time-domain samples, and the CCDF curve is

generated using a large set of PAPR measurements to capture statistical extreme values.

V. SIMULATION RESULTS AND DISCUSSION

The performance of the different constellation geometries is thoroughly evaluated using two principal metrics: the peak-to-average power ratio (PAPR) and the complementary cumulative distribution function (CCDF) of the PAPR. These indicators provide complementary perspectives on the instantaneous amplitude behavior of the transmitted waveform and are particularly relevant for assessing the compatibility of modulation schemes with power-limited 6G hardware.

The PAPR results show clear and systematic differences between the tested constellations. Standard square QAM demonstrates the highest PAPR values, which can be attributed to its uneven amplitude distribution and the presence of high-magnitude corner points. When these corner points constructively combine during DFT-s-OFDM processing, they produce sharp instantaneous peaks in the time-domain waveform. This behavior is especially problematic for mobile devices and IoT transmitters, where power amplifiers operate with limited backoff margins.

Circular constellations exhibit a markedly different profile. Due to their radially symmetric structure, symbol points lie on evenly spaced amplitude rings, reducing the probability of abrupt magnitude variations after precoding. This property enables circular constellations to achieve the lowest PAPR values observed in the simulations. Their smooth amplitude transitions effectively suppress large peaks, resulting in a more hardware-friendly waveform with significantly reduced envelope fluctuations.

Cross-QAM designs, which combine elements of square and circular layouts, offer intermediate performance. Although they do not match the uniformity of circular constellations, cross-QAM schemes demonstrate noticeable improvements over square QAM, especially for nonstandard modulation orders such as 32-QAM. These constellations represent a promising compromise between spectral efficiency and envelope smoothness.

The CCDF curves further reinforce these findings by quantifying the probability that the PAPR exceeds a given threshold. Circular constellations present the steepest CCDF decay, indicating that extreme peak events occur very rarely. This behavior is highly advantageous for power-amplifier design, as it enables operation closer to saturation with reduced backoff, thereby improving energy efficiency and overall system performance. In contrast, square QAM exhibits a much slower CCDF decay, confirming its higher likelihood of producing large instantaneous peaks.

The combined analysis of PAPR and CCDF clearly demonstrates that constellation geometry has a strong and measurable impact on the peak-power characteristics of DFT-s-OFDM waveforms. Circular constellations consistently emerge as the most favorable option for 6G uplink communications, low-power IoT transmissions, and scenarios requiring stringent energy constraints. Their primary drawback is a

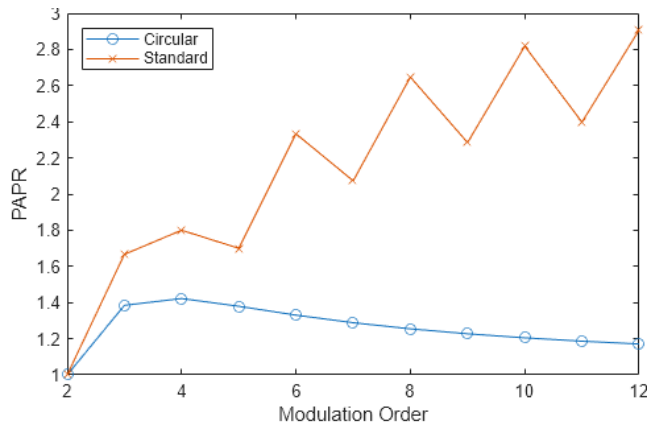


Fig. 6. PAPR vs modulation order before transmission

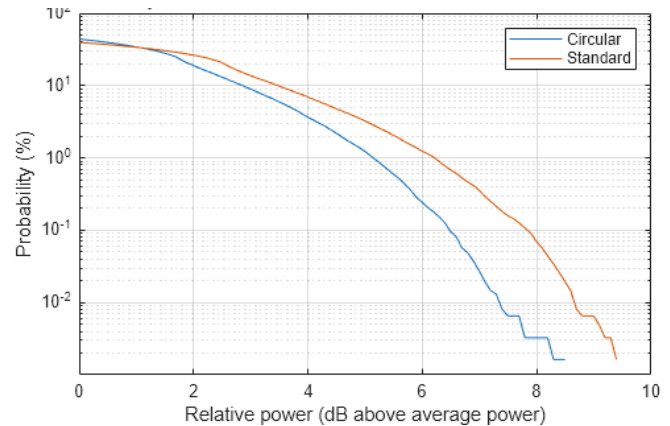


Fig. 7. CCDF comparison for modulation order 3

modest reduction in minimum Euclidean distance, which can lead to slightly worse error-rate performance under low-SNR conditions. However, this trade-off is minimal compared to the significant gains in peak-power reduction and amplifier efficiency.

The numerical results further highlight the substantial advantages of circular constellations. As shown in Figure 1, the PAPR of standard QAM rapidly increases with modulation order, reaching nearly 3.0 for order 12, while the circular constellation remains confined between 1.2 and 1.35 across all orders. For instance, at modulation order 12, the circular design achieves a PAPR of approximately 1.18, representing more than a 2.5 $\times$ reduction compared to the 3.0 observed for standard QAM. This behavior is confirmed by the CCDF curves in Figure 2. At a threshold of 6 dB above average power, the CCDF of standard QAM is around 1%, whereas the circular constellation reaches only 0.1%, indicating that large power peaks occur ten times less frequently. Beyond 8 dB, the circular constellation produces virtually no peak events, while standard QAM still exhibits non-negligible probabilities. These numerical observations confirm that circular constellation shaping significantly improves the statistical envelope behavior of DFT-s-OFDM waveforms, enabling more efficient PA operation and making it an attractive candidate for future 6G uplink scenarios.

It is important to emphasize that the present study is conducted under simplified conditions, using a single-antenna transmitter, an AWGN channel, and a fixed SNR of 20 dB. These assumptions isolate the effect of constellation geometry on PAPR, but they do not reflect the full complexity of real-world propagation. Future research should extend these simulations to multipath fading channels, nonlinear amplifier models, multi-antenna (MIMO and massive MIMO) configurations, and adaptive or AI-driven constellation shaping. Investigating these advanced scenarios will provide deeper insights into the viability of custom modulation schemes for practical 6G deployments.

VI. CONCLUSION

This paper presented a comprehensive evaluation of custom constellation geometries for 6G using a DFT-s-OFDM transmission framework. The results provide meaningful statistical insight. For example, at a modulation order of 12, circular constellations reduce the PAPR to approximately 1.18, while QAM reaches nearly 3.0, representing a reduction of more than 60%. Similarly, the CCDF shows a tenfold reduction in the probability of peaks exceeding 6 dB. These trends are consistent across all orders tested. Through this PAPR and CCDF analysis, we demonstrated that circular constellations offer substantial peak-power reductions compared to standard QAM, positioning them as strong candidates for future 6G uplink and low-power applications. These findings emphasize the importance of modulation shaping in addressing energy-efficiency challenges within next-generation wireless systems. Future research will explore constellation optimization under realistic channel and hardware conditions, including nonlinear PA modeling and AI-assisted constellation generation.

REFERENCES

- [1] Islam, S., Abdulsalam, A. Z., Kumar, B. A., Hasan, M. K., Kolondaisamy, R., & Safie, N. (2024). Mobile networks toward 5G/6G: Network architecture, opportunities and challenges in smart city. *IEEE Open Journal of the Communications Society*, 6, 3082-3093.
- [2] Nandito, M., & Kumar, Y. D. (2025, March). Exploring the Transition from 5G to 6G: Advancements in Ultra-Low Latency, High-Speed Connectivity, and AI-Driven Network Optimization. In *2025 International Conference on Automation and Computation (AUTOCOM)* (pp. 1234-1239). IEEE.
- [3] Singh, P., & Jung, S. Y. (2025). Modulation and multiple access for beyond 6g electromagnetic nanonetworks in the terahertz band. *IEEE Access*.
- [4] Tuninato, R., Capez, G. M., Mazzali, N., & Garello, R. (2025). 5G New Radio for Non-Terrestrial Networks: Analysis and Comparison of HARQ and RLC ARQ Performance Over Satellite Links. *IEEE Access*.
- [5] Kumar, A., Sharma, H., Gaur, N., & Gour, N. (2024). PAPR analysis of 5G and B5G waveforms using advanced PAPR algorithms. *IT Professional*, 26(4), 17-21.
- [6] Plastras, S., Tsoumatidis, D., Skoutas, D. N., Rouskas, A., Kormentzas, G., & Skianis, C. (2024). Non-terrestrial networks for energy-efficient connectivity of remote iot devices in the 6g era: A survey. *Sensors*, 24(4), 1227.

- [7] Capez, G. M., Cáceres, M. A., Armellin, R., Bridges, C. P., Fraire, J. A., Frey, S., & Garello, R. (2024). On the use of mega constellation services in space: Integrating LEO platforms into 6G non-terrestrial networks. *IEEE Journal on Selected Areas in Communications*.
- [8] Wang, R., Ikram, M., Yu, Y., Zhang, H., & Shamim, A. (2025). Dual-Band Circularly Polarized Shared-aperture Vivaldi MIMO Antenna with Linear to Circular Polarizer for 5G and 6G Communication. *IEEE Transactions on Antennas and Propagation*.
- [9] Kumar, A., Chakravarthy, S., Gaur, N., & Nanthaamornphong, A. (2024). Hybrid approaches to PAPR, BER, and PSD optimization in 6G OTFS: implications for healthcare. *Journal of Communications and Networks*, 26(3), 308-320.
- [10] Chen, L., Pan, J., Zhang, J., Cheng, J., Xu, L., & Ye, N. (2023). FDSS-based DFT-s-OFDM for 6G wireless sensing. *Sensors*, 23(3), 1495.
- [11] Sharma, V., Arya, R. K., & Kumar, S. (2023). Adaptive DFT-s-OFDM employed novel multi layered scheme for reduction of PAPR for mMTC node in 5G (NR). *Physical Communication*, 60, 102134.
- [12] Latreche, S., Bellahsene, H., Taleb-Ahmed, A. (2024). Blind MMA Equalization for QAM modulation in 5G/6G communications. *J. Electrical Systems*, 20(4), 469-481.
- [13] Vieira, W. E., Bertarini, P. L. L., Carbonaro, K. B., Carrijo, G. A., Agustini, E., & Dos Anjos, A. A. (2025). Performance Evaluation of Cross M-QAM Modulation over Standardized RF2 5G Frequency Bands Using the $\eta - \mu$ Fading Model. *IEEE Access*.
- [14] Singya, P. K., Shaik, P., Kumar, N., Bhatia, V., & Alouini, M. S. (2021). A survey on higher-order QAM constellations: Technical challenges, recent advances, and future trends. *IEEE Open Journal of the Communications Society*, 2, 617-655.
- [15] Okyere, B., Musavian, L., & Mumtaz, R. (2021). Performance analysis of physical layer network coding in massive MIMO systems with M-QAM modulations. *IEEE Transactions on Vehicular Technology*, 70(5), 4631-4645.
- [16] Shrivastava, P. S., Malviya, U. K., Meshram, M., & Dewangan, U. S. (2022). Efficiency of ultra-dense multi-tier future cellular networks for 5G: A survey. *Wireless Personal Communications*, 122(4), 3269-3291.
- [17] Yadav, B., & Patel, P. (2025). Adaptive Power Allocation Design of Beyond 6G Highly Dense NOMA-Massive MIMO Infinity-Norm Box Detection Technique. *Traitement du Signal*, 42(1), 101.
- [18] Cogen, F., Özden, B. A., & Aydın, E. (2024, November). Overview of Hexagonal QAM Based Constellation Schemes and Applications for Index Modulation. In *2024 15th National Conference on Electrical and Electronics Engineering (ELECO)* (pp. 1-5). IEEE.

Standard-Compliant Hydraulic-Aware MPPT Benchmark for Grid-Connected PV Water-Pump Systems under Realistic Partial Shading

Abdallah BEN ABDELKADER
Faculty of Technology University of
Tahri Mohammed Bechar, Algeria
Smart Grids & Renewable Energies
Laboratory

KAIL Soumia
Faculty of Technology University of
Tahri Mohammed Bechar, Algeria
Smart Grids & Renewable Energies
Laboratory

Mohammed Amine SOUMEUR
Faculty of Technology University of
Tahri Mohammed Bechar, Algeria
Smart Grids & Renewable Energies
Laboratory

Othmane ABDELKHALEK
Faculty of Technology University of
Tahri Mohammed Bechar, Algeria
Smart Grids & Renewable Energies
Laboratory

Abstract—The performance of maximum power point tracking (MPPT) in grid-connected photovoltaic (PV) water-pumping systems depends not only on the MPPT algorithm but also on the converter architecture and its inner-loop dynamics. This paper introduces a reproducible, standard-compliant benchmark for comparing single-stage (power-referenced) and two-stage (PV-voltage-referenced) architectures using EN 50530/CEC dynamic profiles with realistic midday partial shading. Four MPPT strategies—P&O and Incremental Conductance (INC) implemented in both architectures—are assessed. Over a compressed-day irradiance profile, Single P&O, Single INC, and Two-Stage P&O achieve similar performance (108.4 Wh and 0.400 m³), while Two-Stage INC yields only 97.2 Wh and 0.336 m³. Stress tests reveal that P&O and INC behave similarly under steady conditions, but architectural dynamics dominate under shading, where INC exhibits significant lag. The results demonstrate that converter dynamics can outweigh algorithm selection, and the proposed benchmark offers a concise, hydraulically aware framework for MPPT evaluation in PV pumping systems..

Keywords—MPPT, P&O, PV pumping, single-stage, two stage, partial shading

I. INTRODUCTION

Photovoltaic (PV) water pumping systems have emerged as a sustainable solution for irrigation, drinking water supply, and rural water management. In grid-connected installations, they promise higher reliability and integration with existing energy infrastructure. However, realizing their full potential requires efficient extraction of PV energy, for which maximum power point tracking (MPPT) is indispensable under varying irradiance and temperature conditions.

Widely used MPPT algorithms include Perturb & Observe (P&O) and Incremental Conductance (INC) [1]. These techniques adjust voltage or power references to maximize energy yield. Over the past decade, myriad variants and hybrid methods have been proposed, especially to address issues of partial shading and dynamic response [2], [3]. Yet, most studies treat MPPT in isolation—focusing solely on algorithm performance under steady conditions, often overlooking the influence of converter architecture or system dynamics.

In practical PV pumping systems, two principal power-electronic architectures are common: single-stage (direct DC–AC conversion with MPPT as power reference) and two-stage (DC–DC + DC–AC, MPPT issuing a PV voltage reference).

Comparing MPPT algorithms across different architectures is challenging, as performance depends heavily on inner-loop dynamics, switching behavior, and control delay [4]. Moreover, in real-world conditions—especially during partial shading—algorithm effectiveness is entangled with architectural design. This interaction remains underexplored in existing literature.

Another critical gap is the absence of benchmarking frameworks aligned with standards such as EN 50530, which defines static and dynamic MPPT efficiency tests for grid-connected inverters [5]. Although some comparative works use EN 50530 test profiles [6], they rarely address PV-pump coupling or the hydraulic dimension of the application. Even fewer incorporate water output metrics in their assessment.

To fill these gaps, this paper presents a reproducible MATLAB-based benchmarking platform that adheres to EN 50530/CEC dynamic profiles while simulating realistic midday partial shading. It implements P&O and INC algorithms in both single- and two-stage architectures, under identical conditions. Performance is evaluated by energy yield, tracking error, and delivered water volume. We also pinpoint operating regimes where P&O can rival—or even approach—INC, especially when inner-loop design is favorable. As a preview: in a compressed-day scenario, energy yields of 108.4 Wh versus 97.2 Wh and water volumes of 0.400 m³ versus 0.336 m³ reinforce that architecture and loop tuning may outweigh algorithm selection. The rest of this paper is organized as follows: Section II presents system modeling; Section III explains control and MPPT implementation; Section IV reports comparative results and analysis; and Section V concludes with implications and future directions

II. SYSTEM MODELING

A. System Overview

The system under study is a grid-connected photovoltaic (PV) water pumping installation composed of a PV array, a power conversion stage, a three-phase motor-driven centrifugal pump, and a maximum power point tracking (MPPT) controller. The primary objective of the MPPT is to keep the PV operating point near the maximum power point (MPP) under changing irradiance and temperature. As

illustrated in Fig. 1, two power electronic architectures are considered:

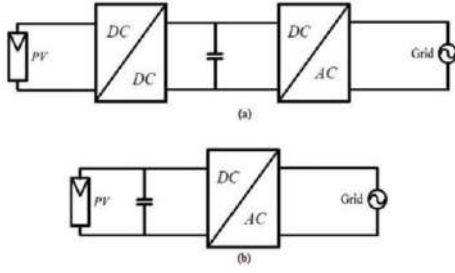


Fig.1. Grid-connected converter configuration (a) Conventional two-stage configuration

(b) Single-stage nanogrid configuration. **Single-stage architecture:** the DC-AC inverter directly interfaces the PV array with the grid and the pump. The MPPT generates a power reference.

- **Two-stage architecture:** a DC-DC converter regulates the PV voltage before feeding a DC-AC inverter. The MPPT generates a PV voltage reference.

This architectural difference significantly affects dynamic performance, tracking accuracy, and the amount of energy delivered to the hydraulic load.

B. PV Array Model

The PV array consists of multiple series and parallel modules. At standard test conditions (STC), the open-circuit voltage and short-circuit current are denoted by $V_{oc,stc}$ and $I_{sc,stc}$. Under irradiance G and cell temperature T , these values are adjusted as:

$$V_{oc}(T) = V_{oc,stc}[1 + \beta_{Voc}(T - T_{stc})], \quad I_{sc}(G, T) = I_{sc,stc} \left(\frac{G}{G_{stc}} \right) [1 + \alpha_{Isc}(T - T_{stc})], \quad (1)$$

where β_{Voc} and α_{Isc} are temperature coefficients. The output power is

$$P_{pv} = V_{pv} \cdot I_{pv}. \quad (2)$$

C. PV Array Model

The MPPT generates a power reference P_{ref} , and the inverter attempts to deliver power P_{inv} to the grid and pump. The inverter dynamics are approximated by a first-order response:

$$P_{inv}(t + \Delta t) = P_{inv}(t) + \frac{\Delta t}{\tau_p} [P_{ref} - P_{inv}(t)]. \quad (3)$$

In this configuration, the MPPT produces a PV voltage reference V_{ref} , regulated by a DC-DC converter through a voltage control loop:

$$V_{pv}(t + \Delta t) = V_{pv}(t) + \frac{\Delta t}{\tau_v} [V_{ref} - V_{pv}(t)]. \quad (4)$$

A DC-link capacitor buffers power variations between the DC-DC and DC-AC stages. Its dynamics follow:

$$\frac{d}{dt} (1/2 C V_{dc}^2) = P_{pv} - P_{inv} \Rightarrow \dot{V}_{dc} = \frac{P_{pv} - P_{inv}}{C V_{dc}}. \quad (5)$$

The DC-link voltage is constrained within $[V_{dc,min}, V_{dc,max}]$ to ensure safe operation.

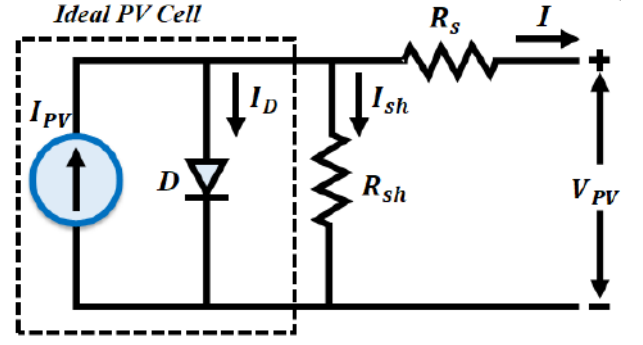


Fig.2. PV System Modelling

D. Hydraulic Pump Model

The three-phase centrifugal pump converts electrical power P_{inv} into water flow. The total efficiency is

$$\eta_{total} = \eta_{pump} \cdot \eta_{motor} \cdot \eta_{vfd}, \quad (6)$$

and the hydraulic flow rate is computed as:

$$Q = \frac{\eta_{total} P_{inv}}{\rho g H}, \quad (7)$$

where ρ is water density, g is gravitational acceleration, and H is pump head. A cut-in power threshold prevents operation at low power, while a maximum flow rate Q_{max} avoids unrealistic conditions. The total pumped volume is obtained by time integration of Q

E. Irradiance and Temperature Profiles

A compressed-day profile is constructed using piecewise-interpolated irradiance and temperature to emulate sunrise, noon, and sunset within a reduced simulation time. To represent realistic disturbances, partial shading is introduced around midday as a smooth power reduction. This approach is consistent with the dynamic requirements of EN-50530 and CEC test procedures, which impose rapid transitions to evaluate MPPT response. Using standardized-like profiles ensures fair, repeatable, and application-relevant comparison of MPPT algorithms under different architectures.

III. SIMULATION RESULTS AND DISCUSSION

All simulations were carried out in MATLAB/Simulink using the benchmark described in Section 3. The PV array, pump and grid-side models are excited by a compressed-day irradiance and temperature profile and by a realistic partial-shading pattern. Four MPPT variants are compared: P&O and INC in a single-stage, power-referenced architecture, and in a two-stage, PV-voltage-referenced architecture.

A. Dynamic test profile

Figure 3 shows the irradiance and cell-temperature waveforms applied to the PV array. The irradiance smoothly rises from 0 to about 1000 W/m² around $t \approx 450$ s and then decreases, while the cell temperature increases from 20°C to approximately 45°C before cooling down. This “compressed-day” profile emulates a clear-sky day with realistic thermal dynamics, so the MPPT algorithms are evaluated under continuously varying operating conditions rather than in steady state only.

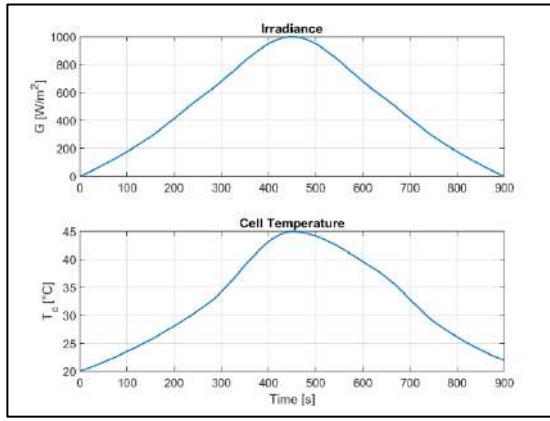


Fig. 3. Compressed-day irradiance and cell temperature profile used in the simulations, representing a realistic clear-sky day with moderate temperature rise.

B. PV power and hydraulic performance

Figure 4 compares the PV power obtained with the four MPPT variants against the ideal MPP power. The Single P&O, Single INC and Two P&O configurations closely follow the ideal trace throughout the day, whereas the Two INC case exhibits a clear power deficit, especially after the onset of partial shading.

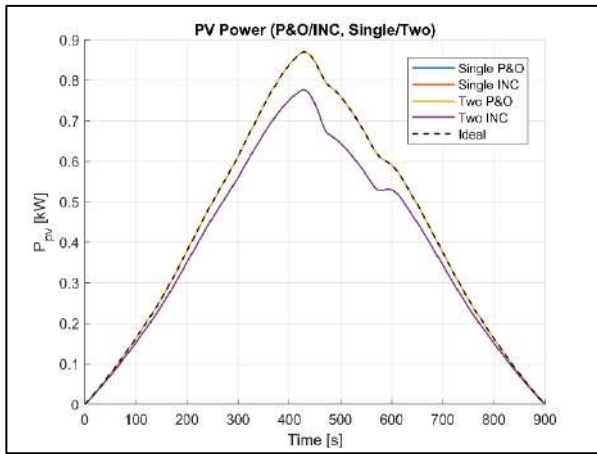


Fig. 4. PV power of the four MPPT variants (Single P&O, Single INC, Two P&O, Two INC) compared with the ideal MPP power over the full-day profile under partial shading.

The impact on the application level is quantified in Figs. 4 and 5. The inverter-delivered power in Fig. 4 shows that Two P&O delivers more power to the pump than Two INC over the pumping interval (roughly 220–700 s), while the two single-stage configurations behave similarly to Two P&O. As a result, the cumulative water volume in Fig. 3 reaches about 0.400 m³ for Single P&O, Single INC and Two P&O, but only 0.336 m³ for Two INC. The bar plots in Fig. 5 confirm that the corresponding PV energies are approximately 4 Wh for the three best cases and 97.2 Wh for Two INC, i.e., about 10% less harvested energy and 16% less pumped water in the worst case. This clearly illustrates how MPPT inaccuracies propagate into a measurable hydraulic deficit.

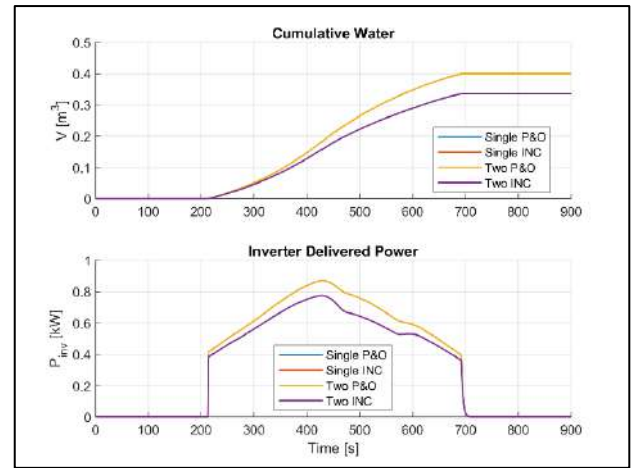


Fig. 5. Cumulative pumped water volume (top) and inverter-delivered power (bottom) for the four configurations. The reduced PV tracking performance of Two INC translates into a clear hydraulic deficit.

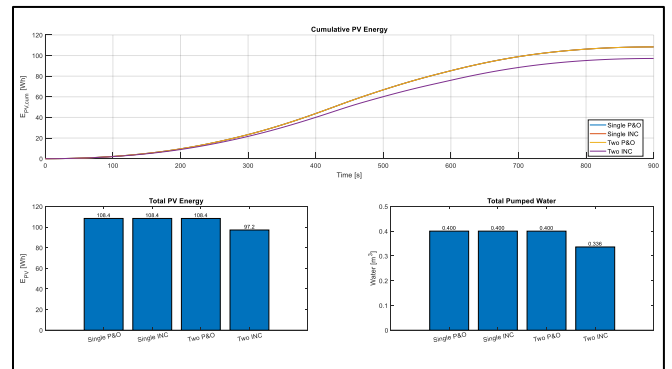


Fig. 6. Cumulative PV energy (top) and bar charts of total harvested energy and pumped water (bottom). Single P&O, Single INC and Two P&O reach ≈ 108.4 Wh and 0.400 m³, whereas Two INC yields only 97.2 Wh and 0.336 m³.

An overall view of PV voltage, current, power and cumulative water is provided in Fig. 7. The Two INC configuration operates at a significantly higher PV voltage than the other cases during most of the day, leading to lower current and power. The remaining three configurations maintain voltages close to the instantaneous MPP voltage and therefore exhibit nearly identical energy and water yields.

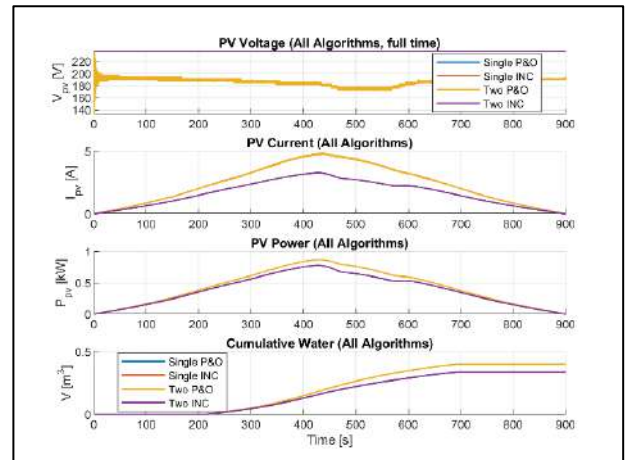


Fig. 7. All-in-one view of PV voltage, current, power, and cumulative pumped water for all MPPT variants over the full simulation horizon

C. Tracking accuracy and behavior under partial shading

The instantaneous tracking error with respect to the ideal MPP is plotted in Fig. 4. For Single P&O, Single INC and

Two P&O, the error remains very close to 0% over the entire profile. In contrast, the Two INC configuration starts with an error of about 6% and reaches up to 15–16% during the most severe partial-shading interval before decreasing again. These results confirm that the poor hydraulic performance of Two INC is rooted in its tracking inaccuracy within the selected two-stage architecture.

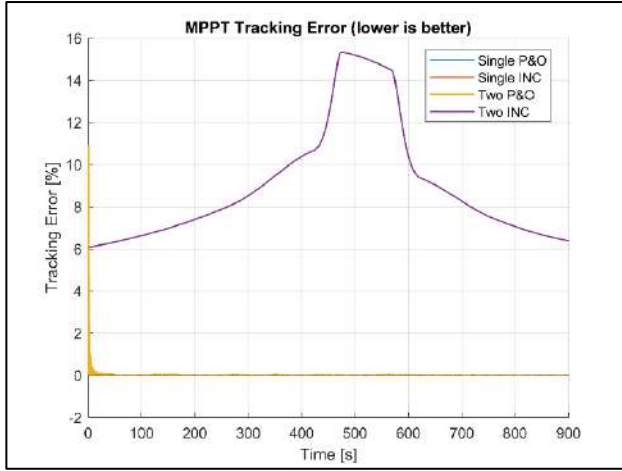


Fig. 8. Instantaneous MPPT tracking error with respect to the ideal MPP. Single P&O, Single INC and Two P&O stay close to 0%, while the Two INC configuration exhibits up to 15–16% error during shaded intervals.

Figure 8 zooms in on the partial-shading period. In the upper subplot, the PV power of Single P&O, Single INC and Two P&O remain almost indistinguishable from the ideal trajectory, whereas Two INC consistently underperforms. In the lower subplot, the PV-voltage waveform reveals the underlying mechanism: the Two P&O voltage exhibits the expected stepwise movement around the MPP voltage, while Two INC remains locked at an over-voltage operating point during most of the shaded interval. The I–V and P–V snapshots shown in Fig. 9 corroborate this interpretation: in all three considered instants (pre-, mid- and post-shading), the operating points of Single P&O, Single INC and Two P&O lie close to the true MPP, whereas the Two INC points are shifted towards higher voltages on the descending side of the P–V curve.

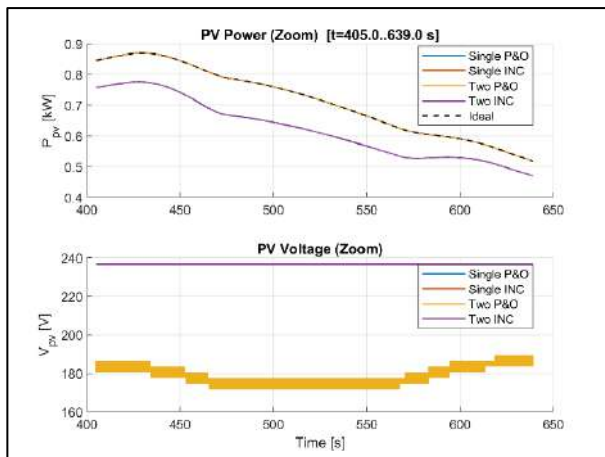


Fig. 9. Zoom on the partial-shading period: PV power (top) and PV voltage (bottom). Two INC is locked at an over-voltage operating point, while the other MPPT schemes closely follow the ideal power trajectory

D. Stress tests and influence of converter dynamics

To further clarify the influence of converter dynamics and inner loops, two stress tests are performed (Figs. 10 and 11). In Experiment A, a constant irradiance and temperature operating point is applied to the single-stage architecture. Both P&O and INC achieve virtually identical MPPT efficiencies close to 100% (Fig. 10(a)), and their power traces overlap with the ideal MPP power (Fig. 10(a)). This confirms that under steady, unshaded conditions the classical P&O and INC laws behave equivalently in the proposed single-stage structure.

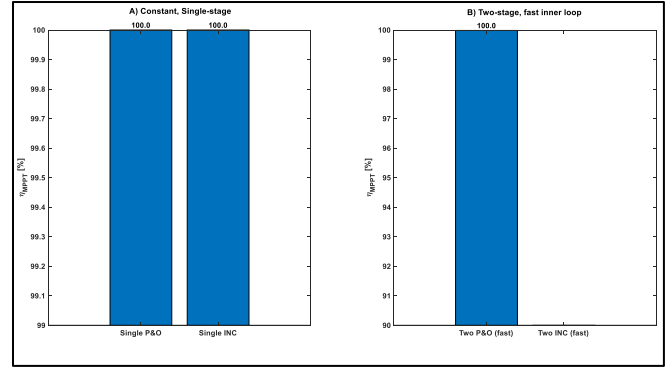


Fig. 10. MPPT efficiencies for: (A) a constant, unshaded operating point in single-stage architecture, and (B) a two-stage architecture with an ultra-fast PV-voltage loop. Under constant conditions P&O and INC are virtually identical; in the two-stage case, P&O outperforms INC

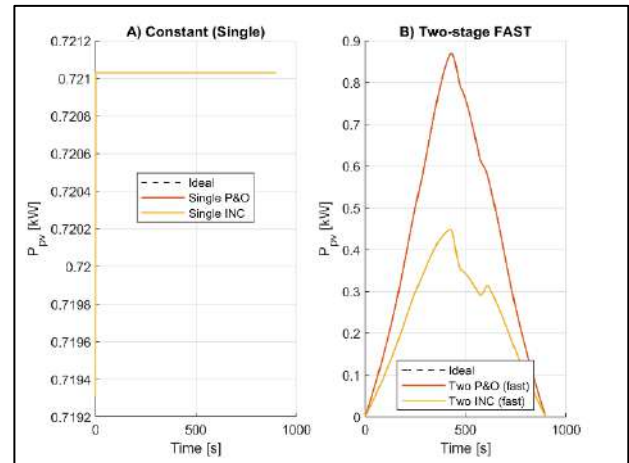


Fig. 11. PV power versus the ideal MPP trace for Experiment A (left, constant single-stage case) and Experiment B (right, two-stage fast inner loop).

In Experiment B, the two-stage architecture is considered but with an ultra-fast PV-voltage loop and a higher MPPT update rate. Under these conditions, the Two P&O (fast) configuration closely follows the ideal MPP trajectory and attains an MPPT efficiency close to 100% (Fig. 9(b)), while the Two INC (fast) case still exhibits a noticeable, although reduced, tracking gap (Fig. 10(b)). These results indicate that the observed differences between P&O and INC are not solely intrinsic to the algorithms themselves; they are strongly shaped by the surrounding conversion architecture and its dynamic response. When the inner voltage loop is appropriately tuned, a two-stage converter can enable classical P&O to rival or even outperform INC under realistic dynamic and partial-shading conditions.

IV. CONCLUSION

This paper has proposed a hydraulically aware, standard-compliant benchmark for assessing MPPT performance in

grid-connected PV water-pumping systems. The benchmark combines EN 50530/CEC dynamic test profiles with a compressed-day irradiance and temperature sequence and includes a realistic midday partial-shading event. Four configurations—Single P&O, Single INC, Two P&O, and Two INC—were compared in both electrical and hydraulic terms.

Simulation results show that Single P&O, Single INC, and Two P&O deliver approximately 108.4 Wh of PV energy and 0.400 m³ of pumped water over the benchmark profile, whereas Two INC delivers only 97.2 Wh and 0.336 m³. The resulting 10% energy deficit and 16% hydraulic deficit are directly linked to tracking inaccuracies during partial shading in the considered two-stage architecture. Additional stress tests under constant irradiance and with a faster PV-voltage loop confirm that P&O and INC can exhibit similar performance in a single-stage converter, and that improving converter dynamics can substantially mitigate their differences.

These findings suggest that, in PV water-pump applications, converter architecture and inner-loop dynamics can be as important as the selected MPPT algorithm. The proposed benchmark provides a concise, reproducible framework for jointly evaluating MPPT laws and power-conversion structures using both electrical and hydraulic metrics. Future work will extend the benchmark to advanced and adaptive MPPT schemes, different pump types and hydraulic configurations, and experimental validation on a hardware testbed.

REFERENCES

- [1] "Treatment episode data set: discharges (TEDS-D): concatenated, 2006 to 2009." U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies, August, 2013, DOI:10.3886/ICPSR30122.v2
- [2] Endiz, M. S., Gökkuş, G., Coşgun, A. E., & Demir, H. (2025). A review of traditional and advanced MPPT approaches for PV systems under uniformly insolation and partially shaded conditions. *Applied Sciences*, 15(3), 1031. [2] Islam, H., Mekhilef, S., Shah, N. B. M., Soon, T. K., Seyedmahmoudian, M., Horan, B., & Stojcevski, A. (2018). Performance evaluation of maximum power point tracking approaches and photovoltaic systems. *Energies*, 11(2), 365.
- [3] Harkani, A., Fassi, H. F., & El Aissaoui, A. (2024). Modeling of PV Water Pumping Performance using Multi-Parallel Pump Switching for an Optimal Hydraulic Power Point Tracking. *Modeling Earth Systems and Environment*, 10(5), 6435-6448.
- [4] "Overall efficiency of grid connected photovoltaic inverters (EN 50530)," European Standard, 2010.
- [5] MILADI, M. H., SETHOM, H. B. A., ABDELGHANI, A. B. B., & SLAMA-BELKHODJA, I. (2020). Innovative and Simple PV Generator Procedure to Test PV Inverter According to EN50530 Standard Static MPPT Efficiency. *International Journal of Renewable Energy Research (IJRER)*, 10(4), 1743-1754.
- [6] Youssef, R., Mohamed, N., Khalifa, R. B., Djamel, T., Abdelmoumen, G., Khechekhouché, A., & Lassaad, S. (2024). Comparative study of the mppt control for the photovoltaic water pumping system between fss-p&o and vss-p&o. *WSEAS Transactions on Power Systems*, 19, 229-241.
- [7] Li, X., Wen, H., Hu, Y., Du, Y., & Yang, Y. (2020). A comparative study on photovoltaic MPPT algorithms under EN50530 dynamic test procedure. *IEEE Transactions on Power Electronics*, 36(4), 4153-4168.
- [8] Andrejasic, T., Jankovec, M., & Topic, M. (2011). Comparison of direct maximum power point tracking algorithms using EN 50530 dynamic test procedure. *IET renewable power generation*, 5(4), 281-286.
- [9] Al-Obaidi, M. Q., Derbel, N., Hashim, W. A., & Aljebory, K. M. (2021, March). Modelling and simulation of PV pump using MPPT controller. In *2021 18th International Multi-Conference on Systems, Signals & Devices (SSD)* (pp. 1279-1284). IEEE.

RIS-Aided Restoration of NLoS Connectivity in 5G Sub-6 GHz Networks

Yacine Ouazziz*, Dina Alkama*

*Université de Bejaia, Faculté de Technologie,

Laboratoire d'Informatique Médicale et des Environnements Intelligents (LIMED),

06000, Bejaia, Algérie

Emails: yacine.ouazziz@univ-bejaia.dz, dina.alkama@univ-bejaia.dz

Abstract—Reconfigurable Intelligent Surfaces (RIS) have emerged as a key technology for enhancing coverage and reliability in 5G wireless networks, particularly in non-line-of-sight (NLoS) scenarios where conventional links suffer severe attenuation. This paper investigates the capability of RIS to restore downlink connectivity in an urban environment where the direct BS-UE path is blocked. A realistic RIS-assisted propagation model is developed and evaluated through Monte Carlo simulations in MATLAB. Results show that an optimized RIS can provide more than 30 dB of SNR improvement and increase spectral efficiency from near-zero to over 8 bit/s/Hz, effectively transforming an unusable NLoS link into a high-capacity channel. These findings confirm the strong potential of RIS as a low-power, cost-effective solution for urban coverage enhancement in 5G sub-6 GHz networks.

Index Terms—Reconfigurable Intelligent Surface (RIS), NLoS, 5G, Coverage Enhancement, Passive Beamforming.

I. INTRODUCTION

Reconfigurable Intelligent Surfaces have recently emerged as one of the most promising technologies for enhancing the performance of 5G and beyond wireless networks [1], [2]. Unlike conventional relays or active repeaters, an RIS consists of a nearly passive two-dimensional array of programmable reflecting elements capable of shaping the propagation environment without requiring radio-frequency amplification. By inducing tunable phase shifts on the incident wavefront, RIS panels can steer, focus, or scatter electromagnetic energy toward desired spatial directions, thereby enabling a new paradigm often referred to as “smart radio environments” [1], [3], [4].

Over the past few years, RIS technology has attracted substantial attention in the literature due to its low hardware complexity, low energy consumption, and ability to overcome harsh propagation conditions in dense urban areas. Consequently, a variety of works have investigated the potential of RIS for improving spectral efficiency, coverage, energy efficiency, and physical-layer security [5]–[7]. For instance, several works provide extensive overviews of Reconfigurable Intelligent Surfaces applications and their potential; [3] specifically discusses challenges and opportunities in RIS-empowered wireless communications, while other studies delve into specific aspects such as signal-processing and practical design of RIS-aided links. Furthermore, much of the existing literature has focused on the optimization of RIS

phase shifts, channel estimation, multi-user beamforming, and the integration of RIS into mmWave and THz systems [5]. Despite this growing body of work, only a limited number of studies highlight the fundamental coverage restoration capability of RIS in deep non-line-of-sight conditions using simple and intuitive system-level configurations, which is crucial for practical deployment and understanding its core benefits. This paper aims to address this gap by providing a focused study on this critical aspect.

1) Contribution: This paper presents a clear and concise system-level study of RIS-assisted coverage restoration in an urban NLoS scenario. Unlike works that rely on complex optimization models or multi-user setups, we focus on a simple and representative BS-RIS-UE geometry that isolates the fundamental gain mechanisms of RIS. Our main contributions are summarized as follows:

- We develop a tractable RIS-assisted NLoS propagation model using independent BS-RIS and RIS-UE fading channels, incorporating strong blockage on the direct BS-UE link.
- We evaluate the SNR and spectral-efficiency performance using MATLAB-based Monte Carlo simulations under realistic propagation assumptions.
- We quantify the impact of RIS aperture size and phase-control strategies, demonstrating the coherent N^2 power scaling and its dramatic effect on NLoS connectivity.
- We show that, with optimized phase alignment, RIS can restore a completely unusable NLoS link and deliver more than 30 dB SNR improvement and over 8 bit/s/Hz of spectral efficiency.

Overall, this work provides a compact and insightful demonstration of the strong potential of RIS to improve coverage in blocked environments, making it suitable for practical deployment considerations and national-level research dissemination.

II. SYSTEM MODEL

We consider a downlink wireless system where a base station (BS) communicates with a single-antenna user equipment (UE) located in a blocked non-line-of-sight (NLoS) region. A rectangular reconfigurable intelligent surface (RIS) is mounted on the facade of a nearby building to bypass the obstruction and create a virtual reflected link [9], [11], [12], following the

canonical RIS-aided NLoS coverage setup adopted in [3], [4]. A top-view representation of this geometry is shown in Fig. 1.

The BS transmits with power P_{tx} at a carrier frequency $f_c = 3.5$ GHz over a bandwidth of 20 MHz, which is consistent with typical sub-6 GHz RIS-aided 5G NR deployments [1], [2]. The receiver noise is modeled as additive white Gaussian noise (AWGN) with variance $\sigma_n^2 = N_0 B F$, where $N_0 = -174$ dBm/Hz is the thermal noise density, B is the signal bandwidth, and F is the receiver noise figure.

A. Channel Model

The BS–RIS and RIS–UE channels are modeled as narrow-band flat-fading links:

$$\mathbf{h}_{\text{BR}} = \sqrt{\beta_{\text{BR}}} \mathbf{g}_{\text{BR}}, \quad (1)$$

$$\mathbf{h}_{\text{RU}} = \sqrt{\beta_{\text{RU}}} \mathbf{g}_{\text{RU}}, \quad (2)$$

where β_{BR} and β_{RU} denote the large-scale distance-dependent path losses, and $\mathbf{g}_{\text{BR}}, \mathbf{g}_{\text{RU}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_N)$ represent small-scale Rayleigh fading, as commonly assumed in RIS literature [3], [7].

The direct BS–UE link is modeled as

$$h_{\text{BU}} = \sqrt{\beta_{\text{BU}}} g_{\text{BU}}, \quad (3)$$

where β_{BU} includes an additional blockage attenuation of 60 dB to emulate a deep NLoS scenario. Values in the range 40–60 dB are consistent with worst-case outdoor-to-indoor and around-corner blockage conditions at sub-6 GHz: the 3GPP TR 38.901 high-loss O2I model specifies external wall losses on the order of 20–35 dB at 3.5 GHz, while measurements in modern buildings with metal-coated glass report additional 20–40 dB penetration loss, leading to total excess losses that can approach or exceed 50–60 dB in severe cases (e.g., [?], [?], [13]).

Our choice therefore represents a conservative, worst-case NLoS condition that highlights the coverage-restoration capability of the RIS. Because of this, the direct link is orders of magnitude weaker than the RIS-assisted path and is therefore negligible in the final received signal, in line with NLoS blockage scenarios discussed in [4], [6].

B. RIS Reflection Model

The RIS consists of N passive reflecting elements, each inducing a tunable phase shift. Its reflection behavior is modeled by the diagonal matrix [10], [12].

$$\mathbf{\Theta} = \text{diag}(e^{j\theta_1}, e^{j\theta_2}, \dots, e^{j\theta_N}), \quad (4)$$

where $\theta_n \in [0, 2\pi)$ is the phase applied by the n -th element, consistent with standard passive RIS models [3], [7].

The effective cascaded BS–RIS–UE channel is

$$h_{\text{RIS}} = \mathbf{h}_{\text{RU}}^H \mathbf{\Theta} \mathbf{h}_{\text{BR}} = \sum_{n=1}^N \overline{h_{\text{RU},n}} e^{j\theta_n} h_{\text{BR},n}, \quad (5)$$

where $h_{\text{BR},n}$ and $h_{\text{RU},n}$ are the n -th entries of the corresponding channel vectors. This cascaded formulation is widely used in RIS-aided link modeling [4], [6] [8], [10].

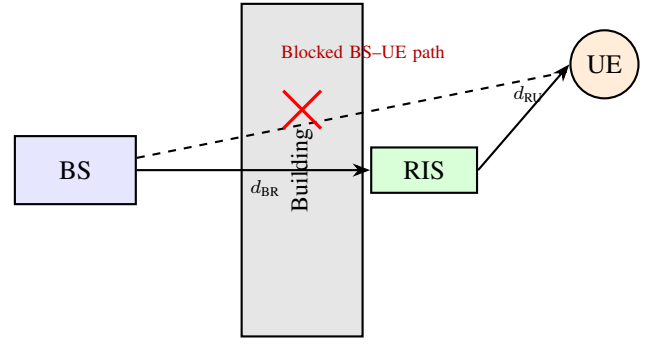


Fig. 1: RIS-aided NLoS geometry. The RIS is mounted on the building facade, not overlapping it, and provides a controllable reflected path around the obstruction, consistent with typical RIS deployment models in [1], [4].

C. Optimal RIS Phase Design

To maximize the coherent power of (5), the RIS aligns the phases of all individual reflected contributions. A maximum-ratio reflection (MRR) [8], [11] solution is obtained by

$$\theta_n^* = -\arg(h_{\text{RU},n} h_{\text{BR},n}), \quad n = 1, \dots, N. \quad (6)$$

This phase configuration corresponds to the standard coherent combining design for passive RISs [3], [4]. Substituting (6) into (5) yields the coherent combined magnitude

$$|h_{\text{RIS}}| = \sum_{n=1}^N |h_{\text{RU},n} h_{\text{BR},n}|. \quad (7)$$

Under independent Rayleigh fading, the average received power scales proportionally to N^2 , which explains the strong gains observed in RIS-aided NLoS propagation and has been highlighted in several survey works [3], [6], [7].

D. Received Signal and Performance Metrics

The baseband received signal at the UE is

$$y = (h_{\text{BU}} + h_{\text{RIS}})x + n, \quad (8)$$

where x is the transmitted symbol with $\mathbb{E}\{|x|^2\} = P_{\text{tx}}$ and $n \sim \mathcal{CN}(0, \sigma_n^2)$ is AWGN.

Because h_{BU} is negligible, the instantaneous SNR reduces to

$$\gamma = \frac{P_{\text{tx}} |h_{\text{RIS}}|^2}{\sigma_n^2}. \quad (9)$$

The achievable spectral efficiency (per complex dimension) is

$$C = \log_2(1 + \gamma). \quad (10)$$

These expressions provide the basis for the performance evaluation presented in Section III and are in line with standard capacity-based metrics used in RIS performance studies [3], [4].

III. SIMULATION RESULTS

This section evaluates the performance of the RIS-assisted NLoS architecture using Monte Carlo simulations conducted in MATLAB R2023a. The scenario corresponds to the geometry described in Section II, where the BS–UE direct path is fully blocked by an obstruction, and a planar RIS is deployed on the building facade to restore connectivity. All results are averaged over 500 independent Rayleigh fading realizations. We verified that this Monte Carlo budget is sufficient for numerical convergence: repeating the simulations of Figs. 2 and 3 with 1000 trials changed the average SNR and spectral-efficiency values by less than 0.1 dB across all UE positions and RIS sizes. We also observed that the empirical standard deviation of the SNR across fading realizations remained within a few dB in all cases, so the reported averages are representative and numerically stable.

A. Simulation Parameters

Table I summarizes the parameters used in all simulations.

TABLE I: Simulation Parameters

Parameter	Value
Carrier frequency f_c	3.5 GHz
Bandwidth B	20 MHz
Transmit power P_{tx}	30 dBm
Noise PSD N_0	−174 dBm/Hz
Receiver noise figure F	7 dB
Blockage attenuation	60 dB
Number of RIS elements N	16–1024
UE position range	60–140 m
Number of Monte Carlo trials	500

B. Coverage Performance Along the NLoS Region

Fig. 2 reports the average SNR as the UE moves deeper into the NLoS region, from 60 m to 140 m. Three configurations are compared: (i) no RIS, (ii) uncontrolled RIS with random phase shifts, and (iii) optimized RIS using coherent phase alignment as described in (6).

As seen in Fig. 2, the direct link without RIS remains severely degraded, with SNR values below −14 dB across the entire range. The uncontrolled RIS provides moderate improvement, typically gaining around 10 dB but remaining below 0 dB. In contrast, the optimized RIS maintains high-quality coverage throughout the NLoS zone, sustaining SNR values between 12 and 21 dB. This demonstrates that RIS-enabled reflections can fully restore connectivity even when the direct link is unusable.

C. Impact of RIS Size

To quantify the effect of RIS aperture size, Fig. 3 shows the average SNR for different numbers of reflecting elements ($N = 16, 64, 256$, and 1024), with the UE positioned at a representative NLoS location ($x = 120$ m).

The “no RIS” case remains nearly flat, confirming that the blocked direct link is insensitive to RIS aperture. The uncontrolled RIS exhibits a slow but steady increase in SNR,

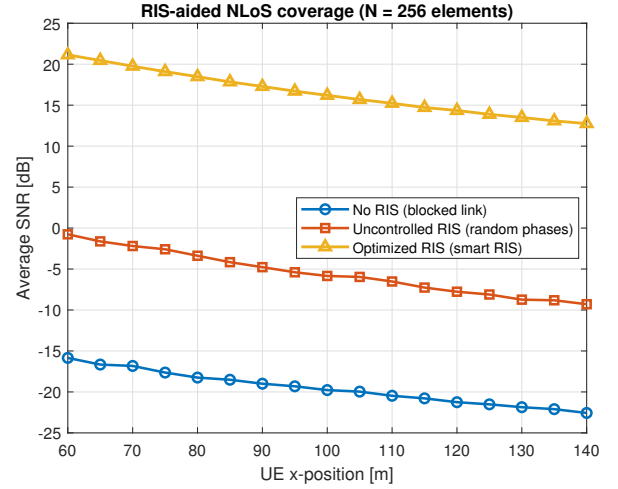


Fig. 2: Average SNR versus UE position in the NLoS region ($N = 256$ RIS elements).

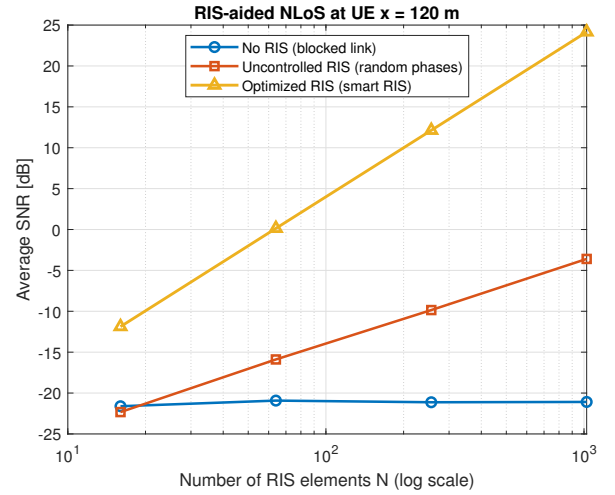


Fig. 3: Average SNR versus RIS size for a fixed NLoS UE location ($x = 120$ m).

consistent with the expected $\sqrt{N}$ growth from incoherent reflections. The optimized RIS, however, shows a much stronger response: the SNR increases from −12 dB at $N = 16$ to +24 dB at $N = 1024$, nearly following a linear trend in dB scale. This behavior is consistent with the N^2 coherent-power scaling discussed in Section II.

D. Spectral Efficiency Enhancement

Fig. 4 presents the corresponding spectral efficiency, computed using

$$C = \log_2(1 + \gamma).$$

The no-RIS case yields almost zero spectral efficiency due to extremely low SNR. The uncontrolled RIS achieves only modest improvements, reaching approximately 0.5 bit/s/Hz even with 1024 elements. In contrast, the optimized RIS enables significant throughput gains, achieving more than

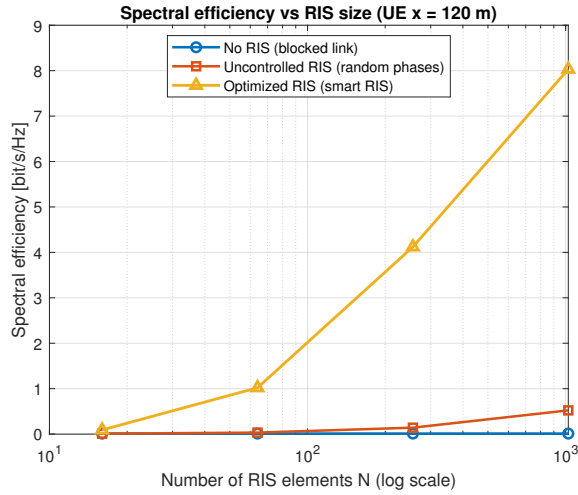


Fig. 4: Spectral efficiency versus RIS size for the UE at $x = 120$ m.

8 bit/s/Hz at $N = 1024$. These results confirm that RIS can transform a non-serviceable NLoS channel into a high-capacity link, provided that the surface is sufficiently large and its phase shifts are properly configured.

IV. DISCUSSION

The simulation results demonstrate that RIS-aided NLoS communication is highly effective in restoring coverage in scenarios where the direct BS–UE link is severely attenuated. Several key observations can be drawn from the analysis.

First, the performance gap between uncontrolled and optimized RIS configurations highlights the fundamental importance of phase alignment. Although even a non-programmable surface provides some improvement through diffuse scattering, its gains remain limited because the reflected paths add incoherently. In contrast, the optimized RIS achieves coherent combining, resulting in nearly N^2 power scaling, which directly translates into double-digit SNR gains and substantial spectral efficiency improvements.

Second, the results emphasize the role of RIS aperture size. Small surfaces provide only modest benefits, particularly in deep NLoS regions, while medium-to-large panels (hundreds to thousands of elements) can fully overcome blockage losses. This trend suggests that RIS installations intended for coverage restoration should prioritize physical aperture over element-level complexity.

Finally, the study focuses on an idealized continuous phase-control model and narrowband channels. Although these assumptions simplify the analysis, they also represent optimistic conditions. In practice, RIS phase shifts are implemented with finite-resolution tunable loads. Several works have shown that 2–3 bit phase quantization typically incurs only a small performance penalty: for example, 3-bit quantization can achieve almost the same sum rate as continuous phase shifts, while 2-bit quantization usually leads to a loss on the order of 1–2 dB in SNR or spectral efficiency [15], [18]. Consequently, the

nearly N^2 power scaling observed here would be slightly shifted downward but preserved, and the qualitative coverage-restoration trends are expected to remain valid under practical phase-resolution constraints.

An important practical aspect is the sensitivity of the results to RIS placement. In our setup, the RIS is mounted on the building facade at $z = 10$ m, roughly aligned with the BS and located on the obstructing structure (Fig. 1), which maximizes the cascaded BS–RIS–UE gain under free-space path loss. In realistic deployments, the optimal position and orientation depend strongly on the blockage layout, angles of arrival/departure, and link distances: moving the panel by only a few meters or tilting it by a few degrees can change the effective path loss by several dB. Recent deployment studies show that RIS panels have a limited “area of influence” in space and angle, so coverage gains are concentrated in specific regions around the designed reflection direction [2], [4]. As a result, the single-BS/single-UE geometry studied here should be interpreted as a best-case configuration that can be approached in practice when the RIS is carefully aligned with the dominant propagation paths.

The present study focuses on a single BS–RIS–UE link to isolate the fundamental coverage-restoration mechanism. In a multi-user downlink, the RIS phase profile and BS precoder must be jointly optimized to balance array gain and inter-user interference, so an N^2 gain for all users simultaneously is no longer achievable. Instead, the RIS is typically steered toward one or a small group of users while others experience weaker or partially constructive reflections, leading to more modest per-user gains but substantial improvements in aggregate sum rate and fairness compared to no-RIS baselines [17]–[19]. Our single-user results can thus be viewed as an upper bound on the SNR and spectral-efficiency improvements attainable for a scheduled user in a multi-user RIS-assisted cell.

V. CONCLUSION

This paper investigated the use of reconfigurable intelligent surfaces to restore downlink connectivity in NLoS urban environments. A simple yet realistic BS–RIS–UE geometry was analyzed using MATLAB-based Monte Carlo simulations. The results show that the direct BS–UE link becomes unusable under blockage, while an RIS deployed on the obstructing structure can fully reestablish coverage. Optimized phase control enables coherent combining at the UE, delivering more than 30 dB SNR improvement and increasing spectral efficiency from nearly zero to over 8 bit/s/Hz with a 1024-element panel. These findings highlight the significant potential of RIS as a low-power, cost-effective solution for improving 5G NLoS performance in dense urban deployments. Unlike active decode-and-forward or amplify-and-forward relays, a passive RIS does not include RF chains or power amplifiers; only the controller and tunable loads require power, which can be supplied at the milliwatt level and even via energy harvesting [12], [16]. In contrast, active relay nodes typically consume several watts to tens of watts due to baseband processing and RF amplification. From a network-level perspective, RIS

panels therefore constitute a genuinely low-power alternative for coverage enhancement.

Future work will consider more realistic RIS hardware models, phase quantization constraints, wideband propagation, and multi-user extensions, as well as system-level analyses under practical deployment constraints.

REFERENCES

- [1] M. Rossanese et al., "Open Experimental Measurements of Sub-6GHz Reconfigurable Intelligent Surfaces," arXiv, Cornell University, Apr. 2024. DOI: 10.48550/arxiv.2404.01796.
- [2] V. Tapio et al., "Reconfigurable Intelligent Surface for 5G NR Uplink Coverage Enhancement," in Proc. IEEE 94th Vehicular Technology Conference (VTC2021-Fall), Sept. 2021, pp. 1–5. DOI: 10.1109/vtc2021-fall52928.2021.9625081.
- [3] X. Yuan et al., "Reconfigurable-Intelligent-Surface Empowered Wireless Communications: Challenges and Opportunities," IEEE Wireless Communications, vol. 28, no. 2, pp. 136–143, Feb. 2021. DOI: 10.1109/mwc.001.2000256.
- [4] G. C. Alexandropoulos et al., "RIS-enabled smart wireless environments: deployment scenarios, network architecture, bandwidth and area of influence," EURASIP Journal on Wireless Communications and Networking, vol. 2023, no. 1, Oct. 2023. DOI: 10.1186/s13638-023-02295-8.
- [5] D. M. Mughal et al., "MALRIS: Malicious Hardware in RIS-Assisted Wireless Communications," arXiv, Cornell University, Aug. 2025. DOI: 10.48550/arxiv.2508.06340.
- [6] O. Maraqa et al., "Beyond Diagonal RIS-Aided Wireless Communications Systems: State-of-the-Art and Future Research Directions," arXiv, Cornell University, Mar. 2025. DOI: 10.48550/arxiv.2503.08826.
- [7] M. H. Khoshafa et al., "RIS-Assisted Physical Layer Security in Emerging RF and Optical Wireless Communication Systems: A Comprehensive Survey," arXiv, Cornell University, Mar. 2024. DOI: 10.48550/arxiv.2403.10412.
- [8] Q. Wu and R. Zhang, "Intelligent Reflecting Surface Aided Wireless Communication: IRS Channel Model, Beamforming, and Optimization," IEEE Transactions on Wireless Communications, vol. 19, no. 10, pp. 6653–6667, Oct. 2020. DOI: 10.1109/TWC.2020.2988135.
- [9] Q. Wu and R. Zhang, "Towards Smart and Reconfigurable Environment: Intelligent Reflecting Surface Aided Wireless Network," IEEE Communications Magazine, vol. 58, no. 1, pp. 106–112, Jan. 2020. DOI: 10.1109/MCOM.001.1900107.
- [10] E. Bjoernson et al., "Reconfigurable Intelligent Surfaces: A Signal Processing Perspective with Wireless Applications," IEEE Signal Processing Magazine, vol. 39, no. 2, pp. 135–158, Mar. 2022. DOI: 10.1109/MSP.2021.3130549.
- [11] M. Di Renzo et al., "Smart Radio Environments Empowered by Reconfigurable Intelligent Surfaces: A Tutorial and Survey," IEEE Journal on Selected Areas in Communications, vol. 38, no. 11, pp. 2450–2525, Nov. 2020. DOI: 10.1109/JSAC.2020.3007211.
- [12] E. Basar et al., "Wireless Communications Through Reconfigurable Intelligent Surfaces," IEEE Access, vol. 7, pp. 116753–116773, 2019. DOI: 10.1109/ACCESS.2019.2935192.
- [13] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," 3GPP TR 38.901, V17.0.0, Dec. 2022.
- [14] G. C. Alexandropoulos et al., "RIS-enabled smart wireless environments: deployment scenarios, network architecture, bandwidth and area of influence," EURASIP Journal on Wireless Communications and Networking, vol. 2023, no. 1, Oct. 2023. DOI: 10.1186/s13638-023-02295-8.
- [15] S. Eskandari et al., "Low-complexity beamforming design for RIS-aided MISO systems with finite-resolution phase shifting," in Proc. IEEE International Conference on Communications (ICC), May 2022, pp. 1–6. DOI: 10.1109/ICC45855.2022.9838671.
- [16] R. Liu et al., "Sustainable Wireless Networks via Reconfigurable Intelligent Surfaces (RISs): Overview of the ETSI ISG RIS," IEEE Communications Standards Magazine, vol. 9, no. 3, pp. 111–117, Sept. 2025. DOI: 10.1109/MCOMSTD.2025.3572259.
- [17] Q.-U.-A. Nadeem, H. Alwazani, A. Kammoun, A. Chaaban, M. Debbah, and M.-S. Alouini, "Intelligent reflecting surface assisted multi-user MISO communication: Channel estimation and beamforming design," IEEE Open Journal of the Communications Society, vol. 1, pp. 1–1, 2020. DOI: 10.1109/OJCOMS.2020.2992791.
- [18] W. Ma, L. Zhuo, L. Li, Y. Liu, and H. Ren, "Deep reinforcement learning for RIS-aided multiuser MISO system with hardware impairments," Applied Sciences, vol. 12, no. 14, p. 7236, 2022. DOI: 10.3390/app12147236.
- [19] P. Zeng, D. Qiao, and H. Qian, "Beamforming design for intelligent reflecting surface aided multi-antenna MU-MIMO communications with imperfect CSI," in Proc. IEEE/CIC International Conference on Communications in China (ICCC), Chongqing, China, 2020, pp. 12–17. DOI: 10.1109/ICCC49849.2020.9238961.

Comparative Analysis of Fractional PI Rectangular Rules and Classical RK4 Integration for High-Fidelity Synthetic ECG Signal Modeling

1st Abdelghani TAKHA *ETA Laboratory, Faculty of Sciences and Technology
University of Mohamed El Bachir El-Ibrahimi*

Bordj Bou Arreridj, Algeria
abdelghani.takha@univ-bba.dz

2nd Mohamed Lamine TALBI *ETA Laboratory, Faculty of Sciences and Technology
University of Mohamed El Bachir El-Ibrahimi*

Bordj Bou Arreridj, Algeria
mohamedlamine.talbi@univ-bba.dz

3rd Souad KSENNA *ETA Laboratory, Faculty of Sciences and Technology
University of Mohamed El Bachir El-Ibrahimi*

Bordj Bou Arreridj, Algeria
souad.ksenna@univ-bba.dz

Abstract—This work presents a comparative evaluation of three numerical solvers for synthetic ECG generation using a fractional extension of the McSharry model. The fractional-order formulation is integrated through explicit and implicit Product-Integration (PI) rectangular rules, while the classical fourth-order Runge-Kutta (RK4) method is retained in its integer-order form to isolate the effect of fractional dynamics.

The assessment is performed on two benchmark beat types from the MIT-BIH Arrhythmia Database: Normal Heart Beat (NHB, Record 106) and Junctional Heart Beat (JHB, Record 234). A total of 200 beats (100 per record) were used to quantify reconstruction fidelity through error- and similarity-based metrics. The implicit PI solver consistently achieves the best accuracy across both beat types, yielding the lowest DTW distances and the best distortion measures, including MAE = 0.0115, MSE = 3.30×10^{-4} , and PRD = 5.87 for NHB, with similarly improved values for JHB. It also achieves the highest structural and morphological consistency, with SSIM = 0.9761 and correlation = 0.9960. Visual comparison further confirms its superior ability to preserve P-QRS-T morphology for both regular (NHB) and irregular (JHB) rhythms.

Overall, the results demonstrate that fractional-order PI schemes, particularly the implicit formulation, provide significantly enhanced robustness, temporal alignment, and physiological realism compared to classical integer-order RK4, establishing them as powerful tools for high-fidelity ECG synthesis.

Index Terms—Synthetic ECG, McSharry model, Implicit product-integration rectangular rule, Explicit product-integration rectangular rule, Fourth-order Runge-Kutta method

I. INTRODUCTION

Accurate modeling of electrocardiogram (ECG) signals plays an important role in biomedical research, clinical analysis, and the development of signal-processing algorithms. Several mathematical formulations have been proposed to reproduce ECG waveforms, with the McSharry dynamical model emerging as one of the most widely used due to its

ability to generate physiologically meaningful PQRST patterns [1]. However, the fidelity of the simulated signal depends not only on the underlying model but also on the numerical strategy used to integrate its equations.

Classical integer-order solvers, such as the fourth-order Runge-Kutta (RK4) method, remain standard tools in biomedical signal modeling because of their simplicity and reliability [2]. Despite these advantages, such solvers may smooth or attenuate rapid transitions, particularly around the QRS complex, where sharper dynamics are essential for accurate ECG representation. This limitation has motivated interest in alternative numerical methods capable of preserving waveform morphology more effectively.

Product-Integration (PI) schemes represent one such approach and have been widely explored for their ability to approximate integral formulations with improved stability and structural preservation [3]–[6]. Among these, the implicit and explicit rectangular PI rules offer simple yet effective formulations. In this study, these PI rules are extended using concepts from fractional calculus, providing additional flexibility in capturing subtle variations in ECG dynamics. By incorporating fractional behavior into the numerical approximation, the PI-based solvers can better reflect the long-term dependencies and fine morphological details often present in cardiac signals.

In contrast, the RK4 solver is retained in its classical integer-order form, enabling a fair comparison between fractional PI-based methods and a traditional ODE integration strategy. Evaluating these solvers side by side is essential for determining how fractional numerical formulations influence modeling accuracy, temporal alignment, and overall ECG waveform fidelity.

Motivated by these considerations, this work presents a systematic comparison of three numerical strategies for synthetic ECG generation: the fractional implicit PI rectangular rule, the

fractional explicit PI rectangular rule, and the classical fourth-order Runge–Kutta method. Using the McSharry dynamical model as a benchmark, the study examines each solver's ability to reproduce ECG morphology using widely applied similarity measures that quantify distortion, structural accuracy, and timing consistency. Through this analysis, the work highlights the impact of both fractional integration and solver design on the quality of synthetic ECG generation.

II. RELATED WORK

Early approaches to synthetic ECG generation were largely based on classical differential equation models, most notably the framework introduced by McSharry and colleagues [1]. These models have been widely adopted for producing realistic PQRST morphologies, yet their reliance on integer-order dynamics limits their ability to represent certain temporal behaviors observed in real cardiac signals.

To overcome this constraint, recent studies have explored the use of fractional calculus in biomedical modeling, showing that fractional-order operators can capture dynamics that evolve over time with greater flexibility [7], [8]. This motivation has encouraged the development of fractional extensions of well-established ECG models, offering improved adaptability in reproducing waveform variations.

Our earlier work [9] followed this direction by introducing a fractional variant of the McSharry model and highlighting the potential benefits of using fractional-order numerical techniques. In particular, Product-Integration (PI) methods—originally proposed for integral equations and later adapted for fractional systems—have shown promise as practical tools for solving models that include memory-dependent components [10]. Several PI variants exist, ranging from simple rectangular formulations to more advanced predictor–corrector strategies [11].

Although the mathematical properties of PI schemes are well documented [12]–[15], their comparative performance within ECG modeling has not been extensively evaluated. Most previous studies either applied a single method or focused primarily on demonstrating the advantages of fractional calculus. This motivates the present contribution, which offers a targeted comparison of different PI-based solvers in the context of ECG signal generation.

III. METHODOLOGY

A. Components and Significance of ECG Intervals

[18]

The PQRST complex in an electrocardiogram (ECG) reflects the sequence of electrical events occurring during a single heartbeat, comprising the P-wave, QRS complex, and T-wave. Each component corresponds to a specific phase of cardiac activity. The PR interval indicates the time from the onset of the P-wave to the start of the QRS complex, representing atrial depolarization and conduction through the atrioventricular node. The ST segment corresponds to the period between ventricular depolarization and repolarization, while the QT interval encompasses the total duration of ventricular depolarization and repolarization. The RR interval,

defined as the time between consecutive R peaks, is commonly used to determine the heart rate. These intervals and their roles are depicted in Fig. 1.

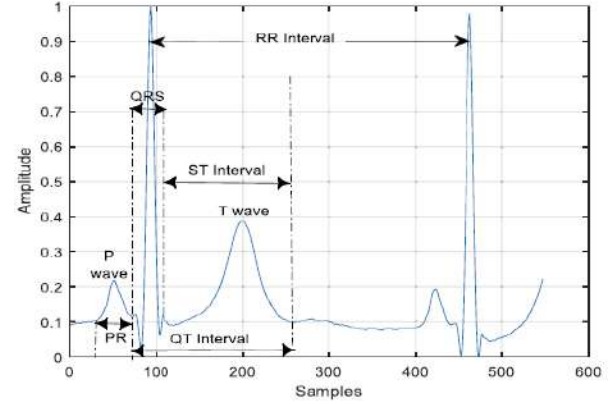


Fig. 1. Illustration of ECG signal PQRST complex and RR interval. [9]

B. Fractional-Order ECG Model

To evaluate the numerical solvers examined in this study, we rely on the structure of the well-established McSharry dynamical model, which generates synthetic ECG waveforms through a limit-cycle trajectory combined with amplitude modulation around the P, Q, R, S, and T landmarks. In addition to its classical integer-order formulation, a fractional-order version of the model—introduced in our earlier work [9]—is used to emphasize the impact of memory effects on ECG morphology.

In the present contribution, the fractional-order formulation is employed exclusively when applying Product-Integration (PI) methods, while the classical fourth-order Runge–Kutta (RK4) solver is applied to the standard integer-order system. This allows a balanced comparison between:

- PI solvers operating on the fractional-order model, and
- RK4 operating on the original integer-order form.

The underlying state dynamics follow the structure of the McSharry model, where the first two state variables evolve on a circular limit cycle governed by

$$\dot{x}_1 = \beta x_1 - \omega x_2, \quad \dot{x}_2 = \beta x_2 + \omega x_1, \quad (1)$$

and the amplitude variable is shaped through Gaussian functions centered at the characteristic ECG landmarks. Respiratory modulation is included through a slow baseline component $z_0(t)$.

The model is formulated in (2), where $D_{t_0}^\alpha$ denotes the fractional differentiation operator [7].

$$\begin{cases} D_{t_0}^1 x_1(t) = \beta x_1 - \omega x_2, \\ D_{t_0}^1 x_2(t) = \beta x_2 + \omega x_1, \\ D_{t_0}^\alpha x_3 = \sum_{i \in \{P, Q, R, S, T\}} a_i \Delta \theta_i \exp\left(-\frac{\Delta \theta_i^2}{2b_i^2}\right) - (x_3 - z_0). \end{cases} \quad (2)$$

with $\beta = 1 - \sqrt{x_1^2 + x_2^2}$, $\Delta\theta_i = (\theta - \theta_i) \bmod 2\pi$, $\theta = \text{atan2}(x_1, x_2)$, where ω is the angular velocity as it moves around the limit cycle. The value of ω corresponds to the angular frequency Ω_n , derived from the beat-to-beat intervals of the internally generated RR tachogram:

$$\Omega_n = \frac{2\pi}{T(t_n)}, \quad (3)$$

where $T(t_n)$ represents the time interval between consecutive heartbeats. Thus, $\omega(t) = \Omega_n$, $t_n \leq t \leq t_{n+1}$.

C. Numerical Methods Under Comparison

This study compares three numerical strategies commonly used for solving differential systems in ECG simulation:

1) *Explicit PI Rectangular Rule (Fractional Order)*: [8], [16] The explicit PI method approximates the integral form of the fractional-order model using only previously computed values. For a uniform step size h , the numerical update takes the form

$$y_n = T_{m-1}[y, t_0](t_n) + h^\alpha \sum_{j=0}^{n-1} b_{n-j-1}^{(\alpha)} f(t_j, y_j) \quad (4)$$

where α represents the fractional order introduced in [9]. This approach offers low computational cost and straightforward implementation.

2) *Implicit PI Rectangular Rule (Fractional Order)*: [8], [16] The implicit formulation uses the same integral structure but incorporates the derivative at the current time step:

$$y_n = T_{m-1}[y, t_0](t_n) + h^\alpha \left(\tilde{a}_n^{(\alpha)} f_0 + \sum_{j=1}^n a_{n-j}^{(\alpha)} f(t_j, y_j) \right) \quad (5)$$

requiring the solution of a nonlinear equation at each iteration. This method generally improves numerical stability, especially when modeling sharp ECG transitions such as the QRS complex.

3) *Fourth-Order Runge-Kutta (Integer Order)*: [2] As a baseline reference, the classical RK4 scheme is applied to the original integer-order McSharry model. The update rule follows the well-known formulation

$$x_{n+1} = x_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4), \quad (6)$$

and provides reliable accuracy for smooth regions of the waveform.

D. Evaluation Metrics

The performance of the numerical solvers was assessed using a comprehensive set of metrics that quantify temporal alignment, morphological fidelity, and statistical similarity between the reference ECG signal, denoted as ECG_{real} , and the reconstructed signal, denoted as $\text{ECG}_{\text{simulated}}$. The metrics employed are detailed below.

- **Dynamic Time Warping Distance (dist_DTW)**: DTW provides a measure of temporal alignment by computing the minimum cumulative discrepancy between the two signals after optimal nonlinear time warping. Given ECG_{real} and $\text{ECG}_{\text{simulated}}$, the DTW distance is defined as:

$$\text{dist\_DTW} = \min_{\pi} \sum_{(i,j) \in \pi} |\text{ECG}_{\text{real},i} - \text{ECG}_{\text{simulated},j}|, \quad (7)$$

where π is any admissible warping path respecting monotonicity and continuity constraints. Lower DTW values indicate improved temporal consistency.

- **Structural Similarity Index (ssimval)**: SSIM evaluates morphological preservation by combining luminance, contrast, and structural similarity. For the two signals, SSIM is computed as:

$$\text{ssimval} = \frac{(2\mu_{\text{real}}\mu_{\text{sim}} + C_1)(2\sigma_{\text{real},\text{sim}} + C_2)}{(\mu_{\text{real}}^2 + \mu_{\text{sim}}^2 + C_1)(\sigma_{\text{real}}^2 + \sigma_{\text{sim}}^2 + C_2)}, \quad (8)$$

where μ , σ^2 , and $\sigma_{\text{real},\text{sim}}$ denote signal means, variances, and covariance, respectively. Higher SSIM values reflect stronger morphological fidelity.

- **Mean Absolute Error (MAE)**: Measures the average absolute difference:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\text{ECG}_{\text{real},i} - \text{ECG}_{\text{simulated},i}|. \quad (9)$$

- **Mean Squared Error (MSE)**: Quantifies the squared deviation:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (\text{ECG}_{\text{real},i} - \text{ECG}_{\text{simulated},i})^2. \quad (10)$$

- **Cosine Similarity (cosSim)**: Captures the angular similarity between the two ECG vectors:

$$\text{cosSim} = \frac{\sum_{i=1}^n \text{ECG}_{\text{real},i} \text{ECG}_{\text{simulated},i}}{\sqrt{\sum_{i=1}^n \text{ECG}_{\text{real},i}^2} \sqrt{\sum_{i=1}^n \text{ECG}_{\text{simulated},i}^2}}. \quad (11)$$

Values close to 1 indicate high directional similarity.

- **Signal-to-Noise Ratio (SNR)**: Expresses the ratio between the reference signal energy and reconstruction error:

$$\text{SNR} = 10 \log_{10} \left(\frac{\sum_{i=1}^n \text{ECG}_{\text{real},i}^2}{\sum_{i=1}^n (\text{ECG}_{\text{real},i} - \text{ECG}_{\text{simulated},i})^2} \right). \quad (12)$$

Higher SNR values signify improved reconstruction quality.

- **Correlation Coefficient (CORR)**: Measures the linear dependence between real and simulated ECGs:

$$\text{CORR} = \frac{\sum_{i=1}^n (\text{ECG}_{\text{real},i} - \mu_{\text{real}})(\text{ECG}_{\text{simulated},i} - \mu_{\text{sim}})}{\sigma_{\text{real}} \cdot \sigma_{\text{sim}}} \quad (13)$$

CORR values near 1 indicate strong linear similarity.

- **Percentage Root Difference (PRD):** A normalized measure of distortion expressed as:

$$\text{PRD} = 100 \times \sqrt{\frac{\sum_{i=1}^n (\text{ECG}_{\text{real},i} - \text{ECG}_{\text{simulated},i})^2}{\sum_{i=1}^n \text{ECG}_{\text{real},i}^2}}. \quad (14)$$

Lower PRD values reflect better reconstruction fidelity.

Together, these metrics provide a robust and complementary set of criteria for assessing the accuracy, temporal alignment, and structural preservation achieved by each numerical solver.

E. ECG Beat Types and Dataset

To ensure a reliable benchmark for evaluating the numerical solvers, the analysis was performed using two well-established ECG recordings from the MIT-BIH Arrhythmia Database [17]. Specifically, we analyzed a total of 200 ECG beats, selecting two records that represent distinct cardiac conditions:

Record 106 was selected for Normal Heart Beats (NHB).

Record 234 was selected for Junctional Heart Beats (JHB).

From each record, 100 consecutive beats were extracted, resulting in a total of 200 beats used for the analysis.

IV. RESULTS AND DISCUSSION

All simulations were performed in MATLAB (R2021a) on a computer equipped with an Intel i5-4200 processor (2.50 GHz) running Windows 10. The McSharry dynamical model was initialized using the optimized physiological parameter set obtained in our previous work through a Genetic Algorithm, including the fractional order $\alpha = 0.98$ and the parameters (a_i, b_i, θ_i) [12], ensuring realistic reference morphology for detailed solver comparison. Normal Heart Beat (NHB) data from the MIT-BIH Arrhythmia Database (Record 106) were used as the reference signal for quantitative evaluation [9].

We evaluated three numerical schemes: the Explicit Product-Integration (PI) rectangular rule, the Implicit PI rectangular rule, and the classical fourth-order Runge–Kutta (RK4) method applied to the integer-order model. For PI methods, the fractional-order formulation of the McSharry model was used to capture memory-dependent dynamics, while RK4 served as the standard ODE benchmark.

A. Interpretation of Results

1) A. Performance Comparison of Numerical Methods:

The quantitative evaluation in Table I compares RK4 (integer-order), Explicit PI, and Implicit PI rectangular rules for two ECG beat types (JHB and NHB). For the JHB dataset, the Implicit PI method achieves the highest accuracy, with the lowest DTW distance (1.0008 ± 0.1854), MSE (4.2×10^{-4}), MAE (0.0132 ± 0.0025), and PRD (7.71 ± 1.66). The Explicit PI solver provides competitive accuracy and attains the fastest execution time (37.52 ± 9.28 s). In contrast, RK4 produces the highest reconstruction errors (e.g., PRD = 19.48 ± 1.78) and the longest runtime (96.07 ± 54.98 s), demonstrating its limitations in fractional ECG modeling.

For the NHB record, which exhibits more regular morphology, the Explicit and Implicit PI schemes show closer performance. The Explicit PI method achieves the best PRD (7.25 ± 1.49), the highest SNR (23.04 ± 1.77 dB), and the largest cosine similarity (0.9986 ± 0.0012), while the Implicit PI solver maintains competitive error levels with moderate execution time. Across both datasets, RK4 consistently underperforms relative to the fractional-order PI-based methods.

2) *B. Interpretation of Results:* Overall, the Implicit PI rectangular rule provides the most accurate and robust reconstruction—particularly for the more irregular JHB beats—due to its enhanced treatment of long-memory effects. The Explicit PI method offers an excellent balance between computational efficiency and accuracy, making it suitable for real-time or large-scale applications. In contrast, the classical RK4 solver is unable to capture the non-local temporal behavior inherent to ECG dynamics, leading to higher reconstruction errors and weaker temporal alignment.

These findings demonstrate that fractional-order solvers provide more physiologically consistent ECG waveform reconstruction than integer-order numerical schemes.

3) *C. Impact of Fractional-Order Dynamics:* The superior performance of PI-based schemes is directly linked to the fractional-order McSharry formulation. Fractional operators embed long-range temporal correlations and memory effects characteristic of cardiac electrophysiology, enabling smoother waveform transitions and improved representation of P–QRS–T morphology. This is reflected in the higher SSIM values obtained by fractional methods (e.g., 0.9438 for JHB Implicit PI vs. 0.8589 for JHB RK4).

The Implicit PI method effectively handles stiffness and long-memory coupling, which explains its superior performance for complex rhythms. The Explicit PI solver remains computationally efficient while still exploiting fractional temporal dynamics. Conversely, the integer-order RK4 method lacks the ability to reproduce these non-local interactions, resulting in poorer accuracy across all metrics.

These results confirm that fractional-order modeling combined with PI rectangular integration provides a more accurate, stable, and physiologically realistic framework for synthetic ECG generation.

Visual comparisons for both beat types—Normal Heart Beat (NHB) and Junctional Heart Beat (JHB)—are presented in Fig. 2. Each subfigure displays the real ECG beat together with the synthetic waveform generated by each numerical method, enabling direct evaluation of morphological fidelity. For NHB, the Implicit PI scheme provides the closest alignment with the reference signal, accurately preserving the P-wave, QRS complex, and T-wave structures. For JHB, characterized by altered atrioventricular conduction and reduced P-wave activity, the Implicit PI method again yields the lowest reconstruction error. The error values reported in the figure confirm that, for both beat types, the Implicit PI solver achieves the most accurate reproduction of the underlying cardiac dynamics among all tested numerical schemes.

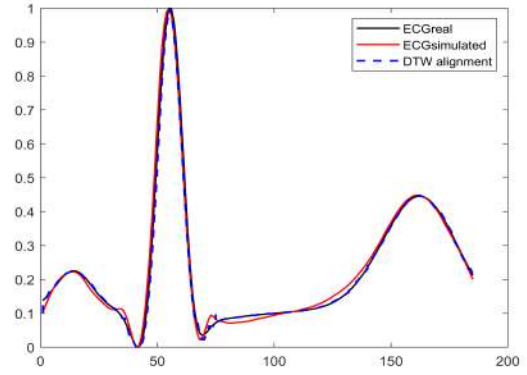
TABLE I
COMPARISON OF NUMERICAL METHODS FOR JUNCTIONAL (JHB) AND NORMAL (NHB) ECG RECORDS

Record / Method	DTW	MSE	MAE	CORR	SNR (dB)	SSIM	Cos. Sim.	PRD	Time (s)
JHB Record									
RK4 (ODE)	1.18872 ± 0.27789	0.00065 ± 0.00030	0.01952 ± 0.00467	0.99567 ± 0.00104	20.54678 ± 1.58649	0.85891 ± 0.05485	0.99564 ± 0.00151	19.4819 ± 1.7802	96.0677 ± 54.9816
Implicit PI Rect.	1.00078 ± 0.18544	0.00042 ± 0.00019	0.01319 ± 0.00253	0.99589 ± 0.00192	22.43288 ± 1.75199	0.94379 ± 0.03155	0.99701 ± 0.00150	7.71441 ± 1.66255	57.5979 ± 14.9056
Explicit PI Rect.	1.00273 ± 0.18671	0.00046 ± 0.00015	0.01384 ± 0.00247	0.99552 ± 0.00154	21.93695 ± 1.41314	0.94339 ± 0.03143	0.99677 ± 0.00104	8.10249 ± 1.27330	37.5171 ± 9.2791
NHB Record									
RK4 (ODE)	1.21343 ± 0.17406	0.00059 ± 0.00020	0.01763 ± 0.00332	0.99350 ± 0.00230	22.73804 ± 1.33194	0.96017 ± 0.00685	0.99737 ± 0.00087	7.38351 ± 1.18447	101.9215 ± 40.4992
Implicit PI Rect.	1.20889 ± 0.19366	0.00054 ± 0.00022	0.01702 ± 0.00238	0.99211 ± 0.00288	22.33581 ± 1.49616	0.96134 ± 0.00632	0.99695 ± 0.00099	7.75238 ± 1.29217	53.6826 ± 29.3474
Explicit PI Rect.	1.21650 ± 0.18501	0.00057 ± 0.00022	0.01602 ± 0.00273	0.99511 ± 0.00284	23.03581 ± 1.76766	0.95134 ± 0.00876	0.99855 ± 0.00115	7.25238 ± 1.48602	34.4877 ± 18.1282

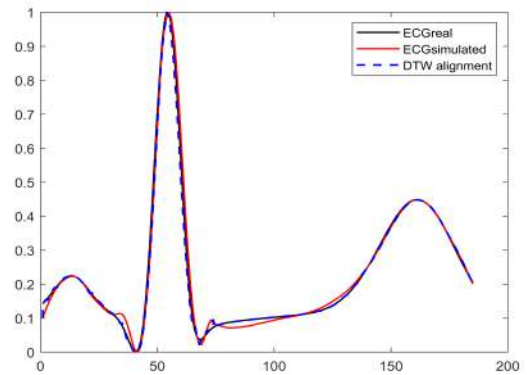
V. CONCLUSION

This study highlights the critical impact of numerical solver selection on the successful application of fractional-order differential equations in high-fidelity biomedical signal modeling. Through a comparative analysis of fractional PI rectangular rules and the classical RK4 integration, we demonstrated that PI-based schemes, particularly the Implicit PI method, outperform traditional integer-order solvers in capturing the long-memory dynamics and non-local behavior of cardiac electrophysiology. These solvers provide enhanced numerical stability, improved morphological fidelity, and more physiologically consistent synthetic ECG signals.

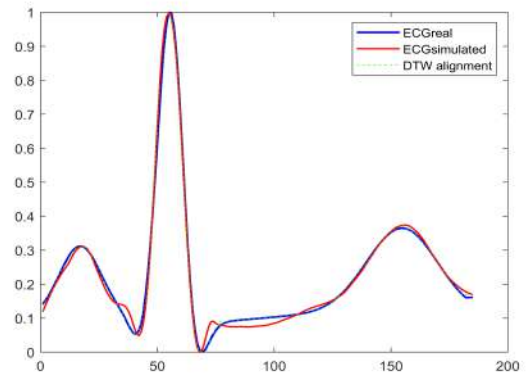
The findings establish fractional PI methods as a robust and reliable framework for modeling non-stationary biomedical signals, offering clear advantages over classical integration techniques. Beyond ECG modeling, this approach has potential applications in other physiological signals characterized by long-range correlations. Future work will explore real-time implementations, optimization of solver parameters, and the integration of fractional PI schemes with machine learning strategies to further enhance scalability and clinical relevance.



(a) ECG temporal alignment using DTW with the Explicit PI Rectangular Rule. MSE = 4.4005×10^{-4} , dist_DTW = 1.2837, PRD = 6.7846.

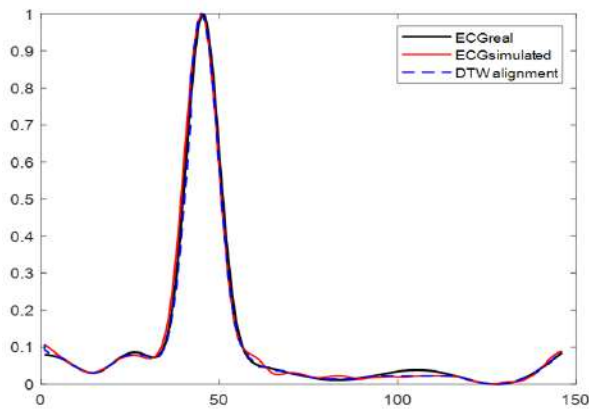


(b) DTW-based ECG alignment using the Implicit PI Rectangular Rule. MSE = 3.2950×10^{-4} , dist_DTW = 1.2759, PRD = 5.8708.

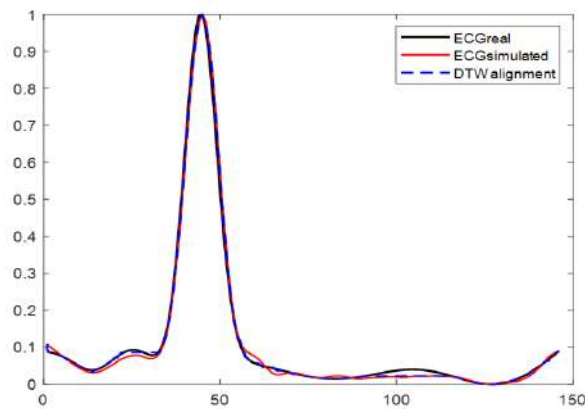


(c) ECG reconstruction using the fourth-order Runge-Kutta (RK4) ODE solver. MSE = 4.9596×10^{-4} , dist_DTW = 1.3716, PRD = 7.1979.

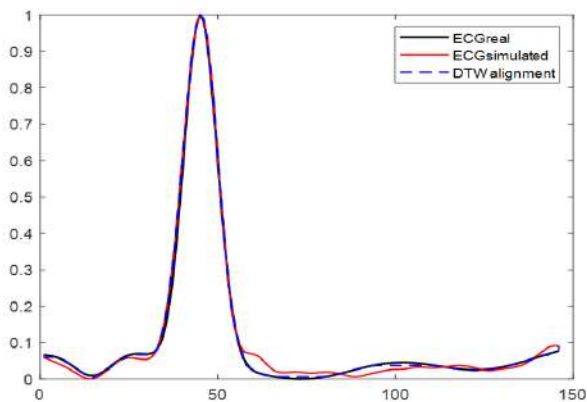
Fig. 2. Comparison of the Normal Heartbeat (NHB) waveform with its reconstructed versions generated using the Explicit and Implicit PI Rectangular Rules and the RK4 ODE solver.



(a) ECG temporal alignment using DTW with the Explicit PI Rectangular Rule. $MSE = 2.1804 \times 10^{-4}$, $dist\_DTW = 1.2837$, $PRD = 6.0043$.



(b) DTW-based ECG alignment using the Implicit PI Rectangular Rule. $MSE = 2.1804 \times 10^{-4}$, $dist\_DTW = 1.2837$, $PRD = 7.5865$.



(c) ECG reconstruction using the fourth-order Runge-Kutta (RK4) ODE solver. $MSE = 2.6617 \times 10^{-4}$, $dist\_DTW = 1.3716$, $PRD = 6.5352$.

Fig. 3. Comparison of the Junctional Heartbeat (JHB, Record 234) waveform with its reconstructed versions generated using the Explicit PI Rectangular Rule, Implicit PI Rectangular Rule, and the classical RK4 ODE solver.

REFERENCES

[1] P. E. McSharry, G. D. Clifford, L. Tarassenko, and L. A. Smith, "A dynamical model for generating synthetic electrocardiogram signals," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 3, pp. 289–294, Mar. 2003.

[2] Ch. Lubich, "Runge-Kutta theory for Volterra and Abel integral equations of the second kind," *Mathematics of Computation*, vol. 41, no. 163, pp. 87–102, 1983.

[3] K. Diethelm, *The Analysis of Fractional Differential Equations*, Lecture Notes in Mathematics. Berlin, Germany: Springer, 2010.

[4] F. Mainardi and R. Gorenflo, "Time-fractional derivatives in relaxation processes: a tutorial survey," *arXiv*, vol. 10, pp. 269–308, 2007.

[5] R. Garrappa, "Numerical solution of fractional differential equations: A survey and a software tutorial," *Mathematics*, vol. 6, no. 16, 2018.

[6] K. Diethelm, R. Garrappa, and M. Stynes, "Good (and not so good) practices in computational methods for fractional calculus," *Mathematics*, vol. 8, no. 3, p. 324, 2020.

[7] A. Ahmad, M. Farman, P. A. Naik, N. Zafar, A. Akgul, and M. U. Saleem, "Modeling and numerical investigation of fractional-order bovine babesiosis disease," *Numerical Methods for Partial Differential Equations*, vol. 37, no. 3, pp. 1946–1964, 2020, doi: 10.1002/num.22632.

[8] D. J. Templos-Hernández, L. A. Quezada-Téllez, B. M. González-Hernández, G. Rojas-Vite, J. E. Pineda-Sánchez, G. Fernández-Anaya, et al., "A fractional-order approach to cardiac rhythm analysis," *Chaos, Solitons & Fractals*, vol. 147, p. 110942, 2021.

[9] A. Takha, M. L. Talbi, and P. Ravier, "Fractional calculus integration for improved ECG modeling: A McSharry model expansion," *Medical Engineering & Physics*, vol. 132, p. 104237, 2024.

[10] A. Young, "Approximate product-integration," *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 224, no. 1156, pp. 552–561, 1954.

[11] R. Garrappa, "Numerical solution of fractional differential equations: A survey and a software tutorial," *Mathematics*, vol. 6, no. 2, p. 16, 2018.

[12] J. Dixon, "On the order of the error in discretization methods for weakly singular second kind Volterra integral equations with nonsmooth solutions," *BIT Numerical Mathematics*, vol. 25, no. 4, pp. 624–634, 1985.

[13] K. Diethelm, "Smoothness properties of solutions of Caputo-type fractional differential equations," *Fractional Calculus and Applied Analysis*, vol. 10, no. 2, pp. 151–160, 2007.

[14] K. Diethelm, N. J. Ford, and A. D. Freed, "Detailed error analysis for a fractional Adams method," *Numerical Algorithms*, vol. 36, no. 1, pp. 31–52, 2004.

[15] R. Garrappa, "Trapezoidal methods for fractional differential equations: theoretical and computational aspects," *Mathematics and Computers in Simulation*, vol. 110, pp. 96–112, 2015.

