

Nur Syazreen Ahmad
Junita Mohamad-Saleh
Jiashen Teh *Editors*

Proceedings of the 12th International Conference on Robotics, Vision, Signal Processing and Power Applications

Lecture Notes in Electrical Engineering

Volume 1123

Series Editors

- Leopoldo Angrisani, Department of Electrical and Information Technologies Engineering, University of Napoli Federico II, Napoli, Italy
- Marco Arteaga, Departamento de Control y Robótica, Universidad Nacional Autónoma de México, Coyoacán, Mexico
- Samarjit Chakraborty, Fakultät für Elektrotechnik und Informationstechnik, TU München, München, Germany
- Jiming Chen, Zhejiang University, Hangzhou, Zhejiang, China
- Shanben Chen, School of Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
- Tan Kay Chen, Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore
- Rüdiger Dillmann, University of Karlsruhe (TH) IAIM, Karlsruhe, Baden-Württemberg, Germany
- Haibin Duan, Beijing University of Aeronautics and Astronautics, Beijing, China
- Gianluigi Ferrari, Dipartimento di Ingegneria dell'Informazione, Sede Scientifica Università degli Studi di Parma, Parma, Italy
- Manuel Ferre, Centre for Automation and Robotics CAR (UPM-CSIC), Universidad Politécnica de Madrid, Madrid, Spain
- Faryar Jabbari, Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA, USA
- Limin Jia, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China
- Janusz Kacprzyk, Intelligent Systems Laboratory, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
- Alaa Khamis, Department of Mechatronics Engineering, German University in Egypt El Tagamoa El Khames, New Cairo City, Egypt
- Torsten Kroeger, Intrinsic Innovation, Mountain View, CA, USA
- Yong Li, College of Electrical and Information Engineering, Hunan University, Changsha, Hunan, China
- Qilian Liang, Department of Electrical Engineering, University of Texas at Arlington, Arlington, TX, USA
- Ferran Martín, Departament d'Enginyeria Electrònica, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain
- Tan Cher Ming, College of Engineering, Nanyang Technological University, Singapore, Singapore
- Wolfgang Minker, Institute of Information Technology, University of Ulm, Ulm, Germany
- Pradeep Misra, Department of Electrical Engineering, Wright State University, Dayton, OH, USA
- Subhas Mukhopadhyay, School of Engineering, Macquarie University, Sydney, NSW, Australia
- Cun-Zheng Ning, Department of Electrical Engineering, Arizona State University, Tempe, AZ, USA
- Toyooki Nishida, Department of Intelligence Science and Technology, Kyoto University, Kyoto, Japan
- Luca Oneto, Department of Informatics, Bioengineering, Robotics and Systems Engineering, University of Genova, Genova, Genova, Italy
- Bijaya Ketan Panigrahi, Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, Delhi, India
- Federica Pascucci, Dipartimento di Ingegneria, Università degli Studi Roma Tre, Roma, Italy
- Yong Qin, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China
- Gan Won Seng, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore
- Joachim Speidel, Institute of Telecommunications, University of Stuttgart, Stuttgart, Germany
- Germano Veiga, FEUP Campus, INESC Porto, Porto, Portugal
- Haitao Wu, Academy of Opto-electronics, Chinese Academy of Sciences, Haidian District Beijing, China
- Walter Zamboni, Department of Computer Engineering, Electrical Engineering and Applied Mathematics, DIEM—Università degli studi di Salerno, Fisciano, Salerno, Italy
- Junjie James Zhang, Charlotte, NC, USA
- Kay Chen Tan, Department of Computing, Hong Kong Polytechnic University, Kowloon Tong, Hong Kong

The book series *Lecture Notes in Electrical Engineering* (LNEE) publishes the latest developments in Electrical Engineering—quickly, informally and in high quality. While original research reported in proceedings and monographs has traditionally formed the core of LNEE, we also encourage authors to submit books devoted to supporting student education and professional training in the various fields and applications areas of electrical engineering. The series cover classical and emerging topics concerning:

- Communication Engineering, Information Theory and Networks
- Electronics Engineering and Microelectronics
- Signal, Image and Speech Processing
- Wireless and Mobile Communication
- Circuits and Systems
- Energy Systems, Power Electronics and Electrical Machines
- Electro-optical Engineering
- Instrumentation Engineering
- Avionics Engineering
- Control Systems
- Internet-of-Things and Cybersecurity
- Biomedical Devices, MEMS and NEMS

For general information about this book series, comments or suggestions, please contact leontina.dicecco@springer.com.