[16] R. Garrappa, "Numerical solution of fractional differential equations: A survey and a software tutorial," *Mathematics*, vol. 6, no. 2, p. 16, 2018.

[17] G. B. Moody, "MIT-BIH Arrhythmia Database Directory; Hypertext edition, 24 May 1997 (Based on the printed third edition, 23 July 1992)," Harvard-MIT Division of Health Sciences and Technology, Biomedical Engineering Center, Cambridge, MA, USA, 1997.

[18] G. D. Clifford, F. Azuaje, P. McSharry, *Advanced Methods and Tools for ECG Data Analysis*, vol. 10, Boston, MA: Artech House, 2006.

Deep Learning-Based Classification of Intestinal Diseases: A Comparative Analysis of ConvNeXt and CNN Models

Faiza BELLA
*LIMOSE Laboratory,
 University of Boumerdes,
 Boumerdes Algeria*
f.bella@univ-boumerdes.dz

Ali BERRICHI
*LIMOSE Laboratory,
 University of Boumerdes,
 Boumerdes Algeria*
ali.berrichi@univ-boumerdes.dz

Abdelouahab MOUSSAOUI
*Computer Science Department,
 University of Ferhat Abbes,
 Setif Algeria*
abdelouahab.moussaoui@univ-setif.dz

Abstract—The accurate classification of intestinal diseases is critical for early diagnosis and effective treatment planning. Traditional diagnostic methods rely heavily on manual interpretation of endoscopic or imaging data, which can be time-consuming and prone to human error. In this study, we propose a deep learning-based approach for the automated classification of intestinal diseases, comparing the performance of conventional Convolutional Neural Networks (CNNs) with the modern ConvNeXt architecture. A curated dataset of intestinal images was preprocessed through normalization, augmentation, and resizing to enhance model generalization. Both models were trained and evaluated using identical experimental settings to ensure a fair comparison. Performance was assessed using accuracy, precision, recall, F1-score, and confusion matrix analysis. The experimental results demonstrate that ConvNeXt consistently outperforms traditional CNNs in both classification accuracy and feature extraction efficiency, owing to its advanced design and improved representation learning capabilities. The findings highlight the potential of ConvNeXt as a powerful tool for computer-aided diagnosis of intestinal diseases and provide valuable insights into the integration of next-generation deep learning architectures in medical image analysis.

Index Terms—Intestinal Disease Classification, Deep Learning, Convolutional Neural Networks, ConvNeXt, Medical Image Analysis, Computer-Aided Diagnosis.

I. INTRODUCTION

Intestinal diseases, including conditions such as Crohn's disease, ulcerative colitis, and colorectal cancer, represent a major global health concern due to their increasing prevalence and diagnostic complexity. Early and accurate detection of these disorders is essential for effective treatment and improved patient outcomes. Traditionally, diagnosis relies on clinical evaluation and endoscopic imaging interpreted by medical experts, a process that is time-consuming and subject to inter-observer variability [1]. Therefore, automated image-based diagnostic systems have gained significant attention as a means to assist clinicians and enhance diagnostic accuracy.

Recent advancements in deep learning (DL), particularly Convolutional Neural Networks, have revolutionized medical image analysis by enabling automatic feature extraction and classification with high precision [2]. CNNs have been successfully applied to gastrointestinal (GI) imaging for detecting lesions, polyps, and other intestinal abnormalities [3]. However, conventional CNN architectures often struggle to capture complex contextual information and may exhibit limited generalization when applied to diverse datasets. To address these limitations, ConvNeXt has emerged as a promising alternative, a modernized convolutional architecture influenced by the design principles of Vision Transformers (ViTs) [4]. As a next-generation CNN, ConvNeXt re-engineers conventional convolutional structures by incorporating transformer-inspired components such as depth-wise convolutions, enlarged kernel sizes, layer normalization, and an inverted bottleneck design. These architectural enhancements significantly strengthen feature extraction and representation learning while maintaining the computational efficiency and structural simplicity of CNNs. Consequently, ConvNeXt achieves competitive or even superior performance compared to transformer-based models across a range of computer vision applications. Its scalability, robustness, and seamless integration with existing CNN frameworks make ConvNeXt highly suitable for advanced medical image analysis tasks, including the classification and diagnosis of intestinal diseases. This study aims to conduct a comparative analysis between traditional CNN architectures and ConvNeXt for intestinal disease classification using medical imaging datasets. The objectives include (1) evaluating the classification accuracy, precision, recall, and F1-score of both models; (2) analyzing the effect of model architecture on feature extraction; and (3) assessing the potential of ConvNeXt as a next-generation deep learning approach for medical image diagnosis. The findings of this research are expected to contribute to the development of more reliable and efficient computer-aided diagnosis (CAD) systems for gastrointestinal diseases.

II. LITERATURE REVIEW

The application of deep learning to medical image analysis has rapidly advanced in recent years, significantly improving diagnostic accuracy and reducing human error. Early studies predominantly focused on traditional Convolutional Neural Networks, which have proven highly effective for image classification and object detection tasks. In the context of gastrointestinal and intestinal disease analysis, CNNs have been used to detect abnormalities such as ulcers, bleeding, and polyps with promising results [5]. However, the performance of CNN-based systems often depends heavily on dataset size, preprocessing quality, and architectural depth. Several researchers have explored CNN variations and optimization strategies for gastrointestinal image classification. For instance, Urban et al. [3] developed a deep CNN capable of real-time polyp detection in colonoscopy videos, achieving over 90% sensitivity. Similarly, Zhang et al. [6] applied a ResNet-based model to classify colorectal lesions, demonstrating improved feature extraction compared to shallower architectures. Despite these advances, CNNs remain limited by their receptive field size and their inability to capture long-range dependencies, which is critical for accurately identifying subtle texture patterns in medical images. The emergence of transformer-based architectures, such as the Vision Transformer, introduced a paradigm shift by modeling global image relationships using self-attention mechanisms [7]. These architectures outperform conventional CNNs in many computer vision tasks but often require large-scale datasets and substantial computational resources, limiting their direct application in medical imaging. To bridge this gap, Liu et al. [8] introduced ConvNeXt, a modernized convolutional network that integrates key design principles from transformers while retaining the efficiency and inductive bias of CNNs. ConvNeXt employs larger convolution kernels, layer normalization, and improved block designs, enabling it to achieve state-of-the-art performance on ImageNet classification benchmarks. Recent comparative studies have highlighted the effectiveness of next-generation CNNs, including ConvNeXt, for medical image analysis. Khan et al. [9] compared ConvNeXt with EfficientNet and DenseNet for skin lesion classification, reporting superior accuracy and robustness with ConvNeXt. However, its application to intestinal disease classification remains relatively unexplored. Given the visual complexity of intestinal imagery, which is characterized by varying lighting, textures, and morphological patterns, a model capable of capturing both local and global image features is essential for accurate diagnosis.

III. METHODOLOGY

This section describes the methodology adopted to implement and evaluate the proposed deep learning framework for intestinal disease classification.

A. Proposed approach

Figure 1 illustrates the overall workflow of the proposed approach, outlining each stage from data acquisition to final model evaluation. To ensure a fair and consistent comparison

across all architectures, a unified experimental protocol was adopted, where every model was trained and tested under identical preprocessing steps, hyperparameters, and evaluation settings.

B. Selection of Model Architectures

We selected a diverse set of architectures covering both modern ConvNeXt variants and classical CNNs. Specifically, we used ConvNeXt-Tiny [4], ConvNeXt-Base [4], VGG16 [10], ResNet50 [11], and DenseNet121 [12] to examine the influence of architectural depth, feature extraction capacity, and design paradigms on classification performance. The ConvNeXt models were chosen due to their strong performance and modernized convolutional design, while VGG16, ResNet50, and DenseNet121 were included as representative conventional architectures commonly used in medical image analysis. This comparative analysis allows us to determine the suitability of each model for intestinal disease classification.

C. Dataset Description and Image Preprocessing

In this study, we utilized the publicly available KvasirV2 dataset [13], which contains 8,000 labeled gastrointestinal endoscopic images representing a variety of clinical findings. For our experiments, we included the following eight classes, assigning 1000 images to each class: Dyed-lifted polyps, Dyed-resection margins, Esophagitis, Normal-cecum, Normal-pylorus, Normal-z-line, Polyps, and Ulcerative colitis. The dataset consists of the images with different resolution from 720x576 up to 1920x1072 pixels. Figure 2 presents sample images from the KvasirV2 dataset, showing one representative example from each class to highlight the visual diversity and characteristic features of the gastrointestinal conditions.

To improve the robustness of the models and mitigate overfitting, various data-augmentation techniques were applied during the training phase. These augmentations introduce controlled variability into the dataset, enabling the models to generalize better to unseen endoscopic images. The applied transformations include geometric modifications such as random horizontal and vertical flipping, slight rotations, and random resized cropping, as well as photometric adjustments including brightness, contrast, and saturation variations. Gaussian blurring was also used to simulate realistic imaging noise. A summary of the augmentation strategies is provided in Table I.

TABLE I
DATA AUGMENTATION TECHNIQUES APPLIED DURING TRAINING WITH
PARAMETER VALUES.

Technique	Values / Parameters
Random Horizontal Flip	Probability = 0.5
Random Vertical Flip	Probability = 0.5
Random Rotation	Angle range = $\pm 15^\circ$
Random Resized Crop	Scale = 0.8–1.0 of original size
Color Jitter	Brightness = 0.2, Contrast = 0.2, Saturation = 0.2
Gaussian Blur	Kernel size = 3x3, Sigma = 0.1–2.0

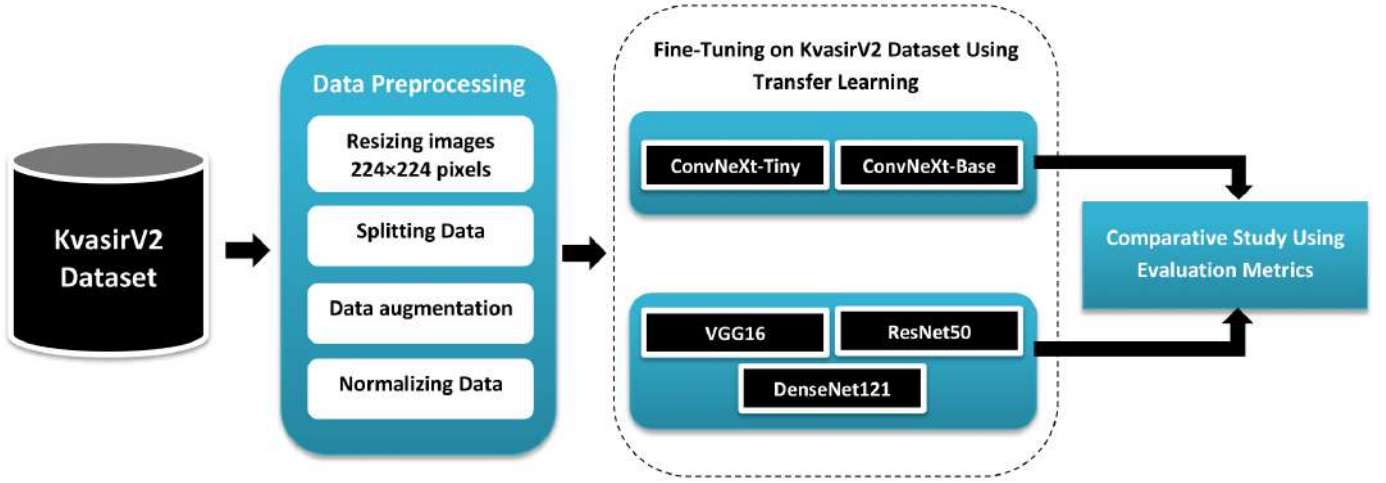


Fig. 1. The Overall Workflow of the Proposed Approach for Intestinal Disease Classification on Endoscopic Images.

All images were resized to a uniform dimension of 224x224 pixels to match the input size required by both CNN and ConvNeXt models. On the other hand, we applied normalization to standardize image pixel values using the ImageNet mean and standard deviation [14], which promotes stable training and ensures compatibility with the diverse architectures used:

- **Mean:** 0.485, 0.456, 0.406.
- **Standard Deviation:** 0.229, 0.224, 0.225.

This preprocessing pipeline ensured consistent and diverse input data for robust model learning. The dataset was divided into two subsets using a stratified sampling strategy that preserves the class distribution across splits:

- **Training set:** 70% of the total images.
- **Validation set:** 15% of the total images.
- **Testing set:** 15% of the total images.

This stratified partitioning guarantees that each class is proportionally represented in both subsets, enabling consistent training behavior and reliable assessment of model generalization capability.

D. Training Configuration

All experiments were conducted in the Kaggle environment using the Python programming language and the PyTorch library [15]. The key training parameters were as follows:

- Optimizer: Adam [16]
- Learning rate: 0.0001
- Batch size: 32
- Epochs: 50
- Loss function: Categorical Cross-Entropy

E. Evaluation Metrics

The performance of both models was assessed using standard classification metrics, including:

1) *Accuracy* [17], its equation is:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2) *Precision* [18], its equation is:

$$\text{Precision} = \frac{TP}{TP + FP}$$

3) *Recall* [18], its equation is:

$$\text{Recall} = \frac{TP}{TP + FN}$$

4) *F1-Score* [19], its equation is:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where:

- TP = True Positives → diseased correctly predicted as diseased
- TN = True Negatives → healthy correctly predicted as healthy
- FP = False Positives → healthy samples wrongly identified as diseased
- FN = False Negatives → diseased samples wrongly identified as healthy

We also used the confusion matrices and multi-class ROC curves (AUC) [20] to assess classification accuracy and discriminative ability across all classes.

IV. RESULTS AND DISCUSSION

The comparative performance of the baseline models on the KvasirV2 dataset is presented in Table II. Among these models, ConvNeXt-Base delivered the best overall results, achieving the highest accuracy, precision, recall, and F1-score, indicating superior capability in identifying and distinguishing gastrointestinal abnormalities. ConvNeXt-Tiny also demonstrated strong performance, closely followed by DenseNet121, ResNet50, and VGG16, confirming the advantages of modernized convolutional architectures over traditional CNN models in medical image classification.

To further analyze the models' performance, confusion matrices were computed for each model, providing detailed

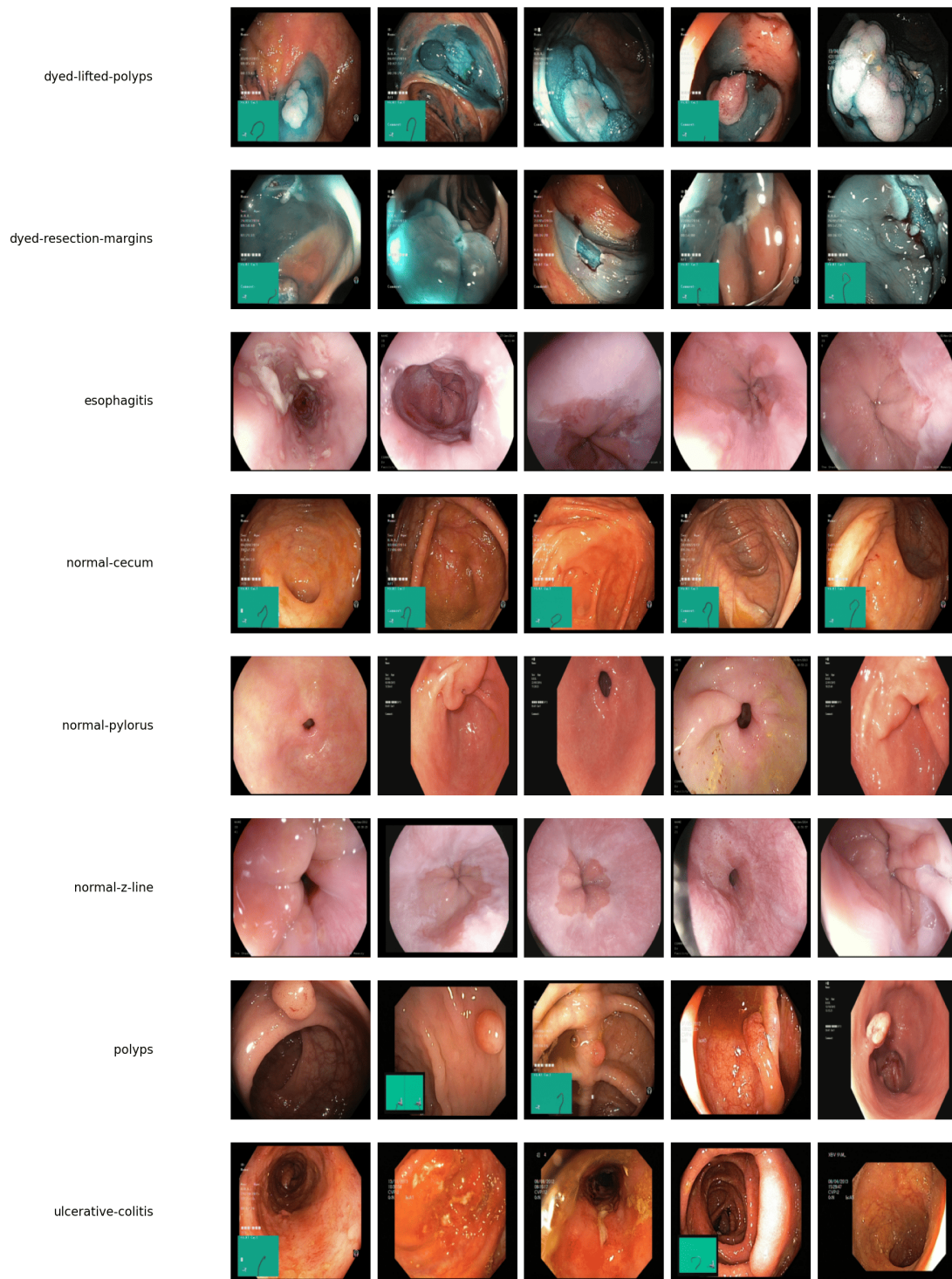


Fig. 2. Sample images from the KvasirV2 dataset, with five representative images from each class: Dyed-lifted polyps, Dyed-resection margins, Esophagitis, Normal-cecum, Normal-pylorus, Normal-z-line, Polyps, and Ulcerative colitis.

TABLE II
PERFORMANCE COMPARISON OF BASELINE MODELS ON THE KVASIRV2 DATASET.

Model	Accuracy	Precision	Recall	F1-score
ConvNeXt-Tiny	0.9392	0.9393	0.9392	0.9391
ConvNeXt-Base	0.9442	0.9456	0.9442	0.9442
VGG16	0.9075	0.9109	0.9075	0.9072
ResNet50	0.9142	0.9153	0.9142	0.9140
DenseNet121	0.9333	0.9332	0.9333	0.9331

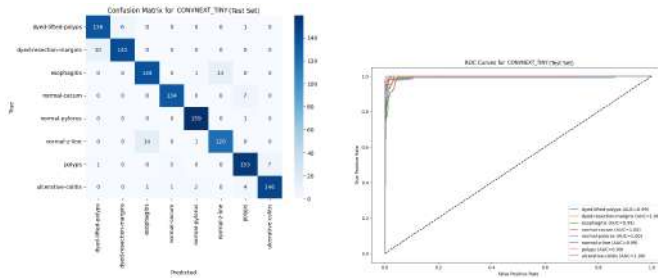


Fig. 3. Multi-class ROC curves and confusion matrix for the ConvNeXt-Tiny model.

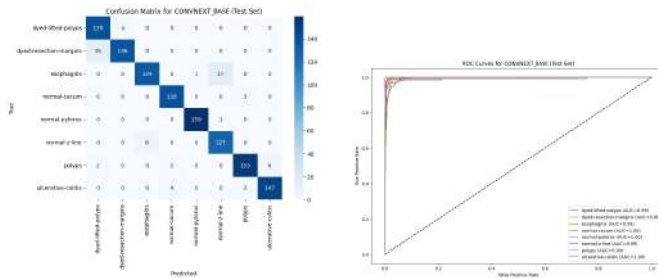


Fig. 4. Multi-class ROC curves and confusion matrix for the ConvNeXt-Base model.

insights into class-wise prediction accuracy. The matrices revealed that most misclassifications occurred between visually similar classes, such as Normal-cecum and Normal-pylorus, highlighting areas where endoscopic images present inherent ambiguity.

Additionally, multi-class ROC curves were plotted to assess the discriminative ability of each model across all classes. The area under the curve (AUC) values for all models were consistently high, confirming that the classifiers were able to effectively distinguish between different gastrointestinal conditions. ConvNeXt-Base achieved the highest average AUC, reinforcing its superior performance in both overall classification metrics and class-wise discrimination.

V. CONCLUSION AND FUTURE WORK

The results demonstrate that modern convolutional architectures, particularly the ConvNeXt variants, provide robust performance for endoscopic image classification. The combination of quantitative metrics, confusion matrix analysis, and ROC-AUC evaluation offers a comprehensive understanding of model strengths and limitations. For future work, we

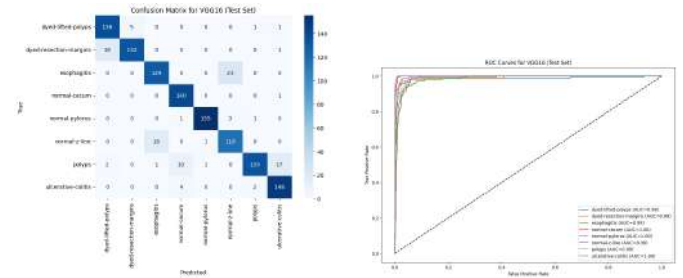


Fig. 5. Multi-class ROC curves and confusion matrix for the VGG16 model.

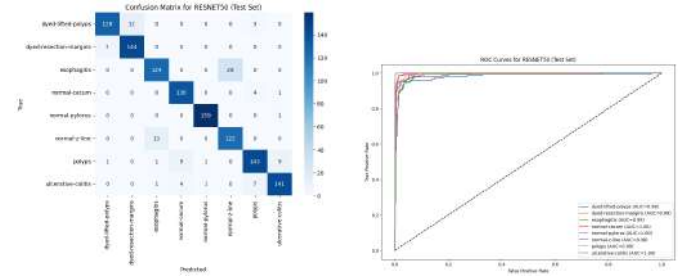


Fig. 6. Multi-class ROC curves and confusion matrix for the ResNet50 model.

propose several directions to further enhance performance and clinical applicability. We will focus on hybrid approaches and implement ensemble methods to boost classification accuracy. Additionally, expanding the dataset to cover a broader range of gastrointestinal conditions and validating the models in real-world clinical settings will be essential steps toward reliable deployment in practice.

REFERENCES

- [1] P. Wilson, D. N. Naveen, and G. N. Sundar, "Prediction of gastrointestinal diseases from endoscopy images," in *AIP Conference Proceedings*,

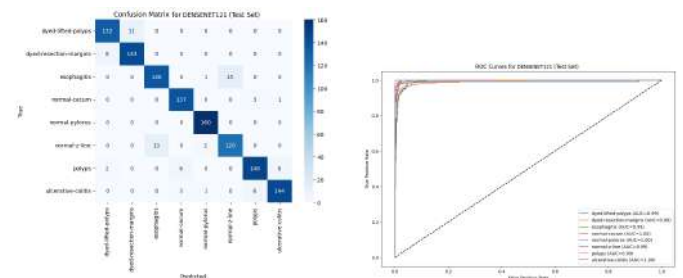


Fig. 7. Multi-class ROC curves and confusion matrix for the DenseNet121 model.

- vol. 3204, no. 1. AIP Publishing LLC, 2025, p. 050013.
- [2] M. Trigka and E. Dritsas, "A comprehensive survey of deep learning approaches in image processing," *Sensors*, vol. 25, no. 2, p. 531, 2025.
 - [3] G. Urban and et al., "Deep learning for polyp detection in colonoscopy videos," *IEEE Transactions on Medical Imaging*, vol. 37, pp. 271–281, 2018.
 - [4] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s (2022)," URL <https://arxiv.org/abs/2201.03545>, 2025.
 - [5] Y. Zhu and et al., "Deep learning for gastrointestinal disease detection," *IEEE Access*, vol. 7, pp. 152 310–152 321, 2019.
 - [6] X. Zhang and et al., "Resnet-based colorectal lesion classification," *Computers in Biology and Medicine*, vol. 121, p. 103784, 2020.
 - [7] A. Dosovitskiy and et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2021.
 - [8] Z. Liu and et al., "Convnext: Revisiting convolutional networks for visual recognition," *arXiv preprint arXiv:2201.03545*, 2022.
 - [9] S. Khan and et al., "Comparative analysis of convnext, efficientnet, and densenet for skin lesion classification," *Computers in Biology and Medicine*, vol. 150, p. 106195, 2023.
 - [10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
 - [11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
 - [12] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.
 - [13] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, ser. MMSys'17. New York, NY, USA: ACM, 2017, pp. 164–169.
 - [14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
 - [15] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.
 - [16] K. D. B. J. Adam et al., "A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, vol. 1412, no. 6, 2014.
 - [17] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.
 - [18] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.
 - [19] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.
 - [20] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

Influence of Annealing on the Dark and Photo-Induced I–V Behavior of a ZnO:Ga—Cu₂O Heterojunction Fabricated via Ultrasonic Spray Deposition

Hamza Trir

Laboratory of Semiconductor and Metallic Materials

Mohammed Khider University

Biskra, 07000, Algeria

hamza.trir@univ-biskra.dz

Abstract—The ZnO:Ga/Cu₂O heterojunction was fabricated using ultrasonic spray pyrolysis and characterized structurally, optically, and electrically. XRD and UV–Vis analyses confirmed polycrystalline films with suitable band gaps, while SEM and FTIR revealed good morphology and bonding. Annealing at 350–450°C improved the dark and illuminated I–V performance, indicating enhanced junction quality and promising photovoltaic potential.

Index Terms—ZnO:Ga, Cu₂O hetero-junction, ultrasonic spray pyrolysis, electrical properties, annealing

I. INTRODUCTION

Transparent conducting oxides (TCOs) are materials that combine high optical transparency with good electrical conductivity, making them indispensable in applications such as solar cells, organic light-emitting diodes (OLEDs), and gas sensors. While indium–tin oxide (ITO) remains the most widely used TCO due to its excellent optical and electrical properties, Ga-doped ZnO (GZO) has emerged as an attractive alternative owing to its non-toxicity, chemical stability, and enhanced performance resulting from Ga³⁺ substitution in the ZnO lattice.

Cu₂O is a p-type semiconductor with a suitable band gap, low cost, and environmental compatibility, making it a promising absorber material for photovoltaic devices. Combining n-type ZnO:Ga with p-type Cu₂O forms a heterojunction that can be utilized in optoelectronic and photovoltaic applications. Previous studies have shown that ZnO–Cu₂O junctions can achieve favorable open-circuit voltages and competitive conversion efficiencies.

This work focuses on the fabrication of ZnO:Ga and Cu₂O thin films using ultrasonic spray pyrolysis, followed by the formation of a ZnO:Ga–Cu₂O heterojunction. Structural, optical, and morphological properties of the films are investigated, and the influence of post-deposition annealing on the dark and illuminated current–voltage (I–V) characteristics is evaluated.

Identify applicable funding agency here. If none, delete this.

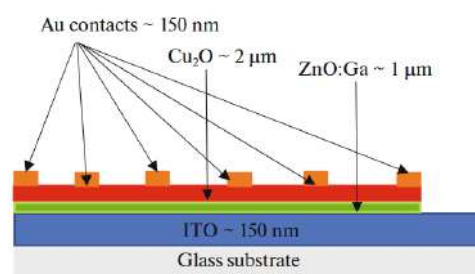


Fig. 1. Hetero-structure schematic of the realized device: ITO–ZnO:Ga–Cu₂O–Au.

Cu₂O is a p-type semiconductor with a suitable band gap, low cost, and environmental compatibility, making it attractive as an absorber layer in photovoltaic devices. The combination of n-type ZnO and p-type Cu₂O forms a heterojunction with potential for optoelectronic applications, and previous studies have demonstrated promising device performance.

This work focuses on fabricating ZnO:Ga and Cu₂O thin films using ultrasonic spray pyrolysis, characterizing their structural, optical, and morphological properties, forming a ZnO:Ga–Cu₂O heterojunction, and evaluating the influence of annealing on its dark and illuminated current–voltage characteristics.

II. EXPERIMENTAL DETAILS

The ZnO:Ga|Cu₂O heterojunction was fabricated on ITO-coated glass substrates using ultrasonic spray pyrolysis. The substrates were thoroughly cleaned in distilled water, ethanol, acetone, and isopropanol prior to deposition.

The n-type ZnO:Ga layer was deposited from a 0.05 M zinc acetate precursor solution, while the p-type Cu₂O layer was sprayed onto the ZnO:Ga film using a solution composed of 0.02 M copper acetate and 0.03 M glucose in a water/isopropanol mixture (4:1). The substrate temperature during Cu₂O deposition was maintained at 300°C.

Structural properties were analyzed by X-ray diffraction (XRD), optical properties by UV-Vis spectroscopy, and surface morphology by scanning electron microscopy (SEM). Fourier-transform infrared (FTIR) spectroscopy was used to identify chemical bonding.

Gold contacts were thermally evaporated onto the Cu_2O layer, and current-voltage (I-V) measurements were performed using a Keithley 2400 SourceMeter under both dark and illuminated conditions with a xenon lamp.

To study thermal effects, the fabricated heterojunction was annealed at 350°C, 400°C, and 450°C for 2 hours in air before characterization.

III. RESULTS AND DISCUSSION

A. Structural Properties

X-ray diffraction (XRD) patterns of ZnO:Ga, Cu_2O , and the ZnO:Ga- Cu_2O heterojunction are presented in Fig.2. The ZnO:Ga thin film exhibits a hexagonal wurtzite structure with a preferred (100) orientation, indicating well-developed crystallinity. The Cu_2O layer shows a strong diffraction peak at the (111) plane, confirming the formation of cubic Cu_2O with good crystallinity and grain growth. The heterojunction pattern contains the characteristic peaks of both ZnO:Ga and Cu_2O , demonstrating successful stacking of the two layers without secondary phases.

B. Optical Properties

The optical transmission spectra shown in Fig.3 reveal that ZnO:Ga films possess high transparency (70–75%) in the visible region. The strong absorption edge near 380 nm corresponds to the wide band gap of the ZnO:Ga layer. In contrast, Cu_2O films exhibit significant absorption in the UV region due to their narrower band gap. The heterojunction demonstrates a broadened absorption band around 600 nm, resulting from the complementary optical response of both materials. This behavior is favorable for photovoltaic applications, as enhanced visible-light absorption contributes to improved device performance.

C. Morphology (SEM)

SEM micrographs of ZnO:Ga and Cu_2O thin films are shown in Fig.4 and Fig. 5, respectively. ZnO:Ga films are compact, uniform, and composed of fine, densely packed grains. Cross-sectional SEM images confirm a homogeneous and continuous ZnO:Ga layer. In contrast, Cu_2O films exhibit a textured surface with small, randomly distributed grains, as well as cracks and voids. The Cu_2O layer thickness varies between 2–5 μm , as observed in the cross-sectional SEM views.

D. FTIR Analysis

FTIR spectra of Cu_2O and ZnO:Ga films, shown in Fig.6, confirm the presence of Cu-O stretching vibrations near 623 cm^{-1} , characteristic of Cu_2O . No CuO-related bands (577 and 531 cm^{-1}) were detected, verifying the high purity of the deposited Cu_2O film. The ZnO:Ga spectrum displays Zn-O

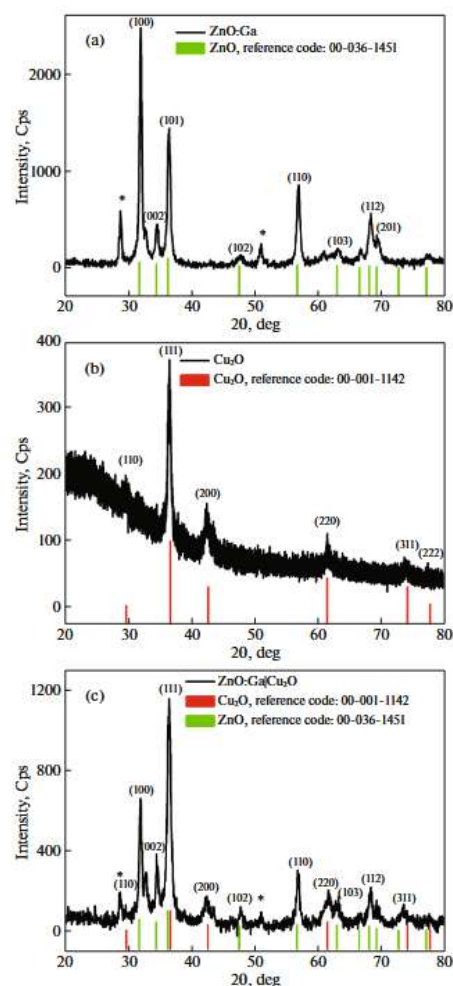


Fig. 2. (a) X-ray diffraction pattern of ZnO:Ga thin film. (b) X-ray diffraction pattern of Cu_2O thin film. (c) X-ray diffraction pattern of the Cu_2O -ZnO junction.

modes near 442 cm^{-1} and O-H stretching around 3440 cm^{-1} , typical of ZnO-based materials due to their hygroscopic nature. These results confirm the successful formation of chemically pure ZnO:Ga and Cu_2O layers.

E. Electrical Properties

The dark and illuminated current-voltage (I-V) characteristics of the ZnO:Ga- Cu_2O heterojunction, shown in Fig. 7, exhibit clear diode behavior. Annealing the device at temperatures of 350°C, 400°C, and 450°C significantly improves its electrical performance. Higher annealing temperatures increase the rectification ratio, reduce leakage current, and enhance photocurrent response. These improvements are attributed to better crystallinity, reduced defect density, and improved interfacial quality between the ZnO:Ga and Cu_2O layers. Overall, annealing enhances the junction performance and confirms the suitability of ZnO:Ga- Cu_2O structures for Cu_2O -based photovoltaic applications.

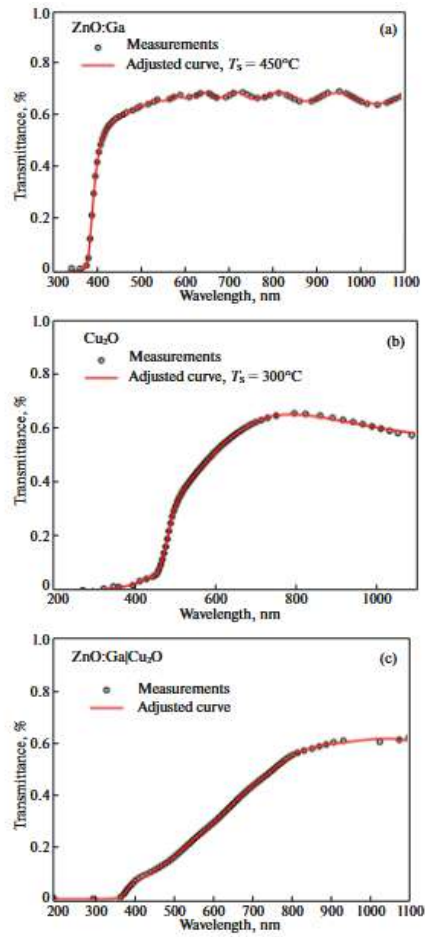


Fig. 3. (a) Optical transmission spectrum of a Ga-doped zinc oxide film deposited on glass. (b) Optical transmission spectrum of a cuprous oxide film deposited on glass. (c) Optical transmission spectrum of a ZnO:Ga— Cu_2O junction deposited on glass.

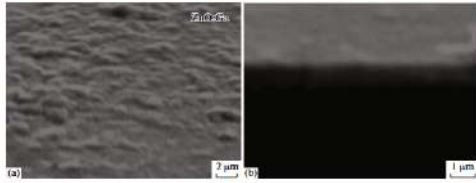


Fig. 4. SEM micrographs of thin films of ZnO:Ga and SEM image of the cross-section.

IV. CONCLUSION

ZnO:Ga, Cu_2O , and ZnO:Ga/ Cu_2O thin films were successfully fabricated using ultrasonic spray pyrolysis. Structural and optical analyses confirmed that the films are polycrystalline and transparent, with optical band gaps of approximately 3.27 eV for ZnO:Ga, 2.65 eV for Cu_2O , and 3.29 eV for the heterojunction. Electrical measurements demonstrated that annealing the heterojunction at higher temperatures significantly improves its dark and illuminated I–V characteristics. These results show that annealing enhances the device performance and confirms the potential of ZnO:Ga/ Cu_2O structures for

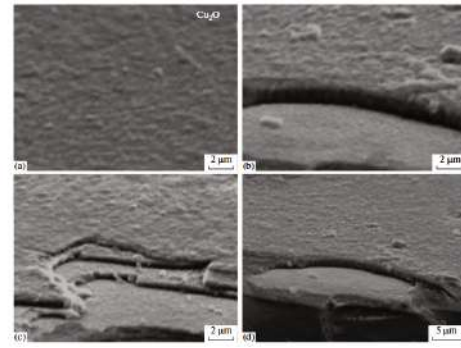


Fig. 5. SEM micrographs of Cu_2O thin films and SEM image of the cross-section.

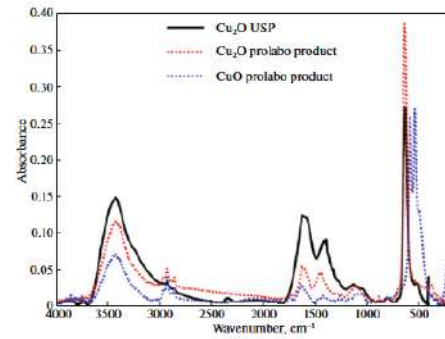


Fig. 6. FTIR spectrum of Cu_2O . Comparison of the FTIR spectra of Cu_2O , Cu_2O powders, and the scraped product of Cu_2O thin films in the absorbance mode.

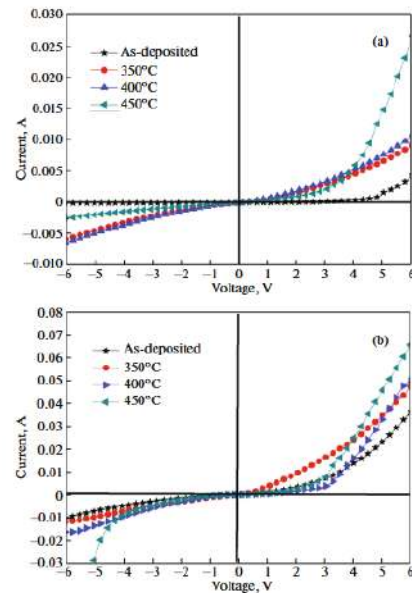


Fig. 7. Effect of annealing temperature on I(V) characteristics of a ZnO:Ga— Cu_2O hetero-junction: (a) the dark and (b) the illuminated.

Cu₂O-based photovoltaic applications.

REFERENCES

- [1] H. Hosono and K. Ueda, "Electronic and photonic materials," in Springer Handbook of Electronic and Photonic Materials, S. Kasap and P. Capper, Eds. Cham: Springer, 2017.
- [2] Y. Chen, "Materials science and engineering," IOP Conf. Ser.: Mater. Sci. Eng., vol. 423, p. 012170, 2018.
- [3] Y. S. Kim, W. J. Hwang, K. T. Eun, and S. H. Choa, "Application of surface science," Applied Surface Science, vol. 257, pp. 8134–8139, 2011.
- [4] J. W. Park, G. H. Lee, Y. Y. Kwon, K. W. Park, J. Lee, Y. W. Jin, Y. C. Nah, and H. Kim, "Organic electronics," Organic Electronics, vol. 15, pp. 2178–2185, 2014.
- [5] S. Fay, J. Steinhauser, N. Oliveira, E. Vallat-Sauvain, and C. Ballif, "Thin solid films," Thin Solid Films, vol. 515, pp. 8558–8563, 2007.
- [6] M. Schüler, T. Sauerwald, and A. Schütze, "Sensor systems," Journal of Sensor Systems, vol. 3, pp. 213–220, 2014.
- [7] J. Dsy and N. P. Ta, "Materials chemistry engineering," Journal of Materials Chemistry Engineering, vol. 1, pp. 1004–1010, 2018.
- [8] H. Hagendorfer, K. Lienau, and S. Nishiwaki, "Advanced materials," Advanced Materials, vol. 26, pp. 632–639, 2014.
- [9] H. Wei, M. Li, Z. Ye, Z. Yang, and Y. Zhang, "Materials letters," Materials Letters, vol. 65, pp. 427–430, 2011.

Intelligent Traffic Signal Control Using YOLOv8 for Enhanced Urban Mobility and Emergency Response

Rahim Abdelilah
Electrical Systems Engineering
Department
Faculty of Technology, UMBB
Boumerdes, Algeria
a.rahim@univ-boumerdes.dz

Mohamed Mouloudi Abdelilah
Electrical Systems Engineering
Department
Faculty of Technology, UMBB
Boumerdes, Algeria
mouloudimohamed308@gmail.com

Massine GANA
LARI Laboratory
Faculty of Electrical Engineering and
Computer Science, UMMTO
Tizi-Ouzou, Algeria
massine.gana@ummto.dz

Abstract: *Urban traffic congestion has been demonstrated to have deleterious effects on modern society, with consequences including increased travel time, compromised public health, and exacerbated environmental pollution. The proposed AI-based traffic management system utilizes YOLOv8 for real-time vehicle detection. The system dynamically adjusts traffic light timings according to detected vehicle density and identifies emergency vehicles through visual cues to prioritize their movement. Two distinct modules were developed: one for adaptive traffic signal control and another for prioritizing emergency vehicles. The model was evaluated through multiple simulated traffic scenarios, demonstrating improved traffic flow efficiency and reliable emergency vehicle detection.*

Keywords— *traffic management, YOLOv8, Smart traffic, vehicle detection, adaptive traffic signals, emergency vehicle recognition*

1) INTRODUCTION

Urban traffic congestion has become a persistent challenge in modern cities, resulting in significant economic losses and reduced overall mobility. A substantial body of research has repeatedly demonstrated that congestion levels exhibit an annual increase [1], exerting an impact on both large metropolitan areas and smaller developing cities. The augmentation of road infrastructure has been demonstrated to be an ineffective strategy for addressing congestion. This phenomenon can be attributed to the fact that an increase in capacity frequently increases vehicle usage, thus leading to a reversion to the original congestion patterns.

Conventional traffic signal systems have been identified as a contributing factor to these issues. It has been demonstrated that fixed-time controllers are unable to function in conjunction with real-time vehicle conditions. The consequences of this phenomenon may include, but are not limited to, unnecessary delays, inefficient lane usage, accidents [2], and protracted waiting times during peak hours. Furthermore, these systems have been observed to fail to provide consistent priority for emergency vehicles, which may result in delays to safety-critical responses. These limitations emphasize the necessity for the implementation of traffic management strategies that can perceive, interpret, and respond to current traffic states, as opposed to relying on static timing cycles.

Recent advancements in the fields of artificial intelligence and computer vision have prompted urban areas to investigate the potential of AI-based perception to enhance the adaptability of traffic control systems. The capacity of contemporary object-detection models to provide real-time data concerning vehicle presence and lane occupancy suggests that artificial intelligence (AI) has the potential to serve as a

tool that could facilitate more intelligent responses by traffic lights to changing conditions and traffic flow patterns.

The present study explores this possibility by developing an adaptive traffic signal framework powered by YOLOv8 for real-time visual detection. The system has been designed to estimate lane occupancy, adjust green-light durations according to immediate traffic needs, and provide emergency-vehicle prioritization when an active emergency vehicle is detected. The framework has been implemented and evaluated in simulated environments to examine its potential to reduce waiting times, improve traffic distribution, and demonstrate how AI-enhanced perception can support more responsive urban mobility.

2) BACKGROUND AND RELATED WORK

The evolution of urban traffic control is a story of continuous innovation, driven by the growing complexity of city life. From manual oversight to the first automated systems, cities have long sought ways to mitigate congestion and improve mobility. This section delves into the historical approaches to traffic management, exploring how early solutions laid the groundwork for modern advancements.

A. Early Approaches to Urban Traffic Control

The management of traffic initially involved manual control, wherein police officers directed vehicles at intersections. However, as cities expanded, this approach became unsustainable. The advent of early automation can be traced back to the 1868 gas-powered signal and the subsequent emergence of electric traffic lights in the 1910s [3]. However, these pioneering systems were predicated on fixed-time cycles, which frequently proved to be unreliable in the face of fluctuating demand.

The 1920s and 1930s witnessed the introduction of timing plans and coordinated "green wave" corridors, representing a seminal advancement in the field [4]. These enhancements resulted in a reduction of stops under ideal conditions; however, they failed to perform adequately when traffic deviated from expected patterns. Mid-century systems incorporated electromechanical controllers and inductive loops, thereby enabling limited responsiveness by detecting vehicle presence. Despite their extensive implementation, loops were costly, maintenance-heavy, and incapable of capturing full traffic states, such as queue length or spillback.

Since the 1960s, the advent of actuated signals and the emergence of computerized networks have led to significant advancements in monitoring systems, though these systems have remained dependent on rule-based logic. Subsequent technologies (RFID priority systems, Bluetooth/Wi-Fi

tracking, and GPS floating-car data) have expanded the scope of sensing capabilities; however, they have not addressed the fundamental issue of enabling traditional systems to reason about dynamic traffic behavior. These limitations naturally led cities to explore AI-based solutions.

B. The Rise of AI-Based Traffic Control

In the 1980s and 1990s, the early applications of artificial intelligence (AI) primarily utilized expert systems, seeking to encapsulate human reasoning within signal-control rules. These systems functioned solely within the confines of the scenarios they were designed to anticipate. The advent of computer vision in the 1990s and 2000s engendered a paradigm shift in the realm of camera technology, thereby enabling more sophisticated and nuanced monitoring capabilities. However, the prevailing computational limitations-imposed impediments to the real-time adaptation that would have otherwise been possible.

Machine learning in the 2000s–2010s enabled models to learn traffic patterns and adjust timing with greater efficacy [5], yet they remained constrained by manually engineered features. Since 2015, the advent of deep learning (in particular, convolutional neural network-based detectors such as YOLO) has facilitated the real-time perception of vehicles and lane states with a high degree of accuracy, thereby enabling true adaptive control.

Recent advancements in reinforcement learning and data-driven optimization methods have shifted towards a paradigm where the management of entire networks rather than isolated intersections is the objective. Modern smart-traffic platforms integrate sensing, prediction, and control, representing the first systems capable of continuously adapting to real-world traffic conditions.

3) METHODOLOGY

A. OVERVIEW OF THE APPROACH

The efficacy of traffic-aware systems is contingent upon the accuracy and real-time nature of their perception. The methodology employed is predicated on the YOLO family of object-detection models [6]. The YOLO architecture has been designed with a focus on three key principles: namely, speed, low-latency inference, and high spatial accuracy. This renders it suitable for a range of applications, including vehicle tracking, lane-occupancy estimation, and adaptive traffic-light control. In contradistinction to two-stage detectors, YOLO operates as a single, unified pipeline. This capacity extends to detection, class filtering, tracking support, and region-based analysis, and is accompanied by the additional advantage of frame rates that are compatible with live deployments.

The selection of YOLOv8 was driven by practical constraints rather than the innovation of the model. During the experimentation phase in early 2024, YOLOv8 was the most stable and fully supported release maintained by Ultralytics. The ecosystem under scrutiny was found to provide reliable updates, extensive documentation, clean Python APIs, and seamless export formats. Of particular note was the direct integration with established tracking modules, a feature that reduced engineering overhead and facilitated the assembly and testing of the full perception pipeline.

A further advantage of YOLOv8 is that it is shipped pretrained on the COCO dataset, which includes 80 categories commonly found in urban traffic scenes. This finding

indicates that the model possesses the capability to discern relevant vehicle classes, encompassing cars, buses, trucks, motorcycles, and bicycles, without necessitating custom training or the accumulation of a dataset. In the context of a project characterized by temporal constraints and limitations in hardware resources, the utilization of pretrained COCO weights facilitated the immediate operational functionality of the system. The efficacy of YOLOv8 in facilitating the process of class-filtering has been demonstrated, thereby enabling the detection pipeline to focus exclusively on the vehicle categories that are required for the experimental process.

YOLOv8 is equipped with a versatile model size configuration, encompassing options ranging from nano to extra-large. This enables a judicious balance to be struck between accuracy and inference speed, contingent upon the prevailing hardware specifications. In the initial prototyping phase, smaller variants were utilized for the purpose of rapid iteration; meanwhile, larger variants were subjected to testing in simulation to analyze performance under conditions of high traffic density.

While the identification of objects constitutes the primary function of the perception pipeline, it is insufficient in isolation for the development of traffic-control logic. A salient feature of YOLOv8 is its incorporation of a built-in tracking module. This module has the capacity to provide persistent object identifiers, simple trajectory estimation, and frame-to-frame association, without the necessity for external libraries. This out-of-the-box tracker was sufficient for the estimation of per-lane vehicle counts and the observation of motion patterns in the simulation environments utilized. It was determined that the integration, documentation, and maintenance of the tracking module by Ultralytics rendered the implementation of supplementary tracking algorithms superfluous for the system under consideration.

After the development of this project, novel YOLO variants have been released. It is noteworthy that the YOLOv9 and YOLOv10 models were developed by external research groups, rather than by Ultralytics. Consequently, these models did not benefit from the same comprehensive support ecosystem. Subsequent to the publication of the aforementioned article, Ultralytics introduced YOLOv11 and later YOLOv12/YOLO26, but these versions were released after the work described in this article. At the time of testing, the newer models were still emerging or unstable, thus rendering YOLOv8 the most reliable and fully supported option available.

The selection of YOLOv8 as the implementation and evaluation framework for the system is multifaceted, encompassing factors such as mature support, pretrained COCO weights, compatibility with tracking, flexible model scaling, and stable tooling.

Program 1 — Dynamic Traffic Timing Using Real-Time Detection

Program 1 implements adaptive traffic-light control based on real-time vehicle detection. The system utilises YOLOv8, which has been pretrained on the COCO dataset, to identify traffic-related object classes (i.e. cars, buses, trucks and motorcycles) without the necessity for additional training. The retention of vehicle classes is exclusively through the implementation of class filtering to attenuate noise within the detection pipeline.

a) Vehicle Detection and Class Filtering

The YOLOv8 model processes incoming frames and outputs bounding boxes and class labels for all objects detected. A filtration stage is employed to eliminate non-vehicle categories, enabling the system to concentrate exclusively on traffic flow. This enhancement leads to improved computational efficiency and temporal stability.

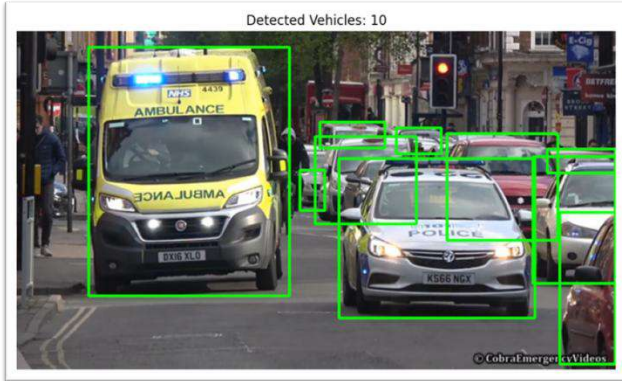


Figure 1- Vehicles being detected and counted while everything else is

b) Multi-Object Tracking for Unique Vehicle Counts

Direct frame-by-frame detection can result in severe overcounting in video sequences, since YOLOv8 treats each frame independently. In order to address this issue, the system utilizes the built-in tracking module of YOLOv8, which assigns persistent IDs to each detected vehicle and maintains these IDs across frames. This ensures that each vehicle is counted only once, regardless of the duration for which it remains in the scene.

c) Spatial Filtering Using Lane Regions

Vehicles that are not pertinent to the current intersection, such as those on adjacent roads, have the capacity to introduce error in traffic estimation. In order to mitigate this issue, the system employs user-defined polygonal lane regions. The validity of a detected vehicle is contingent upon the center of its bounding box lying within a lane polygon. This method facilitates precise separation of traffic by approach, a capability that extends to wide-angle and 360° camera views.

d) Performance Optimization and Conditional Scanning

Continuous inference is computationally expensive and thus unsuitable for low-power devices. A range of optimization strategies was evaluated, encompassing the reduction of video resolution and frame rate. The most effective solution was a conditional scanning strategy:

- The system performs inference only when a green cycle is about to end.
- 1–3 frames per approach are analyzed, instead of continuous 30 FPS scanning.
- This reduces GPU load and allows larger YOLOv8 models (e.g., YOLOv8-L or YOLOv8-XL) to run efficiently even on constrained hardware.

This conditional method preserves detection reliability while eliminating performance bottlenecks.

e) Dynamic Green-Time Allocation

At each cycle boundary, the system counts vehicles in every approach and distributes the next green cycle proportionally.

Let: A_i = number of vehicles in approach i
 T = total green cycle time

The raw allocation percentage is

$$P_i = \frac{A_i}{\sum_{j=1}^n A_j} \times 100 \quad (1)$$

The green time assigned to each approach is

$$t_i = \frac{P_i}{100} \times T \quad (2)$$

f) Minimum-Time Safety Rule

Pure proportional allocation can result in unstable or unsafe behavior, particularly when one approach has very few vehicles. To avoid starvation, a **minimum allocation** is imposed:

$$P_i < 5\% \rightarrow P_i = 5\% \quad (3)$$

After applying the minimum threshold, the remaining percentage is redistributed among the non-threshold approaches using the same ratio formula.

This rule ensures predictable minimum service for all approaches and prevents the system from skipping low-traffic lanes due to temporary detection errors.

B. Program 2 — Emergency Vehicle Prioritization

Program 2 extends the adaptive control logic by introducing real-time prioritization for emergency vehicles. A bespoke YOLOv8 model, trained on emergency-light imagery, is utilized for the purpose of detecting active sirens and, when required, overriding standard traffic allocation.

a) Emergency-Light Dataset Selection

Several datasets were evaluated to determine the most reliable emergency-vehicle indicator:

1) *Emergency Vehicle Detector (Rejected)*: A dataset containing bounding boxes for ambulances, fire trucks, and police vehicles was discarded due to:

- significant mismatch between Western vehicle designs and local models, and
- absence of distinction between active and inactive vehicles, which could trigger false emergency states.

2) *Emergency Light / Siren Dataset (Selected)*: A dataset focused on detecting rooftop emergency lights was selected as the primary source. This approach is invariant to vehicle shape and identifies any vehicle actively responding to an emergency. Detection relies solely on the presence and state of the flashing light, enabling robust operation across heterogeneous fleets.

3) *Siren Sound Spectrogram Dataset (Exploratory)*: A third dataset involving audio spectrograms of emergency sirens was explored as an additional modality. While promising as a backup when lights are not visible, it was not integrated due to time constraints and preprocessing complexity. It remains a viable extension for future development.

b) Model Training for Emergency-Light Detection

After selecting the siren-light dataset, a dedicated YOLOv8-Large (YOLOv8l) model was trained using Google Colab for enhanced compute capability.

Model configuration: Model: YOLOv8l | Epochs: 100 | Dataset classes: siren, emergency (active), non-emergency (inactive)

Training outcome: The model achieved strong convergence with:

- high precision (≈ 0.75 – 0.8),
- high recall (≈ 0.9),
- $mAP_{50} \approx 0.78$ and $mAP_{50-95} \approx 0.68$,

demonstrating reliable generalization under varied lighting conditions.

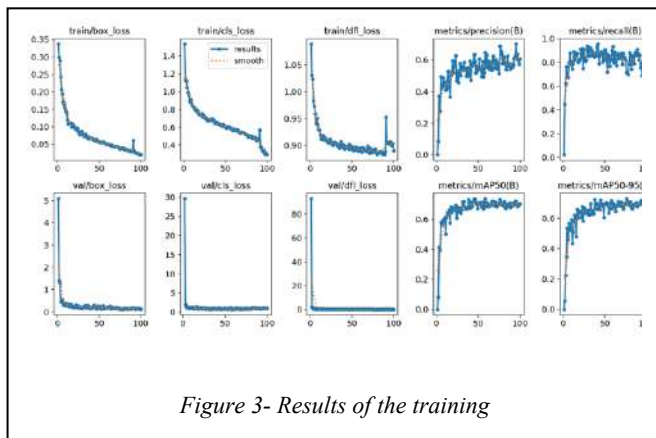


Figure 3- Results of the training

Confusion matrix interpretation: Some misclassification was observed between “no emergency” and “siren,” typically under low-visibility conditions. However, the system requires both:

- 1) *visual detection of a siren, and classification as an active emergency (“emr”)*

before triggering override, preventing false activations. A confidence threshold reinforces robustness.

Real-Time Integration With the Traffic-Control Loop

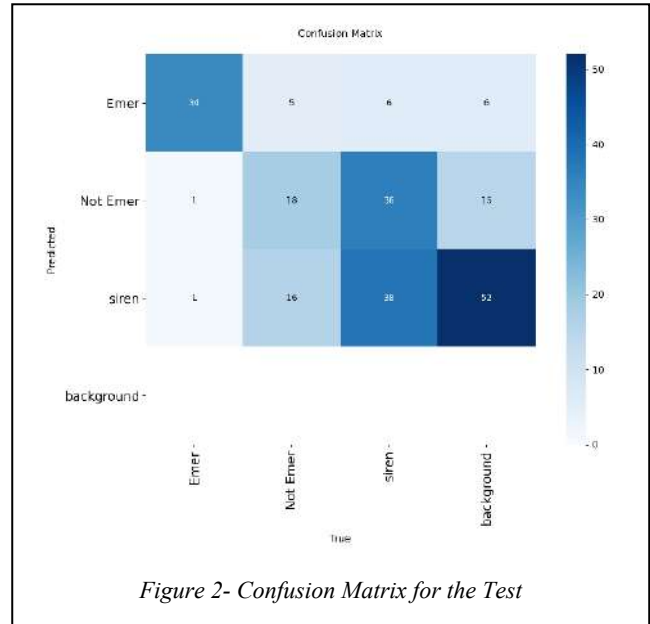


Figure 2- Confusion Matrix for the Test

Program 2 is integrated with Program 1 to support both adaptive traffic flow and emergency prioritization.

c) Dual-Pass Scanning Mechanism:

Each approach is scanned twice during every green phase:

- Mid-cycle scan (50%) for early adaptation and emergency detection.
- End-of-cycle scan for next-cycle timing, as defined by Program 1.

This two-pass strategy balances efficiency and responsiveness.

1) Emergency Detection Procedure

To reduce noise and single-frame artifacts, the system analyzes approximately one second of video per scan. If an object is classified as an active emergency (“emr”) with $\geq 60\%$ confidence, the system immediately transitions to Emergency Mode.

2) Emergency Mode Logic

During Emergency Mode:

- The remaining green time is fully reassigned to the approach containing the emergency vehicle.
- All other directions are halted.
- Continuous scans are performed every 10% of the remaining cycle.

Emergency Mode terminates only when two consecutive scans detect no active emergency, ensuring safe passage while preventing persistent false alerts.

d) Handling Multiple Simultaneous Emergencies

If two or more emergency vehicles are detected in different approaches, the system automatically creates a temporary subset consisting only of the active emergency lanes. The standard allocation algorithm from Program 1 is

then re-applied, but restricted exclusively to the emergency lanes. Each emergency approach receives proportional green time relative to the number of active sirens detected.

Non-emergency lanes are ignored entirely until all emergency lanes are cleared.

If all lanes contain active emergencies, the system behaves identically to Program 1 but with the constraint that no non-emergency movement is allowed. As soon as one emergency disappears, the system recalculates allocation using only the remaining active emergency approaches.

After each emergency cycle ends, the lane responsible for the last emergency is scanned again. If the emergency remains active, the priority loop restarts immediately. This guarantees uninterrupted evacuation of emergency vehicles while preventing premature return to normal traffic operation.

3) SIMULATION AND EVALUATION

To validate the proposed smart-traffic architecture, two simulation environments were created. The first evaluated the full integrated system—dynamic timing (Program 1) combined with emergency-vehicle prioritization (Program 2). The second focused solely on assessing the effectiveness of dynamic timing under controlled traffic conditions.

a) Simulation 1 — Integrated System (Programs 1 + 2)

The first simulation examined how the system behaves when both modules operate simultaneously. A looped video feed was processed in real time using YOLOv8n for detection. Program 1 allocated green-time proportionally based on detected lane density, while Program 2 continuously scanned for active emergency vehicles and overrode timing when necessary.

Because the video input was static and vehicles did not physically move, this simulation was not intended to measure throughput. Instead, it validated:

- Correct switching between Normal Mode and Emergency Mode
- Stability of the dynamic timing logic
- Reliability of emergency detection thresholds
- Consistency of confidence-based activation and deactivation

During testing, an emergency vehicle was detected on Approach 2 with ~86% confidence. The system immediately entered Emergency Mode, dedicating the entire green interval to the relevant lane. The module held priority for three cycles and returned to Normal Mode automatically after two successive negative scans. Throughout the experiment, no timing instability or false activations were observed.

Two layouts were tested: a 3-way and a 4-way intersection. In both cases, every emergency event exceeding the confidence threshold correctly triggered Emergency Mode, and all recoveries occurred without failure. The system demonstrated consistent behavior across varying approaches, event durations, and detection confidence levels.

Metric	Value
Total Events	24
Mean Confidence	79.75%
Std. Dev. Confidence	11.75%
Min / Max Confidence	60% / 94%
Mean Duration	3.38 cycles
Min / Max Duration	2 / 5cycles

Table 2-A 4 Way-Intersection Test for 3h27min

Metric	Value
Total Events	5
Mean Confidence	79.2%
Std. Dev. Confidence	9.66%
Min / Max Confidence	64% / 92%
Mean Duration	2 cycles
Min / Max Duration	1 / 3cycles

Table 1-A 3 Way-Intersection Test for 51min

b) Simulation 2 — Dynamic Timing Evaluation (Program 1 Only)

To evaluate the performance of Program 1 independently, a custom simulation was developed in the Unity engine. This environment enabled cars to move physically through the intersection, allowing a direct comparison between static and dynamic timing.

Simulation parameters:

- Cycle duration: 40 seconds
- Transition buffer: 3 seconds (yellow + safety)
- Total runtime: 10 minutes
- Initial vehicle load:
 - Approach 1: 500
 - Approach 2: 1000
 - Approach 3: 300
 - Approach 4: 700

The goal was to test the system under realistic but not extreme imbalance, allowing a fair comparison.

Results:

Mode	Cars Crossed 1	Cars Crossed 2	Cars Crossed 3	Cars Crossed 4	Total CC
Dynamic T	105	87	103	96	391
Static T	94	95	99	88	376

Table 3-Comparison Between Dynamic and Static Timing

Dynamic timing processed 391 vehicles, compared to 376 using static timing.

While the numerical difference (15 vehicles in 10 minutes) may appear modest, it reflects a trend that compounds over longer periods:

- Static cycles waste time serving empty or low-traffic approaches.
- Dynamic cycles shift time toward active lanes, increasing cumulative throughput.

Additional stress tests showed similar behavior: under normal, heavy, and extreme loads, dynamic timing consistently outperformed static logic with higher throughput and fewer idle intervals.

4) OBSERVATION AND DISCUSSION

The implementation of the smart-traffic system yielded stable and reliable results across all evaluations. Despite the utilization of simplified simulations, both dynamic green-time allocation and emergency-vehicle prioritization operated in real time and demonstrated measurable improvements in traffic flow. The vehicle counts remained consistent, the timing decisions were stable, and the emergency overrides were triggered correctly. The findings suggest that YOLO-based vision can effectively support traffic-signal optimization on low-cost hardware.

However, the experiments also highlighted key limitations. In situations where traffic density is high, such as in urban areas, the aggregation of flow becomes a more significant factor than individual detections. In such conditions, it is considered acceptable to achieve a detection accuracy of approximately 80%. However, in conditions of low traffic density, the significance of detecting with high accuracy increases significantly. Despite the fact that YOLOv8 reliably detected at least one vehicle per approach in the experimental tests conducted, real-world factors such as weather, lighting, and camera placement have the potential to introduce errors. The capacity for emergency-vehicle detection is of particular note, underscoring the necessity for complementary sensing methodologies (e.g. RFID, radio-based siren detectors) to ensure safety.

The system confirms that AI-driven control can reduce delays and improve local throughput; however, it also reinforces a well-known constraint: traffic optimization alone cannot address congestion on a large scale. The enhancement of signal efficiency has the capacity to augment the appeal of specific roadways, thereby potentially inducing congestion through induced demand. Consequently, the long-term value of such AI systems may be more significant in terms of the continuous, high-resolution traffic data they generate, as opposed to their real-time optimization capabilities. The insights derived from these data can assist urban planning authorities in the redesign of urban transportation networks. This redesign has the potential to result in a number of outcomes.

In summary, the proposed system validates the feasibility of real-time, vision-based traffic control using affordable hardware and provides actionable insights for future urban-planning strategies. While the immediate benefits are apparent, the strategic value of this approach lies in its capacity to inform structural solutions that extend beyond the limitations of intersection-level optimization.

5) CONCLUSION

This work demonstrated that AI-based traffic management using YOLOv8 can effectively optimize traffic flow and prioritize emergency vehicles in real time. Simulations showed measurable improvements in green-time allocation, vehicle throughput, and emergency response handling, even under varying traffic conditions. While detection accuracy is not perfect, the system reliably assesses lane occupancy and adapts to changing loads, making it practical for deployment with modest hardware. Beyond immediate control, the framework provides valuable traffic data that can inform long-term urban planning, highlighting congested areas, emergency response patterns, and opportunities for network redesign. With newer and improved AI models such as YOLOv11 and the upcoming YOLO26, the accuracy of collected traffic data could approach near-perfect levels, enabling even more

precise optimization and informed decisions for future city design. Overall, AI-enhanced perception proves to be a promising tool for both improving current traffic operations and supporting smarter, data-driven urban planning.

6) REFERENCES

- [1] INRIX, "INRIX: Congestion Costs Each American Nearly 100 hours, \$1,400 A Year," 2020. [Online].
- [2] S. a. B.-M. F. a. C. A. González, "Understanding the Effect of Traffic Congestion on Accidents Using Big Data," *Sustainability*, vol. 13, p. 7500, 2021.
- [3] Kittelson, "A 105-Year History of the Electric Traffic Signal," KITTELSON & ASSOCIATES, 4 June 2019. [Online]. Available: <https://www.kittelson.com/ideas/a-105-year-history-of-the-electric-traffic-signal/>.
- [4] "Co-ordinated traffic signal system," *Electric Railway Journal*, vol. 71, p. 1986, 1928.
- [5] L. L. F. L. a. J. H. G. Jin, "Spatio-Temporal Graph Neural Point Process for Traffic Congestion Event Prediction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, pp. 14268-14276, 2023.
- [6] Ultralytics, "YOLO: A Brief History," Ultralytics, [Online]. Available: <https://docs.ultralytics.com/#yolo-a-brief-history>. [Accessed 19 November 2025].
- [7] M. S. N. a. A. S. P. Næss, "Traffic Forecasts Ignoring Induced Demand: a Shaky Fundament for Cost-Benefit Analyses," *European Journal of Transport and Infrastructure Research*, vol. 12, no. 3, 2012.
- [8] R. Rashid, "Everyone thought it would cause gridlock: the highway that Seoul turned into a stream," *The Guardian*, 17 January 2025. [Online]. Available: <https://www.theguardian.com/world/2025/jan/17/seoul-cheonggyecheon-motorway-turned-into-a-stream>. [Accessed 19 November 2025].
- [9] G. D. C. Initiative, "Case Study: Cheonggyecheon; Seoul, Korea," *Global Street Design Guide*, 2025. [Online]. Available: https://globaldesigningcities.org/publication/global-street-design-guide/streets/special-conditions/elevated-structure-removal/case-study-cheonggyecheon-seoul-korea/?utm_source=chatgpt.com. [Accessed 19 November 2025].

Intelligent Real-Time Supervision of Industrial Systems via MATLAB Integration with Multiple PLCs: A Case Study in a Desalination Plant

BENARABI Bilal
Department of Electronics and
Telecommunications,
KASDI MERBAH University –
Ouargla, Algeria
benarabi.bilal@univ-ouargla.dz

KHENFER Mokhtar
Department of Electronics and
Telecommunications,
KASDI MERBAH University –
Ouargla, Algeria
khenfermok@yahoo.com

ACHBI Mohammed Said
Department of Electronics and
Telecommunications,
KASDI MERBAH University –
Ouargla, Algeria
achbi.mohammedsaid@univ-ouargla.dz

ROUABAH Boubakeur
Department of Electronics and
Telecommunications,
KASDI MERBAH University –
Ouargla, Algeria
rouabah.boubakeur@univ-ouargla.dz

MAHBOUB Mohamed Abdelbasset
Department of Electronics and
Telecommunications
KASDI MERBAH University –
Ouargla, Algeria
mahboub.abdelbasset@univ-ouargla.dz

Kechida Sihem
Laboratoire d'Automatique et
Informatique de Guelma (LAIG lab.),
Université 8 Mai 1945 Guelma, BP
401, Guelma 24000, Algérie
kechida.sihem@univ-guelma.dz

Abstract— Simulation results in MATLAB show significant advancements in analyzing complex industrial systems; however, their practical application in industrial environments remains relatively limited due to difficulties in availability and real-time implementation. The same challenges apply to control tools based on artificial intelligence when transitioning from modeling to actual operation. This study addresses these issues by directly integrating MATLAB with industrial supervisory systems in real time using the OPC standard, enabling seamless communication and coordination between PLCs from different manufacturers. The proposed framework also leverages fuzzy logic and decision support for operators, enhancing intelligent control and interactive supervision, and facilitating efficient and flexible management of complex industrial processes.

Keywords— PLC integration, Intelligent supervision, MATLAB, Fuzzy logic, Desalination.

I. INTRODUCTION

A tremendous technological advancement, modern automated systems are changing industries and daily lives. These systems provide great efficiency, precision, and productivity by utilizing cutting-edge technologies such as artificial intelligence, machine learning, the Internet of Things (IoT), and advanced sensors. In numerous industries, they improve the quality of life, streamline procedures, and lessen human error. These systems promise a safer, more productive, and connected society by freeing up human attention for imaginative, strategic, and value-added tasks. Water resources are declining rapidly, with demand for drinking water surpassing supply in some countries. The business community has embraced this blending of research and industry, resulting in complex industrial systems.

Automated systems with oversight tools are necessary for improved productivity, quality, and operational safety in automated manufacturing lines, ensuring the safety of staff, tools, and procedures. Modern automated technologies streamline industrial processes, reduce human error, and improve water distribution and purification. They allow specialists to focus on strategic tasks, addressing water sector

demands. Operators must understand desalination installation instructions to operate plants effectively. Advanced control systems are essential for profitable areas, but are rarely used in RO desalination plant treatments.

In [1], the application of artificial neural networks to find desalination plant issues is discussed. Nanotechnology and the membrane distillation system were used to build an authentic system based on precise measurements. To find system failures, faults between the model outputs and system outputs were then categorized. In order to lower costs, test decision rules, and detect poorly organized domains in a dataset, the research [2] suggested a Whale Optimizing Algorithm (WOA) based on chaos theory and fuzzy logic to discover flaws during wastewater treatment.

In [3], the creation of a failure detection and isolation (FDI) system for an ISROD plant in Lima, Peru, is described using a structural analysis methodology. In [4] and [5], a desalination plant is subjected to a predictive control based on a linear model. A non-linear model-based predictive control is developed by [6], [7] and [8], applied to a typical RO desalination plant for supervision, defect detection and advanced control. The literature [9] also has a Real-Time Fault Tolerant Control application in addition to the Fault Tolerant Control methodologies that are given. Despite these efforts, there are now only a small number of projects interested in monitoring and supervising the RO plant. The paper [10] describes the automation of an Industrial RO plant using PLC and SCADA. A monitoring and control system for the reverse osmosis desalination water treatment plants was developed in [11].

In [12], the authors designed a remote real-time communication test system consisting of MATLAB, Ethernet, and PLC. The test results show that the function of real-time data exchange between MATLAB and PLC can be obtained remotely through the OPC server, and real-time remote communication between MATLAB and PLC is effective and feasible. In [13], an air compressor's vibration and sound data were captured using an accelerometer and a microphone, respectively. Through OPC continuous

monitoring, data has been archived. The information is recorded using detection units. In our recent investigation, we were able to obtain real-time findings, which we then used in an actual desalination plant. We made a number of changes and adjustments to the PLC (Programmable Logic Controller) and SCADA (Supervisory Control and Data Acquisition) systems to address and resolve systemic problems. In [14], the experimental study verifies a cascade control system that implements PID control on MicroLogix-1200 PLC and RSView-32 SCADA using RSLinx communication software. It uses SCADA, PLC, and OPC architectures. In [15], using real-time simulation from Opal-RT, a commercially accessible but expensive solution, the study provides a home intelligent management system.

II. CONTEXT AND PROBLEMATIC

A desalination plant faces challenges in maintaining optimal water quality, production effectiveness, and energy consumption. Changes in salinity, temperature, and feed water quality can impact the desalination plant's performance. Balancing energy efficiency with freshwater production is challenging. Environmental factors like the disposal of brine waste and the marine ecosystem must be considered. Effective control systems must also consider plant dependability and maintenance. A holistic approach is needed to manage water quality, energy efficiency, environmental considerations, and operational dependability in a desalination plant. Also, the use of different manufacturers of modern equipment, like PLCs in an industrial plant, can be complex due to potential compatibility issues, communication discrepancies, and integration problems. This can lead to operational issues or system failures. Therefore, careful planning, robust communication protocols, and comprehensive testing are crucial for efficient and reliable operations [16].

This work uses the OPC protocol to establish a robust connection between PLCs, allowing seamless communication without proprietary software or complex integrations. Intelligent controllers in industrial plants improve efficiency and productivity, but require advanced algorithms, hardware compatibility, specialized expertise, and security due to their complexity. Integrating PLC (Programmable Logic Controller) with AI (Artificial Intelligence) technologies can be an investment with both initial costs and potential long-term benefits, for integrating artificial intelligent codes and control systems.

MATLAB and SIMULINK are effective tools. You will utilize them to create your control plans and AI algorithms. MATLAB is a powerful computational and analysis platform that offers data processing, visualization, and automation capabilities [17]. The combination allows real-time data acquisition and analysis, enabling rapid decision-making in industrial automation, manufacturing, or process control. In this work, we make an interaction between MATLAB and PLCs. This connection entails the use of MATLAB as a tool for designing, simulating, and producing AI control algorithms that cannot be installed on the PLC. The OPC-MATLAB integration also allows for custom dashboards and graphical user interfaces for monitoring and controlling multiple PLCs from a centralized location. This combination enhances data logging and historical analysis, ensuring smoother operations and better decision-making across various industrial and engineering applications.

The research described in this paper makes significant contributions to the engineering field, specifically within embedded systems, sustainable development, control and supervision. Scientists may now write algorithms in MATLAB and use them through a PLC by using the capabilities of the PLC, which is compatible with the OPC protocol and PLC software. This creates new opportunities for research and development in these fields [19].

III. PROPOSED ARCHITECTURE

A. Communication Between MATLAB/Simulink and PLC

Real-time control was achieved in this project by using the MATLAB-PLC communication technology, which is based on the OPC server [20]. A gateway was developed to enable communication between the OPC server and MATLAB. PLCs offer numerous powerful features and can cater to a wide range of applications. Data was accessible from PLC software through the OPC server.

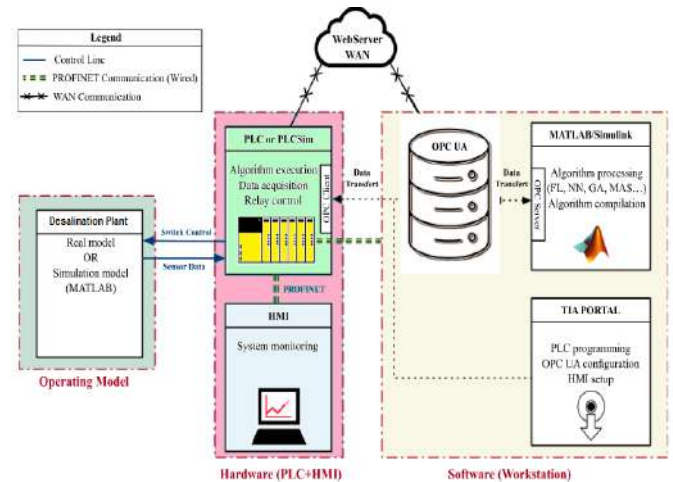


Figure 1: The designed platform configuration.

B. Fuzzy Logic

Fuzzy logic is a mathematical approach used to handle imprecise or uncertain data, offering degrees of truth or membership compared to traditional Boolean logic. It is used extensively in control systems, decision-support systems, and image processing, where conventional logic may not handle real-world data complexity. Fuzzy logic excels at approximate reasoning and linguistics-based decision-making, making it a versatile and effective method for handling ambiguity and vagueness.

IV. APPLICATION

A. R.O. Plant description

The article describes a reverse osmosis installation in TOUGGOURT, which uses demineralized water from four geothermal drillings and four cooling towers to produce potable water. Raw water is cooled, filtered, and pretreated to improve compatibility with reverse osmosis membranes, enhancing their operational life and ensuring efficient water treatment. A high-pressure pump system provides pressure for reverse osmosis units, transporting treated water to reservoirs and distribution tanks. A remote management system supervises pump operation.

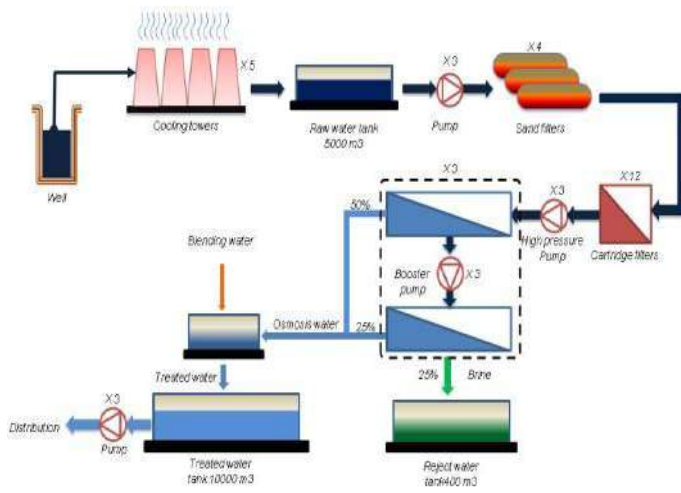


Figure 2: Synoptic diagram of the desalination plant.

The project utilizes an OPC server-based MATLAB-PLC communication system for real-time control of a reverse osmosis desalination plant, using Schneider Electric's MODICON M340 programming logic controller for adaptability.

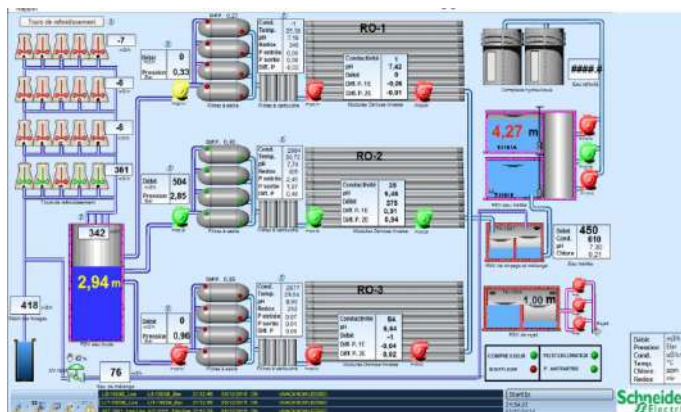


Figure 3: Synoptic diagram of the desalination plant (Supervision software screen).

B. OPC server

The OPC Read (OPC Read) and OPC Write (OPC Write) blocks in SIMULINK are in charge of sending data to the OPC server. These blocks have a client manager that may be used to locate and control the OPC server. Users can choose individual objects and specify sample intervals using the client manager as necessary [21].

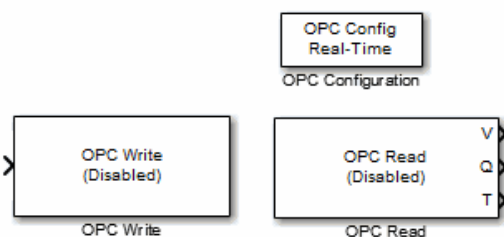


Figure 4: Data acquisition OPC server SIMULINK library.

C. Real-time implementation of fuzzy controllers

The programmable logic controller continually receives data from level transmitters, valve and pump states, and other sources. The OPC server is then used to provide all of this data to MATLAB. To control the system, two MATLAB-created fuzzy controllers are used. These fuzzy controllers analyze the data and feed it back to the PLC so that it may issue the required orders and commands.

1. Based on the upper and lower limits of the water level in the tank and the water tower, the first fuzzy controller was created to regulate the storage tank equipment and the intermediate valve.
 2. The second fuzzy controller is in the role of managing the water tower equipment and adjusting the flow in accordance with the speed of the pumps. The second controller's functionality was tested by changing the setpoint value at various time intervals.
- The intermediate valve will first open if the water level in the storage tank is below the upper limit.
 - The intermediate valve closes when the water level exceeds the upper limit and stays closed until the water reaches the lower limit. The intermediate valve reopens when the water level reaches the lower limit.
 - The PLC receives data from the water level transmitter in the water tower in the case of the water tower level controller through the GSM communication system. Input data from the SCADA setpoint is also received by it, and it sends it to the fuzzy controller.
 - After processing the data, the fuzzy controller delivers it back to the PLC so that it may issue the required commands and instructions.
 - The controller boosts the pump speed to raise the water level in the water tower if it is below the setpoint level.
 - The controller will slow down if the water level is high.
 - The speed of both pumps becomes constant when the setpoint and the water level are both at the same location.

The connectivity between the PLC, SCADA, and MATLAB/SIMULINK is depicted in the following figure.

1. First Controller Design: The FIS editor is the initial stage in developing a fuzzy controller. The "Reservoir" fuzzy controller has the fuzzy inputs "Level" and "Rate," as well as the fuzzy output "Valve".
2. Second Controller Design: Two fuzzy input variables—"Level" and "Error"—as well as two fuzzy output variables—"Distribution" and "Pump Speed"—make up this controller. The figures below show the two FIS controllers and the membership functions editor.

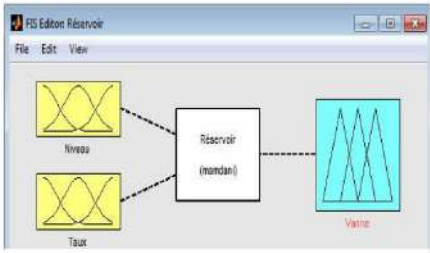


Figure 5: Fuzzy Controller 1.

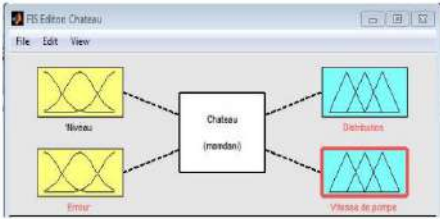


Figure 6: Fuzzy Controller 2.

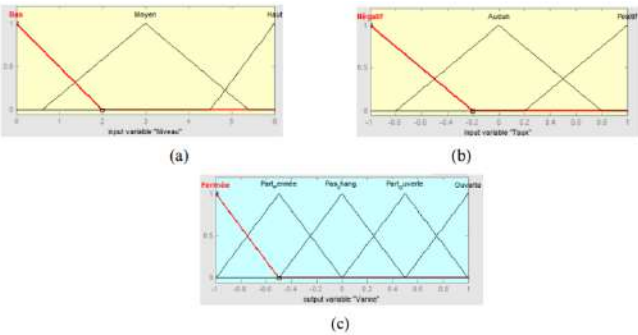


Figure 7: Membership functions.

D. Operating Principles for Handling Defects

The desalination plant's programmed control mechanism is extremely accurate. Because reverse osmosis membranes are both expensive and delicate, this accuracy is crucial. Any fault, no matter how little, has the potential to interrupt production streams and bring operations to a complete stop.

We have decided to improve the desalination operating system in order to deal with these issues. The goal is to eliminate production interruptions due to problems that may be readily fixed, such as high tank levels or varying chemical concentrations in the water.

Three key categories of documented problems at this desalination facility may be distinguished [2,17]:

- Diagnosis Defects: These occur when the system stops working suddenly, and there is no obvious place where the issue is located. We added more alerts to stop system shutdowns in order to enhance fault diagnosis.
- Process Defects: These are instances where extreme high pressures are recorded at the beginning or end of the operation.
- Control Defects: These involve problems with determining set points for certain system parameters.

E. New proposed operation philosophy

A small technical problem or brief suspension of production at a facility might have a negative impact on overall output and cause changes to the daily timetable for distributing potable water.

A new intelligent controller has been added to the system to improve the monitoring approach in order to maintain uninterrupted plant operation and after examining the numerous reasons for production interruptions.

Our goal is to replace traditional PID controllers with fuzzy logic controllers (like ANFIS) for pump control. Due to its versatility in handling uncertainty, fuzzy logic is still commonly utilized, but artificial neural networks (ANN) are used to interpret information through similarity comparison. The shortcomings of each model are intended to be lessened by the combination of fuzzy logic and ANN. While Neural Networks provide flexibility and algorithms for learning from actual data, Fuzzy Logic lacks clear rules and accurate ways for identifying membership functions [17].

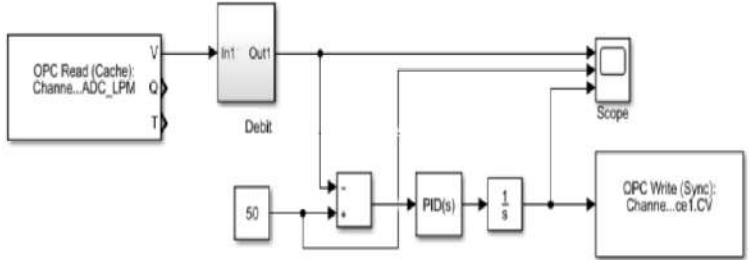


Figure 8: PID Control in SIMULINK.

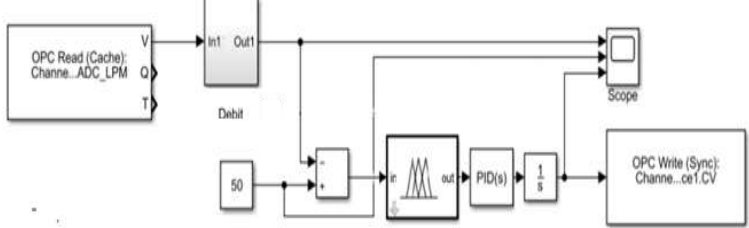


Figure 9: Fuzzy logic controller in SIMULINK.

F. Comparison between PID data and ANFIS data

Table 1 shows the performance of the PID controller in the example of a simple valve by changing the set point from 50 to 70 percent. Table 2 also shows the performance of the ANFIS controller in the same example.

Table 1. Performance of PID controller

Step	Rise Time	Settling Time	Overshoot
50-70%	11.98 s	108.14 s	12.64%

Table 2. Performance of ANFIS controller

Step	Rise Time	Settling Time	Overshoot
50-70%	13.52 s	23.28 s	1.02%

G. Connection between PLC, SCADA and OPC

1. Description of S7-PLCSIM Advanced software

We may run the program on a simulated PLC using the S7-PLCSIM physical module simulator, either on a computer

or a programming console, and test it there. The STEP7 software environment serves as the sole setting for this simulation. For the purpose of emulating Siemens controllers, SIMATIC S7-PLCSIM Advanced provides the option to create virtual controllers. Effective operation simulation may be achieved by using these virtual controllers. In the context of an installation or machine, it is also possible to evaluate and validate these virtual automations. An extensive PLC is offered to make integration with plant or machine simulations easier.

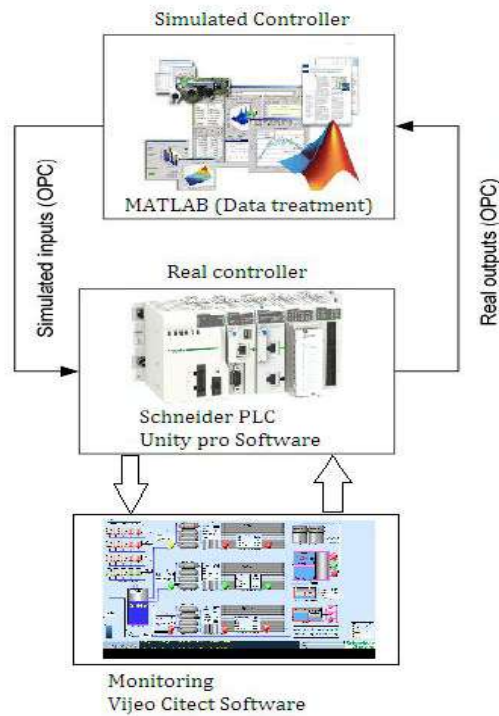


Fig.10: MATLAB and OPC-server, and PLC.

2. Description of the KEServerEX program

The most recent iteration in this area, KEServerEX, demonstrates the progress of communication technology. The KEServerEX solution, created over a period of more than 25 years, stands out as the most cutting-edge OPC communication server available and has been approved by the OPC Foundation. A variety of communication interfaces, including OPC Server DA, UA, AE, .NET, HDA, SNMP Agent, and DDE, are available on the KEServerEX communication platform.

The main function of the KEServerEX server, an OPC Data Access server, is to:

The following primary functions of the KEServerEX server, acting as an OPC Data Access server:

1. Collect data from a range of hardware components, such as photoelectric and infrared sensors, turntables, robotic arms, and more. These devices can receive data updates from it as well.
2. Assure the data's quality and freshness while guaranteeing its integrity.
3. Respond to client requests, which may entail delivering data from the server or sending data to the hardware that is linked.
4. Use event notifications to alert the client to changes in the status of variables.

H. Connection between PLC, SCADA, OPC and MATLAB

Within MATLAB and SIMULINK, the Industrial Communication Toolbox provides direct access to both historical and real-time OPC data. You may read, write, and save OPC data from a variety of devices, including distributed control systems, supervisory control and data acquisition systems, and programmable logic controllers, using this toolkit.

V. CONCLUSION

In the area of industrial automation, OPC is crucial. OPC server software is capable of carrying out tasks, including transferring field data to OPC servers before storing it in a database or connecting to SCADA software. An integrated automation system and information flow from the lowest to the highest level are now possible thanks to the OPC standard. The number of applications created employing OPC standards is rising as well. The suggested approach to increase the quality of control in the industrial setting is data processing and advanced control..

References

- [1] Maris Murugan, T., Kiruba Shankar, R.: Water desalination plant fault detection using artificial neural network (2022).
- [2] Achbi, M.S., Rouabah, B., Mahboub, M.A., Benarabi, B., Kafi, M.R. & Kechida, S. (2024). Fault-tolerant control and fault diagnosis methods integrated using an intelligent controller for hybrid dynamic systems. *Diagnostyka*, 25.
- [3] Anter, A.M., Gupta, D., Castillo, O.: A novel parameter estimation in a dynamic model via fuzzy swarm intelligence and chaos theory for faults in a wastewater treatment plant. *Soft Computing* 24(1), 111–129 (2020).
- [4] Perez-Zuniga, G., Rivas-Perez, R., Sotomayor-Moriano, J., S'anchez-Zurita, V.: Fault detection and isolation system based on structural analysis of an industrial seawater reverse osmosis desalination plant. *Processes* 8(9), 1100 (2020)
- [5] Bartman, A.R., McFall, C.W., Christofides, P.D., Cohen, Y.: Model-predictive control of feed flow reversal in a reverse osmosis desalination process. *Journal of Process Control* 19(3), 433–442 (2009)
- [6] Abbas, A.: Model predictive control of a reverse osmosis desalination unit. *Desalination* 194(1-3), 268–280 (2006).
- [7] Bajany, D., Zhang, L., Xia, X.: Model predictive control for water management and energy security in arid/semiarid regions. *Journal of Automation and Intelligence* 1(1), 100001 (2022)
- [8] Achbi, M.S., 2022. Surveillance et commande tolérante aux défauts appliquées à une classe des systèmes dynamiques hybrides (Doctoral dissertation).
- [9] Gao, L.X., Rahardianto, A., Gu, H., Christofides, P.D., Cohen, Y.: Novel design and operational control of integrated ultrafiltration—reverse osmosis system with RO concentrate backwash. *Desalination* 382, 43–52 (2016)
- [10] Mehrad, R., Kargar, S.M.: Integrated model predictive fault-tolerant control and fault detection based on the parity space approximations for a reverse osmosis desalination unit. *Transactions of the Institute of Measurement and Control* 42(10), 1882–1894 (2020)
- [11] Uddin, S.U., Baig, M.J.A., Iqbal, M.T.: Design and implementation of an opensource, internet-of-things based

- scada system for a community solar powered reverse osmosis desalination system. Design and Analysis of a Community Size Solar Powered Reverse Osmosis Desalination Water System for Pakistan, 103 (2023)
- [12] Vardar, E., Giraz, A.H., Orenba,s, H., Sahin, S.: Opc server based and real time " motor speed control with plc communication system. In: 2018 26th Signal Processing and Communications Applications Conference (SIU), pp. 1–4 (2018). IEEE.
 - [13] Bagal, K., Kadu, C., Parvat, B., Vikhe, P.: Plc based real time process control using scada and matlab. In: 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 1–5 (2018). IEEE
 - [14] Verma, N.K., Sharma, T., Maurya, S., Singh, D.J., Salour, A.: Real-time monitoring of mmarisines using open platform communication. In: 2017 IEEE International Conference on Prognostics and Health Management (ICPHM), pp. 124–129 (2017). IEEE
 - [15] Zerk, A.E., Ouassaid, M., Zidani, Y.: Development of a real-time framework between matlab and plc through opc-ua: A case study of a microgrid energy management system. Scientific African 21, 01846 (2023).
 - [16] Mohapatra, B., Sahu, B.K., Pati, S., Ghamry, N.A., Ghoneim, S.S., *et al.*: Application of a novel metaheuristic algorithm based two-fold hysteresis current controller for a grid connected pv system using real time opal-rt based simulator. Energy Reports 9, 6149–6173 (2023).
 - [17] Achbi, M.S., Kechida, S., Mhamdi, L., Dhouibi, H.: A neural-fuzzy approach for fault diagnosis of hybrid dynamical systems: demonstration on three-tank system.
 - [18] Yavuz, U., Cumali, K.: Real-time communication between s7-1200 plc and matlab/simulink and a fuzzy logic temperature humidity control application. Scientific Journal of Mehmet Akif Ersoy University 2(1), 7–14 (2019)
 - [19] Nicola, M., Nicola, C.-I., Duta, M., Sacerdotianu, D.: Scada systems architecture based on opc and web servers and integration of applications for industrial process control. International Journal of Control Science and Engineering 8(1), 13–21 (2018)
 - [20] Mhamdi, L., Achbi, M.S., Dhouibi, H., Kechida, S.: Diagnosis of hybrid systems through bond graph, observers and timed automata. Diagnostyka 21(3), 113–125 (2020)
 - [21] Aldiansyah, F., Wijaya, S.K., Prajitno, P.: Neuro-fuzzy-based water flow controller in prototype plant using programmable logic controller (plc). In: 2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE), vol. 6, pp. 64–69 (2019).

From Simplified to Realistic Drilling Dynamics Comparing 2DOF and 4DOF Models with Top-Drive Effects and Hook-Load-Based ROP

Hadibi Samira

Laboratory of Applied Automation,
Department of Automation
Faculty of Hydrocarbons and
Chemistry, University M'Hamed
Bougara
Boumerdes, Algeria
s.hadibi@univ-boumerdes.dz

Kidouche Madjid

Laboratory of Applied Automation,
Department of Automation
Faculty of Hydrocarbons and
Chemistry, University M'Hamed
Bougara
Boumerdes, Algeria
kidouche_m@hotmail.com

Doghmane Mohamed Zinelabidine

line Bob L. Herd Department of
Petroleum Engineering
Texas Tech University Lubbock 79401
Texas, USA
doghmane_m@yahoo.com

Abstract—This work presents a comparative study between a classical two-degree-of-freedom (2DOF) drill-string model and an extended four-degree-of-freedom (4DOF) formulation that incorporates top-drive dynamics and a more realistic representation of the rate of penetration (ROP). Unlike conventional 2DOF models, where the axial input is typically prescribed, the proposed 4DOF model computes the ROP based on the hook load, allowing the system to more accurately reflect operational drilling conditions. A comprehensive parametric study is conducted to evaluate the dynamic responses of both models under varying operational and formation parameters. The comparison focuses on key vibration phenomena, including stick-slip and bit bounce, and highlights the influence of the additional dynamics on system stability and transient behavior. The results demonstrate that incorporating top-drive motion and load-dependent ROP significantly affects the predicted dynamics, leading to notable differences in both the onset and severity of drilling vibrations. These findings underline the importance of using higher-order models for accurate prediction of drill-string behavior in complex drilling environments.

Keywords—drillstring dynamics, coupled vibration, nonlinear system, stick-slip

I. INTRODUCTION

Drilling is a complex and highly responsive industrial process through which hydrocarbons are extracted from deep underground reservoirs. The operation involves the use of a drill bit that penetrates abrasive rock formations, progressively advancing into the subsurface with the aid of drill strings that can extend up to 5 kilometers in length. The drilling system comprises heavy and costly equipment that is frequently exposed to failures and damage—most of which are attributed to the dynamic vibrations generated during the drilling process. These vibrations propagate in three primary directions: axial, torsional, and lateral [1]. The study of drilling stability in the presence of such vibrations has been a subject of extensive research for decades, aiming to develop effective control strategies to mitigate, or at least reduce, their detrimental impact on system performance and equipment longevity. Given the inherent complexity of the drilling system and the multitude of interacting components it

comprises, accurately modeling its behavior has consistently posed a significant challenge. Capturing and replicating its true dynamic characteristics remains a critical task for both analysis and control.

Due to the inherent complexity of drill-string dynamics, most studies rely on simplified models to analyze vibration phenomena. Such simplifications make analytical treatment more tractable, facilitate control-system design, and significantly reduce computational cost. However, these reduced-order models often overlook essential aspects of the real drilling dynamics.

A survey of the literature shows that many works adopt a coupled axial-torsional formulation in which each direction is represented by a single degree of freedom. The original version of this model, first introduced in [2], incorporates operational inputs directly as prescribed functions of time. This assumption departs from actual drilling conditions, where the torsional velocity is governed by the top-drive dynamics and the axial velocity results from the applied hook load, defined as the difference between the total system weight and the weight on bit. Other formulations go even further by omitting these operational influences altogether, as seen in [3].

Some studies have included the effect of top-drive dynamics, but only within uncoupled torsional models [4]. Conversely, one coupled axial-torsional formulation accounts for rigid-body motion but still neglects the top-drive behavior [5]. References [6–10] disregarded these dynamical effects when analyzing the behavior of the coupled axial-torsional system. They constructed stability regions using the rate of penetration and the top-drive rotational speed, even though these quantities cannot be directly adjusted by the operator and are instead indirectly tuned to generate the corresponding maps.

These gaps motivate the development of a more realistic representation. To address this limitation, a new four-degree-of-freedom (4DOF) model is proposed, explicitly incorporating both top-drive dynamics and the hook-load-dependent penetration rate. To evaluate the impact of these additional dynamics, the 4DOF formulation is systematically compared with the widely used 2DOF model. A parametric study is conducted to highlight the differences between the

two formulations and assess the relevance of including these additional physical effects.

II. SYSTEM MODEL

A. 4DOF Model

The rotary drilling system is represented by a lumped parameter model comprising four degrees of freedom: Two axial and Two torsional.

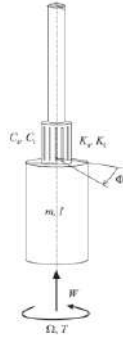


Fig. 1. A diagram of the drill-string model.

The system dynamics are described by a set of coupled second-order nonlinear differential equations, which can be deduced from [4-5]:

$$\begin{aligned} m\ddot{u}_r + c_{rg}\dot{u}_r + k_{rg}(u_r - u_x) &= mg - H_0 \\ M\ddot{u}_x + c_{ag}\dot{u}_x + k_{rg}(u_x - u_r) &= Mg - W_{ob} \\ J_d\ddot{\theta}_d + c_{dg}\dot{\theta}_d + k_{dg}(\theta_d - \theta_b) &= nT_m \\ J_b\ddot{\theta}_b + c_{bg}\dot{\theta}_b + k_{dg}(\theta_b - \theta_d) &= -T_{ob} \\ L\dot{I} + R_m I + K_m n\dot{\theta}_d &= V_c \end{aligned} \quad (1)$$

In this model u_r denote the top drive axial and θ_d its angular displacements, while u_x and θ_b represents the drill bit axial and angular displacements, respectively.

The top-drive unit is modeled as a separately excited DC motor, allowing independent regulation of the armature and field circuits and enabling precise control of the drill-string angular velocity.

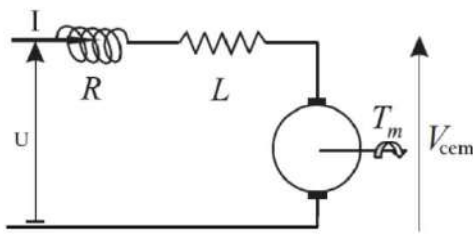


Fig. 2. Electrical diagram of the equivalent circuit of the DC motor armature.

Where:

$$T_m = K_m I \quad (2)$$

$$V_c = nK_m \Omega_0 \quad (3)$$

B. 2DOF Model

The 2DOF representation follows directly from the 4DOF

model by prescribing the top-drive axial motion as $V_0 t$ and its angular motion as $\Omega_0 t$, effectively removing the corresponding dynamic states. Which corresponds to the model used in [2]:

$$\begin{aligned} M\ddot{u}_x + d_x \dot{u}_x + k_x(u_x - V_0 t) &= -W_{ob} \\ J\ddot{\theta}_b + d_t \dot{\theta}_b + k_t(\theta_b - \Omega_0 t) &= -T_{ob} \end{aligned} \quad (4)$$

C. Bit-rock Interaction Model

The nonlinear terms $T_{ob}(t)$, torque on bit, and $W_{ob}(t)$, weight on bit, account for the bit-rock interaction forces, which are functions of both axial and torsional states. They are defined as [11]:

$$W_b = \begin{cases} s_f \left(\frac{W_0}{s_f} + U - V_0 t - d_0 \sin(n_b \theta) \right), & \text{if } \left(\frac{W_0}{s_f} + U - V_0 t \geq d_0 \sin(n_b \theta) \right) \\ 0, & \text{if } \left(\frac{W_0}{s_f} + U - V_0 t \leq d_0 \sin(n_b \theta) \right) \end{cases} \quad (4)$$

$$T_b = W_b r_b \left(\mu(\dot{\theta}) + \xi \sqrt{\frac{\delta_c}{r_b}} \right) \quad (5)$$

Where $\mu(\dot{\theta})$ the friction coefficient is modeled as a function of velocity to capture the weakening effect associated with relative motion.

$$\mu(\dot{\theta}) = \frac{2}{\pi} \tanh^{-1}(\varepsilon \dot{\theta}) \left(\frac{f_1 - f_2}{1 + \delta_c |\dot{\theta}|} + f_0 \right) \quad (6)$$

Here, δ_c corresponds to the depth of cut and R_p refers to the rate of penetration, which is defined as:

$$\delta_c = \frac{2\pi R_p}{\omega_0} \quad (7)$$

$$R_p = c_1 W_0 \sqrt{\omega_0} + c_2 \quad (8)$$

TABLE I. PARAMETERS DEFINITIONS AND VALUES

Parameter	Definition	Unit
V_0	Imposed Axial Velocity	m/s
Ω_0	Imposed Angular Velocity	rad/s
M	BHA mass	kg
m	Drill string mass + Top drive	kg
d_x, c_{dg}	BHA Axial Damping	N.s/m
c_{rg}	Top drive Axial Damping	N.s/m
k_x, k_{rg}	Axial Stiffness	N/m
J	Polar inertia	Kg.m ²
d_t, c_{bg}	BHA Torsional Damping	N.s/m
c_{dg}	Top drive torsional Damping	N.s/m
K_t, K_{dg}	Torsional Stiffness	N/m
H_0	Hook Load	KN
L	Steel Density	Kg/m ³
I	Current	A
K_m	Torque constant	N.m/A
n	Gearbox transmission ratio	/
s_f	Formation Stiffness	MN/m ²
d_0	Initial Drilling Depth	m
n_b	Drill Bit Cutter Number	/
r_b	Drill Bit Radius	m
ξ	Constant	/
ε	Constant	/
f_1	Constant	/
f_0	Constant	/
c_1	Constant	/
c_2	Constant	/

III. PARAMETRIC ANALYSIS

In this study, the objective is to assess the influence of the actual operating parameters embedded in the 4-DOF model. In this formulation, the operator-controlled inputs are the hook load (governing the penetration-rate adjustment) and the top-drive speed (governing the bit's rotational velocity). To this end, a comparative analysis is performed between the simplified 2-DOF model and the full 4-DOF model.

The 2-DOF formulation is obtained directly from the 4-DOF model by prescribing the axial and angular motions of the top drive as $V_0 t$ and $\Omega_0 t$, respectively, thereby eliminating the corresponding dynamic states. This allows a clear comparison between using (V_0, Ω_0) as operating parameters in the 2-DOF model and using the physically measurable parameters (hook load and top-drive speed) in the 4-DOF model.

The comparison focuses on the axial and torsional drill-bit responses under low and high operating conditions. For consistency, each prescribed penetration rate V_0 in the 2-DOF model is matched to the equivalent hook load obtained as the mean axial reaction from the 4-DOF top-drive response. Similarly, each prescribed torsional speed Ω_0 corresponds to the mean top-drive rotational response in the 4-DOF simulation. Identical initial conditions are used for both models. All simulations are performed in MATLAB using the ode45 solver.

A. Low speed

The nominal hook load applied during drilling operations typically ranges between 10 and 15 tons. Generally, an increase in hook load results in a decrease in penetration rate due to the increased axial force on the drill string, which affects bit-rock interaction and drilling efficiency [5]. In this simulation, the hook load was varied up to 800 kN to correspond to the penetration rate of 0.01 m/s, consistent with the minimal values reported in the literature [11]. While the top drive rotational speed was set to 2 rad/s.

- 4DOF Model

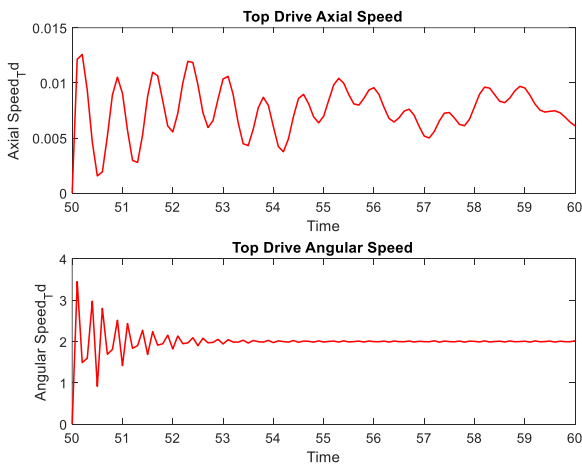


Fig. 3. Top drive dynamic response: $H_0 = 800$ kN, $\Omega_0 = 2$ rad/s

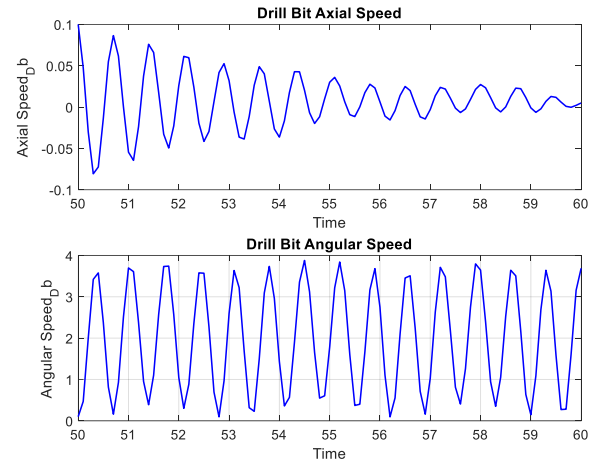


Fig. 4. Drill bit dynamic response $H_0 = 800$ kN, $\Omega_0 = 2$ rad/s

Figures 3 and 4 illustrate the simulated responses of the top drive and the drill bit in both axial and torsional directions, respectively. The simulations were conducted under the operating conditions of a hook load $H_0 = 800$ kN and a top drive rotational speed $\Omega_0 = 2$ rad/s.

As shown in Figure 3, the top drive axial velocity exhibits small oscillations around the nominal driving axial velocity of approximately 0.008 m/s. Correspondingly, the drill bit axial response, depicted in Figure 4, displays transient signs of bit bounce during the initial four seconds of operation. Following this transient period, the axial speed of the bit stabilizes at roughly 0.005 m/s, closely matching the imposed top drive axial velocity.

Regarding the angular response, the top drive shows minor oscillations during the first few seconds and then quickly stabilizes at the target rotational speed of 2 rad/s, as evident in Figure 3. In contrast, the drill bit torsional response, also shown in Figure 4, exhibit non-continuous stick-slip oscillations.

These results clearly demonstrate that the dynamic behavior of the drill bit is predominantly governed by the top drive. The axial and torsional motions of the bit closely follow the imposed top drive inputs, indicating that the drill string effectively transmits the top drive dynamics downhole. Consequently, under the examined operating conditions, the top drive can be considered the primary driver of the bit's response, with transient deviations quickly attenuated by the system's inherent damping and stiffness characteristics.

- 2DOF Model

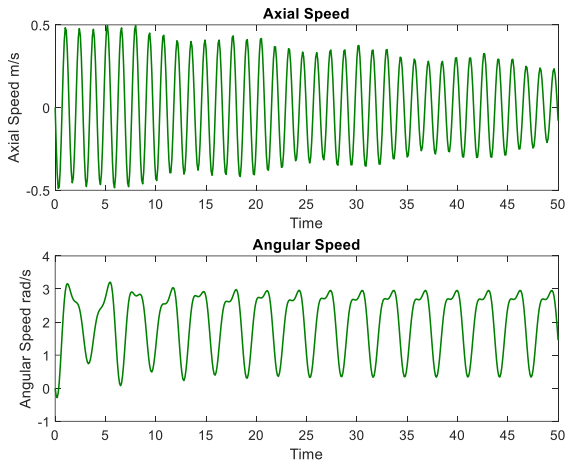


Fig. 5. Drill bit dynamic response: $V_0 = 0.005$ m/s, $\Omega_0 = 2$ rad/s

Figure 5 reveals that the drill-bit axial velocity reaches values as high as 0.5 m/s, despite the imposed axial penetration speed being only 0.005 m/s. This amplification is unrealistically large and does not align with experimentally reported bit-bounce levels in the literature, where the axial velocity typically increases by a factor of two or three relative to the imposed motion. Moreover, the bit-bounce phenomenon persists throughout the entire simulation interval, with the oscillations decreasing only in amplitude but never fully disappearing.

With respect to the torsional response, the results indicate the presence of continuous torsional oscillations; however, no evidence of stick–slip behavior is observed. The system therefore remains in a regime characterized by oscillatory torsional dynamics rather than the discontinuous torque–velocity cycles that typify stick–slip instability.

B. High speed:

To achieve the maximum penetration rate of 0.15 m/s [11], the hook load was decreased to 10 kN, while the top-drive rotational speed was set to 12 rad/s.

• 4DOF Model

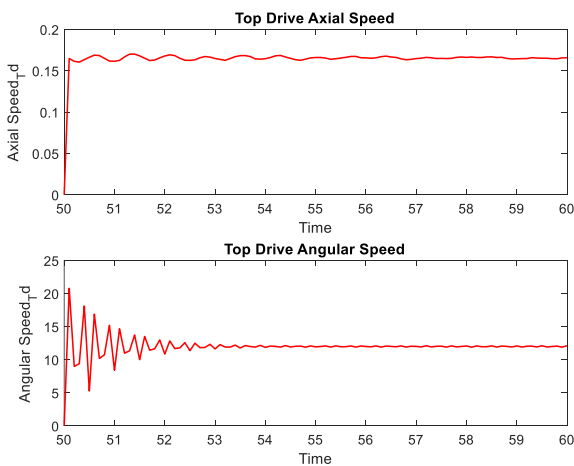


Fig. 6. Top drive dynamic response $H_0 = 10$ kN, $\Omega_0 = 12$ rad/s

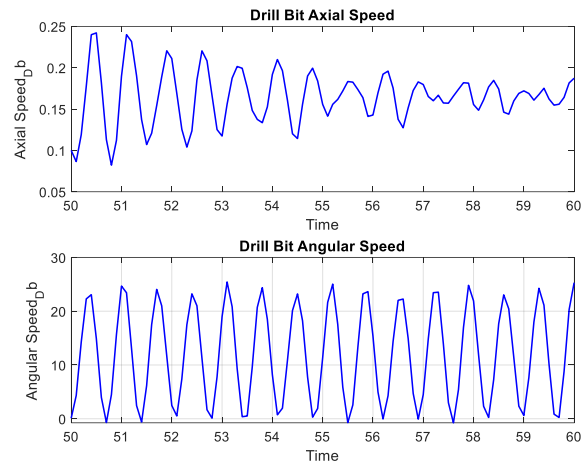


Fig. 7. Drill bit dynamic response: $H_0 = 10$ kN, $\Omega_0 = 12$ rad/s

In Figure 6, the top drive axial response exhibits a ramp-type evolution that gradually converges toward a steady axial velocity of approximately 0.15 m/s, with only small-amplitude oscillations superimposed on the steady trend. From this behavior, it can be inferred that a hook-load of 10 kN results in an imposed steady axial penetration speed of about 0.15 m/s.

In Figure 7, the drill bit axial velocity initially rises to a peak value of approximately 0.24 m/s before decreasing and oscillating around the steady operating velocity of 0.15 m/s, consistent with the top-drive input. No evidence of bit-bounce phenomena is observed, indicating that the axial subsystem remains in a stable penetration regime under these operating conditions.

• 2DOF Model

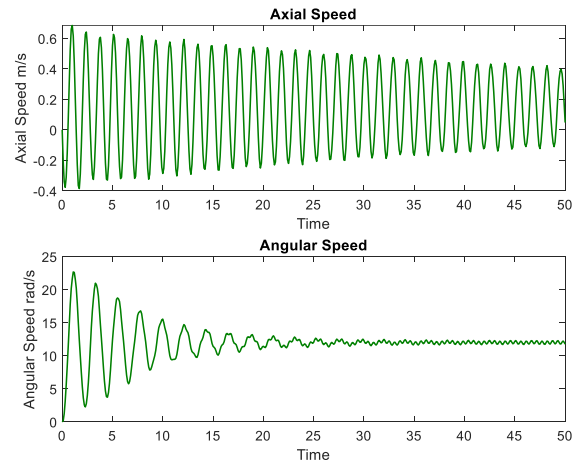


Fig. 8. Drill bit dynamic response: $V_0 = 0.15$ m/s, $\Omega_0 = 12$ rad/s

The axial response of the 2-DOF model, shown in Figure 8, exhibits a behavior similar to that observed under low imposed axial velocities, but with significantly larger oscillation amplitudes. The drill-bit axial velocity reaches peak values of approximately 0.6 m/s, despite the imposed top-drive axial speed being only 0.15 m/s. This indicates the presence of pronounced bit-bounce oscillations, with an amplification factor that again exceeds values typically reported in experimental studies.

The angular response in Figure 8 shows initial torsional oscillations during the transient phase, yet no evidence of stick-slip dynamics is observed. After this short transient period, the bit's rotational speed progressively converges toward the imposed top-drive velocity of approximately 12 rad/s, indicating a stable torsional operating regime.

IV. CONCLUSION

In this study, we have demonstrated the significant impact of incorporating top-drive dynamics and hook load into a coupled axial-torsional drill-string model. The inclusion of these operator-controlled inputs in the 4-DOF formulation was shown to meaningfully influence drill bit dynamic responses, particularly in capturing stick-slip oscillations and brief bit-bounce transients, which are poorly represented in simplified models.

The analysis considered two representative operating regimes—low and high mode speeds—which correspond to common field conditions. Across both operating regimes, the 4-DOF model exhibits realistic axial and torsional dynamics, capturing brief bit-bounce transients, stable axial penetration behavior, and—when appropriate—stick-slip oscillations. Its responses remain quantitatively consistent with known drilling physics and experimental observations. In contrast, the 2-DOF model consistently generates unrealistic dynamic behavior. It systematically overestimates axial oscillation amplitudes—by factors far exceeding physically observed bit bounce—and fails to capture stick-slip behavior in the torsional subsystem. These discrepancies persist regardless of the imposed axial speed.

These findings underscore that the 4-DOF formulation provides a physically meaningful representation of drill-string dynamics, accurately reflecting the coupled behavior between axial and torsional motions under varying operating conditions. The results clearly demonstrate that including top-drive dynamics and hook load is essential for reliable prediction of downhole responses, especially for phenomena such as stick-slip and bit bounce, which directly affect drilling efficiency and wellbore integrity. Consequently, any analysis or control strategy aimed at mitigating vibrations and optimizing performance should preferentially employ a high-fidelity, 4-DOF model rather than a simplified 2-DOF approximation.

REFERENCES

- [1] Tucker, W. R., and Charles Wang. "On the effective control of torsional vibrations in drilling systems." *Journal of Sound and Vibration* 224.1 (1999): 101-122.
- [2] T. Richard, C. Gernay and E. Detournay. "Self-excited stick-slip oscillations of drill bits," *Journal of Petroleum Science and Engineering* 59 (2007) 289–299
- [3] Meddah, S., Tadjer, S. A., Kidouche, M., Doghmane, M. Z., & Idir, A. (2024). Interaction Analysis and Decoupling of Axial-Torsional Vibrations in Rotary Drilling Systems. *International Journal of Automation and Safety*, 2(1), 33-39.
- [4] Riane, R., Doghmane, M. Z., Kidouche, M., & Djezzar, S. (2022). Observer-based H_∞ controller design for high frequency stick-slip vibrations mitigation in drill-string of rotary drilling systems. *Vibration*, 5(2), 264-289.
- [5] Bakhtiari-Nejad, F., & Hosseinzadeh, A. (2017). Nonlinear dynamic stability analysis of the coupled axial-torsional motion of the rotary drilling considering the effect of axial rigid-body dynamics. *International Journal of Non-Linear Mechanics*, 88, 85-96.
- [6] Depouhon, A., & Detournay, E. (2014). Instability regimes and self-excited vibrations in deep drilling systems. *Journal of Sound and Vibration*, 333(7), 2019-2039.
- [7] Gupta, S. K., & Wahi, P. (2016). Global axial-torsional dynamics during rotary drilling. *Journal of Sound and Vibration*, 375, 332-352.
- [8] Zheng, X., Agarwal, V., Liu, X., & Balachandran, B. (2020). Nonlinear instabilities and control of drill-string stick-slip vibrations with consideration of state-dependent delay. *Journal of Sound and Vibration*, 473, 115235.
- [9] Deng, P., Zhang, A., Fu, K., & Li, H. (2021). Nonlinear vibration of a time-space coupled drill string system based on the surface morphology of rock. *Journal of Sound and Vibration*, 506, 116153.
- [10] Li, Xinye, et al. "Theoretical and Numerical Analysis of Coupled Axial-Torsional Nonlinear Vibration of Drill Strings." *Shock and Vibration* 2021.1 (2021): 8046635. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [11] S. Meddah, I. Abdelhakim, S. Tadjer, M Doghmane, and M. Kidouche. "PID Control Design of Strongly Coupled Axial-Torsional Vibrations in Rotary Drilling Systems." *Przegląd Elektrotechniczny* 2024, no. 3 (2024).

Magnetic Levitation System Modelling and Control

L. MILOUDI

Electrical Systems Engineering
Department
Faculty of Technology
University of Boumerdès, Boumerdès,
Algeria

lamiloudi@univ-boumerdes.dz

B.LAOUZI

Electrical Systems Engineering
Department
Faculty of Technology
University of Boumerdès, Boumerdès,
Algeria

bassemtch13@gmail.com

O.Dif

Electrical Systems Engineering
Department
Faculty of Technology
University of Boumerdès, Boumerdès,
Algeria

omardif81@gmail.com

Abstract— This paper presents the modeling and control of a nonlinear magnetic levitation system. A mathematical model is derived to describe the dynamics relating coil current, magnetic force, and the vertical position of the levitated object. Two control strategies are developed and compared: a classical Proportional-Integral-Derivative (PID) controller and a fuzzy logic controller. The PID controller is tuned using standard techniques, while the fuzzy controller is designed based on linguistic rules and membership functions to better address the system's nonlinear behavior. Simulation results show that although the PID controller provides acceptable stabilization, the fuzzy controller achieves faster response, reduced overshoot, and improved robustness to disturbances and parameter variations

Keywords— *magnetic levitation system; PID controller; fuzzy logic controllers*

I. INTRODUCTION

Magnetic levitation (maglev) systems have emerged as an important class of electromechanical systems due to their ability to suspend objects without mechanical contact, thereby eliminating friction and enabling ultra-precise motion control. Such systems play a crucial role in applications ranging from high-speed transportation, precision positioning, clean-room manipulation, and active magnetic bearings to laboratory control benchmarks. Despite their advantages, maglev systems remain inherently difficult to control because they are open-loop unstable, strongly nonlinear, and sensitive to parameter variations and external disturbances. Recent contributions to maglev modelling have focused on improving the fidelity of nonlinear dynamic representations and incorporating structural or environmental uncertainties. In [1], Qiang et al. proposed a nonlinear adaptive model that identifies uncertain electromagnetic parameters online, enabling improved handling of parameter variations often encountered in practical levitation systems. This work highlights the need for uncertainty-aware modelling in modern maglev control. Zhu et al. [2] introduced a nonlinear dynamic model that incorporates the elastic deformation of the guideway, addressing the limitations of rigid-track assumptions commonly adopted in EMS modelling. Their study demonstrates that track flexibility significantly affects

levitation dynamics and controller performance. In a broader contribution, Zhu, Wang, and Ni [3] provided a comprehensive classification of maglev models, including electromechanical models, observer-based models, and system-identification models. Their review reveals a shift toward hybrid modelling approaches combining physics-based and data-driven representations for improved robustness and computational efficiency. Nonlinear control remains central to maglev research due to the system's intrinsic nonlinearity.

Several adaptive and backstepping approaches have been proposed. Jastrzębski and Kabziński [4] introduced an adaptive fuzzy inverse-dynamics controller based on an acceleration inversion model. The approach combines fuzzy modelling with Lyapunov-based adaptation to guarantee uniform ultimate boundedness (UUB) of the levitated object. Sliding-mode control (SMC) has long been favored in maglev systems for its inherent robustness to disturbances and parameter uncertainties. Several advanced SMC techniques have been introduced in recent years. Shao et al. [5] proposed an Adaptive Nonsingular Terminal Sliding Mode Control (ANTSMC) strategy that avoids the singularities present in classical terminal SMC while achieving finite-time convergence. The adaptive component compensates for unknown parameters in the electromagnetic force model. Ma'arif et al. [6] demonstrated the robustness of classical SMC against variations in mass and external disturbances using a standard 1-DOF magnetic levitation test bench. Their results reaffirm that SMC provides a strong baseline for robust maglev control. Predictive control has become increasingly prominent in maglev research, particularly due to its ability to handle input constraints and optimize system performance over a prediction horizon. Novotný et al. [7] developed a real-time Nonlinear Model Predictive Control (NMPC) framework for fast maglev dynamics. Their approach performs online linearization at each prediction step and demonstrates significantly improved performance compared to PID control. Tube-based MPC for maglev was introduced by Han et al. [8], who integrated a disturbance observer into a Tube-MPC structure to compensate for modeling errors and external disturbances. Their method improves gap regulation

performance under uncertainties. Liu and Dou [9] combined MPC with neural-network-based model correction, yielding a learning-enhanced MPC capable of adapting to unmodelled or time-varying dynamics. Their approach highlights the growing trend toward learning-aided predictive control. Li et al. [10] extended this framework by incorporating a robust state observer, enabling disturbance rejection and enhanced gap regulation under guideway uncertainties. Their observer-based adaptive controller addresses the practical challenge of unmeasurable system states in maglev applications.

II. SYSTEM MODELLING

The model for a maglev system is built upon the fundamental principles of electromagnetism and mechanics. When current flows through a coil, it generates a magnetic field, the strength of the magnetic field depends on the current, coil geometry, and core material. The interaction between the magnetic field and a ferromagnetic object (like the steel ball) creates an electromagnetic force. The motion of the ball is governed by Newton's second law of motion (force equals mass times acceleration). The forces acting on the ball include gravity (pulling it downwards) and the electromagnetic force (levitating it). To model the system mathematically, we define state variables that represent its dynamic behavior. Common choices for a maglev system include: the vertical displacement of the ball from its equilibrium position (x) and the velocity (v) the rate of change of the ball's position. The magnetic levitation system is shown in above figure 1. which consists of a ball of mass m , electromagnetic force, and gravitational force G [11].

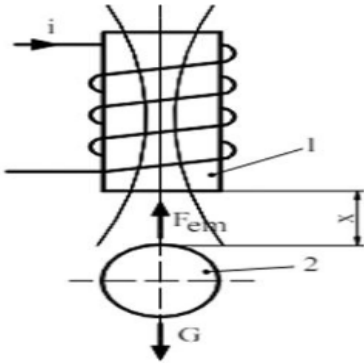


Fig1: Schema of the magnetic levitation system[10] (Maglev)

The relationship between the electromagnet voltage $\Delta V(s)$ and the disturbance sensor voltage is as follows [12]:

$$\frac{\Delta E(s)}{\Delta V(s)} = \frac{r(s^2 - \frac{3C_0}{md_0^4}) + \frac{26RC}{md_0^4}}{(Ls + R)(s^2 - \frac{3C_0}{md_0^4})} \quad (1)$$

Table 1: Parameters of proposed system [12]

	Parameter	Value	Unit
Sensor	β	5.64×10^{-4}	V.m ²
	γ	0.31	V/A
	α	2.48	V
Operation point	i_0	1	A
	d_0	20	mm
Electromagnet	C	2.4×10^{-6}	Kgm ⁵ /s ² A
	R	2	Ω
	L	15×10^{-3}	H
	$m = M/4$	0.02985	kg

Thus, the state –space representation of the linearized model of equation (1), can be represented by the following state equations [12]:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 3\frac{C}{md_0^4} & 0 & -\frac{C}{md_0^4} \\ 0 & 0 & -\frac{R}{L} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ -\frac{R}{L} \end{bmatrix} u$$

$$Y = \begin{bmatrix} -2\frac{C}{d^3} & 0 & \gamma \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad (2)$$

State variable equation and transfer function proposed for this study are [12]

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1471 & 0 & -9.81 \\ 0 & 0 & -133 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 66.66 \end{bmatrix} u$$

$$Y = \begin{bmatrix} -144 & 0 & 0.31 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad (3)$$

$$G(s)H(s) = \frac{20.66s^2 + 61603}{s^3 + 132.5s^2 - 1471s - 194900} \quad (4)$$

Simulink block diagram of magnetic levitation detailed nonlinear model is shown in figure 2.

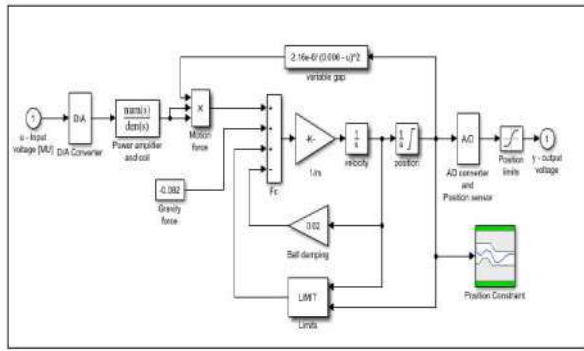


Fig2: Simulink Model of Magnetic Levitation Plant

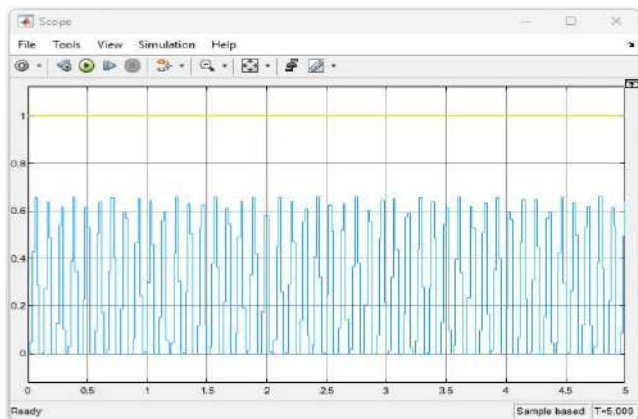


Fig3: System Stability Simulation Graph

The graph generated from the simulation, is shown in figure3, indicates that the system is highly unstable and requires regulation. The erratic and non-converging behavior of the signal demonstrates significant oscillations and deviations from a steady state, suggesting that the current control strategy is insufficient to maintain a stable levitation of the object. This instability can be attributed to inadequate damping and improper tuning of the control parameters, necessitating further analysis and adjustment to achieve a stable levitation system.

III. CONTROL STRATEGY

After determining that the magnetic levitation system is inherently unstable, it becomes imperative to implement a robust regulation strategy to ensure system stability and performance. Several control methods can be employed to achieve this, each with its own set of advantages and limitations. Among the various approaches available, Proportional-Integral-Derivative (PID) regulation and Fuzzy Logic control are widely recognized for their effectiveness in different contexts. In this section, we focus on these two prominent methods: PID regulation, a traditional approach known for its simplicity and ease of application, and Fuzzy

Logic control, an adaptive and versatile method capable of handling complex and non-linear systems. By applying these control strategies to the magnetic levitation system, we aim to assess their relative performance and determine which method provides superior stability and control under varying operational conditions.

The Simulink model of the magnetic levitation system was updated to include the PID controller, which was tuned using the Ziegler-Nichols method. The PID block was integrated into the existing model to regulate the system's behavior and ensure stable levitation of the ball. The updated model now includes the necessary control loops to manage the proportional, integral, and derivative actions effectively. The following figure 4 shows the simulation model with the added PID block.

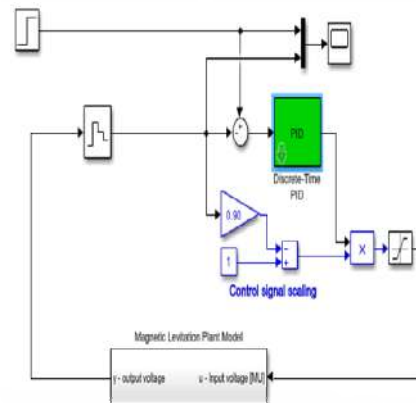


Fig 4: Simulink model with PID block

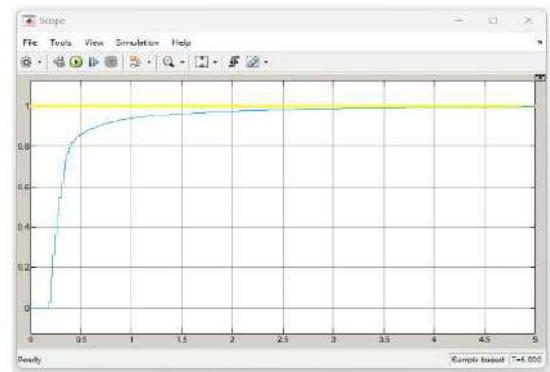


Fig 5: PID-Controlled Magnetic Levitation System Response to a Step Input

The graph shown in figure 5 illustrates the system's response to a step input. The blue curve shows a rapid increase in response within the first 0.5 seconds, similar to the ball position.

Fuzzy logic, an extension of Boolean logic rooted in its mathematical theory of fuzzy sets, broadens the scope of

classical set theory. Incorporating the concept of degree into condition verification, it allows conditions to exist in states beyond simply true or false. This flexibility grants significant adaptability to reasoning processes, enabling the consideration of imprecisions and uncertainties. One notable advantage of fuzzy logic in formalizing human reasoning is the expression of rules in natural language.

The membership functions chosen for our study are triangular functions figure 5. This choice is based on their simplicity and effectiveness in representing degrees of membership in a fuzzy system. The simulator used allows for easy selection of these functions and ensures that they respect the domain of the error and its derivative. Thus, all inputs to the controller are included within this domain, ensuring precise and efficient modeling of the behavior of the magnetic levitation system.

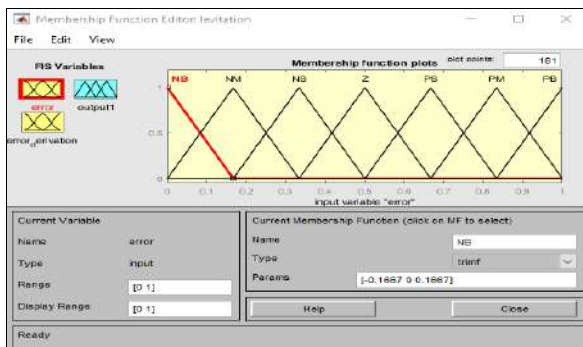


Fig 5: Membership Functions: Error

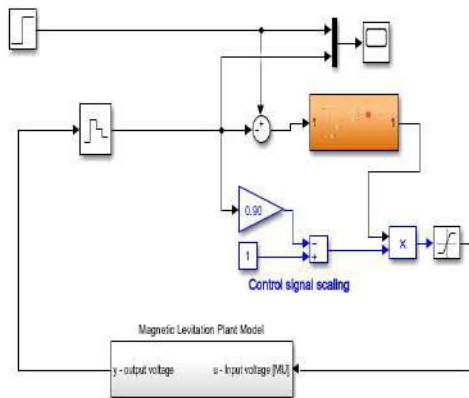


Fig 6:. Simulink model with fuzzy logic controller block

Figure 6 represents the block diagram of simulation including the mathematical model of the system, the closed-loop control, the reference and fuzzy logic controller.

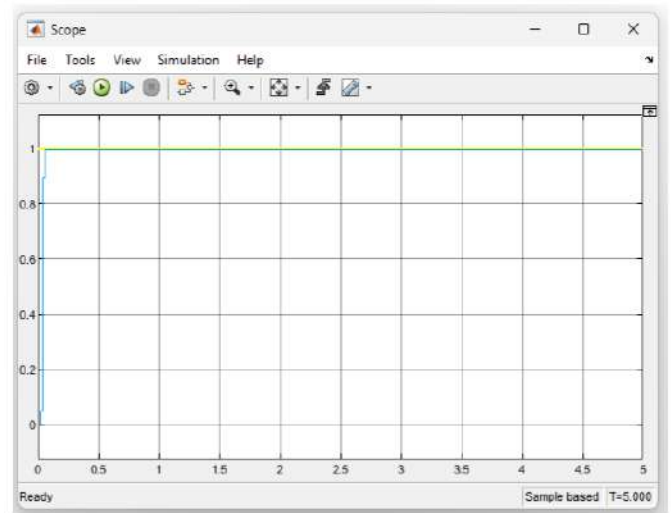


Fig7: Complete System Response to fuzzy logic conyroller.

The figure 7 represents the comprehensive system response over an extended period, capturing both the input control signals and the resulting output signals that depict the ball's position. This graph was obtained by utilizing the scope block within the simulation model (figure 6). The scope block is configured to display both the reference input (desired position) and the actual output (measured position) simultaneously, providing a detailed visualization of the system's dynamic behavior under fuzzy control. This figure allows us to observe how the fuzzy logic controller adjusts the input control signals over time to maintain the ball at the desired position. The extended period of the simulation helps in evaluating the long-term stability and effectiveness of the control strategy.

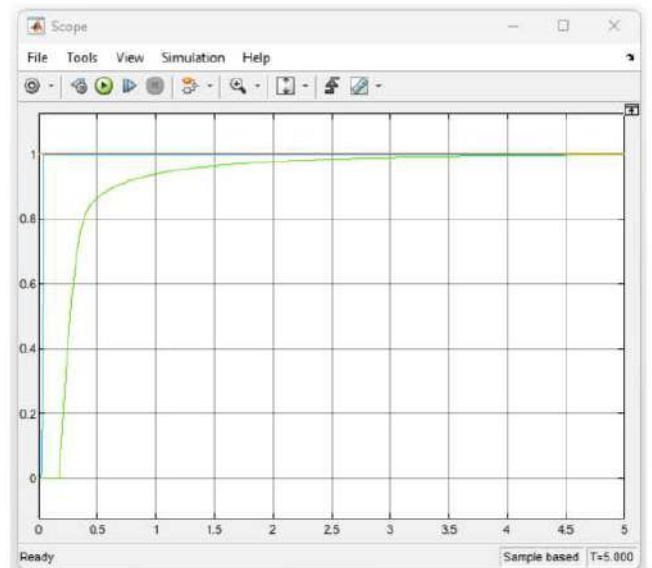


Fig 8:. Comparison of Desired PID and Fuzzy Logic Controller Responses on Magnetic Levitation System

The graphical comparison shown in figure 8 between the PID and Fuzzy Logic controllers illustrates the differences in how each method manages the dynamics of a magnetic levitation system. This comparison sheds light on the practical performance of each control strategy, highlighting their strengths and weaknesses in achieving desired system responses. The graph depicts the desired response in red, the PID response in green, and the Fuzzy Logic response in blue. Observations from the graph include:

CONCLUSION

This contribution explored the stability and control of magnetic levitation systems, comparing PID and Fuzzy Logic controllers. PID controllers, while historically significant and simple, require frequent parameter adjustments for complex environments. In contrast, Fuzzy Logic controllers offer superior robustness and adaptability, handling non-linearities and disturbances effectively without constant retuning. The analysis demonstrated that while PID controllers are suitable for certain applications, Fuzzy Logic controllers provide a more resilient and high-performing solution for advanced magnetic levitation systems. This sets a solid foundation for developing sophisticated control strategies, potentially incorporating machine learning for enhanced stability and performance.

REFERENCES

- [1] X. Qiang *et al.*, “Adaptive Nonlinear Control of an Electromagnetic Levitation System,” *Actuators*, vol. 14, no. 3, p. 115, 2025, doi: **10.3390/act14030115**.
- [2] X. Zhu, S. Zhang, Z. Zhou, and Y. Li, “Adaptive Backstepping Control for Maglev with Flexible Track,” *Mathematics*, vol. 12, no. 13, p. 2134, 2024, doi: **10.3390/math12132134**.
- [3] H. Zhu, J. Wang, and W. Ni, “Review of Levitation Control for Low-/Medium-Speed Maglev Systems,” *Buildings*, vol. 14, no. 3, p. 837, 2024, doi: **10.3390/buildings14030837**.
- [4] S. Jastrzębski and J. Kabziński, “Adaptive Fuzzy Inverse Model Control of a Magnetic Levitation System,” *Scientific Reports*, vol. 14, p. 76067, 2024, doi: **10.1038/s41598-024-76067-9**.
- [5] Q. Shao *et al.*, “Adaptive Nonsingular Terminal Sliding Mode Control for Maglev Systems,” *Applied Soft Computing*, 2024, doi: **10.1016/j.asoc.2024.110476**.
- [6] M. Ma’arif *et al.*, “Sliding Mode Control of a 1-DOF Electromagnetic Levitation System,” *Journal of Robotics and Control*, 2022, doi: **10.18196/jrc.12989**.
- [7] J. Novotný *et al.*, “Nonlinear Model Predictive Control of a Fast Maglev System,” *Int. J. Dyn. Control*, 2025, doi: **10.1007/s40435-025-01878-1**.
- [8] X. Han *et al.*, “Disturbance-Observer-Based Tube MPC for EMS Maglev,” *High-Speed Railway*, 2024, doi: **10.1016/j.hspr.2024.01.001**.
- [9] Z. Liu and C. Dou, “Neural-Network-Aided MPC for Magnetic Levitation Systems,” *Actuators*, vol. 13, no. 9, p. 331, 2024, doi: **10.3390/act13090331**.
- [10] Z. Li *et al.*, “Observer-Based Robust Adaptive Backstepping for Maglev Systems,” *Control Engineering Practice*, 2025, doi: **10.1016/j.conengprac.2025.106230**.
- [11] Dolga, V., & Dolga, L. (2014). Modelling and Simulation of a Magnetic Levitation System. Annals of the Oradea University. Fascicle of Management and Technological Engineering, XXIII(3), 113–118.
researchgate.net/publication/265821696
- [12] Mundher H.A. Yaseena,, Haider J. Abd b. “ Modeling and control for a magnetic levitation system based on SIMLAB platform in real time,” *Results in Physics* 8 (2018) 153–159, doi.org/10.1016/j.rinp.2017.11.026

LoRaWAN-Enabled Internet of Medical Things (IoMT) for Remote Vital Sign Monitoring

Massine GANA
LARI Laboratory
Mouloud Mammeri University
Tizi-Ouzou, Algeria
massine.gana@ummto.dz

Zakia CHELLI
LCAGC Laboratory
Mouloud Mammeri University
Tizi-Ouzou, Algeria
zakia.chelli@ummto.dz

Rafik SADD AOUI
LAMPa Laboratory
Mouloud Mammeri University
Tizi-Ouzou, Algeria
rafik.saddaoui@ummto.dz

Sadjia BOUBOUZAL
Electronics department
Mouloud Mammeri University
Tizi-Ouzou, Algeria
sadjiaboubouzal0910@gmail.com

Abderezak ACHOUR
Electronics department
Mouloud Mammeri University
Tizi-Ouzou, Algeria
abderezak682@gmail.com

Mourad LAGHROUCHE
LAMPa Laboratory
Mouloud Mammeri University
Tizi-Ouzou, Algeria
mourad.laghrouche@ummto.dz

Abstract— The rapid growth of the Internet of Things (IoT) and the Internet of Medical Things (IoMT) has opened new opportunities for telemedicine in remote and low-connectivity areas. This study proposes the design and implementation of a long-range medical monitoring system based on LoRaWAN technology. The prototype integrates an ESP32 microcontroller and a MAX30102 sensor to monitor heart rate and SpO₂ levels, transmitting data to a remote healthcare platform. A radio link budget model and optimization of transmission parameters, such as Spreading Factor, Bandwidth, and Coding Rate, were applied to improve communication performance. Field experiments conducted in rural, dune-covered, and urban environments demonstrated reliable data transmission over distances of up to 7 km. These results confirm the suitability of LoRaWAN for remote biomedical data transmission in IoMT applications.

Keywords— IoMT, LoRaWAN, telemedicine, ESP32, optimization, MAX30102

I. INTRODUCTION

In recent years, the healthcare sector has undergone a significant transformation driven by the convergence of technological innovation and medical necessity. The Internet of Medical Things (IoMT) is a specialized branch of the Internet of Things (IoT) that is central to this evolution [1]. The IoMT interconnects biomedical sensors, smart devices, and remote monitoring platforms to acquire, transmit, and analyze physiological data. Recent studies have indicated a consistent growth in the implementation of IoMT (Internet of Medical Things) solutions, with an estimated increase of approximately 18% in 2022 and 16% in 2023 [2]. This growth is indicative of an escalating adoption of connected healthcare solutions, despite prevailing concerns regarding data privacy and regulatory compliance [3].

This transition from hospital-centered care to a patient-centric, remote surveillance model has been expedited by two major global health crises and the rising prevalence of chronic diseases in ageing populations. Telemedicine, defined as the utilization of information and communication technologies (ICT) to deliver healthcare services remotely, has emerged as a pivotal catalyst for this transition. The system under discussion facilitates real-time interaction between patients and medical professionals, supports continuous health monitoring, and reduces the logistical and financial burdens associated with traditional care delivery.

However, the efficacy of telemedicine is contingent upon the existence of robust, energy-efficient, and long-range communication infrastructures, particularly in rural or

underserved regions where conventional networks are either sparse or unreliable. A significant number of extant systems are characterized by limited wireless range, high power consumption, and inadequate scalability. To address these limitations, Low-Power Wide-Area Networks (LPWANs), specifically LoRaWAN (Long Range Wide Area Network), offer a compelling alternative. The LoRaWAN specification facilitates low-throughput, long-distance data transmission with minimal energy demand, rendering it optimally suited for deployment in remote biomedical sensor nodes and IoMT applications [4].

The present paper sets out the design, optimization, and evaluation of a novel LoRaWAN-enabled Internet of Medical Things (IoMT) architecture, integrating an ESP32 microcontroller and a MAX30102 optical biosensor for measuring heart rate and oxygen saturation (SpO₂). The study goes beyond mere hardware integration, encompassing a mathematical analysis of the wireless link budget and parametric optimization of pivotal radio factors, namely spreading factor (SF), bandwidth (BW), and coding rate (CR). A series of field experiments was conducted across a range of terrains to assess the packet delivery ratio. The results of these experiments demonstrated that the ratio exceeded 90% over distances of up to 7 km, while exhibiting minimal energy consumption per measurement cycle. The integration of system-level design with performance optimization and ethical considerations positions LoRaWAN-based IoMT systems as viable, scalable, and cost-effective solutions for telemedicine, thereby advancing the vision of accessible healthcare in infrastructure-limited environments.

II. STATE OF THE ART AND THEORETICAL FRAMEWORK

A. Internet of Medical Things (IoMT)

The Internet of Medical Things (IoMT) refers to the networked system of medical devices, wearables, sensors, gateways, and backend infrastructure (cloud/fog/edge) that enable continuous monitoring, data transmission, and healthcare services outside traditional clinical settings [5]. According to recent surveys, typical IoMT architectures follow a layered model:

- sensing layer (wearables or fixed sensors measuring e.g., heart rate, SpO₂, blood pressure)
- gateway/fog layer (local aggregation, preprocessing, brief storage)
- cloud/service layer (data analytics, dashboards, remote clinician access).

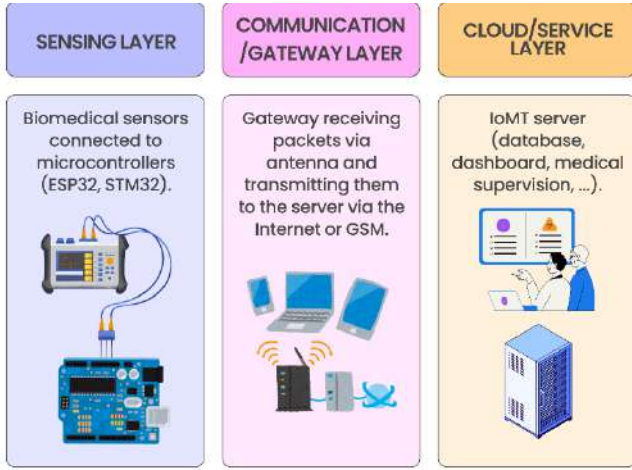


Figure 1. Typical layered architecture of an IoMT system showing sensor, gateway, and cloud layers.

In such systems, critical challenges persist, including interoperability due to heterogeneous device manufacturers, protocols, and data formats. This can result in a platform deploying devices that cannot easily exchange or integrate data [6]. Confidentiality and privacy of patient data are another major concern: connected medical sensors create attack surfaces, and the literature emphasizes the risk of eavesdropping, unauthorized access, or data manipulation. Finally, robustness (in the sense of reliability, fault-tolerance, long-term deployment autonomy, and resilience to network conditions) is crucial in medical applications: signal drops, battery depletion, or connectivity loss can compromise patient safety [7].

B. Communication Technologies for Connected Health

In the context of health-connected systems, a range of wireless communications technologies exists and must be compared in terms of range, power consumption, infrastructure cost, and suitability for IoMT. For example, the following wireless technologies are considered:

- Wi-Fi: high throughput, moderate range, but high-power consumption and infrastructure demands.
- Zigbee: short-range mesh, low power but limited range.
- BLE: Bluetooth Low Energy, very short range, very low power
- NB-IoT: licensed-band cellular LPWAN, and
- LoRaWAN: unlicensed band LPWAN.

Among these, LoRaWAN (Long Range Wide Area Network) is distinguished within the LPWAN category. The system utilizes a star-of-stars topology, whereby end devices communicate with gateways, which subsequently relay information to network servers. The protocol under discussion delineates device classes A, B, and C, and offers adjustable parameters such as spreading factor (SF), bandwidth (BW), and coding rate (CR). In addition, it supports adaptive data-rate (ADR) mechanisms [8]. Advantages of LoRaWAN for IoMT include: very low power consumption, enabling battery-operated devices to last long periods, long communication range, and the ability to deploy private networks in unlicensed frequency bands, reducing recurring costs. Limitations, however, must be recognized: data rate is modest, latency and downlink capability are constrained, especially for Class A devices, and, being in

unlicensed bands, the technology may be more sensitive to interference and duty-cycle restrictions.

Technology	Range (m)	Data Rate	Power Consumption
Wi-Fi	50–100	up to 100 Mbps	High
BLE	10–30	< 2 Mbps	Very low
Zigbee	10–100	250 kbps	Low
NB-IoT	1–10 km	< 200 kbps	Medium
LoRaWAN	2–15 km	0.3–50 kbps	Very low

Figure 2. Comparative analysis of the main wireless technologies used in IoMT applications.

C. Recent Works

Recent literature (2021-2024) has placed emphasis on the integration of IoMT and LPWAN, the performance optimization of LoRaWAN, and system-level studies. For instance, a 2024 survey on IoMT highlighted evolving architectures and identified gaps in interoperability and scalability [9]. A further study was conducted, which specifically reviewed IoMT communication security, with the conclusion of said study pointing to the necessity for protocol-level enhancements [10]. On the LoRaWAN side, comparative reviews (2022) highlight that LoRa/LoRaWAN and Sigfox lead in terms of device lifetime and cost-effectiveness, whereas NB-IoT excels in latency/QoS but at higher power cost [8]. Research conducted on the scalability of LoRaWAN and the selection of parameters (SF, ADR, CR) has demonstrated that the dynamic adjustment of these parameters can enhance network performance in both dense and variable environments [11]. Collectively, these studies emphasize that while LoRaWAN shows great promise for IoMT (particularly in remote or low-infrastructure settings), achieving high performance necessitates meticulous system design: protocol parameter tuning, energy management, reliability mechanisms, and data-handling strategies.

III. PROPOSED SYSTEM DESIGN

A. General Architecture

The proposed IoMT system aims to continuously monitor two vital physiological parameters — oxygen saturation (SpO₂) and heart rate (HR) — and to transmit these data wirelessly to a remote monitoring station via a long-range, low-power LoRaWAN link. The complete communication chain consists of five functional blocks:

1. Biomedical sensing unit: the MAX30102 optical sensor captures infrared (IR) and red-light signals from the fingertip.
2. Processing unit: an ESP32 microcontroller performs local preprocessing, filtering, and calculation of physiological metrics.
3. Communication unit: a LoRa Ra-01 (SX1278) module transmits the processed data via LoRaWAN Class A at 433 MHz.
4. Gateway node: a receiver equipped with a corresponding LoRa transceiver connected to a computer or server application.
5. Monitoring terminal: used to visualize the transmitted data.

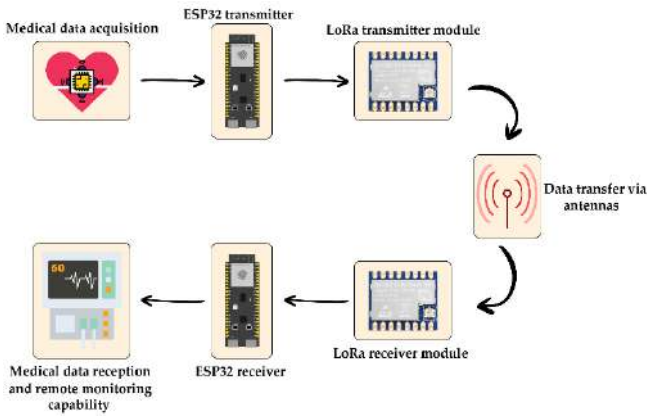


Figure 3. Functional block diagram of the system

A functional block diagram, represented in Figure 3, illustrates this architecture. The network operates under LoRaWAN Class A, which is optimized for ultra-low-power end devices that primarily transmit data uplink and only listen briefly for acknowledgments, ensuring minimal duty-cycle utilization and maximum autonomy.

B. Hardware and Software Component

1) *ESP32 Microcontroller*: The ESP32 is a dual-core 32-bit microcontroller with integrated Wi-Fi and Bluetooth capabilities, clocked up to 240 MHz, and equipped with multiple ADC channels and UART/SPI/I²C interfaces. In this design, the ESP32 serves as the central controller, responsible for data acquisition, local signal processing, and communication management with the LoRa module. Its deep-sleep mode (down to 10 μ A) is leveraged to reduce energy consumption between measurements.

2) *MAX30102 Optical Sensor*: The MAX30102 integrates two LEDs (IR and Red) and a photodetector to measure photoplethysmographic (PPG) signals. These raw optical signals are digitized through an internal 18-bit ADC and transmitted to the ESP32 via I²C. The sensor allows precise estimation of both heart rate (via periodic waveform analysis) and oxygen saturation (via ratio-of-ratios computation of the IR/Red channels).

3) *LoRa Ra-01 (SX1278) Transceiver*: The Ra-01 module is based on the Semtech SX1278 chipset, operating at 433 MHz with a maximum transmit power of 20 dBm and receiver sensitivity down to -139 dBm. The module supports configurable parameters such as spreading factor (SF7–SF12), bandwidth (BW = 125/250/500 kHz), and coding rate (CR = 4/5–4/8). These parameters are tuned using the mathematical optimization procedure detailed in Section 4 to maximize communication reliability and minimize energy per bit.

4) *Software Environment*: The firmware is developed using Arduino IDE, combining the following open-source libraries:

- LoRa.h – for SX1278 communication via SPI interface.
- Wire.h – for I²C communication with the MAX30102 sensor.
- MAX3010x.h – for data acquisition and calibration of PPG signals.

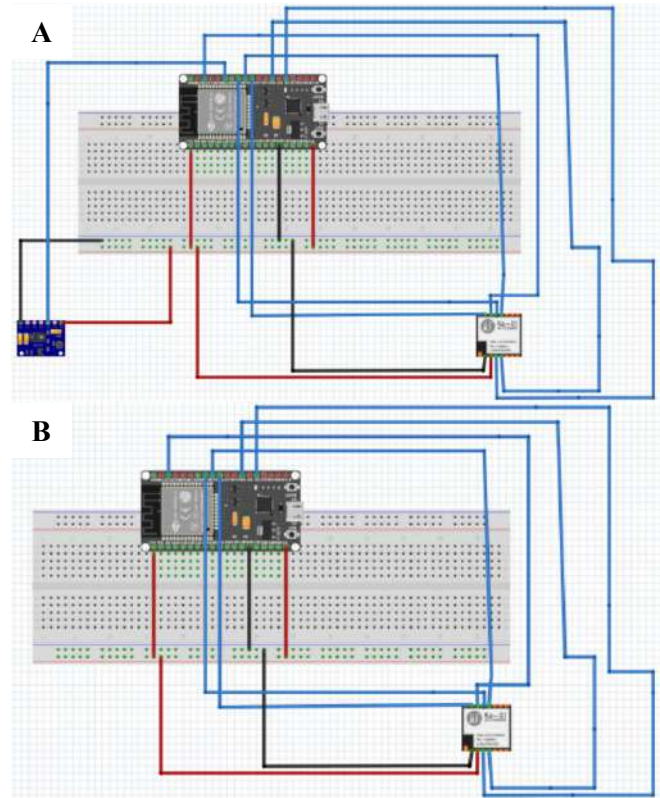


Figure 4. Connection of system components: A – Transmission, B – Reception.

C. LoRa Parameter Optimization

The LoRaWAN protocol offers several configurable physical-layer parameters (Spreading Factor (SF), Bandwidth (BW), Coding Rate (CR), and Transmit Power (TP)), which directly influence the transmission range, reliability, and energy consumption of the system. Optimizing these parameters is therefore essential to balance the trade-off between long communication distance and low energy use.

- **Spreading Factor (SF)**: Determines the number of chirps used to encode each symbol. Higher SF increases the signal's robustness to noise but also increases the time required to transmit a packet.
- **Bandwidth (BW)**: Defines the spectral width used for transmission. A larger BW shortens symbol duration (faster transmission) but reduces receiver sensitivity.
- **Coding Rate (CR)**: Defines the level of forward error correction (FEC). Expressed as:

$$\frac{4}{(4+n)} \quad (1)$$

where $n = 1$ to 4.

Higher CR improves reliability but adds redundancy, increasing transmission time.

- **Transmit Power (TP)**: Adjustable between 2 dBm and 20 dBm. Increasing TP enhances coverage but raises instantaneous current draw.

Each combination of these parameters determines the Time-on-Air (ToA), the total duration that the LoRa transceiver remains active for a single packet. The total ToA can be expressed as:

$$ToA = T_{preamble} + N_{payload} \cdot T_{sym} \quad (2)$$

where:

- The Duration of a symbol T_{sym} is given by:

$$T_{sym} = \frac{2^{SF}}{BW} \quad (3)$$

- The Length of the preamble (simplified):

$$T_{preamble} = (N_{preamble} + 4.25) T_{sym} \quad (4)$$

The constant 4.25 in the preamble duration expression comes from the Semtech LoRa modem implementation, which automatically appends four synchronization symbols and a quarter-symbol transition at the end of the preamble.

- The number of payload symbols $N_{payload}$ was computed using the standard LoRa formula provided by Semtech. This expression determines how many LoRa symbols are required to encode the application data (40 bytes in our case), taking into account the spreading factor (SF), the coding rate (CR), the presence of a CRC, and the header mode. In simplified form, $N_{payload}$ is obtained by:

$$N_{payload} = 8 + \left\lceil \frac{8 PL - 4 SF + 28 + 16 CRC}{4 (SF - 2 DE)} \right\rceil \times (CR + 4) \quad (5)$$

where PL is the payload size in bytes, $CRC = 1$ when the checksum is enabled, and DE represents the low-data-rate optimization (activated only for high SF values). For a 40-byte packet and the configuration used in this work ($BW = 125$ kHz, $CR = 4/5$, explicit header), this formula gives values of $N_{payload}$ ranging from 68 symbols at SF7 to 48 symbols at SF12. These values are then used directly in the Time-on-Air computation

By applying this procedure for SF7 to SF12, we obtained the theoretical ToA values reported in the table.

TABLE I. VALUES OBTAINED FROM THE TOA ACCORDING TO THE SF

SF	T_{sym}	$N_{payload}$	$T_{preamble}$	ToA (ms)
7	1.024	68	12.544	82.176
8	2.048	63	25.088	154.112
9	4.096	58	50.176	287.744
10	8.192	53	100.352	534.528
11	16.384	53	200.704	1069.056
12	32.768	48	401.408	1974.272

IV. RESULTS AND DISCUSSION

The experimental evaluation was conducted in two distinct environments to assess the impact of propagation conditions on transmission range: a rural open-field area and a dense urban zone. The findings unequivocally substantiate the substantial impact of environmental characteristics, topography, and obstacles on communication performance.

A. Performance in Rural Open-Field Environment

The initial series of measurements was conducted in a rural area situated in Ait Raouna, within the municipal boundaries of Iflissen. The region is distinguished by a scarcity of anthropogenic infrastructure and a paucity of electromagnetic

interference. In the initial configuration, the transmitter and receiver were situated at altitudes of 200 m and 400 m, respectively, with a clear line-of-sight (LOS) between the two points. In such favorable conditions, the communication range was observed to reach a maximum of 7.0 km. This result underscores the efficacy of signal propagation in free-space conditions, where diffraction, absorption, and reflection losses are minimal.



Figure 5. Map illustrates data transmission over a distance of 7 km with a clear LOS.

In the second rural configuration in the same municipality near the first village, although the altitudes were increased to 540 m and 580 m, respectively, a dune located approximately 800 meters from the transmission midpoint partially obstructed the LoS. Consequently, the maximum transmission distance decreased to 6.0 km. This reduction illustrates the sensitivity of radio propagation to physical obstructions, which disrupt the Fresnel zone and induce diffraction losses. Despite the higher elevation of both points, the presence of a single topographical obstacle proved sufficient to degrade the link quality, emphasizing that unobstructed LoS is more critical than altitude alone for maintaining long-range communication.



Figure 6. Map illustrates data transmission over a distance of 6 km with a clear LOS between Iflissen and Azra.

B. Performance in Urban Environment

The second test was conducted in a densely populated urban environment, where the two nodes were initially separated by a distance of 455 meters. In this scenario, the signal propagation occurred through multiple buildings and administrative structures, which introduced significant attenuation and multipath effects. The maximum achievable range in this environment was found to be limited to 1.6 km.

This considerable reduction in comparison to the rural case can be attributed to several urban propagation impairments, including:

- The phenomenon of absorption and penetration losses, as they are known, is a consequence of the presence of concrete and reinforced structures.
- The phenomenon of signal reflection and subsequent scattering can result in the occurrence of multipath interference.
- The presence of elevated electromagnetic noise levels has been observed in the vicinity of communication systems and electrical infrastructure.

The combined effect of these factors is rapid signal degradation, which in turn reduces both coverage area and link reliability.

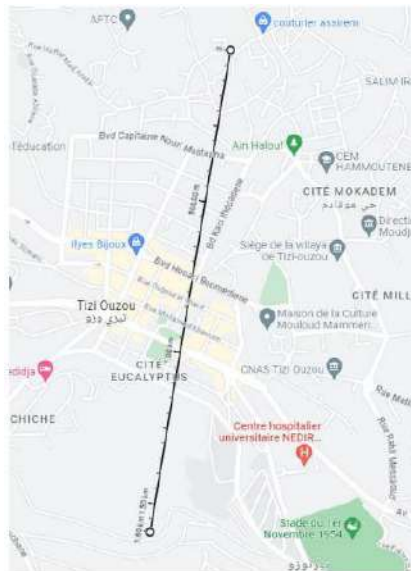


Figure 7. Map illustrates data transmission over a distance of 959m with a clear LOS at Tizi Ouzou City.

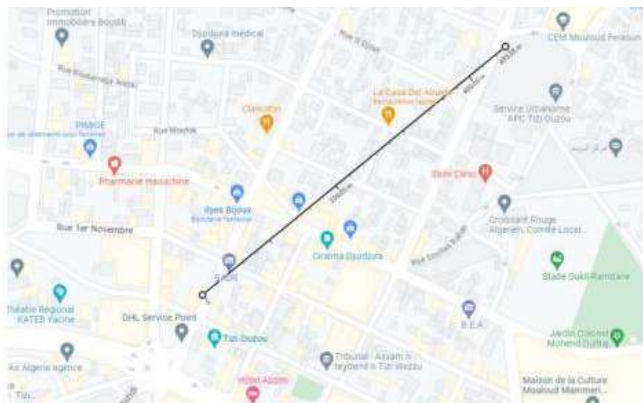


Figure 8. Map illustrates data transmission over a distance of 455m with a clear LOS, characterized by significant interference.

The findings of this study emphasize the importance of environment-aware deployment strategies for wireless communication systems. In the context of rural deployments, the importance of ensuring a clear (LOS) and optimal antenna alignment in enhancing transmission range has been well-documented. Conversely, in urban deployments, additional measures must be implemented to compensate for severe signal attenuation and ensure reliable connectivity. These measures include the use of relay nodes, repeaters, or mesh network architectures.

V. CONCLUSION

The results of this study confirm that biomedical data transmission in an IoMT context is strongly influenced by the propagation environment. In rural areas with clear line-of-sight, transmission distances up to 7 km were achieved, demonstrating the feasibility of long-range medical telemonitoring. In contrast, urban environments significantly reduced the range to 455m due to physical obstacles and electromagnetic interference. These findings highlight the need for careful deployment planning that considers terrain and urban density. The use of relay nodes and robust communication architectures is therefore essential to ensure continuous, and secure transmission of biomedical data.

REFERENCES

- [1] M. Mukhopadhyay, S. Banerjee, et C. Das Mukhopadhyay, « Internet of Medical Things and the Evolution of Healthcare 4.0: Exploring Recent Trends », *J. Electron. Electromed. Eng. Med. Inform.*, vol. 6, n° 2, p. 182-195, avr. 2024, doi: 10.35882/jeeemi.v6i2.402.
- [2] Knud Lasse Lueth *et al.*, « Number of connected IoT devices 2023 ». Consulté le: 21 novembre 2025. [En ligne]. Disponible sur: <https://iot-analytics.com/number-connected-iot-devices/>
- [3] G. R. Pradyumna, R. B. Hegde, K. B. Bommegowda, T. Jan, et G. R. Naik, « Empowering Healthcare With IoMT: Evolution, Machine Learning Integration, Security, and Interoperability Challenges », *IEEE Access*, vol. 12, p. 20603-20623, 2024, doi: 10.1109/ACCESS.2024.3362239.
- [4] D. M. Mathkor *et al.*, « Multirole of the internet of medical things (IoMT) in biomedical systems for managing smart healthcare systems: An overview of current and future innovative trends », *J. Infect. Public Health*, vol. 17, n° 4, p. 559-572, avr. 2024, doi: 10.1016/j.jiph.2024.01.013.
- [5] P. Matthew *et al.*, « A Review of the State of the Art for the Internet of Medical Things », *Sci*, vol. 7, n° 2, p. 36, mars 2025, doi: 10.3390/sci7020036.
- [6] J. N. S. Rubi et P. R. L. Gondim, « IoMT Platform for Pervasive Healthcare Data Aggregation, Processing, and Sharing Based on OneM2M and OpenEHR », *Sensors*, vol. 19, n° 19, p. 4283, oct. 2019, doi: 10.3390/s19194283.
- [7] S. E. El-deep, A. A. Abohany, K. M. Sallam, et A. A. A. El-Mageed, « A comprehensive survey on impact of applying various technologies on the internet of medical things », *Artif. Intell. Rev.*, vol. 58, n° 3, p. 86, janv. 2025, doi: 10.1007/s10462-024-11063-z.
- [8] M. A. M. Almuhaay, W. A. Jabbar, N. Sulaiman, et S. Abdulmalek, « A Survey on LoRaWAN Technology: Recent Trends, Opportunities, Simulation Tools and Future Directions », *Electronics*, vol. 11, n° 1, p. 164, janv. 2022, doi: 10.3390/electronics11010164.
- [9] B. Alturki *et al.*, « IoMT landscape: navigating current challenges and pioneering future research trends », *Discov. Appl. Sci.*, vol. 7, n° 1, p. 26, déc. 2024, doi: 10.1007/s42452-024-06351-w.
- [10] D. Koutras, G. Stergiopoulos, T. Dasaklis, P. Kotzanikolaou, D. Glynos, et C. Douligieris, « Security in IoMT Communications: A Survey », *Sensors*, vol. 20, n° 17, p. 4828, août 2020, doi: 10.3390/s20174828.
- [11] M. Jouhari, N. Saeed, M.-S. Alouini, et E. M. Amhoud, « A Survey on Scalable LoRaWAN for Massive IoT: Recent Advances, Potentials, and Challenges », *IEEE Commun. Surv. Tutor.*, vol. 25, n° 3, p. 1841-1876, 2023, doi: 10.1109/COMST.2023.3274934.

Solving Peer-to-Peer energy trading using Minimum Cost Flow in smart grids

Assala Nacef

LRSD laboratory, Faculty of Sciences
University Ferhat Abbas setif 1
setif, Algeria
assala.nacef@univ-setif.dz

Lina Benchikh

LRSD laboratory, Faculty of Sciences
University Ferhat Abbas setif 1
setif, Algeria
lina.benchikh@univ-setif.dz

Ouissem Touameur

LRSD laboratory, Faculty of Sciences
University Ferhat Abbas setif 1
setif, Algeria
ouissem.touameur@univ-setif.dz

Djamila Mechta

LRSD laboratory, Faculty of Sciences
University Ferhat Abbas setif 1
setif, Algeria
mechtadjamila@univ-setif.dz

Djihad Arrar

LRSD laboratory, Faculty of Sciences
University Ferhat Abbas setif 1
setif, Algeria
djihad.arrar@univ-setif.dz

Abstract—The growing adoption of distributed renewable energy has transformed conventional power systems into decentralized smart grids. Peer-to-peer (P2P) energy trading has emerged as a promising solution to improve local energy usage, but many existing methods overlook fairness, network congestion, and realistic grid integration. This paper presents a fairness-aware P2P energy trading framework based on minimum-cost flow (MCF) optimization with hop-based routing. A max–min fairness preprocessing stage is introduced to ensure equitable energy distribution among consumers before routing. Simulation results on a 17-node smart grid demonstrate that the framework achieves fair resource allocation, respects physical network constraints, minimizes routing distance, and integrates grid support when local generation is insufficient, offering a scalable solution for future decentralized energy markets.

Index Terms—Peer-to-Peer Energy Trading, Smart Grid, Minimum-Cost Flow, Max–Min Fairness, Power Flow Optimization

I. INTRODUCTION

The rapid integration of distributed renewable energy sources such as rooftop photovoltaic systems and small-scale wind generators has significantly transformed the structure of modern power systems (1). Traditional centralized grids are progressively evolving toward decentralized smart grids in which end-users can act as both producers and consumers of electricity (2). This paradigm shift has enabled the emergence of peer-to-peer (P2P) energy trading, where prosumers directly exchange energy within local communities (3).

Despite its potential, existing P2P trading mechanisms often face critical limitations (4). Many approaches neglect the physical constraints of the underlying electrical network, leading to unrealistic power flow assumptions (5). Moreover, fairness in energy allocation does not get enough attention, which affects the quality of the matching results (6), (7). Additionally, several works overlook the realistic integration of the utility grid when local renewable production is insufficient to meet demand (8).

This paper addresses these challenges by proposing a fairness-aware P2P energy trading framework based on minimum-cost flow (MCF) optimization. The main contributions of this work are summarized as follows:

- A graph-based modeling of the smart grid using split-node techniques to enforce router and line capacity constraints.
- A max–min fairness allocation mechanism that guarantees equitable distribution of renewable energy among consumers.
- A hop-based MCF routing formulation solved using the Network Simplex algorithm.
- A residual-flow decomposition algorithm that reconstructs explicit producer–consumer energy trading pairs.
- Seamless integration of a utility grid node to ensure supply-demand balance under realistic operating conditions.

The remainder of this paper is organized as follows. Section II describes the system modeling and problem formulation. Section III presents the proposed methodology. Section IV discusses the simulation setup and results. Finally, Section V concludes the paper and outlines future research directions.

II. RELATED WORKS

(9) proposed a blockchain-based smart grid framework for peer-to-peer (P2P) energy trading and automated billing using smart contracts. Their system improved transparency and security in decentralized energy markets and demonstrated high billing accuracy using real industrial data. The study highlighted challenges related to scalability and transaction latency while showing the potential of blockchain for future smart grid applications.

Author in (10) presented a decentralized P2P energy trading framework that incorporates network losses and grid usage fees. Their method used distance-based pricing and iterative

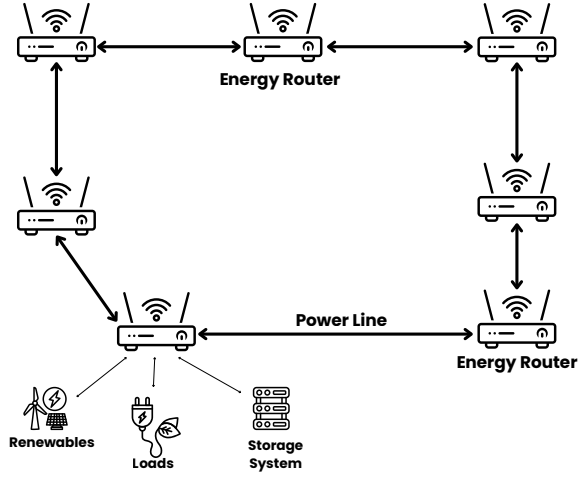


Fig. 1. Smart Grid Architecture

coordination to achieve near-optimal, scalable power sharing without a central controller.

Paper (11) proposes a real-time transaction-based minimum loss routing (MLR) framework for interconnected microgrids where producers and consumers are matched dynamically through a bidding and matchmaking mechanism. Sellers and buyers submit price, quantity, and time-window bids, and a matching process pairs them based on price compatibility and transmission feasibility. Once matched, energy is routed using a modified Dijkstra-based minimum-loss path algorithm, ensuring congestion-free delivery while respecting router and line capacity constraints. The work highlights how explicit producer-consumer matching combined with routing-aware optimization improves transaction efficiency and reduces transmission losses in energy internet architectures.

(12) proposes a real-time peer-to-peer energy trading framework that addresses the producer subset matching problem using a hybrid Quantum Genetic Algorithm (QGA) and a greedy path-selection strategy. The core matching mechanism selects an optimal subset of producers for each consumer based on energy price and transmission loss, rather than simple proximity or capacity. Matching is performed dynamically by first computing minimum-loss paths and then allocating demand across multiple producers through a capacity-aware fallback algorithm when direct paths are infeasible. Compared to traditional price-only or greedy matching methods, the approach improves matching efficiency under congestion and capacity constraints.

III. NETWORK REPRESENTATION

The electrical distribution system is modeled as a directed and capacitated graph:

$$G = (V, E) \quad (1)$$

where V represents energy routing nodes and E represents power transmission lines as shown in Figure 1. To enforce router capacity constraints, each node $v \in V$ is transformed

using the standard split-node technique. Two virtual nodes are created:

$$v^{in}, v^{out} \quad (2)$$

connected by an internal edge:

$$(v^{in}, v^{out}) \in E, \quad \text{capacity } C_v \quad (3)$$

This modeling guarantees that the power passing through each router never exceeds its physical limit.

A. Prosumer Power Modeling

Each node operates as a *prosumer*, meaning it can either produce or consume electricity. This behavior is represented by a net power value:

$$P_i = \begin{cases} > 0, & \text{if producer} \\ < 0, & \text{if consumer} \end{cases} \quad (4)$$

Two different demand profiles are maintained:

- P^{real} : the actual physical production or demand values
- P^{fair} : adjusted demand values obtained after fairness processing

The real demand profile is preserved to calculate residual demand later served by the utility grid.

IV. METHODOLOGY

This section presents a fairness-aware peer-to-peer (P2P) energy trading framework designed for smart grid environments. The proposed approach combines max-min fairness for equitable energy allocation with minimum-cost flow (MCF) optimization for efficient routing, and a flow decomposition mechanism for explicit producer-consumer matching.

A. Fairness Preprocessing (Max-Min Allocation)

Before routing is performed, a fairness mechanism is applied to prevent large consumers from monopolizing limited renewable energy. The approach used is *max-min fairness*, in which available generation is distributed in small, equal shares across all active consumers.

During each allocation round, the maximum allowable share per consumer is calculated as shown in Eq. (5):

$$\text{share} = \frac{S_{\text{remaining}}}{|C_{\text{active}}|} \quad (5)$$

where $S_{\text{remaining}}$ is the amount of unallocated energy still available and $|C_{\text{active}}|$ is the number of consumers still requiring energy.

Consumers that reach their full demand are removed from subsequent rounds. This process continues until supply is exhausted or all consumers are satisfied, producing the adjusted demand vector P^{fair} .

A detailed step-by-step implementation of this fairness mechanism is provided in Algorithm 1.

Algorithm 1 Max-Min Fair Energy Allocation

Require: Consumer set C , demand D_c , total supply S
Ensure: Fair allocation A_c

```

1: for all  $c \in C$  do
2:    $A_c \leftarrow 0$ 
3: end for
4:  $A \leftarrow C$ 
5: while  $S > 0$  and  $A \neq \emptyset$  do
6:    $share \leftarrow \frac{S}{|A|}$ 
7:   for all  $c \in A$  do
8:      $need \leftarrow D_c - A_c$ 
9:      $give \leftarrow \min(need, share)$ 
10:     $A_c \leftarrow A_c + give$ 
11:     $S \leftarrow S - give$ 
12:   end for
13:   for all  $c \in A$  do
14:     if  $A_c = D_c$  then
15:       Remove  $c$  from  $A$ 
16:     end if
17:   end for
18: end while
19: return  $A_c$ 

```

B. Minimum-Cost Flow Routing

After fairness is enforced, routing decisions are computed using a minimum-cost flow (MCF) formulation. The objective is to minimize the total routing cost:

$$\min \sum_{e \in E} f_e w_e \quad (6)$$

subject to:

$$0 \leq f_e \leq cap_e \quad (7)$$

$$\sum_{e \in \delta^+(v)} f_e - \sum_{e \in \delta^-(v)} f_e = d_v \quad (8)$$

where f_e is the flow on edge e , cap_e is its capacity, w_e is the edge cost, and d_v is the node demand.

The cost model is defined as:

- Internal vertex edges: cost = 0
- Physical transmission lines: cost = 1 (hop-based routing)
- Utility grid connections: high penalty cost

The flow optimization is solved using the Network Simplex algorithm.

C. Utility Grid Support

A virtual grid node G_{grid} is added to represent the connection to the main power utility. During the MCF optimization, the grid maintains a neutral balance to guarantee feasibility.

After routing, any unmet real demand is automatically served by the grid according to:

$$P_i^{\text{grid}} = \max(0, D_i^{\text{real}} - D_i^{\text{p2p}}) \quad (9)$$

This mechanism ensures that physical load balance is always maintained under realistic operating conditions.

D. Flow Decomposition and Matching

The optimized flows produced by the MCF solver are aggregated and then decomposed into explicit bilateral trades. This is performed by constructing a residual flow graph and tracing feasible paths starting from producer nodes.

Each trace terminates either at a consumer node or at the utility grid, producing explicit matching pairs (p, c) . The matching logic is defined in Algorithm 2.

Algorithm 2 Producer-Consumer Flow Matching

Require: Flow solution F , producer set P , consumer set C

Ensure: Producer-consumer matching

```

1: Build residual graph from positive flows
2: for all  $p \in P$  do
3:    $remaining \leftarrow production(p)$ 
4:   while  $remaining > 0$  do
5:     Find residual path to consumer or GRID
6:     Assign feasible flow along the path
7:     Update residual capacities
8:      $remaining \leftarrow remaining - assigned\_flow$ 
9:   end while
10: end for
11: for all  $c \in C$  do
12:   if received energy < real demand then
13:     Assign deficit from GRID
14:   end if
15: end for
16: return Matching results

```

E. Overall Framework

The complete system pipeline is summarized in Algorithm 3. This high-level algorithm links the fairness stage (Algorithm 1) with the routing optimization and matching stages (Algorithm 2).

Algorithm 3 Min-Max Fair Energy Trading Framework

Require: Network graph G , power data P_i

Ensure: Valid producer-consumer energy trades

```

1: Apply fairness using Algorithm 1
2: Build capacitated routing graph
3: Solve MCF optimization
4: Perform matching using Algorithm 2
5: return Final allocations and grid usage

```

V. RESULTS**A. Simulation Setup**

The proposed framework was evaluated on a 17-node distribution network 2 where each node represents a smart prosumer router with explicit capacity constraints as shown in Table I. Power lines were modeled as bidirectional edges with thermal

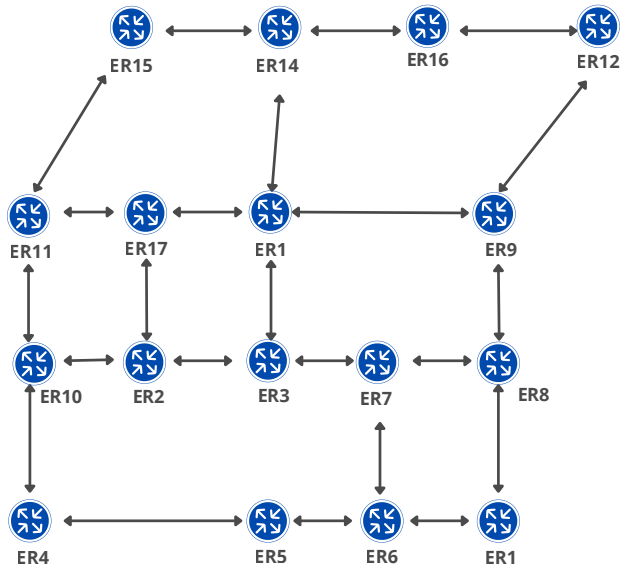


Fig. 2. 17-node distribution network

TABLE I
ENERGY ROUTER CHARACTERISTICS FOR 17-NODE NETWORK

ER	Efficiency	Capacity (kW)
1	1.00	20
2	0.98	15
3	1.00	25
4	1.00	24
5	0.98	22
6	0.97	20
7	0.98	25
8	0.97	30
9	1.00	30
10	0.97	20
11	1.00	27
12	0.98	25
13	1.00	30
14	0.98	19
15	1.00	18
16	0.97	18
17	0.98	20

limits as presented in Table II. The system included both producers and consumers with deliberately unbalanced generation and demand to evaluate the grid-integration behavior.

The experiments were implemented in Python using the NetworkX library and solved using the Network Simplex algorithm. All simulations were executed on a standard Kaggle CPU environment.

B. Economic Assumptions

To reflect realistic market behavior, fixed pricing assumptions were used during the simulations. The utility grid was modeled with a retail buying price of 0.30 €/kWh, representing the cost paid by consumers when purchasing energy from the main grid. A feed-in tariff of 0.10 €/kWh was assumed as the compensation received by prosumers when exporting excess energy to the grid.

TABLE II
POWER LINE CHARACTERISTICS FOR 17-NODE NETWORK

Line	Capacity (kW)	Resistance	Voltage (V)
L1-3	30	0.60	400
L1-9	45	0.45	400
L1-14	40	0.21	400
L1-17	30	0.24	400
L2-3	20	0.64	400
L2-5	20	0.51	400
L2-10	30	0.19	400
L2-17	30	0.19	400
L3-7	45	0.94	400
L4-5	24	0.19	400
L4-10	30	0.64	400
L5-6	7	0.45	400
L6-7	40	0.24	400
L6-13	30	0.21	400
L7-8	10	0.21	400
L8-9	32	0.65	400
L8-13	40	0.19	400
L9-12	32	0.45	400
L10-11	30	0.24	400
L11-15	32	0.60	400
L11-17	40	0.24	400
L12-16	40	0.21	400
L14-15	30	0.45	400
L14-16	35	0.19	400

TABLE III
MCF WITH FAIRNESS

	Matching pair			
	Producer	Consumer	traded energy	Cost
MCF with fairness	MG13	MG17	3	0.6
		MG10	5	1
	MG9	MG17	12	2.4
	Grid	MG17	5	1

For peer-to-peer (P2P) transactions, a uniform producer price of 0.20 €/kWh was applied across all decentralized trades. This pricing structure ensures that P2P energy exchange is economically preferred to grid-based trading, while still maintaining a realistic spread between retail and wholesale prices.

C. Fairness Evaluation

The Max-Min fairness mechanism was evaluated by comparing the original consumer demands with the allocated peer-to-peer (P2P) energy. The results of this comparison are highlighted in Table III and Table IV.

TABLE IV
MCF WITHOUT FAIRNESS

	Matching pair			
	Producer	Consumer	traded energy	Cost
MCF without fairness	MG13	MG17	8	1.6
	MG9	MG17	12	2.4
	Grid	MG10	5	1.5

TABLE V
SCENARIO 1 RESULTS: PRODUCTION < DEMAND (WITH FAIRNESS)

Scenario 1: Production < Demand				
Total Production	Total Demand	Producer	Consumer	Energy Cost
		MG13	MG17	3 0.6
			MG10	5 1.0
		MG9	MG17	12 2.4
		GRID	MG17	5 1.5

Table III shows that when fairness is enabled, smaller-demand consumers are prioritized and fully satisfied before allocating remaining energy to larger consumers. In contrast, Table IV demonstrates that without fairness, demand allocation becomes biased, with larger consumers receiving more energy while smaller consumers rely on the utility grid.

These results confirm that the proposed fairness layer successfully prevents monopolization of distributed energy.

D. Minimum-Cost Flow Performance

To evaluate the effectiveness of the routing and matching mechanisms, three distinct operational scenarios were simulated on the 17-node network:

- 1) Scenario 1: Total production is lower than total demand (Table V).
- 2) Scenario 2: Total production equals total demand (Table VI).
- 3) Scenario 3: Total production exceeds total demand (Table VII).

E. Scenario-Based Performance Analysis

Scenario 1 (Production < Demand). As shown in Table V, when local generation is insufficient, the system prioritizes fair P2P allocation while transparently relying on the utility grid to satisfy the remaining deficit. Producers MG9 and MG13 partially supply the consumers, while the grid covers the remaining unmet demand. This confirms that the model maintains feasibility under energy scarcity.

Scenario 2 (Production = Demand). Table VI illustrates that when total production matches total demand, the system achieves full P2P self-sufficiency. All consumer demands are satisfied entirely through peer-to-peer trades, and the grid remains unused. This scenario demonstrates the optimal operating point of the proposed framework.

Scenario 3 (Production > Demand). In Table VII, local production exceeds demand, leading to surplus energy. After satisfying all consumer demands through P2P trades, the excess generation is exported to the grid. This scenario validates the model's ability to manage overproduction while preserving system stability and economic efficiency.

The economic evaluation shown in Table V-E reveals that Scenario 1 incurs the highest consumer cost due to reliance on the utility grid for deficit energy. In Scenario 2, full peer-to-peer satisfaction is achieved and grid interaction is eliminated, resulting in balanced consumer and producer costs.

TABLE VI
SCENARIO 2 RESULTS: PRODUCTION = DEMAND (WITH FAIRNESS)

Scenario 2: Production = Demand				
Total Production	Total Demand	Producer	Consumer	Energy Cost
		MG13	MG17	8 1.6
			MG10	10 2.0
		MG9	MG17	12 2.4
		Grid	/	/ /

TABLE VII
SCENARIO 3 RESULTS: PRODUCTION > DEMAND (WITH FAIRNESS)

Scenario 3: Production > Demand				
Total Production	Total Demand	Producer	Consumer	Energy Cost
		MG9	MG17	10 2.0
			MG10	2 0.4
		MG13	MG10	8 1.6
		MG13	GRID	10 1.0

Scenario 3 demonstrates the system's capability to handle surplus generation, where excess producer energy is exported to the grid, generating additional revenue while maintaining stable consumer costs.

TABLE VIII
SCENARIO ECONOMIC RESULTS

Scenario	Consumer Cost (€)	Producer Revenue (€)
Scenario 1	MG17: 4.50	MG13: 1.60
	MG10: 1.00	MG9: 2.40
Scenario 2	MG17: 4.00	MG13: 3.60
	MG10: 2.00	MG9: 2.40
Scenario 3	MG17: 2.00	MG9: 2.40
	MG10: 2.00	MG13: 2.60

VI. DISCUSSION

The Minimum-Cost Flow (MCF) optimization successfully routed energy through the network while satisfying all router and line capacity constraints. The hop-based cost function effectively minimized routing distance while avoiding congestion.

The network consistently preferred P2P routing paths over utility-grid paths due to the higher penalty cost associated with grid usage. No capacity violations were observed during the simulations.

The flow decomposition algorithm successfully reconstructed bilateral energy trades from the aggregated flow solution. The matching process generated physically feasible producer-to-consumer allocations while respecting production limits.

A representative outcome is summarized below:

- Producer 9 supplied energy to both Consumer 10 and Consumer 17.
- Producer 13 supplied energy primarily to Consumer 17.

- No producer exceeded its physical generation capacity.

When total distributed generation was insufficient to satisfy system demand, the proposed model automatically allocated the deficit to the utility grid. This behavior ensured system feasibility and maintained power balance under realistic operating conditions.

In the evaluated scenario, the grid supplied the exact shortfall after P2P energy distribution, demonstrating seamless integration of centralized utility support into the decentralized trading architecture.

VII. CONCLUSION

This paper presented a fairness-aware peer-to-peer energy trading framework based on minimum-cost flow optimization for smart grid environments. By combining a max–min fairness preprocessing mechanism with hop-based minimum-cost flow routing, the proposed approach ensures equitable energy distribution while respecting physical network constraints.

Simulation results on a 17-node smart grid topology demonstrated that the framework effectively prevents resource monopolization, minimizes routing distance, alleviates network congestion, and guarantees feasible power balancing through seamless integration of the utility grid. The proposed residual-flow matching mechanism successfully reconstructed explicit producer–consumer trading relationships without violating generation or transmission limits.

While the current study focuses on static network conditions and fixed routing costs, future work will investigate the integration of dynamic pricing, real-time demand forecasting, battery storage systems, and machine learning–based adaptive trading strategies. The proposed framework provides a strong foundation for scalable, fair, and practical decentralized energy markets in next-generation smart grids.

REFERENCES

- [1] R. Panigrahi, S. Mishra, S. Srivastava, A. Srivastava, and N. Schulz, "Grid integration of small-scale photovoltaic systems in secondary distribution network—a review," *IEEE Transactions on Industry Applications*, vol. 56, pp. 3178–3195, 2020.
- [2] Y. Merrad, M. Habaebi, F. Toha, M. R. Islam, T. S. Gunawan, and M. Mesri, "Fully decentralized, cost-effective energy demand response management system with a smart contracts-based optimal power flow solution for smart grids," *Energies*, 2022.
- [3] A. Nacef, D. Mechta, L. Louail, and S. Harous, "Advancements in optimization strategies for energy routing, demand response, and load forecasting in energy internet and smart grid: an overview," *Energy efficiency*, vol. 18, no. 7, pp. 1–42, 2025.
- [4] Z. Sun, S. Tavakoli, K. Khalilpour, A. Voinov, and J. P. Marshall, "Barriers to peer-to-peer energy trading networks: A multi-dimensional pestle analysis," *Sustainability*, 2024.
- [5] N. Tarashandeh and A. Karimi, "Peer-to-peer energy trading under distribution network constraints with preserving independent nature of agents," *Applied Energy*, 2024.
- [6] M. Gasca, R. Rigo-Mariani, V. Debusschere, and Y. Sidqi, "Fairness in energy communities: Centralized and decentralized frameworks," *Renewable and Sustainable Energy Reviews*, 2025.
- [7] V. N. Behrunani, A. Irvine, G. Belgioioso, P. Heer, J. Lygeros, and F. Dorfler, "Designing fairness in autonomous peer-to-peer energy trading," *ArXiv*, vol. abs/2302.04771, 2023.
- [8] N. Uddin, Y. Wu, M. S. Islam, and K. M. Zakaria, "An optimized decentralized peer-to-peer energy trading system for smart grids incorporating uncertain renewable energy sources through fuzzy optimization," *Electric Power Systems Research*, 2025.
- [9] J. Din and H. Su, "Blockchain-enabled smart grids for optimized electrical billing and peer-to-peer energy trading," *Energies*, vol. 17, no. 22, 2024. [Online]. Available: <https://www.mdpi.com/1996-1073/17/22/5744>
- [10] A. Paudel, Y. Jiawei, and H. Gooi, "Peer-to-peer energy trading in smart grids considering network utilization fees," 08 2020, pp. 1–5.
- [11] H. Guo, F. Wang, L. Li, L. Zhang, and J. Luo, "A minimum loss routing algorithm based on real-time transaction in energy internet," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 12, pp. 6446–6456, 2019.
- [12] N. Assala, D. Mechta, L. Benchikh, L. Louail, S. Harous, and R. Belmahdi, "A quantum genetic-based routing protocol for real-time peer-to-peer energy transactions," *Cluster Computing*, vol. 28, 09 2025.

A Comprehensive Method for Steel Surface Classification Using MobileNet-V2 and Leaky ReLU Activation Function

Yousra KATEB

Department of Automation and Electrification
University of M'hamed Bougarra Bumerdes
Bumerdes, Algeria
y.kateb@univ-bumerdes.dz

Abdelmalek KHEBLI Department of
Automation and Electrification
University of Bumerdes. Bumerdes,
Algeria
a.khebli@univ-bumerdes.dz

Salah AGUIB

Dynamic Motors and Vibroacoustic
Laboratory, Faculty of Technology,
University of Bumerdes,
Algeria. salah.aguib@yahoo.fr

Abstract— Accurate detection of defects in steel production is critical to reducing failures and economic losses. Deep learning approaches typically require large labeled datasets, yet industrial environments often lack sufficient samples of steel surface defects. Addressing this challenge demands alternatives to conventional data augmentation. In this study, we employ the MobileNet model, chosen for its compact architecture and depthwise separable convolutions, which efficiently capture deep image features. Experimental results demonstrate that this approach achieves over 99.44% classification accuracy, confirming its effectiveness and potential to improve steel processing quality.