To submit a proposal or request further information, please contact the Publishing Editor in your country:

China

Jasmine Dou, Editor (jasmine.dou@springer.com)

India, Japan, Rest of Asia

Swati Meherishi, Editorial Director (Swati.Meherishi@springer.com)

Southeast Asia, Australia, New Zealand

Ramesh Nath Premnath, Editor (ramesh.premnath@springernature.com)

USA, Canada

Michael Luby, Senior Editor (michael.luby@springer.com)

All other Countries

Leontina Di Cecco, Senior Editor (leontina.dicecco@springer.com)

**** This series is indexed by EI Compendex and Scopus databases. ****

Nur Syazreen Ahmad · Junita Mohamad-Saleh ·
Jiashen Teh
Editors

Proceedings of the 12th International Conference on Robotics, Vision, Signal Processing and Power Applications

 Springer

Editors

Nur Syazreen Ahmad
School of Electrical and Electronic
Engineering
Universiti Sains Malaysia
Nibong Tebal, Penang, Malaysia

Junita Mohamad-Saleh
School of Electrical and Electronic
Engineering
Universiti Sains Malaysia
Nibong Tebal, Penang, Malaysia

Jiashen Teh
School of Electrical and Electronic
Engineering
Universiti Sains Malaysia
Nibong Tebal, Penang, Malaysia

ISSN 1876-1100

ISSN 1876-1119 (electronic)

Lecture Notes in Electrical Engineering

ISBN 978-981-99-9004-7

ISBN 978-981-99-9005-4 (eBook)

<https://doi.org/10.1007/978-981-99-9005-4>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.

The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Paper in this product is recyclable.

List of Reviewers (RoViSP 2023)

Abdul Sattar Din, USM
Aeizal Azman Abdul Aahab, USM
Azniza Abd Aziz, USM
Bakhtiar Affendi Rosdi, USM
Bee Ee Khoo, USM
Chia Ai Ooi, USM
Dahaman Ishak, USM
Dzati Athiar Ramli, USM
Haidi Ibrahim, USM
Intan Zainal Abidin, USM
Jagadheswaran Rajendran, USM
Jiashen Teh, USM
Junita Mohamad-Saleh, USM
Mohamad Kamarol Mohd Jamil, USM
Mohamad Khairi Ishak, USM
Mohamad Tarmizi Abu Seman, USM
Mohamed Fauzi Packeer Mohamed, USM
Mohamed Salem, USM
Mohd Fadzli Mohd Salleh, USM
Mohd Khairunaz Mat Desa, USM
Mohd Nadhir Ab Wahab, USM
Mohd Tafir Mustaffa, USM
Muhammad Firdaus Akbar, USM
Muhammad Hafeez Mohamed Hariri, USM
Muhammad Najwan Hamidi, USM
Muhammad Nasiruddin Mahyuddin, USM
Nor Ashidi Mat Isa, USM
Nor Azlin Ghazali, USM
Nor Muzlifah Mahyuddin, USM
Noramalina Abdullah, USM
Nur Syazreen Ahmad, USM

Nur Zatil 'Ismah Hashim, USM
Soo Siang Teoh, USM
Tay Lea Tien, USM

Contents

Electrical Power, Energy and Industrial Applications	
Finite Element Modeling of Electric Field Distribution in a Defective XLPE Cable Insulation Under Different Magnitudes of Stressing Voltage	3
S. H. Sulaiman, Mohamad N. K. H. Rohani, A. Abdulkarim, A. S. Abubakar, G. S. Shehu, U. Musa, A. A. Mas'ud, N. Rosle, and F. Muhammad-Sukki	
Optimal Switching Sequence of Urban Power System Based on Dynamic Thermal Rating Parameter Adjustment	11
Yi Su and Jiashen Teh	
Improved DC-Link Voltage Control Scheme for Standalone PV Renewable Energy System	19
Muhammad Najwan Hamidi	
Wind Energy Distributions for Integration with Dynamic Line Rating in Grid Network Reliability Assessment	27
Olatunji Ahmed Lawal and Jiashen Teh	
Modeling of Multijunction Solar Cells InGaAs/InGaP/GaAs/GeSi for Improving the Efficiency of PV Modules by 43%	35
Muhammad Shehram, Muhammad Najwan Hamidi, Aeizaal Azman A.Wahab, and M. K. Mat Desa	
Control Method of Two-Stage Grid-Connected PV Inverter System	43
Wang Zhe, Dahaman Ishak, and Muhammad Najwan Hamidi	
Methodological Comparison and Analysis for Six-Switching PMSM Motor Control	53
Musa Mohammed Gujja, Dahaman Ishak, and Muhammad Najwan Hamidi	