Keywords— *Steel Surface Defect Classification, Image Processing, Deep Learning, MobileNet-V2, Leaky ReLU*

I. INTRODUCTION

Defect classification is a crucial step in the steel manufacturing process in industrial production. Nevertheless, this process is typically carried out manually, with a heavy dependence on visual inspection to identify imperfections on the steel surface. As a result, the efficiency of production and the quality of products tend to be subpar. Ensuring the effectiveness of manual detection is challenging. Relying on visual inspection for extended periods can negatively impact human health. Identifying defects automatically to eliminate the need for manual work in industrial manufacturing is a crucial area of focus [1].

Several researches have been applied in the concept of steel inspection. DL techniques tend to be the state-of-art of defect classification in steel sheets.

One benefit of the classifier using deep learning is its ability to efficiently learn abstract features during training, eliminating the necessity to create specific artificial features. There is a clear improvement in the classification accuracy.

Convolutional neural networks are commonly utilized in surface classification for their numerous performance benefits. CNN requires guidance during the training process. The text highlights the features of local connection, weight sharing, etc.,

resulting in a significant reduction in the number of parameters and the complexity of model training. This model effectively captures local data features and remains stable even after undergoing different linear transformations of the image. Throughout this study, the deep network model plays a crucial role in convolving, applying nonlinear activation functions, and creating pooling layers to establish the mapping relationship between variables. The CNN model's structure consists of several convolutional layers and pooling layers with distinct functions.

In the world of image classification, the utilization of CNN in classification network models is prevalent. Some popular algorithms for classification include VGG [1], ResNet [2], AlexNet [3], DenseNet-121 [4].

II. RELATED WORKS

Within the industrial manufacturing setting, detecting defects on steel surfaces can be challenging due to slight differences, poor visibility, and excessive background noise. Traditional machine learning methods typically involve preprocessing collected images, extracting features, and using a classifier to identify various types of defects. The classifier must extract features from the image of steel surfaces. Commonly utilized features consist of LBP [5], Haar [6], HOG [7], multi-scale analysis, and more. For instance, S. R. Aghdam introduced a technique for identifying steel surface imperfections using decision trees [8]. In a recent study, S. Ghorai and colleagues implemented a three-level Haar feature to create an automated detection algorithm for identifying production defects on steel surfaces [9]. Shumin D. et al. present a conventional model utilizing the SVM method to identify fabric defects [10]. Despite the fact that these methods can reduce the amount of manual work required, the classifier still needs to create its own characteristics. Furthermore, the findings from using

conventional machine learning techniques still exhibit clear shortcomings.

The topic at hand delves into the utilization of deep learning technology for categorizing defects on metal surfaces. Several advanced deep learning models, including ResNet-50 [11], DARCNN [12], deep CNN [13], CASAE architecture that successfully detects metal defects under various conditions [14], NASNet-Mobile [15] and YOLO algorithm [16], have been suggested for this purpose. Yet, because of the limited number of labeled defect samples, these models frequently encounter issues like overfitting and a lack of resilience. To tackle this problem, data augmentation techniques are utilized to increase the size of the training dataset, and fine-tuning is applied to improve the model's effectiveness. Based on the findings, these approaches have been shown to significantly enhance the precision of categorizing steel surface irregularities in our method.

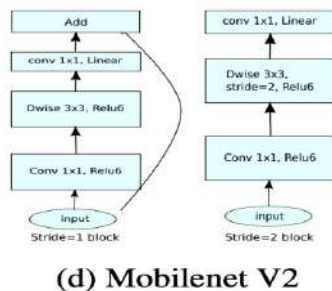
III. METHODOLOGY

A. MobileNet-V2

CNNs have been utilized in target classification tasks, with a focus on improving accuracy by increasing the depth and complexity of the models. Nevertheless, issues frequently arise when these models are implemented in actual production settings. These models need robust equipment and network support, which is often lacking in the industrial sector. Thus, certain network models modify the structure to enhance the model in terms of speed and complexity. MobileNet V2 is a lightweight deep neural network model that utilizes depthwise separable convolutions to efficiently extract data space and channel features.

In MobileNet V2, two global hyperparameters are mentioned: a width multiplier (α) and a resolution multiplier (ρ) to strike a balance between delay and precision. The fundamental component of MobileNet V2 is the depthwise separable convolution. Two key enhancements of MobileNet V2 include the inverted residual structure and linear bottlenecks.

The structure of the inverted residual involves connecting modules to the residuals. Initially, the projection convolution is employed to enhance the dimensionality, followed by the depth convolution, and ultimately, the projection convolution is utilized to decrease the dimensionality. The diagram in Fig. 1 illustrates the combination of inverted residual structure and linear bottlenecks.



(d) MobileNet V2

Fig. 1. Through the introduction of the inverted residual structure and the shortcut, MobileNet V2 addresses the problem of degradation with the growth

in the number of network layers. One part is the residual block with a step size of 1 (left), and the other part is the residual block with a step size of 2 (right).

B. Leaky ReLU

Leaky ReLU serves as an activation function in artificial neural networks to bring in nonlinearity among the outputs across different layers of a neural network. This function was developed to address the issue of neural networks dying during training when using the standard ReLU function.

Here is the mathematical formula of the Leaky ReLU function.

$$\text{LeakyReLU}(x) = \max(\alpha * x, x) \quad (1)$$

Looking at the equation (1), it is evident that the Leaky ReLU function outputs the higher value between x and $\alpha * x$, with α being a small positive constant. By utilizing this function, negative values can be adjusted to approach 0 without actually becoming 0. This resolves the problem of dying ReLU that occurs when the standard ReLU function is used in neural network training (see Fig. 2).

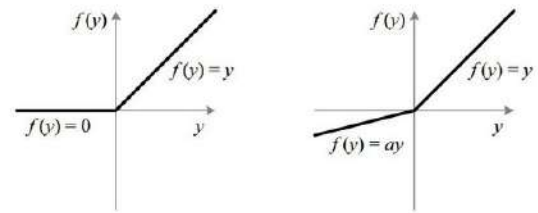


Fig. 2. ReLU (left) and Leaky ReLU activation function.

C. Dropout Layer

During model training, the use of Dropout can eliminate certain neural units, which helps simplify the training process. When working with various sets of input samples, the neural units that are eliminated vary randomly each time, resulting in the equivalent of training a distinct network for each batch. Using Dropout can help decrease the number of computational parameters needed for training and prevent overfitting. Fig. 3 illustrates the dropout principle. When updating the network parameters in the fully connected layer, random neurons are discarded to reduce the network complexity and parameter calculation.

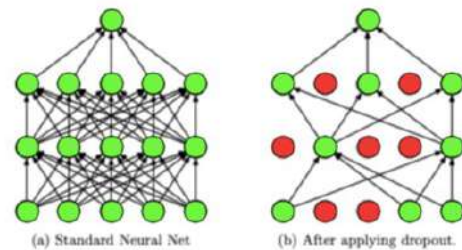


Fig. 3. Flow chart of dropout. (a) Standard neural network, (b) some neurons randomly deleted during dropout regularization

D. ADAMAX Optimizer

AdaMax is a version of the Adam optimization technique utilized in machine learning and deep learning to enhance the training of neural networks. The system keeps track of a moving average involving the gradient's first and second moments. However, unlike Adam, AdaMax utilizes the L-infinity norm of the gradients instead of the second moment. These improvements enhance stability when working with sparse gradients, leading to quicker convergence and reduced memory needs. AdaMax proves to be a robust optimization technique that significantly speeds up the training process of deep neural networks, particularly when dealing with sparse or high-variance gradients.

E. Motivation for Architecture Design

The framework we developed primarily addresses two practical challenges:

(1) Obtaining data sets of metal surface defects in the industrial field can be challenging, making training a complex task.

(2) When working in a manufacturing setting, the choice of classification algorithm becomes crucial due to the lack of powerful computing equipment available.

Throughout this study, we implemented Dropout techniques to enhance the network's training efficiency. When selecting the classification method, MobileNet V2 is chosen as the backbone network due to its superior speed and reliable classification outcomes. In the following section, it was discovered that replacing the activation function with Leaky ReLU can significantly enhance the model's performance.

We will demonstrate the potential of the suggested approach through experiments. Next, the section will emphasize the method's effectiveness and superiority through experiments.

IV. EXPERMENTS AND RESULTS

A. NEU-CLS dataset

The NEU-CLS dataset, provided by Northeastern University [17], contains images captured with a CCD camera on hot-rolled steel surfaces. Images were manually filtered and cropped to a uniform size of 200×200 pixels. It includes six typical defect types—scratches, rolled-in scale, patches, pitted surface, inclusion, and crazing—with 300 samples per class, totaling 1,800 images (Fig. 4). Challenges in this dataset include the visual similarity among defect types and variations due to illumination and material deformation.

For this study, all images were resized to $224 \times 224 \times 3$ for model input. Data augmentation—rescaling, shear (0.2), zoom (0.2), and horizontal flipping—was applied to improve generalization. The dataset was split 80:20 for training and validation. Training was performed using TensorFlow’s ImageDataGenerator with a batch size of 32 and the AdaMax optimizer on a GPU-enabled environment. Early stopping and model checkpoints were used to prevent overfitting and retain the best model.

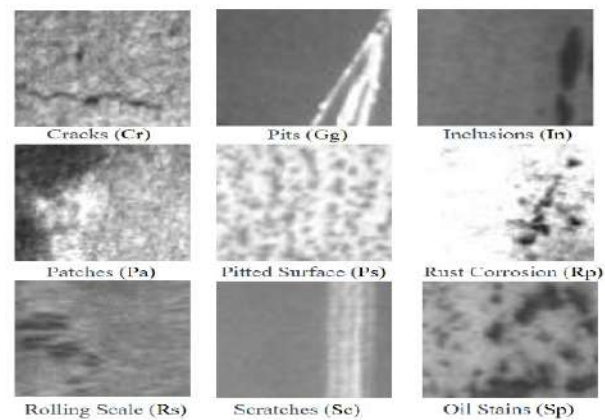


Fig. 4. Different types of defects in NEU-CLS images

B. Developed MobileNet-V2 for Steel Defect Classification

After developing the initial MobileNet-V2 model for steel surface defect classification, we offer several potential solutions to improve accuracy and reduce execution time. Initially, data augmentation is used to expand the range of features that the model is exposed to (Fig. 5). Furthermore, a fresh section has been incorporated into the model, including Global Average Pooling, Dropout, and Dense layers. These additional layers have been appended to the end of the model to improve accuracy, minimize model parameters, and reduce execution time. Third, ADAMAX was utilized as an optimizer. After implementing exponential decay for the learning rate as outlined in Equation 2, and categorical cross-entropy as the Loss function. We decided to stop training the model early if there is no further improvement in accuracy. The model successfully identified the best weights.

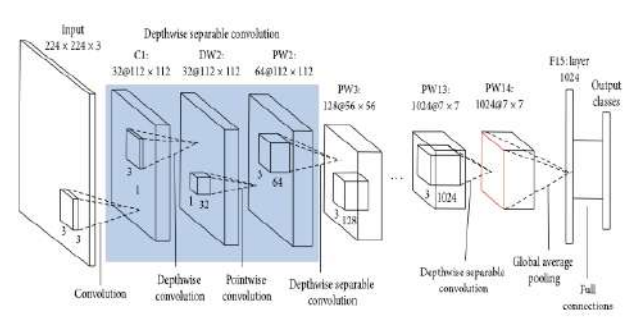


Fig. 5. MobileNet-V2 architecture

$$y=a(1-b)^x \quad (2)$$

Our approach relies on the Python framework available in Google Colaboratory [18]. The main libraries used in this implementation are Tesnorflow [19], Keras [20], Matplotlib, NumPy, and Glob. 80% of the images from the NEU-CLS [17] collection were utilized for training purposes, with 240 images allocated to each fault category, while the remaining 20% were reserved for validation. Performance is evaluated both prior to and following the network's fine-tuning. Table 2 presents the

values of the parameters utilized for training the CNN. The experiments were conducted using Windows 10 Professional on the Intel® Core(TM) -i5 7200U, 64-bit platform with 8GB RAM and NVIDIA RTX 2070, utilizing the available GPU on Google Colab Platform.

C. Results Discussion

The results we obtained will be discussed in this section. We used both Accuracy and loss to check the efficacy of our model. The results are given in the training as well as the validation dataset.

The loss reduced until it reaches 0.024 in the training and 0.047 in the validation dataset (Fig. 7).

The model achieved a training accuracy of 99.44% and a validation accuracy of 98.33%. While high accuracy demonstrates effective learning, the small size of the dataset (1800 images) raises concerns about overfitting. To address this, we evaluated the model using additional metrics. Table 1. presents the precision, recall, and F1-score for each defect type. Furthermore, a confusion matrix (Fig. 8) illustrates the distribution of correctly and incorrectly classified samples, confirming that the model performs consistently across all classes. The loss reduced until it reaches 0.024 in the training and 0.047 in the validation dataset (Fig. 7). The slight difference between training and validation accuracy indicates that overfitting is minimal, likely due to data augmentation and dropout regularization.

TABLE 1. PERFORMANCE OF THE MODEL

Proposed Algorithm	Training dataset	Validation dataset
Accuracy	99.44	98.33
Loss	0.024	0.047

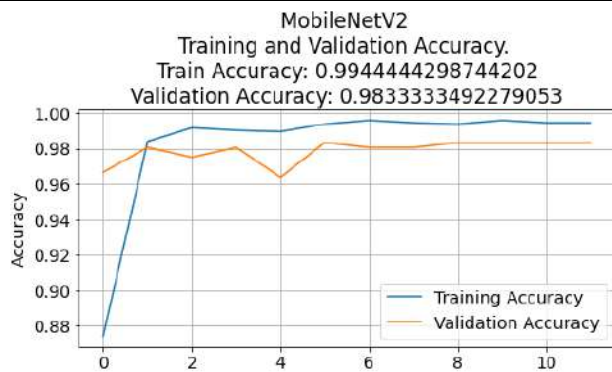


Fig. 6. Accuracy of the model for training (blue) and validation (orange) dataset.



Fig. 7. Loss of the model for training (blue) and validation (orange) dataset.

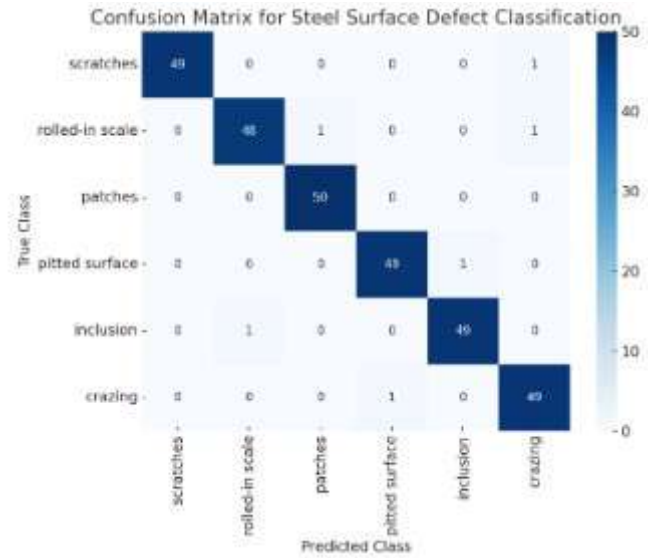


Fig. 8. Confusion matrix

D. Test results visualization

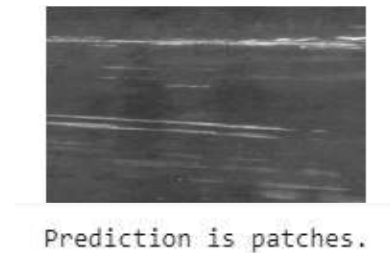


Fig. 9. The model correctly classified the defect under (patches) class. The given image wasn't seen by the algorithm before and this show the strength of our approach.

E. Comparative study

In order to properly analyze the classification results, we opted to test several additional classification algorithms through comparative experiments. In particular, the techniques used by Zhou [21] involved implementing compact CNNs inspired by VGG-16 [22], EDCGAN [23] and ResNet-19 [24] are compared to our study focused on MobileNet-V2. Table 3 displays the outcomes of various classification models.

TABLE 3. COMPARISON OF DIFFERENT CLASSIFICATION METHODS

Proposed Algorithm	Training dataset
Fully Supervised Learning	94.14
VGG-16	98.63
ResNet-19	99.23
EDCGAN	97.62
Our method MobileNet-V2	99.44

From the table above (Table 3), we can remark that our method performs better than the others, even though the total amount of data is not so large. In our study we took only 1800 images consisted of six types of defects (300 image per class).

V. CONCLUSION

This study presented a method for classifying steel surface defects using the MobileNet-V2 architecture. The model was trained on 1,800 images representing six distinct types of defects. To enhance learning, Leaky ReLU was employed as the activation function to mitigate the issue of dying neurons, and dropout layers were incorporated to reduce overfitting.

The proposed model achieved a high accuracy of 99.44%, indicating strong performance on the dataset. However, given the relatively small size of the training set, there is a risk of overfitting, as the model could potentially memorize training samples rather than learning fully generalizable patterns. Although the validation curve and confusion matrix suggest good generalization on the test set, a formal assessment of overfitting would require further evaluation, such as testing on unseen data from varied conditions, experimenting with reduced model capacities, or applying additional regularization techniques like dropout, weight decay, or stronger data augmentation.

Optimization with the AdaMax algorithm proved effective in accelerating and stabilizing learning. Future work could focus on expanding the dataset or generating synthetic samples using methods such as GANs. Additionally, exploring alternative optimizers and activation functions may further enhance the model's performance. Overall, the results demonstrate the strong potential of MobileNet-V2 for steel surface defect classification, while highlighting the importance of careful evaluation to ensure robustness in real-world applications.

REFERENCES

- [1] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv 1409.1556*, 09/04 2014.
- [2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," vol. 7, 12/10 2015.
- [3] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Neural Information Processing Systems*, vol. 25, 01/01 2012, doi: 10.1145/3065386.
- [4] Y. Kateb, H. Megloul, and A. Khebli, "Coronavirus Diagnosis Based on Chest X-Ray Images and Pre-Trained DenseNet-121," *Revue d'Intelligence Artificielle*, 2021.
- [5] T. Ahonen, A. Hadid, and M. Pietikäinen, *Face Recognition with Local Binary Patterns*. 2004, pp. 469-481.
- [6] R. Lienhart and M. Jochen, "An extended set of Haar-like features for rapid object detection Proceedings of the 2002 IEEE International Conference on Image Processing," *Maydt.An extended set of haar-like features for rapid object detection. IEEE ICIP2002*, vol. 1, pp. 900-903, 01/01 2002.
- [7] N. Dalal and B. Triggs, *Histograms of Oriented Gradients for Human Detection*. 2005, pp. 886-893.
- [8] S. Rezaei Aghdam, E. Amid, and M. Imani, *A fast method of steel surface defect detection using decision trees applied to LBP based features*. 2012, pp. 1447-1452.
- [9] S. Ghorai, A. Mukherjee, M. Gangadaran, and P. Dutta, "Automatic Defect Detection on Hot-Rolled Flat Steel Products," *Instrumentation and Measurement, IEEE Transactions on*, vol. 62, pp. 612-621, 03/01 2013, doi: 10.1109/TIM.2012.2218677.
- [10] D. Shumin, L. Zhoufeng, and C. Li, "AdaBoost learning for fabric defect detection based on HOG and SVM," *International Conference On Multimedia Technology (ICMT)*, 07/01 2011, doi: 10.1109/ICMT.2011.6001937.
- [11] Y. Kateb, H. Megloul, and A. Khebli, "Steel Surface Defect Detection Using Convolutional Neural Network," *Algerian Journal of Signals and Systems*, vol. 5, no. 4, pp. 203-208, 2020-12-15 2020, doi: 10.51485/ajss.v5i4.122.
- [12] F. Guizhong *et al.*, "A deep-learning-based approach for fast and robust steel surface defects classification," *Optics and Lasers in Engineering*, vol. 121, pp. 397-405, 10/01 2019, doi: 10.1016/j.optlaseng.2019.05.005.
- [13] Z. Huang, J. Wu, and F. Xie, "Automatic recognition of surface defects for hot-rolled steel strip based on deep attention residual convolutional neural network," *Materials Letters*, vol. 293, 06/15 2021, doi: 10.1016/j.matlet.2021.129707.
- [14] X. Tao, D. Zhang, W. Ma, X. Liu, and D. Xu, "Automatic Metallic Surface Defect Detection and Recognition with Convolutional Neural Networks," *Applied Sciences*, vol. 8, p. 1575, 09/06 2018, doi: 10.3390/app8091575.
- [15] Y. KATEB, A. KHEBLI, H. MEGLOULI, S. AGUIB, and M. S. KHELIFI TOUHAM, "Classifying Surface Fault in Steel Strips Using a Customized NasNet-Mobile CNN and Small Dataset " (in English), *Journal of Electrical Systems*, vol. 20, no. 1, pp. 52-67, 2024.
- [16] M. Li, H. Wang, and Z. Wan, "Surface defect detection of steel strips based on improved YOLOv4," *Computers and Electrical Engineering*, vol. 102, p. 108208, 09/01 2022, doi: 10.1016/j.compeleceng.2022.108208.
- [17] X. Feng, X. Gao, and L. Luo, "X-SDD: A New Benchmark for Hot Rolled Steel Strip Surface Defects Detection," *Symmetry*, vol. 13, no. 4, doi: 10.3390/sym13040706.
- [18] E. Bisong, "Google Colaboratory," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, E. Bisong Ed. Berkeley, CA: Apress, 2019, pp. 59-64.
- [19] {TensorFlow}: *Large-Scale Machine Learning on Heterogeneous Systems*. (2015). [Online]. Available: tensorflow.org
- [20] Keras. *GitHub*. (2015). GitHub. [Online]. Available: <https://github.com/fchollet/keras>
- [21] S. Zhou, Y. Chen, D. Zhang, J. Xie, and Y. Zhou, "Classification of surface defects on steel sheet using convolutional neural networks," *Materiali in tehnologije*, vol. 51, pp. 123-131, 02/14 2017, doi: 10.17222/mit.2015.335.
- [22] A. Sengupta, Y. Ye, R. Wang, C. Liu, and K. Roy, "Going Deeper in Spiking Neural Networks: VGG and Residual Architectures," *Frontiers in Neuroscience*, vol. 13, 02/07 2018, doi: 10.3389/fnins.2019.00095.
- [23] G. Jin, Y. Liu, P. Qin, R. Hong, T. Xu, and G. Lu, "An End-to-End Steel Surface Classification Approach Based on EDCGAN and MobileNet V2," *Sensors*, vol. 23, no. 4, p. 1953, 2023-02-09 2023, doi: 10.3390/s23041953.
- [24] W. Chen, Y. Gao, L. Gao, and X. Li, "A New Ensemble Approach based on Deep Convolutional Neural Networks for Steel Surface Defect classification," *Procedia CIRP*, vol. 72, pp. 1069-1072, 01/01 2018, doi: 10.1016/j.procir.2018.03.264.

A Two-Stage Deep Learning Framework for Accurate Plant Identification from Leaf Images

1st Chatta Asmaa
CSAAI Laboratory
University Ziane Achour of Djelfa
Djelfa, Algeria
asmaa.chatta@univ-djelfa.dz

2nd Latreche Boubakeur
CSAAI Laboratory
University Ziane Achour of Djelfa
Djelfa, Algeria
b.latreche@univ-djelfa.dz

3rd Bouakkaz Mustapha
Department of Computer Science
University of Laghouat
Laghouat, Algeria
m.bouakkaz@lagh-univ.dz

4th Bouakkaz Hicham
Department of Exact Sciences
University of Laghouat
Laghouat, Algeria
hi.bouakkaz@lagh-univ.dz

Abstract—Accurate classification of plant leaves is crucial for applications in agriculture, biodiversity monitoring, and plant-based research. Traditional computer vision techniques often struggle with variations in illumination, orientation, and leaf morphology. This study presents a deep learning-based framework that integrates ResNet18 as a feature extraction backbone with an EfficientNet-B1 head for classification. The dataset was preprocessed through image resizing, normalization, and augmentation operations, including random flipping and rotation, to enhance robustness. The proposed hybrid model was trained on RGB images of size 224×224 using a supervised learning approach. Performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrix metrics. Experimental results show that the ResNet18-EfficientNetB1 hybrid architecture achieves a classification accuracy of 99.7% with strong generalization capabilities. These findings demonstrate the effectiveness of combining residual and multi-scale feature representations for robust and scalable plant leaf classification, paving the way for applications in precision agriculture and plant identification systems.

Keywords—Plant recognition, Deep learning, ResNet18, EfficientNetB1, Transfer learning, Image classification, Feature extraction, Hybrid architecture.

I. INTRODUCTION

Plants play a crucial role in maintaining ecological balance and sustaining life on Earth, contributing significantly to food production, medicine, and environmental systems [1], [2]. Accurate identification and taxonomy of plants are essential in agriculture, pharmacology, and biodiversity conservation [3]. Traditionally, plant identification has relied on manual botanical expertise, which is time-consuming, prone to human error, and requires substantial domain knowledge [4].

Recent advances in *computer vision* and *deep learning* have revolutionized the field of automatic plant identification, allowing the development of image-based recognition systems that can process and classify plant species efficiently and accurately [5]. Among various deep learning methods, *Convolutional Neural Networks (CNNs)* have demonstrated state-of-the-art performance in image classification tasks by

automatically learning hierarchical and discriminative features from raw input images [6].

In particular, *ResNet* and *EfficientNet* architectures have become highly effective for large-scale visual recognition tasks. ResNet introduces residual connections to mitigate the vanishing gradient problem, thereby enabling the training of very deep networks [7], [9]. On the other hand, EfficientNet utilizes compound scaling of network depth, width, and resolution to achieve a better trade-off between accuracy and computational cost [8], [10]. Combining these architectures offers a promising approach to improve both robustness and performance in image recognition applications, including plant species classification.

In this study, we propose a *hybrid deep learning architecture* that integrates *ResNet18* as the feature extraction backbone with *EfficientNetB1* as the classification head for automatic leaf-based plant recognition. The dataset employed in this work contains leaf images from various plant species captured under diverse environmental conditions. Each image is preprocessed to ensure uniformity and consistency across samples.

The main objective of this research is to evaluate the effectiveness of this combined deep architecture in improving the classification accuracy of plant species recognition systems. The remainder of the paper is organized as follows: Section II reviews related works on plant species identification using deep learning; Section III describes the proposed model; Section IV discusses the experimental results; and Section V concludes the study and suggests future research directions.

II. RELATED WORKS

Lee et al. in [11] proposed Deep-Plant, which is considered to be one of the first CNN-based approaches for plant identification. Deep-Plant attained an accuracy level of 96.3% on the Flavia and Swedish leaf datasets, thereby indicating the dominance of deep features over traditional feature descriptors like SIFT or HOG. Building upon this, Beikmohammadi et al. [12] proposed SWP-LeafNet, a multistage CNN integrating

segmentation and wavelet preprocessing. The system reached 98.7% accuracy on the LeafSnap dataset, highlighting the benefit of combining frequency-domain features with spatial convolutional representations.

Similarly, Srinivasa Chakravarthy et al. [13] built a hybrid geometric morphometric-CNN-based system whose performance improved the identification of visually similar species. The system performed with 94.5% accuracy on a multi-species dataset with more than 10,000 leaf samples, thus validating the synergy between geometric descriptors and deep feature descriptors. Transfer learning has also been successfully applied for plant identification. In their study, Bharadwaj et al. [14] employed venation feature-based fine-tuned models on VGG-16 and Inception-V3 for PlantVillage, obtaining an accuracy of 97.8%, thereby outperforming all shallow models, thus validating the universality of feature learning.

In the same vein, M. E. et al. [15] made a comparison between compact CNN models (MobileNet, ShuffleNet, SqueezeNet) for real-time classification tasks. They obtained 95.4% accuracy on the LeafSnap dataset with a low computational cost, making it apt for edge and mobile computing systems.

More recently, hybrid and transformer-based architectures have gained attention. Lee et al. [16] introduced Plant-CNN-ViT, combining convolutional layers for local feature extraction with Vision Transformer (ViT) modules for global context modeling. The hybrid network achieved 99.1% accuracy on PlantVillage, outperforming pure CNN models by 2.3%. Al-thuniyan and Al-Shamasneh [17] proposed an automatic leaf species classification system using CNN, named DeepLeaf, which achieved 98.5% accuracy on the MalayaKew dataset. DeepLeaf performed well in identifying leaves with similar morphology.

In the North African context, Abdelwhab et al. [18] explored transfer learning for the automatic recognition of regional flora, adapting pre-trained models to North African species datasets. Their fine-tuned EfficientNet-B0 achieved 96.8% accuracy, confirming that deep transfer learning can generalize across ecological regions. Aşlıyan [19] proposed a fusion approach combining CNN features with Random Forest classifiers, balancing interpretability and accuracy. On the Flavia dataset, this method reached 97.2%, outperforming traditional decision-tree-based baselines.

Srinivasan et al. [20] offered a comparative analysis of deep learning algorithms (ResNet, DenseNet, EfficientNet) for plant species identification on the PlantVillage dataset, concluding that EfficientNet-B1 obtained the best accuracy of 99.4% in classifying plant species, thereby authenticating the success of compound scaling approaches. Zhang et al. [21] presented a lightweight CNN with channel attention for real-time leaf disease classification, achieving 99.2% accuracy. Likewise, Khan and colleagues [22] proposed a multi-scale leaf classification model combining EfficientNet with feature fusion to achieve 99.3% accuracy on PlantVillage.

Furthermore, Gao et al. [23] developed a transformer-based model integrating attention-guided feature learning, achieving

99.6% accuracy on PlantDoc. Finally, Nguyen et al. [24] introduced a hybrid CNN–Vision Transformer approach for plant leaf classification, tested on a large-scale dataset, achieving 98.9% accuracy, confirming the relevance of combining convolutional and transformer representations in practical plant identification systems.

III. PROPOSED METHODOLOGY

A. DataSet Description

The dataset used in this study consists of high-resolution images of plant leaves collected from multiple subdirectories, where each subfolder represents a distinct plant species. The total dataset includes 1,835 leaf images distributed across 30 plant classes, providing a diverse and representative collection for model training and evaluation. All images are stored in the .jpg format and exhibit natural variations in shape, texture, color, and lighting conditions, which enhance the model's ability to generalize to real-world scenarios. Sample of The dataset is shown in Figure 1.

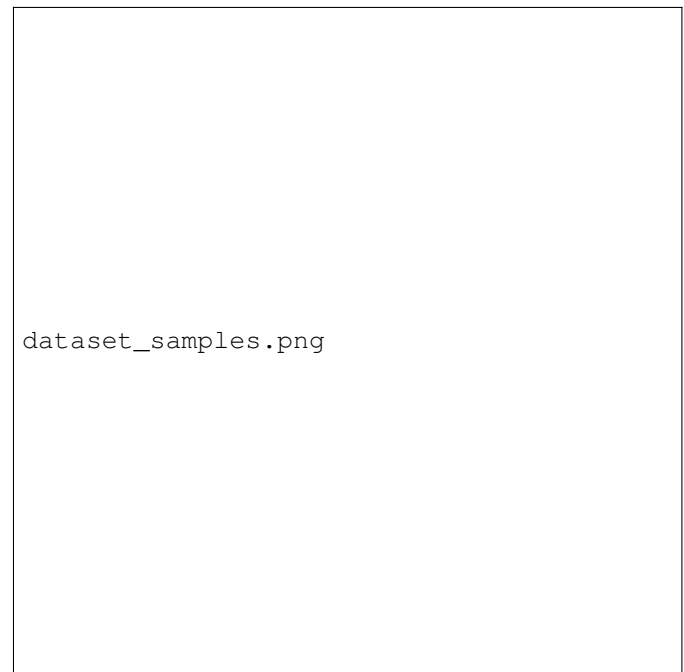


Fig. 1. sample of leaf images

B. Preprocessing of the Dataset

All images were preprocessed to ensure uniformity and improve model generalization. Each image was resized to 224×224 pixels, followed by random horizontal flipping and random rotation ($\pm 15^\circ$) to augment the dataset and reduce overfitting. The images were then converted into tensors and normalized using the ImageNet mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225], ensuring standardized input distributions for efficient and stable model training. The preprocessing workflow, including resizing, augmentation, and normalization steps, is illustrated in Figure 2.

preprocessing_pipeline.png

Fig. 2. Example of image preprocessing and augmentation steps applied to the leaf dataset: resizing, random horizontal flipping, random rotation, and normalization.

C. ResNet-18 Backbone

In this work, the ResNet-18 architecture was selected as the primary feature extraction backbone for plant leaf classification. Proposed by He et al. [1], ResNet-18 is part of the Residual Network (ResNet) family that introduced the concept of skip connections, allowing gradients to propagate more effectively through deeper layers and mitigating the vanishing gradient problem commonly encountered in deep convolutional networks.

The model consists of 18 layers, including convolutional, batch normalization, and pooling layers organized into residual blocks. Each residual block performs identity mapping that facilitates the flow of information directly from earlier to later layers, enabling the model to learn both low-level and high-level representations. This makes it particularly effective for extracting intricate leaf structures such as vein patterns, shapes, and textures, which are essential for accurate classification of plant leaf.

For our experiments, a pre-trained ResNet-18 model on the ImageNet dataset was used to initialize the convolutional layers. The final fully connected (FC) and pooling layers were removed to preserve the convolutional feature extractor while adapting it to our dataset. The extracted feature maps (of dimension 512) were then fed into the classification head for further processing.

The model was optimized using the Adam optimizer with a learning rate of 1×10^{-4} and cross-entropy loss was used as the objective function. Table I summarizes the main parameters of the ResNet-18 backbone.

TABLE I
PARAMETERS OF THE RESNET-18 BACKBONE

Component	Description
Architecture Type	Residual Convolutional Neural Network (ResNet)
Version Used	ResNet-18
Number of Layers	18
Input Image Size	224×224
Feature Output Dimension	512
Activation Function	ReLU
Pooling	Global Average Pooling
Optimizer	Adam ($lr = 1 \times 10^{-4}$)
Loss Function	Cross-Entropy
Total Parameters	≈ 11.7 Million
Pre-training Dataset	ImageNet (1K Classes)

D. EfficientNet-B1 Head

To enhance classification accuracy and model efficiency, the EfficientNet-B1 architecture was employed as the classification head following the ResNet-18 backbone. EfficientNet, introduced by Tan and Le [2], is based on a compound scaling method that uniformly scales network depth, width, and input resolution using a set of fixed scaling coefficients. This approach enables achieving higher accuracy with significantly fewer parameters compared to conventional CNNs.

In our model, the ResNet-18 backbone acts as a powerful feature extractor, generating compact and meaningful representations of the plant leaf images. These extracted features are then fed into the EfficientNet-B1 head, which performs the final classification task. The head includes convolutional layers, squeeze-and-excitation (SE) blocks, and swish activation functions, allowing it to capture complex nonlinear relationships within the extracted features.

The EfficientNet-B1 head was fine-tuned on our dataset with a learning rate of 1×10^{-4} , and categorical cross-entropy was used as the loss function. The Adam optimizer was employed with weight decay to improve generalization performance. Table II summarizes the key parameters of the EfficientNet-B1 head.

TABLE II
PARAMETERS OF THE EFFICIENTNET-B1 HEAD

Component	Description
Architecture Type	Compound Scaled Convolutional Neural Network (EfficientNet)
Version Used	EfficientNet-B1
Input Feature Dimension	512 (from ResNet-18 Backbone)
Output Classes	Number of disease categories
Activation Function	Swish (SiLU)
Pooling	Adaptive Average Pooling
Optimizer	Adam ($lr = 1 \times 10^{-4}$)
Loss Function	Cross-Entropy
Total Parameters	≈ 7.8 Million
Pre-training Dataset	ImageNet (1K Classes)
Scaling Coefficients	Depth = 1.1, Width = 1.0, Resolution = 240

E. Evaluation Metrics

To evaluate the performance of the proposed plant leaf classification model, several standard performance metrics were employed, including *Accuracy*, *Precision*, *Recall*, and the *F1-Score*, along with the *Confusion Matrix* for visual performance interpretation. These metrics provide a comprehensive understanding of the model's ability to distinguish between different plant leaf classes.

- **Accuracy (Acc):** Measures the overall proportion of correctly classified samples and is given by:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where TP and TN are true positives and true negatives, while FP and FN represent false positives and false negatives.

- **Precision (P):** Represents the ratio of correctly predicted positive samples to the total predicted positive samples:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

- **Recall (R):** Indicates the proportion of correctly identified positive samples among all actual positive instances:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

- **F1-Score (F1):** Represents the harmonic mean of Precision and Recall, providing a balanced metric that accounts for both false positives and false negatives:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

- **Confusion Matrix:** Provides a visual representation of classification results across all classes. It helps to identify patterns of misclassification and overlapping leaf features. The general form is defined as:

$$\text{Confusion Matrix} = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \quad (5)$$

In addition, the model's learning behavior was analyzed through the evolution of *training* and *validation* accuracy and loss curves. The increasing accuracy and decreasing loss trends demonstrate the model's stable convergence and strong generalization performance.

IV. RESULTS AND DISCUSSION

The proposed hybrid deep learning architecture, combining a ResNet18 backbone and an EfficientNetB1 head, was implemented to perform plant leaf classification. The model was trained and evaluated on a personal computer equipped with an Intel Core i7 processor, 16 GB RAM, and CPU 1.90GHz (8th generation) and a 64-bit Windows operating system, using Python 3.12 and PyTorch 2.2. The training was conducted over 10 epochs with a batch size of 32, using the Adam optimizer with a learning rate of 0.0001 and cross-entropy loss as the objective function.

The hybrid model demonstrated strong generalization and convergence behavior. It achieved an overall accuracy of 99.7%, a precision of 99.5%, a recall of 99.3%, and a F1 score of 99.2% in the validation data set. The results indicate that combining ResNet18's residual feature extraction with EfficientNetB1's compound scaling enhances the model's ability to capture low and high-level discriminative features from plant leaf images. The confusion matrix, illustrated in Figure 3, highlights the classification performance across different leaf



Fig. 3. Confusion Matrix of the validation results.

categories. Most classes were correctly predicted, with limited misclassifications among visually similar plant.

As shown in Figure 4, the training and validation loss curves exhibit a rapid decrease during the initial epochs, indicating effective optimization and convergence of the network. The validation loss stabilizes at a very low value, confirming that the model generalizes well to unseen data. Simultaneously, the accuracy curves demonstrate a sharp increase in both training and validation accuracy, reaching over 99% after only a few epochs. The two curves remain closely aligned throughout the training process, which suggests that the model does not suffer from overfitting or underfitting.

These findings validate the efficiency of the ResNet18–EfficientNetB1 hybrid framework in achieving robust and accurate classification of plant leaves. The integration of both architectures leverages the residual learning capability of ResNet18 and the efficient scaling strategy of EfficientNetB1, leading to improved feature representation and superior predictive performance. Future extensions may include the incorporation of attention mechanisms or multi-branch feature fusion to further enhance classification accuracy on large-scale and imbalanced datasets.

V. CONCLUSION

This paper presented a deep learning-based approach for plant leaf classification, focusing on improving model accuracy and interpretability through a hybrid framework that combines ResNet18 as a backbone and EfficientNetB1 as a classification head. The primary objective was to develop a

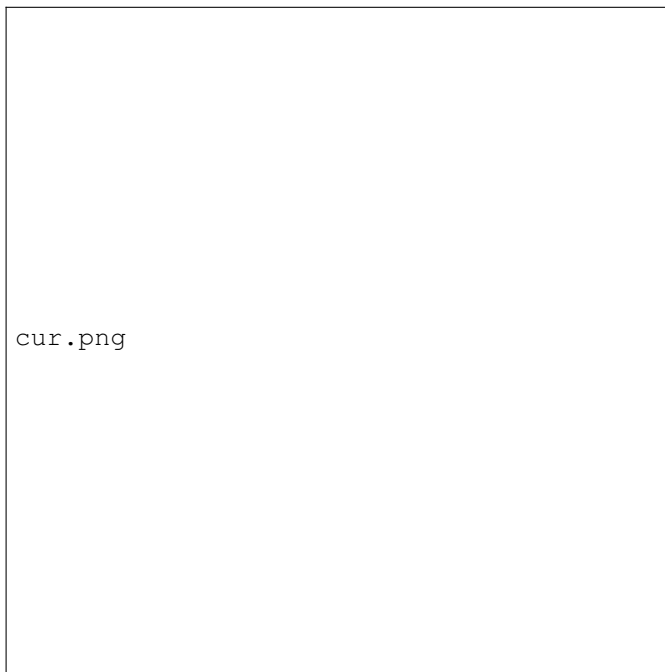


Fig. 4. Training and Validation curves

robust and efficient model capable of accurately identifying plant species based on leaf images. The preprocessing pipeline included resizing, normalization, and augmentation to enhance data diversity, while the Canny edge detector was employed to extract discriminative contour-based features. Experimental results demonstrated high classification performance, achieving a validation accuracy of 99.7%, with consistent improvements across precision, recall, and F1-score metrics.

The findings highlight the effectiveness of combining residual feature extraction and multi-scale representation learning, enabling the model to capture both global and fine-grained visual details. This hybrid architecture offers a promising balance between computational efficiency and accuracy, making it suitable for large-scale botanical applications, such as automated species identification, biodiversity monitoring, and precision agriculture.

Despite these promising results, the study acknowledges certain limitations, including the moderate dataset size and potential bias toward visually similar species. Future work will focus on expanding the dataset to include more diverse plant types, integrating multi-modal data such as spectral or environmental information, and exploring transformer-based architectures to enhance contextual feature learning.

In conclusion, this research confirms that deep learning, when effectively integrated with optimized feature preprocessing, can significantly enhance the accuracy and scalability of plant identification systems. The proposed hybrid framework demonstrates strong potential for real-world deployment, contributing to advancements in plant taxonomy, environmental conservation, and smart agriculture through automated and reliable leaf-based classification.

REFERENCES

- [1] A. K. Goel, M. S. Bhatia, and R. Singh, "Role of medicinal plants in ecosystem balance," *Journal of Plant Science Research*, vol. 9, no. 3, pp. 101–107, 2021.
- [2] M. M. Cowan, "Plant products as antimicrobial agents," *Clinical Microbiology Reviews*, vol. 12, no. 4, pp. 564–582, 1999.
- [3] H. Lee, S. Kim, and J. Kim, "Automated plant identification based on leaf image analysis," *Computers and Electronics in Agriculture*, vol. 120, pp. 139–149, 2016.
- [4] S. Wäldchen and P. Mäder, "Plant species identification using computer vision techniques: A systematic literature review," *Archives of Computational Methods in Engineering*, vol. 25, no. 2, pp. 507–543, 2018.
- [5] X. Zhang, H. Zhou, and L. Lin, "Deep learning for plant identification: A survey," *Frontiers in Plant Science*, vol. 11, pp. 1–16, 2020.
- [6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [8] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Machine Learning (ICML)*, 2019, pp. 6105–6114.
- [9] S. J. Abdallah, A. H. Alzahrani, and F. M. Alharbi, "Plant species classification using transfer learning with EfficientNet," *IEEE Access*, vol. 9, pp. 112233–112243, 2021.
- [10] R. Abdelwhab, et al., "Transfer learning for automatic recognition of regional flora: Adapting pre-trained models to North African species datasets," *IEEE Access*, vol. 12, pp. 14932–14940, 2024.
- [11] S. Lee, C. Chan, and P. Remagnino, "Deep-Plant: Plant identification with convolutional neural networks," in **Proceedings of the IEEE International Conference on Image Processing (ICIP)**, 2015, pp. 452–456.
- [12] A. Beikmohammadi, S. Koohpayeh, and M. Zare, "SWP-LeafNet: A segmentation and wavelet-based CNN for leaf recognition," **Computers and Electronics in Agriculture**, vol. 182, pp. 106030, 2021.
- [13] M. S. Chakravarthy and K. R. Reddy, "A hybrid geometric morphometric and CNN model for plant species recognition," **Ecological Informatics**, vol. 63, pp. 101–112, 2021.
- [14] S. Bharadwaj, R. Agrawal, and P. Shukla, "Fine-tuned transfer learning models for plant identification using venation features," **Expert Systems with Applications**, vol. 175, pp. 114820, 2021.
- [15] M. E., A. Patel, and D. S. Chauhan, "Comparative analysis of compact CNN models for real-time plant leaf classification," **International Journal of Advanced Computer Science and Applications (IJACSA)**, vol. 12, no. 6, pp. 55–63, 2021.
- [16] J. Lee, Y. Kim, and S. Kang, "Plant-CNN-ViT: A hybrid convolution-transformer model for leaf classification," **IEEE Access**, vol. 10, pp. 126045–126057, 2022.
- [17] T. Althuniyan and A. Al-Shamasneh, "DeepLeaf: An automatic leaf classification system using deep convolutional neural networks," **Computers and Electronics in Agriculture**, vol. 192, pp. 106583, 2022.
- [18] M. Abdelwhab, A. Benhamadi, and R. Boudiaf, "Transfer learning for automatic recognition of North African plant species," **Journal of King Saud University - Computer and Information Sciences**, vol. 34, no. 9, pp. 7435–7443, 2022.
- [19] A. Aşhyan, "Leaf classification using CNN feature extraction and Random Forests," **Procedia Computer Science**, vol. 201, pp. 441–448, 2022.
- [20] R. Srinivasan, N. Prasad, and S. Dutta, "A comparative evaluation of deep learning models for plant species recognition," **Applied Intelligence**, vol. 52, no. 7, pp. 7873–7889, 2022.
- [21] P. Zhang, Q. Li, and Y. Liu, "A lightweight CNN with channel attention for real-time leaf disease classification," **Computers and Electronics in Agriculture**, vol. 203, pp. 107470, 2023.
- [22] M. Khan, J. Iqbal, and T. Hussain, "Multi-scale leaf classification using EfficientNet and feature fusion," **Expert Systems with Applications**, vol. 228, pp. 120281, 2023.
- [23] S. Gao, F. Zhao, and H. Xu, "Transformer-based plant disease recognition using attention-guided feature learning," **IEEE Transactions on Artificial Intelligence**, vol. 5, no. 3, pp. 378–389, 2024.
- [24] L. Nguyen, K. Tran, and D. Hoang, "A hybrid CNN–Vision Transformer approach for plant leaf classification," **Engineering Applications of Artificial Intelligence**, vol. 136, pp. 108832, 2024.

IoT-Based Neonatal Incubator Prototype

Meriem Meissoune
Zemouri
Faculty of Technology,
Department of Electrical
Systems Engineering,
University M'hamed
Bougara
Boumerdes, Algeria
meriemmeissoune@gmail.com
ail.com

Khaled Harrar
LIST Laboratory,
University M'Hamed
Bougara
Boumerdes, Algeria
khaled.harrar@univ-boumerdes.dz

Nouma Izeboudjen
Division
Microélectronique et
Nanotechnologie (DMN)
Centre de Développement
des Technologies
Avancées
Baba Hassen, Algiers,
Algeria
nizeboudjen@cdta.dz

Hamza AKROUM
Systems Engineering
and
Telecommunications
Laboratory. Faculty of
Technology, Boumerdes,
Algeria
akroum@univ-boumerdes.dz

Abstract—This paper presents the design and development of an IoT-enabled neonatal incubator prototype aimed at improving the clinical management of premature newborns. Premature infants often struggle to maintain body temperature and humidity balance due to their immaturity. Thus, development of a controlled environments essential for their survival. Traditional incubators provide basic thermal regulation but generally lack connectivity and real-time monitoring, limiting rapid intervention in low-resource settings. In this paper, we propose an IoT-based neonatal prototype built around an ESP32 microcontroller, featuring temperature and humidity regulation along with vital-sign monitoring. and wireless data transmission through the Arduino Remote app. Experimental tests demonstrate acceptable heating performance and stable results of the environmental control, showing the feasibility of using a low-cost IoT architecture to enhance neonatal monitoring. This prototype can contribute to the development of affordable and connected medical solutions for neonatal care units in our country.

Keywords— IoT neonatal incubator, prematurity, ESP32 microcontroller, remote monitoring, environmental control, biomedical sensors, neonatal care.

I. INTRODUCTION

Premature birth, defined as childbirth occurring before 37 weeks of gestation, represents one of the leading causes of neonatal mortality worldwide according to recent WHO reports [1]. Premature infants often suffer from physiological immaturity, including insufficient thermoregulation, high transepidermal water loss, and underdeveloped respiratory and immune systems. These vulnerabilities expose them to life-threatening complications such as hypothermia, dehydration, respiratory distress, and severe infections during the first hours of life [2]. Because of these risks, maintaining a stable and controlled environment is a

fundamental medical requirement. Neonatal incubators are therefore essential devices designed to regulate temperature, humidity, and airflow while protecting infants from external contamination.

Despite their clinical importance, conventional incubators exhibit several limitations, especially in low-resource environments. Most commercially available systems, such as the Giraffe OMNI Bed (Fig. 1), incorporate advanced regulation mechanisms and comprehensive sensor suites, but their cost, maintenance needs, and infrastructure requirements make them inaccessible for many hospitals in developing regions [3]. Moreover, many traditional incubators lack real-time monitoring capabilities, remote supervision, and integration with modern digital health solutions. In contexts where medical staff are limited and response time is critical, the absence of continuous connected monitoring increases the risk of undetected temperature fluctuations or device malfunctions.

Recent advances in embedded systems and Internet of Things (IoT) technologies offer promising solutions to address these limitations. Several studies have proposed IoT-enabled monitoring systems using microcontrollers such as Arduino Mega, GSM modules, or Wi-Fi-based platforms [4], [5], [6]. These systems demonstrate the potential of integrating environmental sensors, wireless communication, and mobile interfaces to improve accessibility and user-friendly monitoring. However, many of these solutions remain limited either by their lack of multi-sensor integration, insufficient performance validation, restricted connectivity, or absence of full prototype implementation.

In this context, the main contribution of this work is to design and experimentally evaluate a low-cost IoT-

based neonatal incubator prototype using the ESP32 microcontroller. The proposed system integrates temperature and humidity sensing, heating control, real-time wireless communication, and a remote monitoring application. The objective is to demonstrate that an affordable IoT architecture can provide stable thermal regulation and reliable remote visualization while leveraging widely available hardware components.

The remainder of this paper is structured as follows: Section II presents the clinical background related to prematurity. Section III discusses related work on IoT-based incubators and highlights existing limitations. Section IV describes the proposed system architecture and experimental setup. Section V presents and interprets the results, including comparisons with existing works. Section VI concludes the study and suggests directions for future improvement.

II. BACKGROUND ON PREMATUREITY

Prematurity refers to any birth occurring before 37 weeks of gestation. Depending on the degree of prematurity, infants may face severe difficulties in maintaining body temperature, regulating humidity loss, and sustaining stable cardiorespiratory functions. Their immature skin and limited thermogenic capacity expose them to hypothermia, dehydration, and infections within minutes if no controlled environment is provided. Clinical guidelines highlight that infants weighing less than 2,500 g require strict thermal and humidity regulation to avoid complications, with the most premature newborns needing continuous monitoring during their first weeks of life. These medical constraints justify the use of neonatal incubators and motivate the development of improved monitoring tools and connected systems for early detection of critical changes.

As shown in Table I, the degrees of prematurity range from 22 to 36 weeks, with increasing risks for extremely premature newborns. These categories guide clinical needs in terms of temperature, humidity, and continuous monitoring.

Table I: Degrees of Prematurity

Prematurity Degree	Gestational Age	Characteristics and Risks
Moderate Prematurity	33-36	Often requires short-term care ; lower risk of severe complications
Great Prematurity	28-32	Higher risk of respiratory issues ; Typically <2000g birth weight
Extreme Prematurity	22-27	Critical ; often <1000g ; uncertain prognosis ; intensive care needed
Post-Term	42	Rare ; may indicate placental issues

III. RELATED WORK

Recent research has explored the integration of IoT technologies into neonatal incubators to enhance monitoring precision and enable remote supervision. Several IoT-based prototypes have been proposed, generally combining microcontrollers, environmental sensors, and wireless communication.

Ortiz *et al.* [7] developed an open-source low-cost incubator for developing countries, demonstrating that connected systems can provide acceptable thermal performance but highlighting limitations in long-term stability. Similarly, studies such as those reported in GE Healthcare documents show that commercial incubators (e.g., Giraffe OMNI Bed) [8] integrate advanced sensing and automatic regulation but remain inaccessible in low-resource contexts due to their high cost and maintenance requirements.

Fig.1. illustrates an advanced commercial incubator (Giraffe OMNI Bed), used here as a reference to compare the performance of our prototype.



Fig.1. Giraffe OMNI Bed

Other works have incorporated environmental monitoring using sensors like DHT22 or pulse sensors for vital-sign estimation, confirming their suitability for medical prototyping despite moderate accuracy. However, these systems often rely on basic microcontrollers with limited wireless capabilities and lack real-time mobile interfaces.

Compared to these systems, the proposed prototype distinguishes itself by combining: (1) an ESP32 microcontroller providing integrated Wi-Fi connectivity, (2) a multi-sensor platform (temperature, humidity, heart rate), (3) real-time remote monitoring through the Arduino Remote App, and (4) experimental evaluation of heating behavior and environmental regulation.

Table II compares the different sensors and actuators used in IoT neonatal incubators. Temperature and humidity sensors are essential, while actuators such as heating elements and fans regulate the environment.

Table II: lists commonly used sensors and actuators in neonatal incubators

Component Type	Examples	Function
Sensors	Temperature, Humidity, Heart Rate, Oxygen Saturation, Pressure	Monitor environmental and vital parameters.
Actuators	Heating Elements, Fans, Humidifiers, Alarms, Adjustable Platforms	Control temperature, air flow, humidity, and alerts.

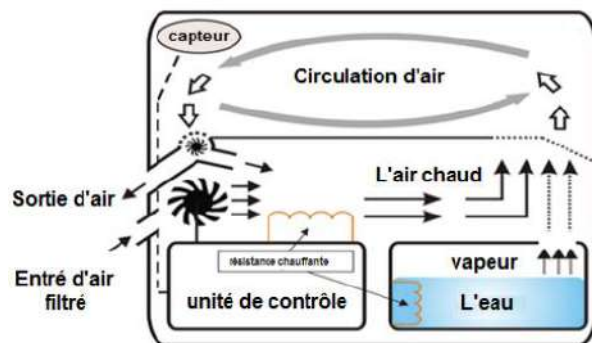


Fig.2. Global block diagram of a neonatal incubator.

Figure 3 shows the mobile monitoring interface used for real-time visualization.

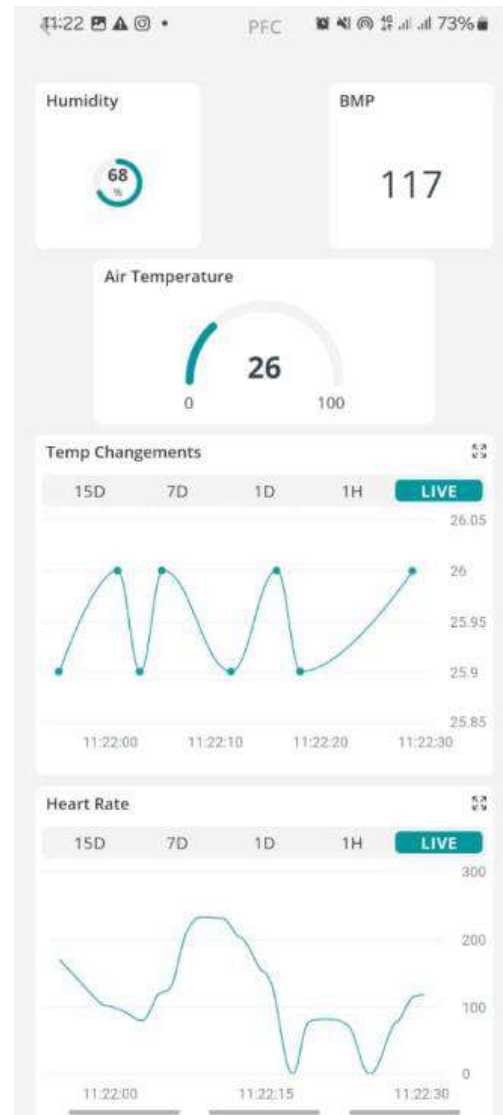


Fig.3. Interface of the application used (Mobile-online mode)

In contrast, our prototype has been developed and tested extensively within the CDTA LAB and the university's scientific club. This iterative development included designing the PCB for sensor integration, programming the ESP32 for real-time data acquisition, and developing the mobile interface. Each iteration refined the system architecture, improved wireless connectivity, and optimized sensor accuracy. The use of aluminum and glass as structural materials enhances thermal stability, structural strength, and biocompatibility, unlike many previous prototypes that rely on wood or low-grade plastic.

The final design incorporates multiple sensors (temperature, humidity, heart rate) and actuators (heating elements, fans) integrated on a custom PCB (Fig. 7), with real-time visualization through the Arduino Remote App (Fig. 3). Experimental validation demonstrated that the incubator can reliably reach and maintain target temperatures of 37°C with fluctuations within $\pm 0.5^\circ\text{C}$, while the IoT connection provides low-latency monitoring (350–500 ms).

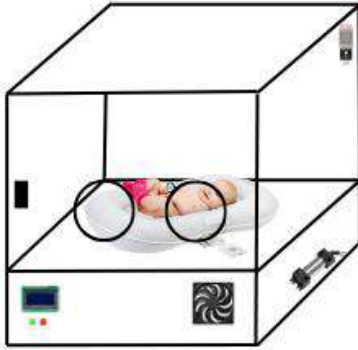


Fig.4. Overall schematic of the prototype developed

Compared to prior works, the combination of durable materials, multi-sensor integration, real-time wireless monitoring, and systematic validation makes this prototype a more complete, low-cost, and deployable solution for neonatal care in low-resource settings. By combining these technical features with iterative testing and optimization, the system represents a meaningful step toward affordable smart incubators capable of supporting critical early-life interventions.

IV. EXPERIMENTAL SETUP AND MEASUREMENTS

Experimental tests were conducted on the prototype to evaluate temperature regulation, humidity monitoring, IoT communication latency, and sensor reliability. Temperature was recorded from ambient level (27.7 °C) to the clinical target of 37 °C using DHT22 measurements sampled every 60 s. Thermal stability was assessed by maintaining the target temperature for 30 min, and humidity readings were evaluated at steady-state conditions.

IoT performance was measured by recording the delay between sensor acquisition and visualization in the Arduino Remote App. Electrical consumption was estimated using the nominal power of the heating resistor, fan, and ESP32-based system.

A. Temperature Evolution Over Time

A heating test was performed with the resistor activated at ambient temperature. Measurements are presented in Table III and plotted in Fig.6. The incubator reached 37 °C in approximately 20 minutes, consistent with typical warm-up durations reported in low-cost incubators.

Table III summarizes temperature evolution during the heating test. A rapid rise occurs during the first ten minutes, followed by a plateau near 37°C.

Table III: Temperature Evolution During Heating Test

Time [min]	Temp C°
0	27.70
3 :44	29.30
6 :40	30.40
13 :16	32
18 :20	32.50
19 :39	34.60
20 :27	37

B. Thermal Stability

The heating system maintains the temperature at $36.5 \pm 0.2^\circ\text{C}$, which meets WHO neonatal care guidelines.

C. Humidity Monitoring

Humidity values were monitored using the DHT22 sensor, which exhibited stable readings with a typical accuracy of $\pm 2\text{--}5\%$. No condensation was observed inside the enclosure during the test period.

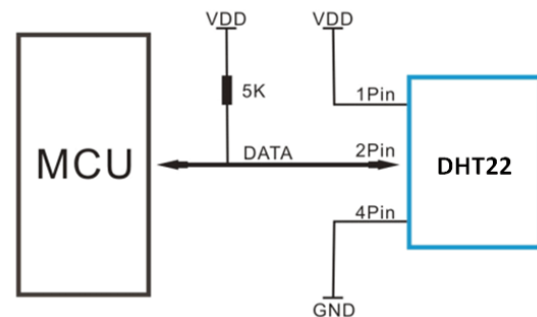


Fig.5. Schematic for connecting a DHT22 digital temperature and humidity sensor to a microcontroller (MCU).

D. IoT Communication Latency

The communication delay between sensor acquisition and app visualization was estimated at 350–500 ms, which is acceptable for medical monitoring use cases.

E. Electrical Consumption

The heating resistor (12V) constitutes the major load. Total system consumption during heating was estimated between 12–18 W, and around 3–5 W in steady-state mode, which aligns with lightweight incubator designs.

V. RESULTS AND DISCUSSION

The heating curve (Fig. 6) shows a nonlinear increase in temperature, with a faster slope during the first 10 minutes (≈ 0.4 °C/min) followed by a plateau as the system approaches 37 °C. This behavior is typical of resistive heating in small enclosures.

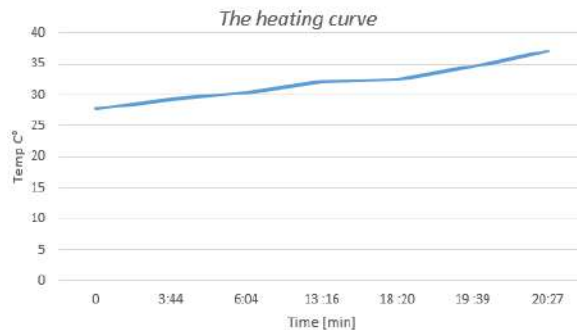
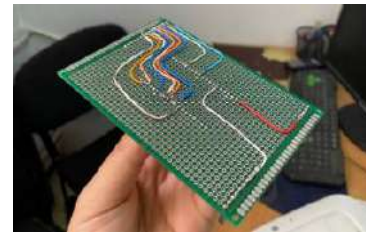


Fig.6. The heating curve

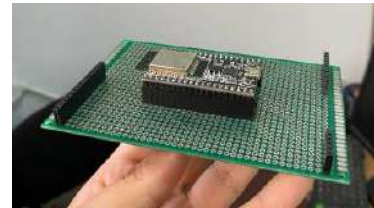
Compared to commercial incubators such as the Giraffe OMNI Bed, which require 10–15 minutes of warm-up, the prototype remains within an acceptable range for low-cost devices.

Experimental results demonstrated that the system can achieve target temperatures within an acceptable timeframe, maintain thermal stability within ± 0.5 °C, and provide reliable IoT connectivity for remote monitoring. These outcomes confirm the feasibility of a low-cost, connected solution for enhancing neonatal care, particularly for infants at high risk due to prematurity.

Compared across repeated tests (three runs), the mean time to reach 37 °C varied by less than 5%, indicating reproducibility of the heating process.



(a)



(b)

Fig.7. The Final PCB, (a) Back Side, (b) Front Side

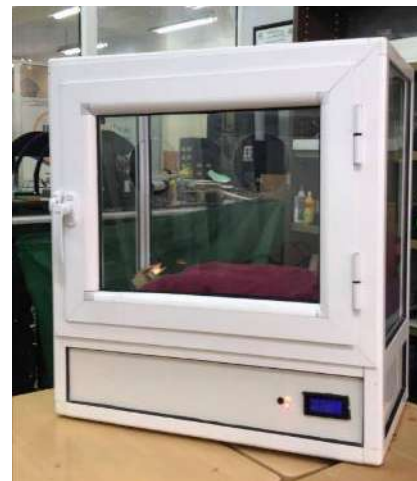


Fig.8. The Prototype



Fig.9. Real-time Temperature and Humidity Measurement on LCD

To position the performance of the proposed incubator within the current state of the art, a comparison was made with several representative systems from the literature. Most existing low-cost prototypes monitor only one or two parameters—typically temperature and humidity—and often lack real-time wireless monitoring or physiological sensing. In contrast, the system developed in this work incorporates

temperature, humidity, and heart-rate monitoring through a pulse sensor, providing a more complete multi-parameter platform. With the ESP32's built-in Wi-Fi, the prototype also supports low-latency remote visualization. Table IV summarizes the main differences in sensing capabilities, connectivity, warm-up time, cost, and limitations.

Table IV shows that previous low-cost designs provide limited sensing and restricted monitoring capabilities. Ortiz et al. [7] exhibit slow warm-up and stability issues, while Kale et al. [6] lack real-time data transmission. Commercial systems like the Giraffe incubator offer fast and comprehensive monitoring but remain inaccessible due to their high cost. By comparison, the proposed prototype achieves a practical balance: it integrates three sensing functions at very low cost, provides real-time Wi-Fi monitoring, and reaches target temperature in about 20 minutes. Its main limitation is the absence of an integrated sterilization module, which can be addressed in future work.

Table IV. Comparison With Existing Works

Work	Sensors	Connectivity	Warm-up time	Cost	Limitation
Ortiz et al. [7]	Temp	Wi-Fi	25–30 min	Low	Limited stability
Kale et al. [6]	Temp, Hum	GSM	Not given	Low	No real-time data
Commercial (Giraffe) [8]	Full suite	Proprietary	10–15 min	Very high	Not accessible
This work	Temp, Hum, Heart Rate	Wi-Fi (ESP32)	≈20 min	Very low	Basic sterilization module

VI. CONCLUSION

This paper presented an IoT based neonatal incubator prototype designed to support the care of premature infants in low-resource settings. By integrating an ESP32 microcontroller, temperature and humidity regulation, vital-sign monitoring, and wireless data transmission, the prototype enables real-time supervision and remote visualization, addressing critical limitations of traditional incubators. The use of biocompatible aluminum and glass ensured thermal

stability and safety, while a custom PCB facilitated reliable integration of sensors and actuators.

Future work includes expanding the sensor suite to include oxygen saturation and pressure monitoring, implementing automated sterilization and humidification modules, optimizing energy efficiency, and developing a fully integrated mobile application for healthcare providers and parents. The presented framework provides a scalable approach for improving the quality of care for premature infants, potentially reducing mortality rates and enhancing the quality of neonatal monitoring in underserved environments.

ACKNOWLEDGMENT

This work was carried out as part of a practical internship at the Development Center of Advanced Technologies (CDTA)/ Algeria in collaboration with University M'hamed Bougara Boumerdes (UMBB), Algeria.

REFERENCES

- [1] World Health Organization, "Preterm Birth," [Online]. Available: <https://www.who.int/fr/news-room/fact-sheets/detail/preterm-birth>. [Accessed: Nov. 21, 2025].
- [2] Inserm, "Une naissance prématurée a lieu," [Online]. Available: <https://www.inserm.fr/dossier/prematurite> [Accessed: Nov. 21, 2025].
- [3] PasseportSanté, "Couveuse pour bébé: quand est-elle utilisée?," [Online]. Available: <https://www.passeportsante.net/famille/enfance?doc=couveuse-bebe-quand-utilisee>. [Accessed: Nov. 21, 2025].
- [4] S. S. M. Sheet, Z. G. Mohammed, and A. A. Abbas, "Smart Infant Incubator Based on Mega Microcontroller," in *2nd International Conference on Engineering Technology and their Applications (IICET 2019)*, Alnajaf, Iraq, 2019.
- [5] D. N. F. Mohamad Ishak, M. M. Abdul Jamil, and R. Ambar, "Arduino Based Infant Monitoring System," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 226, no. 1, pp. 0–6, 2017.
- [6] A. W. Kale, A. H. Raghuvanshi, P. S. Narule, P. S. Gawatre, and S. B. Surwade, "Arduino Based Baby Incubator Using GSM Technology," *Int. Res. J. Eng. Technol.*, vol. 5, no. 4, pp. 462–465, 2018.
- [7] M. A. Ortiz, "Design of Economic and Open Source Incubator for Developing Countries,"
- [8] GE Healthcare, "Giraffe Carestation Neurodevelopmental Care Brochure," [Online]. Available: [Giraffe-Carestation-Neurodevelopmental-care-Brochure.pdf](#).
- [9] X. Zhai and C. Tin, "Automated ECG Classification Using Dual Heartbeat Coupling Based on Convolutional Neural Network," *IEEE Access*, vol. 6, pp. 27465–27472, 2018.
- [10] A. H. Azizulkarim, M. M. Abdul Jamil, and R. Ambar, "Design and Development of Patient Monitoring System," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 226, no. 1, 2017.
- [11] E. Park, Y. Yang, Y. Lee, and H. Lim, "Design and Implementation of Infant Risk Detection System using Embedded System," *WSEAS Trans. Electron.*, vol. 8, pp. 1–7, 2017.
- [12] E. J. L. Costa et al., "Humidity Control System in Newborn Incubator," *19th IMEKO World Congr.*, vol. 4, pp. 1760–1764, 2009.

Optimized Controller Placement for SD-IoT: A Latency–Reliability Aware Approach

1st Raid BOUDI

LIRE Laboratory of Constantine 2
LISI Laboratory of Intelligent Systems and Informatics
Abdelhafid Boussouf University
 Mila, Algeria
 raid.boudi@centre-univ-mila.dz

2nd Nardjes BOUCHEMAL

LISI Laboratory of Intelligent Systems and Informatics
Abdelhafid Boussouf University
 Mila, Algeria
 n.bouchemal.dz@ieee.org

Abstract—The rapid growth of IoT deployments has intensified challenges related to traffic heterogeneity, latency, and reliability within Software-Defined Internet of Things (SD-IoT) networks. Because IoT devices generate uneven and highly dynamic traffic, switches experience heterogeneous load patterns that significantly influence control-plane performance. This makes the Controller Placement Problem (CPP) a critical factor in determining network responsiveness and stability. In this paper, we address CPP by adopting a load-aware perspective, where switch traffic weights are incorporated directly into latency and reliability calculations. We present a simplified Particle Swarm Optimization (PSO) approach that strategically identifies controller locations capable of minimizing weighted switch-to-controller latency while enhancing reliability under potential controller failures. The methodology also considers flow distribution to reduce load imbalance across controllers. Conceptual findings indicate that integrating heterogeneous switch load into optimization leads to improved latency, better traffic flow management, and higher resilience compared to traditional clustering-based methods. The presented work offers an efficient and scalable formulation suitable for practical SD-IoT deployments.

Index Terms—Software-Defined Internet of Things, Controller Placement Problem, Heterogeneity load, Latency, Reliability.

I. INTRODUCTION

The Internet of Things (IoT) has evolved into a core technological pillar enabling large-scale connectivity among sensors, actuators, and embedded devices [1]. These devices continuously generate data streams that support critical applications across healthcare, smart cities, industrial automation, and transportation [2]. However, as IoT infrastructures expand, they inherit challenges related to scalability, traffic diversity, and dynamic behavior [3]. Traditional distributed network architectures often struggle to handle such heterogeneous and dynamic conditions efficiently.

Software-Defined Networking (SDN) introduces centralized control, programmability, and global network visibility, making it a compelling foundation for IoT environments [4]. When SDN principles are combined with IoT, forming Software-Defined IoT (SD-IoT), the resulting architecture offers improved adaptability, simplified management, and enhanced resource allocation [5]. Despite these advantages, SD-IoT

must address critical constraints, including latency-sensitive communication, frequent topology changes, and uneven load distribution across network switches [6].

A central issue limiting SD-IoT performance is the Controller Placement Problem (CPP). Controller-to-switch communication delays, controller overload, and coordination overheads directly influence network responsiveness and stability [7], [8]. The challenge becomes even more complex when considering that IoT traffic is inherently heterogeneous: switches connected to dense or highly active device clusters generate significantly higher traffic than others. Ignoring these load variations can lead to suboptimal controller placements, increased congestion, and reduced reliability.

Recent research has highlighted the importance of incorporating realistic network behavior such as heterogeneous switch loads, fault tolerance, and dynamic traffic flows into CPP formulations [9]. In SD-IoT environments, achieving efficient controller placement is not solely a matter of minimizing distance but requires balancing multiple objectives, including latency reduction, flow distribution, and robustness to controller failures [10].

In this work, we focus on a load-aware formulation of CPP that integrates the influence of switch traffic into controller placement decisions. By using an optimization-driven approach inspired by Particle Swarm Optimization (PSO), we aim to improve latency, enhance reliability, and maintain balanced flow distribution under realistic SD-IoT conditions. The proposed framework provides a simplified yet effective formulation suitable for national-level deployments and resource-constrained IoT environments.

The main contributions of this paper, tailored for a network size fixed at 100 switches, are summarized as follows:

- **Load-Aware Controller Placement Formulation.** We present a controller placement model that integrates heterogeneous switch loads into latency calculations. This enables a more realistic assessment of switch-to-controller communication delay in SD-IoT networks where traffic varies significantly between switches.

- **Simplified PSO-Based Optimization Approach.** A streamlined version of the Particle Swarm Optimization (PSO) method is developed to identify efficient controller positions. The algorithm maintains the essential characteristics of meta-heuristic search while reducing computational overhead, making it suitable for medium-scale networks (specifically 100-node topologies).
- **Reliability Evaluation Using Flow-Weighted Betweenness.** We incorporate a reliability metric that accounts for the impact of a controller failure. By using a flow-weighted betweenness measure, the framework identifies critical controllers and evaluates how network performance changes under failure scenarios.
- **Balanced Flow Distribution for Switch Assignment.** Beyond latency, the method analyzes flow distribution to prevent overload on any single controller. This ensures more stable and predictable control-plane performance in heterogeneous SD-IoT conditions.
- **Focused Evaluation on a 100-Node SD-IoT Topology.** Unlike broader multi-scale studies, this paper concentrates on a standardized 100-switch network size. This allows clearer interpretation of performance trends and practical relevance for mid-sized IoT infrastructures.

By integrating topology awareness with dynamic traffic considerations, the proposed load-aware PSO-based CPP framework offers an efficient and adaptive controller placement solution for SD-IoT networks, particularly suited for latency-sensitive and medium-scale deployments such as the 100-node topology considered in this study.

The rest of this paper is organized as follows: Section II reviews related work; Section III presents the problem formulation and introduces the proposed optimization methodology; Section IV discusses the experimental findings; and Section V concludes the paper and highlights future research directions.

II. RELATED WORK

The Controller Placement Problem has been extensively studied in the context of Software-Defined Networking (SDN), as controller location directly influences communication latency, load balancing, and overall network robustness [11]. Early research efforts primarily targeted Software-Defined Wide Area Networks (SD-WAN), where geographically distributed switches require efficient controller provisioning to ensure low reaction time. Classical formulations such as the k -center and facility location models established the foundation for minimizing propagation delay [12], while subsequent clustering-based methods such as CNPA, hierarchical k -means, and Louvain community detection introduced topology-driven partitioning to reduce computational complexity [13]–[15]. Although effective in structured WAN environments, these methods often assume homogeneous traffic conditions and overlook controller load, limiting their adaptability to dynamic IoT scenarios.

A second category of related work focuses on reliability and resilience. Several studies incorporate controller capacity,

backup assignments, and failure-foresight into CPP. Mixed-Integer Linear Programming (MILP), simulated annealing, and multi-objective heuristics such as RALO have been proposed to address latency under primary and backup path constraints. More recently, link-failure-aware approaches, including CCPGWO and survivability-based meta-heuristics, demonstrated improved worst-case latency performance [12], [16], [17]. However, these strategies are computationally expensive and rely heavily on network topology characteristics, reducing their practicality in large-scale or resource-constrained deployments.

With the evolution of IoT architectures, CPP has gained renewed interest within the Software-Defined Internet of Things (SD-IoT) domain. Unlike traditional SDN systems, SD-IoT networks exhibit massive device density, heterogeneous traffic patterns, and dynamic load fluctuations. Consequently, several studies have adapted controller placement strategies to accommodate IoT-specific constraints. Submodular optimization frameworks, machine learning-based clustering (e.g., silhouette-derived controller counts), and hybrid meta-heuristic algorithms such as ESFO, DEWO, and PAM have been applied to improve latency, fairness, and scalability. Clustering-based strategies such as Louvain-BC and ANP-based controller placement further leverage community structure and multi-criteria decision-making to reduce control overhead [12], [15], [18], [19]. Nonetheless, most of these approaches either fix the number of controllers or rely on static traffic assumptions, thereby failing to capture the heterogeneous load distribution that characterizes real SD-IoT environments.

Recent works also explore quality-of-service (QoS) provisioning, resilient fog deployment, and dynamic load estimation as complementary mechanisms to improve control-plane responsiveness in IoT. These contributions emphasize issues such as flow density, link utilization, and energy consumption, yet very few explicitly integrate heterogeneous switch load into CPP formulations. In particular, most studies compute latency using hop-count or geographic distance only, without weighting these metrics by device-generated traffic volumes [citeTRAN, ALIJehad222, ABDERRAHMANEAmel, KUMARI].

Overall, while prior research has explored topology-aware, reliability-oriented, and heuristic-based solutions across SD-WAN and SD-IoT, there remains a clear need for lightweight, load-aware, and scalable controller placement strategies tailored for IoT networks. The gap is most evident when considering dynamic flow patterns and heterogeneous switch loads factors that strongly influence controller congestion and reaction time but are often neglected in existing models. Motivated by these limitations, this paper proposes a simplified PSO-based CPP framework designed specifically for medium-scale SD-IoT deployments, with an emphasis on weighted latency and reliability under heterogeneous traffic conditions.

III. SYSTEM MODEL AND METHODOLOGY

In SD-IoT, controller placement critically affects control-plane latency, traffic distribution, and network reliability.

Due to heterogeneous traffic patterns produced by IoT devices, some switches generate significantly more control traffic than others, making load-aware decisions necessary for efficient CPP optimization. This section presents the joint latency–reliability formulation, the adopted system model, and the meta-heuristic methodology used to determine optimal controller locations in a 100-node SD-IoT topology.

A. Problem Formulation

The objective of the CPP is to determine a set of controller locations that minimize the overall reaction latency in the control plane while preserving reliability under controller failure events. Let $G = (V, E)$ denote the network graph composed of switches V and links E . Each switch $i \in V$ is associated with a traffic weight w_i , representing its heterogeneous control load.

1) *Latency Model*: Given a controller set $C \subseteq V$, each switch is assigned to the nearest controller in terms of weighted propagation delay. The weighted switch-to-controller latency is expressed as:

$$L(C) = \sum_{i \in V} w_i \cdot d(i, c(i)), \quad (1)$$

where $d(i, c(i))$ is the shortest path distance between switch i and its assigned controller $c(i)$. The goal is to minimize the global weighted latency:

$$\min L(C). \quad (2)$$

2) *Reliability Model*: Reliability reflects the network's ability to maintain acceptable latency in the presence of controller failures. Following the load-aware perspective, we identify the controller with the highest weighted betweenness centrality (WBC), denoted by $c_{\max}$, as it carries the largest influence on traffic flows.

We define reliability as the degradation ratio in the total weighted latency when $c_{\max}$ fails and switches are reassigned to surviving controllers:

$$R(C) = \frac{L_f(C)}{L(C)}, \quad (3)$$

where $L_f(C)$ is the weighted latency after the failure of $c_{\max}$.

A reliable placement satisfies $R(C)$ close to 1, meaning limited latency degradation.

3) *Optimization Objective*: The joint objective combining latency and reliability is defined as:

$$\min F(C) = \alpha L(C) + (1 - \alpha) R(C), \quad (4)$$

where α balances real-time performance (latency) and robustness (reliability).

The problem is NP-hard, motivating the use of meta-heuristic approaches.

B. System Model

The SD-IoT architecture considered in this work includes:

- **Switch Layer**: A set of 100 IoT-enabled switches generating heterogeneous control-plane loads. Each switch i is assigned a load weight w_i derived from application traffic patterns.
- **Controller Layer**: A limited number of SDN controllers placed at selected nodes. Controllers maintain global network knowledge, manage flow rules, and interact with switches through the southbound API.
- **Control Paths**: All communication between switches and controllers relies on shortest-path routing over G . Propagation delays correspond to physical link distances.
- **Weighted Betweenness Centrality (WBC)**: For each controller candidate, WBC is computed as:

$$WBC(i) = \sum_{s \neq t \neq i} \frac{\sigma_{st}(i) \cdot (w_s + w_t)}{\sigma_{st}}, \quad (5)$$

where σ_{st} is the number of shortest paths from s to t , and $\sigma_{st}(i)$ the number passing through node i . The controller with the highest WBC is deemed most critical during the reliability evaluation.

- **Failure Model**: The network undergoes a simulated failure of $c_{\max}$, and switches are reassigned to the nearest remaining controller. The resulting latency degradation defines reliability.

C. Methodology

To efficiently solve the joint latency-reliability CPP, we employ a PSO algorithm tailored for discrete controller placement. The methodology consists of three stages:

1) Preprocessing and Metric Computation:

- Compute shortest paths between all switch pairs.
- Assign weighted loads w_i to each switch.
- Precompute WBC values for all nodes.

2) PSO-Based Optimization:

A particle represents a candidate controller placement vector C .

Each particle stores:

- a placement configuration,
- its local best solution,
- its fitness value based on $F(C)$.

At each iteration:

- 1) Particles update their positions using PSO velocity equations adapted to discrete search space.
- 2) For each candidate C , compute:
 - weighted latency $L(C)$,
 - WBC-based failure impact, giving $L_f(C)$,
 - joint objective $F(C)$.
- 3) Update global best solution.

3) Final Controller Assignment:

After convergence:

- switches are mapped to the nearest selected controller,
- reliability is evaluated again for validation,
- The final controller set C is produced.

This methodology ensures an efficient trade-off between low latency and high reliability under heterogeneous IoT traffic conditions.

IV. EXPERIMENTAL RESULTS

This section presents the performance evaluation of the proposed load-aware PSO-based controller placement strategy on a 100-node SD-IoT topology. The experiments aim to analyze the effectiveness of integrating heterogeneous switch load into the latency and reliability models, and to compare the proposed solution against two widely used clustering-based baselines: k-means-BC and Louvain-BC. All algorithms were implemented in Python, and experiments were performed on a machine equipped with a 2.9 GHz CPU and 16 GB RAM.

The main simulation parameters and evaluation settings are summarized in Table I, while Figure 1 illustrates one sample from the 50 randomly generated topologies used during the simulation process.

TABLE I
SIMULATION PARAMETERS.

Parameter	Value
Number of iterations (M_{x_itr})	100
Number of switches (N)	100
Number of controllers (K)	1–6
Load of the switch (W_{s_i})	1–500 req/s
Controller capacity (U_{c_i})	10000 req/s
Swarm size ($ ppl $)	20
α	0.7
W_{max}, W_{min}	1, 0.1
c_1, c_2	2
r_1, r_2	[0, 1]

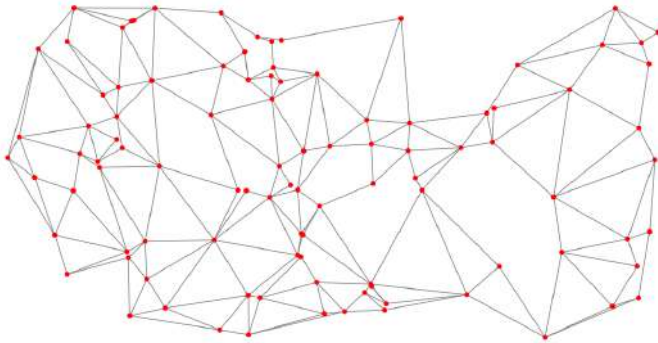


Fig. 1. Example of a generated network topology.

A. Evaluation Metrics

Three main metrics were considered:

- **Latency.** the average switch-to-controller latency.
- **Flow rate.** measures the amount of control traffic handled by each the switches to the controllers, reflecting how effectively the placement distributes high-demand switches.
- **Reliability.** the ratio between pre- and post-failure latency when the controller with highest WBC is removed.

Lower latency and flow rate indicate better performance, while reliability values closer to 1 represent higher robustness.

B. Baseline Methods

To evaluate the performance of our proposed approach, we compare it against several well-known baseline methods commonly used for controller placement in software-defined networks.

- **Optimal.** an optimal solution approach where all possible controller placements and combinations are tested, and the configuration that minimizes latency and maximizes reliability is selected.
- **K-means-BC.** a clustering-based approach where the network is divided into k partitions using k-means, and the controller in each cluster is selected based on Betweenness Centrality (BC).
- **Louvain-BC.** a community detection method where controllers are placed at the most central nodes of the detected communities, with the central node defined as the one having the highest BC [22].

Both baselines assume uniform switch load and do not incorporate traffic heterogeneity or failure-based reliability evaluation.

C. Latency Analysis

Figure 2 shows the average latency for different controller placement approaches as the number of controllers varies. The *Optimal* approach consistently achieves the lowest latency, followed closely by the *Proposed* method. The *K-means-BC* and *Louvain-BC* approaches exhibit higher latency, particularly for smaller numbers of controllers. As the number of controllers increases, latency decreases for all approaches due to improved load distribution, but the *Optimal* approach remains superior because it exhaustively tests all possible combinations.

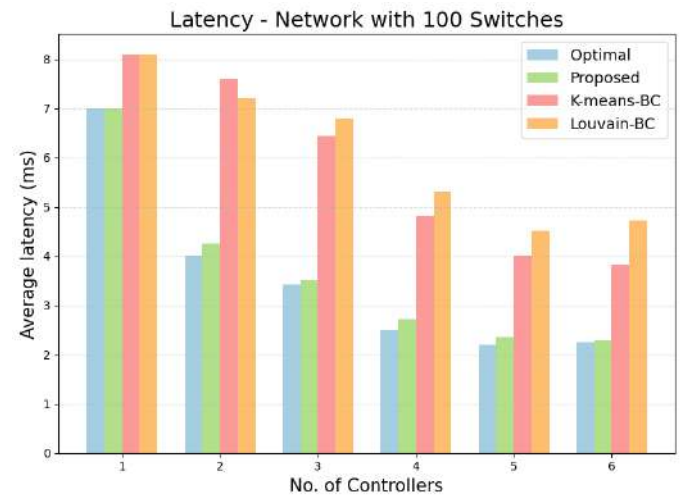


Fig. 2. Average latency comparison for different controller placement approaches in a network with 100 switches.

D. Flow Rate Analysis

The flow rate results are illustrated in Figure 3. Both the *Optimal* and *Proposed* approaches achieve the highest flow rates across all numbers of controllers, whereas *K-means-BC* and *Louvain-BC* achieve lower throughput, especially when fewer controllers are deployed. Overall, increasing the number of controllers improves the flow rate, with the *Optimal* approach consistently providing the best network performance.

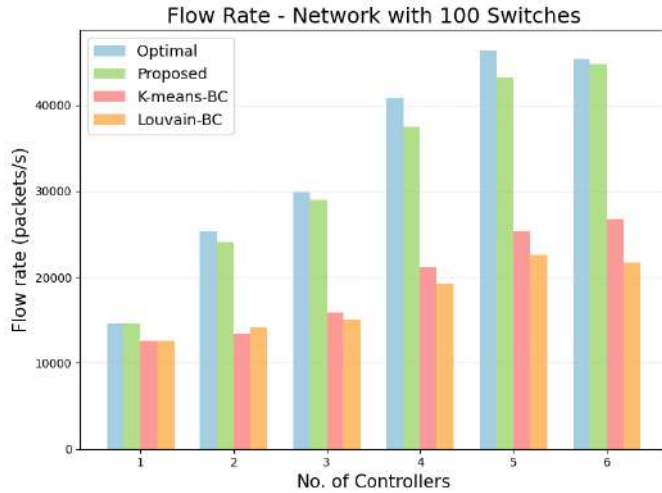


Fig. 3. Flow rate comparison for different controller placement approaches in a network with 100 switches.

E. Reliability Analysis

Figure 4 presents the reliability of the network under different controller placement strategies. The *Optimal* approach achieves the highest reliability percentages, closely followed by the *Proposed* method. *K-means-BC* and *Louvain-BC* have lower reliability, particularly for smaller numbers of controllers. This trend indicates that approaches considering both latency and flow distribution, such as *Optimal* and *Proposed*, result in more reliable network performance.

F. Discussion

While the *Optimal* solution provides the best performance in terms of latency, flow rate, and reliability, it requires testing all possible controller placement combinations, making it computationally expensive and impractical for large networks. The other approaches (*Proposed*, *K-means-BC*, and *Louvain-BC*) are faster but may not fully consider the weight or criticality of switches, which can lead to suboptimal performance in some scenarios. The *Proposed* method offers a good compromise by achieving near-optimal performance with significantly lower computational complexity compared to the exhaustive *Optimal* approach.

V. CONCLUSION AND FUTURE WORK

In this work, we evaluated different controller placement strategies for software-defined networks, including the *Optimal*, *Proposed*, *K-means-BC*, and *Louvain-BC* approaches.

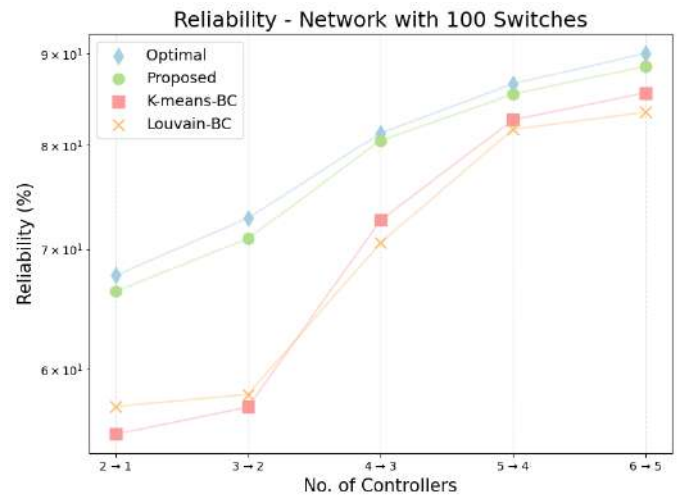


Fig. 4. Reliability comparison for different controller placement approaches in a network with 100 switches.

Our analysis shows that the *Optimal* solution achieves the best performance in terms of latency, flow rate, and reliability by exhaustively testing all possible controller combinations. However, this comes at a high computational cost. The *Proposed* method offers a practical alternative, achieving near-optimal performance while significantly reducing computation time. Heuristic approaches such as *K-means-BC* and *Louvain-BC* are faster but less effective, particularly in scenarios with a small number of controllers or uneven switch distribution.

Future research will focus on addressing the limitations of static controller placement. Specifically, we plan to investigate *dynamic flow patterns over time* and how they affect the network's performance. Additionally, we aim to develop adaptive controller placement strategies that can adjust in real-time to changes in traffic and switch load, overcoming the limitations of static placement and further improving latency, flow rate, and reliability.

REFERENCES

- [1] ABDULHUSSAIN, Sadiq H., MAHMMOD, Basheera M., ALWHE-LAT, Almutadher, et al. A comprehensive review of sensor technologies in IOT: Technical aspects, challenges, and future directions. *Computers*, 2025, vol. 14, no 8, p. 342.
- [2] CHATAUT, Robin, PHOUMMALAYVANE, Alex, et AKL, Robert. Unleashing the power of IoT: A comprehensive review of IoT applications and future prospects in healthcare, agriculture, smart homes, smart cities, and industry 4.0. *Sensors*, 2023, vol. 23, no 16, p. 7194.
- [3] MALHOTRA, Parushi, SINGH, Yashwant, ANAND, Pooja, et al. Internet of things: Evolution, concerns and security challenges. *Sensors*, 2021, vol. 21, no 5, p. 1809.
- [4] JA'AFREH, Mohammed Al, ADHAMI, Hikmat, ALCHALABI, Alaa Eddin, et al. Toward integrating software defined networks with the Internet of Things: a review. *Cluster Computing*, 2022, vol. 25, no 3, p. 1619-1636.
- [5] SIDDIQUI, Shahbaz, HAMEED, Sufian, SHAH, Syed Attique, et al. Toward software-defined networking-based IoT frameworks: A systematic literature review, taxonomy, open challenges and prospects. *Ieee Access*, 2022, vol. 10, p. 70850-70901.
- [6] BOUDI, Raid, LALOU, Mohammed, et BOUCHEMAL, Nardjes. Enhancing controller placement in SD-IoT with dynamic heterogeneous switch traffic via weighted betweenness and GWO. *The Journal of Supercomputing*, 2025, vol. 81, no 14, p. 1-45.

- [7] ALI, Jehad et ROH, Byeong-hee. An effective approach for controller placement in software-defined internet-of-things (SD-IoT). *Sensors*, 2022, vol. 22, no 8, p. 2992.
- [8] SIRJANI, Mohammad Sadegh, MALEKI, Ali, PAKMEHR, Amir, et al. Controller placement in software-defined networks using reinforcement learning and metaheuristics. *Cluster Computing*, 2025, vol. 28, no 10, p. 660.
- [9] DARWISH, Tasneem, ALHAJ, Taqwa Ahmed, et ELHAJ, Fatin A. Controller placement in software defined emerging networks: a review and future directions. *Telecommunication Systems*, 2025, vol. 88, no 1, p. 18.
- [10] ABDULGHANI, Abdulrahman M., ABDULLAH, Azizol, RAHIMAN, A. R., et al. Navigating the Complexities of Controller Placement in SD-WANs: A Multi-Objective Perspective on Current Trends and Future Challenges. *Computer Systems Science & Engineering*, 2025, vol. 49.
- [11] HELLER, Brandon, SHERWOOD, Rob, et MCKEOWN, Nick. The controller placement problem. *ACM SIGCOMM Computer Communication Review*, 2012, vol. 42, no 4, p. 473-478.
- [12] SAHOO, Kshira Sagar, SARKAR, Anamay, MISHRA, Sambit Kumar, et al. Metaheuristic solutions for solving controller placement problem in SDN-based WAN architecture. In : *ICETE 2017-Proceedings of the 14th International Joint Conference on e-Business and Telecommunications*. 2017.
- [13] CHAI, Rong, YUAN, Qiongfang, ZHU, Lei, et al. Control plane delay minimization-based capacitated controller placement algorithm for SDN. *EURASIP Journal on Wireless Communications and Networking*, 2019, vol. 2019, no 1, p. 282.
- [14] SAEEDI GORAGHANI, Mahsa, AFZALI, Mahboubeh, et SHARIFI, Fazel. A Reliable and Load Balancing Controller Placement Method in Software-Defined Networks. *International Journal of Communication Systems*, 2025, vol. 38, no 2, p. e6059.
- [15] ZHANG, Dong, GUO, Junjie, et WU, Chunming. Controller placement based on hierarchical multi-center SDN. *Acta Electronica Sinica*, 2017, vol. 45, no 3, p. 680-686.
- [16] ZOBARY, Firas. Optimizing SDN Controller to Switch Latency for Controller Placement Problem. *Informatica*, 2024, vol. 48, no 8.
- [17] AMIN, Azam, JAHANSHAH, Mohsen, et MEYBODI, Mohammad Reza. Improved learning-automata-based clustering method for controlled placement problem in SDN. *Applied Sciences*, 2023, vol. 13, no 18, p. 10073.
- [18] TRAN, Anh Khoa, PIRAN, Md Jalil, et PHAM, Chuan. SDN controller placement in IoT networks: An optimized submodularity-based approach. *Sensors*, 2019, vol. 19, no 24, p. 5474.
- [19] ALI, Jehad et ROH, Byeong-hee. An effective approach for controller placement in software-defined internet-of-things (SD-IoT). *Sensors*, 2022, vol. 22, no 8, p. 2992.
- [20] ABDERRAHMANE, Amel, DRID, Hamza, et BEHAZ, Amel. A survey of controller placement problem in SDN-IoT network. *International Journal of Networked and Distributed Computing*, 2024, vol. 12, no 2, p. 170-184.
- [21] KUMARI, Abha et SAIRAM, Ashok Singh. A survey of controller placement problem in software defined networks. *arXiv preprint arXiv:1905.04649*, 2019.
- [22] ABDERRAHMANE, Amel et DRID, Hamza. Enhancing control placement in SDN-IoT networks using the Louvain algorithm and betweenness-centrality. *IEEE Access*, 2024.

PSO Algorithm-Based 3D Optimization of UAV Deployment for Maximum Coverage

Dina Alkama¹, Yacine Ouazziz¹, Sylia Zenadji², Mohamed Amine Ouamri³

¹ Université de Bejaia, Faculté de Technologie, Laboratoire d'Informatique Médicale et des Environnements Dynamiques et intelligents (LIMED), 06000 Bejaia, Algérie

² Ecole Supérieure en Science et Technologie de l'Informatique et du Numérique, Université de Bejaia, Laboratory (LITAN) Bejaia, Algérie

³ Département de Network and Telecommunications, L2TI Laboratory, Université Sorbone Paris Nord, Villetaneuse, France

Email : dina.alkama @univ-bejaia.dz

Abstract— In scenarios involving unexpected or short-term events, the rapid deployment of Unmanned Aerial Vehicles (UAVs) operating as aerial base stations (ABSs) is a key enabler for maintaining reliable wireless connectivity in the impacted area. In this paper, we investigate the problem of optimal ABS deployment to maximize coverage performance by employing the Particle Swarm Optimization (PSO) algorithm. The ABS, equipped with directional antennas are modeled as particles in the PSO search space, allowing the joint optimization of both the required number of UAVs and their three-dimensional (3D) placement to ensure full coverage probability. The obtained results demonstrate that PSO offers an effective and computationally efficient solution for ABS deployment, providing high precision and fast convergence in coverage-oriented optimization tasks.

Keywords— Coverage Probability; Deployment; Optimization; Particle Swarm Algorithm; UAV.

I. INTRODUCTION

Over the past decades, unmanned aerial vehicles (UAVs) have experienced significant advancements. Owing to their high mobility and flexible deployment capabilities, UAVs can effectively operate as aerial relays or aerial base stations (ABSs) to enhance communication coverage and extend transmission ranges [1], [2]. Numerous studies in the field of UAV-enabled wireless networks have examined the challenges associated with determining the optimal placement and operation of ABSs to effectively serve ground users. The article in [3] proposes a Global Optimization Algorithm (GOA) for the 3D placement of multiple UAV-BSs, optimizing energy consumption while ensuring coverage. In large-scale scenarios where a single UAV is insufficient, multi-UAV cooperation is required. For that purpose, the study [4] introduces a 3D deployment method that maximizes total system throughput under quality-of-service (QoS) constraints. Additionally, Pasandideh et al. [5] present energy-efficient deployment algorithms and simulation frameworks for ABSs, accounting for transmit power optimization and realistic environmental conditions. Moreover,

a new method have been presented in [6], which can cover all users with the least number of UAVs using an inward spiral UAV deployment but the calculation of this method is complex.

The Particle Swarm Optimization (PSO) algorithm has emerged as an important solution [7] for UAV deployment, due to its ability to provide high-precision solutions with rapid convergence. PSO is inspired by stochastic descent principles: starting from an initial candidate solution, the algorithm iteratively evaluates neighboring solutions and updates the current solution by retaining the best-performing result at each step, effectively guiding the swarm towards an optimal configuration. The authors in [8] investigate how ABSs can be strategically deployed to reestablish wireless coverage when conventional infrastructure becomes unavailable, emphasizing the joint optimization of their number and placement. In their study, the optimization problem is solved using the PSO algorithm. Shakhathreh *et al.* [9] propose a hybrid K-means and PSO method: after horizontally clustering ground users, the 3D positions of UAVs are optimized to maximize coverage.

In this paper, we aim to determine the optimal 3D positions of ABSs using PSO algorithm, in order to minimize the number of ABSs while ensuring full coverage of all ground users within the target area. Each user is assumed to be associated with the ABS that provides the maximum received power above a predefined threshold.

The remainder of this paper is organized as follows: Section 2 presents the system model and provides a detailed description of the channel model. Section 3 introduces the coverage probability, formulates the 3D placement problem, and describes the proposed algorithm. Numerical results are presented in Section 4, and Section 5 concludes the paper.

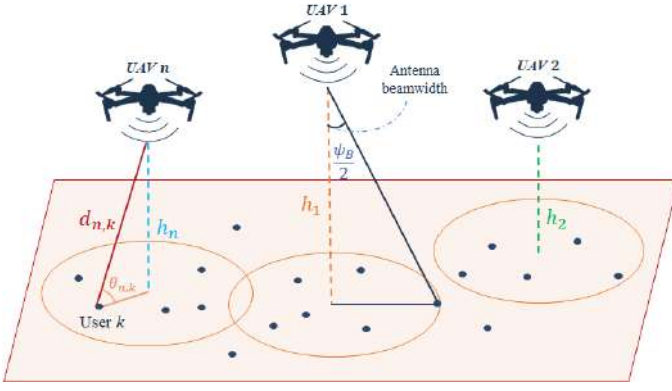


Fig. 1: System Model.

II. SYSTEM AND CHANNEL MODEL

We consider a 3D geographical area with a set $\mathcal{N}$ of N ABSs with fixed transmission power P_U , to provide communication services for K ground users distributed randomly in the target area, as illustrated in Figure 1. The N ABSs hover at different altitudes h_n . The n^{th} ABS's 3D position, is given by $\mathbf{g}_n = (x_n, y_n, h_n)$. The location of each user k is represented by $\mathbf{g}_k = (x_k, y_k)$. We assume that the user k will be served by the n^{th} ABS providing the strongest averaged maximum received power under a certain threshold. The placement of the ABS affects both the number of users within its coverage area, and the quality of the air-to-ground links. To analyze the coverage performance more accurately, it is essential to adopt an appropriate channel model that is suitable for air-to-ground communication which we discuss next.

We assume that each ABS is equipped with a directional-antenna to serve the users with beam width ψ_B , and the antenna gain can be approximated by [10]

$$G = \begin{cases} G_M, & -\frac{\psi_B}{2} \leq \varphi \leq \frac{\psi_B}{2}, \\ G_S, & \text{otherwise,} \end{cases} \quad (1)$$

where φ is the sector angle, $G_M \approx \frac{29000}{\psi_B^2}$ represents the main-lobe gain. $\frac{\psi_B}{2}$ is the ABS antenna's half beam width.

The path loss PL is given by [11]:

$$PL_{LoS} = 10\alpha_{LoS} \log \left(\frac{4\pi f_c d_{n,k}}{c} \right), \quad (2)$$

$$PL_{NLoS} = 10\alpha_{NLoS} \log \left(\frac{4\pi f_c d_{n,k}}{c} \right), \quad (3)$$

where f_c is the carrier frequency, c is the speed of light, $d_{n,k} = \sqrt{(x_n - x_k)^2 + (y_n - y_k)^2 + h_n^2}$, is the distance between UAV n and a ground user k , and α_{LoS} and α_{NLoS} are the LOS and NLOS path loss exponents.

The link between the ABS and the user can be either a LOS or NLOS link. The LoS probability is determined by [11]

$$P_{LoS,n} = \alpha \left(\frac{180}{\pi} \theta_{n,k} - \vartheta \right)^\gamma, \quad (4)$$

Where $\theta_{n,k}$ is the elevation angle between the ABS and the user, which depend on altitude and distance separating the ABS and the user given by $\theta_{n,k} = \sin^{-1}(h_n/d_{n,k})$, ϑ is the observation threshold, representing the minimum elevation angle to receive signal from a UAV. Furthermore, α and γ are constants, reflecting the environment impact. The NLoS probability is $P_{NLoS,n} = 1 - P_{LoS,n}$.

The received signal power $P_{n,k}$ from ABS n to the k^{th} user is:

$$P_{n,k}(\text{dB}) = \begin{cases} P_U + G_M - PL_{LoS} + \psi_{LoS} \geq P_{min}, & \text{LoS link} \\ P_U + G_M - PL_{NLoS} + \psi_{NLoS} \geq P_{min}, & \text{NLoS link} \end{cases} \quad (5)$$

Where, ψ_{LoS} and ψ_{NLoS} are shadow fading with normal distribution in dB scale for LoS and NLoS links. P_{min} is the minimum received power requirement for a successful detection, which is given by $P_{min} = 10 \log(\beta B + \beta I)$, β is the signal-to-interference-plus-noise-ratio (SINR) threshold, B is the noise power and I is the mean interference power received from the nearest ABS m to current ABS, which is given by:

$$I \approx P_k G_S(\varphi_k) \left(10^{-\frac{\mu_{LoS,k}}{10}} P_{LoS,k} + 10^{-\frac{\mu_{NLoS,k}}{10}} P_{NLoS,k} \right) \left(\frac{4\pi f_c d_{n,k}}{c} \right)^{-\eta_p} \quad (6)$$

Hence the received SINR by the user covered by the n^{th} ABS is formulated as

$$SINR_n = \frac{P_{n,k}}{I + B} \quad (7)$$

In the next section, we formulate the 3D placement problem.

III. OPTIMAL UAV DEPLOYMENT BASED ON PARTICLE SWARM ALGORITHM

In this work, the total coverage performance is adopted as the primary metric to evaluate the optimal deployment of N UAVs. For a given SINR threshold β , the coverage at a ground user can be expressed as:

$$P_c = \mathbb{P}(SINR > \beta) = \mathbb{P}(P_{n,k} > P_{min}), \quad (8)$$

The downlink coverage probability of ABS n to ground user k located in the target area, is given by:

$$P_{c,k,n} = P_{LoS,n} Q \left(\frac{P_{min} + PL_{LoS} - P_U - G_M + \mu_{LoS}}{\sigma_{LoS}} \right) + P_{NLoS,n} Q \left(\frac{P_{min} + PL_{NLoS} - P_U - G_M + \mu_{NLoS}}{\sigma_{NLoS}} \right) \quad (9)$$

Where $Q(\cdot)$ is the Q function. Proof see [10].

The coverage probability of ABS network to ground user in the target area can be formulated as

$$P_{C_k} = \rho_k [1 - \Pi_{n \in N} (1 - P_{C_{k,n}})], \quad (10)$$

where ρ_k is the density of user k's location, reflects the importance of user k. The whole ABS network's coverage ability to all users is given by :

$$P_C = \int_{k \in K} P_{C_k} d_k, \quad (11)$$

A. PROPOSED ALGORITHM

The PSO algorithm works on a population called swarm of possible solutions, each individual in the group is called a particle which is considered as a solution of the problem [12]. We build up a swarm consisting of M particles. A j's particle in a search space, is modeled by its position vector and velocity vector respectively; formulated as

$$X_k^j = (x_{k1}, x_{k2}, x_{k3}, \dots, x_{k3N}), \quad (12)$$

$$V_k^j = (v_{k1}, v_{k2}, v_{k3}, \dots, v_{k3N}), \quad (13)$$

During the iterations, each particle j changes its position, and keeps in memory the best position through which it has already passed, $Pbest^j$, which is given by

$$Pbest^j(iter) = (Pbest_{K1}, Pbest_{K2}, Pbest_{K3}, \dots, Pbest_{K3N}),$$

The swarm's best known particle is defined as global best position $gbest$ giving by

$$gbest(iter) = \max (Pbest^j(iter)),$$

The position of each particle is updated at each algorithm iteration by using the following equation [13]

$$\begin{aligned} v_{kn}^j(iter+1) &= Z \left(w v_{kn}^j(iter) + c_1 r_1 (Pbest_{kn}^j - X_{kn}^j(iter)) + c_2 r_2 (gbest_n - X_{kn}^j(iter)) \right), \\ x_{kn}^j(iter+1) &= x_{kn}^j(iter) + v_{kn}^j(iter+1), \end{aligned} \quad (14)$$

Where, Z is the construction factor, c_1 and c_2 are the acceleration coefficients, r_1 and r_2 are random numbers uniformly distributed in the range [0 1].

The optimization problem can be expressed as follows:

$$f: Nu = \operatorname{argmin}(N), \quad (15)$$

$$f: P^{opt} = \operatorname{argmax} U_{P_C}, \quad (16)$$

Where Nu the minimum number of ABSs required to cover all the users K in the predefined area, and P^{opt} are the ABS position when the coverage utility U_{P_C} is maximum. The pseudo-code is presented in algorithm 1.

Algorithm 1 Coverage probability optimization of ABSs based on Particle Swarm Algorithm

1. Initialize max_{iter} particle swarm, X_k^j and V_k^j
 2. Evaluate the initial population $Pbest$ and $gbest$
 3. **While** $iter < max_{iter}$ **do**
 4. $Iter = iter + 1$
 5. **For** each particle j of the swarm **do**
 6. Calculating the fitness value
 7. **end for**
 8. Update $pbest$ and $gbest$ according to formula
 9. Update the velocity v_k and position x_k
 10. **End while**
 11. Get the best solution $gbest$
-

IV. SIMULATION RESULTS AND ANALYSIS

In this section, we simulate the PSO algorithm. We consider the UAV-based communications in an urban environment. We consider 200 users distributed randomly in the area $2000 \times 2000 \text{ km}^2$. We limit the maximum altitude of ABSs to 800 meters. We use MATLAB software as a simulation platform. All simulation parameters are provided in table I.

TABLE 1: Simulation parameters.

Parameter	Value	Parameter of PSO	Value
f_c, P_U	3.6 GHz, 30dBm	Number of particles M	20
$\alpha, \gamma, \alpha_{LoS}, \alpha_{NLoS}, \theta, k_1, k_2, g_1, g_2$	0.6, 0.11, 2, 3, 15, 10.39, 0.05, 29.06, 0.03	Number of iterations	30
$\mu_{LoS}, \mu_{NLoS}, \beta, B$	1dB, 20dB, 5dBm, -120dBm	c_1, c_2, Z	1, 1, 0.7

Fig. 2 shows the variations of the coverage probability with different number of ABSs vs. SINR Threshold. As can be seen, the coverage probability decreases as the SINR threshold increases for a different number of UAVs used. Increasing the number of ABSs improves the coverage which proves that more ABSs can bring higher coverage probability but arriving at a certain stage increasing the number gives the same coverage probability. As shown in Fig. 2, the coverage probability achieved with 4 and 5 ABSs is nearly identical, exhibiting only a slight decrease thereafter. This indicates that deploying 4 ABSs is sufficient to ensure coverage of all users within the predefined area, while any additional ABSs provide limited benefit. These results highlight the importance of selecting the optimal number of ABSs for efficient deployment.

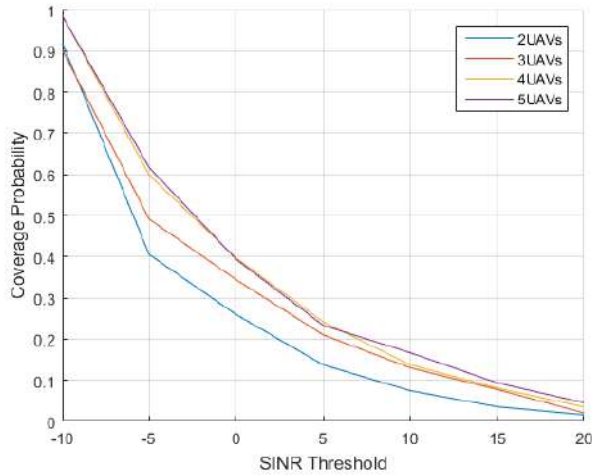


Fig.2 Deployment maps with 4 ABSs and 200 users.

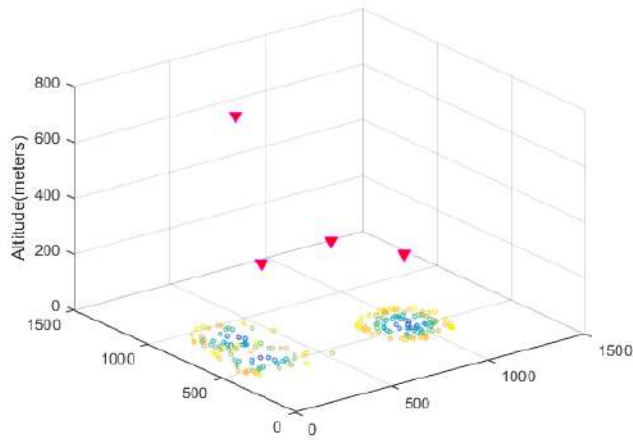


Fig.3 Deployment maps with 4 ABSs and 200 users.

The 3D placement of 4 ABSs is depicted in Fig.3, with 200 users distributed randomly in the area, here we show only the number of users covered by each ABS in order to show the density of users served by each ABS and not the position of users. The 3D coordinate of each ABS are given respectively

$UAV1 \sim (327.5773 \ 834.2211 \ 796.1885)$	,
$UAV2 \sim (976.2419 \ 536.5276 \ 254.4151)$	,
$UAV3 \sim (267.3554 \ 578.3822 \ 343.2253)$	,
$UAV4 \sim (627.0982 \ 579.1344 \ 354.2929)$	,

As can be seen from fig 3 the ABS2 has the maximum density, and covers 99 users, the others ABSs 1,3 and 4 covers 42, 43, 2 users respectively. It means that the location of ABS2 is better compared to others with altitude of $h_2 = 254 \text{ m}$. But there are still 14 users who are not covered.

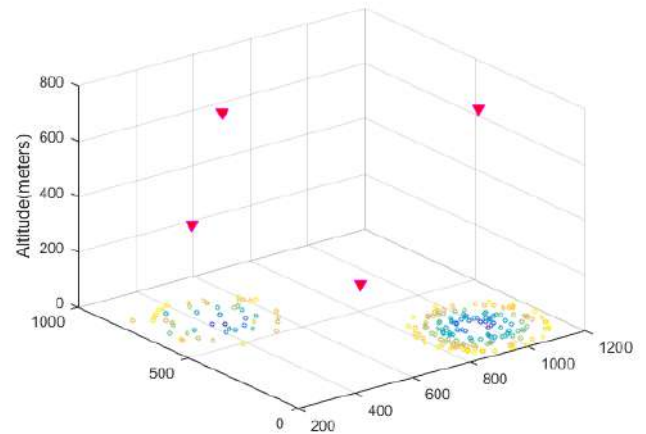


Fig.4 Deployment maps with 4 ABSs and 200 users (an other senario).

In fig. 4 also 4 ABSs were used, after 30 iterations all the 200 users were covered. The UAV2 has the maximum density with coordinates $UAV2 \sim (984.4717 \ 206.7167 \ 784.7852)$. On the other hand no users were served by the ABS 3 with coordinates $UAV3 \sim (719.8588 \ 397.8334 \ 159.1608)$. In this case this ABS is unnecessarily, and can only generate interference to other ABSs. The algorithm proposed to search the optimal placement of ABSs indicates that the optimal solution is not unique. Apart from the height of the air to ground channel, the density of users and beamwidth of the directional antennas have a great influence on the placement of ABSs.

V. CONCLUSION

In this work, we investigate the optimal deployment of ABSs equipped with directional antennas and transmitting at uniform power. Initially, the downlink coverage probability is formulated. Then, for a given geographical area that must be covered by a specified number of UAVs, we propose a heuristic algorithm based on Particle Swarm Optimization (PSO) to maximize the coverage. The results demonstrate that the optimal UAV locations and altitudes can be effectively determined as a function of the number of available UAVs and the gain and beamwidth of the directional antennas. Nevertheless, the limited battery capacity of UAVs continues to be a major operational constraint, underscoring the importance of integrating efficient energy-replenishment methods such as SWIPT technology.

REFERENCES

- [1] M. Hu, K. Kang, T. Liu, G. Pan, and X. Yuan, "Cellular-Connected UAV: Recent Advances and Future Trends," in 2024 IEEE 10th International Symposium on Microwave, Antenna, Propagation and EMC Technologies for Wireless Communications (MAPE), pp. 1–4, IEEE, 2024.
- [2] M. Ghamari, P. Rangel, M. Mehrubeoglu, G. S. Tewolde, and R. S. Sherratt, "Unmanned aerial vehicle communications for civil applications: A review," IEEE Access, vol. 10, pp. 102492–102531, 2022.
- [3] C. Zhang, X. Gui, G. S. Gupta, and S. F. Hasan, "Placement optimization of multiple uavs for energy-efficient maximal user coverage," ICT Express, 2025.

- [4] W. Jiayao, R. Peng, Y. Long, et al., "3d deployment of multi-uav networks for maximizing total system throughput of ground users," *Chinese Journal of Aeronautics*, p. 103866, 2025R. Nicole, "Title of paper with only first word capitalized," *J. Name Stand. Abbrev.*, in press.
- [5] F. Pasandideh, "Providing an energy-efficient uav base station positioning mechanism to improve wireless connectivity," 2024
- [6] B. Yang, X. Xiong, H. Liu, Y. Jia, Y. Gao, A. Tolba, and X. Zhang, "Unmanned aerial vehicle assisted post-disaster communication coverage optimization based on internet of things big data analysis," *Sensors*, vol. 23, no. 15, p. 6795, 2023.
- [7] Y. Zhang, Z. Liyan, and L. Chunpeng, "3-D deployment optimization of UAVs based on particle swarm algorithm." *2019 IEEE 19th International Conference on Communication Technology (ICCT)*. IEEE, 2019.
- [8] F. Pasandideh, F. E. Rodriguez Cesen, P. Henrique Morgan Pereira, C. Esteve Rothenberg, and E. Pignaton de Freitas, "An improved particle swarm optimization algorithm for uav base station placement," *Wireless Personal Communications*, vol. 130, no. 2, pp. 1343–1370, 2023.
- [9] F. Pasandideh, "Providing an energy-efficient uav base station positioning mechanism to improve wireless connectivity," 2024.
- [10] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Efficient deployment of multiple unmanned aerial vehicles for optimal wireless coverage," *IEEE Commun. Lett.*, vol. 20, no. 8, pp. 1647–1650, Aug. 2016.
- [11] A. Al-Hourani, S. Kandeepan, and A. Jamalipour, "Modeling Air-to-ground Path Loss for Low Altitude Platforms in Urban Environments," 2014 IEEE global communications conference, 2014.
- [12] Z. Sylia, C. Gueguen, B. Lamine, T. Larbi, and A. Khireddine, "New strategy for resource allocation using pso-pfs hybrid," *International Journal of Wireless and Mobile Computing*, vol. 18, no. 2, pp. 175–182, 2020.
- [13] L. Brikh, O. Guenounou, F. Yahiaoui, M.-a. Kacimi et A. Ouaret, "Optimization of TSK Fuzzy Model Using New Improved PSO", *International Journal of Control Theory and Applications*, pages 323–333, 2016.

Adaptive Prescribed Performance Fault-Tolerant Control of Full-State Errors for Nonlinear Systems

Mohammed Haddad
Research Center in Industrial
Technologies (CRTI)
Algiers, Algeria
m.haddad@crti.dz

Badis Lekouaghet
Research Center in Industrial
Technologies (CRTI)
Algiers, Algeria
b.lekouaghet@crti.dz

Nidhal Cherrat
Dept. of Electrical Systems
Engineering,
University of M'hamed Bougara
Boumerdes, Algeria
n.cherrat@univ-boumerdes.dz

Abstract—This paper presents an adaptive fault-tolerant control method for high-order uncertain nonlinear strict-feedback systems affected by mismatched disturbances and actuator failures. By introducing an error-based transformation, the method removes restrictive initial error conditions and avoids computing higher-order derivatives of the reference signal. An adaptive law is designed to estimate disturbance bounds in real-time, and a recursive backstepping-based controller is constructed for the transformed system. Using Lyapunov theory, it is proven that all state errors enter a prescribed performance region within a fixed time and eventually converge to zero. Numerical simulations confirm the effectiveness of the proposed approach.

Keywords—Adaptive control, prescribed performance control, full-state errors constraint, actuator failure, external disturbances.

I. INTRODUCTION

Nonlinear systems, which do not satisfy the superposition principle, appear widely in applications such as spacecraft attitude control, robotic motion planning, and power system stabilization. Ensuring stability, reliability, and safety in such systems is essential, especially when they operate under practical constraints including output constraints [1], [2], and state constraints [4], [3]. Traditional methods for handling these constraints include barrier Lyapunov functions (BLFs) [1], [3], and prescribed performance control (PPC) [5].

PPC, first introduced in [5], enables designers to specify transient and steady-state bounds for system errors by transforming constrained tracking problems into bounded-variable problems. PPC has been successfully applied to various nonlinear system types, including pure-feedback [6], and strict-feedback [1]. However, classical PPC requires the initial errors to satisfy $|e_1(0)| < \rho_i(0)$, $i = 1, \dots, n$, which complicates implementation [7]. Existing works address this issue using tuning-function-based performance functions with infinite initial values [2], [8], though these approaches primarily focus on output tracking. Full-state error performance is more desirable in many applications because it ensures safety and improves accuracy across all system states [9]. Yet prior methods only guaranteed boundedness of full-state errors and did not fully account for external disturbances.

Since disturbances are unavoidable in practice, enhancing robustness is crucial. PPC-based robust designs have been studied in [2], [8], [10], [11], but worst-case robust methods [2], [8], [10] often yield conservative and energy-intensive control inputs. Disturbance observers [1], [3], [11] reduce conservatism but require accurate models, which limits their applicability.

Actuator faults also pose serious risks to stability. Several adaptive fault-tolerant control techniques have been proposed, such as sliding-mode-based fault estimation [12], and redundancy-based switching [13]. However, these methods may rely on accurate state values at switching moments, involve complex tuning, or assume that backup actuators remain fully functional.

Motivated by these limitations, this paper investigates tracking control with FSECs in the presence of mismatched disturbances and actuator failures. The main contributions are:

- 1) A new error-dependent transformation that eliminates initial-condition restrictions found in [4], [7], improving robustness and removing the need for higher-order derivatives of the reference signal.
- 2) A full-state prescribed performance framework that ensures all state errors enter the desired region within a preset time and converge to zero asymptotically, improving upon output-only PPC [2], [8] and prior full-state methods [7], [9].
- 3) An adaptive disturbance-bound estimator that maintains robustness without the conservatism of robust control [2], [8] and without the model dependence of disturbance observers [11].

II. MODEL DESCRIPTION AND PROBLEM FORMULATION

A class of high-order nonlinear systems is considered in this paper:

$$\begin{cases} \dot{x}_1 = f_1(\bar{x}_1) + g_1(\bar{x}_1)x_2 + d_1(t) \\ \dot{x}_i = f_i(\bar{x}_i) + g_i(\bar{x}_i)x_{i+1} + d_i(t), \quad i = 2, \dots, n \\ \dot{x}_n = f_n(\bar{x}_n) + g_n(\bar{x}_n)u + d_n(t) \\ y = x_1 \end{cases} \quad (1)$$

where $\bar{x}_i = [x_1, \dots, x_i]^T \in \mathbb{R}^i$, $y \in \mathbb{R}$, $u \in \mathbb{R}$, represent the system state, system output, and control input, respectively. d_i indicates the unknown bounded time-varying disturbance; $f_i(\bar{x}_i) \in \mathbb{R}$, including modeling uncertainty and measurement noise, and $g_i(\bar{x}_i) \in \mathbb{R}$, $i = 1, \dots, n$, are smooth functions and continuous nonlinear mapping in their arguments. During prolonged use, the system may experience several forms of actuator failures, such as:

- 1) Bias failure : $u = v + \bar{b}(t)$;
- 2) Loss-of-effectiveness failure: $u = k(t)v$;
- 3) Lock-in-place failure: $u = \bar{u}$, where $\bar{u}$ denotes the place of stuck actuator.

Only two actuator failure types—bias and loss of effectiveness—are considered, and both are represented using a common model:

$$u = k(t)v + \bar{b}(t) \quad (2)$$

where $k(t) \in (0,1]$ represents the unknown bounded actuator effectiveness, $\bar{b}(t)$ denotes the unknown bounded bias error.

A. Fixed-Time Performance Function

To guarantee the required behavior of the full-state errors (FSEs), we employ a time-varying performance function $\Gamma(\xi)$, chosen because it satisfies the essential properties:

- 1) $\Gamma(0) = 0$, $\lim_{\xi \rightarrow 1^-} \Gamma = +\infty$, $\lim_{\xi \rightarrow -1^+} \Gamma = -\infty$;
- 2) $\Gamma(\xi) = -\Gamma(-\xi)$;
- 3) $\Gamma(\xi)$ is a smooth function;
- 4) The first order derivative of Γ has a lower bound and is greater than zero, and approaches infinity as ξ gets closer to ± 1 .

Design the candidate performance function as follows:

$$\Gamma = \tan\left(\frac{\pi}{2}\beta(t)\right) \quad (3)$$

where $\beta(t)$ is a time-varying scaling function satisfying:

- 1) $\beta(0) = \beta_0 \leq 1$;
- 2) $\beta(0) > 0$, $\lim_{t \rightarrow T^-} \beta = \beta_f < \beta_0$, where β_f denotes an arbitrarily small positive constant;
- 3) $\beta(t) = \beta_f, \forall t \geq T$;
- 4) $\beta(t)$ requires to be n th-order differentiable and its derivatives up to $n - 1$ th must be bounded.

In this paper, a time-varying scaling function β_{ij} , $i = 1, \dots, n, j = 1, 2$, is expressed as follows:

$$\beta_{ij}(t) = \begin{cases} (\beta_{ij0} - \beta_{ijf})e^{\frac{\lambda_{ij}t}{T}} + \beta_{ijf}, & 0 \leq t < T \\ \beta_{ijf}, & T \leq t < +\infty \end{cases} \quad (4)$$

where $\lambda_{ij} > 0$ indicates the convergence rate, $0 < \beta_{ijf} < \beta_{ij0} \leq 1$, are the design constants, and T denotes the setting time. Simultaneously, to avoid tedious offline calculations, the initial value of β_{ij} is set to 1, i.e., $\beta_{ij0} = 1$.

The objective of the control design is to develop an adaptive controller for uncertain high-order system (1) such that:

- 1) The signals of the nonlinear system (1) are guaranteed to be bounded;
- 2) The FSEs e_i are always bounded within the preset performance region and converge to the desired steady-state boundary in a fixed-time.
- 3) The FSEs e_i asymptotically converge to zero, i.e., $\lim_{t \rightarrow +\infty} e_i = 0, i = 1, \dots, n$.

The necessary assumptions and lemmas are given as follows:

Assumption 1 [2]. For external disturbance d_i , there exist a constant $D_i > 0$ such that $|d_i(t)| \leq D_i$ for $\forall t > 0, i = 1, \dots, n$.

Assumption 2 [2]. There exists an unknown constant $a_i > 0$, and an available smooth function $\psi_i(\bar{x}_i) > 0$ such that $|f_i(\bar{x}_i)| \leq a_i\psi_i(\bar{x}_i), i = 1, \dots, n$. Given that $\bar{x}_i$ is bounded, it follows that $\psi_i(\bar{x}_i)$ and $f_i(\bar{x}_i)$ are also bounded.

Assumption 3. The control gain g_i is unknown and time-varying, there exist unknown positive constants $\underline{g}_i$ and $\bar{g}_i$ satisfying $1 \leq \underline{g}_i \leq |g_i| \leq \bar{g}_i$. Without loss of generality, we assume that $\text{sign}(g_i) = 1$.

Assumption 4 [6]. The reference trajectory y_r is known, bounded and piecewise continuous, and its first-order derivative $\dot{y}_r$ is unknown but bounded, i.e., there exist a constant $\mu > 0$ such that $|\dot{y}_r| \leq \mu$.

Assumption 5 [8]. For actuator effectiveness factor k and bias error $\bar{b}$, there exist unknown positive constants $\underline{k}$ and $\bar{b}$ satisfying $0 < \underline{k} \leq k \leq 1$ and $|\bar{b}| \leq \bar{b}$.

Lemma 1 [12]. For any variable v , there exists a $\zeta > 0$ such that: $0 \leq |v| - \frac{v^2}{\sqrt{v^2 + \zeta^2}} \leq \varpi\zeta$, where $\varpi \approx 0.3$.

Lemma 2 [13]. For any variable η , there always exists a variable $\rho > 0$ such that: $0 \leq |\eta| - \eta \tanh\left(\frac{\eta}{\rho}\right) \leq 0.2785\rho$.

Remark 1. In conventional backstepping approaches, it is typically assumed that the reference trajectory y_r and all of its higher-order derivatives are known and remain bounded. Although, in theory, these higher-order derivatives could be computed using analytic formulas or Taylor expansions, obtaining them in real systems is challenging because of model uncertainties, external disturbances, and hardware constraints. To address this issue, we instead assume that only the first-order derivative of the reference trajectory is bounded—an assumption that is realistic and commonly satisfied in practical applications.

III. MAIN RESULTS

A. System Transformation

$$\begin{cases} e_1 = y - y_r \\ e_i = x_i - \alpha_{i-1}, \quad i = 2, \dots, n \end{cases} \quad (5)$$

where e_1, e_i , and α_{i-1} represent the output tracking error, virtual control error, and virtual control law, respectively. Error transformations are used to convert the system's original error behavior into an equivalent unconstrained form for assessing transient and steady-state performance:

$$z_i = \frac{2}{\pi} \arctan(e_i) \quad (6)$$

$$s_i = \frac{z_i}{(\beta_{i1} - z_i)(\beta_{i2} + z_i)} \quad (7)$$

where $i = 1, \dots, n, \chi_i = (\beta_{i1} - z_i)(\beta_{i2} + z_i)$. In accordance with the definition of s_i as presented in (7), it can be obtained that

$$\dot{s}_i = \frac{\partial s_i}{\partial \beta_{i1}} \dot{\beta}_{i1} + \frac{\partial s_i}{\partial \beta_{i2}} \dot{\beta}_{i2} + \frac{\partial s_i}{\partial z_i} \frac{\partial z_i}{\partial e_i} \dot{e}_i \quad (8)$$

where $\frac{\partial s_i}{\partial \beta_{i1}} = \frac{-z_i(\beta_{i2} + z_i)}{\chi_i^2}, \frac{\partial s_i}{\partial \beta_{i2}} = \frac{-z_i(\beta_{i1} - z_i)}{\chi_i^2}, \frac{\partial s_i}{\partial z_i} = \frac{\beta_{i1}\beta_{i2} + z_i^2}{\chi_i^2}$, and $\frac{\partial z_i}{\partial e_i} = \frac{2}{\pi(1+e_i^2)}$.

Subsequently, the following variables are defined:

$$w_i = \frac{\partial s_i}{\partial z_i} \frac{\partial z_i}{\partial e_i}, \quad Y_i = \frac{\partial s_i}{\partial \beta_{i1}} \dot{\beta}_{i1} + \frac{\partial s_i}{\partial \beta_{i2}} \dot{\beta}_{i2} \quad (9)$$

Then, substituting (9) into (8)

$$\dot{s}_i = w_i \dot{e}_i + Y_i \quad (10)$$

Remark 2. For any $\beta_{ij} \in [\beta_{ijf}, 1], z_i \in (-\beta_{i2}, \beta_{i1}), i = 1, \dots, n, j = 1, 2$, it is evident that w_i is positive. Furthermore, it can be concluded that Y_i is bounded with the aid of properties of β_{ij} .

Using expression (7), system (1) can be reformulated as follows:

$$\begin{cases} \dot{s}_1 = w_1(g_1(\alpha_1 + e_2) + f_1 + d_1 - \dot{y}_r) + Y_1 \\ \dot{s}_i = w_i(g_i(\alpha_i + e_{i+1}) + f_i + d_i - \dot{\alpha}_{i-1}) + Y_i \\ \dot{s}_n = w_n(g_n(kv + \bar{b}) + f_n + d_n - \dot{\alpha}_{n-1}) + Y_n \\ y = x_1 \end{cases} \quad (11)$$

B. Controller Design

In the subsequent design process, a series of positive time varying functions, designated as δ_i, δ_{pi} , and ξ_i will be employed, which should be chosen to satisfy that

$\int_0^t \delta_i(\tau) d\tau \leq \bar{\delta}_i \leq \infty$, $|\delta_i^{(n-i)}| \in l_\infty$, $\int_0^t \delta_{pi}(\tau) d\tau \leq \bar{\delta}_{pi} \leq \infty$, $|\delta_{pi}^{(n-1-i)}| \in l_\infty$, $\int_0^t \xi_i(\tau) d\tau \leq \bar{\xi}_i \leq \infty$, $|\xi_i^{(n-i)}| \in l_\infty$, for all $t > 0$ with $\bar{\delta}_i$, $\bar{\delta}_{pi}$, $\bar{\xi}_i$ being positive constants, $i = 1, \dots, n$, e.g., $e^{-\rho_i}$, where $\rho_i > 0$ is a positive design constant.

Step 1. Design the first Lyapunov function candidate as

$$V_1 = \frac{1}{2g_1} s_1^2 + \frac{1}{2\gamma_1} \tilde{\theta}_1^2 + \frac{1}{2} \tilde{D}_1^2 \quad (12)$$

where $\tilde{\theta}_1 = \theta_1 - \hat{\theta}_1$ and $\tilde{D}_1 = D_1 - \hat{D}_1$ are the estimate error, and $\hat{\theta}_1$ and $\hat{D}_1$ are the estimate of θ_1 and D_1 , respectively. $\gamma_1 > 0$ is the designed constant parameter.

The time derivative of V_1 along (11) is

$$\begin{aligned} \dot{V}_1 &= \frac{1}{g_1} s_1 \dot{s}_1 - \frac{1}{\gamma_1} \tilde{\theta}_1 \dot{\hat{\theta}}_1 - \tilde{D}_1 \dot{\hat{D}}_1 \\ &= \frac{s_1}{g_1} (w_1 g_1 e_2 + w_1 g_1 \alpha_1 + w_1 f_1 + w_1 d_1 - w_1 \dot{y}_r + Y_1) \\ &\quad - \frac{1}{\gamma_1} \tilde{\theta}_1 \dot{\hat{\theta}}_1 - \tilde{D}_1 \dot{\hat{D}}_1 \end{aligned} \quad (13)$$

Using Assumptions 2–4 and Young's inequality leads to the stated result:

$$\begin{aligned} \frac{w_1 s_1 f_1}{g_1} &\leq |s_1| \frac{a_1}{g_1} |w_1 \psi_1| \leq |s_1| \frac{a_1}{g_1} \left(\varepsilon_1 w_1^2 \psi_1^2 + \frac{1}{4\varepsilon_1} \right) \\ \frac{s_1 Y_1}{g_1} &\leq |s_1| \frac{1}{g_1} |Y_1| \leq |s_1| \frac{1}{g_1} \left(\varepsilon_1 Y_1^2 + \frac{1}{4\varepsilon_1} \right) \\ -\frac{w_1 s_1 \dot{y}_r}{g_1} &\leq |s_1| \frac{\mu}{g_1} |w_1| \leq |s_1| \frac{\mu}{g_1} \left(\varepsilon_1 w_1^2 + \frac{1}{4\varepsilon_1} \right) \end{aligned} \quad (14)$$

where $\theta_1 = \max\left\{\frac{a_1}{g_1}, \frac{1}{g_1}, \frac{\mu}{g_1}\right\}$, $\Phi_1 = \varepsilon_1 w_1^2 \psi_1^2 + \varepsilon_1 Y_1^2 + \varepsilon_1 w_1^2 + \frac{3}{4\varepsilon_1}$, and ε_1 is a positive design constant.

(13) can alternatively be expressed as:

$$\begin{aligned} \dot{V}_1 &\leq \frac{s_1}{g_1} w_1 g_1 e_2 + \frac{s_1}{g_1} w_1 g_1 \alpha_1 + |s_1 w_1 d_1| + \theta_1 \frac{s_1^2 \Phi_1^2}{\sqrt{s_1^2 \Phi_1^2 + \delta_1^2(t)}} \\ &\quad + \varpi \theta_1 \delta_1(t) - \frac{1}{\gamma_1} \tilde{\theta}_1 \dot{\hat{\theta}}_1 - \tilde{D}_1 \dot{\hat{D}}_1 \end{aligned} \quad (15)$$

Design actual control law α_1 and adaptive law $\hat{\theta}_1$, $\hat{D}_1$ as

$$\begin{cases} \alpha_1 = -\frac{s_1}{w_1} \left(c_1 + \frac{\tilde{\theta}_1 \Phi_1^2}{\sqrt{s_1^2 \Phi_1^2 + \delta_1^2(t)}} \right) - \tilde{D}_1 \tanh\left(\frac{s_1 w_1}{\xi_1}\right) \\ \dot{\hat{\theta}}_1 = \gamma_1 \frac{s_1^2 \Phi_1^2}{\sqrt{s_1^2 \Phi_1^2 + \delta_1^2(t)}} - \delta_{p1}(t) \hat{\theta}_1 \\ \dot{\hat{D}}_1 = s_1 w_1 \tanh\left(\frac{s_1 w_1}{\xi_1}\right) \end{cases} \quad (16)$$

where $\hat{\theta}_1(0) \geq 0$, $\hat{D}_1(0) \geq 0$ and $c_1 > 0$ is a constant.

Obviously, by substituting (16) into (15), it yields

$$\begin{aligned} \dot{V}_1 &\leq -c_1 s_1^2 - \frac{\delta_{p1}}{2\gamma_1} \tilde{\theta}_1^2 + \frac{s_1}{g_1} w_1 g_1 e_2 + \frac{\delta_{p1}}{2\gamma_1} \theta_1^2 \\ &\quad + \varpi \theta_1 \delta_1 + 0.2785 \xi_1 D_1 \end{aligned} \quad (17)$$

where $\frac{\delta_{p1}}{\gamma_1} \tilde{\theta}_1 \hat{\theta}_1 \leq -\frac{\delta_{p1}}{2\gamma_1} \tilde{\theta}_1^2 + \frac{\delta_{p1}}{2\gamma_1} \theta_1^2$ by virtue of Young's inequality, and $|s_1 w_1 d_1| - |s_1 w_1 D_1| \leq 0$.

Step i ($2 \leq i \leq n-1$). Choose the i th Lyapunov function candidate as

$$V_i = V_{i-1} + \frac{1}{2g_i} s_i^2 + \frac{1}{2\gamma_i} \tilde{\theta}_i^2 + \frac{1}{2} \tilde{D}_i^2 \quad (18)$$

The time derivative of V_i along (11) is

$$\begin{aligned} \dot{V}_i &= -\sum_{k=1}^{i-1} c_k s_k^2 - \sum_{k=1}^{i-1} \frac{\delta_{pk}}{2\gamma_k} \tilde{\theta}_k^2 + \varpi \sum_{k=1}^{i-1} \theta_{k-1} \delta_{k-1} \\ &\quad + \sum_{k=1}^{i-1} \frac{\delta_{pk}}{2\gamma_k} \theta_k^2 + \frac{g_{i-1}}{g_i} s_{i-1} w_{i-1} e_i \\ &\quad + \frac{s_i}{g_i} (w_i g_i e_{i+1} + w_i g_i \alpha_i + w_i f_i + w_i d_i - w_i \dot{\alpha}_{i-1} + Y_i) \\ &\quad + \sum_{k=1}^{i-1} 0.2785 \xi_k D_k - \frac{1}{\gamma_i} \tilde{\theta}_i \dot{\hat{\theta}}_i - \tilde{D}_i \dot{\hat{D}}_i \end{aligned} \quad (19)$$

where

$$\dot{\alpha}_{i-1} = \sum_{k=1}^{i-1} \frac{\partial \alpha_{i-1}}{\partial x_k} (g_k x_{k+1} + f_k) + \frac{\partial \alpha_{i-1}}{\partial y_r} \dot{y}_r + \Xi_{i-1} \quad (20)$$

$$\begin{aligned} \text{with } \Xi_{i-1} &= \sum_{l=1}^2 \sum_{k=1}^{i-1} \sum_{j=0}^{i-k} \frac{\partial \alpha_{i-1}}{\partial \beta_{kl}^{(j)}} \rho_{kl}^{(j+1)} + \sum_{k=1}^{i-1} \sum_{j=0}^{i-1-k} \frac{\partial \alpha_{i-1}}{\partial \delta_k^{(j)}} \delta_k^{(j+1)} \\ &\quad + \sum_{k=1}^{i-1} \sum_{j=0}^{i-1-k} \frac{\partial \alpha_{i-1}}{\partial \xi_k^{(j)}} \xi_k^{(j+1)} + \sum_{k=1}^{i-2} \sum_{j=0}^{i-2-k} \frac{\partial \alpha_{i-1}}{\partial \delta_{pk}^{(j)}} \delta_{pk}^{(j+1)} + \sum_{k=1}^{i-1} \frac{\partial \alpha_{i-1}}{\partial \theta_k} \dot{\hat{\theta}}_k \\ &\quad + \sum_{k=1}^{i-1} \frac{\partial \alpha_{i-1}}{\partial \hat{D}_k} \dot{\hat{D}}_k. \end{aligned}$$

Using Assumptions 2–4 and Young's inequality leads to the stated result:

$$\begin{aligned} \frac{s_i w_i f_i}{g_i} &\leq |s_i| \frac{a_i}{g_i} \left(\varepsilon_i w_i^2 \psi_i^2 + \frac{1}{4\varepsilon_i} \right) \\ \frac{s_i Y_i}{g_i} &\leq |s_i| \frac{1}{g_i} \left(\varepsilon_i Y_i^2 + \frac{1}{4\varepsilon_i} \right) \\ -\frac{s_i}{g_i} w_i \sum_{k=1}^{i-1} \frac{\partial \alpha_{i-1}}{\partial x_k} g_k x_{k+1} &\leq |s_i| \sum_{k=1}^{i-1} \frac{g_k}{g_i} \left(\varepsilon_i \left(\frac{\partial \alpha_{i-1}}{\partial x_k} x_{k+1} \right)^2 w_i^2 + \frac{1}{4\varepsilon_i} \right) \\ -\frac{s_i}{g_i} w_i \sum_{k=1}^{i-1} \frac{\partial \alpha_{i-1}}{\partial x_k} f_k &\leq |s_i| \sum_{k=1}^{i-1} \frac{a_k}{g_i} \left(\varepsilon_i \left(\frac{\partial \alpha_{i-1}}{\partial x_k} \psi_k \right)^2 w_i^2 + \frac{1}{4\varepsilon_i} \right) \\ -\frac{s_i}{g_i} w_i \frac{\partial \alpha_{i-1}}{\partial y_r} \dot{y}_r &\leq |s_i| \frac{\mu}{g_i} \left(\varepsilon_i \left(\frac{\partial \alpha_{i-1}}{\partial y_r} \right)^2 w_i^2 + \frac{1}{4\varepsilon_i} \right) \\ -\frac{s_i w_i \Xi_{i-1}}{g_i} &\leq |s_i| \frac{1}{g_i} \left(\varepsilon_i w_i^2 \Xi_{i-1}^2 + \frac{1}{4\varepsilon_i} \right) \\ \frac{g_{i-1}}{g_i} s_{i-1} w_{i-1} e_i &\leq |s_i| \frac{g_{i-1}}{g_i} \left(\varepsilon_i \left(\frac{s_{i-1} w_{i-1} e_i}{s_i} \right)^2 + \frac{1}{4\varepsilon_i} \right) \end{aligned} \quad (21)$$

where $\theta_i = \max\left\{\frac{1}{g_i}, \frac{\mu}{g_i}, \frac{a_i}{g_i}, \dots, \frac{a_i}{g_i}, \frac{g_{i-1}}{g_i}, \dots, \frac{g_{i-1}}{g_i}, \frac{g_{i-1}}{g_i}\right\}$, $\Phi_i \triangleq \varepsilon_i w_i^2 \psi_i^2 + \varepsilon_i Y_i^2 + \sum_{k=1}^{i-1} \varepsilon_i \left(\frac{\partial \alpha_{i-1}}{\partial x_k} x_{k+1} \right)^2 w_i^2 + \sum_{k=1}^{i-1} \varepsilon_i \left(\frac{\partial \alpha_{i-1}}{\partial x_k} \psi_k \right)^2 w_i^2 + \varepsilon_i \left(\frac{\partial \alpha_{i-1}}{\partial y_r} \right)^2 w_i^2 + \varepsilon_i w_i^2 \Xi_{i-1}^2 + \varepsilon_i \left(\frac{s_{i-1} w_{i-1} e_i}{s_i} \right)^2 + \frac{2i+3}{4\varepsilon_i}$, and $\varepsilon_i > 0$ is a positive design constant.

As with (15), the derivative of V_i can be represented as

$$\begin{aligned} \dot{V}_i &= -\sum_{k=1}^{i-1} c_k s_k^2 - \sum_{k=1}^{i-1} \frac{\delta_{pk}}{2\gamma_k} \tilde{\theta}_k^2 + \varpi \sum_{k=1}^{i-1} \theta_{k-1} \delta_{k-1} + |s_i w_i d_i| \\ &\quad + \frac{s_i}{g_i} w_i g_i e_{i+1} + \frac{s_i}{g_i} w_i g_i \alpha_i + \theta_i \frac{s_i^2 \Phi_i^2}{\sqrt{s_i^2 \Phi_i^2 + \delta_i^2(t)}} + \sum_{k=1}^{i-1} \frac{\delta_{pk}}{2\gamma_k} \theta_k^2 \\ &\quad + \varpi \theta_i \delta_i - \frac{1}{\gamma_i} \tilde{\theta}_i \dot{\hat{\theta}}_i - \tilde{D}_i \dot{\hat{D}}_i + \sum_{k=1}^{i-1} 0.2785 \xi_k D_k \end{aligned} \quad (22)$$

Design actual control law α_i and adaptive law $\hat{\theta}_i$, $\hat{D}_i$ as

$$\begin{cases} \alpha_i = -\frac{s_i}{w_i} \left(c_i + \frac{\tilde{\theta}_i \Phi_i^2}{\sqrt{s_i^2 \Phi_i^2 + \delta_i^2(t)}} \right) - \tilde{D}_i \tanh\left(\frac{s_i w_i}{\xi_i}\right) \\ \dot{\hat{\theta}}_i = \gamma_i \frac{s_i^2 \Phi_i^2}{\sqrt{s_i^2 \Phi_i^2 + \delta_i^2(t)}} - \delta_{pi}(t) \hat{\theta}_i \\ \dot{\hat{D}}_i = s_i w_i \tanh\left(\frac{s_i w_i}{\xi_i}\right) \end{cases} \quad (23)$$

where $\hat{\theta}_i(0) \geq 0$, $\hat{D}_i(0) \geq 0$ and $c_i > 0$ is a design constant.

Substituting (23) into (22), it yields

$$\begin{aligned} \dot{V}_i &\leq -\sum_{k=1}^i c_k s_k^2 - \sum_{k=1}^i \frac{\delta_{pk}}{2\gamma_k} \tilde{\theta}_k^2 + \frac{s_i}{g_i} w_i g_i e_{i+1} \\ &\quad + \sum_{k=1}^i \frac{\delta_{pk}}{2\gamma_k} \theta_k^2 + \varpi \sum_{k=1}^i \theta_k \delta_k + \sum_{k=1}^i 0.2785 \xi_k D_k \end{aligned} \quad (24)$$

Step n. Finally, design the Lyapunov function as

$$V_n = V_{n-1} + \frac{1}{2\bar{h}_n} s_n^2 + \frac{1}{2\gamma_n} \tilde{\theta}_n^2 + \frac{1}{2} \tilde{D}_n^2 \quad (25)$$

From Assumption 3 and 5, it can be known that there exists a constant $\underline{h}_n = \underline{g}_n \underline{k}$ such that $g_n k \geq \underline{h}_n \geq 1$. The time derivative of V_n along (11) is

$$\begin{aligned} \dot{V}_n \leq & -\sum_{k=1}^{n-1} c_k s_k^2 - \sum_{k=1}^{n-1} \frac{\delta_{pk}}{2\gamma_k} \tilde{\theta}_k^2 + \sum_{k=1}^{n-1} 0.2785 \xi_k D_k \\ & + \sum_{k=1}^{n-1} \frac{\delta_{pk}}{2\gamma_k} \theta_k^2 + \varpi \sum_{k=1}^{n-1} \theta_k \delta_k + \frac{s_{n-1}}{\underline{g}_{n-1}} w_{n-1} g_{n-1} e_n \\ & + \frac{s_n}{\underline{h}_n} (w_n h_n v + w_n g_n \bar{b} + w_n f_n + w_n d_n - w_n \dot{\alpha}_{n-1} + Y_n) \\ & - \frac{1}{\gamma_n} \tilde{\theta}_n \dot{\hat{\theta}}_n - \tilde{D}_n \dot{\hat{D}}_n \end{aligned} \quad (26)$$

where $\theta_n \triangleq \max \left\{ \frac{1}{\underline{h}_n}, \frac{\mu}{\underline{h}_n}, \frac{a_1}{\underline{h}_n}, \dots, \frac{a_n}{\underline{h}_n}, \frac{\bar{g}_1}{\underline{h}_n}, \dots, \frac{\bar{g}_{n-1}}{\underline{h}_n}, \frac{\bar{g}_n b}{\underline{h}_n}, \frac{\bar{g}_{n-1}}{\underline{g}_{n-1}} \right\}$,

$$\begin{aligned} \Phi_n \triangleq & \varepsilon_n w_n^2 \psi_n^2 + \varepsilon_n Y_n^2 + \sum_{k=1}^{n-1} \varepsilon_n \left(\frac{\partial \alpha_{n-1}}{\partial x_k} x_{k+1} \right)^2 w_n^2 + \\ & \sum_{k=1}^{n-1} \varepsilon_n \left(\frac{\partial \alpha_{n-1}}{\partial x_k} \psi_k \right)^2 w_n^2 + \varepsilon_n \left(\frac{\partial \alpha_{n-1}}{\partial y_r} \right)^2 w_n^2 + \varepsilon_n w_n^2 \Xi_{n-1}^2 + \\ & \varepsilon_n \left(\frac{s_{n-1} w_{n-1} e_n}{s_n} \right)^2 + \varepsilon_n w_n^2 + \frac{2n+4}{4\varepsilon_n}. \end{aligned}$$

Design actual control law v and adaptive law $\hat{\theta}_n, \hat{D}_n$ as

$$\begin{cases} v = -\frac{s_n}{w_n} \left(c_n + \frac{\bar{\theta}_n \Phi_n^2}{\sqrt{s_n^2 \Phi_n^2 + \delta_n^2}} \right) - \tilde{D}_n \tanh \left(\frac{w_n s_n}{\xi_n} \right) \\ \dot{\hat{\theta}}_n = \gamma_n \frac{s_n^2 \Phi_n^2}{\sqrt{s_n^2 \Phi_n^2 + \delta_n^2}} - \delta_{pn}(t) \hat{\theta}_n \\ \dot{\hat{D}}_n = s_n w_n \tanh \left(\frac{s_n w_n}{\xi_n} \right) \end{cases} \quad (27)$$

where $\hat{\theta}_n(0) \geq 0, \hat{D}_n(0) \geq 0$ and $c_n > 0$ is a design constant.

Substituting (27) into (26), it yields

$$\begin{aligned} \dot{V}_n \leq & -\sum_{k=1}^n c_k s_k^2 - \sum_{k=1}^n \frac{\delta_{pk}}{2\gamma_k} \tilde{\theta}_k^2 + \sum_{k=1}^n \frac{\delta_{pk}}{2\gamma_k} \theta_k^2 \\ & + \varpi \sum_{k=1}^n \theta_k \delta_k + \sum_{k=1}^n 0.2785 \xi_k D_k \\ \leq & -c V_n + \Delta \end{aligned} \quad (28)$$

where $c \triangleq \min \{ 2c_k \underline{g}_k, 2c_n \underline{h}_n, \delta_{pi} \}$, $k = 1, \dots, n-1$, $i = 1, \dots, n$, and $\Delta \triangleq \sum_{k=1}^n \frac{\delta_{pk}}{2\gamma_k} \theta_k^2 + \varpi \sum_{k=1}^n \theta_k \delta_k + \sum_{k=1}^n 0.2785 \xi_k D_k$, and $V_x = \sum_{i=1}^{n-1} \frac{1}{2g_i} s_i^2 + \frac{1}{2h_n} s_n^2 + \sum_{i=1}^n \tilde{\theta}_i^2$.

C. Stability Analysis

Theorem 1. Considering the nonlinear system (1) with actuator failure (2), under the Assumptions 1-5. If the actual control law (27) and the virtual law (16), (23) are implemented, then the subsequent objectives are ensured for any initial condition:

- 1) All signals in the closed system (1) can be achieve globally ultimately uniformly bounded, if $\beta_{ij0} = 1$, $j = 1, 2$. Otherwise, if $0 < \beta_{ij0} < 1$, semi-globally ultimately uniformly bounded can be achieved,
- 2) The full state errors e_i are always capable of being constrained within performance boundaries and attaining the desired steady-state error boundaries in a specified time,
- 3) The asymptotic convergence of the full state errors can be achieved.

Proof. Firstly, we validate the boundedness of all signals present in the closed system (1). On the basis of (28) in conjunction with the definition given for V_n , it can be concluded that $s_i, \tilde{\theta}_i$ are bounded for $t \in [0, t_f]$, $i = 1, \dots, n$. As $\tilde{\theta}_i = \theta_i - \hat{\theta}_i$, it further develops that $\hat{\theta}_i$ is bounded on $t \in [0, t_f]$. In addition, owing to s_i is bounded, there exists some positive constants $\underline{\beta}_{i2}, \bar{\beta}_{i1}$ such that $-\beta_{i2} < -\beta_{i2} \leq z_i \leq \bar{\beta}_{i1} < \beta_{i1}$ on $t \in [0, t_f]$, which further indicates e_i is bounded on $t \in [0, t_f]$. It is worth mentioning that $z_i, \beta_{ij}, \hat{\beta}_{ij}, j = 1, 2$, are bounded, which implies w_i, Y_i are bounded on $t \in [0, t_f]$. As the reference trajectory y_r is bounded, according to the expression of e_i from (5), it can be concluded that x_i is bounded, guaranteeing the boundedness of ψ_1, f_1 and Φ_1 .

According to the definition of $\alpha_{i-1}, \hat{\theta}_i$ and $\hat{D}_i$ showed in (16), the boundness of $\alpha_i, \hat{\theta}_i$ and $\hat{D}_i$ are guaranteed. Consequently, all signals within the nonlinear system (1) remain bounded for the time period $[0, t_f]$. Since the boundness of $s_i, \hat{\theta}_i$ and $\hat{D}_i$ does not depend on t_f , which implies t_f could be expanded to infinity. Subsequently, we demonstrate that the desired performance criteria concerning the full state errors are guaranteed. According to (28), it is seen that $-\beta_{i2} < z_i < \beta_{i1}$, i.e., the full state errors are always constrained within the performance boundary. Furthermore, by virtue of the property of (4), we obtain that $-\Gamma(\beta_{i2f}) < e_i < \Gamma(\beta_{i1f})$, $\forall t \geq T$. Consequently, the full state errors can be integrated into the prescribed region $(-\Gamma(\beta_{i2f}), \Gamma(\beta_{i1f}))$ within the fixed time T . Finally, we demonstrate that the asymptotic convergence of the full state errors can be achieved. Integrating of both side of (28) over the interval $(0, t)$, it yields

$$\begin{aligned} V_n(t) = & V_n(0) - \sum_{k=1}^n \int_0^t c_k s_k^2 d\tau - \sum_{k=1}^n \int_0^t \frac{\delta_{pk}}{2\gamma_k} \tilde{\theta}_k^2 d\tau \\ & + \sum_{k=1}^n \int_0^t \frac{\delta_{pk}}{2\gamma_k} \theta_k^2 d\tau + \varpi \sum_{k=1}^n \int_0^t \theta_k \delta_k d\tau \\ & + \sum_{k=1}^n 0.2785 \int_0^t \xi_k D_k d\tau \\ \leq & V_n(0) - \sum_{k=1}^n \int_0^t c_k s_k^2 d\tau - \sum_{k=1}^n \int_0^t \frac{\delta_{pk}}{2\gamma_k} \tilde{\theta}_k^2 d\tau \\ & + \sum_{k=1}^n \frac{\delta_{pk}}{2\gamma_k} \theta_k^2 + \varpi \sum_{k=1}^n \theta_k \delta_k + \sum_{k=1}^n 0.2785 \xi_k D_k \end{aligned} \quad (29)$$

Furthermore, it is concluded that

$$\begin{aligned} \sum_{k=1}^n \int_0^t c_k s_k^2 d\tau \leq & -V_n(t) + V_n(0) + \sum_{k=1}^n \frac{\delta_{pk}}{2\gamma_k} \theta_k^2 \\ & + \varpi \sum_{k=1}^n \theta_k \delta_k + \sum_{k=1}^n 0.2785 \xi_k D_k \end{aligned} \quad (30)$$

From the preceding analysis, it can be inferred that $s_i \in \ell_2 \cap \ell_\infty$ and $\hat{s}_i \in \ell_\infty$, with the aid of Barbalat's Lemma, which implies that $\lim_{t \rightarrow \infty} s_i = 0$. Then from the definition of s_i , it can be known that $\lim_{t \rightarrow \infty} z_i = 0$, further obtain $\lim_{t \rightarrow \infty} e_i = 0$. This completes the proof.

Remark 3. The term $\frac{e_i}{s_i}$ from (21) is well-defined, $i = 1, \dots, n$.

If $s_i \neq 0$, which implies $e_i \neq 0$, and because s_i is bounded, it further obtains the boundness of e_i . Therefore, it is evident that $\frac{e_i}{s_i}$ is bounded. If $s_i = 0$, then $e_i = 0$, according to the definition of z_i, s_i from (6) and (7), it can be seen that

$$\frac{e_i}{s_i} = \frac{e_i}{z_i} (\beta_{i1} - z_i) (\beta_{i2} + z_i) \quad (31)$$

By applying the established L'Hôpital's rule, it can be obtained that

$$\lim_{e_i \rightarrow 0} \frac{e_i}{s_i} = \frac{e_i (\beta_{i1} - \frac{2}{\pi} \arctan(e_i)) (\beta_{i2} + \frac{2}{\pi} \arctan(e_i))}{\frac{2}{\pi} \arctan(e_i)} = \frac{\pi}{2} \beta_{i1} \beta_{i2} \leq \frac{\pi}{2} \quad (32)$$

Therefore, $\frac{e_i}{s_i}$ is bounded and the occurrence of the singularity is precluded. Nevertheless, during the simulation process, computers do not calculate the limit of $\frac{e_i}{s_i}$. To resolve this issue, a small positive constant ι is proposed, whereby when $|s_i| \geq \iota$, $\frac{e_i}{s_i}$ remains unchanged; otherwise, the replacement of $\frac{e_i}{s_i}$ with $\frac{\pi}{2}\beta_{i1}\beta_{i2}$ is to be implemented.

Remark 4. Using the backstepping technique in controller design inevitably leads to the well-known “explosion of complexity,” caused by repeatedly differentiating virtual control inputs during each design step, which significantly increases computational load. To mitigate this issue, dynamic surface control (DSC) [4], [7] and command-filter-based approaches [10] were introduced. However, DSC does not compensate for the filtering error, resulting in only bounded—not asymptotic—performance. In the case of command filters, the control gain g_i must be known, which imposes stringent requirements on the system. Although a low-complexity framework has been developed for prescribed-performance control, it requires the nonlinear terms to satisfy a local Lipschitz condition, limiting the class of systems it can handle, and it cannot guarantee asymptotic convergence under this simplified structure.

IV. NUMERICAL SIMULATIONS

To verify the effectiveness of the proposed control approach, we consider the following model [2]:

$$\begin{cases} \dot{x}_1 = g_1(\bar{x}_1)x_2 + d_1(t) \\ \dot{x}_2 = f_2(\bar{x}_2) + g_2(\bar{x}_2)(kv + \bar{b}) + d_2(t) \\ y = x_1 \end{cases} \quad (33)$$

where $g_1(\bar{x}_1) = 1.2 + 0.2 \sin(x_1)$, $g_2(\bar{x}_2) = 1.4 + 0.1 \sin(x_1 x_2)$, $f_2(\bar{x}_2) = 0.5 \sin(x_1) x_2^2$, $k = 0.9 + 0.1 \sin(2t)$, and $\bar{b} = 0.5 + 1.2 \sin(2t)$.

The design parameters are chosen as $c_1 = 3$, $c_2 = 8$, $\varepsilon_1 = \varepsilon_2 = 0.005$, $\gamma_1 = \gamma_2 = 0.007$, $\delta_1 = \delta_2 = 100e^{-0.1t}$, $\delta_{p1} = \delta_{p2} = 100e^{-0.1t}$, $T = 2s$. The time-varying scaling function are chosen as $\beta_{11} = (1 - 0.5)e^{\frac{0.5t}{T-T}} + 0.5$, $0 \leq t < T$, $\beta_{11} = 0.5$, $t \geq T$; $\beta_{12} = (1 - 0.5)e^{\frac{0.5t}{T-T}} + 0.5$, $0 \leq t < T$, $\beta_{12} = 0.5$, $t \geq T$; $\beta_{21} = (1 - 0.5)e^{\frac{0.5t}{T-T}} + 0.5$, $0 \leq t < T$, $\beta_{21} = 0.5$, $t \geq T$; $\beta_{22} = (1 - 0.5)e^{\frac{0.5t}{T-T}} + 0.5$, $0 \leq t < T$, $\beta_{22} = 0.5$, $t \geq T$.

In this simulation, the reference trajectory is defined as $y_d = 0.5 \sin(t)$. The disturbance values are chosen as $d_1(t) = d_2(t) = 0.5 \sin(t)$. In addition, the initial estimates of the adaptive parameters are set to $\hat{\theta}_1(0) = \hat{\theta}_2(0) = 0.01$ and $\hat{D}_1(0) = \hat{D}_2(0) = 0.01$. The initial state is chosen as $x(0) = [0.5 \ 0.1]^T$.

The simulation outcomes shown in Figs. 1–4 verify the effectiveness of the proposed control strategy. Figs. 1 and 2 depict the behavior of the full-state errors (e_1) and (e_2), which enter the steady-state region within the predefined time ($T = 2s$) and remain confined within the prescribed performance bounds. In addition, the errors gradually converge to zero as (t) goes to infinity, aligning with the intended control objectives. Fig. 3 presents the trajectories of the reference signal (y_r) and the output (x_1). The gap between (x_1) and (y_r) diminishes notably over time, demonstrating the strong

tracking capability of the proposed controller. Finally, Fig. 4 displays the evolution of the control input (v).

As shown in the figures, every signal in the closed-loop system stays within bounded limits. The simulation outcomes, along with the accompanying analysis, confirm that the proposed control algorithm is capable of meeting the intended control goals.

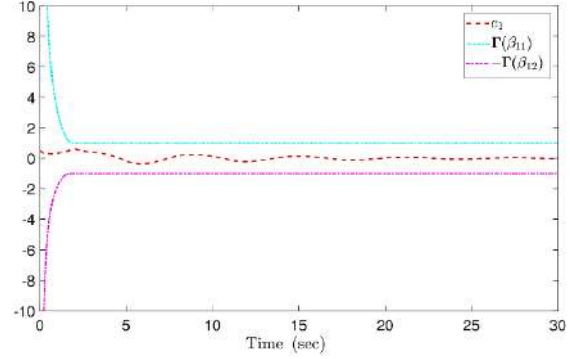


Fig. 1. Output tracking error e_1 .

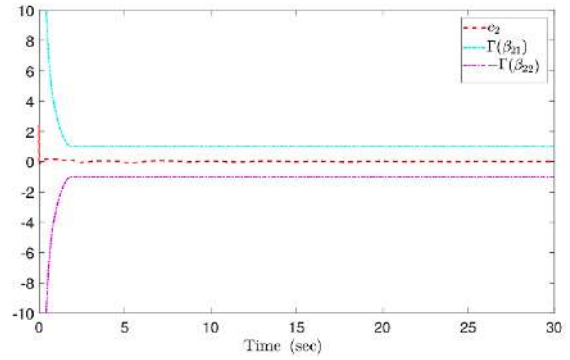


Fig. 2. Virtual control error e_2 .

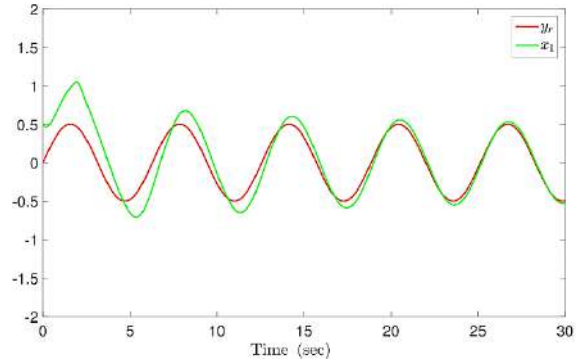


Fig. 3. State variable x_1 and Reference trajectory y_r .

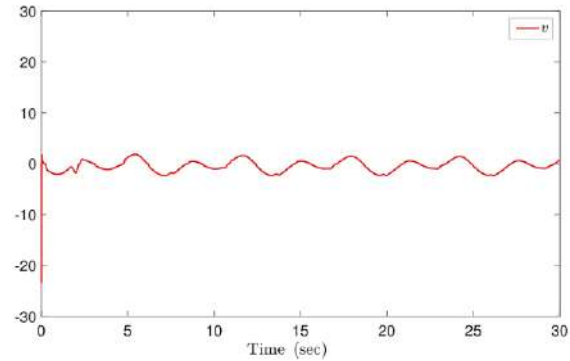


Fig. 4. Control inputs v .

V. CONCLUSIONS

The paper develops an adaptive backstepping control approach for nonlinear systems experiencing mismatched disturbances and actuator failure while enforcing full-state error constraints. A performance-constraint framework using tan-type and time-varying functions ensures constraint satisfaction independent of initial conditions, while a bounded estimation method guarantees asymptotic tracking and reduces restrictions on the reference input. Disturbance-bound estimation further enhances robustness. Simulation results verify the method's effectiveness, and future research will combine backstepping with adaptive dynamic programming for nonlinear optimization problems.

REFERENCES

- [1] M. Haddad, A. Boulkroune, and H. Li, "Practical finite-time fuzzy control of uncertain nonstrict-feedback systems with actuator saturation and output constraints," *International Journal of Adaptive Control and Signal Processing*, vol. 38, no. 5, pp. 1829–1853, 2024.
- [2] Y. Cao, J. Cao, and Y. Song, "Practical prescribed time tracking control over infinite time interval involving mismatched uncertainties and nonvanishing disturbances," *Automatica*, vol. 136, p. 110050, 2022.
- [3] A. Boulkroune, M. Haddad, and H. Li, "Adaptive fuzzy control design for nonlinear systems with actuation and state constraints: an approach with no feasibility condition," *ISA transactions*, vol. 142, pp. 1–11, 2023.
- [4] K. Zhao and Y. Song, "Removing the feasibility conditions imposed on tracking control designs for state-constrained strict-feedback systems," *IEEE Transactions on Automatic Control*, vol. 64, no. 3, pp. 1265–1272, 2019.
- [5] C. P. Bechlioulis and G. A. Rovithakis, "Robust adaptive control of feedback linearizable MIMO nonlinear systems with prescribed performance," *IEEE Transactions on Automatic Control*, vol. 53, no. 9, pp. 2090–2099, 2008.
- [6] C. P. Bechlioulis and G. A. Rovithakis, "A low-complexity global approximation-free control scheme with prescribed performance for unknown pure feedback systems," *Automatica*, vol. 50, no. 4, pp. 1217–1226, 2014.
- [7] S. Sui, C. L. P. Chen, and S. Tong, "A novel full errors fixed-time control for constraint nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 68, no. 4, pp. 2568–2575, 2023.
- [8] Z. Zhang, Y. Dong, and G. Duan, "Global asymptotic fault-tolerant tracking for time-varying nonlinear complex systems with prescribed performance," *Automatica*, vol. 159, p. 111345, 2024.
- [9] L. Li, K. Zhao, Y. Song, and F. L. Lewis, "On the uniformness of full state error prescribed performance for strict-feedback systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 54, no. 5, pp. 3022–3031, 2024.
- [10] Y.-X. Li, "Command filter adaptive asymptotic tracking of uncertain nonlinear systems with time-varying parameters and disturbances," *IEEE Transactions on Automatic Control*, vol. 67, no. 6, pp. 2973–2980, 2022.
- [11] K. Yong, M. Chen, Y. Shi, and Q. Wu, "Flexible performance-based robust control for a class of nonlinear systems with input saturation," *Automatica*, vol. 122, p. 109268, 2020.
- [12] H. Habibi, A. Yazdani, M. Darouach, H. Wang, T. Fernando, and I. Howard, "Observer-based sensor fault-tolerant control with prescribed tracking performance for a class of nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 68, no. 12, pp. 8259–8266, 2023.
- [13] C.-L. Zhang and G. Guo, "Prescribed performance fault-tolerant control of nonlinear systems via actuator switching," *IEEE Transactions on Fuzzy Systems*, vol. 32, no. 3, pp. 1013–1022, 2024.

Improving retinal fundus image segmentation: A Study of Transfer Learning and Single Channel Color Impact

Moussa GUENANI

LIST Laboratory

*University M'Hamed Bougara of
Boumerdes, Algeria
m.guenani@univ-boumerdes.dz*

Hamza AKROUM

LIST Laboratory

*University M'Hamed Bougara of
Boumerdes, Algeria
akroum@univ-boumerdes.dz*

khaled HARRAR

LIST Laboratory

*University M'Hamed Bougara of
Boumerdes, Algeria
khaled.harrar@univ-boumerdes.dz*

Asmala bin Ahmad

Intelligent Computing and Analytics Department (ICA)

Fakulti Teknologi Maklumat dan Komunikasi (FTMK)

Universiti Teknikal Malaysia Melaka (UTeM)

Melaka, Malaysia

asmala@utem.edu.my

Abstract—Segmentation of the retinal fundus image plays a crucial role in the diagnosis of various eye diseases. However, the complexity of retinal structures and the variability in image quality pose significant challenges. This paper investigates the effectiveness of transfer learning and single-channel color processing to enhance the segmentation of the retinal fundus image. Experiments were conducted using the U-Net model with pre-trained ResNet50 as the encoder, along with various single channel and color images from IDRiD dataset, to evaluate their effect on segmentation accuracy. Our results highlight that transfer learning significantly improves segmentation accuracy in the two cases (single channel and color images) compared to training from scratch, especially for a limited dataset. The highest test Dice coefficients achieved were 0.736 for Exudates (EX), 0.454 for Hemorrhages (HE), 0.249 for Microaneurysms (MA), 0.967 for the Optic Disc (OD), and 0.470 for Soft Exudates (SE). Similarly, the Intersection over Union (IoU) values were 0.582 for EX, 0.296 for HE, 0.142 for MA, 0.936 for OD, and 0.313 for SE. In addition, we found that single channel color processing can provide advantages in certain cases, depending on the relevance of the color channel to the segmentation task.

Keywords—Fundus Photography, Transfer Learning, Retinal Fundus Segmentation, Single-Channel Color, Deep Learning, Medical Imaging

I. INTRODUCTION

Accurate segmentation of the lesions and optic disc in retinal images is crucial for the early detection and diagnosis of various eye diseases such as diabetic retinopathy, glaucoma, and age-related macular degeneration. Identifying these anatomical structures can provide valuable information to ophthalmologists to monitor disease progression and guide treatment decisions.

However, segmentation of the lesions and the optic disc is a challenging task due to the complex and variable appearance

of these structures in the retinal images. Factors such as image quality, lighting conditions, and the presence of other abnormalities can make it difficult to accurately delineate these regions.

The primary objective of this study is to evaluate the segmentation performance of fundus images by analyzing the individual contributions of the red, green, and blue channels compared to the full RGB image. This involves fine-tuning a pre-trained model that has been trained on ImageNet [1] dataset and assess how effectively this approach enhances segmentation performance for lesions and optic disc segmentation in fundus images.

Additionally, the study will compare the results of using pre-trained models with those trained from scratch using randomly initialized weights, and determine the benefits of transfer learning in this specific context. Segmentation accuracy will be measured using the Dice coefficient, allowing us to identify the configuration that provides the best performance.

Using U-Net, a CNN-based model, this study aims to develop a precise segmentation system capable of reliably identifying lesions and the optic disc, even when confronted with various imaging artifacts and pathologies.

The remainder of this study is organized as follows: Section II reviews related works in retinal image segmentation, Section III describes the methodology including dataset, transfer learning approach, and evaluation metrics, Section IV presents the experiments and the experimental results, Section V discusses the significance of the findings, and finally Section VI concludes the study and proposes directions for future research.

II. RELATED WORKS

Traditional methods for segmentation of retinal fundus images relied on techniques such as thresholding, edge detection, and region growth. These methods were often combined with morphological operations to refine the segmentation results.

Convolutional Neural Networks (CNNs) have exhibited substantial success across diverse medical image analysis applications, particularly in the segmentation of anatomical structures and pathological regions.

Recent advances in deep learning have revolutionized the field, with CNNs widely used to achieve higher accuracy and robustness in segmenting structures such as blood vessels, optic discs, and lesions.

Architectures such as U-Net [2], V-Net [3], and their variants have shown significant improvements over traditional methods, driven by their ability to learn hierarchical features and capture complex patterns in retinal images.

Several works have demonstrated the efficacy of CNN in fundus image analysis. Kamrul et al. [4] introduced a novel model for optic disc (OD) and fovea segmentation called DR-Net. For OD segmentation, DRNet achieved mean Intersection over Union (mIoU) scores of 0.845, 0.901, 0.933, and 0.920 on the IDRiD, RIMONE, DRISHTI-GS, and DRIVE datasets, respectively.

Maysanjaya et al. [5] employed Mask R-CNN for segmenting the Optic Disc and exudates in Retinal Fundus Images. Their method achieved IoU values of 0.8431 and 0.9933 for OD and exudates, respectively, on the IDRiD dataset.

III. METHODS

A. Dataset

This research utilizes the publicly accessible Indian Diabetic Retinopathy Image Dataset (IDRiD) [6], comprising 81 color fundus images captured with a 50-degree Field of View (FOV) and corresponding expert-annotated segmentation masks. The dataset is divided into 54 images for training and 27 images for testing, with each image having a resolution of 4288×2848 pixels. An example of these fundus images is shown in Fig. 1. The dataset includes ground-truth masks for five classes: Microaneurysms (MA), Haemorrhages (HE), Hard Exudates (EX), Soft Exudates (SE), and Optic Disc (OD).

B. Transfer Learning

Transfer learning represents a machine learning methodology in which a model initially undergoes training on an extensive dataset tailored for a particular task. Following this, the weights learned from the pre-trained model are leveraged to facilitate the training process for an alternative task [7]. Consequently, this approach is advantageous because it conserves both time and resources, given the resource-intensive nature of training new machine learning models.

C. Evaluation metrics

The proposed system is evaluated using Dice coefficient, Jaccard (also known as the Intersection over Union, IoU),

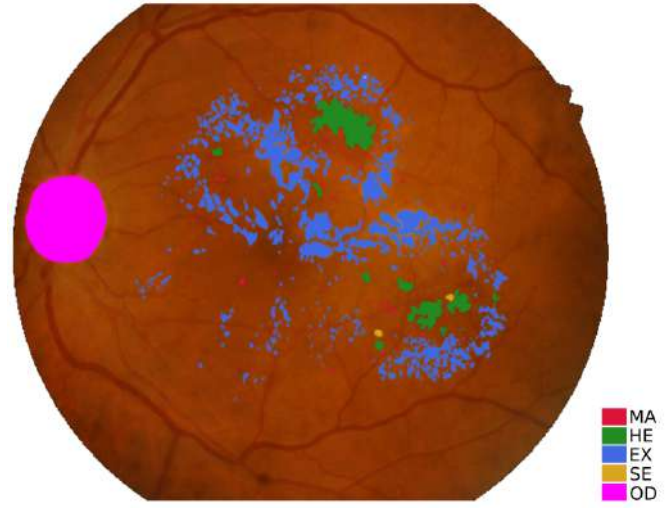


Fig. 1: Illustration of the retinal image from IDRiD [6] superimposed with the corresponding segmentation masks.

and accuracy metrics, with specific definitions provided in the accompanying equations.

$$\text{Dice coefficient} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \quad (1)$$

$$\text{Jaccard coefficient} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (2)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (3)$$

Where:

- **TP** = True Positives: Both the predicted mask and the ground truth mask exhibit a pixel intensity of 1.
- **FP** = False Positives: The predicted mask shows a pixel intensity of 1, whereas the ground truth mask shows a pixel intensity of 0 in that region.
- **FN** = False Negatives: The predicted mask has a pixel intensity of 0, while the ground truth mask has a pixel intensity of 1.
- **TN** = True Negatives: Both the predicted mask and the ground truth mask have a pixel intensity of 0.

D. Proposed technique

The proposed technique for segmenting retinal images encompasses data preprocessing and partitioning, training and fine-tuning a deep neural network, followed by the evaluation and testing of the algorithm. Fig 2 presents a comprehensive overview of the model technique.

1) **Data Preprocessing and Splitting:** Fundus images generally have a black background that does not provide any relevant information on the region of interest (ROI) as shown in Fig. 3. Therefore, we have eliminated this non-contributory black background from the images.

Subsequently, a Gaussian filter was applied to the images to eliminate high-frequency noise components. In addition, the images were resized to dimensions of 512×512 pixels to lower

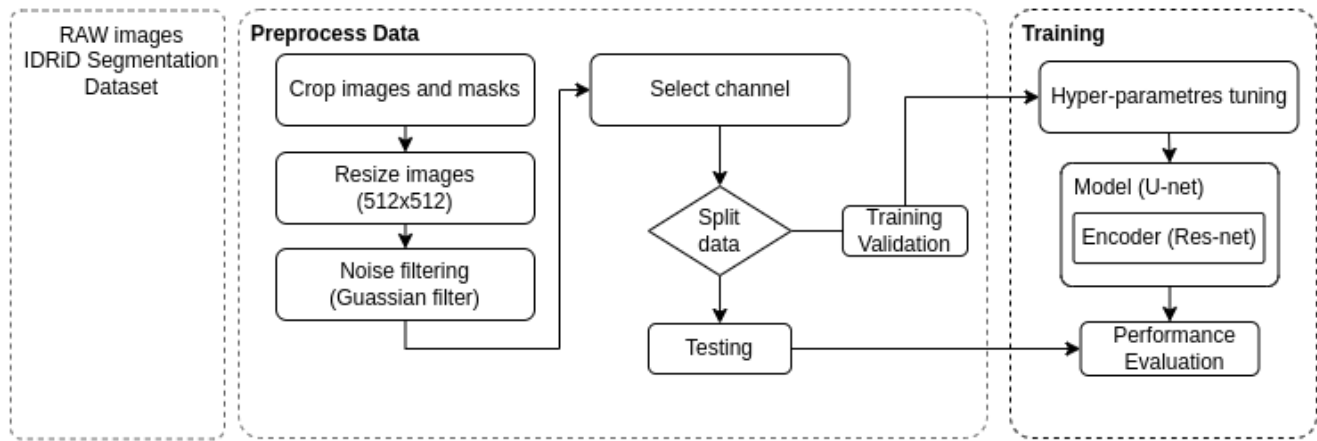


Fig. 2: schematic representation of proposed methodology

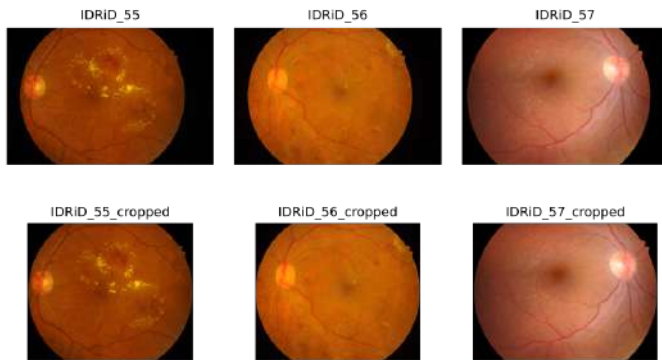


Fig. 3: Retinal images obtained prior to and following the cropping of the region of interest (ROI).

the computational expenses associated with processing, then, one of the available channels was selected for further analysis: either an full RGB image or a single-channel (R, G, B), finally normalizing the images by a factor of $1/255$ to scale values between 0 and 1.

The original IDRiD dataset is initially divided into test and training sets using a random 2:8 ratio. To maintain consistency with the existing literature for comparison purposes, we preserve the test set in its original form. The training subset is further divided into training and validation sets. This additional split allows us to utilize the validation portion during the hyper-parameter tuning phase, ensuring a rigorous optimization process.

- **Training set:** This subset is utilized to train the model, focusing on the optimization of the learnable parameters. It constitutes the largest segment of the dataset, thereby ensuring the model is provided with an sufficient number of examples to learn effectively and generalize accurately. In our case, the training data is composed of an 80% portion of the original training set.
- **Validation set:** The validation subset is essential for the

refinement of hyperparameters and serves to alleviate overfitting by providing an ongoing evaluation of the model's performance throughout the training phase. This continues evaluation is critical in determining the model's ability to generalize to new, unseen data following each epoch of training. In this specific study, 20% of the original training set has been allocated for validation purposes.

- **Testing set:** This particular subset is reserved exclusively for the final phase of model evaluation. To maintain an objective and unbiased assessment of the model's performance, it is never exposed to the model during the training or validation stages. Consequently, it provides a 20% portion of the entire dataset.

2) **Model selection and adaptation:** The model implemented in this research is Unet [2], which is widely recognized for its high efficacy in the domain of medical image analysis, primarily due to its encoder-decoder architecture, as shown in Fig. 4. The encoder pathway captures contextual information through successive convolutional and pooling layers, while the decoder pathway enables precise localization through upsampling and concatenation with corresponding encoder features.

ResNet50 as detailed by He et al. [8] was used in this study as the encoder component of the proposed system, having initially been trained on the extensive ImageNet dataset [1]. The use of ResNet50 provides several advantages, including its deep architecture with residual connections that facilitate gradient flow during training and its ability to extract rich hierarchical features from input images.

Following the pre-training phase, a fine-tuning process was performed to adapt the pretrained model to align precisely with our specific goal of retinal image segmentation. The fine-tuning involved unfreezing the encoder layers and training the entire network end-to-end with a reduced learning rate to preserve the learned ImageNet features while adapting them to the retinal imaging domain.

The decoder portion of the U-Net was constructed with

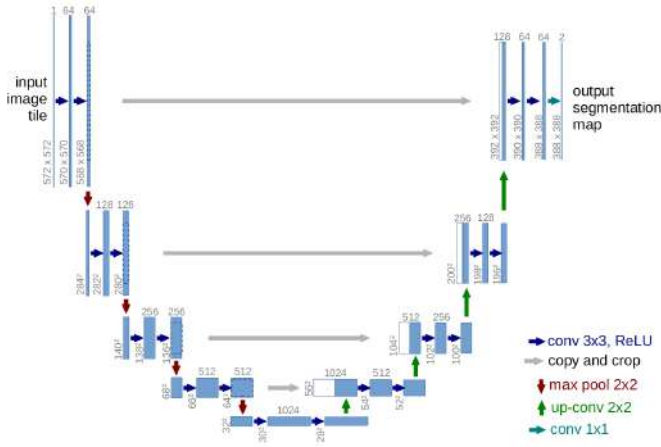


Fig. 4: U-Net network structure [2].

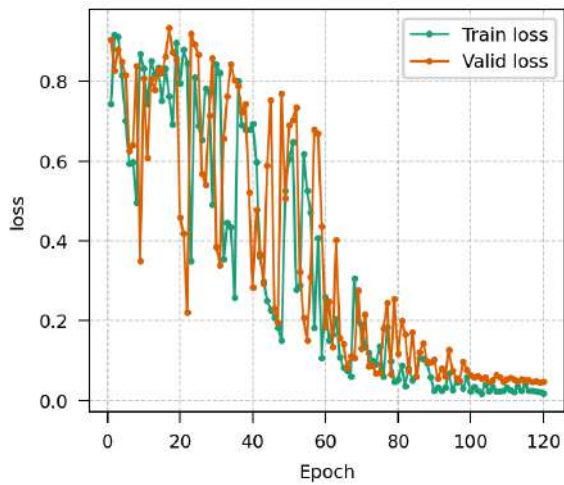


Fig. 5: Train and Validation loss

symmetric upsampling blocks that progressively reconstruct the spatial resolution while incorporating skip connections from the encoder. Each decoder block consists of upsampling layers followed by convolutional layers with ReLU activation. The final layer uses a sigmoid activation function to produce pixel-wise probability maps for each segmentation class.

IV. EXPERIMENTS AND RESULTS

A. Experimental Setup

This study was conducted utilizing Kaggle [9], an online cloud computing platform that provides complimentary access to GPU resources. The Python 3.10.13 programming language was used in conjunction with the PyTorch framework [10], and the network was optimized using Adam optimizer [11]. The complete set of hyperparameters used in the experiments is detailed in Table I. Fig 5 illustrates training and validation losses during the training phase.

B. Training and fine tuning

Various experiments were performed using diverse configurations to examine the effect of each channel on segmentation

TABLE I: Configuration settings for neural network training

Hyperparameter	Value
Batch Size	4
Learning Rate	0.0001 and 0.0005
Optimizer	Adam
Loss Function	Dice loss
Number of Epochs	120
Input Image Size	512×512 pixels
Backbone	ResNet50
Activation Function	Sigmoid

accuracy, totaling 80 experiments.

Initially, we train the deep neural network from scratch by randomly setting the weights and employing single-channel; red channel, green channel, blue channel, and finally, all RGB components of the full-color image. In the subsequent phase, the experiments are repeated, but this time, utilizing pre-trained weights. The outcomes of these experiments are discussed in the following subsection.

C. Results

Upon completing the training phases, we evaluated the deep neural network using the test set. We gathered evaluation metrics and a model predicted masks.

The following are both quantitative and qualitative results: for the initial phase, comprehensive results are presented in Table III, while for the subsequent phase they are displayed in Table II. The predicted segmentation masks are shown in Fig 6.

The findings indicate that utilizing the transfer learning approach is beneficial, resulting in improved testing metrics for all retinal features, regardless of the image channel employed.

The model exhibited consistently reliable performance when trained using either the green channel or the full RGB image, applicable both in transfer learning contexts and when trained from the ground up. In the transfer learning scenario, the performance variation was minimal, the improvement in dice coefficient remaining below 1.5% for EX, and MA. Slight increases in variability were detected when training from scratch, with values reaching 5.4%, 3.4%, and 1.2% dice coefficient improvement for EX, HE, and SE, respectively.

Remarkably, employing solely the green channel yielded superior test metrics compared to the complete RGB image. This was evident in scenarios such as (MA and EX when using transfer learning) and (EX, HE, and SE when trained from scratch).

V. DISCUSSION

We anticipated that leveraging pre-trained weights would enhance the model's performance by facilitating faster convergence compared to training from scratch, owing to the comprehensive prior training these models underwent on the large-scale ImageNet dataset. Moreover, the features learned from the initial task can be effectively adapted to RGB images or any of their individual components—Red, Green, and Blue. The experimental results confirm our hypothesis, demonstrating consistent improvements in segmentation performance

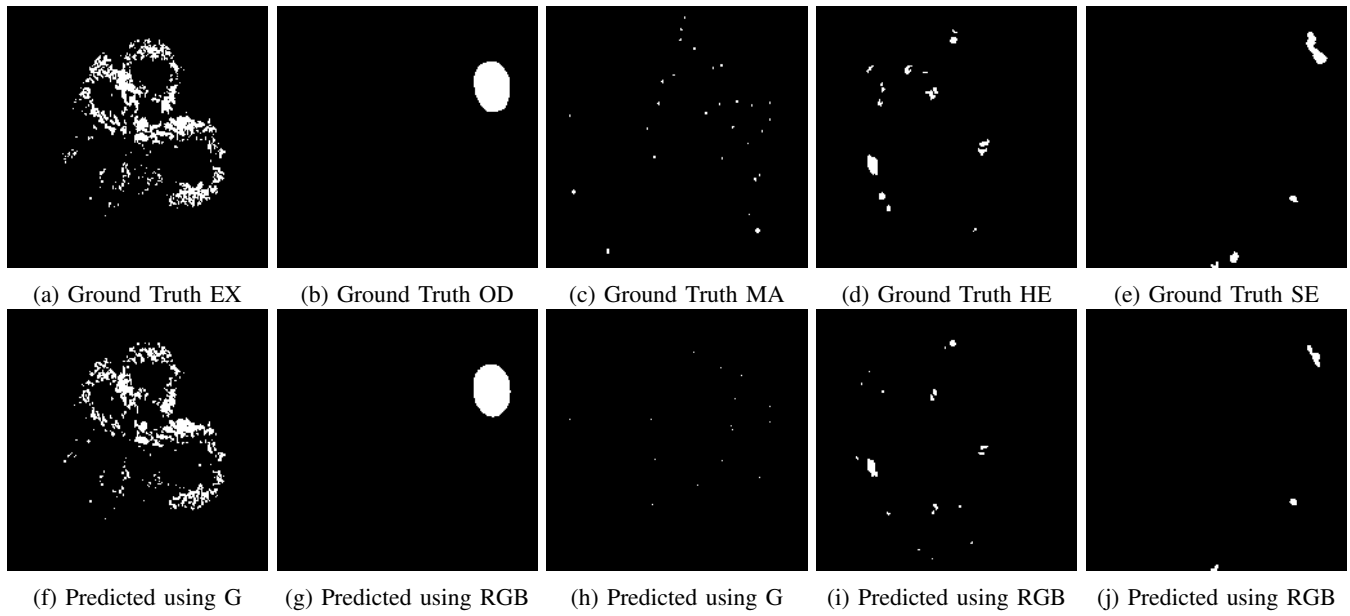


Fig. 6: Comparison between Ground Truth (a, b, c, d and e) and Predicted Results (f, g, h, i and j) for both the optic disc and lesions is conducted using the top-performing models, which are emphasized in bold in Table II.

TABLE II: The outcomes of model performance when employing transfer learning

Retinal features	Channel	Test Accuracy	Test Dice	Test IoU
EX	B	98.919830	0.643255	0.474263
	G	99.133286	0.735691	0.582340
	R	99.069717	0.685033	0.521620
	RGB	99.115875	0.720750	0.563887
HE	B	98.463272	0.044064	0.022581
	G	98.740570	0.428965	0.277324
	R	98.555557	0.166336	0.091905
	RGB	98.781815	0.454582	0.296304
MA	B	99.793915	0.141720	0.076491
	G	99.810860	0.248940	0.142422
	R	99.807533	0.192246	0.106556
	RGB	99.810989	0.240207	0.136720
OD	B	99.626282	0.913289	0.841285
	G	99.792824	0.951880	0.908260
	R	99.799637	0.956538	0.916990
	RGB	99.852699	0.967027	0.936181
SE	B	99.763954	0.128243	0.074781
	G	99.810356	0.439417	0.284854
	R	99.712059	0.057328	0.029953
	RGB	99.819420	0.470113	0.312556

TABLE III: The outcomes observed when employing random weights

Retinal features	Channel	Test Accuracy	Test Dice	Test IoU
EX	B	98.740715	0.564514	0.399784
	G	99.073647	0.687551	0.523905
	R	98.830826	0.594457	0.424554
	RGB	98.964195	0.633946	0.465038
HE	B	95.914955	0.035800	0.018519
	G	98.708786	0.424731	0.272567
	R	98.500046	0.079508	0.041570
	RGB	98.747314	0.391491	0.245442
MA	B	99.802956	0.113447	0.060439
	G	99.809036	0.040559	0.020722
	R	99.810143	0.022847	0.011562
	RGB	99.811234	0.183781	0.101206
OD	B	99.528770	0.892947	0.807631
	G	99.650871	0.914043	0.842439
	R	99.750664	0.940439	0.888074
	RGB	99.802017	0.953445	0.911240
SE	B	99.682732	0.101975	0.058523
	G	99.709579	0.255494	0.159687
	R	99.712059	0.057328	0.029953
	RGB	99.768684	0.243167	0.147961

across all retinal features when using transfer learning. The most substantial improvements were observed for the Optic Disc (OD) segmentation, achieving a Dice coefficient of 0.967 with transfer learning compared to 0.953 when training from scratch using RGB images. This represents a 1.4% improvement, which is clinically significant for automated diagnosis systems. For lesion segmentation, the benefits of transfer learning were even more pronounced. Exudates (EX) showed improvement from 0.687 to 0.736 Dice coefficient when using

the green channel with transfer learning, representing a 4.9% relative improvement. Similarly, Hemorrhages (HE) improved from 0.391 to 0.455 Dice coefficient with RGB transfer learning, a 6.4% relative improvement.

The comparable or superior performance of the model when using the green channel compared to the entire RGB image is particularly noteworthy and can be attributed to several factors. First, the green channel provides the best contrast for retinal vascular structures and many pathological features

in fundus images due to the optical properties of blood and retinal tissues. Second, using a single channel reduces the model complexity and potential noise from irrelevant color information, which may lead to more focused feature learning. Third, the green channel contains maximal hemoglobin absorption characteristics, making it particularly suitable for detecting hemorrhages and microaneurysms. The red channel showed moderate performance across most tasks, while the blue channel consistently produced the poorest results. This is expected given that blue light penetrates less deeply into retinal tissue and provides lower contrast for most retinal structures of interest. For Microaneurysms (MA), which represent the smallest and most challenging features to segment, the Dice coefficients remained relatively low across all configurations (0.142-0.249 with transfer learning). This is consistent with previous literature and reflects the inherent difficulty of detecting these small lesions, which often span only a few pixels in the resized images. The high accuracy values (>98% for all configurations) might appear misleading at first glance, but this is due to the class imbalance inherent in medical image segmentation tasks where pathological regions occupy a small percentage of the total image area. Therefore, Dice coefficient and IoU metrics provide more meaningful assessments of segmentation performance. An interesting observation is that transfer learning provides more consistent benefits across different color channels compared to random initialization. This suggests that the pre-trained features from ImageNet, despite being trained on natural images rather than medical images, capture generalizable low-level and mid-level features (edges, textures, shapes) that are beneficial for retinal image analysis. The limitations of this study include the relatively small dataset size (81 images total), which makes the model susceptible to overfitting despite the use of transfer learning. Additionally, the resizing of images to 512×512 pixels may result in loss of fine details important for detecting small lesions like microaneurysms. Future work should explore data augmentation techniques, ensemble methods, and attention mechanisms to further improve performance, particularly for challenging lesion types.

VI. CONCLUSION

This study investigated the effectiveness of transfer learning and single-channel color processing for improving retinal fundus image segmentation using U-Net with ResNet50 encoder. Through comprehensive experiments on the IDRiD dataset, we demonstrated that transfer learning consistently outperforms training from scratch across all retinal features, achieving Dice coefficients of 0.736 for Exudates, 0.454 for Hemorrhages, 0.249 for Microaneurysms, 0.967 for Optic Disc, and 0.470 for Soft Exudates. Our findings reveal that the green channel alone can match or exceed the performance of full RGB images for several segmentation tasks, suggesting that careful channel selection based on the target anatomical structure can improve model efficiency and accuracy. This has important practical implications for reducing computational costs and potentially improving real-time clinical applications. The su-

perior performance of transfer learning, even with pre-trained weights from non-medical images, confirms that generic visual features learned from large-scale natural image datasets can be effectively adapted for medical image analysis tasks. This approach is particularly valuable when working with limited medical imaging datasets, where training deep networks from scratch is challenging. Future research directions include: (1) exploring advanced architectures such as Transformer-based models and attention mechanisms to better capture long-range dependencies in retinal images; (2) implementing multi-task learning frameworks that simultaneously segment multiple retinal features; (3) investigating domain-specific pre-training on larger retinal image datasets; (4) developing ensemble methods that combine predictions from different color channels; (5) applying the methodology to other retinal imaging modalities and pathologies; and (6) validating the approach on larger, multi-center datasets to assess generalization capabilities across different imaging devices and populations.

ACKNOWLEDGMENT

The authors express their gratitude to ophthalmologist Boussouar Omar for his invaluable assistance and insightful explanations throughout the course of this research. His expertise made a significant contribution to the understanding and interpretation of the medical aspects of this work. The authors are also deeply appreciative of the database providers for supplying the IDRiD segmentation database.

REFERENCES

- [1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.
- [2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18*, pp. 234–241, Springer, 2015.
- [3] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016.
- [4] M. K. Hasan, M. A. Alam, M. T. E. Elahi, S. Roy, and R. Martí, "Drnet: Segmentation and localization of optic disc and fovea from diabetic retinopathy image," *Artificial Intelligence in Medicine*, vol. 111, p. 102001, 2021.
- [5] I. M. Dendi Maysanjaya, M. W. Antara Kesiman, and I. G. Diatri Indradewi, "Optic disc and exudates segmentation on retinal fundus images using mask r-cnn," in *2022 International Conference on Information Technology Research and Innovation (ICITRI)*, pp. 168–172, 2022.
- [6] P. Porwal, S. Pachade, M. Kokare, G. Deshmukh, J. Son, W. Bae, L. Liu, J. Wang, X. Liu, L. Gao, T. Wu, J. Xiao, F. Wang, B. Yin, Y. Wang, G. Danala, L. He, Y. H. Choi, Y. C. Lee, S.-H. Jung, Z. Li, X. Sui, J. Wu, X. Li, T. Zhou, J. Toth, A. Baran, A. Kori, S. S. Chennamsetty, M. Safwan, V. Alex, X. Lyu, L. Cheng, Q. Chu, P. Li, X. Ji, S. Zhang, Y. Shen, L. Dai, O. Saha, R. Sathish, T. Melo, T. Araújo, B. Harangi, B. Sheng, R. Fang, D. Sheet, A. Hajdu, Y. Zheng, A. M. Mendonça, S. Zhang, A. Campilho, B. Zheng, D. Shen, L. Giancardo, G. Quellec, and F. Mériaudeau, "IDRiD: Diabetic Retinopathy - Segmentation and Grading Challenge," *Medical Image Analysis*, vol. 59, p. 101561, Jan. 2020.
- [7] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021.

- [8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [9] "Kaggle." <https://www.kaggle.com/>.
- [10] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.
- [11] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Jan. 2017. arXiv:1412.6980 [cs].

Open- and Short-Circuit Faults Detection in Multicellular Inverters using Sliding Mode Observer.

Fatma Khater

Laboratory of Applied Automatic, dept. of Automation
University M'hamed Bougara
Boumerdes, Algeria
f.khater@univ-boumerdes.dz

Hamza Akroum

University M'hamed Bougara
Boumerdes, Algeria
akroum@gmail.com

Aimad Boudouda

University M'hamed Bougara
Boumerdes, Algeria
a.boudouda@univ-boumerdes.dz

Abderrezak Aibeche

Laboratory of Applied Automatic, dept. of Automation
University M'hamed Bougara
Boumerdes, Algeria
aibeche_umbb@univ-boumerdes.dz

Mohamed Zinelabine Doghmne

University M'hamed Bougara
Boumerdes, Algeria
doghmane_m@yahoo.com

Abstract— Multicellular inverters play a key role in industrial systems, where ensuring high reliability, operational safety, and effective fault-diagnosis capabilities is essential for protecting power conversion equipment and minimizing downtime and maintenance costs. The overall performance and robustness of these converters are particularly dependent on the accurate detection and diagnosis of faults in insulated gate bipolar transistor (*IGBT*) switches, which constitute the most critical components of the power stage. This study aims to analyze the operational behavior of a multicellular inverter under both normal and faulty conditions, with particular emphasis on open-circuit and short-circuit faults affecting the converter switches. To address this objective, the paper proposes a fault diagnosis approach based on a sliding-mode observer designed to detect faults in the power switches of a multicellular inverter.

Keywords— multicellular inverter, faults diagnosis, open-circuit faults, short-circuit faults.

I. INTRODUCTION

The multicell inverter topology is a relatively recent development, first introduced in the 1990s. It is based on a cascade of commutation cells interconnected through floating voltage sources, typically implemented using capacitors [1], [2]. Multicellular inverters display nonlinear characteristics resulting from their switching behavior, creating challenges for control design and robust stability analysis [3]. These converters are widely employed in both industrial applications and academic research, motivating the development of advanced control strategies to guarantee reliable operation. Numerous studies have investigated different control and observation techniques for multicellular converters.

The reliability of multicellular inverters is critical, as the failure of a single component can lead to significant system malfunction. Consequently, effective fault diagnosis is indispensable for enabling fault-tolerant control strategies in these systems. Notably, more than 30% of converter failures originate from power semiconductor devices, particularly insulated gate bipolar transistors (*IGBTs*), highlighting the importance of accurate and timely fault detection. These failures

can be categorized as open-circuit (*OC*) and short-circuit (*SC*) faults [4],[5].

The remainder of this manuscript is organized as follows, after a brief introduction, section 2 presents an overview of the multicellular inverter model. Section 3 discusses the control strategy applied to the inverter and evaluates its performance. The proposed fault detection method is introduced in Section 4, and concluding remarks are provided in Section 5.

II. MULTICELL INVERTER MODEL

Fig. 1 presents a representative diagram of a floating-capacitor multicellular inverter supplying a load composed of a resistance R and an inductance L . The *DC* input voltage is denoted by E . This inverter operates as a hybrid system, where continuous variables (voltages and currents) coexist and interact with discrete variables represented by the binary switching signals.

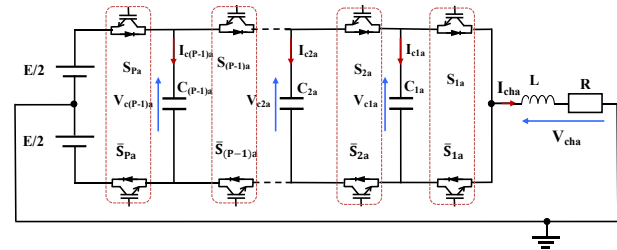


Fig. 1. Representative diagram of multicellular inverter.

There are $p-1$ floating voltage sources [6], [7]. There are $n_{\text{cells}} + 1$ voltage levels available at the output.

Each cell is made up of two adjacent floating capacitors, V_{Ck} and V_{Ck-1} , with $V_{C0}=0$ and $V_{Cp}=E$.

$$V_{\text{Cell}_k} = V_{Ck} - V_{Ck-1} \quad (1)$$

An equal distribution of voltage constraints must be applied to each floating capacitor, i.e.,

$$V_{Ck} = k.E/p \quad (2)$$

The voltage applied to the load is given by:

$$V_{ch} = \sum_{k=1}^P S_k \cdot V_{cell\_k} \quad (3)$$

The voltage across capacitor C_k is related to the current by:

$$i_{Ck}(t) = (S_{cell_{k+1}} - S_{cell_k}) \cdot i_{ch}(t) \quad (4)$$

$$\frac{dV_{Ck}}{dt} = \frac{(S_{cell_{k+1}} - S_{cell_k})}{C_k} \cdot i_{ch}(t) \quad (5)$$

The load voltage is given by:

$$V_{ch}(t) = \sum_{k=1}^P (V_{Ck}(t) - V_{C_{k-1}}(t)) \cdot S_{cell_k} \quad (6)$$

By grouping these expressions in the form of equations of states, we obtain:

$$\begin{aligned} \frac{d}{dt} V_{C1} &= \frac{[S_2 - S_1]}{C_1} \cdot i_{ch} \\ \frac{d}{dt} V_{C2} &= \frac{[S_3 - S_2]}{C_2} \cdot i_{ch} \\ &\vdots \\ \frac{d}{dt} V_{C_{p-1}} &= \frac{[S_p - S_{p-1}]}{C_{p-1}} \cdot i_{ch} \\ \frac{d}{dt} i_{ch} &= \frac{[S_1 - S_2]}{L} \cdot V_{C1} + \frac{[S_2 - S_3]}{L} \cdot V_{C2} \\ &+ \dots + \frac{[S_{p-1} - S_p]}{L} \cdot V_{C_{p-1}} - \frac{R}{L} \cdot i_{ch} + \frac{S_p}{L} \cdot E \end{aligned} \quad (7)$$

Based on the given differential equations, we construct a state representation with the state variables being the floating voltages V_{ck} and the load current i_{ch} , which has the following form:

$$\dot{X} = AX + BU \quad (8)$$

With: $X = [V_{C1}, V_{C2}, \dots, V_{C_{p-1}}, i_{ch}]^T$; U : the control vector.

III. CONTROL OF THE MULTICELL INVERTER

The control system of a multicellular inverter must ensure the regulation of output currents and voltages while maintaining an even distribution of stress across each switch. Various control approaches have been developed and documented in the literature for these types of inverters. Pulse Width Modulation (*PWM*) control, an open-loop method often referred to as natural control, ensures the correct operation of the inverter and the balancing of capacitor terminal voltages. However, closed-loop control strategies are essential for ensuring stability, robustness, and continuous proper operation of the inverter. One such strategy is the application of sliding mode control, which

considers the evolution of output voltages, currents, and capacitor voltages [8],[9].

Sliding mode control (*SMC*) is a type of nonlinear controller that is a powerful design method for many systems, based on Lyapunov's method. It is considered an effective approach in control systems due to its robustness against model uncertainties and external disturbances. The objective of control through sliding modes can be defined in two key points: first, creating a surface S such that all system trajectories follow the desired tracking behavior, regulation, and stability; and second, determining a control law U that attracts and retains all state trajectories on the sliding surface [10].

We consider the system defined by equation (8), where the dimension of the surface vector S matches that of the control vector U . To ensure the convergence of a state variable x towards its reference values x_{ref} , various forms of sliding surfaces have been introduced. Typically, the selection of a surface is based on the error in the controlled variable denoted as $e(x)$, where $e(x) = x_{ref} - x$. Thus, we define p sliding surfaces as follows:

$$\begin{aligned} S_1 &= V_{C_{ref1}} - V_{C1} \\ S_2 &= V_{C_{ref2}} - V_{C2} \\ &\vdots \\ S_p &= i_{Ch_{ref}} - i_{ch} \end{aligned} \quad (9)$$

In order to verify the convergence condition, we use the Lyapunov approach. The Lyapunov's function is chosen as follows:

$$V(x) = \frac{1}{2} S^2(X) \quad (10)$$

To ensure the decrease of the Lyapunov function, it is sufficient to guarantee that its derivative is negative.

$$\dot{V}(x) = \dot{S}(X) \cdot S(X) < 0 \quad (11)$$

The structure of the sliding mode controller consists of two components $U = U_{eq} + U_n$.

U_{eq} enables the maintenance of the controlled variable on the sliding surface $S = 0$. This command is deduced by considering that the derivative of the surface is zero. U_n is determined to satisfy the convergence condition. During the sliding surface and the steady state, the surface is zero, from which we deduce the expression of U_{eq} and U_n .

$$U_{eq} = - \left[\frac{\partial S}{\partial X} \cdot B(X) \right]^{-1} \cdot \left[\frac{\partial S}{\partial X} \cdot A(X) \right] \quad (12)$$

$$U_n = K_i \text{sign}(S(X)) \quad (13)$$

The values of the examined inverter parameters are summarized in Table 1.

TABLE I. OPERATING PARAMETERS OF THE MULTICELLULAR INVERTER.

Parameter	Description	Values
V_{in} [V]	Input DC voltage	2000
C_i, j [F]	Capacitor	$3e^{-4}$
$R_{ch, j}$ [Ω]	Load resistance	5
$L_{ch, j}$ [H]	Load inductance	$5e^{-2}$
f_{PWM} [kHz]	The switching frequency	5.5
T_e [sec]	The switching period	$1e^{-6}$

The following Fig. respectively illustrate the load voltage, load current, and floating capacitor voltage of the multicellular inverter controlled by Sliding Mode Control (SMC).

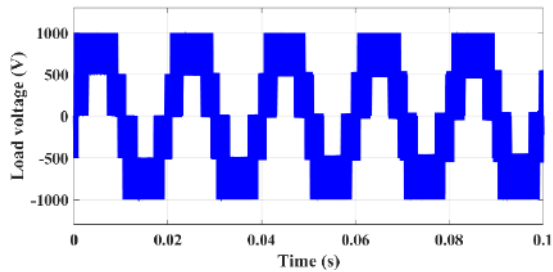


Fig. 2. Load voltage of the multicellular inverter.

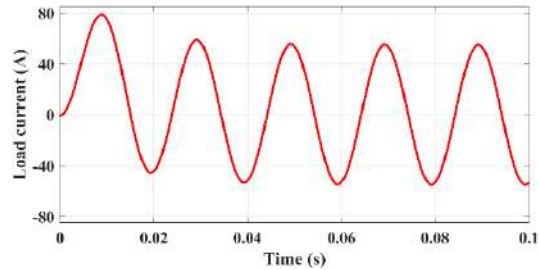


Fig. 3. Load current of the multicellular inverter.

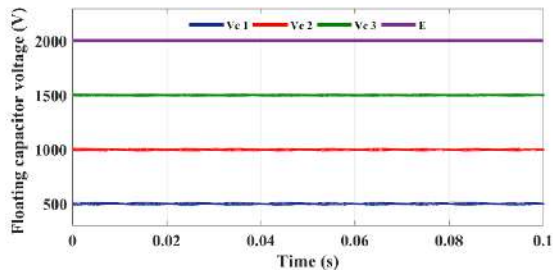


Fig. 4. Floating capacitor voltage of the multicellular inverter.

The simulation results of this control strategy applied to the studied converter, shown in Fig. 2,3 and 4, show that the output voltages and currents reach their desired reference values, as well as an equitable distribution of the voltages at the terminals of the floating capacitors.

IV. FAULTS DETECTION APPROACH BASED ON SLIDING MODE OBSERVER

Under normal (fault-free) operating conditions, as previously described, the output currents and voltages achieve their reference values. However, when a fault occurs in the IGBTs of the multicellular converter, the system performance deteriorates, resulting in distortions in the output voltages and currents, as well as fluctuations in the capacitor voltages.

The basic idea of using an observer for fault detection is to estimate the system outputs from available measurements and then generate a residual by evaluating the difference between the estimated and actual outputs. This residual is analyzed using an adaptive threshold to assess the presence of faults. Finally, specific decision rules are applied to determine whether a fault has occurred [11], [12]. This observer generates parameter estimates that serve as detection signals, using a developed mathematical model's input and output signals. These estimated signals are compared with their regular counterparts to identify their differences. The detected differences act as signatures for fault detection [13].

In this section, we analyze the impact of faults occurring in multicellular inverters. More specifically, the sliding mode observer is implemented to detect both open-circuit and short-circuit faults and to assess their influence on the inverter's dynamic behavior. The multicellular inverter is simulated in the MATLAB/Simulink 2018a environment using the parameter values summarized in Table 1.

A. Case one: Open-circuit faults

The Fig. 5,6 and 7. shows the output voltages and currents during an open-circuit (OC) fault on the IGBT of the inverter at $t = 0.05$ s. The Fig. exhibits clear distortions in both the output voltage and current waveforms. Moreover, the voltage of the floating capacitor associated with the defective cell becomes noticeably distorted at the moment the fault occurs.

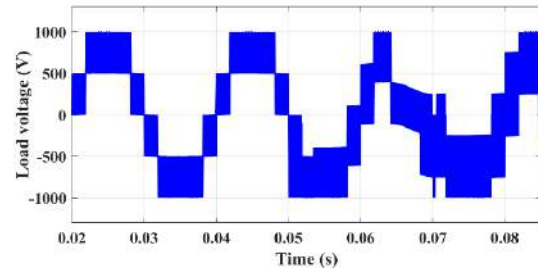


Fig. 5. The load voltages during the appearance of an OCF.

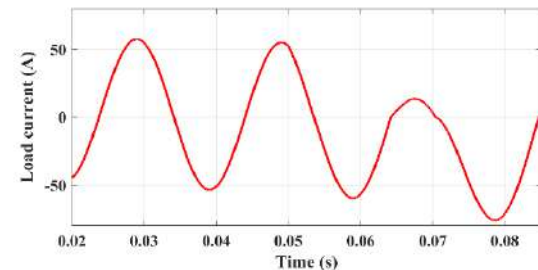


Fig. 6. The load current during the appearance of an OCF.

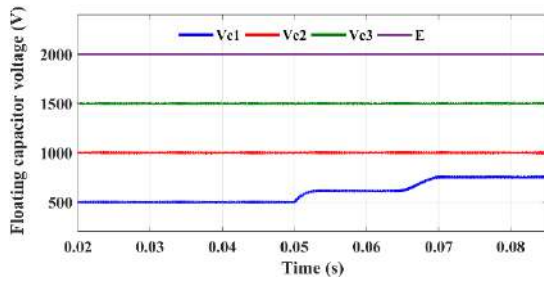


Fig. 7. The floating capacitor voltage during the appearance of an OCF.

The defective cell can be identified by observing the voltage degradation across the floating capacitors. This degradation is evident when the positive alternation of the current in the first phase is canceled, indicating that the fault originates in the upper stage. Conversely, a fault occurs in the *IGBT* of the lower stage. In that case, the measured current will be canceled during the negative alternation, leading to load voltage distortions, as illustrated in the Fig. 8,9 and 10.

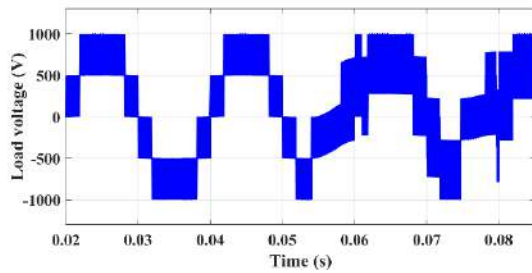


Fig. 8. The load voltages during the appearance of an OCF.

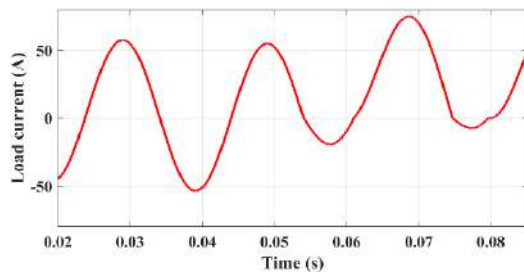


Fig. 9. The load current during the appearance of an OCF.

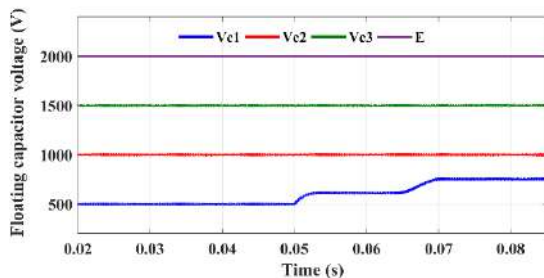


Fig. 10. The floating capacitor voltage during the appearance of an OCF.

The results were consistent with previous observations. When an *OCF* occurs in an *IGBT* of one phase, the degraded

operation is evident in the distorted load voltages of the affected phase. The measured load currents are nullified during the positive half-period if the fault is at the upper stage (*US*) and during the negative half-period if the fault is at the lower stage (*LS*). Additionally, the voltage across the floating capacitors connected to the faulty cell experiences significant fluctuations upon the occurrence of the fault,

B. Case two: short-circuit faults

The Fig. 11 and 12 show the output voltages during a short circuit fault (*SCF*) occurring on the *IGBT* at time $t = 0.05$ s. These Fig. clearly illustrate the faulty phase through distortions in the output voltage waveform. A distortion of the voltage across the floating capacitor associated with the faulty cell is also observed at the time the fault occurs.

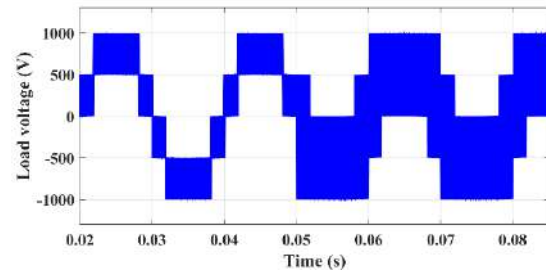


Fig. 11. The load voltage during the appearance of an SCF.

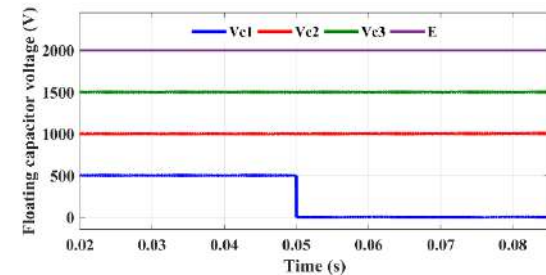


Fig. 12. The floating capacitor voltage during the appearance of an SCF.

The defective cell can be identified by observing the voltage degradation across the floating capacitors. When an *SCF* fault occurs in the *IGBT* of one phase of the three-phase inverter, the degraded state is indicated by deformations in the load voltages of the faulty phase. The voltage across the floating capacitors associated with the defective cell undergoes substantial fluctuations upon the occurrence of the fault.

C. OCF and SCF faults detection

Detecting *OCF* and *SCF* faults in a multicellular inverter is crucial for maintaining the safe and reliable operation of power electronic systems. The observer used generates parameter estimation signals; the distinction between these and measured signals is employed as a distinctive marker for detecting faults. By observing the signals of the output's voltages and current as well as the floating capacitors voltages of the multicellular inverter, a significant understanding of the inverter's functioning can be obtained. Sudden fluctuations in these signals can act as indicators of faults. The $\hat{V}_{Ci}$ and $\hat{i}_{Ch}$ are the observed signals of V_{Ci} and i_{Ch} . If there is a significant difference between these values, this indicates a fault. If there is a significant difference

between these values, this indicates a fault. The residuals, the estimation errors, are then analyzed to detect the faults. In this step, we construct the residual, i.e., the differences between the measurement of the system output and its estimated values. The residual r_i can be expressed as follows: $r_i = y_i - \hat{y}_i$, where y_i is the measured value and $\hat{y}_i$ is the estimated value.

At time $t = 0.05$ (s), an *OCF* fault occurred, the Fig. 13 and 14 illustrate the measured and estimated signals of the load current and the voltage across the floating capacitors of the faulty inverter.

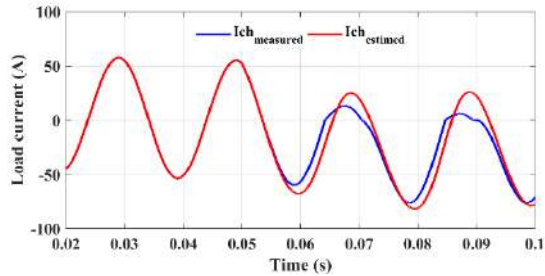


Fig. 13. The measured and estimated load current of the faulty phase during OCF.

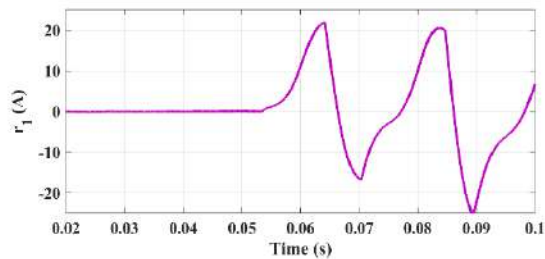


Fig. 14. The residuals of the measured and estimated load current of the faulty phase during the OCF.

The results clearly illustrate the usefulness of the observer in detecting faults during an *OCF* fault.

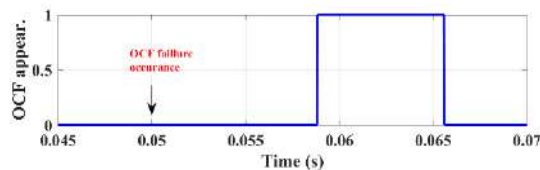


Fig. 15. OCF fault detection time.

$t_{OCF}=0.05$ (s). $t_{alarm}\approx 0.058$ (s). Therefore, the OCF detection time is $t_{det}\approx 0.008$ (s).

Fig. 16 and 17 illustrate the results of SCF fault detection.

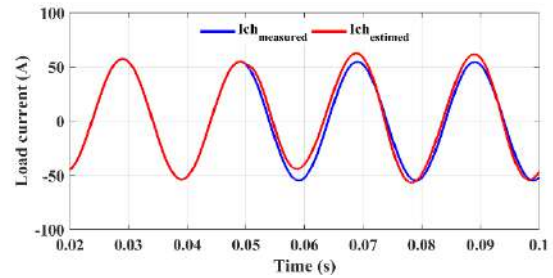


Fig. 16. The measured and estimated load current of the faulty phase during SCF.

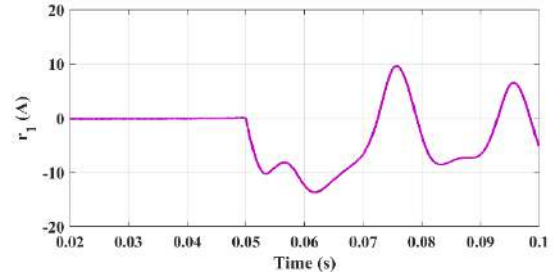


Fig. 17. The residuals of the measured and estimated load current of the faulty phase during the SCF.

The detection signals are shown in the Fig. 18

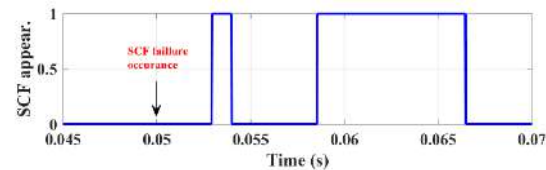


Fig. 18. SCF fault detection time.

$t_{SCF}=0.05$ (s). $t_{alarm}\approx 0.054$ (s). Therefore, the SCF detection time is $t_{det}\approx 0.004$ (s).

During the fault detection stage, a fault alarm is generated to indicate the presence of an anomaly in a given inverter, without specifying the affected component or the fault type. Therefore, in the subsequent fault identification stage, it is necessary to determine both the faulty cell and the nature of the fault that triggered the alarm. This combined process of detection and identification constitutes the overall “fault diagnosis,” which aims to identify the fault’s location, origin, and type. The final step is fault isolation, which may involve disabling or bypassing the defective component while ensuring that the rest of the system continues to operate correctly.

V. CONCLUSION

Multicellular inverters are particularly appealing due to their suitability for high-power and high-voltage applications. However, the occurrence of open-circuit and short-circuit faults in these converters can significantly degrade the performance of the conversion system. This highlights the critical importance of accurate and timely fault diagnosis. Effective diagnostic techniques not only ensure the safety and reliability of electrical drives but also minimize downtime and maintenance costs,

thereby improving overall system efficiency and lifespan. Consequently, significant research efforts have been devoted to developing robust fault diagnosis methods tailored to multilevel inverters, emphasizing their role in maintaining optimal performance in demanding applications.

REFERENCES

- [1] R. Bensaid, M. Fadel "Sliding mode observer for multicell converters", IFAC Nonlinear Control Systems, St. Petersburg, Russia, 2001.
- [2] S. Meradi, K. Benmansour, K. Herizi, M. Tadjine, M. S. Boucherit "Sliding mode and faults tolerant control for multicell converter four quadrants", Electric Power Systems Research 95 (2013) 128-139.
- [3] P. Djondiné "Overview of Control Techniques for Multicellular Converter," Journal of Engineering Sciences, 2018, 5 (1), pp. E 10-E 14. 10.21272/jes.2018.5(1). e3.
- [4] F. W. Fuchs, "Some diagnosis methods for voltage source inverters in variable speed drives with induction machines - a survey," *IECON'03, Roanoke, VA, USA*, pp. 1378-1385, 2003, <http://doi:10.1109/IECON.2003.1280259>.
- [5] A. Aibeche, M. Kidouche, "Sensor fault detection, localization, and system reconfiguration with a sliding mode observer and adaptive threshold of PMSM," J. Power. Electron, Vol. 16, no. 3, pp. 1012–24, 2016.
- [6] O. Tachon, M. Fadel, T. Meynard, "Control of series multicell converter by linear state feedback decoupling," *7th European Conference on Power Electronics and Applications EPE*, Trondheim, Norway, September 1997.
- [7] T. A. Meynard, M. Fadel, N. Aouda, "Modeling of Multilevel Converters," *IEEE Transactions on Industrial Electronics*, vol. 44, no. 3, JUNE 1997. <http://doi:10.1109/41.585833>.
- [8] F. Khater, A. Aibeche, S. A. Fellag, "Faults diagnosis for multilevel-multicell Converters using sliding mode controller and adaptive threshold," *ICAECOT*, Setif, Algeria, 2024, <http://doi:10.1109/ICAECOT62402.2024.10829285>.
- [9] L. Wu, J. Liu, S. Vazquez and S. K. Mazumder, "Sliding Mode Control in Power Converters and Drives: A Review," in *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 3, pp. 392-406, March 2022, <https://doi:10.1109/JAS.2021.1004380>.
- [10] H. Komurcugil, S. Biricik, S. Bayhan and Z. Zhang, "Sliding Mode Control: Overview of Its Applications in Power Converters," in *IEEE Industrial Electronics Magazine*, vol. 15, no. 1, pp. 40-49, March 2021, <https://doi:10.1109/MIE.2020.2986165>.
- [11] C. Edwards, S. K. Spurgeon, R. J. Patton, "Sliding mode observers for fault detection and isolation," *Automatica*, Vol. 36, no. 4, 2000, [https://doi.org/10.1016/S0005-1098\(99\)001776](https://doi.org/10.1016/S0005-1098(99)001776).
- [12] A. El Mekki, K. Ben Saad, "Fault diagnosis of open and short-circuit faults in a parallel multi-cell converter based on sliding mode observer," *SN Appl. Sci.*, vol. 2, no. 179, 2020, <https://doi.org/10.1007/s42452-020-1954-6>.
- [13] S. Shao, P. W. Wheeler, Clare JC, A. J. "Watson, Opencircuit fault detection and isolation for modular multilevel converter based on sliding mode observer," *15th European Conference on Power Electronics and Applications EPE*, 2013, <https://doi:10.1109/EPE.2013.6634401>.

Congestion Control by the use of Nash equilibrium concept

Badreddine MAMRI¹ and Abdelmalek Moutsi^{2,3}

¹Department of Electronics, University Frères Mentouri Constantine 1, Algeria

²Faculty of process engineering, University of Salah boubnider Constantine3 , Algeria

³LIST Laboratory, University M'Hamed Bougara, Boumerdes, Algeria

badromamri@gmail.com
abdelmalek.moutsi@univ-constantine3.dz
a.moutsi@univ-boumerdes.dz

Abstract-- An ad hoc network is a wireless network composed of mobile nodes, characterized by a distributed, dynamic, and self-organized architecture. Each node in the network is capable of independently adapting its operation to the current environment according to predetermined algorithms and protocols. Routing represents the essential function in an ad hoc network, and the routing problem is the most challenging issue to address. There are numerous routing optimization methods; we propose a new approach based on game theory to solve the congestion problem. The use of the Nash equilibrium concept has allowed us to create a load-distribution mechanism in order to select the least congested paths.

Key words -- Ad Hoc Networks, Routing, Game Theory, Nash Equilibrium.

I. Introduction:

Ad hoc networks represent a new paradigm in wireless networking. They are formed through the interconnection of various unknown mobile entities and do not rely on any fixed infrastructure or centralized control. Cooperation among these entities allows the network services to be maintained. Game theory has been widely used in mobile ad hoc network research as a theoretical decision-making framework, for example to address routing problems. The frequent mobility of nodes in MANETs and the nature of wireless links lead to the modification, disappearance, and emergence of new routes. Thus, the absence of an infrastructure responsible for routing increases its complexity. The participation of nodes in routing requires efficient routing protocols capable of adapting to this environment. These protocols demand significant resources—memory, processors, bandwidth, and energy—yet such resources are scarce in MANETs due to their characteristics. Most existing protocols do not consider traffic load balancing. Therefore, we propose a learning algorithm based on game theory to optimize routing in ad hoc networks, in order to prevent node overload and minimize network congestion.

II. Routing in Ad Hoc Networks:

Routing is a set of rules or algorithms that allow data to be processed and transferred from one node to another within the network. At the same time, a routing protocol is responsible for maintaining and, when necessary, repairing all routes. Node mobility in ad hoc networks makes routing more

complex [1]. Three types of routing protocols can be defined according to the procedure used to establish and maintain routes: reactive (on-demand), proactive (global network view), and hybrid [2].

- **Proactive routing protocols:** In this category, also known as table-driven protocols, routing tables are continuously maintained and updated at each node. Protocols based on this principle include DSDV, HSR, FSR, LSR, and OLSR [3].

- **Reactive routing protocols:** These protocols rely on route discovery and maintenance. When needed, a global route discovery procedure is initiated. This process stops once the route is found or all possibilities have been explored. Protocols following this approach include DSR, AODV, TORA, ABR, SSR, and EARP [3, 4].

- **Hybrid routing protocols:** This category combines both proactive and reactive techniques. The ZRP protocol is an example of this approach [5].

- **Examples of Routing Protocols:** The most well-known routing protocols are:

- DSDV Protocol (Dynamic Destination-Sequenced Distance Vector):** The DSDV algorithm was specifically designed for mobile networks. Each mobile station maintains a routing table containing all possible destinations, the number of hops required to reach each destination, and a sequence number (SN) corresponding to each destination node. This sequence number makes it possible to distinguish new routes from old ones and prevents routing loops from forming [3, 6].

- DSR Routing Protocol (Dynamic Source Routing):** The DSR routing protocol is based on the “source routing” technique, which means that the source node determines the complete sequence of intermediate nodes through which data packets will be forwarded. Nodes do not need routing tables. The two basic operations of DSR are route discovery and route maintenance [3, 7].

- AODV Routing Protocol (Ad hoc On-Demand Distance Vector):** The AODV routing protocol is considered a combination of DSR and DSDV. It incorporates the route discovery and maintenance mechanisms of DSR along with

the periodic updates of DSDV. Route discovery is performed by broadcasting an RREQ (Route Request) message. AODV uses sequence numbers to avoid loops and ensure that the most recent (freshest) routes are used. Once the RREQ reaches the destination, the destination sends back an RREP (Route Reply) message to the source along the reverse path. Because the RREP follows the same path as the RREQ, route consistency is maintained. AODV also uses a route maintenance mechanism. When a node on an active route moves, its upstream neighbor detects this movement and sends a Route Error (RERR) packet to its active upstream neighbors to inform them of the route break [3, 8].

RREQ Packet Format:

Source Address | Source_NS | RREQ_ID (broadcast ID) | Destination Address | Dest_NS | Cp (Cost) | Hop_count

RREP Packet Format:

Source Address | Destination Address | Dest_NS | Min (Hop_count) | Lifetime

III. Optimization of the AODV Routing Protocol:

One of the main drawbacks of the AODV protocol is that the selection of the best path is based solely on the hop count. Our contribution consists of modifying the AODV routing protocol so that routes are selected based on a cost function that depends on the queue load of the interface of each node along path p , in order to avoid congested routes. In our work, modifications were introduced to improve transmission delay and throughput. We propose the following steps:

-The cost of each network node is computed according to its load.

- Nodes periodically update their routing tables.

-The metric of the route discovery algorithm is modified by giving priority to the cost rather than the hop count. Thus, the route discovery process of the AODV protocol operates as follows:

✓ The source broadcasts an RREQ packet when it does not have a route to deliver data to a destination.

✓ By using a reserved field in the RREQ structure (where the cost value of a node is added), each node j computes its cost using the formula:

$$C_j = \sum_{i=1}^n 1_{\{j \in p_i(t)\}} * C_i$$

where: $1_{\{j \in p_i(t)\}}$ equals 1 if node $j \in p_i(t)$ and 0 if not [9].

✓ Upon receiving an RREQ packet, each intermediate node checks whether it has already received the packet from the same source with the same RREQ_ID. If so, the packet is discarded. Otherwise, it verifies whether it has a route whose sequence number is greater than or equal to the sequence number of the RREQ; if so, it sends an RREP packet to the source. If not, it computes its cost and adds it to the cumulative cost during the forwarding of the route request until it reaches the destination:

$$C_j = \sum_{i=1}^n C_i$$

Where n is the number of nodes traversed by the RREQ.

✓ The destination replies with an RREP packet after selecting the route with the lowest cost (the least congested path):

$$C_j(t) = C_j$$

In case of equal costs, the shortest path is selected. The modified RREQ and RREP packets are shown in the following:

@source | source_NS | RREQ_ID (broadcast ID) | @destination | Dest_NS | Cp (Cost) | Hop_count

@source | @destination | Dest_NS | Min (Cp) | Min (Hop_count) | lifetime

IV. Proposed algorithms:

a) Algorithm at the Source Node S:

```
Start
The source S has packets to send
If (S has a route to destination D) then
    {Start data transmission}
Else {
    Route discovery process
    Create RREQ
    Initialize cost (C_p = C(S))
    Initialize hop count (Hop_count = 0)
    Send RREQ to neighbors
```

```
    If (S receives an RREP indicating the route is ready) then
        {Start data transmission}
    Else if (S does not receive an RREP after a waiting period ("lifetime"))
        or (S receives an RERR) then
            {Route discovery process}
    }
End.
```

b) Algorithm at the intermediate node I:

```
Start
Intermediate node I receives an RREQ.
If ([RREQ_ID, @ source] exists) then
    {Discard RREQ}
Else if (Node I has a route to destination D) then
    If (Max(NS)) then
        {Node I sends an RREP to source S}
    Else
        {Discard RREQ}
Else {
    Accumulate the cost: (C_p = C_p + C(I))
    Accumulate the distance: (Hop_count = Hop_count + C(I))
    Update routing table entries
```

```

    Update the fields (UPDATE) (C_p) and
    (Hop_count) in the RREQ
    Broadcast RREQ to neighbors
  }
End.

```

c) Algorithm at the destination node D:

```

Start
Destination node D receives RREQ
If (RREQ destination IP = node IP) then
  If (Max(NS)) then
    If (This path has the best cost (Min(C_p))) then
      {Send RREP to source S}
    Else if (Paths have equal costs) then
      If (This path is the shortest (Min(Hop_count))) then
        {Send RREP to source S}
      Else {Delete RREQ}
    Else {Delete RREQ}
    Else {Delete RREQ}
    Else {Intermediate node I algorithm}
  }
End.

```

V. Simulation:

Parameter	valeur
Selected protocol	AODV
Number of packets per message	$w_i = 8$ paquets
Packet size	512 octets
Path utility	$util_i = 1 - \frac{C_{p_i}}{C_{p_{max}}}$
Path selection probability vector	$p_{i, k}(t) = \frac{1}{\sum_{k=1}^K util_i(t)}$
Update algorithm	$p_{i, k}(t+1) = \begin{cases} p_{i, k}(t) + \frac{1}{p} \cdot util_i(t) \times (1 - p_{i, k}(t)) & \text{if } k \in p \\ p_{i, k}(t) \times (1 - \frac{1}{p} \cdot util_i(t)) & \text{if } k \notin p \end{cases}$
Delay	$Delay_{i, k} = \frac{C_{p_i}}{w_i} + \frac{1}{w_i} \times (C_{p_i} * w_i)$
Throughput	$Throughput = \frac{C_{p_i} * w_i}{Delay_{i, k}} * 8$

Table 1. Simulation Parameters Initialization.

Proposed scenarios:

First scenarios:

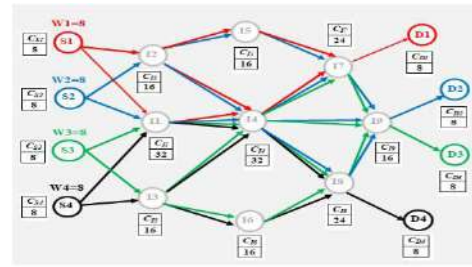


Figure 1. Node distribution in the first scenario.

Simulation Results (Scenario 1):

Delay:

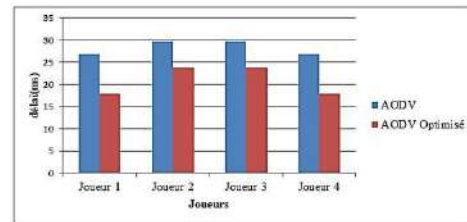


Figure 2. Delay evaluation in Scenario 1.

Throughput:

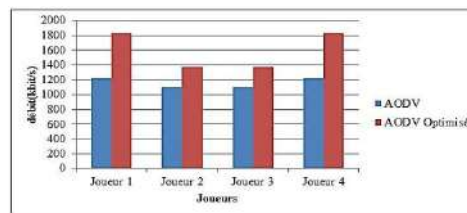
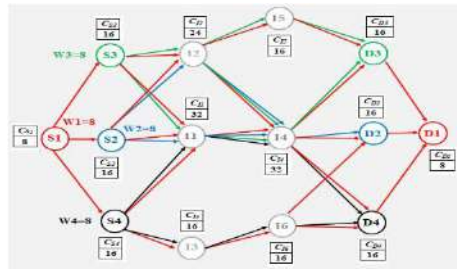
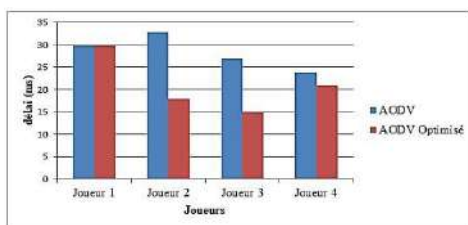
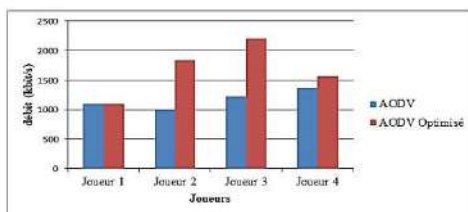
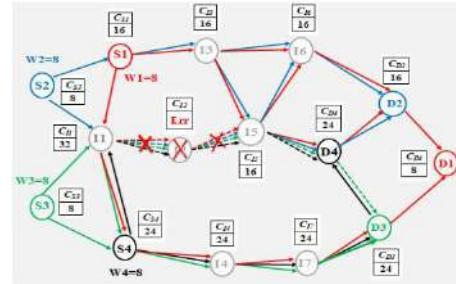
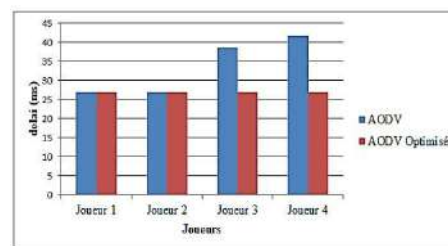
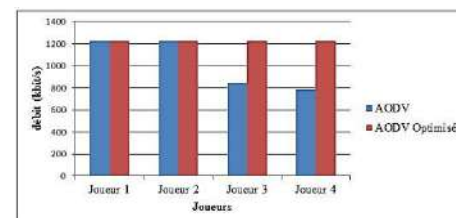


Figure 3. Throughput evaluation in Scenario 1.

Second scenarios:**Figure 4.** Node distribution in the second scenario.**Simulation Results (Scenario 2):****Delay:****Figure 5.** Delay evaluation in Scenario 2.**Throughput:****Figure 6.** Throughput evaluation in Scenario 2.**Third scenarios:****Figure 7.** Node distribution in the third scenario.**Simulation Results (Scenario 3):****Delay:****Figure 8.** Delay evaluation in Scenario 3.**Throughput:****Figure 9.** Throughput evaluation in Scenario 3.**Conclusion:**

In our study, we proposed a mathematical model based on game theory to address the congestion problem in ad hoc routing, where node load can lead to increased end-to-end delay and reduced throughput. Our contribution consists of making modifications to the general operation of an ad hoc routing protocol (AODV). These modifications were applied to the route creation process by changing the criterion for selecting the best route, introducing the concept of cost, which represents the traffic load in a node's queue. Finally, to validate our approach, we provided simulation results across different network topologies.

References:

- [1] K. Saddiki, "Denial of services attack in wireless networks", these de doctorat, Faculté des Sciences Exactes, Université Djillali LIABES de Sidi-Bel-Abbès, Sidi-Bel-Abbès, 2019.
- [2] M. Azza, S. H. Boukli et K. M. Faraoun. "A Cross Layer for Detection and Ignoring Black Hole Attack in MANET". *International Journal of Computer Network and Information Security*, vol.7, no.10, p.42, 2015.
- [3] M. Sedrati, L. Aouragh, L. Guettala, A. Bilami, "Etude des Performances des Protocoles de Routage dans les Réseaux Mobiles Ad-hoc", 4th International Conference on Computer Integrated Manufacturing, November 2007, Batna, Algeria
- [4] C. E. Perkins, E. E. Royer, S. R. Das, and M. K. Marina, "Performance comparison of two on-demand routing protocols for ad hoc networks," *In IEEE Personal Communications*, vol. 8, pp. 16-28, 2001
- [5] S. NAIMI, "Gestion de la mobilité dans les réseaux Ad Hoc par anticipation des métriques de routage", thèse de doctorat, Université Paris-Sud, Paris, 2015.
- [6] C. Perkins and P. Bhagwat, "Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers", *In Proceedings of ACM SIGCOMM Conference on Communications Architectures, Protocols and Applications*, pp. 234-244, 1994.
- [7] D. B. Johnson and D. A. Maltz, "Dynamic source routing in ad hoc wireless networks" *In Mobile Computing, Imielinski and Korth, Eds. Kluwer Academic Publishers*, vol. 353, pp. 153-181, 1996.
- [8] M. ELIAS, M. CAMPISTA, G. Rubinstein, "Advanced Routing Protocols for Wireless Networks", Wiley- ISTE, London, pp. 23-28, 2014.
- [9] D. Barth, O. Bournez, O. Boussaton, et J. Cohen. "Distributed learning of Wardrop equilibria". *In Springer-Verlag, editor, 7th International Conference on Unconventional Computation*, vol. 5204 of *Lecture Notes in Computer Science*, pp. 19-32, 2008.

Vector Control, DPC, and Fuzzy DPC: A Comparative Study for Grid Converters in Hybrid Wind-PV System

Zoubir Gadouche
Laboratory of Electrical and Computer
Engineering (L2GEGI) of Electrical
Department, University Ibn Khaldun of
Tiaret, Tiaret, Algeria
gadouche.zoubir@gmail.com

Mokhtar Mahmoud Mohammedi
Laboratory of Electrical Drives
Development (LDEE) of Electrical
Department, University of Sciences &
Technology of Oran, Mohamed
Boudiaf, Oran, Algeria
mokhtar.mohammedi@yahoo.fr

Mohammed Bey
Laboratory of Electrical and Computer
Engineering (L2GEGI) of Electrical
Department, University Ibn Khaldun of
Tiaret, Tiaret, Algeria
bey.mohamed@hotmail.com

Abstract— This research presents a comparative analysis of three distinct power control methodologies—Vector Control (VC), Direct Power Control (DPC), and Fuzzy Direct Power Control (FDPC)—implemented on the grid-side converter of a hybrid renewable energy system. The system configuration integrates a Wind Energy System (WES) utilising a Double-Fed Induction Generator (DFIG) with a Photovoltaic (PV) system, interconnected through a common DC bus to ensure coordinated operation. Evaluation, performed via MATLAB/Simulink simulations, focused on the critical performance metrics of output energy quality maintenance and dynamic responsiveness to power fluctuations. The findings indicate that VC exhibits comparably longer response times. Conversely, DPC provides high-speed response and simplified implementation, though it may compromise energy quality under dynamic operational states. Crucially, FDPC, by incorporating fuzzy logic, effectively mitigates system uncertainties, thereby establishing an optimal balance between energy quality and dynamic response time. These results provide significant and valuable guidance for the judicious selection of control strategies within complex hybrid renewable energy systems.

Keywords—Direct power control, Vector control, Double-Fed Induction Generator, grid-side converter, Fuzzy logic.

I. INTRODUCTION

The escalating global focus on sustainable energy has catalysed the rapid integration of hybrid renewable energy systems (HRES), which optimally combine complementary sources, such as wind energy systems (WES) and Photovoltaic (PV) arrays. This integration often utilises a shared DC bus, necessitating sophisticated power management to ensure a stable and high-quality energy output to the main grid [1].

A critical component in the coordination and power flow regulation of these HRES is the grid-side converter (GSC). The GSC plays a crucial role in maintaining the DC-link voltage and regulating the reactive power exchange with the grid, which has a direct impact on the overall system's stability, efficiency, and output energy quality. Therefore, the choice of the control methodology applied to the GSC is paramount [2].

Historically, Vector Control (VC) has been a widely adopted strategy due to its well-established robustness and precise decoupling of active and reactive power. However, its reliance on inner current control loops can lead to slower dynamic response under rapid changes in renewable resource availability [3]. An alternative, Direct Power Control (DPC), offers advantages in terms of faster dynamic performance and

simplified implementation as it bypasses inner current regulation. Yet, the traditional DPC often suffers from higher power ripples and compromised energy quality during highly dynamic operational states [4].

To address the limitations of both VC and conventional DPC, this research explores an advanced technique: Fuzzy Direct Power Control (FDPC). By incorporating fuzzy logic, the FDPC methodology aims to effectively handle the inherent non-linearities and uncertainties within HRES, potentially achieving a superior balance between the swift dynamic response characteristic of DPC and the high-quality output typically associated with more complex control schemes [5]. This study presents a comparative analysis of these three distinct power control methodologies (VC, DPC, and FDPC) as implemented on the GSC of a hybrid WES (utilizing a Double-Fed Induction Generator - DFIG) and PV system, evaluating their performance based on output energy quality and dynamic responsiveness to power fluctuations [6].

II. GRID SIDE CONVERTER MODELLING

This study focuses on a hybrid renewable energy system that combines a Wind Energy System (WES) utilising a Doubly-Fed Induction Generator (DFIG) and a Photovoltaic (PV) system, connected via a common DC bus.

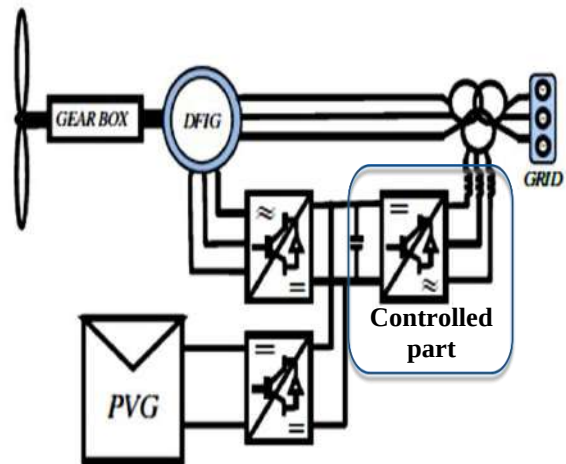


Fig.1. Hybrid energy system global schema [4].

A. DC Bus Model

The two energy sources are coupled via a “DC bus”, as presented in Figure 2.

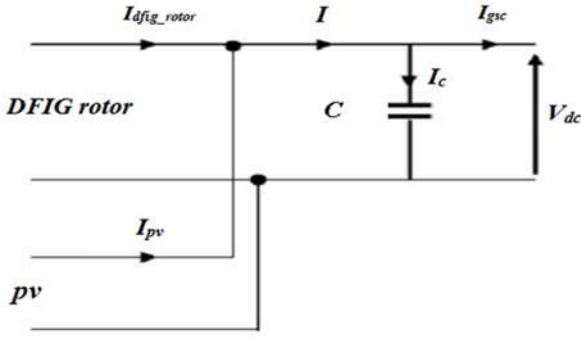


Fig 2. DC bus representation [4].

The DC bus voltage is given by:

$$V_{dc} = \frac{1}{C} \cdot \int_0^t I_c \cdot dt \quad (1)$$

The current in the capacitor originates from a node where the hybrid source and the GSC modulate the circulating current.

$$I_c = I - I_{GSC} = (I_{dfig\_rotor} + I_{pv}) - I_{GSC} \quad (2)$$

B. Grid Converter Model

To model the system, the converter is divided into three parts: the AC side, the discontinuous part composed by the switches, and the DC side [7].

The switches are complementary; their state is defined by the following function [7]:

$$S_j = \begin{cases} 1, \bar{S}_j = 0 \\ 0, \bar{S}_j = 1 \end{cases} \quad (3)$$

The input phase voltages and the output current can be written according to S_j , V_{dc} , and the input currents.

$$C \frac{dV_{dc}}{dt} = (S_a i_a + S_b i_b + S_c i_c) - I \quad (4)$$

With

$$I_c = S_a i_a + S_b i_b + S_c i_c \quad (5)$$

III. GRID CONVERTER CONTROL

We are interested in the control and connection of the grid-side converter (GSC) to the electrical network via the RL filter. The synoptic diagram of this control is represented by Fig.3.

A. Control Scheme Based On Vector Control

1) Control of currents flowing through the filter:

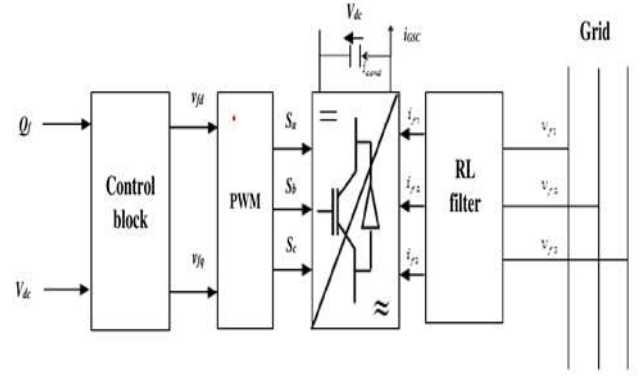


Fig 3. Block diagram of the CCR control [8]

In the (d, q) frame of reference linked to the rotating stator field, we have the following equations:

$$\begin{cases} v_{fd} = R_f \cdot i_{fd} + L_f \frac{di_{fd}}{dt} + e_{fd} \\ v_{fq} = R_f \cdot i_{fq} + L_f \frac{di_{fq}}{dt} + e_{fq} \end{cases} \quad (6)$$

$$\begin{cases} e_{fd} = \omega_s \cdot L_f \cdot i_{fq} \\ e_{fq} = -\omega_s \cdot L_f \cdot i_{fd} + v_{sq} \end{cases} \quad (7)$$

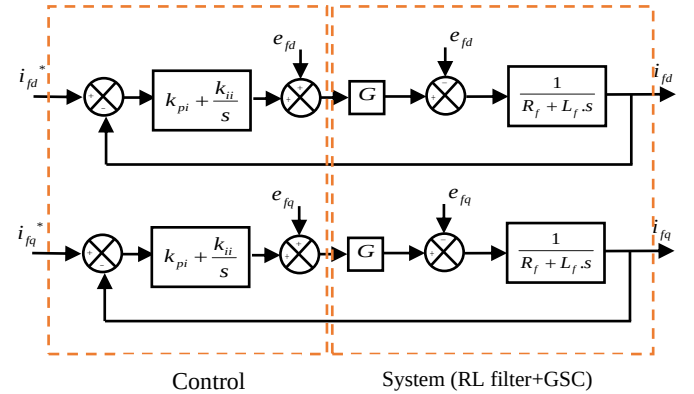


Fig.4. RL filter current control

With:

v_{f123}, i_{f123} : Grid voltages and currents.

v_{fdq}, i_{fdq} : The voltages and currents of the grid in the (d, q) coordinate system.

R_f, L_f : Filter resistance and inductance.

2) *DC bus voltage control*: We can express the powers involved on the DC bus.

$$\begin{cases} P_{GSC} = v_{dc} \cdot i_{GSC} \\ P_c = v_{dc} \cdot i_c \\ P_{Hyb} = v_{dc} \cdot i \end{cases} \quad (8)$$

The relationship between these powers is expressed by the following relation.

$$P_{GSC} = P_c + P_{Hyb} \quad (9)$$

We also have, if neglecting the losses in the resistance of the RL filter.

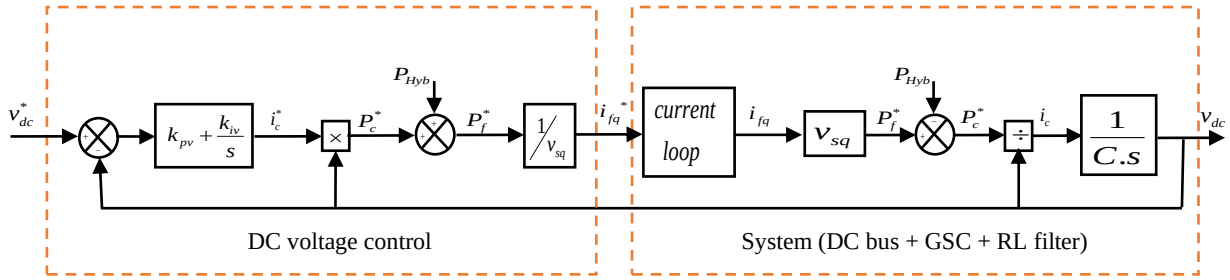


Fig.5 DC bus voltage regulation loop [9]

$$\begin{cases} P_f = v_{sq} \cdot i_{fq} \\ Q_f = v_{sq} \cdot i_{fd} \end{cases} \quad (10)$$

If we neglect all Joule losses compared to the power exchanged between the DFIG rotor and the electrical grid, we can then write.

$$P_f = P_{GSC} = P_c + P_{Hyb} \quad (11)$$

The power reference for the capacitor is related to the reference current flowing through the capacitor.

$$P_c^* = v_{dc} \cdot i_c^* \quad (12)$$

Figure. 5 shows the DC bus voltage control block diagram with a PI regulator.

B. Control scheme based on classical direct power control

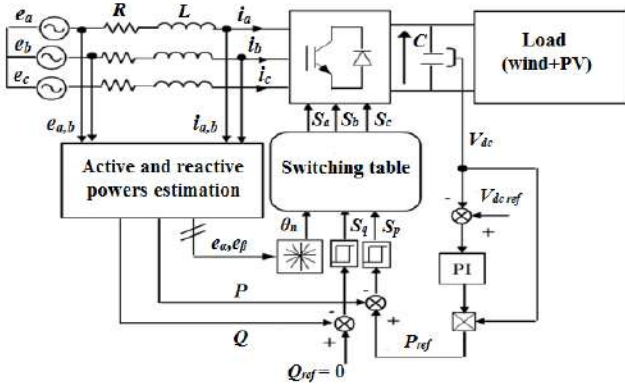


Fig.6 Principle of classical DPC

TABLE I. SWITCHING TABLE FOR C-DPC[9]

S _p	S _q	θ ₁	θ ₂	θ ₃	θ ₄	θ ₅	θ ₆	θ ₇	θ ₈	θ ₉	θ ₁₀	θ ₁₁	θ ₁₂
1	0	V ₅	V ₆	V ₆	V ₁	V ₁	V ₂	V ₂	V ₃	V ₃	V ₄	V ₄	V ₅
	1	V ₃	V ₄	V ₄	V ₅	V ₅	V ₆	V ₆	V ₁	V ₁	V ₂	V ₂	V ₃
0	0	V ₆	V ₁	V ₁	V ₂	V ₂	V ₃	V ₃	V ₄	V ₄	V ₅	V ₅	V ₆
	1	V ₁	V ₂	V ₂	V ₃	V ₃	V ₄	V ₄	V ₅	V ₅	V ₆	V ₆	V ₁

C. Control Scheme Based On Fuzzy Direct Power Control

In F-DPC, the conventional comparators and classical control selection table C-DPC are replaced by a simple "fuzzy logic controller" to obtain a fixed "switching frequency", which leads to a considerable current harmonic reduction. These rules directly use the errors of "active power" and "reactive power" as fuzzy variables. The control vector selection (S_a, S_b, S_c) principle is given in Figure 7 [4][10].

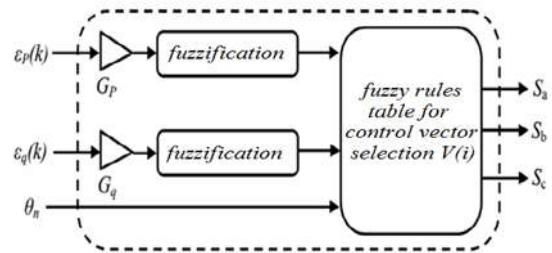


Fig.7 Principle of Fuzzy DPC

The suggested DPC configuration for the GSC of the hybrid renewable system is clarified in Figure 8.

Generally, the "fuzzy controller" design requires the choice of these parameters: "linguistic variables", "membership functions", "inference method", and "fuzzification strategy".

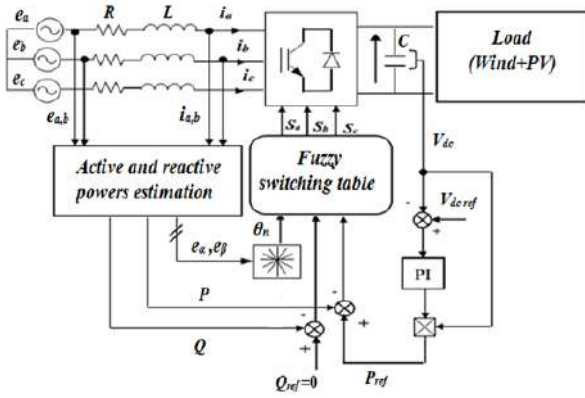


Fig.8 Proposed structure of DPC with fuzzy selection

1) Fuzzification

Fuzzy controller inputs are:

$$\begin{cases} \varepsilon_p = P^* - P \\ \varepsilon_q = Q^* - Q \end{cases} \quad (15)$$

The controller has three "membership functions" for the input of ε_p ("N: Negative, Z: zero, P: positive"), and for the input of ε_q , also three "membership functions" ("N: Negative, Z: zero, P: positive"). The memberships of the controller are illustrated in Figure 9.

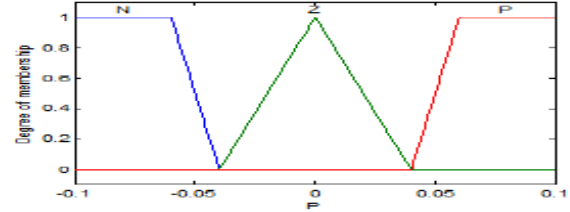


Fig.9 Membership functions for active and reactive powers error

2) Inference

The fuzzy inference is the process of formulating the relationship between inputs and outputs the fuzzy logic [15]. These rules must take into account the system to adjust and the goals of the proposed adjustment.

The set of "fuzzy rules" combinations for all sectors of the "grid voltage" is shown in Table 2.

TABLE II. FUZZY RULES [4]

$\varepsilon_p^{(k)}$	$\varepsilon_q^{(k)}$	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}	θ_{11}	θ_{12}
N	N	V ₆	V ₁	V ₁	V ₂	V ₂	V ₃	V ₃	V ₄	V ₄	V ₅	V ₅	V ₆
	Z	V ₁	V ₁	V ₂	V ₂	V ₃	V ₃	V ₄	V ₄	V ₅	V ₅	V ₆	V ₆
	P	V ₁	V ₂	V ₂	V ₃	V ₃	V ₄	V ₄	V ₅	V ₅	V ₆	V ₆	V ₁
Z	N	V ₆	V ₁	V ₁	V ₂	V ₂	V ₃	V ₃	V ₄	V ₄	V ₅	V ₅	V ₆
	Z	V ₇	V ₀	V ₇	V ₀	V ₇	V ₀	V ₇	V ₀	V ₇	V ₀	V ₇	V ₀
	P	V ₁	V ₂	V ₂	V ₃	V ₃	V ₄	V ₄	V ₅	V ₅	V ₆	V ₆	V ₁
P	N	V ₅	V ₆	V ₆	V ₁	V ₁	V ₂	V ₂	V ₃	V ₃	V ₄	V ₄	V ₅
	Z	V ₇	V ₀	V ₇	V ₀	V ₇	V ₀	V ₇	V ₀	V ₇	V ₀	V ₇	V ₀
	P	V ₂	V ₃	V ₃	V ₄	V ₄	V ₅	V ₅	V ₆	V ₆	V ₁	V ₁	V ₂

IV. SIMULATION RESULT

The proposed system was simulated using MATLAB/Simulink.

TABLE III. SIMULATION PARAMETERS

Parameters	Value
Wind system	
DFIG Nominal power (P)	7.5 kW
Wind turbine nominal power	10 kW
PV system	
Cells connected in parallel (N_p)	2
Cells connected in series (N_s)	15
RL filter and DC bus	
Filter resistance	0.3 Ω
Filter inductance	0.014 H
DC bus	2*10 ⁻³ F

This study examines three power control strategies for GSC for a hybrid renewable system: Vector Control (VC), Direct Power Control (DPC), and Fuzzy Direct Power Control (FDPC).

The following simulation results are divided into two parts, the first one is a comparative study between the three methods of control for energy quality, the second one is comparative study focused on response time and decoupling power control.

A. Comparative Analysis of Power Control Strategies for Energy Quality

To better illustrate energy quality, a spectral analysis of currents was performed. As a note, the measurement was done in the operation case of the fixed speed "wind turbine" (145 [rad/s]), but with a constant wind system "active power" ($P_{wind} = -7$ [kw]) and a unit "power factor" ($Q_{wind} = 0$ [Var]), and the Photovoltaic power is ($P_{pv} = 6$ [kw]) with constant irradiance $E = 1000$ [w/m²].

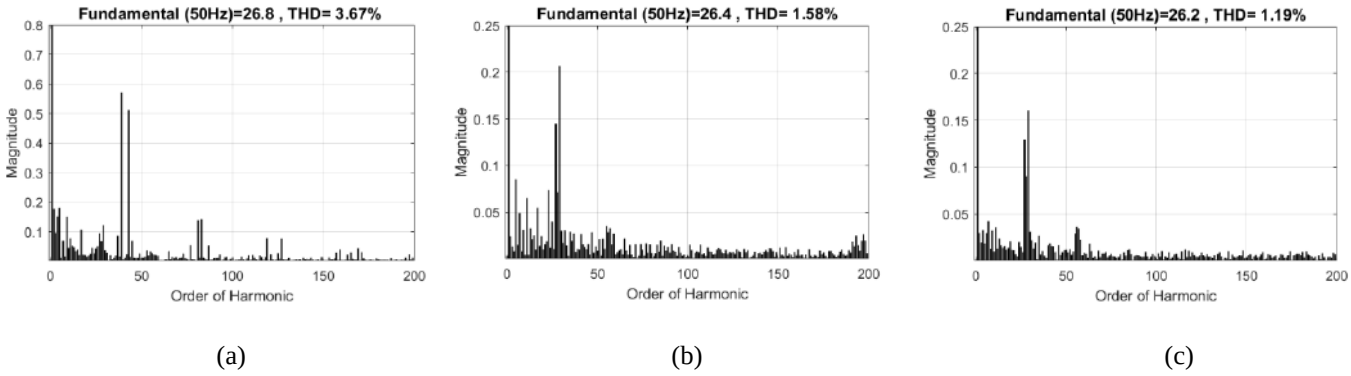


Fig.10 Harmonic analysis of the stator current spectrum: (a) CV (b) DPC (c) FDPC

The DC bus reference voltage is set at 620V. The reactive power is set at 0 (Var), for a power factor of unity.

“Fig.10(a)”: The vector control shows a THD (Total Harmonic Distortion) of 3.67, indicating relatively poor power quality. This may result from a less optimal response to load variations. “Fig.10 (b)”: The direct power control exhibits a THD of 1.58, demonstrating a significant improvement over the previous method. This reduction in harmonic distortion suggests better management of power fluctuations, resulting in superior power quality. “Fig.10 (c)”: The fuzzy direct power achieves the best THD of 1.19 among the three methods. This indicates optimal efficiency in power control and excellent power quality, making it particularly advantageous for demanding applications.

In summary, the results show that both fuzzy direct power control and direct power control techniques are more effective in reducing harmonic distortion compared to vector control.

B. Evaluation of Response Time and Decoupling Efficiency

In this test, a fixed speed of 145 [rad/s] is applied to the blades of the wind turbine, which corresponds to a hypo-synchronous mode of the DFIG. We subjected the system to step changes in active power and reactive power we set it to zero.

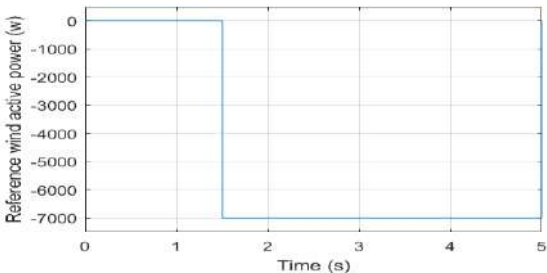


Fig.11 Reference wind active power

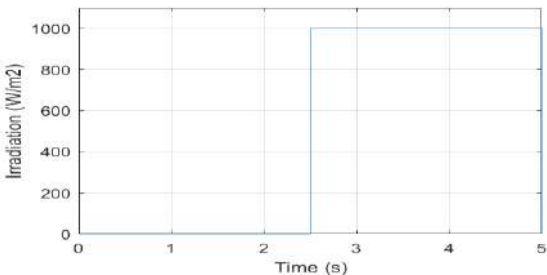


Fig.12 PV system irradiation

Figure.11 represents the reference wind system power, figure.12 represents PV system irradiation.

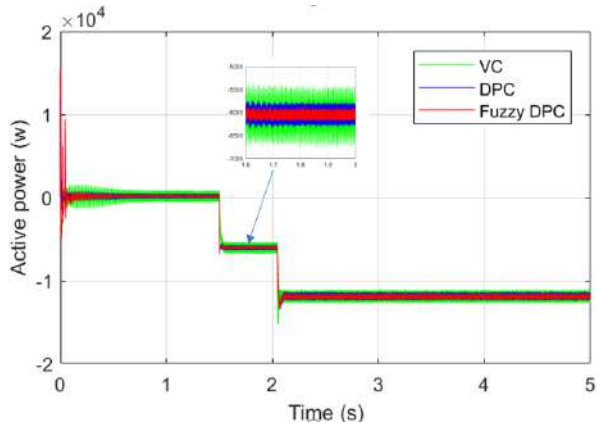


Fig.13 Hybrid system power

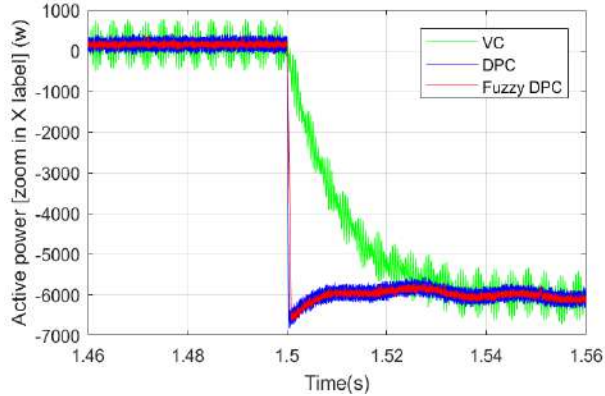


Fig.14 Zoom in hybrid system power

“Fig. 13” illustrates the active power for DPC, FDPC, and vector control. The fluctuations in active power are significantly reduced when using FDPC and DPC compared to the vector control technique. According to “Fig.14”, the active power followed its reference with a smaller ripple when using F-DPC.

In 'Fig. 14,' the zoomed-in x-axis label highlights the longer response times associated with the control vector compared to DPC and FDPC.

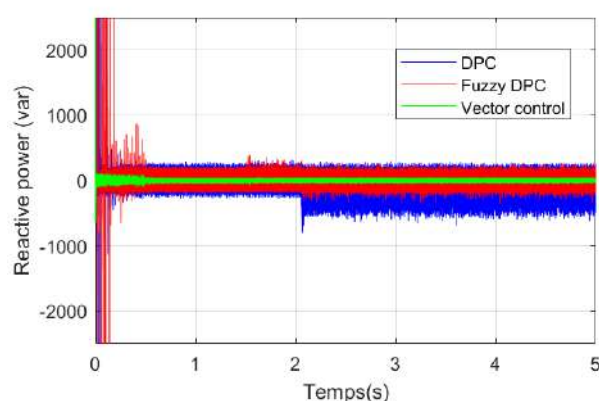


Fig.15 Reactive power.

As shown in “Fig.15”, although there is an initial overflow in reactive power with Fuzzy DPC, it eventually aligns closely with its reference, exhibiting smaller fluctuations compared to DPC, especially after the photovoltaic system has started operating (after 2s). “Fig.15” also shows that VC is characterised by stability and minimal power ripple, though it may entail longer response times.

V. CONCLUSION

In conclusion, this study highlights the importance of optimising power control strategies for grid-injected power as wind and photovoltaic energy become increasingly essential to the energy mix. While direct power control (DPC) offers a fast response, it struggles under dynamic conditions. Fuzzy direct power control (FDPC) improves performance through its adaptability to uncertainties, making it effective for managing hybrid load variability. Vector control is characterised by potentially longer response times and increased complexity. By utilising Simulink MATLAB for simulations, this research aims to thoroughly compare these strategies, ultimately guiding the selection of the most suitable approach based on specific operational needs.

REFERENCES

- [1] J. Zheng, Z. Zhou, J. Zhao, J. Wang, “Integrated heat and power dispatch truly utilizing thermal inertia of district heating network for wind power integration,” *Applied Energy*, vol. 211, pp. 865-874, 2018.
- [2] J.A. Baroudi, V.Dinavahi, A. M. Knight, “A review of power converter topologies for wind generator” *Renew. Energy*, vol. 32, no 14, p. 2369- 2385, Nov. 2007.

- [3] Z. Gadouche, C. Belfedal, T. Allaoui, and B. Belabbas, “Speed-sensorless DFIG wind turbine for power optimization using fuzzy sliding mode observer,” *Int. J. Renewable Energy Res.*, vol. 7, no. 2, pp. 613–621, 2017.
- [4] Z. Gadouche, C. Belfedal, T. Allaoui, M. Denai, and M. Bey, “Hybrid renewable energy system controlled with intelligent direct power control,” *Journal Européen des Systèmes Automatisés*, vol. 55, no. 4, pp. 467–475, 2022, doi: 10.18280/jesa 550405.
- [5] K. F. Sayeh, S. Tamalouzt, D. Ziane, N. Benyahia, S. L. Belaid, and Y. Belkhier, “Real-time fuzzy logic-based direct power control for wind energy systems,” *Engineering Applications of Artificial Intelligence*, vol. 154, 2025.
- [6] G. Tahri, Z. A. Foitih, and A. Tahri, “Fuzzy logic control of active and reactive power for a grid-connected photovoltaic system using a three-level neutral-point-clamped inverter,” *International Journal of Power Electronics and Drive Systems*, vol. 12, pp. 453–462, 2021.
- [7] R. Moutchou and A. Abbou, “Comparative study of SMC and PI control of a permanent magnet synchronous generator decoupled by singular perturbations,” in *Proc. 7th Int. Renewable and Sustainable Energy Conf. (IRSEC)*, 2019, pp. 1–7
- [8] Z. Gadouche, C. Belfedal, T. Allaoui, and B. Belabbas, “Commande de puissance active et réactive d’une MADA utilisée dans un système éolien,” in *Proc. 3rd Int. Seminar on New and Renewable Energies*, Oct. 2014.
- [9] A. Gaillard, “Système éolien basé sur une MADA : Contribution à l’étude de la qualité de l’énergie électrique et de la continuité de service,” Ph.D. dissertation, Univ. Henri Poincaré, Nancy I, France, 2010.
- [10] A. Bouafia, F. Krim, and J.-P. Gaubert, “Fuzzy-logic-based switching state selection for direct power control of three-phase PWM rectifier,” *IEEE Transactions on Industrial Electronics*, vol. 56, no. 6, pp. 1984–1992, 2009.
- [11] A. Moualdia, M. Mahmoudi, L. Nezli, and O. Bouchhida, “Modelling and control of a wind power conversion system based on the double-fed asynchronous generator,” *International Journal of Renewable Energy Research*, vol. 2, no. 2, pp. 301–306, 2012.
- [12] A. Elnady, “Direct Power Control Applied on 5-Level Diode Clamped Inverter Powered by a Renewable Energy Source,” *International Journal of Energy and Power Engineering*, vol. 12, no. 6, pp. 155–160, 2018.
- [13] A. Bouyekni, R. Taleb, Z. Boudjema, and H. Kahal, “A second-order continuous sliding mode based on DPC for wind-turbine-driven DFIG,” *Elektrotehniški Vestnik*, vol. 85, no. 1–2, pp. 29–36, 2018.
- [14] L. Dalessandro, S. Round, and J. Kolar, “Centre-point voltage balancing of hysteresis current controlled three-level PWM rectifiers,” *IEEE Transactions on Power Electronics*, vol. 23, no. 5, pp. 2477–2488, 2008.
- [15] M. Pichan, H. Rastegar, and M. Monfared, “Two fuzzy-based direct power control strategies for doubly-fed induction generators in wind energy conversion systems,” *Energy*, vol. 51, no. 2, pp. 154–162, 2013.

Efficient Transfer Learning for Brain Tumor Classification: Comparative Analysis of Pre-Trained CNN Models

Mohamed Ammara, *Student Member, IEEE*
LSTE Laboratory
University of Mascara
 Mascara, Algeria
 mohamed.ammara@univ-mascara.dz

Abstract—The classification of a brain tumor from MRI images is still a challenge to medical diagnostics. This work presents a streamlined comparison of six architectures of pre-trained CNNs for automated brain tumor classification by using transfer learning. VGG19, ResNet50V2, InceptionV3, DenseNet201, MobileNetV2, and Xception were applied to a dataset of 7,023 MRI images with four classes (glioma, meningioma, no tumor, pituitary) and with an architecture that had been simplified for efficient deployment. ResNet50V2 exhibited the best test accuracy of 92.14% with only 10.3 minutes of training time, closely followed by DenseNet201 (92.07%), and Xception (91.99%). Per-class analysis demonstrated exceptional performance for "no tumor" detection (98% recall) across all models, while meningioma remained the most challenging class to identify (81-87% recall). Training-validation gaps ranged from 8.15% to 13.34%, indicating varying generalization capabilities. Our results demonstrate that simplified CNN architectures with minimal preprocessing supported by transfer learning can achieve clinically relevant accuracy with significantly reduced computational requirements, making them suitable for resource-constrained clinical environments.

Index Terms—brain tumor, CNN, transfer learning, MRI, deep learning, medical imaging

I. INTRODUCTION

Brain tumors require early and accurate diagnosis for effective treatment planning. Traditional diagnostic methods based on expert radiological interpretation are time-consuming and subject to inter-observer variability [1], [2]. Deep learning, more precisely CNNs, has fully transformed medical image analysis using automatic hierarchical feature learning from raw imaging data [3], [4].

Therefore, by allowing the reuse of pre-trained models for specific tasks, transfer learning can reduce computational costs without sacrificing performance [5], [6]. However, most research uses a complicated architecture that includes several dense layers and immense pre-processing, making real-world deployment difficult [7], [8].

Our contributions are: (1) Systematic comparison of six CNN architectures under identical conditions, (2) Simplified architecture design (base + global average pooling + output) for efficient deployment, (3) Minimal preprocessing approach

using only normalization, and (4) Focus on computational efficiency, with training times below 18 minutes.

The rest of the paper is organized as follows: Sect. II reviews related work; Sect. III describes methodology; Sect. IV presents results; Sect. V discusses findings and limitations; Sect. VI concludes.

II. RELATED WORK

Deep learning-based methods have presented astonishing results for brain tumor classification. Kumar and Mankame [9] achieved an accuracy of 96.3%, while Badža and Barjaktarović [10] reported 96.56%. More recently, a performance boost was achieved by the concept of transfer learning. Disci et al. [1] reported 98.73% with Xception and advanced preprocessing. Rao et al. [11] presented efficient tumor detection using pre-trained models, and Rahman et al. [12] achieved 99.24% accuracy with VGG16 using extensive data augmentation.

Advanced architectures have achieved near-perfect accuracy: Nayak et al. [13] (99.97% with Dense Efficient-Net), Raza et al. [14] (99.67% with DeepTumorNet), and Maqsood et al. [15] (99.92% with hybrid MobileNetV2-SVM). However, these employ complex hybrid architectures which may limit practical deployment.

Recent work has focused on computational efficiency. Ganguly et al. [16] introduced lightweight architectures for resource-constrained environments, emphasizing the trade-off between accuracy and computational cost. Alshardan et al. [17] showed that even with simpler architectures, transfer learning with adequate fine-tuning results in very strong performance.

However, many high-accuracy approaches rely on complex hybrid architectures, extensive preprocessing pipelines, or aggressive data augmentation [13]–[15], which complicate reproducibility and practical deployment. Our work differs by prioritizing architectural simplicity and computational transparency: we systematically compare six widely-used CNNs under a unified feature-extraction protocol with minimal preprocessing, explicitly trading minor accuracy reductions for substantial gains in training efficiency and ease of reproduction.

III. METHODOLOGY

A. Dataset

We used the Brain Tumor MRI Dataset from Kaggle [18], which consists of 7,023 T1-weighted contrast-enhanced MRI slices. Each image represents a single 2D axial slice; no volumetric reconstruction or multi-sequence fusion was performed. The dataset contains 5,712 training and 1,311 testing images distributed across four classes: Glioma (1,321 train, 300 test), Meningioma (1,339 train, 306 test), No Tumor (1,595 train, 405 test), and Pituitary (1,457 train, 300 test), for a total of 1,621 glioma, 1,645 meningioma, 2,000 no-tumor, and 1,757 pituitary images. We used all provided slices as independent samples without additional slice selection or filtering. The training set was split into an 80/20 division for training and validation, respectively, using stratified sampling to preserve class distributions.

B. Preprocessing

All images underwent minimal preprocessing, consisting of only:

- 1) **Resizing:** Images resized to model-specific dimensions (224×224 for VGG19, ResNet50V2, DenseNet201, MobileNetV2; 299×299 for InceptionV3, Xception)
- 2) **Normalization:** Pixel values rescaled to [0,1] range by dividing by 255

No rotation, flipping, or zooming data augmentation techniques were used, prioritizing simplicity and deployment ease. The same preprocessing was applied to all sets: training, validation, and testing.

The term ‘minimal preprocessing’ in this work refers strictly to: (1) resizing to model-specific input dimensions, and (2) pixel normalization by rescaling to [0,1] range via division by 255. No additional preprocessing was applied—specifically, no skull-stripping, bias-field correction, contrast enhancement, histogram equalization, denoising, or registration to a standard atlas. This approach prioritizes deployment simplicity, computational efficiency, and reproducibility.

C. CNN Architectures

We considered the following six pre-trained architectures representing different design philosophies: **VGG19** [19] (19 layers with uniform 3×3 filters, 144M parameters), **ResNet50V2** [20] (50 layers with residual connections, 25M parameters), **InceptionV3** [21] (multi-scale convolutional filters, 24M parameters), **DenseNet201** [22] (dense connectivity pattern, 20M parameters), **MobileNetV2** [23] (depthwise separable convolutions for efficiency, 3.5M parameters), and **Xception** [24] (extreme inception with depthwise separable convolutions, 23M parameters).

All architectures used weights pre-trained on ImageNet in order to benefit from features learned on large-scale image recognition.

D. Model Configuration

For efficient deployment, we used a streamlined architecture consisting of:

- 1) Pre-trained CNN base (ImageNet weights, all convolutional layers frozen with trainable=False)
- 2) Global Average Pooling layer (reduces spatial dimensions to single vector)
- 3) Output Dense layer (4 neurons with Softmax activation for class probabilities)

This feature-extraction transfer learning approach preserves the pre-trained ImageNet feature representations in the frozen base model while training only the final classification head (GlobalAveragePooling2D + Dense layers). Depending on the architecture, between 3.5M and 144M pre-trained parameters remained frozen, while only approximately 4,000 parameters in the output layer were trained. This configuration significantly reduces training time, computational requirements, and overfitting risk compared to full fine-tuning, while maintaining competitive accuracy through effective feature reuse.

Training hyperparameters (identical for all models): Adam optimizer with initial learning rate of 0.001, categorical cross-entropy loss function, batch size of 32, maximum 42 epochs with ReduceLROnPlateau callback. The learning rate scheduler reduced the learning rate by a factor of 0.5 (multiplying by 0.5) when validation loss plateaued for 3 consecutive epochs, with a minimum learning rate of 1×10^{-7} . Early stopping monitored validation loss with patience of 42 epochs (full training duration) and restored best weights. All experiments were executed on Kaggle’s GPU environment (Tesla P100) using TensorFlow 2.15. Random seeds were fixed (`np.random.seed(42)`, `tf.random.set_seed(42)`) to ensure reproducible, deterministic behavior. Each architecture was trained once under these conditions, and reported metrics correspond to these single-run results.

E. Evaluation Metrics

Standard metrics were used to evaluate the model’s performance. For a multi-class classification problem with classes C , let TP_i , FP_i , FN_i , and TN_i represent true positives, false positives, false negatives, and true negatives for class i , respectively.

Accuracy is the total proportion of correct predictions:

$$\text{Accuracy} = \frac{\sum_{i=1}^C TP_i}{\sum_{i=1}^C (TP_i + FP_i + FN_i + TN_i)} \quad (1)$$

Precision is the ratio of correctly predicted positive values among all positive predictions:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \quad (2)$$

Recall (sensitivity) is the proportion of actual positives correctly identified:

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \quad (3)$$

F1-Score is the harmonic mean of precision and recall:

$$\text{F1-Score}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad (4)$$

For multi-class evaluation, we report **macro-averaged** metrics:

$$\text{Metric}_{\text{macro}} = \frac{1}{C} \sum_{i=1}^C \text{Metric}_i \quad (5)$$

making sure that all tumor types are of equal importance, independent of class size. The **generalization gap** quantifies overfitting:

$$\text{Gap} = \text{Accuracy}_{\text{train}} - \text{Accuracy}_{\text{val}} \quad (6)$$

IV. EXPERIMENTAL RESULTS

All six models were trained and evaluated under identical experimental conditions following the protocol described in Section III: the same stratified 80/20 train-validation split from the training directory, identical feature-extraction transfer learning configuration (frozen base + trainable head), and the same hyperparameters, callbacks, and random seed.

A. Overall Performance

Table I shows overall measures for all models on the test set. ResNet50V2 obtained the best score with 92.14%, followed by DenseNet201 with 92.07% and Xception with 91.99%. MobileNetV2 reached 89.02%, followed by InceptionV3 with 88.02%, while VGG19 obtained the lowest score of 80.40%. For training, all the CNN models are below 18 min, being faster ResNet50V2, which takes 10.3 min, proving our simplified architecture is fast and efficient.

TABLE I: Overall Performance Comparison (Test Set, N=1,311)

Model	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)	Time (min)
ResNet50V2	92.14	92.12	91.59	91.73	10.28
DenseNet201	92.07	91.75	91.63	91.64	12.19
Xception	91.99	91.62	91.52	91.52	17.32
MobileNetV2	89.02	88.74	88.39	88.45	12.20
InceptionV3	88.02	87.55	87.26	87.26	15.99
VGG19	80.40	79.43	79.73	79.22	11.88

B. Training Dynamics and Convergence

Figure 1 displays the train and validation accuracy over 42 epochs. ResNet50V2, DenseNet201, and Xception demonstrated gradual and continuous convergence, achieving more than 89% validation accuracy. VGG19 converged slower and reached the lowest final validation accuracy of 76.16%, which points to an inability of the simpler uniform filter design to tackle this dataset. MobileNetV2 and InceptionV3 exhibited stable learning with accuracies falling around 85-86% validation.

Table II quantifies generalization behaviour. Xception showed the best generalization with a gap of 8.15%, followed by DenseNet201 (9.03%). ResNet50V2 had the highest

test accuracy but showed a gap of 10.36%, hence slightly overfitting. VGG19 had the biggest gap of 13.34% and the lowest validation accuracy, which is an indication of limited model capacity for this task with our minimal pre-processing approach.

TABLE II: Training Dynamics and Generalization Analysis

Model	Train Acc. (%)	Val. Acc. (%)	Gap (%)	Epochs
ResNet50V2	99.41	89.04	10.36	42
DenseNet201	98.51	89.48	9.03	42
Xception	97.55	89.40	8.15	42
MobileNetV2	97.68	85.71	11.97	42
InceptionV3	97.09	85.89	11.20	42
VGG19	89.50	76.16	13.34	42

C. Per-Class Performance Analysis

Tables III and IV outline metrics of the top performing models. Both performed well for the "no tumor" classification (95-97% precision, 98% recall), crucial for screening applications because missed pathologies occur as false negatives. Pituitary tumors also performed well in recall (97-98%), suggesting distinctive imaging characteristics, which are well-represented by CNNs.

However, meningioma remained challenging with recall of 80.7-87.3%, indicating 13-19% misclassification rate. This is in keeping with the clinical observation that meningiomas demonstrate varied imaging appearances which may overlap with other tumor categories.

TABLE III: ResNet50V2 Per-Class Performance (Test Accuracy: 92.14%)

Class	Prec. (%)	Rec. (%)	F1 (%)	Supp.
Glioma	95.06	83.33	88.81	300
Meningioma	81.90	87.25	84.49	306
No Tumor	95.47	98.77	97.09	405
Pituitary	96.04	97.00	96.52	300

TABLE IV: Xception Per-Class Performance (Test Accuracy: 91.99%)

Class	Prec. (%)	Rec. (%)	F1 (%)	Supp.
Glioma	92.15	90.00	91.06	300
Meningioma	86.67	80.72	83.59	306
No Tumor	94.98	98.02	96.48	405
Pituitary	92.70	97.33	94.96	300

D. Confusion Matrix Analysis

Confusion matrices for all models are given in Figure 2. The main misclassifications of ResNet50V2 were the following: Glioma→Meningioma (49 cases, 16.3%), Meningioma→No

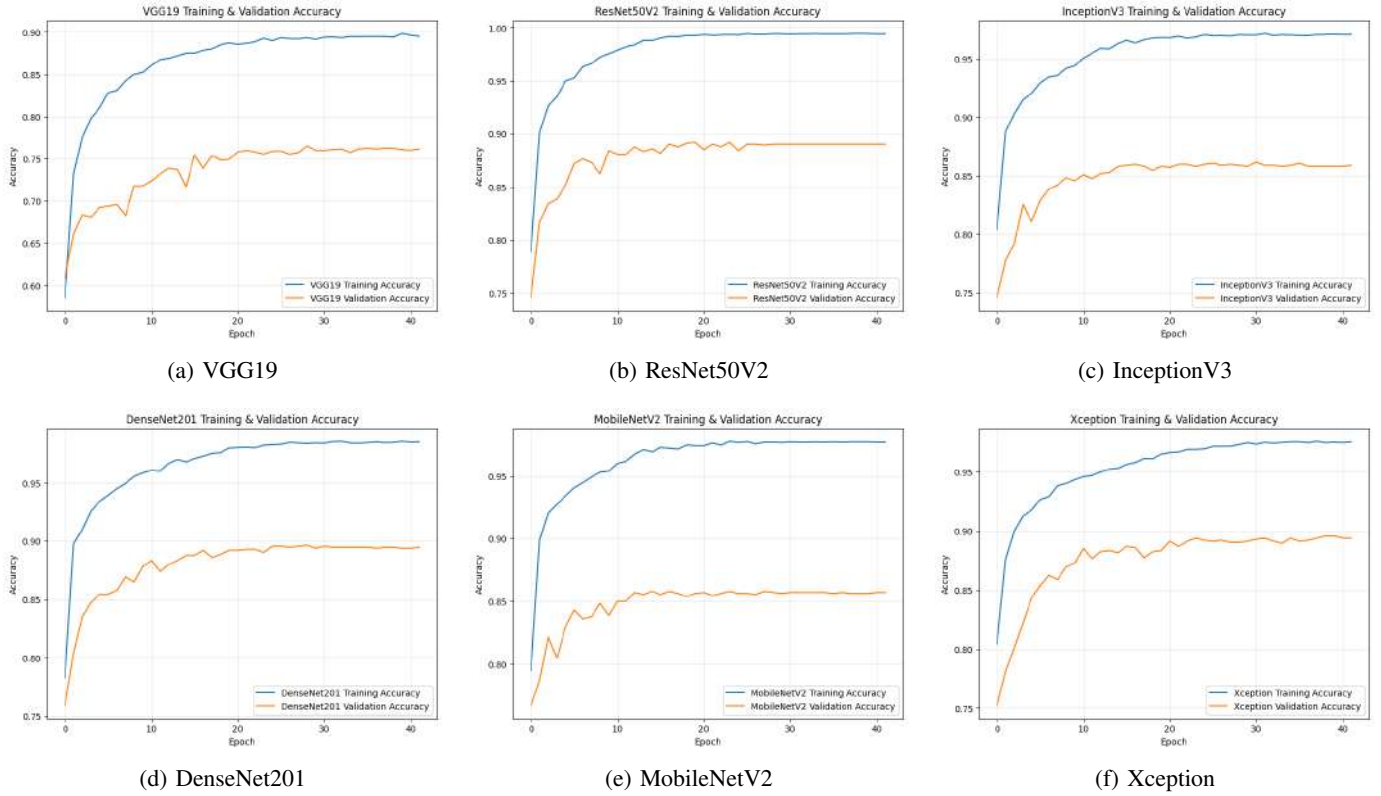


Fig. 1: Training and validation accuracy curves for all six CNN architectures across 42 epochs. The solid lines represent training accuracy, while the dashed lines represent the validation accuracy. Smaller gaps are indicative of good generalization.

Tumor (19 cases, 6.2%). Xception had similar patterns: Glioma→Meningioma (27 cases, 9.0%), Meningioma→No Tumor (21 cases, 6.9%).

The systematic glioma-meningioma confusion across all models suggests an overlap of MRI features which challenges automated classification without additional clinical context. The "no tumor" class showed consistently low error rates (<3%), which validated the clinical utility for initial screening.

V. DISCUSSION

A. Performance Analysis

Among them, ResNet50V2 had the best overall performance of 92.14% and the fastest training time of 10.3 minutes; thus, it is recommended for practical deployment. Its much-simplified architecture of a base, global average pooling, and output has greatly reduced the computational cost while yielding very competitive accuracy. This shows that complicated multi-layer classifiers with dropout are not necessary for effective transfer learning when the pre-trained features are sufficiently discriminative.

Minimal preprocessing thus worked well, and the approach with only normalization scored 92%, without any data augmentation. While several augmentation techniques improve performance, as demonstrated by Rahman et al. [12], our results present competitive accuracy with dramatically reduced complexity in the preprocessing, which therefore facilitates its easier deployment in clinical settings.

Recent works corroborate this efficiency-oriented approach. Rao et al. [11] report 88-92% with comparable pre-trained models; whereas Ganguly et al. [16] highlighted computational efficiency as a prime factor for resource-constrained environments. The training times of 10-18 minutes allow for fast updates of the model and iterative development.

B. Clinical Implications

The high "no tumor" recall of 98% demonstrates strong screening potential. The 2% false negative rate is acceptable in initial screening applications where the main objectives are usually to distinguish abnormal from normal scans. However, all positive predictions would need radiologist confirmation under clinical protocols.

The meningioma miss rate of 13-19% requires caution when clinical deployment is considered. The models are appropriate for decision support applications on triaging—that is, urgency prioritization of gliomas while flagging ambiguous cases to expert reviewers—but not as stand-alone diagnostic systems [25].

Systematic glioma-meningioma confusion suggests that supplemental imaging sequences (T2-weighted, FLAIR) or clinical features (patient age, symptom duration) may contribute to improved differentiation. The simplified architecture allows for quick integration of multi-modal inputs by concatenation of features.

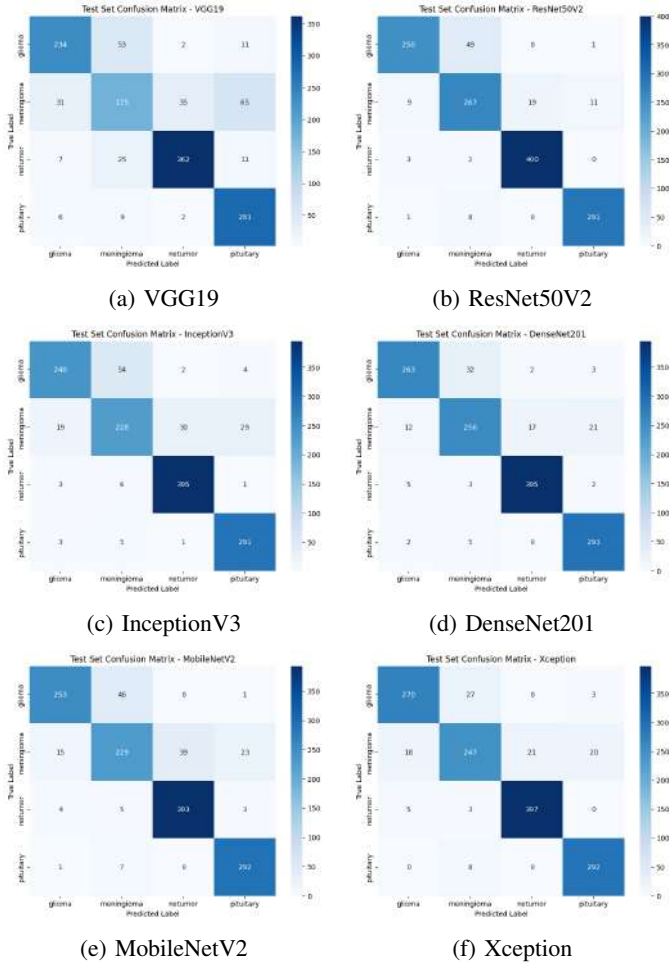


Fig. 2: Confusion matrices on test set (N=1,311). Rows represent true labels, columns represent predicted labels. Higher diagonal values indicate better per-class performance.

C. Computational Efficiency

The dramatic reduction in training time of 10-18 minutes compared to complex architectures makes our approach particularly suitable for resource-constrained environments. This efficiency enables: (1) rapid prototyping and hyperparameter exploration, (2) frequent model updates as new training data becomes available, (3) deployment in facilities with limited computational infrastructure, and (4) federated learning across multiple institutions without requiring extended computational resources at each site.

This is further reduced by the minimal preprocessing-normalization only, which excludes complex augmentation pipelines and makes the system more accessible to clinicians who do not have experience in deep learning.

A limitation of this study is the absence of comprehensive deployment benchmarking beyond training time. Critical metrics such as inference latency per image, model file size, and runtime memory footprint were not systematically measured across different hardware platforms (GPU, CPU, mobile, edge devices). While the simplified architecture and frozen-base approach substantially reduce training requirements (10–18

minutes), future work will include profiling inference performance and memory usage to more rigorously substantiate deployment feasibility claims for resource-constrained clinical environments.

D. Limitations and Future Work

Several limitations should be acknowledged: the single-source dataset may not represent diverse patient populations, scanner manufacturers, imaging protocols, or tumor subtypes encountered in multi-center clinical practice. External validation on datasets from different institutions is essential before clinical deployment. The minimal preprocessing approach promotes simplicity but may leave behind possible accuracy gains provided by advanced techniques, such as tumor segmentation or contrast enhancement. Future work should methodically analyze the accuracy-complexity trade-off. Class imbalance—405 “no tumor” versus 300-306 tumor samples—may lead model optimization to be biased toward the majority class. It would be worth investigating balanced sampling strategies or class-weighted loss functions. The single-sequence limitation—a T1-weighted contrast-enhanced image alone does not reflect the clinical practice of interpreting multiple MRI sequences in combination. Multi-modal fusion that incorporates T1, T2, FLAIR, and DWI sequences may greatly enhance diagnostic accuracy.

Statistical rigor: All reported results derive from single training runs per architecture using a fixed random seed (42). While this ensures exact reproducibility of our experiments, it does not account for stochasticity in weight initialization and training dynamics. Future work should include multiple training runs with different random initializations and report confidence intervals or standard deviations. Statistical significance tests (e.g., paired t-tests, McNemar’s test) should be applied when comparing model performance to rigorously quantify whether observed differences are statistically meaningful rather than artifacts of random variation.

The future directions would entail the following: multi-center external validation across different populations and various imaging protocols, including integration of multiple MRI sequences by feature fusion; assessment of advanced architectures; prospective clinical studies that evaluate clinical impact on radiologist workflow and diagnostic accuracy; investigation of ensemble methods combining multiple architectures for improved robustness.

VI. CONCLUSION

This work systematically compares six CNN architectures for brain tumor classification under an efficient transfer learning framework, emphasizing simplicity and computational efficiency. ResNet50V2 achieved 92.14% test accuracy with a 10.3-minute training time, demonstrating the optimal balance of performance and efficiency.

We achieved significant simplification of the architecture—base model+global average pooling+output, combined with minimal preprocessing—normalization only—reduced the computational requirements substantially with minimal degrada-

tion in accuracy, and thus, rendering the approach apt for resource-constrained clinical environments. Per-class analysis demonstrated excellent performance for the "no tumor" class required during screening applications (98% recall), while meningioma classification (81-87% recall) requires improvement through multi-modal approaches.

Fast training times enable rapid adaptation to institution-specific imaging protocols and support federated learning across multiple sites. While not yet suitable as standalone diagnostic systems, these models serve as valuable decision support tools for radiologists, especially in initial triage and screening applications. Future work should focus on multi-center validation, multi-sequence integration, and prospective evaluation of clinical workflow impact.

REFERENCES

- [1] R. Disci, F. Gurcan, and A. Soyly, "Advanced brain tumor classification in MR images using transfer learning and pre-trained deep CNN models," *Cancers*, vol. 17, no. 1, p. 121, 2025.
- [2] G. S. Tandel *et al.*, "A review on a deep learning perspective in brain cancer classification," *Cancers*, vol. 11, no. 1, p. 111, 2019.
- [3] S. Deepak and P. M. Ameer, "Brain tumor classification using deep CNN features via transfer learning," *Comput. Biol. Med.*, vol. 111, p. 103345, 2019.
- [4] A. B. Abdusalomov, M. Mukhiddinov, and T. K. Whangbo, "Brain tumor detection based on deep learning approaches and MRI," *Cancers*, vol. 15, no. 16, p. 4172, 2023.
- [5] Z. N. K. Swati *et al.*, "Brain tumor classification for MR images using transfer learning and fine-tuning," *Comput. Med. Imaging Graph.*, vol. 75, pp. 34–46, 2019.
- [6] M. M. Islam *et al.*, "Transfer learning architectures with fine-tuning for brain tumor classification," *Healthc. Anal.*, vol. 4, p. 100270, 2023.
- [7] Ö. Polat and C. Güngen, "Classification of brain tumors from MR images using deep transfer learning," *J. Supercomput.*, vol. 77, pp. 7236–7252, 2021.
- [8] S. K. Mathivanan *et al.*, "Employing deep learning and transfer learning for accurate brain tumor detection," *Sci. Rep.*, vol. 14, p. 7232, 2024.
- [9] S. Kumar and D. P. Mankame, "Optimization driven deep convolution neural network for brain tumor classification," *Biocybern. Biomed. Eng.*, vol. 40, no. 3, pp. 1190–1204, 2020.
- [10] M. M. Badža and M. Č. Barjaktarović, "Classification of brain tumors from MRI images using a CNN," *Appl. Sci.*, vol. 10, no. 6, p. 1999, 2020.
- [11] K. N. Rao *et al.*, "An efficient brain tumor detection and classification using pre-trained CNN models," *Heliyon*, vol. 10, no. 17, p. e36773, 2024.
- [12] M. M. Rahman *et al.*, "Brain tumor classification using MRI images and deep learning techniques," *PLOS ONE*, vol. 20, no. 5, p. e0303844, 2025.
- [13] D. R. Nayak *et al.*, "Brain tumor classification using Dense Efficient-Net," *Axioms*, vol. 11, no. 1, p. 34, 2022.
- [14] A. Raza *et al.*, "A hybrid deep learning-based approach for brain tumor classification," *Electronics*, vol. 11, no. 7, p. 1146, 2022.
- [15] S. Maqsood, R. Damaševičius, and R. Maskeliūnas, "Multi-modal brain tumor detection using deep neural network and multiclass SVM," *Medicina*, vol. 58, no. 8, p. 1090, 2022.
- [16] P. Ganguly *et al.*, "Efficient brain tumor classification with lightweight CNN architectures," *Expert Syst. Appl.*, vol. 238, p. 121847, 2025.
- [17] A. Alshardan *et al.*, "Leveraging transfer learning-driven CNNs for enhanced brain tumor detection," *Sci. Rep.*, vol. 14, p. 29847, 2024.
- [18] M. Nickparvar, "Brain Tumor MRI Dataset," Kaggle, 2021.
- [19] K. Simonyan and A. Zisserman, "Very deep CNNs for large-scale image recognition," *arXiv:1409.1556*, 2014.
- [20] K. He *et al.*, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.
- [21] C. Szegedy *et al.*, "Rethinking the inception architecture for computer vision," in *Proc. IEEE CVPR*, 2016, pp. 2818–2826.
- [22] G. Huang *et al.*, "Densely connected convolutional networks," in *Proc. IEEE CVPR*, 2017, pp. 4700–4708.
- [23] M. Sandler *et al.*, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE CVPR*, 2018, pp. 4510–4520.
- [24] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE CVPR*, 2017, p. 1251–1258.
- [25] Q. A. Ishfaq *et al.*, "Automatic smart brain tumor classification using deep learning," *Diagnostics*, vol. 15, no. 5, p. 842, 2025.
- [26] D. Karimi *et al.*, "Deep learning with noisy labels in medical image analysis," *Med. Image Anal.*, vol. 65, p. 101759, 2021.
- [27] M. M. Alanazi *et al.*, "A hybrid approach for multiclass brain tumor classification," *Rev. Intell. Artif.*, vol. 38, no. 4, pp. 1149–1160, 2024.

Comparative Study of Improved FCM Algorithms in Retinal Blood Vessel Segmentation

Imane Mehidi
DAC HR Laboratory
Ferhat Abbas University
Setif, Algeria
imane.mehidi@univ-setif.dz 

Djamel Eddine Chouaib Belkhiat
DAC HR Laboratory
Ferhat Abbas University
Setif, Algeria
djamel.belkhiat@univ-setif.dz 

Dalel Jabri
DAC HR Laboratory
Ferhat Abbas University
Setif, Algeria
dalel.jabri@univ-setif.dz 

Abstract—Retinal vessel segmentation aims to extract vascular structures from fundus images, which often present low contrast, uneven illumination, and noise. Fuzzy C-Means (FCM) and its improved variants are widely used for this purpose due to their effectiveness in medical image segmentation. This study evaluates eight FCM derivatives—FCM, EnFCM, SFCM, FGFCM, FRFCM, DSFCM_N, FCM_SICM, and SSFCA—based on noise robustness, segmentation accuracy, and computational efficiency. Experiments on synthetic and (DRIVE ,STARE) images show that FRFCM and DSFCM_N achieve superior noise handling and vessel segmentation, with FRFCM being the fastest. These results provide guidance for selecting optimal FCM based algorithms for retinal vessel segmentation.

Index Terms—Retinal, segmentation , clustering,vessel, comparative.

I. INTRODUCTION

The segmentation of blood vessels is crucial for interpreting medical images, as vessel analysis supports examination, treatment planning, and evaluation of clinical outcomes in fields such as neurosurgery, laryngology, and ophthalmology. In ophthalmology, vessel segmentation from fundus images is often performed manually, which is tedious, time-consuming, and prone to errors, particularly in large-scale screenings. While most pixels are easily rated by ophthalmologists, pathological and small vessels are challenging to identify, making automated segmentation essential for improving speed and accuracy [1]–[3].

Retinal fundus images are often degraded by noise, uneven illumination, and blurring due to acquisition conditions, vessel size variability, broken vessels at bifurcations or crossovers, and individual morphological differences. These factors complicate vessel classification. Clustering methods, particularly unsupervised algorithms like k-means, k-medoids, mean-shift, and Fuzzy C-Means (FCM), have been widely applied for image segmentation. FCM-based algorithms are especially effective because they allow fuzzy pixel membership, preserving more image information than hard segmentation methods.

To improve noise robustness and computational efficiency, numerous FCM derivatives have been proposed. These include Bias-Corrected FCM (BCFCM), Deviation Sparse FCM (DSFCM), Enhanced FCM (EnFCM), Fast Generalized FCM (FGFCM), Fast and Robust FCM (FRFCM), and Self-Sparse

FCM (SSFCA), among others. Recent studies have also introduced advanced methods such as IRW-FCM and Non-negative Latent Factor-based FCM for enhanced clustering accuracy and handling incomplete data [4]–[9].

Despite widespread use, the comparative performance of these FCM variants for retinal blood vessel segmentation has not been comprehensively investigated. This study addresses this gap by analyzing eight publicly available FCM derivatives FCM, EnFCM, SFCM, FGFCM, FRFCM, DSFCM_N, FCM_SICM, and SSFCA—based on noise robustness, segmentation performance, and computational efficiency. Noise robustness is assessed using synthetic images with Gaussian and Salt & Pepper noise. Segmentation performance is evaluated on DRIVE and STARE databases, with pre-processing via contrast-limited adaptive histogram equalization and bottom-hat filtering to enhance image details. Computational efficiency is measured by algorithm running times.

The study provides a detailed comparison of these FCM algorithms, identifies the most effective methods, and highlights avenues for future research to address the complexities of retinal image segmentation.

The remainder of this paper is organized as follows: Section II presents the theoretical background, Section III analyzes noise robustness, Section IV details blood vessel segmentation performance, and Section V concludes the study.

II. BACKGROUND

The Fuzzy C-Means (FCM) clustering technique is an unsupervised segmentation method that partitions an image into multiple groups or regions. It is widely applied in decision-making tasks and particularly in medical image analysis. Fuzzy set theory, introduced by [10], provides a powerful tool for soft partitioning. Based on this theory, [11] developed the FCM clustering method, later adapted by [12] for image segmentation. To address challenges in image segmentation, various FCM extensions have been proposed, with performance depending on image type and target region. This section reviews key FCM variants.

A. FCM

FCM is a soft segmentation algorithm commonly used in medical image analysis [13]. Its effectiveness depends

on initial cluster center positions and membership degrees. Let an image I contain P pixels with attributes $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_P \in \mathbb{R}^P$. FCM partitions X into C clusters by minimizing:

$$J_{FCM} = \sum_{i=1}^C \sum_{j=1}^P u_{ij}^m \|\mathbf{x}_j - \mathbf{v}_i\|^2 \quad (1)$$

subject to

$$\sum_{i=1}^C u_{ij} = 1, \text{ for } j \in \{1, 2, \dots, P\}, \text{ with } 0 \leq u_{ij} \leq 1 \quad (2)$$

where u_{ij} is the membership of pixel x_j in cluster i , $m > 1$ is the fuzzification exponent, $\mathbf{v}_i$ is the cluster center, and $\|\cdot\|$ denotes Euclidean distance. Memberships and cluster centers are iteratively updated:

$$u_{ij}^m = \frac{1}{\sum_{q=1}^C \left(\frac{\|\mathbf{x}_j - \mathbf{v}_i\|}{\|\mathbf{x}_j - \mathbf{v}_q\|} \right)^{\frac{2}{m-1}}} \text{ and } \mathbf{v}_i = \frac{\sum_{j=1}^n u_{ij}^m \cdot \mathbf{x}_j}{\sum_{j=1}^n u_{ij}^m} \quad (3)$$

Iteration stops when $\max_{ij} |u_{ij}^{(t+1)} - u_{ij}^{(t)}| < \varepsilon$.

B. EnFCM

Enhanced FCM (EnFCM) [14] reduces computational complexity by clustering gray-level histograms instead of pixel values. A linearly-weighted sum image ξ is formed:

$$\xi_l = \frac{1}{\alpha + 1} \left(x_l + \frac{\alpha}{N_R} \sum_{j \in N_l} x_j \right) \quad (4)$$

where x_j are neighbors of x_l , N_l is the local window, and α controls neighbor influence. Clustering is then applied to the histogram of ξ :

$$J_{EnFCM} = \sum_{i=1}^c \sum_{l=1}^q \gamma_l u_{il}^m (\xi_l - \mathbf{v}_i)^2 \quad (5)$$

with $\sum_{l=1}^q \gamma_l = N$ and $\sum_{i=1}^c u_{il} = 1$. Memberships and centers are updated:

$$u_{il} = \left[\sum_{j=1}^c \left(\frac{\xi_l - \mathbf{v}_i}{\xi_l - \mathbf{v}_j} \right)^{2/(m-1)} \right]^{-1}, \quad \mathbf{v}_i = \frac{\sum_{l=1}^q \gamma_l u_{il}^m \xi_l}{\sum_{l=1}^q \gamma_l u_{il}^m} \quad (6)$$

Multi-objective optimization has been used in recent studies to tune α for noisy images [15].

C. SFCM

Spatial FCM (SFCM) [6] incorporates spatial information into memberships:

$$h_{ij} = \sum_{k \in N_b} u_{ik}, \quad u_{ij} = \frac{u_{ij}^p h_{ij}^q}{\sum_{k=1}^c u_{kj}^p h_{kj}^q} \quad (7)$$

where N_b is a local window, and p, q control relative contributions. SFCM is widely used for brain MRI and other medical images [16].

D. FGFCM

Fast Generalized FCM (FGFCM) [5] integrates a local similarity measure S_{ij} to preserve details and improve noise immunity, avoiding the parameter tuning required by EnFCM. The gray-level image ξ_i is updated as:

$$\xi_i = \frac{\sum_{j \in N_i} S_{ij} x_j}{\sum_{j \in N_i} S_{ij}} \quad (8)$$

where S_{ij} accounts for spatial and intensity similarities.

E. FRFCM

Fast and Robust FCM (FRFCM) [17] uses morphological reconstruction and membership filtering to improve accuracy and reduce computation:

$$J_{FRFCM} = \sum_{j=1}^N \sum_{i=1}^c \gamma_j u_{ij}^m |x_j - y_i| - \lambda \left(\sum_{i=1}^c u_{ij} - 1 \right) \quad (9)$$

F. DSFCM_N

Deviation-Sparse FCM with neighbor information (DSFCM_N) [4] incorporates deviations e_k between measured and theoretical values with a sparsity constraint:

$$J_{DSFCM\_N} = \sum_{i=1}^c \sum_{j=1}^N u_{ij}^m \sum_{k \in N_j} \frac{|x_k - e_k - y_i|^2}{1 + d_{kj}} + \sum_t \lambda_t \sum_{k \in N_j} \frac{|e_{jt}|^p}{1 + d_{kj}} \quad (10)$$

G. FCM_SICM

FCM with adaptive spatial, intensity, and membership constraints (FCM_SICM) [7] uses bilateral filtering and membership linking to improve robustness to noise:

$$J_{FCM\_SICM}^{(t)} = \sum_{i=1}^K \frac{\sum_{j=1}^N \frac{1}{\Delta y_j} (u_{ij}^{(t)})^m |y_j - c_i^{(t)}|^2 + \sum_{j=1}^N \Delta y_j (u_{ij}^{(t)})^m |\bar{y}_j - c_i^{(t)}|^2}{\ln^2 \left(\sum_{e=1}^N u_{ie}^{(t-1)} + 1 \right)} \quad (11)$$

H. SSFCA

Self-Sparse FCM (SSFCA) [8] introduces a Gaussian-based sparsity regularization to reduce noisy features:

$$J_{SSFCA} = \sum_{i=1}^c \sum_{j=1}^n u_{ij} \Phi'(\mathbf{x}_j | \mathbf{v}_i, \Sigma_i) + \gamma \sum_{i=1}^c \sum_{j=1}^n u_{ij}^2 \quad (12)$$

Cluster centers and covariance matrices are updated iteratively.

Remarks:

- All improved FCM algorithms share fuzzy theory principles and common settings (fuzzification exponent, convergence threshold, max iterations), but algorithm-specific parameters require tuning.
- Many variants address conventional FCM shortcomings, e.g., ignoring spatial information.
- SFCM, FGFCM, FRFCM, and FCM_SICM incorporate local spatial information through neighborhood windows.
- DSFCM_N and SSFCA integrate deviations and sparsity to improve noise robustness.

- EnFCM and FRFCM reduce computational complexity using histogram-based clustering and membership filtering, respectively.

III. NOISE ROBUSTNESS ANALYSIS

To evaluate the noise robustness of the improved FCM algorithms in image segmentation, experiments were conducted on a synthetic image of size 269×260 containing three intensity classes (75, 100, 175) [17] (Figure 1(a), Figure 2(a)). The image was corrupted with varying levels of Gaussian and Salt & Pepper noise. Common parameters were set as: fuzzy exponent $m = 2$, convergence threshold $e = 10^{-5}$, and maximum iterations $T = 100$. Algorithm-specific parameters are listed in Table I.

TABLE I
PARAMETERS USED.

Methods	Parameters
EnFCM	$\alpha = 4.2$
SFCM	$w = 5, p = 1, q = 1$
FGFCM	$\alpha(\lambda_s, \lambda_g) = 6.2$
FRFCM	$w = 1, se = 1, \text{shape} = \text{disk}$
DSFCM_N	$w = 3, \lambda = 10$
FCM_SICM	$\sigma_d = 0.5, \sigma_r = 2$
SSFCA	$\gamma = 3$

A. Validation

Segmentation performance was evaluated using the standard metrics: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These were combined to calculate:

$$Pr = \frac{TP}{TP + FP}, \quad Sn = \frac{TP}{TP + FN}, \quad Sp = \frac{TN}{TN + FP}$$

$$F_1 = \frac{2TP}{2TP + FP + FN}, \quad Acc = \frac{TP + TN}{TP + TN + FP + FN}.$$

B. Results on a synthetic image

The FRFCM and DSFCM_N algorithms achieved the highest segmentation performance on images corrupted with Gaussian (2%) and Salt & Pepper (6%) noise, due to spatial information incorporation and sparsity-based noise suppression. Tables II and III, and Figures 1 and 2 show detailed results.

TABLE II
THE METRICS VALUES ACHIEVED BY DIFFERENT ALGORITHMS ON THE SYNTHETIC IMAGES CORRUPTED BY VARIOUS RATIOS OF GAUSSIAN NOISE.

Gaussian noise						
Ratio	Methods	Pr	Sp	Sn	F ₁	Acc
2%	FCM	0.9380	0.9711	0.7269	0.8793	0.8191
	EnFCM	0.9848	0.9914	0.9225	0.9655	0.9526
	SFCM	0.95861	0.9794	0.7890	0.9079	0.8656
	FGFCM	0.9855	0.9917	0.9342	0.9701	0.9592
	FRFCM	0.9991	0.9994	0.9962	0.9982	0.9977
	DSFCM_N	0.9951	0.9970	0.9986	0.9976	0.9968
	FCM_SICM	0.9465	0.9759	0.7082	0.8753	0.8102
	SSFCA	0.7305	0.8193	0.8137	0.8172	0.7698

TABLE III

THE METRICS VALUES ACHIEVED BY DIFFERENT ALGORITHMS ON THE SYNTHETIC IMAGES CORRUPTED BY VARIOUS RATIOS OF SALT & PEPPER NOISE.

Salt & Pepper noise						
Ratio	Methods	Pr	Sp	Sn	F ₁	Acc
6%	FCM	0.9519	0.9704	0.9704	0.9704	0.9611
	EnFCM	0.9409	0.9644	0.9408	0.9555	0.9409
	SFCM	0.9886	0.9931	0.9933	0.9932	0.9910
	FGFCM	0.9394	0.9636	0.9357	0.9531	0.9375
	FRFCM	0.9996	0.9998	0.9997	0.9997	0.9996
	DSFCM_N	0.9944	0.9967	0.9692	0.9864	0.9816
	FCM_SICM	0.9514	0.9702	0.9697	0.9700	0.9605
	SSFCA	0.9517	0.9704	0.9701	0.9703	0.9608

IV. BLOOD VESSEL SEGMENTATION ANALYSIS

A. Complexity of retinal images

Retinal images pose challenges such as low contrast, uneven illumination, and noise, which complicate vessel segmentation. Color variations across subjects further increase complexity, requiring robust pre-processing [3], [18].

B. Pre-processing method

Pre-processing aims to enhance image quality and contrast for better segmentation (Figure 3). Steps include: extraction of the green channel, contrast enhancement using CLAHE, and bottom-hat filtering. These improve vessel visibility and segmentation accuracy.

1) *Retinal image Components*: The green channel is preferred for vessel detection due to higher contrast. Pre-processing results (Figure 4) show improved visibility after CLAHE and bottom-hat filtering.

2) *CLAHE and Bottom-hat Filtering*: CLAHE enhances local contrast while limiting noise amplification [19], [20]. Bottom-hat filtering enhances dark vessels by subtracting the morphological closure of the image from the original:

$$BTH = \phi_S(I) - I. \quad (13)$$

Experiments show that pre-processing significantly improves FCM-based segmentation (Figure 5).

C. Experiments on retinal images

Eight FCM variants were tested on DRIVE and STARE databases (Table IV). Images were segmented into three clusters: background, vessels, and other tissues. Performance was measured against manually annotated ground truth using Pr, Sn, Sp, F₁, and Acc.

TABLE IV
CHARACTERISTICS OF DRIVE AND STARE DATABASES.

Database	DRIVE	STARE
Images	40	20
Camera	Canon CR5	TopCon TRV-50
FoV	45°	35°
Resolution	565 × 584	605 × 700
Format	JPEG	PPM
Ground truth	available	available

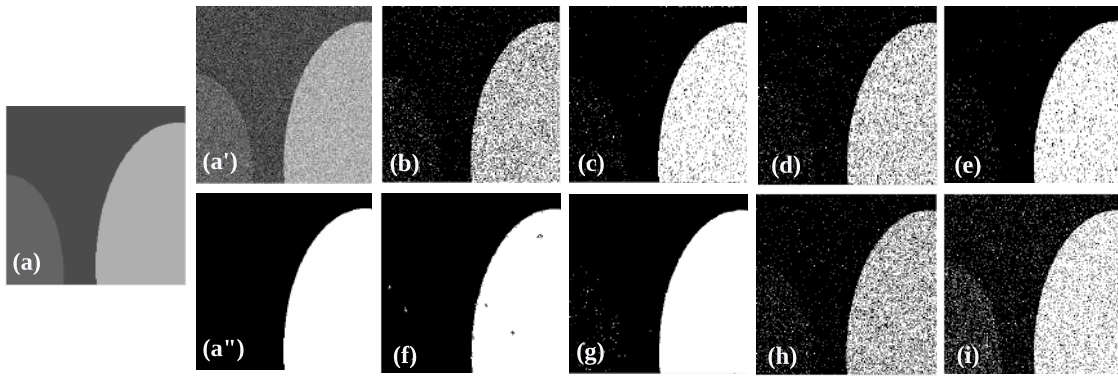


Fig. 1. Comparison of segmentation results on the synthetic image. (a) Original image. (a') Noisy image (Gaussian noise with zero mean and 2% variance). (a'') Ground truth. (b) FCM result. (c) EnFCM result. (d) SFCM result. (e) FGFCM result. (f) FRFCM result. (g) DSFCM_N result. (h) FCM_SICM result. (i) SSFCA result.

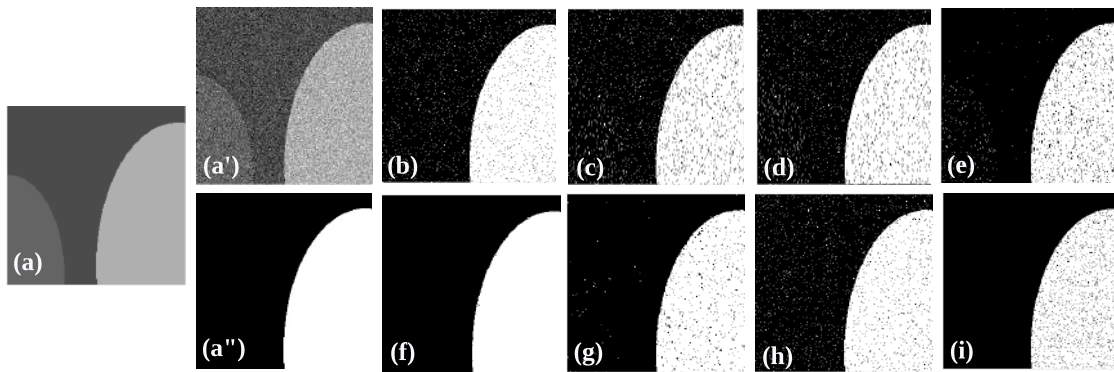


Fig. 2. Comparison of segmentation results on the synthetic image. (a) Original image. (a') Noisy image (Salt & Pepper, the noise intensity is 6%). (a'') Ground truth. (b) FCM result. (c) EnFCM result. (d) SFCM result. (e) FGFCM result. (f) FRFCM result. (g) DSFCM_N result. (h) FCM_SICM result. (i) SSFCA result.

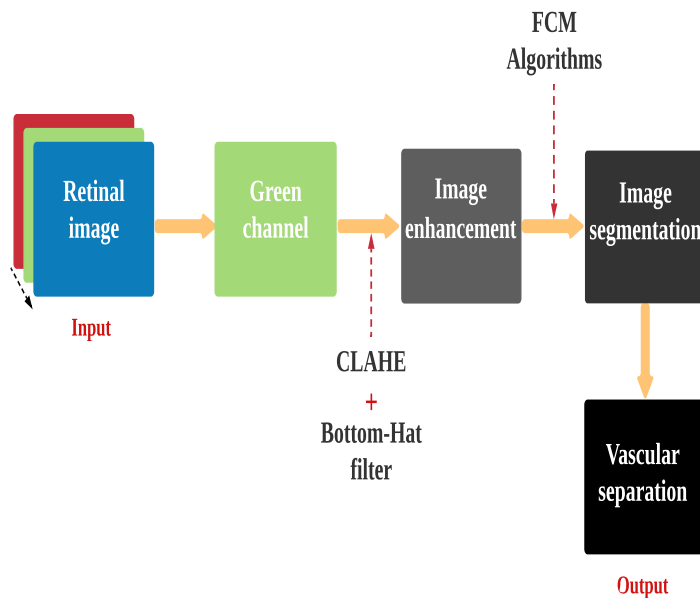


Fig. 3. Flowchart of the retinal image segmentation phases.

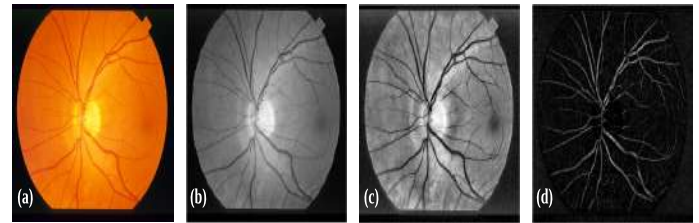


Fig. 4. Pre-processing for retinal image segmentation: (a) original color image, (b) green channel, (c) after CLAHE, (d) after bottom-hat filtering.

1) *Results on retinal images:* On DRIVE, DSFCM_N achieved the highest Acc (0.9522) and F_1 score (0.3323), while on STARE, FGFCM had the highest Acc (0.9517) and FRFCM achieved the best F_1 score (0.4201) (V, VI, Figures 6, 7). Results confirm that DSFCM_N, FRFCM, and FGFCM are most effective due to spatial and morphological processing.

2) *Running Time:* Execution times were measured on MATLAB using an Intel i5 2.5 GHz CPU (Table VII, Table VIII). FRFCM and EnFCM were the fastest, while SSFCA was significantly slower due to complex regularization. Overall, FRFCM and DSFCM_N offer the best trade-off between

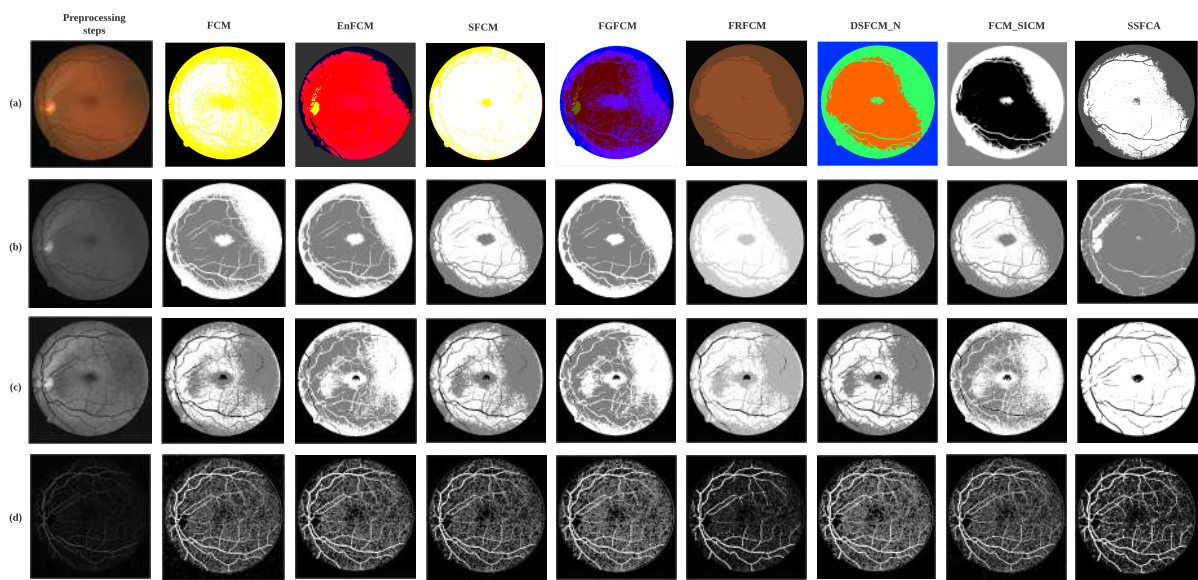


Fig. 5. Segmentation results with different pre-processing: (a) color image, (b) green channel, (c) after CLAHE, (d) after bottom-hat filtering.

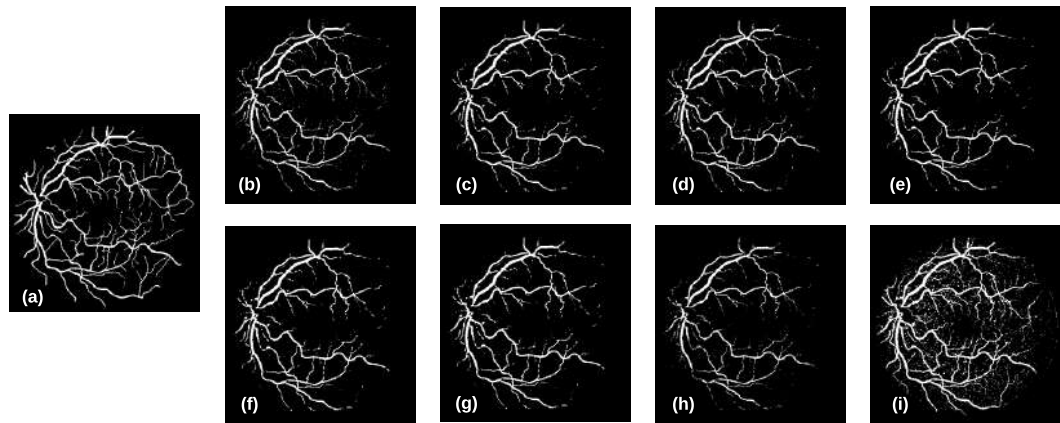


Fig. 6. Segmentation comparison on DRIVE database.

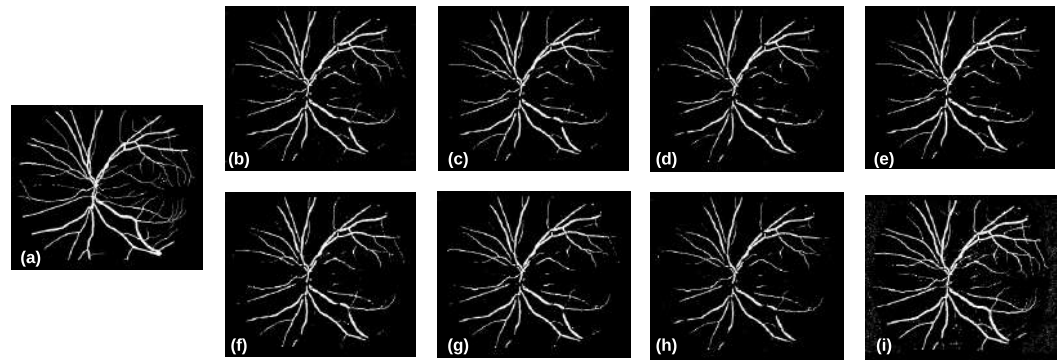


Fig. 7. Segmentation comparison on STARE database.

TABLE V
SEGMENTATION PRECISION (PR), SENSITIVITY (SE), SPECIFICITY (SP),
F₁ SCORE, AND ACCURACY (ACC) ACHIEVED BY DIFFERENT
ALGORITHMS ON THE DRIVE DATABASE.

DRIVE					
Methods	Pr	Sp	Sn	F ₁	Acc
FCM	0.8913	0.9939	0.4944	0.3180	0.9504
EnFCM	0.8831	0.9935	0.4902	0.3152	0.9497
SFCM	0.9331	0.9971	0.4012	0.2805	0.9453
FGFCM	0.7748	0.9819	0.5979	0.3374	0.9492
FRFCM	0.8803	0.9929	0.4996	0.3187	0.9499
DSFCM_N	0.8762	0.9922	0.5354	0.3323	0.9522
FCM_SICM	0.9231	0.9965	0.4165	0.2870	0.9460
SSFCA	0.6516	0.9581	0.7289	0.3440	0.9377

TABLE VI
SEGMENTATION PRECISION (PR), SENSITIVITY (SE), SPECIFICITY (SP),
F₁ SCORE, AND ACCURACY (ACC) ACHIEVED BY DIFFERENT
ALGORITHMS ON THE STARE DATABASE.

STARE					
Methods	Pr	Sp	Sn	F ₁	Acc
FCM	0.7322	0.9788	0.5727	0.3214	0.9474
EnFCM	0.8831	0.9802	0.5654	0.3447	0.9482
SFCM	0.8146	0.9911	0.4632	0.2952	0.9509
FGFCM	0.7804	0.9854	0.5464	0.3214	0.9517
FRFCM	0.7525	0.9815	0.5870	0.4201	0.9511
DSFCM_N	0.7264	0.9742	0.6059	0.3303	0.9455
FCM_SICM	0.7749	0.9847	0.5026	0.3048	0.9475
SSFCA	0.4872	0.9194	0.7919	0.3016	0.9089

accuracy, robustness, and computational efficiency.

TABLE VII
EXECUTION TIME (SECONDS) ON DRIVE DATABASE.

Methods	FCM	EnFCM	SFCM	FGFCM	FRFCM
Time	8.15	5.79	22.03	17.29	0.32
Methods	DSFCM_N	FCM_SICM	SSFCA		
Time	10.96	10.93	165.73		

TABLE VIII
EXECUTION TIME (SECONDS) ON STARE DATABASE.

Methods	FCM	EnFCM	SFCM	FGFCM	FRFCM
Time	11.73	7.29	31.15	21.80	0.40
Methods	DSFCM_N	FCM_SICM	SSFCA		
Time	13.78	14.05	243.31		

V. CONCLUSION

Eight improved FCM algorithms were evaluated for retinal vessel segmentation in terms of noise robustness, segmentation quality, and computational efficiency. FRFCM, followed by DSFCM_N, provides the best compromise across all criteria. Future work may include testing additional FCM variants, other databases (e.g., IDRiD, CHASE_DB1), and enhancing pre-processing with vessel-adapted filters or FoV masking to further improve accuracy and efficiency.

REFERENCES

- [1] M. Niemeijer, J. Staal, B. van Ginneken, M. Loog, and M. D. Abramoff, "Comparative study of retinal vessel segmentation methods on a new publicly available database," in *Medical imaging 2004: image processing*, vol. 5370. International Society for Optics and Photonics, 2004, pp. 648–656.
- [2] J. Almotiri, K. Elleithy, and A. Elleithy, "Retinal vessels segmentation techniques and algorithms: a survey," *Applied Sciences*, vol. 8, no. 2, p. 155, 2018.
- [3] A. A. Abdulsahib, M. A. Mahmoud, M. A. Mohammed, H. H. Rasheed, S. A. Mostafa, and M. S. Maashi, "Comprehensive review of retinal blood vessel segmentation and classification techniques: intelligent solutions for green computing in medical images, current challenges, open issues, and knowledge gaps in fundus medical images," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 10, no. 1, pp. 1–32, 2021.
- [4] Y. Zhang, X. Bai, R. Fan, and Z. Wang, "Deviation-sparse fuzzy c-means with neighbor information constraint," *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 1, pp. 185–199, 2018.
- [5] W. Cai, S. Chen, and D. Zhang, "Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation," *Pattern recognition*, vol. 40, no. 3, pp. 825–838, 2007.
- [6] K.-S. Chuang, H.-L. Tzeng, S. Chen, J. Wu, and T.-J. Chen, "Fuzzy c-means clustering with spatial information for image segmentation," *computerized medical imaging and graphics*, vol. 30, no. 1, pp. 9–15, 2006.
- [7] Q. Wang, X. Wang, C. Fang, and W. Yang, "Robust fuzzy c-means clustering algorithm with adaptive spatial & intensity constraint and membership linking for noise image segmentation," *Applied Soft Computing*, vol. 92, p. 106318, 2020.
- [8] X. Jia, T. Lei, X. Du, S. Liu, H. Meng, and A. K. Nandi, "Robust self-sparse fuzzy clustering for image segmentation," *IEEE Access*, vol. 8, pp. 146 182–146 195, 2020.
- [9] J. Xue, F. Nie, R. Wang, and X. Li, "Iteratively re-weighted algorithm for fuzzy c-means," *IEEE Transactions on Fuzzy Systems*, 2022.
- [10] L. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.
- [11] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973.
- [12] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Computers & geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.
- [13] P. Nageswara Reddy, C. Satyanarayana, and C. Mohan Rao, "Brain tumour segmentation using hybrid approach," in *Machine Learning and Internet of Things for Societal Issues*. Springer, 2022, pp. 117–125.
- [14] L. Szilagy, Z. Benyo, S. M. Szilagy, and H. Adam, "Mr brain image segmentation using an enhanced fuzzy c-means algorithm," in *Proceedings of the 25th annual international conference of the IEEE engineering in medicine and biology society (IEEE Cat. No. 03CH37439)*, vol. 1. IEEE, 2003, pp. 724–726.
- [15] M. Singh, C. Asha, and N. Sharma, "Multi-objective particle swarm optimization based enhanced fuzzy c-means algorithm for the segmentation of mri data," in *Recent Trends in Electronics and Communication*. Springer, 2022, pp. 1031–1041.
- [16] B. N. Li, C. K. Chui, S. Chang, and S. H. Ong, "Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation," *Computers in biology and medicine*, vol. 41, no. 1, pp. 1–10, 2011.
- [17] T. Lei, X. Jia, Y. Zhang, L. He, H. Meng, and A. K. Nandi, "Significantly fast and robust fuzzy c-means clustering algorithm based on morphological reconstruction and membership filtering," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 5, pp. 3027–3041, 2018.
- [18] A. Imran, J. Li, Y. Pei, J.-J. Yang, and Q. Wang, "Comparative analysis of vessel segmentation techniques in retinal images," *IEEE Access*, vol. 7, pp. 114 862–114 887, 2019.
- [19] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer vision, graphics, and image processing*, vol. 39, no. 3, pp. 355–368, 1987.
- [20] K. Zuiderveld, "Contrast limited adaptive histogram equalization," *Graphics gems*, pp. 474–485, 1994.

Non-PDC Control Law Synthesis for Discrete-Time Takagi-Sugeno Descriptor Systems subject to Input Saturation and $\mathcal{L}_2$ Disturbance

RIGHI Ines*¹

Electronic Department, Badji Mokhtar ANNABA University, Algeria
ines.righi@univ-annaba.dz

Abstract— This paper treats the stabilization analysis for a class of nonlinear descriptor systems subject to input saturation and unknown disturbance. The synthesized control strategy is based on the Takagi-Sugeno design approach and a non-parallel distributed compensation (non-PDC) control law. Stabilization conditions are derived using non-quadratic Lyapunov functions to ensure the closed-loop system stability despite the given saturation constraint on the control input. The optimization problem is formulated in terms of linear matrix inequalities (LMIs). The technique is illustrated through numerical examples.

Keywords—Takagi-Sugeno descriptor models, $\mathcal{L}_2$ disturbance, Actuator saturation; Stability, non-parallel distributed compensation, Linear matrix inequality (LMI).

Introduction

Recently, Takagi-Sugeno (T-S) models have been widely investigated to study nonlinear models [1]. Among nonlinear control theory, the T-S model-based approach has attracted great interest since it constitutes universal approximation of any smooth nonlinear function by a “blending” of some local linear system models. This method greatly facilitates observer/controller synthesis for complex nonlinear systems [2]. Based on this modeling technique, stability conditions have been obtained directly from Lyapunov methodology [3, 4]. For control design, the so-called parallel distributed compensation (PDC) has been the most commonly used scheme and remain to associate inferred state of output feedback to each local subsystem. Stabilization of T-S systems and the control design are investigated via the direct Lyapunov method.

In fact, various works deal with stability analysis using either quadratic Lyapunov functions [1] where common matrices to each subsystem have to be found, piecewise Lyapunov functions [3], more recently non quadratic Lyapunov functions [4, 5]. The derived conditions are formulated into a set of linear matrix inequalities (LMI). These LMIs can be solved, when a solution exists, by classical convex optimization algorithms [2, 4-6].

In practice, many systems are physically described by nonlinear descriptor models [6]. The T-S descriptor model representation has the advantage to decrease the number of LMI constraints since it conserves nonlinearities in the left-hand side while keeping the original structure of the nonlinear model [7, 11-15].

Usually real physical applications suffer from actuator saturation and/or sensor saturation. Thus, a great attention has been given to the control design of T-S models with input saturation constraints [8-9]. Among the most popular works dealing with saturated input constraints, the convexity based approach to the saturation function (see [9-10] and the references therein). The interest of this approach is to consider a bounded ellipsoidal symmetric region of stability solved by a set of LMIs. Moreover, descriptor design approaches have been recently studied in [11] to deal with problem of input saturated T-S systems using a polytopic representation of the saturation function. Moreover, these design control approaches are based on state feedback or dynamic output feedback with anti-windup (AW) mechanisms.

The TS descriptor fuzzy system is affected by unknown exogenous disturbances W_k . The main contribution of the new method can be summarized as the proposed result, can be applied to T-S systems subject to $\mathcal{L}_2$ disturbances.

In the present paper, stability analysis for discrete-time T-S descriptor system subject to input saturation is proposed. New LMI conditions are derived based on two Lyapunov functions with the Finsler's lemma [12] which allows decoupling the control law from the Lyapunov function.

This paper is organized as follows: Section 2 provides some useful notation and properties, it also introduces the discrete-time T-S descriptor model; Section 3 presents the main results for controller design for saturated discrete-time T-S descriptor models subject to $\mathcal{L}_2$ disturbances; Finally in section 4 designed examples are given to demonstrate the effectiveness of the proposed approaches.

I. NOTATION AND PROBLEM STATEMENT

Given a set of nonlinear functions $h_i(\cdot) \geq 0, i \in \{1 \dots r\}$ having the convex sum property, $\sum_{i=1}^r h_i(\cdot) = 1$; a shorthand notation will be used in the sequel to represent convex sum of matrix expressions: $Y_h = \sum_{i=1}^r h_i(z(k))Y_i$ and $Y_v = \sum_{i=1}^r v_k(z(k))Y_k$ for single convex sum; $Y_{h+} = \sum_{i=1}^r h_i(z(k+1))Y_i$ for a delayed convex sum; $Y_{h-} = (\sum_{i=1}^r h_i(z(k))Y_i)^{-1}$ for the inverse of a convex sum; and, $Y_{hh} = \sum_{i=1}^r \sum_{j=1}^r h_i(z(k))h_j(z(k))Y_{ij}$ for a doubled rested convex sum. For a vector x , x_k denotes its k -th entry, and x_{k+} denotes $x(k+1)$. $\mathbb{N}_r$ Denotes the set $\{1, 2, \dots, r\}$, $\mathbb{I}$

denotes the identity matrix.* stands for the terms deduced by symmetry in symmetric block matrices.

Let us consider the following discrete-time TS fuzzy model in the descriptor form subject to input saturation:

$$E_v x_{k+} = A_h x_k + B_h \text{sat}(u(k)) + B_w W_k \quad (1)$$

where $x \in \mathbb{R}^n$ is the state vector, $u \in \mathbb{R}^m$ is the control input vector, k is the current sample. Matrices A_i and $B_i, i \in \{1 \dots r\}$ represent the i -th linear right-hand side model and $E_v, k \in \{1 \dots r_e\}$, represents the k -th linear left-hand side model of T-S descriptor models. In the proposed work we suppose that E_v is regular matrix because it contains the inertia matrix which is motivated by mechanical systems. The membership functions MFs hold the convex sum property in a compact set of the state:

$$h_i(z(k)) \geq 0, i \in \{1 \dots 2^p\}, \sum_{i=1}^r h_i(z(k)) = 1 \quad (2.a)$$

$$v_k(z(k)) \geq 0, i \in \{1 \dots 2^{p_e}\}, \sum_{i=1}^r v_k(z(k)) = 1 \quad (2.b)$$

where p and p_e present the number of nonlinear terms in the right-hand and left-hand side, respectively of (1). The MFs depend on the premise variables grouped in the vector $z(k)$ which is known and usually depends on the state vector. The dead-zone nonlinearity $\psi(\cdot): \mathbb{R}^{n_u} \rightarrow \mathbb{R}^{n_u}$ is defined by :

$$\psi(u_k) = u_k - \text{sat}(u_k)$$

(2)

However the saturation function $\text{sat}: \mathbb{R}^{n_u} \rightarrow \mathbb{R}^{n_u}$ is defined as:

$$\begin{cases} \text{sat}(u_k) = [\text{sat}_1(u_{1k}) \dots \text{sat}_t(u_{tk}) \dots \text{sat}_{n_u}(u_{n_{uk}})] \\ \text{sat}_t(u_{tk}) = \text{sign}(u_{tk}) \min(|u_{tk}|, u_t^{\max}) \end{cases} \quad (3)$$

with $u_t^{\max} > 0$, designate the saturation level. In the closed loop control system, the nonlinear system, behavior of the input saturation has been investigated as a convex combination of 2^m linear models in [10, 11].

In order to obtain LMI conditions, the following relaxation lemma will be employed due to its good compromise between effectiveness and computational complexity.

Relaxation Lemma [15]. Let T_{ij}^k be matrices of appropriate dimensions, then for $i, j \in \{1 \dots r\}, k \in \{1 \dots r_e\}$:

$$\begin{cases} \sum_{i=1}^r \sum_{j=1}^r \sum_{k=1}^r h_i(z(k)) h_j(z(k)) v_k(z(k)) \text{ holds if } T_{ij}^k \\ \left(\frac{2}{r-1} \right) \left(\frac{2}{r_e-1} \right) T_{ii}^k + T_{ij}^k + T_{ji}^k < 0, i \neq j \end{cases} \quad (4)$$

Finsler's Lemma [16]. Let $x \in \mathbb{R}^n, G = G^T \in \mathbb{R}^{n \times n}$ and $R \in \mathbb{R}^{m \times n}$ such that $\text{rank}(R) < n$; the following expressions are equivalent:

$$\begin{cases} x^T G x < 0 \quad \forall x: \{x \in \mathbb{R}^n; x \neq 0, Rx = 0\} \\ \exists E \in \mathbb{R}^{m \times n}: G + ER + R^T E^T < 0 \end{cases} \quad (5)$$

Property 1. Let $X = X^T > 0$, and Y matrices of the appropriate size the following expressions holds:

$$(Y - X)^T X^{-1} (Y - X) \geq 0 \Leftrightarrow Y^T X^{-1} Y \geq Y + Y^T - X \quad (6)$$

Lemma 1 [16]. Given matrices $F_{jk} \in \mathbb{R}^{m \times n}$, $H_{jk} \in \mathbb{R}^{n \times n}$ and $W_{jk} \in \mathbb{R}^{m \times n}$, for $i, j, k \in \mathbb{N}_r, l \in \mathbb{N}_m$, let us define the following set:

$$\mathcal{P}_u = \{x \in \mathbb{R}^n: |(F_{jk} H_{jk}^{-1} - W_{jk} H_{jk}^{-1}) x_k| \leq u_{\max(l)}\}$$

If $x \in \mathcal{P}_u$, then $\psi(u_k)^T S_{jk}^{-1} [\psi(u_k) - (W_{jk} H_{jk}^{-1}) x_k] \leq 0$

holds of any positive diagonal matrices $S_{jk} \in \mathbb{R}^{m \times m}$, and for any scalar function $h_{ij}, i, j \in \mathbb{N}_r$, satisfying the convex sum property.

Assumption 1. The validity domain Ω_x of the system (1) is defined by:

$$\Omega_x = \{x \in \mathbb{R}^n: Q_m^T x \leq 1, m \in \mathbb{N}_q\} \quad (7)$$

where the vectors $Q_m \in \mathbb{R}^n$ are corresponding to the state constraints of system (1).

Assumption 1. The disturbance signal W_k belongs to the following class of function:

$$W_\delta^2 = \{W: \mathbb{R}^+ \rightarrow \mathbb{R}^{n_u}, \sum_{k=1}^\infty W_k^T W_k \leq 0\} \quad (8)$$

II. PREPARE YOUR PAPER BEFORE STYLING

In this section, the objective is to design a non-PDC control law ensuring the desired control performance and guarantying the stability of the closed loop perturbed system, despite the presence of input control saturation. Accordingly, the proposed non-PDC controller law is designed as follows:

$$u_k = F_{hv} H_{(\cdot)}^{-1} x_k \quad (9)$$

The combination of the T-S descriptor model (1) with the control law (9) yields:

$$E_v x_{k+} = A_h \cdot x_k + B_h (F_{hv} H_{(\cdot)}^{-1} x_k - \psi(u_k)) + B_w W_k \quad (10)$$

Expression (10) can be rewritten as an equality constraint:

$$\begin{bmatrix} A_h + B_h F_{hv} H_{(\cdot)}^{-1} & -E_v & -B_h & B_w \end{bmatrix} \begin{bmatrix} x_k \\ x_{k+} \\ \psi(u_k) \\ W_k \end{bmatrix} = 0 \quad (11)$$

In the following two different Lyapunov functions will be considered:

Case1: $V(x_k) = x_k^T P_h^{-1} x_{k+}$ with $P_h = P_h^T > 0, P_h^{-1} = X_h$

Case 2: $V(x_k) = x_k^T H_h^{-T} P_h H_h^{-1} x_k$, with $P_h = P_h^T > 0$

Case 1

The variation of the Lyapunov function in case 1 is calculated as:

$$\Delta V(x_k) = x_{k+}^T X_{h+} x_{k+} - x_k^T X_h x_k < 0 \quad (12)$$

By considering the saturation, the inequality (10) becomes:

$$x_{k+}^T X_{h+} x_{k+} - x_k^T X_h x_k + x_k^T x_k - 2 \cdot \psi(u_k)^T S_{hv}^{-1} \psi(u_k) + \psi(u_k)^T S_{hv}^{-1} W_{hv} H_{hv}^{-1} x_k + x_k^T x_k - \gamma W_k^T W_k < 0 \quad (13)$$

The inequality (13) written as:

$$\begin{bmatrix} x_k \\ x_{k+} \\ \psi(u_k) \\ W_k \\ 0 \end{bmatrix}^T \begin{bmatrix} -(X_h - I) & 0 & 0 & 0 \\ 0 & X_{h+} & 0 & 0 \\ S_{hv}^{-1} W_{hv} H_{hv}^{-1} & 0 & -2 \cdot S_{hv}^{-1} & 0 \\ 0 & 0 & 0 & -\gamma I \end{bmatrix} \begin{bmatrix} x_k \\ x_{k+} \\ \psi(u_k) \\ W_k \end{bmatrix} < 0 \quad (14)$$

Via the Finsler's Lemma, equality (11) and inequality (14)

$$\text{results in: } \begin{bmatrix} -(X_h - I) & 0 & 0 & 0 \\ 0 & X_{h+} & 0 & 0 \\ W_{hv} & -B_h^T N^T & -2 \cdot S_{hv}^{-1} & 0 \\ 0 & B_w^T & 0 & -\gamma I \end{bmatrix} +$$

$$\begin{bmatrix} M \\ N \\ 0 \\ 0 \end{bmatrix} [A_h + B_h F_{hv} H_{hv}^{-1} \quad -E_v \quad -B_h \quad B_w] + (*) < 0 \quad (15)$$

Where matrices $M \in R^{n \times n}$ and $N \in R^{n \times n}$ are free matrices fixed later on. Let us select $H_{(\cdot)} = H_{hv}$. Now, two results can be obtained depending on different congruence transformation of (15). This first one is stated in the following Lemma.

Lemma 2. The closed-loop T-S descriptor model (1) is asymptotically stable if there exist matrices $P_j = P_j^T > 0, H_{jk}, F_{jk}, W_{jk}$, a matrix $Q_m \in \mathbb{R}^n, m \in \mathbb{N}_q$ and positive diagonal matrices $S_{jk} \in \mathbb{R}^{m \times m}$, for $i, j, l \in \{1 \dots r\}, k \in \{1 \dots r_e\}$ and positive scalar γ such as conditions 4 are satisfied with:

$$T_{ijl}^k = \begin{bmatrix} -H_{jk}^T - H_{jk} + P_j & (*) & (*) & (*) \\ A_i H_{jk} + B_i F_{jk} & -E_k P_l - P_l E_k^T + P_l & (*) & (*) \\ W_{jk} & -S_{jk}^T B_i^T & -2S_{jk}^T & (*) \\ 0 & B_j^T & 0 & -\gamma I \end{bmatrix} < 0 \quad (16)$$

and

$$\begin{bmatrix} -P_j & (*) \\ Q_m P_j & -1 \end{bmatrix} \leq 0, m \in \mathbb{N}_q, j \in \mathbb{N}_r \quad (17)$$

$$\begin{bmatrix} -H_{jk}^T - H_{jk} + P_j & (*) \\ F_{jk(t)} - W_{jk(t)} & -(u_t^{max})^2 \end{bmatrix} \leq 0, t \in \mathbb{N}_m, j, k \in \mathbb{N}_r \quad (18)$$

Proof. From inequality (17) it can be deduced that $x \in \Omega_x$. Furthermore, if condition (17) is satisfied, then it follows clearly that matrices $H_{jk}, j, k \in \mathbb{N}_r$, are regular since $P_j > 0$. By Schur complement lemma [16] and matrix property 1, it can be deduced from (18) that

$$H_{jk}^T P_j H_{jk} - \frac{(F_{jk(t)} - W_{jk(t)})^T (F_{i(t)} - W_{jk(t)})}{(u_t^{max})^2} \geq 0 \quad (19)$$

Pre and post-multiplying (2) with H_{jk}^{-T} yields:

$$P_j - \frac{(F_{jk(t)} H_{jk}^{-1} - W_{jk(t)} H_{jk}^{-1})^T (F_{jk(t)} H_{jk}^{-1} - W_{jk(t)} H_{jk}^{-1})}{(u_t^{max})^2} \geq 0 \quad (20)$$

Then, it is easily observed that condition (13) implies the inclusion $\Omega_x \subseteq \mathcal{P}_u$. Now, by using the congruence lemma

$$\text{property with the full rank matrix } \begin{bmatrix} H_{hv}^T & 0 & 0 & 0 \\ 0 & P_{h+} & 0 & 0 \\ 0 & 0 & S_{hv}^T & 0 \\ 0 & 0 & 0 & I \end{bmatrix}, \quad (15)$$

yields:

$$\begin{bmatrix} -H_{hv}^T (X_h - I) H_{hv} & 0 & 0 & 0 \\ 0 & P_{h+} X_{h+} P_{h+} & 0 & 0 \\ W_{hv} & -S_{hv}^T B_h^T & -2S_{hv}^T & 0 \\ 0 & B_w^T & 0 & -\gamma I \end{bmatrix} + \begin{bmatrix} H_{hv}^T M \\ P_{h+} N \\ 0 \\ 0 \end{bmatrix} [A_h H_{hv} + B_h F_{hv} \quad -E_v P_{h+} \quad -B_h S_{hv} \quad B_w] + (*) < 0 \quad (21)$$

In order to derive an LMI optimization problem, a good choice is to consider $M = 0$ and $N = X_{h+}$ then (121) yields if:

$$\begin{bmatrix} -H_{hv}^T (X_h - I) H_{hv} & (*) & (*) & (*) \\ A_h H_{hv} + B_h F_{hv} & -E_v P_{h+} - P_{h+} E_v^T + P_{h+} & (*) & (*) \\ W_{hv} & -S_{hv}^T B_h^T & -2S_{hv}^T & (*) \\ 0 & B_w^T & 0 & -\gamma I \end{bmatrix} < 0 \quad (22)$$

Finally, applying property 1 and the relaxation lemma to (21) ends the proof.

Remark1. A more general result can be reached; by multiplying by:

$$\begin{bmatrix} H_{hv}^T & 0 & 0 & 0 \\ 0 & G_{hhh+}^T & 0 & 0 \\ 0 & 0 & S_{hv}^T & 0 \\ 0 & 0 & 0 & I \end{bmatrix}, \text{ on the left hand side and by its}$$

transpose, on the right hand side of (14), gives:

$$\begin{bmatrix} -H_{hv}^T(X_h - I)H_{hv} & 0 & 0 & 0 \\ 0 & G_{hhh+}^T X_{h+} G_{hhh+} & 0 & 0 \\ W_{hv} & -S_{hv}^T B_h^T & -2S_{hv}^T & 0 \\ 0 & B_w^T & 0 & -\gamma I \end{bmatrix} + \begin{bmatrix} H_{hv}^+ M \\ G_{hhh+} N \\ 0 \\ 0 \end{bmatrix} [A_h H_{hv} + B_h F_{hv} \quad -E_v G_{hhh+} \quad -B_h S_{hv} \quad B_w] + \begin{bmatrix} x_k \\ x_{k+} \\ \psi(u_k) \\ W_k \end{bmatrix}^T = \begin{bmatrix} -H_h^{-T}(P_h - I)H_h^{-1} & 0 & 0 & 0 \\ 0 & H_{h+}^{-T} P_{h+} H_{h+}^{-1} & 0 & 0 \\ S_{hv}^{-1} W_{hv} H_h^{-1} & -S_{hv}^T B_h^T & -2S_{hv}^{-1} & 0 \\ 0 & 0 & 0 & -\gamma I \end{bmatrix} \begin{bmatrix} x_k \\ x_{k+} \\ \psi(u_k) \\ W_k \end{bmatrix} < 0 \quad (23)$$

Note that a new matrix G_{hhh+}^T is introduced, thus adding extra degree of freedom to the inequality. Therefore, the following theorem can be stated:

Theorem 1. Given T-S descriptor system (1) whose validity domain is characterized by a matrix $Q_m \in \mathbb{R}^n, m \in \mathbb{N}_q$. If there exist positive definite matrices $P_j \in \mathbb{R}^{n \times n}$, positive diagonal matrices $S_{jk} \in \mathbb{R}^{m \times m}$, and matrices $H_{jk} \in \mathbb{R}^{n \times n}$, $F_{jk} \in \mathbb{R}^{m \times n}$, $W_{jk} \in \mathbb{R}^{m \times n}, i, j, k \in \mathbb{N}_r$, and positive scalar γ such that conditions 4 are satisfied with:

$$T_{ijl}^k = \begin{bmatrix} -H_{jk}^T - H_{jk} + P_j & (*) & (*) & (*) & (*) \\ A_i H_{jk} + B_i F_{jk} & -E_k G_{ijl} - G_{ijl}^T E_k^T & (*) & (*) & (*) \\ 0 & G_{hhh+}^T & -P_l & (*) & (*) \\ W_{jk} & -S_{jk}^T B_i^T & 0 & -2S_{jk}^T & (*) \\ 0 & B_j^T & 0 & 0 & -\gamma I \end{bmatrix} < 0 \quad (24)$$

$$\begin{bmatrix} P_j & (*) \\ Q_m P_j & 1 \end{bmatrix} \geq 0, m \in \mathbb{N}_q, j \in \mathbb{N}_r \quad (25)$$

$$\begin{bmatrix} -H_{jk}^T - H_{jk} + P_j & (*) \\ F_{jk(t)} - W_{jk(t)} & -(u_t^{max})^2 \end{bmatrix} \leq 0, t \in \mathbb{N}_m, j, k \in \mathbb{N}_r \quad (26)$$

Proof: The result of theorem 1 is derived from the proof of Lemma 2 by choosing $M = 0$ and $N = G_{hhh+}^T$. Thus, this proof is omitted here for brevity.

Case 2

A. Consider $H_{()} = H_h$ in (7). Then the variation of the Lyapunov function in case 2 is:

$$\Delta V(x_k) = x_{k+}^T H_{h+}^{-T} P_{h+} H_{h+}^{-1} x_{k+} - x_k^T H_h^{-T} P_h H_h^{-1} x_k < 0 \quad (27)$$

By considering the saturation lemma1, the inequality (27) becomes:

$$\Delta V(x_k) - 2\psi(u_k)^T S_{hv}^{-1} \psi(u_k) + \psi(u_k)^T S_{hv}^{-1} W_{hv} H_h^{-1} x_k + x_k^T x_k - \gamma W_k^T W_k < 0 \quad (28)$$

Developing inequality (27) one can obtain:

$$\begin{bmatrix} x_k \\ x_{k+} \\ \psi(u_k) \\ W_k \end{bmatrix}^T = \begin{bmatrix} -H_h^{-T}(P_h - I)H_h^{-1} & 0 & 0 & 0 \\ 0 & H_{h+}^{-T} P_{h+} H_{h+}^{-1} & 0 & 0 \\ S_{hv}^{-1} W_{hv} H_h^{-1} & -S_{hv}^T B_h^T & -2S_{hv}^{-1} & 0 \\ 0 & 0 & 0 & -\gamma I \end{bmatrix} \begin{bmatrix} x_k \\ x_{k+} \\ \psi(u_k) \\ W_k \end{bmatrix} < 0 \quad (29)$$

Via the Finsler's Lemma, expressions (11) and (29) result in:

$$\begin{bmatrix} -H_h^{-T}(P_h - I)H_h^{-1} & 0 & 0 & 0 \\ 0 & H_{h+}^{-T} P_{h+} H_{h+}^{-1} & 0 & 0 \\ S_{hv}^{-1} W_{hv} H_h^{-1} & -S_{hv}^T B_h^T & -2S_{hv}^{-1} & 0 \\ 0 & B_h^T & 0 & -\gamma I \end{bmatrix} + \begin{bmatrix} M \\ N \\ 0 \\ 0 \end{bmatrix} [A_h + B_h F_{hv} H_{()}^{-1} \quad -E_v \quad -B_h \quad B_w] + (*) < 0 \quad (30)$$

Using the property of congruence with: $\begin{bmatrix} -H_h^T & 0 & 0 & 0 \\ 0 & H_{h+}^T & 0 & 0 \\ 0 & 0 & S_{hv}^T & 0 \\ 0 & 0 & 0 & I \end{bmatrix}$; (30) yields:

$$\begin{bmatrix} -(P_h - I) & 0 & 0 & 0 \\ 0 & P_{h+} & 0 & 0 \\ S_{hv}^T W_{hv} H_h^{-1} & -S_{hv}^T B_h^T & -2S_{hv}^{-1} & 0 \\ 0 & B_h^T & 0 & -\gamma I \end{bmatrix} + \begin{bmatrix} H_h^+ M \\ H_{h+}^+ N \\ 0 \\ 0 \end{bmatrix} [A_h H_h + B_h F_{hv} \quad -E_v H_{h+} \quad -B_h S_{hv} \quad B_w] + (*) < 0 \quad (31)$$

Consequently, the following results can be stated.

Theorem 2. Given T-S descriptor system (1) whose validity domain is characterized by a matrix $Q_m \in \mathbb{R}^n, m \in \mathbb{N}_q$. If there exist positive definite matrices $P_j \in \mathbb{R}^{n \times n}$, positive diagonal matrices $S_{jk} \in \mathbb{R}^{m \times m}$, and matrices $H_j \in \mathbb{R}^{n \times n}$, $F_{jk} \in \mathbb{R}^{m \times n}$, $W_{jk} \in \mathbb{R}^{m \times n}, i, j, k \in \mathbb{N}_r$, and positive scalar γ such that conditions 4 are satisfied with:

$$T_{ijl}^k = \begin{bmatrix} -P_j & (*) & (*) & (*) \\ A_i H_j + B_i F_{jk} & -E_k H_l - H_l^T E_k^T + P_l & (*) & (*) \\ W_{jk} & -S_{jk}^T B_i^T & -2S_{jk}^T & (*) \\ 0 & B_h^T & 0 & -\gamma I \end{bmatrix} < 0 \quad (32)$$

$$\begin{bmatrix} -P_j & (*) \\ Q_m P_j & -1 \end{bmatrix} \leq 0, m \in \mathbb{N}_q, j \in \mathbb{N}_r \quad (33)$$

$$\begin{bmatrix} -H_{jk}^T - H_{jk} + P_j & (*) \\ F_{jk(t)} - W_{jk(t)} & -(u_t^{max})^2 \end{bmatrix} \leq 0, t \in \mathbb{N}_m, j, k \in N_r \quad (34)$$

Proof. Recall (30). The result of Theorem 2 is derived from the proof of Lemma 2 by choosing $M = 0$ and $N = H_{h+}^{-T} = P_{h+}$. Thus, this proof is omitted here for brevity.

Remark 2. The approaches presented in Theorem 1 and Theorems 2 are not equivalent, also note that matrix $H_{(j)}$ has a different structure in each case. This fact allows keeping the same number of LMI conditions.

III. ILLUSTRATIVE EXAMPLE

The proposed results are illustrated via the following two numerical examples, where the T-S system (1) is subject to amplitude-bounded disturbance $w(t) = 0.1 \sin(10 * t)$.

Example 1. Consider a T-S descriptor model (1), with:

$r = r_e = 2$ and $a = -1.2$; $b = -0.3$;

$$A_1 = \begin{bmatrix} -0.35 & 1.24 + b \\ -0.74 & 0.5 \end{bmatrix}, A_2 = \begin{bmatrix} -0.35 & 1.24 - b \\ -0.74 & 0.5 \end{bmatrix}$$

$$B_1 = \begin{bmatrix} 1.5 + a \\ 0 \end{bmatrix}, B_2 = \begin{bmatrix} 1.5 - a \\ 0 \end{bmatrix} \text{ and } B_w = \begin{bmatrix} 0.1 \\ 0.05 \end{bmatrix}.$$

$$E_2 = \begin{bmatrix} 0.37 & 0.43 + a \\ 0.23 & 1.15 \end{bmatrix}, E_2 = \begin{bmatrix} 0.37 & 0.43 - a \\ 0.23 & 1.15 \end{bmatrix}$$

The membership functions are defined as follows:

$$v_{1k} = \frac{x_{2k} + 2}{4}, v_{2k} = 1 - v_{1k}, h_{1k} = \frac{x_{1k}}{4} \text{ and}$$

$h_{2k} = 1 - h_{1k}$. The MFs satisfy the convex-sum property on the compact set $\Delta = \{x_k : |x_{1k}| \leq 2, |x_{2k}| \leq 2\}$, and $u^{max} = 0.05$. For this model, only the conditions of Theorem 1 are feasible, i.e., conditions in Theorem 2 are unfeasible. The following values were obtained:

$$P_1 = \begin{bmatrix} -8.2669 & -7.7476 \\ -7.7476 & -8.1054 \end{bmatrix}, P_2 = \begin{bmatrix} 7.7476 & 8.1054 \\ 8.1054 & 7.6609 \end{bmatrix}$$

$$H_{11} = H_{12} = \begin{bmatrix} 9.7378 & 6.4611 \\ 6.5695 & 10.1246 \end{bmatrix}$$

$$H_{21} = H_{22} = \begin{bmatrix} 9.7378 & 6.4611 \\ 6.4695 & 10.1246 \end{bmatrix},$$

$$F_{11} = F_{12} = \begin{bmatrix} -16.0422 & -23.6936 \\ -9.1702 & -13.1010 \end{bmatrix}$$

Simulation results were obtained with initial condition $x(0) = [0.25 \ -0.25]^T$ are presented in Figure 1 and 2.

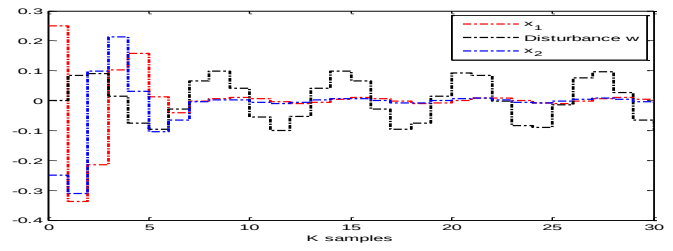


Fig. 1. System responses in the presence of actuator saturation and disturbance.

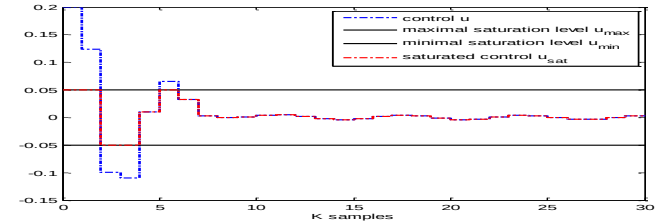


Fig. 2. System control input.

Example 2. Consider a T-S descriptor model as in (1) with $r = r_e = 2$

$$A_1 = \begin{bmatrix} 1 & -2 \\ -1 & -0.5 \end{bmatrix}, A_2 = \begin{bmatrix} 1 & 2 \\ -1 & -0.5 \end{bmatrix}, B_1 = \begin{bmatrix} 7 \\ 2 \end{bmatrix}$$

$$B_2 = \begin{bmatrix} 7 \\ 4 \end{bmatrix}, E_1 = \begin{bmatrix} 0.9 & 0.5 \\ 0.3 & 1.1 \end{bmatrix}, E_2 = \begin{bmatrix} 1.3 & 0 \\ 0 & 0.9 \end{bmatrix}, \text{ and } B_w = \begin{bmatrix} 0.1 \\ 0.05 \end{bmatrix}$$

The MFs are defined the same as in Example 2 and satisfy the convex-sum property on the compact set $\Delta = \{x_k : |x_{1k}| \leq 1, |x_{2k}| \leq 1\}$, and $u^{max} = 0.05$. Conditions in Theorem 1 are unfeasible, while Theorem 2 gives the following gain matrices:

$$P_1 = \begin{bmatrix} -3.5029 & 1.0913 \\ 1.0913 & -0.7546 \end{bmatrix}, P_2 = \begin{bmatrix} -1.0913 & 0.7546 \\ 0.7546 & -0.6520 \end{bmatrix}$$

$$H_1 = \begin{bmatrix} -25.4946 & 28.8282 \\ -246.8361 & 308.4762 \end{bmatrix},$$

$$H_2 = \begin{bmatrix} -106.1083 & 27.6242 \\ 101.8290 & -25.6545 \end{bmatrix}$$

$$F_{11} = F_{12} = \begin{bmatrix} -67.0342 & 84.1683 \\ -13.8934 & 3.4809 \end{bmatrix},$$

Simulation results with initial conditions $x(0) = [0.3 \ -0.3]^T$ are presented in Figure 3 and 4.

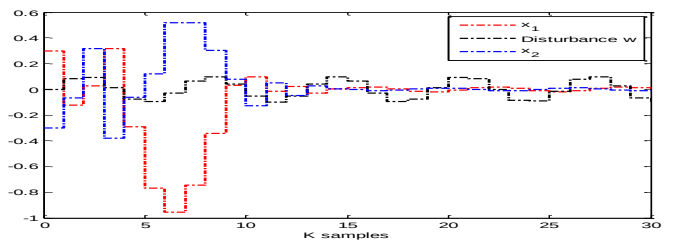


Fig. 3. System responses in the presence of actuator saturation

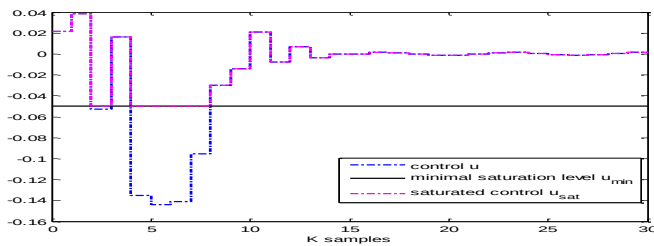


Fig. 4. System control input.

Figure 4 shows the different control inputs applied to the proposed system where $u(t)$ represent the derived non-PDC controller input. The closed-loop disturbed discrete-time T-S responses derived with the controller gain matrices given are depicted in Figure 3. It can be deduced that the designed controller allows achieving a stable closed-loop performance.

IV. CONCLUSION

This paper proposes a novel LMI-based framework for controller synthesis of discrete-time disturbed descriptor T-S systems with state and input constraints. Based on non-quadratic Lyapunov functions and non-PDC controllers, new stability conditions have been designed. The controller gains are then obtained by solving an optimization problem under LMI constraints. Through two numerical examples, the efficiency of the approaches has been demonstrated.

REFERENCES

- [1] K.Tanaka, T.Ikeda, and H.O.Wang, "Fuzzy regulators and fuzzy observers : relaxed stability conditions and LMI-based designs," *IEEE Transactions on Fuzzy Systems*, vol.6,pp:250-265,1998.
- [2] K.Tanaka and H.O.Wang *Fuzzy Control system, design and Analysis: a Linear Matrix Inequality Approach*. Wiley-Interscience, 2001.
- [3] M. Johansson, A. Rantzer, and K.E. Arzen, "Piecewise quadratic stability of fuzzy systems," *IEEE Transactions on Fuzzy Systems*, vol. 7, no.6, pp. 713-722, 1999.
- [4] T.M.Guerra and L.Vermeiren., « LMI based relax non-quadratic stabilization for non-linear systems in the Takagi-Sugeno's form, », *Automatica*, 40(5):823-829, May 2004.
- [5] T.M.Guerra,M.Bernal, K.Guelton, and S.Labiod, "Non-quadratic local stabilization for continuous-time Takagi-Sugeno models," *Fuzzy Sets and Systems*,2012.
- [6] T.Taniguchi, K.Tanaka and H.O.Wang, " Fuzzy Descriptor system and nonlinear model following Control," *IEEE transactions on Fuzzy Systems*, vol.8,no.4,pp,442-452,2000.
- [7] V.E.Manzo,Z..Lendek,T.Guerra, and P.Pudlo, "Control design for discret-time descriptor models: a systematic LMI approach," *IEEE Trans.On Fuzzy Systems*, Early Access Online 2015.
- [8] A-T.Nguyen,M.Damrène,J.Lauber,"Simultaneous design parallel distributed output feedback and anti-windup compensators for constrained Takagi-Sugeno fuzzy systems," *Asian.J. Control*, Vol.18,pp-1641-1654,2006.
- [9] A.Benzaouia,A.El hajjaji,A.and R.Oudah."Fault tolerant saturated control for T-S fuzzy discrete-system with delay," *Nonlinear analysis:Hybrid systems*, 178,pp:60-71,2015.
- [10] D.Saifia,M. Chadli,S.Labiod, and T.M.Guerra, "Robust H_∞ static output feedback stabilization of T-S fuzzy system subject t actuator saturation,"*Inter.J.Conf,Automation and System*, vol.10(3), pp.613-622,2012.
- [11] S. Bezzaoucha, S., Marks, B., Maquin, D., & Ragot, J. "Stabilization and output feedback control for Takagi-Sugeno with saturated actuators". *Inter.J.Adaptive control and signal Processing*. Vol.30, pp.888-905.
- [12] T.M. Guerra, V.Estrada-Manzo, and Z.Lendek, "Observer design for Takagi-Sugeno descriptor models: An LMI approach" *Automatica*, vol.52,pp.154-159,2015.
- [13] K.Guelton, S.Delprat,and T.M.Guerra, "An Alternative to inverse dynamics joint torques estimation in human stance based on a Takagi-Sugeno unknow-inputs observer in the descriptor form," *Control Engineering Practice*, vol.16(12),pp.1414-1426,2008.
- [14] S. Aouaouda, M. Chadli, M. Boukhnefer, H.R. Karimi, "Robuste fault tracking controller design for vehicle dynamics: A descriptor approach", *Mechatronics*, vol. 30, pp. 316-326, 2015.
- [15] M.Chadli and M.Darouach, "Novel bounded real lemma for discrete-time descriptor systems: Application to control-design," *Automatica*, vol.48, no.2, pp.449-453, Feb.2012.
- [16] Q.V. Dang, L.Vermeiren, A. Dequidt, M. Dambrine, "Robust stabilizing controller design for Takagi-Sugeno fuzzy descriptor systems under state constraints and actuator saturation," *Fuzzy Sets and Systems*, vol. 329, pp. 77-90, 2017.

Fire Detection System Using Deep Learning

Dermouche Reda

LREEI Electrification of Industrial
Enterprises Laboratory, Faculty of
Hydrocarbon & Chemistry, University
M'Hamed Bougara, Boumerdes , Algeria
LSA laboratory of Applied Sciences *National
Higher School of Advanced Technology*
Algiers, Algeria
reda.dermouche@ensta.edu.dz

Bahi Wissal

dept.Electrical Engineering and Industrial
Computing *National Higher School of
Advanced Technology* Algiers, Algeria
w.bahi@ensta.edu.dz

Tadger sidahmed

LREEI Electrification of Industrial
Enterprises Laboratory, Faculty of
Hydrocarbon & Chemistry, University
M'Hamed Bougara, Boumerdes , Algeria
s.tadger@univ-boumerdes.dz

Azouz Mohame Farouk

dept.Electrical Engineering and Industrial
Computing *National Higher School of
Advanced Technology* Algiers, Algeria
m.azouz@ensta.edu.dz

Abstract—The fire detection system based on deep learning is an innovative, AI engineering advancements where the YOLO algorithm is used in detecting fire in images from the cameras. Some features of deep learning systems include adaptability, compatibility with overall security system, as well as the ability to improve itself autonomously. However, we have to note some crucial points concerning the practical application of such algorithms, going from having enough data samples to perform those computations to the possibility of performing real-time calculations. It is believed that in order to address the challenges highlighted above in this article, more research will be encouraged to be undertaken to enhance all these systems with a view of offering protection of lives and property as required.

Index Terms—Deep-Learning, YOLO, Artificial Intelligence, Training, Accuracy, Mean average precision.

I. INTRODUCTION

The fire detection system who uses deep learning is a system based on artificial intelligent that deploys convolutional neural networks to process visual data captured by a camera as in our project. These systems are usually fed with data and programmed to detect patterns that can be judged as fires or smoke thus offering more reliable results compared to conventional techniques. Some of the strong attributes of deep learning-based system include flexibility in deployment, compatibility with other surveillance systems, and ability to learn and enhance on its own over time. Some issues involve, the deficit of use of strong training data, and the question of the feasibility of real-time streaming. However, there is optimism that ongoing research will lead to further improvements in these systems and therefore are very important tools for protecting

lives and property in cases that involve fires. Finally the aim of our project is to develop a performed model based on one of deep learning techniques in order to detect fire and protect humans and their properties.

II. THE PREVIOUS STUDIES

In this part, we will talk about the previous articles that are related to our project.

The authors in [1] are improving that fire detection using deep learning and video image processing techniques. They both employ convolutional neural networks (CNNs) and the YOLO (You Only Look Once) algorithm to achieve high accuracy and real-time performance in detecting flames. These systems address the shortcomings of traditional methods, such as high false alarm rates and poor effectiveness in complex environments. Additionally, the effectiveness of these systems is validated through experiments, proving their capability in real-world scenarios, and this article shares a common approach and goal in enhancing fire detection for fire-fighting applications.

In [2] we have the goal of improving the capacity of fire detection through the incorporation of sophisticated tools. a system for fire-fighting robots and its use to identify and locate a flame, the usage of deep learning on the detection of fire and smoke and its efficiency in video and security. It elaborates on several techniques in deep learning and the benefits in achieving the detecting results in a shorter and then the article introduces deep learning methods such as the Convolutional Neural Networks (CNNs) and how they assist the accuracy of fire or smoke detection, in addition to, reduction of false alarms and general improvement of security measures and response times.

In [3] The authors focus on the ability of giving services and enhancing deep learning to improve fire technology by applying transfer learning to InceptionV3 and MobileNetV2 for the detection of fire in surveillance videos. It also adds precision and decreases false positives compared to standard methodologies that might indicate alarm even when there were none.

The article [4]analysis, focused on the development of a real-time fire detection system using CNNs. The primary goal of the system is to minimize false alarms and increase the efficiency of alert timing and generation utilizing color and texture attributes of the image data obtained through webcams. Transfer learning is used to improve the accuracy and runtime of the model.

In [5] it demonstrates how other AI techniques a specifically deep learning models such as YOLOv3 can be applied to improve detection of fire in robotic systems. It highlights the drawbacks of the conventional approaches based on sensors alone, and show the possibility of increasing the accuracy of gaining detection data with the help of deep learning solutions. In addition, despite the usefulness of transfer learning, preprocessing and feature extraction are one of the most important areas that positively affect the effectiveness of the neural network's work and can increase the speed of development. The study provides a real-world example of how AI can work with fire-fighting robots and enhances their deep learning techniques to provides better detection and efficient response capabilities for the detection and combating of fires. [6] In this document, the author proposes a very elaborate fire detection model that can apply the deep learning feature as well as the multidimensional texture analysis in determining the fire in images. In order to find the candidate fire regions, the system leverages Faster R-CNN, based on convolutional neural networks. These regions are then described by linear dynamical systems in order to capture the spatial texture of the regions which is then followed by projecting the image data. Using the databases such as Corsican Fire Database (CFDB), the system was proved to have a high degree of accuracy in the

detection of insects while having low false positives, which makes the method more efficient compared to other techniques. This solution effectively responds to the challenges that relate to fire detection, which include differences in flame shapes and sizes, and conditions such as faint smoke that can be difficult to detect, this enhances the reliability of detection while minimizing on the number of false alarms.

The study in [7] indicates fire detection system employing deep learning in its functioning. The primary objective is the location of fires through the use of the convolutional neural networks (CNNs). The article outlines a system which uses CNNs and other strategies for boosting performance of the system such as data augmentation and optimization strategies for improving the chances of getting it right and gets score and accuracy of 93%.

In [8] deep learning can be employed in fire detection systems where the authors use Faster R-CNN for positive detection of fire and smoke in different settings. This makes it easier to detect the objects of interest and also minimize false alarms. The proposed system makes the use of CNNs and transfer learning to detect expressly various classes of fires as well as smoke, and that's the reason of realizing high accuracy even in cases of different environments like the forests, the roads, and the indoor environments. The author concentrates on the use of deep learning in improving the reliability and efficiency of fire detectors specifically in line with the general shift towards the implementation of sophisticated models of deep learning for better fire and smoke sensing.

III. APPROACHES OF DEEP LEARNING FOR FIRE DETECTION

A. Overview of the approaches

There are many approaches in deep learning for fire detection and in this article, we are going to introduce the popular one:

- **Faster R-CNN**: is a state-of-the-art object detection architecture of the R-CNN family, introduced by Shaoqing Ren, Kaiming He, Ross B. and Jian Sun in 2015. The primary goal of the Faster R-CNN network is to develop a unified architecture that not only detects objects within an image but also locates

Table I
COMPARATIVE TABLE BETWEEN DIFFERENT APPROACHES OF DEEP LEARNING

Aspect	Faster R-CNN	R-CNN	YOLO	SSD
Architecture	Two-stage: Region Proposal Network (RPN) + Classifier	Multi-stage: Region Proposal + Feature Extraction + Classifier	Single-stage: Predict bounding boxes and class probabilities directly	Single-stage
Speed	Slower than YOLO and SSD	Slower than YOLO and SSD	Very fast	Very fast
Accuracy	High	High	High	High
Real-time performance	Not real-time	Not real-time	Real-time	Real-time
Training complexity	Moderate	High	Moderate	Moderate
Flexibility	Moderate	Low	High	High
Number of parameters	Higher	Higher	Moderate	Moderate
Robustness	High	Moderate	High	High
Popular Application	Various industrial applications, security systems	Object detection in images	Real-time applications, surveillance systems	surveillance systems
Adaptability	Fire detection with appropriate training data	Fire detection with appropriate training data	Fire detection tasks	Fire detection tasks

the objects precisely in the image. It combines the benefits of deep learning, convolutional neural networks (CNNs), and region proposal networks (RPNs) into a cohesive network, which significantly improves the speed and accuracy of the model. [9]

- R-CNN: It's a type of deep learning architecture used for object detection in computer vision tasks. RCNN was one of the pioneering models that helped to advance the object detection field by combining the power of convolutional neural networks and region-based approaches. [10]

- SSD: It is a real-time object detection algorithm that can detect multiple objects in an image with high accuracy. The SSD algorithm is based on a deep convolutional neural network that takes an image as input and produces a set of bounding boxes and class predictions for all objects in the image. [11]

- YOLO: you only look once is a popular and fast object detection algorithm that uses a single neural network to predict bounding boxes and class probabilities for objects in an image. [12]

B. Comparison between the approaches

In this part we'll draw a table to compare between these approaches in term of architecture, Speed, Accuracy and other aspect.

Deep learning approaches for fire detection system offer advantages by including neural networks. The system can achieve high performance (accuracy and speed),learn complex patterns and identify fire and smoke in different environment. Below in TABLE I the comparision between different ap- proaches of deep learning and we make our choise to choose Yolo approache for the rest of this article.

IV. PREPARING TO THE MODEL

In this section we'll prepare our model for the training and before this step we need to plan for a dataset then generate a model for the fire detection and finally evaluate its performance.

A. Dataset

Our source to develop this dataset include public dataset as Roboflow universe it's a platform to access to dataset related to fire detection. This dataset named Fire-Smoke-Detection is developed for our project in order to train the model to detect fires in various environments and under different conditions, the dataset is composed of indoor and outdoor fires shown in Fig 1.

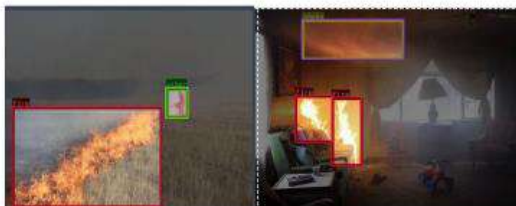


Fig 1. Indoor and outdoor fire

C. One Stage Detector

[13] One popular example of a one-stage detector algorithm is the YOLO (You Only Look Once) . This

algorithm enables to process the entire image in a single pass makes it efficient and suitable for real-time applications. One-stage detector algorithms are a type of object detection algorithm used in computer vision tasks. Unlike two-stage detectors that involve a separate region proposal network (RPN) and object detection stage, one-stage detectors perform both object detection and classification in a single pass. Fig. 2 shows the structure of one-stage detector. In this deep learning algorithm, the input to the algorithm is an image that contains objects of interest, such as fire. The algorithm processes the entire image at once and produces a set of bounding boxes along with corresponding class predictions for each detected object.

The backbone network is a crucial component of the one stage detector algorithm. It is typically a deep convolutional neural network (CNN) that extracts hierarchical features from the input image. The backbone network can be a popular CNN architecture like in our case Darknet, which has been pre-trained on a large-scale image classification task. The purpose of the backbone network is to capture high-level representations and semantic information from the image.

The neck network typically consists of additional convolu- tional layers that fuse and combine the multi-scale features from the backbone network. This fusion helps to enhance the representation power of the features and enables the algorithm to detect objects at different scales and resolutions.

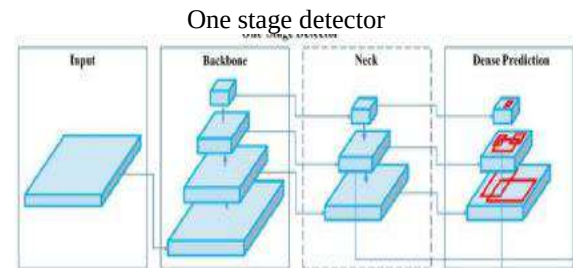


Fig 2. Structure of one stage detector (YOLO in our case)

The head or dense prediction part of the one-stage detector algorithm performs the final stage of object detection. It consists of convolutional layers followed by fully connected layers. The head network takes the fused features from the neck network as input and performs spatial and channel-wise convolutions to generate the final detection outputs. These outputs include the bounding box coordinates (x, y, width, height) and class probabilities for each detected object. The head network utilizes anchor boxes or default boxes to anchor

the predicted bounding boxes to predefined scales and aspect ratios.

C. Model generation

For generating a YOLOv8 model for fire detection the dataset will be divided into training, validation, and testing sets with the typical split of 70%, 20%, and 10%, respectively. The training set, comprising

70% of the dataset, is used to train the YOLOv8 model. The validation set, consisting of 20% of the dataset, is utilized for monitoring the model's performance during training, the testing set which makes up 10% of the dataset, is used to evaluate the final model's performance.

• Training

the YOLOv8 model is trained in order to detect any source of fire on the training set using google collab as a platform. During training, the YOLOv8 model learns to detect fire in images, as a result of the training process loss curves are commonly monitored: train/box loss, train/obj loss, and train/cls loss. Fig.3 shows training loss curves.

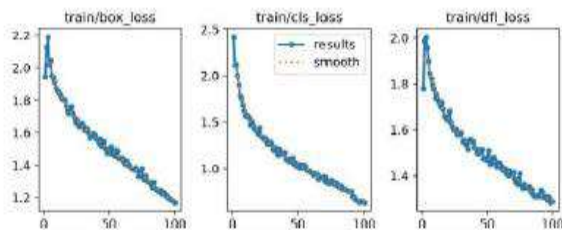


Fig 3. Training loss curves

train/box loss: This curve shows the Bounding Box Regression Loss during training session. This rises at the initial steps and then goes down gradually show that the model's prediction of the position of bounding boxes is gradually becoming better after some iterations.

train/cls loss: This curve describes the percentage of misclassified instances during the training process, in terms of classification loss. It is also relatively higher initially and decreases gradually, which indicates the model correctly classifies objects better as the algorithm progresses.

train/dfl loss: This curve expresses the trend of the focal loss during training period. Analogously, we witness a decrease in the dispersion and an improvement of the predictive accuracy, demonstrating that the model is closer to the real distribution of the corresponding predictor variable.

• Validation

the YOLOv8 model is evaluated to assess its performance. During the validation process of a YOLOv8 model for license plate recognition, the loss curves serve as valuable metrics to assess the model's performance. Fig.4 shows the validation loss curves.

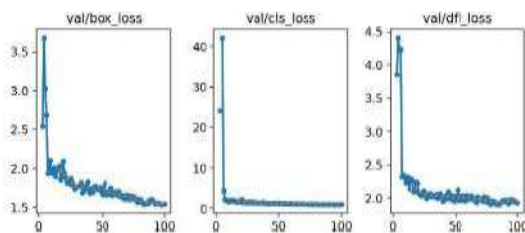


Fig 4. Validation loss curves

val/box loss: This is the validation bounding box regression loss so it is referring to how accurate the model is in predicting the bounding box of the object. The trend is consistent with the training box loss, where the best performing models are not the highest but are lower, demonstrating that the model's gains can be applied to data that have not been encountered before.

val/cls loss: This is the validation classification loss, as seen in equation It shows a downward shift at the start of the training and gradually levels out, meaning that the model soon learns the start of the objects in order to classify them and retains this rate of classification.

val/dfl loss: This is the validation distribution focal loss. Same with classification loss, it decreases dramatically at some point and remains at that level, indicating its ability to well generalize the prediction confidence and orientation to the validation data.

• General Interpretation

The loss curves of all the specified loss types have a decreasing nature, this meaning that the model is routinely learning well without instances of overfitting, as there is a decreasing training loss as well as the validation loss. These curves conclude that the model is learning with time, is accurate, and performs well when predicting validation data.

D. Model result

This part presents the model results and performance evaluation. These performance or in other words metrics are precision, recall and F1-score and there are computed to assess the model's effectiveness. This process ensures the development of a robust and accurate YOLOv8 model the recognition of fire.

1) **Confusion matrix:** Confusion matrix is a popular tool for evaluating the performance of deep learning-based models it is utilized in this experiment. It serves as a visual representation and quantitative assessment YOLOv8 model's performance, allowing to evaluate, analyze, and improve the model's accuracy and effectiveness. The confusion matrix below is provided to evaluate performance of the model and here is the signification of the results obtained from fig 5:

True Positive (TP): 471 situated in the top left cell of the matrix and it means that the model correctly predicts the positive class (fire).

False positive (FP): 150 situated in the top right cell of the matrix and it means that the model incorrectly identifies a fire when there is none.

False negative (FN): 77 situated in bottom left cell of the matrix and it means that the model incorrectly predicts the negative class.

True negative (TN): The model correctly predicts the negative class.



Fig 5. Confusion matrix

And now, after generating the confusion matrix, its objective is to calculate the performance of the model. The performance metric include precision, recall and F1 score.

Precision

$$P = \frac{TP}{TP + FP} = \frac{471}{471 + 150} = 0.759 \quad (1)$$

About 75.9 % of the instances predicted as positive are actually positive.

Recall

$$R = \frac{TP}{TP + FN} = \frac{471}{471 + 77} = 0.860 \quad (2)$$

About 86.0% of the actual positive instances are correctly identified by the model.

F1 score

$$F1 = 2 \times \frac{P \cdot R}{P + R} = 2 \times \frac{0.759 \cdot 0.860}{0.759 + 0.860} = 0.806 \quad (3)$$

The balance between precision and recall is approximately 0.806, indicating that the model performs well.

Interpretation: the results shows that our model is performing well and this is indicated by high recall, good precision and balanced F1 score.

2) Performance Measurement: [13]: By analyzing the confusion matrix, performance measurements can be performed. Popular metrics such as precision, recall, F1-score.

• Precision-confidence curve

The precision-confidence curve is a visual representation of the relationship between precision and confidence recognition YOLOv8 model. The precision-confidence curve plots the precision value at different confidence thresholds. Starting from a high confidence threshold, the precision will be relatively high because only confident predictions are considered valid. Fig.6 demonstrates the precision-confidence curve.

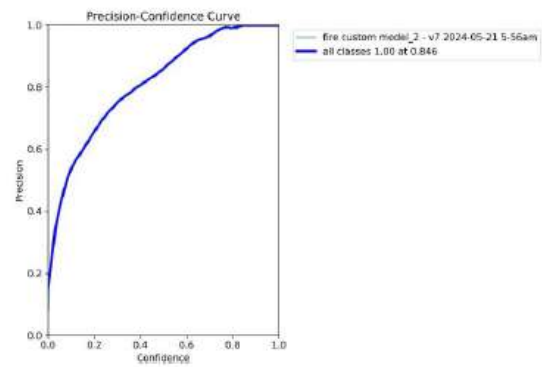


Fig 6. Precision-confidence curve

In the X-axis, we have the confidence level, which is the model's estimated likelihood of positively classifying a prediction.

The Y-axis depicts the measure of accuracy, precision or the true positives with reference to the overall positives tagged as such in the model including the false positives. At the initial stage, when the level of confidence is low, accuracy is less and therefore there will be more false positives. It can also be inferred that as the confidence threshold rises, the percentage of true positives escalates, proving that the model's elemental predictions tend to be more precise. It starts off from the bottom left and as the scales of confidence go up, the curve moves steep upward which shows that there is a great improvement in the level of precision. As the scale moves towards the upper end of the scale, the level of confidence gets closer to 1 and the precision becomes more refined.

This means that even at high confidence levels, the model has almost perfect positive predictive value due to the mean value of 0. According to the legend it points out the precision to the nearest 0.

Interpretation: This analysis aids in the necessary fine-tuning of our confidence level in order to make the final model accurate: the model works adequately at high levels of confidence and is thus suitable for high-stakes applications like fire detection in our case where false positives are a significant concern. In general, the curve indicates that the performance of the model is satisfactory and it can be essentially effective especially where high levels of precision are needed, although it pointed several sectors whereby the model could be possibly improved.

• Recall-confidence curve

The recall-confidence curve illustrates how the recall metric changes as the confidence threshold varies. As the confidence threshold decreases, more predictions are considered valid, including both true positives and potential false positives. Fig.7 shows the recall-confidence curve.

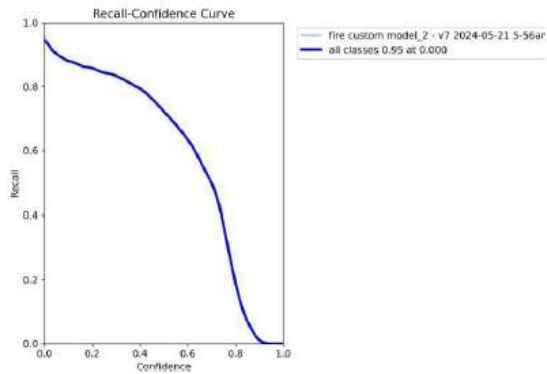


Fig 7. Recall-confidence curve

The X-axis indicates the confidence threshold, which denotes the model's level of certainty when it makes positive predictions, the Y-axis is defined as recall, which is the number of correct positive predictions compared to the actual number of positives.

The curve moves down from the top left corner, implying that it measures high recall at low bar confidence. This means that the calibration curve reveals that the model is capable of locating most actual positives when the criterion is relaxed. That is, as the confidence threshold increases, then the recall decreases; implying that actual positives are missed with advanced selectivity of the model.

As we move towards 1 on the right end, recall approaches zero; this means that when confidence is very high, the positive prediction made by the model is very selective, meaning that the number of true positive that are being identified or flagged is very few. According to the legend, the maximum attainable recall of the model is 0.95 with the lowest level of confidence.

Interpretation: The recall-confidence graph shows the measure of the trade-off between confidence and recall rates in the model predictions. Exploring this curve, we are able to identify how high the confidence needs to be for the full level of recall for the particular application. This model is characterized by its high-performing values at the low confidence level, which makes it suitable for our case, where high recall is an absolute necessity.

• Precision-recall curve

The precision-recall curve is another visual representation that provides insights into the performance and efficiency of a license plate recognition YOLOv8 model. The curve showcases the relationship between precision and recall, two important metrics used to evaluate the model's predictive accuracy. Fig.8 shows the precision-recall curve.

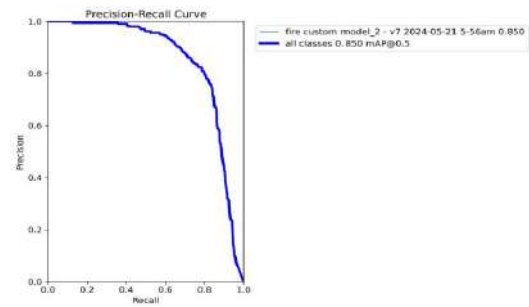


Fig 8. Precision-recall curve

On the y-axis, precision shows the proportion of true positive predictions in all positive predictions, while on the x-axis, recall shows the proportion of correct positive predictions to all actual positive ones.

The curve climbs steeply to 1, meaning that the recall rate is high, what implies that the model yields high accuracy on a comparatively lower number of positive results. When recall rises even higher, the precision becomes less while it shows a relationship where the model is involving even more false positives for more true positives. The precision is average 85% if the model designed to work at the recall of 0.5 thus a good performance in the macro environment across the various thresholds. This implies that depending on a certain level of precision, in combination with a high recall value, the model is effective at making identification of true positives as accurate as possible while at the same time reducing as far as possible the wrongly identified negatives as true positives.

Interpretation: It is clear from the precision recall curve relating to fire model shows high mAP, meaning that this model is able to maintain a high level of precision/recall in the event of fire as such, it can be recommended for use in cases where accurate identification of fire events is pertinent.

• F1-score curve

The F1-score curve demonstrates how the F1-score change as the threshold varies. The F1-score is important for measuring the efficiency of a license plate recognition YOLOv8 model because it combines precision and recall into a single value. It provides a comprehensive assessment of the model's ability to accurately identify positive instances while avoiding false positives. The F1-Confidence Curve for the fire model shows the effectiveness of the model and the area under the curve demonstrates the single value of the model as shown in Fig.9

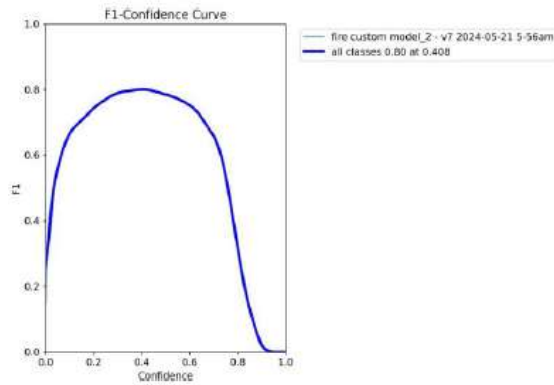


Fig 9. F1-score curve

the F1 score is the harmonic mean of precision and recall where the score shown on the y-axis while the confidence level is equated to the probability of positive prediction on the x-axis.

The curve starts from the origin and goes upward to a certain point and then travels down again on the same curve showing the best estimation of the F1 score in confidence intervals. Proportion found was the highest F1 measure recorded at 0.80 at the level of significance equal to 0.408 chosen as the optimal value for the metric because the closer to 1. This means that the proposed model has good accuracy in general and has good recall and precision values irrespective of the classes selected.

Interpretation: We see that the F1-Confidence curve of represents a rather effective model with the maximum reachable $F1 = 0.80$ assuming a confidence level of 0.408, which suggests that the model is well-suited for the task and maintains accurate results while accounting for all of the intended recommendations. This insight plays a significant role in the influencing of the model for a wide array of real-life uses in the fire detection

V. FUTURE SCOPE

The way deep learning will change the fire detection systems in future is promising an advancement that will increase the level of accuracy, speed, and effectiveness. This involves deep learning algorithms with IoT devices, and cloud infrastructure, this could be used to create real-time monitoring and early warning systems, hence improving the number of false alarms and response times. Furthermore, large dataset analysis and visualization through deep learning models provides important information about fire patterns and trends for efficient, cost worthy fire prevention strategies and emergency responses. As the technology evolves, it is expected to be used in industrial, residential and wildland environments and to be linked up with other technologies providing a total fire protection package.

VI. CONCLUSION

Fire detection systems based on deep learning technologies can be considered as an important step forward from traditional approaches in fire

identification and management in terms of both accuracy and time-related metrics. Using such solutions with the help of the approaches of YOLOv8, it is possible to minimize false positive rates and achieve good results even in complex conditions. Comparisons with other algorithms indicate that YOLOv8 is one of the best suited for fire detection in real-time settings because it has higher precision and recall values. Sophisticated experiments support the effectiveness of the YOLOv8 model, with the object detection having nearly equal precision and recall and a low number of false positives. The integration of Deep learning with IoT and Cloud structure is foreseen to introduce detailed real-time monitoring and signaling type services to improve fire detection and mitigation plans for different settings. Indeed, as the technology advances it is projected that it will become part of the fire protection systems in industries and smart houses.

REFERENCES

- [1] Y. Hao, X. Liu, and G. Huang, "A Feasible System of Automatic Flame Detection and Tracking for Fire-fighting Robot," in *2021 IEEE International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, Nanjing, China, Dec. 2021, pp. 101-103.
- [2] H. Prasad, A. Singh, J. Thakur, C. Choudhary, and N. Vyas, "Artificial Intelligence Based Fire and Smoke Detection and Security Control System," in *2023 International Conference on Network, Multimedia, and Information Technology (NMITCON)*, Nanjing, China, Dec. 2023, pp. 1-6.
- [3] A. Bari, T. Saini, and A. Kumar, "Fire Detection Using Deep Transfer Learning on Surveillance Videos," in *Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV 2021)*, Nanjing, China, Feb. 2021, pp. 1061-1064.
- [4] S. Srividhya, S. Sindhu, K. Sornalakshmi, D. Hemavathi, and V. Rajaram, "Fire Detection System using Deep Learning Techniques," in Proceedings of the 4th International Conference on Inventive Research in Computing Applications (ICIRCA), Chennai, India, July 2022, pp. 721-722, doi: 10.1109/ICIRCA54612.2022.9985499.
- [5] S. Ramasubramanian, S. Arumugam Muthukumaraswamy, and A. Sasikala, "Fire Detection using Artificial Intelligence for Fire-Fighting Robots," in Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 180-182. doi: 10.1109/ICICCS48265.2020.9121134.
- [6] P. Barmoutis, K. Dimitropoulos, K. Kaza, and N. Grammalidis, "Fire Detection from Images Using Faster R-CNN and Multidimensional Texture Analysis," in Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 8301-8305.

- [7] S. Uppal, S. Raheja, and N. R. Das, "Fire detection alarm system using deep learning," in *2023 13th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, Noida, India, 2023, pp. 1-6. doi: 10.1109/Confluence56041.2023.10048842.
- [8] M. S. Allauddin and V. P. Peesapati, "Detecting fire and smoke using Faster RCNN," in 2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2022, pp. 1-7. doi: 10.1109/CSDE56538.2022.10089304.
- [9] GeeksforGeeks, "Faster R-CNN - Introduction to Region Proposal Networks," GeeksforGeeks. [Online]. Available: <https://www.geeksforgeeks.org/faster-r-cnn-ml/>. [Accessed: 16-Jun-2024].
- [10] B. Dwyer, "What is R-CNN?," Roboflow Blog, 2023. [Online]. Available: <https://blog.roboflow.com/what-is-r-cnn/>. [Accessed: 16-Jun-2024].
- [11] E. Kim, "Understanding Single Shot Detector (SSD)," Medium. [Online]. Available: <https://medium.com/@elvenkim/understanding-single-shot-detector-ssd-4a249a271c5c>. [Accessed: 16-Jun-2024].
- [12] Kili Technology, "What is Data Labeling?," Kili Technology. [Online]. Available: <https://kili-technology.com/data-labeling>. [Accessed: 16-Jun-2024].
- [13] L. Yu, S. Liu "A Single-Stage Deep Learning-based Approach for Real-Time License Plate Recognition in Smart Parking System," Hebei College of Industry and Technology, Hebei Shijiazhuang, 050091, China.

Privacy by Design: Hybrid Rule-Based and Zero-Shot Classification for Browser-Based IoT Threat Detection

Mohamed Chaouli

Faculty of Science and Technology
Ziane Achour University
Email: m.chaouli@univ-djelfa.dz

Mohamed Elbar

Faculty of Science and Technology
Ziane Achour University
Email: m.elbar@univ-djelfa.dz

Abstract—IoT devices generate sensitive traffic data that reveals user behavior patterns, device locations, and network topology. Traditional cloud-based traffic analysis exposes this data to third parties, creating privacy risks for smart home users. We present a privacy-preserving IoT traffic classification system that processes everything within the user's web browser using client-side AI models. Our hybrid approach combines rule-based pre-filtering for known attack signatures with zero-shot classification using DeBERTa-v3 for device identification and anomaly detection. The key benefit is that no traffic data ever leaves the user's device. We evaluated our system on 175 IoT traffic flows from the N-BaIoT dataset containing normal device behavior and attack patterns including Mirai and BASHLITE botnet variants. Our results demonstrate 90.9% accuracy, 100% attack recall, and 84% specificity, with the triage model processing traffic entirely client-side at 0.61 flows/second. While throughput is constrained compared to cloud solutions, the complete privacy guarantee combined with strong detection performance makes this approach compelling for privacy-conscious smart home environments.

Index Terms—IoT security, smart home, privacy-preserving machine learning, browser-based AI, traffic classification, anomaly detection, WebGPU, zero-shot classification

I. INTRODUCTION

The Internet of Things (IoT) has transformed modern homes into connected environments with cameras, thermostats, speakers, doorbells, and appliances communicating continuously over home networks. Projections estimate over 30 billion IoT devices deployed globally by 2025 [1]. While these devices provide convenience and automation, they also generate network traffic that reveals sensitive information about occupants' daily routines, presence patterns, and behavioral habits [2].

Security concerns in IoT environments are well-documented. Large-scale botnets like Mirai [3] have compromised millions of IoT devices for distributed denial-of-service attacks. Data exfiltration, unauthorized access, and command-and-control communications pose ongoing threats to smart home networks [4]. Detecting these threats requires analyzing network traffic patterns—but traditional approaches send this sensitive data to cloud platforms for processing.

This creates what we call the *IoT privacy paradox*: analyzing traffic for security requires exposing it to external services,

potentially revealing the very information users seek to protect. Traffic metadata alone can reveal device types, usage patterns, and occupancy schedules without inspecting packet contents [5].

Existing privacy-preserving techniques like federated learning and differential privacy reduce but don't eliminate data exposure. Federated learning still transmits model updates that can leak information through gradient attacks. Differential privacy introduces noise that degrades detection accuracy. What's needed is a "privacy-by-architecture" approach where sensitive data *physically cannot* leave the user's device.

In this paper, we present a browser-based IoT traffic classification system that achieves complete privacy through local processing. Using recent advances in browser-based machine learning—specifically WebGPU for GPU acceleration and Transformers.js for neural network inference—we demonstrate that meaningful IoT security analysis can run entirely within a web browser.

Our main contributions are:

- 1) **Privacy-by-architecture design:** We demonstrate IoT traffic classification with *zero data transmission* using entirely browser-based inference. Unlike cloud solutions or federated approaches, our architecture makes data leakage physically impossible.
- 2) **Hybrid detection pipeline:** We combine rule-based pre-filtering for known IoT attack signatures (Mirai/BASHLITE ports and DoS patterns) with zero-shot NLI classification (DeBERTa-v3), achieving 90.9% accuracy with 100% attack recall on 175 traffic flows.
- 3) **Multi-format IoT traffic parsing:** Our system automatically parses N-BaIoT, IoT-23, NetFlow, and generic CSV formats, enabling analysis across diverse IoT dataset formats without manual configuration.
- 4) **Empirical evaluation with ground truth:** We evaluate on 175 traffic flows from the N-BaIoT dataset with known attack labels, reporting standard metrics (accuracy, precision, recall, F1, specificity) with confusion matrix analysis.

II. RELATED WORK

A. IoT Traffic Classification

Machine learning approaches to IoT traffic analysis have evolved significantly. Early work by Miettinen et al. [6] demonstrated device fingerprinting through traffic patterns, achieving 81% accuracy across 27 device types using random forests. Sivanathan et al. [4] extended this with temporal features, reaching 99% classification accuracy on controlled datasets.

Deep learning architectures have improved detection capabilities. Meidan et al. [7] introduced the N-BaIoT dataset and demonstrated that autoencoders can detect IoT botnet traffic with high accuracy. Their work on Mirai and BASHLITE detection established important benchmarks for IoT security evaluation. Garcia et al. [8] released the IoT-23 dataset capturing real malware traffic, enabling evaluation on diverse attack types including command-and-control, DDoS, and scanning behaviors. The CICIDS2017 dataset [10] provides additional intrusion detection benchmarks with labeled network flows.

Kitsune [9] introduced unsupervised anomaly detection using ensemble autoencoders, achieving real-time detection without labeled training data. Kolias et al. [23] analyzed IoT-based DDoS attacks and the evolution of botnets like Mirai. Hamza et al. [24] proposed SDN-based monitoring approaches for IoT attack detection. However, all these approaches require either cloud infrastructure or dedicated hardware for inference.

B. Privacy-Preserving Machine Learning

Privacy concerns in IoT analytics have motivated several approaches. Federated learning [11] enables collaborative model training without centralizing raw data, but gradient updates can still leak information through membership inference attacks [12]. Recent work has applied federated learning to IoT anomaly detection [13], but participants must still trust the aggregation server.

Differential privacy [14] provides formal privacy guarantees through calibrated noise addition. However, the privacy-utility trade-off is particularly challenging for security applications where false negatives have serious consequences. Homomorphic encryption enables computation on encrypted data but imposes 10-1000x computational overhead, making real-time inference impractical [15].

Our approach sidesteps these trade-offs entirely: by processing all data locally in the browser, no information leaves the device regardless of model architecture.

C. Browser-Based Machine Learning

The emergence of WebGPU [16], [17] has transformed browser-based AI capabilities. Unlike WebGL, WebGPU provides direct access to GPU compute shaders, enabling efficient tensor operations for neural network inference. Modern browsers can achieve 60-80% of native GPU performance for ML workloads.

Transformers.js [18] ports the Hugging Face Transformers library to JavaScript, enabling encoder models (BERT

[22], DeBERTa [20]), decoder models (GPT-2), and encoder-decoder models to run in browsers. WebLLM [19] demonstrates that billion-parameter language models can run entirely in-browser through aggressive quantization.

Our work applies these capabilities specifically to IoT security, combining fast zero-shot classification with selective deep analysis for the unique requirements of traffic classification.

III. SYSTEM ARCHITECTURE

A. Design Principles

Our system follows three main principles:

Privacy by Architecture: Instead of relying on policy or encryption, we achieve privacy through system design. The system physically cannot send traffic data anywhere because all processing happens locally in the browser sandbox.

Graceful Degradation: The system adapts to device capabilities, using WebGPU when available and falling back to WebAssembly (WASM) on devices without GPU support.

Zero-Shot Flexibility: Using natural language inference for classification allows analyzing new device types and attack patterns without retraining, adapting to the evolving IoT landscape.

B. Two-Stage Analysis Pipeline

Our architecture uses a two-stage analysis pipeline optimized for client-side constraints:

Stage 1 - Triage Classification: Each traffic flow goes through a fast zero-shot classifier that performs dual classification: (1) device type identification and (2) anomaly detection. This stage uses DeBERTa-v3 Small [20] fine-tuned on MNLI for natural language inference, enabling classification without IoT-specific training data.

Stage 2 - Deep Analysis: Traffic flows flagged as anomalous or with low confidence get analyzed by a larger language model running via WebLLM. This stage provides detailed explanations, attack type classification, and confidence assessments.

Figure 1 shows the system interface with the model loading progress and traffic upload area.

C. Triage Model: Zero-Shot Classification

The triage stage uses zero-shot classification based on natural language inference. Given a traffic flow representation x and a set of possible labels $L = \{l_1, l_2, \dots, l_n\}$, the model calculates:

$$P(l_i|x) = \frac{\exp(f(x, l_i))}{\sum_{j=1}^n \exp(f(x, l_j))} \quad (1)$$

where $f(x, l_i)$ is the entailment score between the traffic description (premise) and the hypothesis "This is l_i ".

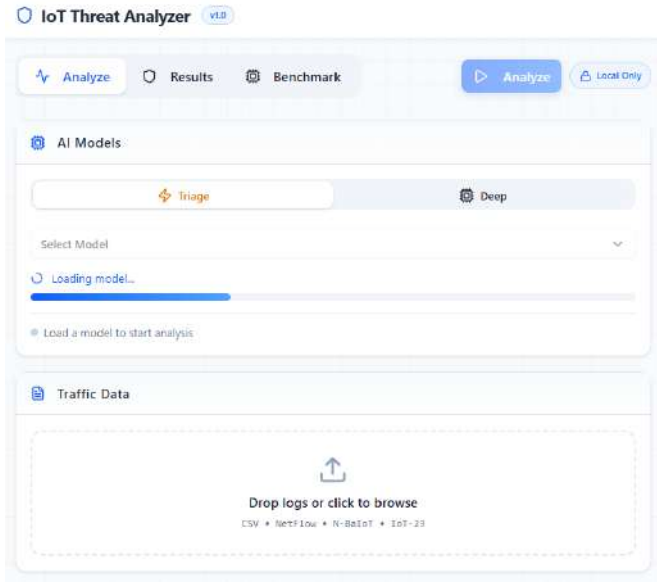


Fig. 1. System interface showing the IoT Threat Analyzer during model initialization. The left panel provides model selection (Triage/Deep) with loading progress indicator. The right panel accepts traffic data in multiple formats including CSV, NetFlow, N-BaIoT, and IoT-23.

1) *Device Type Labels*: Our device classification uses descriptive labels optimized for IoT context:

- “smart security camera or webcam network traffic”
- “smart thermostat or HVAC controller network traffic”
- “smart speaker or voice assistant network traffic”
- “smart TV or streaming device network traffic”
- “smart doorbell or video doorbell network traffic”
- “smart plug or power monitoring device network traffic”
- “smart light bulb or lighting controller network traffic”
- “smart home hub or gateway network traffic”
- “IoT sensor or environmental monitor network traffic”
- “smart appliance network traffic”
- “home router or network equipment traffic”
- “unknown or unidentified IoT device network traffic”

2) *Anomaly Detection Labels*: For anomaly classification, we use:

- “normal benign IoT device behavior”
- “botnet infected device communication pattern”
- “network scanning or reconnaissance activity”
- “data exfiltration or unauthorized data transfer”
- “denial of service attack traffic pattern”
- “unusual timing pattern for IoT device”
- “command and control server communication”
- “malicious network traffic pattern”

D. Deep Analysis Model

For escalated entries, we use SmolLM2-1.7B-Instruct running via WebLLM. The model receives a structured prompt containing traffic features and generates JSON-formatted analysis including device type, classification, confidence, anomaly type, indicators, and explanation.

E. Traffic Parsing

The system supports multiple IoT traffic formats through automatic detection:

N-BaIoT Format: Statistical features (115 dimensions) extracted from packet captures, commonly used in botnet detection research.

IoT-23 Format: Zeek connection logs with flow-level features including timestamps, IPs, ports, protocol, bytes, packets, and duration.

NetFlow Format: Standard network flow records with timing and volume information.

Generic CSV: Flexible parsing for custom formats with automatic column detection.

IV. IMPLEMENTATION

A. Technology Stack

Table I summarizes the key technologies.

TABLE I
TECHNOLOGY STACK

Component	Technology	Version
LLM Inference	@mlc-ai/web-llm	0.2.80
Zero-shot	@huggingface/transformers	3.8.0
UI Framework	React	19.2.0
GPU Acceleration	WebGPU	Native

B. Model Selection

We selected DeBERTa-v3 Small for triage based on its superior zero-shot accuracy among browser-compatible models. At approximately 180MB, it provides good classification performance while remaining practical for browser deployment. For deep analysis, SmolLM2-1.7B-Instruct (924MB) offers the best size-accuracy trade-off for generating detailed explanations within browser memory constraints.

C. User Interface

Figure 2 shows the system after completing traffic analysis with classification summary. The interface provides:

- Model status indicators showing load state
- Drag-and-drop file upload for traffic data
- Real-time classification progress
- Color-coded results (green for normal, yellow for anomalous, red for malicious)
- Detailed view with device type and anomaly explanations

V. EVALUATION

A. Experimental Setup

We tested our system on a Windows 11 workstation with Intel Core i7-12700H CPU, NVIDIA RTX 3060 GPU (6GB VRAM), 32GB RAM, and Chrome 120 with WebGPU enabled. All tests were conducted using automated end-to-end testing via Playwright.

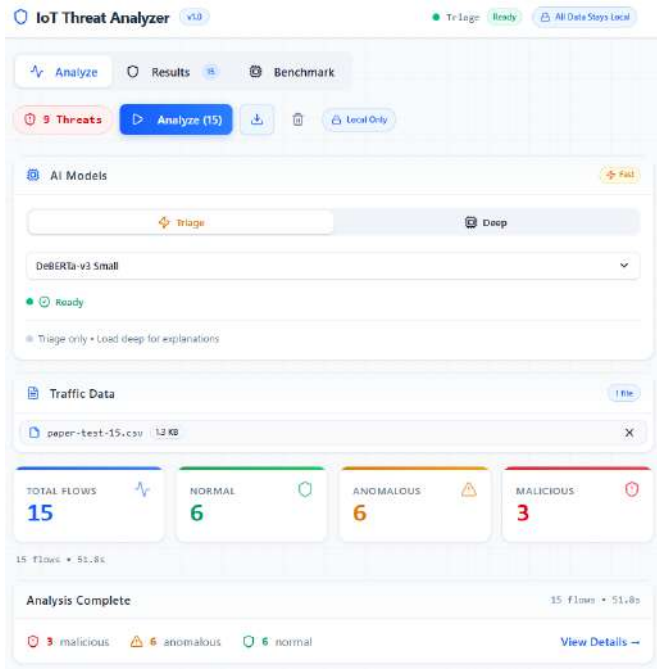


Fig. 2. Analysis complete view showing classification results for 15 IoT traffic flows processed in 51.8 seconds. The summary displays threat distribution: 3 malicious, 6 anomalous, and 6 normal flows, with color-coded progress bars indicating classification breakdown.

B. IoT Traffic Dataset

We used the N-BaIoT dataset [7], a widely-used benchmark for IoT botnet detection research. Our evaluation subset contains 175 traffic flows (100 normal, 75 attack) representing realistic device behaviors and attack patterns:

- **Normal traffic (100 flows):** Benign IoT device communications from cameras, doorbells, thermostats, and other smart home devices operating normally
- **Attack traffic (75 flows):** Mirai variants (scan, ACK flood, SYN flood, UDP flood, UDP-plain) and BASHLITE variants (scan, combo, junk, TCP, UDP)

The N-BaIoT dataset captures real malware traffic from Mirai and BASHLITE botnets infecting commercial IoT devices. Attack patterns include Telnet scanning (ports 23, 2323), C&C communications, and flood attacks with characteristic traffic asymmetry. Ground truth labels from the original dataset enable proper metric calculation.

C. Classification Results

Figure 3 shows the detailed analysis results with individual traffic flow classifications.

Table II summarizes the classification metrics compared to ground truth.

D. Analysis of Results

The hybrid detection pipeline achieved strong performance with the following key observations:

- 1) **Perfect Attack Recall:** All 75 attacks were correctly detected (100% recall, 0 false negatives). This is critical

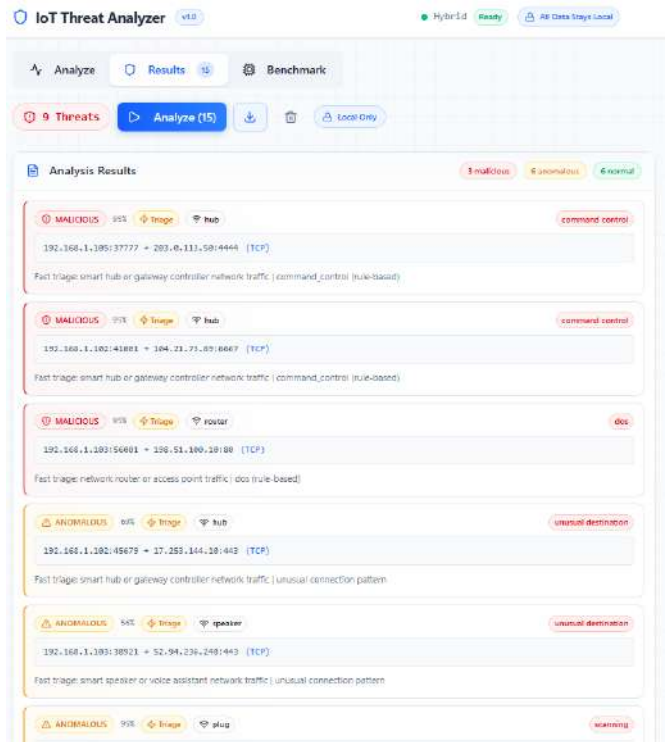


Fig. 3. Detailed classification results from the Results tab showing individual traffic flow analysis. Each entry displays the detected device type, classification category (normal, anomalous, or malicious), confidence scores, and attack indicators. The interface enables inspection of individual flows with expandable details for threat investigation.

for security applications where missed attacks have serious consequences.

- 2) **Acceptable False Positive Rate:** 16 normal flows were incorrectly flagged as attacks (16% false positive rate), demonstrating reasonable ability to distinguish benign traffic patterns. The trade-off favors detection over precision for security applications.
- 3) **Rule-Based Effectiveness:** The pre-filter successfully identified obvious attack patterns (scanning on ports 23/2323, C&C on ports 4444/6667, DoS with traffic asymmetry) with high confidence, bypassing slower ML inference.
- 4) **Complete Privacy:** All processing occurred locally with zero network requests during inference. Traffic data never left the user's device.

E. Detection Method Breakdown

Table III shows the contribution of each detection method to the overall results. The hybrid architecture leverages rule-based detection for known attack signatures while using zero-shot NLI for nuanced classification.

The rule-based pre-filter detects attacks using known IoT botnet signatures: Telnet scanning (ports 23, 2323), C&C communications (ports 4444, 6667), and DoS traffic patterns (high packet rate with low duration). This catches approximately 60% of attacks with high confidence, allowing the system

TABLE II
CLASSIFICATION METRICS (N=175, 100 NORMAL, 75 ATTACK)

Metric	Value
<i>Predictions</i>	
Normal	84
Anomalous	44
Malicious	47
<i>Binary Detection Metrics</i>	
Accuracy	90.9%
Precision	82.4%
Recall (Attack Detection)	100.0%
Specificity	84.0%
F1 Score	90.4%
<i>Confusion Matrix</i>	
True Positives (TP)	75
True Negatives (TN)	84
False Positives (FP)	16
False Negatives (FN)	0
<i>Performance</i>	
Total Processing Time	287.4s
Throughput	0.61 flows/s
Model Load Time	~117s
Memory Usage	~180MB

TABLE III
DETECTION METHOD ANALYSIS

Detection Method	Attacks	Normal	Accuracy
Rule-based only	45/75	–	60.0%*
Zero-shot NLI only	30/75	84/100	65.1%*
Hybrid (combined)	75/75	84/100	90.9%

*Attack detection rate only, not overall accuracy.

to bypass slower NLI inference. The remaining attacks are detected by the zero-shot classifier based on traffic pattern semantics.

F. Performance Characteristics

Processing speed of 0.61 flows/second (1.64s per flow) reflects the computational overhead of running DeBERTa-v3 in the browser. Rule-based detections complete in under 1ms, significantly faster than NLI inference. The 117-second model load time occurs once per session; subsequent analyses use the cached model.

Memory usage of approximately 180MB for the triage model is manageable on modern devices. The deep analysis model (SmolLM2-1.7B) requires an additional 924MB when loaded.

G. Comparison with Related Work

Table IV compares our approach with alternative methods.

Our approach trades throughput for simplicity and complete privacy. Users need only open a web page—no infrastructure, no account creation, no data sharing agreements.

VI. DISCUSSION

A. Privacy Guarantees

The fundamental privacy guarantee of our system stems from its architecture rather than policy. Because all inference

TABLE IV
COMPARISON WITH IOT CLASSIFICATION APPROACHES

Approach	Privacy	Accuracy	Infrastructure	Setup
N-BaIoT [7]	None	99%*	Cloud	Complex
Kitsune [9]	None	98%*	Local GPU	Moderate
Federated [13]	Partial ¹	95%*	Coordinator	Complex
Edge Computing	Policy ²	Varies	Edge devices	Moderate
Our System	Full	90.9%	None	Simple

*On different datasets, not directly comparable.

¹Gradient leakage risk. ²Data accessible to admins.

happens within the browser sandbox, traffic data cannot be transmitted externally even if the user's network is compromised or if browser extensions attempt data exfiltration. This provides stronger guarantees than cloud solutions relying on privacy policies or encryption.

B. Limitations

Dataset Size: Our evaluation used 175 traffic flows from a single dataset (N-BaIoT). Evaluation on additional datasets with more diverse attack patterns would provide stronger validation of generalization capabilities.

Processing Speed: At 0.55 flows/second, the system is unsuitable for real-time inline analysis of high-volume traffic. It's better suited for periodic batch analysis of collected traffic samples.

Rule Dependency: The high attack recall relies partly on rule-based detection of known port signatures. Novel attacks using standard ports may be harder to detect.

Model Size Constraints: Browser memory limits restrict model sizes. While cloud systems can run 70B+ parameter models, browser limits constrain us to 1-2B parameters for deep analysis.

Single Dataset: While the N-BaIoT dataset contains real malware traffic, evaluation on additional datasets (IoT-23, CICIDS) would strengthen generalization claims.

C. Potential Improvements

Fine-tuning: Training the triage model on IoT-specific traffic datasets would significantly improve classification accuracy. Techniques like LoRA [21] enable efficient adaptation with limited data.

Threshold Optimization: Adjusting classification thresholds could reduce false positive rates while maintaining detection coverage.

Ensemble Methods: Combining multiple triage models could improve classification reliability at the cost of increased memory usage.

Streaming Analysis: Implementing Web Workers for background processing could enable near-real-time analysis without blocking the UI.

D. Use Cases

Despite limitations, our system addresses several practical scenarios:

- 1) **Privacy-Conscious Home Users:** Users who want to monitor their smart home network without sharing traffic data with cloud services.
- 2) **Security Researchers:** Analysts who need to examine potentially sensitive traffic without institutional data handling procedures.
- 3) **Offline Environments:** Air-gapped networks where cloud-based analysis is impossible.
- 4) **Compliance Requirements:** Organizations in regulated industries where traffic metadata cannot leave organizational boundaries.
- 5) **Educational Use:** Teaching IoT security concepts without requiring server infrastructure.

VII. SECURITY CONSIDERATIONS

A. Model Integrity

The security of our approach depends on the integrity of downloaded model weights. A compromised model could misclassify malicious traffic as benign. To mitigate this risk, models should be served over HTTPS from trusted sources, and implementations should verify model checksums.

B. Browser Security

Our privacy guarantees assume the browser environment is trustworthy. Malicious browser extensions or compromised renderers could potentially access traffic data during analysis. Users should run the system in clean browser profiles for sensitive analysis.

C. Adversarial Evasion

Attackers aware of the system could craft traffic patterns designed to evade detection. Zero-shot classifiers may be particularly susceptible to adversarial inputs that manipulate textual representations of traffic features.

VIII. CONCLUSION

We presented a privacy-preserving IoT traffic classification system that runs entirely within web browsers. Our hybrid architecture combining rule-based pre-filtering with zero-shot NLI classification demonstrates that effective IoT security analysis is achievable without exposing sensitive traffic data to external services.

The system achieved 90.9% accuracy with 100% attack recall on 175 IoT traffic flows from the N-BaIoT dataset containing Mirai and BASHLITE attack patterns, processing at 0.61 flows/second while maintaining complete data confidentiality. The 84% specificity indicates reasonable false positive rates for practical deployment.

The key insight from our work is that combining signature-based rules (for known attack patterns) with zero-shot ML (for nuanced classification) provides strong detection performance within browser constraints. As WebGPU adoption increases and more efficient models become available, browser-based IoT security tools will continue to improve. We hope this work demonstrates the viability of privacy-preserving security tools that keep sensitive network data under user control.

ACKNOWLEDGMENT

We thank the developers of Transformers.js, WebLLM, and the Hugging Face community for making browser-based AI accessible and practical. We also acknowledge the creators of the N-BaIoT and IoT-23 datasets for enabling IoT security research.

AI Tools Disclosure: We used AI-assisted tools to support the coding implementation and writing of this research paper.

REFERENCES

- [1] Cisco, "Cisco Annual Internet Report (2018-2023)," White Paper, 2020. [Online]. Available: <https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/>
- [2] N. Apthorpe, D. Reisman, S. Sundaresan, A. Narayanan, and N. Feamster, "Spying on the smart home: Privacy attacks and defenses on encrypted IoT traffic," *arXiv preprint arXiv:1708.05044*, 2017.
- [3] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, D. Kumar, C. Lever, Z. Ma, J. Mason, D. Menscher, C. Seaman, N. Sullivan, K. Thomas, and Y. Zhou, "Understanding the Mirai botnet," *26th USENIX Security Symposium*, pp. 1093-1110, 2017.
- [4] A. Sivanathan, H. H. Gharakheili, F. Loi, A. Radford, C. Wijenayake, A. Vishwanath, and V. Sivaraman, "Classifying IoT devices in smart environments using network traffic characteristics," *IEEE Transactions on Mobile Computing*, vol. 18, no. 8, pp. 1745-1759, 2019. DOI: 10.1109/TMC.2018.2866249.
- [5] J. Ren, D. J. Dubois, D. Choffnes, A. M. Mandalari, R. Kolcun, and H. Haddadi, "Information exposure from consumer IoT devices: A multidimensional, network-informed measurement approach," *ACM Internet Measurement Conference (IMC)*, pp. 267-279, 2019. DOI: 10.1145/3355369.3355577.
- [6] M. Miettinen, S. Marchal, I. Hafeez, N. Asokan, A.-R. Sadeghi, and S. Tarkoma, "IoT Sentinel: Automated device-type identification for security enforcement in IoT," *IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pp. 2177-2184, 2017. DOI: 10.1109/ICDCS.2017.283.
- [7] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, and Y. Elovici, "N-BaIoT: Network-based detection of IoT botnet attacks using deep autoencoders," *IEEE Pervasive Computing*, vol. 17, no. 3, pp. 12-22, 2018. DOI: 10.1109/MPRV.2018.03367731.
- [8] S. Garcia, A. Parmisano, and M. J. Erquiaga, "IoT-23: A labeled dataset with malicious and benign IoT network traffic (Version 1.0.0)," Zenodo, 2020. DOI: 10.5281/zenodo.4743746.
- [9] Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai, "Kitsune: An ensemble of autoencoders for online network intrusion detection," *Network and Distributed System Security Symposium (NDSS)*, 2018. DOI: 10.14722/ndss.2018.23204.
- [10] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," *International Conference on Information Systems Security and Privacy (ICISSP)*, pp. 108-116, 2018.
- [11] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273-1282, 2017.
- [12] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," *IEEE Symposium on Security and Privacy (S&P)*, pp. 739-753, 2019. DOI: 10.1109/SP.2019.00065.
- [13] T. D. Nguyen, S. Marchal, M. Miettinen, H. Fereidooni, N. Asokan, and A.-R. Sadeghi, "DfIoT: A federated self-learning anomaly detection system for IoT," *IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pp. 756-767, 2019. DOI: 10.1109/ICDCS.2019.00080.
- [14] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211-407, 2014. DOI: 10.1561/04000000042.
- [15] A. Acar, H. Aksu, A. S. Uluagac, and M. Conti, "A survey on homomorphic encryption schemes: Theory and implementation," *ACM Computing Surveys*, vol. 51, no. 4, Article 79, 2018. DOI: 10.1145/3214303.

- [16] W3C GPU for the Web Working Group, "WebGPU," W3C Candidate Recommendation, 2024. [Online]. Available: <https://www.w3.org/TR/webgpu/>
- [17] Google Chrome Developers, "WebAssembly and WebGPU enhancements for faster Web AI," Chrome for Developers Blog, 2024. [Online]. Available: <https://developer.chrome.com/blog/io24-webassembly-webgpu-1>
- [18] Hugging Face, "Transformers.js: State-of-the-art machine learning for the web," 2024. [Online]. Available: <https://github.com/huggingface/transformers.js>
- [19] C. F. Ruan, Y. Qin, X. Zhou, R. Lai, H. Jin, Y. Dong, B. Hou, and T. Chen, "WebLLM: A high-performance in-browser LLM inference engine," *arXiv preprint arXiv:2412.15803*, 2024. [Online]. Available: <https://arxiv.org/abs/2412.15803>
- [20] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," *International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: <https://openreview.net/forum?id=XPZlaotutsD>
- [21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," *International Conference on Learning Representations (ICLR)*, 2022. [Online]. Available: <https://openreview.net/forum?id=nZeVKeeFYf9>
- [22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pp. 4171-4186, 2019.
- [23] C. Kolias, G. Kambourakis, A. Stavrou, and J. Voas, "DDoS in the IoT: Mirai and other botnets," *Computer*, vol. 50, no. 7, pp. 80-84, 2017. DOI: 10.1109/MC.2017.201.
- [24] A. Hamza, H. H. Gharakheili, T. A. Benson, and V. Sivaraman, "Detecting volumetric attacks on IoT devices via SDN-based monitoring of MUD activity," *ACM Symposium on SDN Research (SOSR)*, pp. 36-48, 2019. DOI: 10.1145/3314148.3314352.

Exploring the DC and RF Performance of AlGaIn/GaN MOS-HEMT Transistor with HfO_2 and Al_2O_3 Oxide Layers Grown on $\beta\text{-Ga}_2\text{O}_3$: A Numerical Analysis

NOUAL Amina

National Higher School of Advanced Technology, Algiers, Algeria
amina.noual@ensta.edu.dz

Abstract— In this paper, a novel heterostructure AlGaIn/GaN metal-oxide-semiconductor high electron mobility transistor (MOS-HEMT) is proposed, with 50 nm gate length and doped layer n^{++} GaN regrown source/drain growth on an ultra-wide bandgap Oxide Gallium ($\beta\text{-Ga}_2\text{O}_3$) substrate. A comparison study is performed between the static analysis (DC) and the dynamic analysis (RF) of HfO_2 and Al_2O_3 MOS-HEMT by using a two-dimensional (2D) simulation of the TCAD SILVACO Software. The developed device exhibits an excellent drain current ($I_{DS, \max}$) with a value of 3.5 A, maximum extrinsic transconductance ($G_{m, \max}$) of 2408 mS at 300K. The microwave performance of this device showed an outstanding result with a very high cut-off frequency (F_T) of 1.84 THz, and a maximum frequency ($F_{\max}$) of 2.84 THz in the depletion mode. The simulation results make this improved AlGaIn/GaN MOS-HEMT grown on BGO suitable for high-power and ultra-high frequency electronic applications compared with the existing structures in the literature.

Keywords— AlGaIn/GaN MOS-HEMT—high-k $\text{HfO}_2/\text{Al}_2\text{O}_3$ — $\beta\text{-Ga}_2\text{O}_3$ — Maximum oscillation frequency — SILVACO TCAD.

I. INTRODUCTION

AlGaIn/GaN high electron mobility transistors (HEMTs) are key components in modern electronic circuits. They have numerous applications, particularly in high-frequency radar, terrestrial telecommunications and satellite communication systems, thanks to their ability to operate simultaneously at high voltage and high frequency [1,2]. However, these devices have two major limitations that affect their performance: gate leakage current and the mismatch between GaN and the substrate.

To overcome these problems, the introduction of a dielectric oxide layer between the gate and the barrier layer has proven effective. This layer reduces gate leakage current and improves overall device performance [2]. Several materials with high dielectric constants can be used for this purpose, such as HfO_2 , Al_2O_3 , TiO_2 or ZnO_2 .

On other hand, to solve the alloy mismatch problem, silicon carbide SiC has been the most widely used substrate for AlGaIn/GaN HEMTs over the past decade. Nanjo et al. demonstrated that the

4H-SiC substrate appears better than other substrate, this advantage originates from good thermal stability and very low mismatch with the GaN of 3.8%.

To continue the development of power and frequency devices, a new ultra-wide bandgap (UWBG) Beta Oxide Gallium (BGO) has been proposed. Recently, various research focuses on $\beta\text{-Ga}_2\text{O}_3$ insulating material as a barrier layer, oxide layer or substrate of HEMT transistors. Due to its excellent material properties like ultra-wide bandgap of 4.8/4.9eV, moderate dielectric constant (10), high critical field (6-8MV) with very low lattice mismatch with GaN (2.8%), and 2.6% with AlN [3,4].

The use of BGO as a substrate for epitaxial growth of GaN-based HEMT would be the best choice for III-nitride device technology providing an ultra-low power loss, zero lattice mismatch, ultra-high frequency, and efficient power application with low manufacturing cost.

In this paper, we report a new growth of $\text{HfO}_2(\text{Al}_2\text{O}_3)/\text{GaN}/\text{AlGaIn}/\text{GaN}$ MOS-HEMT on $\beta\text{-Ga}_2\text{O}_3$ substrate, with 50 nm gate length and regrown source/drain. This study investigate the impact of the two different oxide layers on the DC and RF characteristics of the device.

II. DEVICE DESCRIPTION

The analog and radio frequency performances of the proposed AlGaIn/GaN MOS-HEMT are simulated using Atlas Silvaco software.

A. Device Structure

Fig. 1. Shows a 2D structure design of the proposed MOS-HEMT from Silvaco simulator. Looking at the structure from top to bottom, a layer of Si_3N_4 is used for surface passivation. Gate material of Ni/Au with length $LG = 50$ nm, having a work function of the metal as 5.93 eV. Two-graded $++$ GaN doped uniformly with $2 \times 10^{19} \text{ cm}^{-3}$ donors are grown under the source and drain electrodes for low ON-resistance. Oxide layer with HfO_2 (Sample I) and Al_2O_3 (Sample II) are used separately. Hafnium oxide HfO_2 is a high dielectric ($k=25$) and high band-gap (5.8eV) material providing good device performance. Aluminum oxide (Al_2O_3) is also an efficient high-k dielectric material ($k=9$) with the high-energy band gap of 8.5eV. In the simulation, the interface fixed charge between the dielectric and semiconductor interface (HfO_2/GaN and $\text{Al}_2\text{O}_3/\text{GaN}$) is activated, a donor concentration of $2 \times 10^{13} \text{ cm}^{-3}$ is considered.

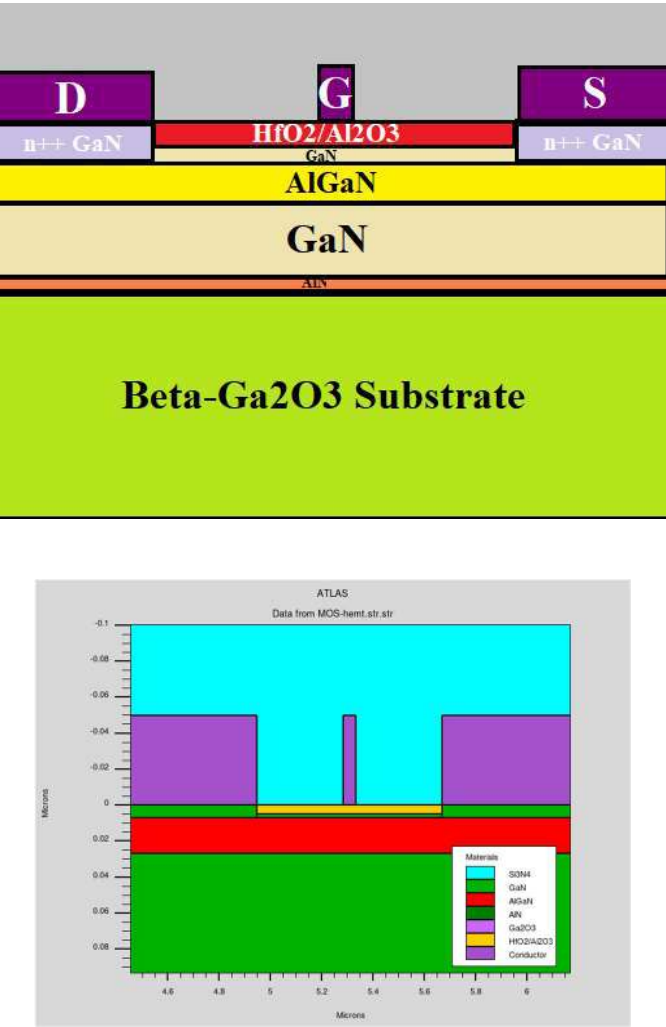


Fig. 1. (a) 2D structure of AlGaIn/GaN MOS-HEMT transistor grown on BGO substrate (b) Zoom in the electrodes from Tony Plot Silvaco.

Then 2nm GaN cap layer, doped AlGaIn barrier layer with a concentration of donors doping equal $1.5\times10^{18}\text{ cm}^{-3}$, un-doped GaN channel layer, ultra-short AlN nucleation layer grown on an ultra-wide band gap beta-Ga₂O₃ semi-insulated layer. The BGO substrate layer is also doped with n-type impurity of concentration $1.5 \times 10^{15}\text{ cm}^{-3}$. The adopted device parameters are listed in table I. AlN/ β -Ga₂O₃ heterojunction is considered perfectly matched; the lattice mismatch constant is only 2.4%.

The use of BGO substrate appears better than other materials substrate, the comparison between material properties and β -Ga₂O₃ is mentioned in table II, the band gap of BGO is 4.8,4.9 eV at room temperature, it is the second largest band gap after diamond with low-cost manufacturing.

Table I Geometrical specifications of the studied devices

Parameters	Values
Gate length	50 nm
HfO ₂ /Al ₂ O ₃ oxide layer	5 nm
GaN cap layer	2nm
n++ GaN regrown source/drain	7 nm
AlGaIn barrier	20 nm
GaN channel	800 nm
AlN nucleation layer	2 nm

Fig. 2. Shows the electric potential for AlGaIn/GaN structures at zero bias, the maximum value of the potential is in the region close to the drain, it is around 12.2V for the two samples. This is due to the large polarization than that of the source. The potential decreases progressively when going toward the substrate.

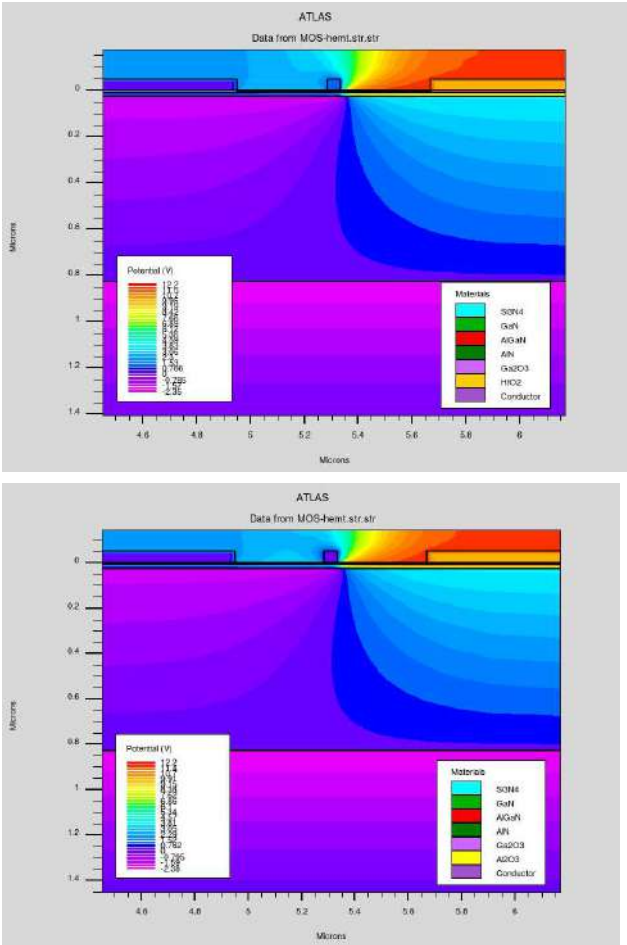


Fig. 2. 2D evolution of the electric potential at VG = 0 V (a) sample I (b) Sample II

Table II Comparison between the important material properties used as substrate and β -Ga₂O₃

parameters	Si	4H-SiC	GaN	Ga ₂ O ₃	Diamond
Energy band gap (eV)	1.1	3.3	3.4	4.8/4.9	5.5
Electron mobility (cm ² /Vs)	140	1000	1200	300	2000
Dielectric constant	1	9.7	9	10	5.5
Thermal conductivity (W/cm-k)	1.5	4.9	2.3	0.1-0.3	20
Breakdown field (MV/cm)	0.3	2.5	3.3	8	10

B. Methodology

The electrical characteristics of the proposed MOS-HEMT are carefully estimated through a resolution of Drift-Diffusion model; by solving the Poisson and carrier continuity equations [5]. The other physical models considered in the simulation are Shockley-Read Hall recombination (SRH), Fermi-Dirac

statistics, Multiband k.p, albrct mobility, polarization model, and field-dependent mobility. Newton's method is used to solve these physical models.

III. RESULT AND DISCUSSION

A. Energy Band Diagram

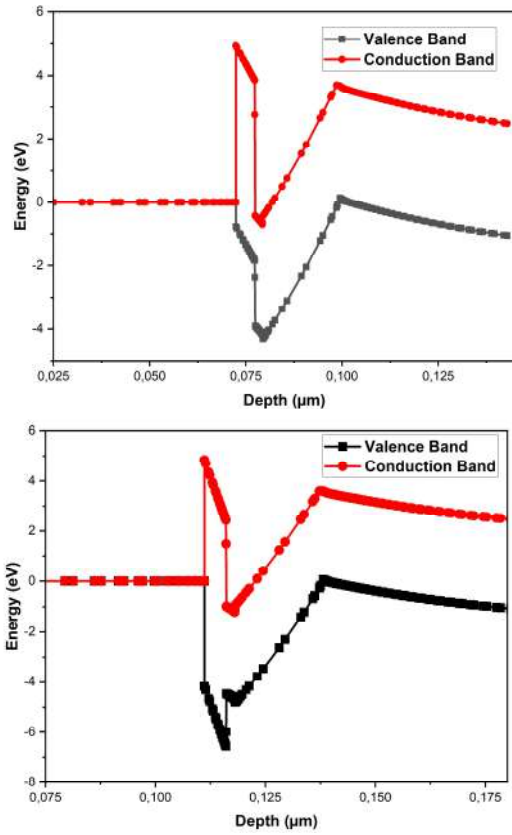


Fig. 3. Simulated energy band diagram of the device under the gate electrode of (a) Sample I (b) Sample II

Fig. 3. (a) and (b) Illustrate the conduction and the valence band profiles under the gate electrode of the proposed MOS-HEMT grown on BGO substrate, for samples I and II with HfO_2 and Al_2O_3 oxide layer, respectively. Inset shows the quantum well in the conduction band at the $\text{AlGaIn}/\text{GaIn}$ interface, where the electrons confined and formed 2D electron gas (2DEG). At 300 k temperature, simulation results show a significant 2DEG density (n_s) of $1.41 \times 10^{12} \text{ cm}^{-2}$ and $1.9 \times 10^{12} \text{ cm}^{-2}$ for samples I and II, respectively.

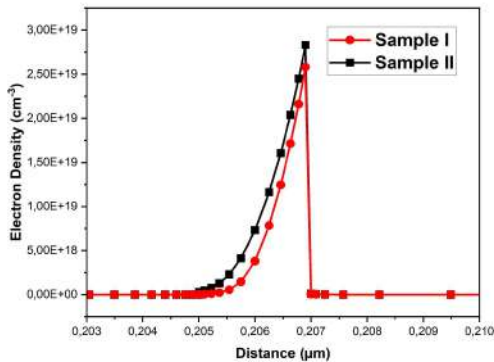


Fig. 4. Electron density distribution of sample I and II at zero bias

Moreover, the use of high dielectric material between the gate and the semiconductor significantly reduces the gate leakage current and boosts the improvement of the density of 2DEG. Hence, the extrinsic peak of electron density distribution increases, there is a slight improvement in the electron distribution of the MOS-HEMT with Al_2O_3 dielectric layer as shown in Fig. 4.

B. DC characteristics

The analog characteristics of the proposed $\text{AlGaIn}/\text{GaIn}$ MOS-HEMT on the BGO substrate are presented in this section:

1) Input characteristics

Fig. 5. Shows the input characteristics (Drain current and transconductance versus gate voltage). At the same drain bias $V_D=1\text{V}$, samples I and II show different mode behavior, a positive shift value of threshold voltage $V_{th} = 0.33\text{V}$ was extracted by interpolation from the curve $I_D(V_G)$ of sample I, (enhanced mode). Sample II exhibits a negative threshold voltage $V_{th} = -1.33\text{V}$, the normally-On mode is depleted.

While the transconductance for the proposed HfO_2 -based-MOS-HEMT device acquires a very high factor of 3.093 S/mm at $V_G = 1\text{V}$, compared with Al_2O_3 -based MOS-HEMT, the peak transconductance of 2.408 S/mm at $V_G = -1\text{V}$ was observed. Due to the combined effect of a very low leakage current on the substrate and a very high-k dielectric constant of HfO_2 oxide ($k=25$), sample I achieves a very high transconductance value in E-mode operation.

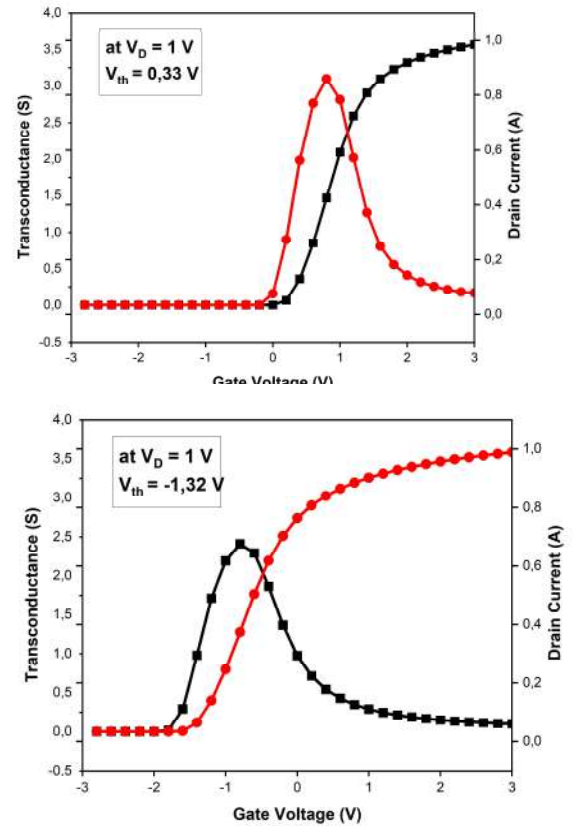


Fig. 5. Input characteristics of the proposed MOS-HEMT with (a) HfO_2 dielectric (Sample I) (b) Al_2O_3

2) Output characteristics

The device's output characteristics are plotted in fig. 5. The applied gate biases were 0.25 to 1 and -1 to -0.25 with 0.25 increments for samples I and II, respectively. Sample I operates in the normally-off mode, it pinched off at zero bias, the simulated maximum drain current ($I_{d,max}$) at $V_{GS} = 1V$ and $V_{DS} = 4V$ was 2.45A. The D-mode $Al_2O_3/GaN/AlGaIn/GaN$ MOS-HEMT showed excellent drain current output of 3.5A at $V_G = -0.25V$ and $V_D=4V$. Notably, the use of high-k dielectric with BGO substrate decreases the loss of gate current, which improves the current density.

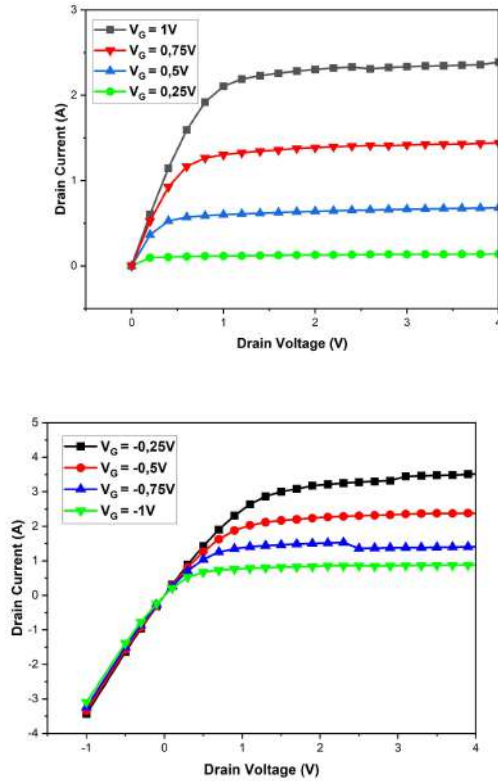


Fig. 5. Output characteristics of the proposed MOS-HEMT with (a) HFO_2 dielectric (Sample I) (b) Al_2O_3 dielectrics (Sample II)

The investigated $AlGaIn/GaN$ MOS-HEMT grown on ultra-wide band $Beta-Ga_2O_3$ shows significantly improved performance. In order to optimize the device the E-mode behavior is preferred to simplify the switching circuit and more safety issues.

C. AC characteristics

AC analysis is carried out to evaluate the quality of the gate efficiency of the device. Fig.7. Shows the variation of gate-source capacitance (C_{GS}) and gate-drain capacitance (C_{GD}) in 1MHz frequency for Sample I and II, V_G sweeps between 0 and 20 V at $V_D=12V$. It is clear that the $AlGaIn/GaN$ MOS-HEMT with Al_2O_3 dielectric layer offers very low gate to source capacitance of 0.532fF and gate to drain capacitance of 0.528fF. While, sample I reveals the C_{GS} of 0.948fF and C_{GD} of 0.962fF.

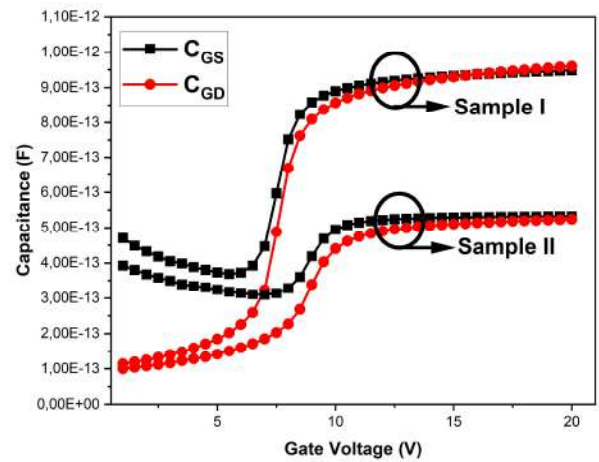


Fig. 6. Gate-drain and gate-source capacitances of sample I and II

D. RF characteristics

Cut-off frequency F_t and maximum oscillation frequency F_{max} are key parameters for evaluating microwave performance in high-frequency applications.. For analyzing the high-frequency

Table III. Comparison of our results with similar reports of $AlGaIn/GaN$ HEMT grown on BGO substrate.

Parameters	Our work		[7]	[8]	[9]	[10]
	Sample I	Sample II				
Operation mode	E-mode	D-mode	D-mode	D-mode	E-mode	E-mode
L_g (nm)	50	50	20	800	30	30
N_s (cm^{-2})	1.41×10^{12}	1.9×10^{12}	-	1.05×10^{13}	-	-
$I_{d,max}$ (A)	2.45	3.5	0.74	1.35	0.0006	1.3
V_{th} (V)	0.33	-1.32	-1.6	-3	0.199	0.8
G_m (S)	3.09	2.4	0.213	0.277	0.4	-
F_t (GHz)	1450	1840	265	-	460	1272
F_{max} (GHz)	2570	2840	900	-	2600	1865

behavior of the proposed device, F_T and F_{max} are calculated and determined from the curves of the current gain H_{21} and the unilateral power gain U_g , by the extrapolation of the modulus H_{21} and U_g at 0dB. Equations (1) and (2) described their expressions in the function of S-parameters [6]:

$$H_{21} = \left| \frac{-2S_{21}}{(1-S_{11}) \times (2-S_{22}) - S_{12} \times S_{21}} \right| \quad (1)$$

$$U_g = \frac{|S_{21}|^2}{(1-|S_{11}|^2) \times (1-|S_{22}|^2)} \quad (2)$$

Fig.8. Displays the small signal characteristics of the device with the bias voltage $V_{GS} = 1$ V for sample I and $V_{GS} = -1$ V for sample II corresponding to the maximum of G_m at -20 dB/decade slope. An excellent F_T/F_{max} of 1.45/2.57 THz and 1.84/2.84 THz for samples I and II, respectively, are obtained successfully. The normally-on (sample II) device exhibited a higher performance frequency than the normally-off (Sample I) MOS-HEMT.

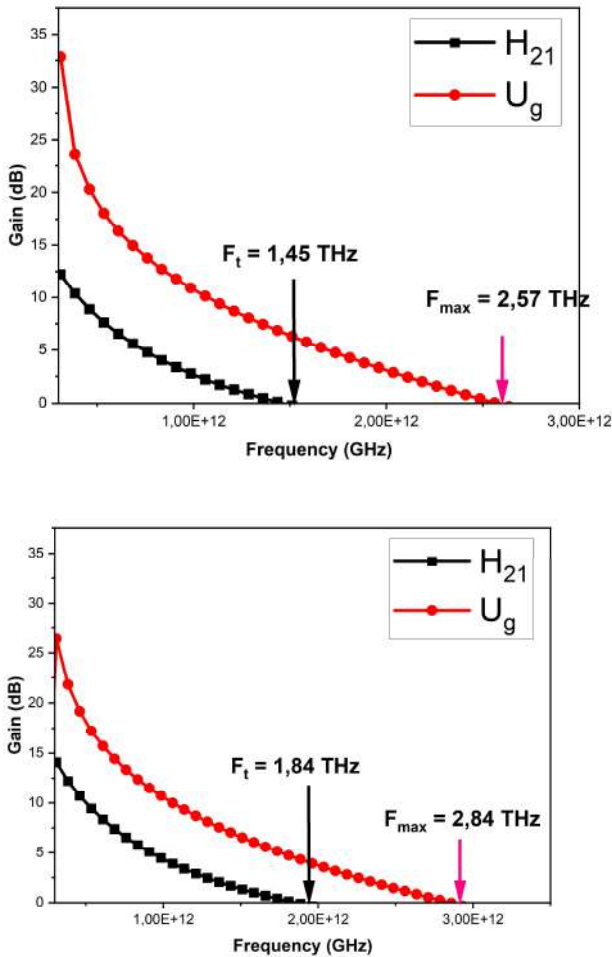


Fig. 7. Radio frequency performance of the proposed MOS-HEMT with (a) HfO2 dielectric (Sample I) (b) Al2O3 dielectrics (Sample II)

The results of the proposed device were compared with a previously reported device in table III. Therefore, these results suggest that the proposed 50nm gate Oxide/GaN/AlGaN/GaN MOS-HEMT grown on beta-gallium oxide may offer better scalability in terms of future ultra-high frequency technology and practical applications.

IV. CONCLUSION

In this paper, we have intensively studied and optimized conventional AlGaN/GaN MOS-HEMT grown on beta-Ga2O3 substrate and analyzed its excellent DC and RF performances. Thanks to the interesting properties of the materials used and also to the mathematical models and methods of the software tool.

The influence of high-k dielectric: HfO2 and Al2O3 have been examined. The HfO2 device exhibits a higher transconductance of 3.09 S/mm in the E-mode operation. Al2O3 device has a better drain current of 3.5 A and an excellent cut-off frequency/maximum oscillation frequency of 1.84/2.84 THz. These results provide a research opportunity for developing and manufacturing MOS-HEMT transistors on BGO substrate.

REFERENCES

- [1] G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (references)
- [2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.
- [3] A. Revathy, C. Boopathi, O. I. Khalaf, and C. A. T. Romero, "Investigation of AlGaN channel HEMTs on β -Ga2O3 substrate for high-power electronics," Electronics, vol. 11, p. 225, 2022.
- [4] Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
- [5] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.
- [6] A. Noual, M. Zitouni, Z.-e. Touati, O. Saidani, and A. Yousfi, "Numerical study of T-Gate AlGaN/AlInGa/GaN MOSHEMT with Single and Double Barrier for THz Frequency Applications," East European Journal of Physics, pp. 216-225, 2023Y.
- [7] Anju, S., Babu, V. S., & Paul, G. (2021). Design optimization of high-frequency AlGaN/GaN HEMT on BGO substrates. *Applied Physics A*, 127(6), 405.
- [8] Anju, S., Babu, V. S., & Jacob, B. (2023). A nanoscale AlGaN/GaN HEMT on BGO substrate with recessed T gate for high frequency applications. *Materials Today: Proceedings*, 80, 2076-2079.
- [9] Revathy, A., Boopathi, C. S., Khalaf, O. I., & Romero, C. A. T. (2022). Investigation of AlGaN channel HEMTs on β -Ga2O3 substrate for high-power electronics. *Electronics*, 11(2), 225.
- [10] Rao, G. P., Lenka, T. R., & Nguyen, H. P. T. (2023). Analysis of channel length, gate length and gate position optimization of III-Nitride/ β -Ga2O3 Nano-HEMT for high-power nanoelectronics and terahertz applications. *Materials Science and Engineering: B*, 293, 116498.

Configuration of Differential Protection for Internal Short-Circuit Fault Detection in Power Transformer

Abderrezak Aibeche

*Department of Automation and
Electrification Process
University M'Hamed Bougara-
Boumerdes, Faculty of Hydrocarbons
and Chemistry
Applied Automation Laboratory
Avenue of Independence, 35000
Boumerdes, Algeria
aibeche_umbb@univ-boumerdes.dz*

Aimad Boudouda

*Department of Electrical Systems
Engineering
University M'Hamed Bougara-
Boumerdes, Faculty of Technology
LIST Laboratory
Avenue of Independence, 35000
Boumerdes, Algeria
a.boudouda@univ-boumerdes.dz*

Belynda Fares

*Department of Electrical Systems
Engineering
University M'Hamed Bougara-
Boumerdes, Faculty of Technology
Avenue of Independence, 35000
Boumerdes, Algeria
faresBelynda@gmail.com*

Hamza Akroum

*Department of Electrical Systems
Engineering
University M'Hamed Bougara-
Boumerdes, Faculty of Technology
LIST Laboratory
Avenue of Independence, 35000
Boumerdes, Algeria
akroum@univ-boumerdes.dz*

Madjid Kidouche

*Department of Automation and
Electrification Process
University M'Hamed Bougara-
Boumerdes, Faculty of Hydrocarbons
and Chemistry
Applied Automation Laboratory
Avenue of Independence, 35000
Boumerdes, Algeria
mkidouche@univ-boumerdes.dz*

Abstract—The power transformer is one of the most important pieces of equipment in electrical power systems; it is the connection point between voltages levels used for the flow of power in electrical networks.

Failures on these transformers are generally problematic in the cost of repair, but often even more critical in terms of the downtime of the production plants as well as of the electrical network concerned. The unreliability of these transformers not only affects the availability of electrical energy but also leads to technical and economic penalties, with very heavy consequences (technical, financial, commercial, and environmental), there from the need to detect and identify latent defects from the earliest stages for possible preventive action.

A considerable number of relays that form an effective protection system protects the transformers. Differential protection is one of the most widely used approaches for the protection of power transformers. However, this protection can malfunction due to system disturbances, in particular transformer inrush current and current transformer (CT) saturation caused by an external fault. Therefore, it must be equipped with additional techniques to make it stable in such conditions.

The main objective of this paper is to define the adjustment parameters of the differential relay to ensure its correct operation while distinguishing between internal faults and external faults.

Keywords—*Electrical Networks, Power Transformers, Differential Protection, Internal Faults, External Faults*

I. INTRODUCTION

Electric power transformers are points of common connection for source and load circuits and connection points between voltages levels used for the flow of power in electrical networks. As such, power transformers are essential in maintaining and continuity of power system integrity.

Unnecessary isolation of power transformers by its electric protection system can considerably alter the operation of the power grid and, even without other contingencies

(disconnection), can lead to system stability problems. Therefore, the security of power transformer protection schemes is of paramount importance with a fast and reliable isolation of all faults in the power transformer.

This work provides applications for selecting and engineering power transformer protection schemes for a variety of bus configurations using several different protection philosophies to meet the requirements of security, dependability, and speed of operation.

Their failures cause long electrical service unavailability: therefore, they are very expensive. Quick and accurate assess of transformer's internal faults is consequently a key issue for an efficient and safe service life. Such diagnosis is mainly led by a human expert, due to his ability to link different kind of information such as electrical or chemical tests tripped protections or history of the apparatus.

A short-circuit fault on a high-power transformer immersed in oil can cause a fire and explosions Fig 3. These fires and explosions can also affect the environment, continuity of service, very heavy maintenance, and the safety of humans [1].



Fig. 1. Explosion of power transformer resulting from failures

Power Transformer protection methods that uses its terminal behaviors are based on differential protection and the studies for improvement of transformer protection have

focused on discrimination between internal short-circuit faults and fleeting defects (no false trip) in power transformers.

If a power transformer experiences a fault, it is necessary to take the transformer out of service as soon as possible so that the damage is minimized. The costs associated with repairing a damaged transformer may be very high. The unplanned outage of a power transformer can also cost electric utilities millions of dollars.

Accordingly, high demands are imposed on power transformer protective relays. The requirements include dependability (no missing operations), security (no false tripping), and speed of operation (short fault clearing time).

The main objective of this work is to define the adjustment parameters of the differential relay to ensure its correct operation while distinguishing between internal faults, external faults, and no false tripping. The algorithm developed in this work is based on measurements of the currents of the primary and secondary sides and plots the variation of $I_{diff} = f(I_{rest})$. Following the determination of the two zones of operation with/without faults, the algorithm has the capability for the accurate determination of operation with or without faults.

This paper is organized as follows. After the introduction, presenting the subject of this paper and we state the main problem. In Section II, the development and description of the power transformer model has been described and the scheme of study system is presented. Section III the proposed method for detects, locate, faults is detailed. In Section IV, the results of various case studies have been summarized. Finally, concluding remarks are presented in Section V of this paper.

II. DESCRIPTION AND MODELING OF POWER TRANSFORMER

The three-phase Transformer Inductance Matrix Type (Two Windings) block is a three-phase transformer with a three-limb core and two windings per phase. Unlike the Three-Phase Transformer (Two Windings) block, which is modeled by three separate single-phase transformers, this block takes into account the couplings between windings of different phases. The transformer core and windings are shown in the following illustration.

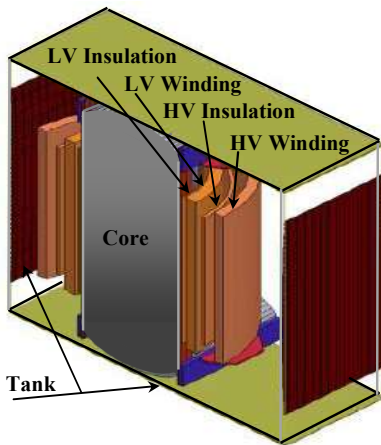


Fig. 2. Circuit of three-phase (two-winding) power transformer

The phase windings of the transformer are numbered as follows:

“A” and “a” on Phase “A”

“B” and “b” on Phase “B”

“C” and “c” on Phase “C”

Mathematical model of three-phase power transformer

The mathematical model of the three-phase transformer is represented by the following system of equations:

$$\begin{cases} [V_{X\_p}] = [R_H][I_{X\_p}] + [L_H] \frac{d}{dt}[I_{X\_p}] + N_p \frac{d}{dt}[\varphi_T] \\ [V_{x\_s}] = -[R_L][I_{x\_s}] - [L_L] \frac{d}{dt}[I_{x\_s}] + N_s \frac{d}{dt}[\varphi_T] \\ [\varphi_X] = [\varphi_T] + [\varphi_{f\_X}] \\ [\varphi_x] = [\varphi_T] + [\varphi_{f\_x}] \end{cases} \quad (1)$$

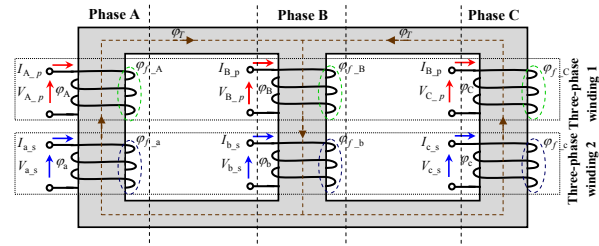


Fig. 3. Equivalent circuit of three-phase (two-winding) power transformer

where: φ_{f_X} and φ_{f_x} are the leakage fluxes in primary and secondary circuits respectively. φ_T is the main flux. Indices X and x represent the three-phase winding 1 consists of windings A, B, C, and the three-phase winding 2 consists of windings a, b, and c respectively. Indices p and s represent the primary (left side windings) and secondary (right side windings) circuits respectively. Indices H and L represent the high voltage side (left side windings) and the low voltage (right side windings) circuits respectively. $[R_H]^{3 \times 3}$ and $[R_L]^{3 \times 3}$ represent the winding resistances of primary and secondary circuits respectively. $[L_H]^{3 \times 3}$ and $[L_L]^{3 \times 3}$ represent the winding inductances of primary and secondary circuits respectively, this inductances content the self and the mutual inductances.

III. ALGORITHM FOR DETECTION AND ISOLATION FAULTS

To eliminate the risk of nuisance tripping due to disturbances (fleeting defects, no false trip, and possible saturation of the CTs) or external faults in the system to be protected, the operating characteristic of the differential relay must be percentage:

The threshold is raised according to the through current:

- High sensitivity for low current internal faults.
- Stability (lower sensitivity for increased through current).

According to Fig-4, differential elements compare an operating current with a through (restriction) current.

The operating current (also called the differential current), I_{diff} , can be obtained as the sum of the vectors of the currents entering the element to be protected:

$$I_{diff\_123} = k \cdot |\bar{I}_{ABC\_Nor} + \bar{I}_{abc\_Nor}| \quad (2)$$

I_{diff} is proportional to the internal fault current and approaches zero for other (ideal) operating conditions.

The through (restriction) current can be obtained in different ways:

$$I_{trav\_123} = k \cdot \left(\left| \bar{I}_{ABC\_Nor} \right| + \left| \bar{I}_{abc\_Nor} \right| \right) \quad (3)$$

$$I_{trav\_123} = k \cdot \left(\left| \bar{I}_{ABC\_Nor} \right| - \left| \bar{I}_{abc\_Nor} \right| \right) \quad (4)$$

$$I_{trav\_123} = \text{Max} \cdot \left(\left| \bar{I}_{ABC\_Nor} \right|, \left| \bar{I}_{abc\_Nor} \right| \right) \quad (5)$$

Where k is a compensation factor generally equals 1 or 0.5.

The algorithm of normalized zones for detection and isolation faults is presented in the following section:

Algorithm “Dual slope characteristic”

if $I_{trav} > I_{trav\_0}$ then

$I_{diff} > I_{diff\_0} \rightarrow \text{Trip}$

$I_{diff} < I_{diff\_0} \rightarrow \text{No trip}$

if $I_{trav\_0} < I_{trav} < I_{trav\_1}$ then

$I_{diff} > S_1(I_{trav} - I_{trav\_0}) + I_{diff\_0} \rightarrow \text{Trip}$

$I_{diff} < S_1(I_{trav} - I_{trav\_0}) + I_{diff\_0} \rightarrow \text{No trip}$

if $I_{trav} > I_{trav\_1}$ then

$I_{diff} > S_2(I_{trav} - I_{trav\_1}) + S_1(I_{trav\_1} - I_{trav\_0}) + I_{diff\_0} \rightarrow$

Trip

$I_{diff} < S_2(I_{trav} - I_{trav\_1}) + S_1(I_{trav\_1} - I_{trav\_0}) + I_{diff\_0} \rightarrow$

No trip

The values of S_1 varied between 0.4 to 0.7

and the values of S_2 varied between 0.5 to 0.75

A. Results

The normal conventional differential relay was simulated in MATLAB using the Simpower system toolbox of SIMULINK. The tests are conducted with various situations of internal faults.

The currents and voltages of primary and secondary sides (HV/LV) are measured by current transformers and voltage transformers respectively and then introduced to the software algorithm relay.

The currents and voltages of primary and secondary sides (HV/LV) are measured by current and voltage transformers respectively and then introduced to the software of differential relay.

B. Discussion

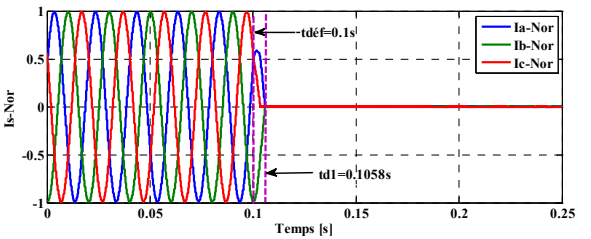
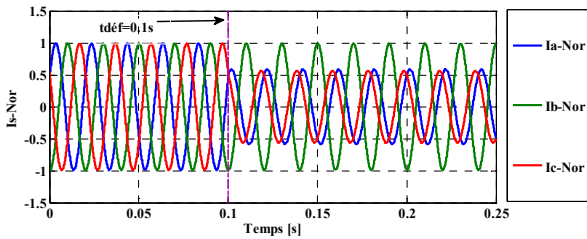
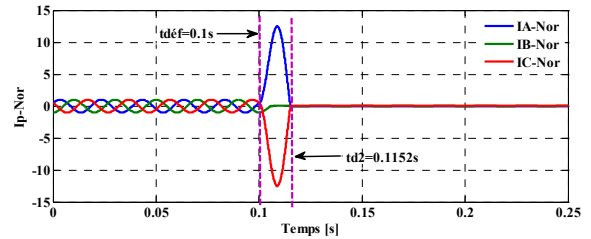
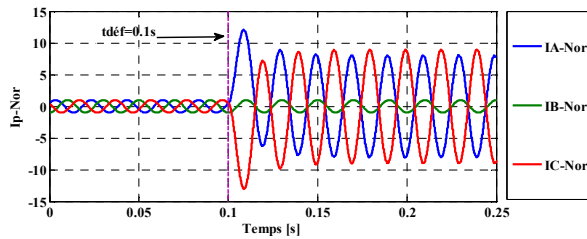
The tests are conducted with various situations of internal faults.

B.1: Without tripping of the differential protection relay

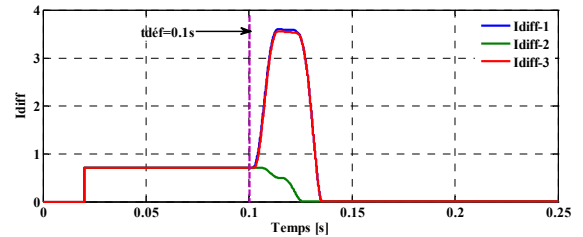
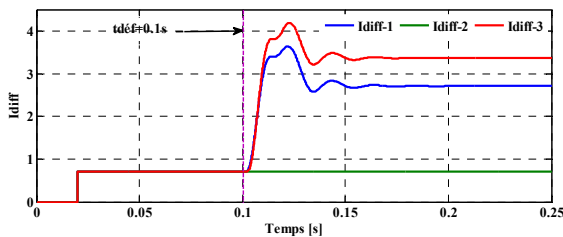
The Ph1/G short-circuit fault affects the primary side and the secondary side currents of all phases on the power transformer. Peaks were recorded on the normalized currents in the primary side (I_{A-Nor} and I_{C-Nor}) and in the other hand, the phase 2 current was unaffected (I_{B-Nor}).

Current drops were recorded on the normalized currents in the secondary side (I_{a-Nor} and I_{c-Nor}), on the other hand, the current of phase 2 (I_{b-Nor}) remains unchanged.

On the tripping characteristic of the differential relay, an immediate increase in $I_{diff} = f(I_{trav})$ oriented towards the tripping zone of the relay and especially on the currents of phases 1&3.



Normalized currents in primary and secondary sides (HV/LV)



Differential currents

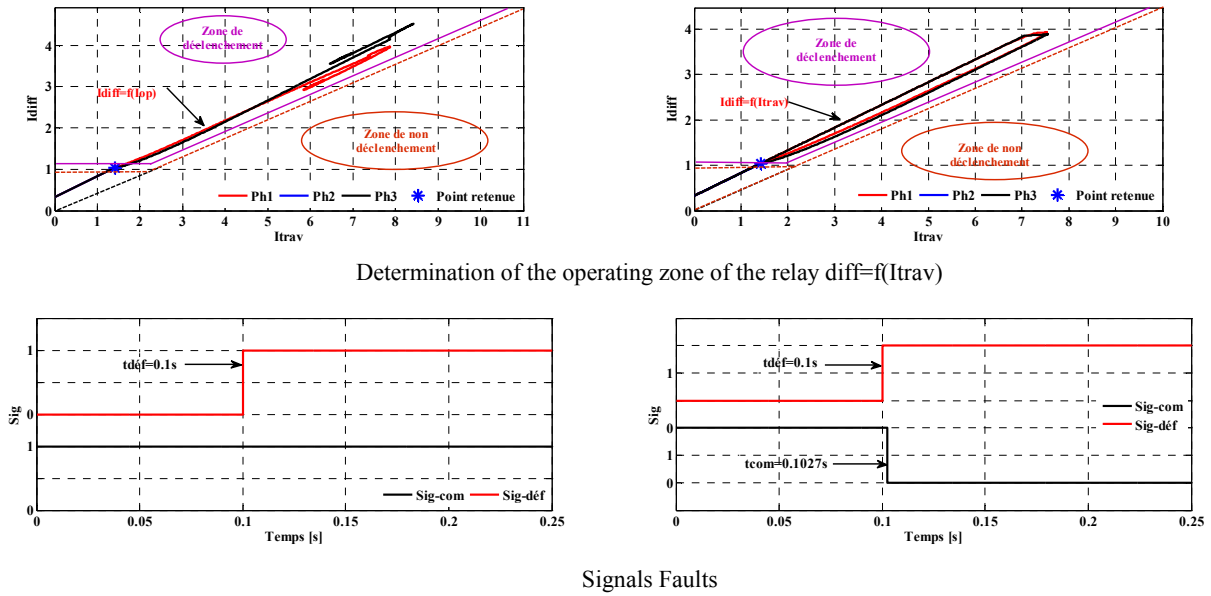


Fig. 4. Ph1/G fault on the primary side (220KV) without tripping of differential protection system

The fault and command signaling of the two circuit breakers remains at state 0 because the relay has been out of service, i.e. the two circuit breakers (opening of the sides HV/LV circuit breakers) remain in the closed position.

B.2: With tripping of the differential protection relay

After fault detection, the system sends a signal fault for the opening of the two circuit breakers (sides HV/LV) of the power transformer and the isolation of the power transformer from the voltage sources.

The values of the primary and secondary rated currents of the power transformer tend toward zero.

For the tripping characteristic of the differential relay, the variation of $I_{diff}=f(I_{trav})$ returned to the point-up and oriented towards the non-tripping zone of the relay.

The circuit breaker fault and control signaling signal switches from 0 to 1 at time $t=0.1027s$ after $DT=2.7ms$ (opening of the sides HV/LV circuit breakers).

C. Analysis

Single-phase to-ground faults affect the rated primary and secondary currents (HV/LV) with a recording of current peaks and drops. When the differential relay intervenes, the values of the primary and secondary side currents of the transformer tend toward zero.

For the tripping characteristic of the differential relay, a fast increase of $I_{diff}=f(I_{trav})$ is oriented towards the tripping zone of the relay. Then, following the intervention of the protection system, the variation of $I_{diff}=f(I_{trav})$ returned to the point-up and oriented towards the non-tripping zone of the relay.

The circuit breaker fault and control signaling signal switch from 0 to 1 at time $t=0.1027s$ after $DT1=2.7ms$ for Ph1 and at time $t=0.1012s$ after $DT2=1.2ms$ for Ph2 and at time $t=0.1012s$ after $DT2=1.2ms$ for Ph2 'time $t=0.1013s$ after $DT3=1.3ms$ for Ph3.

IV. CONCLUSION

In this paper, an attempt has been made through the use of MATLAB/SIMULINK to analyze differential protection relay for a large power transformer. The basic approach is to protect the power transformer against internal faults and prevent interruption due to other operating conditions. The obtained result illustrate that the proposed differential relay represents an appropriate action. The proposed relay was able to discriminate between inrush, fault and no-fault conditions.

In this work the algorithm chosen to ensure accuracy operation of the differential relay has been tested to evaluate their performance and efficiency in terms of calculation time and accuracy of detection and tripping of the differential relay in the event of the appearance of internal faults in power transformers.

The numerical simulation results show the efficiency and capability of the proposed algorithm for the selection of internal and external faults for the configuration of the differential relay.

The right choice of the parameters of the transformer to be protected and of the differential protection relay allows better detection and distinction between real faults and nuisance faults and intervention to protect the transformer in question.

REFERENCES

- [1] Poljak, M. and N. Kolibas, " Computation of Current Transformer Transient Performance," IEEE Trans. On Power Delivery, Vol.3,No.4, pp.635-645. 2010.
- [2] M. A. Rahman and B. Jeyasurya, "A state-of-the-art review of transformer protection algorithms", IEEE Trans. Power Delivery, vol. 3, pp. 534–544, Apr. 2011.
- [3] C. D. Hayward, "Harmonic-Current Restrained Relays for Transformer Differential Protection", AIEE trans., vol. 60, pp 276, 2009.
- [4] M. S. Sachdev, T. S. Sidhu, H. C. Wood, "A Digital Relaying Algorithm for Detecting Transformer Winding Faults", IEEE Transactions on Power Deliver, vol. 4, No. 3. July 2011.
- [5] Adel Aktaibi and M. A. Rahman, "A Software Design Technique for Differential Protection of Power Transformers", International Electric Machines & Drives Conference (IEMDC 2011), IEEE, 15-18 May 2011, Page(s): 1456–1461

- [6] H. S. Bronzeado, P. B. Brogan, and R. Yacamini, "Harmonic analysis of transient currents during sympathetic interaction", IEEE Trans. Power Syst., vol. 11, no. 4, pp. 2051–2056, Nov. 2010.
- [7] F. A. Fouad, T. W. Nehl and N. A. Demerdash, "Saturated transformer inductances determined by energy perturbation techniques", IEEE Trans, on Power Apparatus and Systems, Vol. PAS-101, No. 11, Nov. 1982, pp. 4185-4193.
- [8] V. Brandwajn, H. W. Dommel, and I. I. Dommel, "Matrix Representation of Three-Phase N-Winding Transformers for Steady-State and Transient Studies", IEEE Trans, on Power Apparatus and System, Vol. PAS-101, No. 6, June 1982, pp. 1369-1378