Energy Efficiency Performance Optimization and Surge Prediction of Centrifugal Gas Compressor	61
Mukhtiar Ali Shar, Masdi B Muhammad, Ainul Akmar B Mokhtar, and Mahnoor Soomro	
A Novel Fuzzy PID Control Algorithm of BLDC Motor	69
Zhou Hongqiang and Dahaman Ishak	
Optimized Design Point Model of SGT500 Using GasTurb 14	77
Mahnoor Soomro, Tamiru Alemu Lemma, Syed Ihtsham Ul-Haq Gilani, and Mukhtiar Ali Shar	
Optimal RE-DG and Capacitor Placement for Cost-Benefit Maximization in Malaysia Distribution System	85
Maizatul Shafiqah Sharul Anuar, Mohd Nabil Muhtazaruddin, Mohd Azizi Abdul Rahman, and Mohd Effendi Amran	
Internal Discharge Patterns Identification of Void in High Voltage Solid Insulation Using Phase Resolved Method	93
N. Rosle, M. N. K. H. Rohani, N. A. Muhamad, S. A. Suandi, and M. Kamarol	
Exploring the Impact of Accelerated Thermal Aging on POME-Based MWCNT Nanofluid	101
Sharifah Masniah Wan Masra, Yanuar Zulardiansyah Arief, Siti Kudnie Sahari, Ernieza Musa, Andrew Ragai Henry Rigit, and Md. Rezaur Rahman	
Accelerating Electric Vehicle Adoption on Malaysian Islands: Lessons from Japan's Islands of the Future Initiative	109
M. Reyasudin Basir Khan, Jabbar Al-Fattah, Gazi Md. Nurul Islam, Ahmad Anwar Zainuddin, Chong Peng Lean, Saidatul Izyanie Kamarudin, and Miqdad Abdul Aziz	
Effect of Incidence Angle on the Performance of a Dual Cantilever Flutter Energy Harvester	117
Venod Reddy Velusamy, Muhammad Izzikry Mohd Farid Suhaimi, and Faruq Muhammad Foong	
Temporal Distribution of Thunderstorm Activity in Southern Region of Peninsular Malaysia	125
Shirley Anak Rufus, N. A. Ahmad, Z. A. Malek, Noradlina Abdullah, Nurul 'Izzati Hashim, and Noor Syazwani Mansor	

Electronic Design and Applications

Enhance the AlGaN/GaN HEMTs Device Breakdown Voltage by Implementing Field Plate: Simulation Study 133

Naeemul Islam, Mohamed Fauzi Packeer Mohamed, Firdaus Akbar Jalaludin Khan, Nor Azlin Ghazali, Hiroshi Kawarada, Mohd Syamsul, Alhan Farhanah Abd Rahim, and Asrulnizam Abd Manaf

Design of Low-Power and Area-Efficient Square Root Carry Select Adder Using Binary to Excess-1 Converter (BEC) 141

Poh Yui Lyn, Nor Azlin Ghazali, Mohamed Fauzi Packeer Mohamed, and Muhammad Firdaus Akbar

Simulation of Bottom-Gate Top-Contact Pentacene Based Organic Thin-Film Transistor Using MATLAB 149

Law Jia Wei and Nor Azlin Ghazali

Acoustic Beamforming Using Machine Learning 157

Te Meng Ting and Nur Syazreen Ahmad

A Comparative Analysis on Electrical and Photovoltaic Performances of MIS Structures on High Resistivity Silicon with Tunneling Insulator 165

Nur Bashirouh binti Attaullah, Nur Zatil ‘Ismah Hashim, Chong Kah Hui, Nor Muzlifah Mahyuddin, Alhan Farhanah Abd Rahim, Mohd Marzaini bin Mohd Rashid, and Mundzir Abdullah

Robotics, Control, Mechatronics and Automation

An Analysis of the Performance of the SPRT Chart with Estimated Parameters Under the Weibull Distribution 175

Jing Wei Teoh, Wei Lin Teoh, Laila El-Ghandour, Zhi Lin Chong, and Sin Yin Teh

Semi-Automatic Liquid Filling System Using NodeMCU as an Integrated IoT Learning Tool 183

N. F. Adan, Z. Zainal, M. N. A. H. Sha’abani, M. F. Mohamed Nor, and M. F. Ismail

Short Range Radio Frequency (RF) Data Acquisition Unit for Agricultural Product Monitoring System 191

S. M. N. S. Shatir, A. B. Elmi, M. N. Akhtar, M. N. Abdullah, and A. H. Ismail

Enhanced Parameter Estimation of Solar Photovoltaic Models Using QLESCA Algorithm	199
Qusay Shihab Hamad, Sami Abdulla Mohsen Saleh, Shahrel Azmin Suandi, Hussein Samma, Yasameen Shihab Hamad, and Imran Riaz	
Active Disturbance Rejection Control of Flexible Joint System	207
Li Qiang and Nur Syazreen Ahmad	
Research on Synchronous Control of Double-Cylinder Electro-Hydraulic Position Servo System Based on Active Disturbance Rejection Control	215
Liu Lizhen, Li Qiang, and Nur Syazreen Ahmad	
Passivity-Based Control of Underactuated Rotary Inverted Pendulum System	223
Minh-Tai Vo, Van-Dong-Hai Nguyen, Hoai-Nghia Duong, and Vinh-Hao Nguyen	
Comparative Analysis of Data-Driven Models for DC Motors with Varying Payloads	231
Helen Shin Huey Wee and Nur Syazreen Ahmad	
Development and Control of Underactuated Parallel Rotary Double Inverted Pendulum System	239
Minh-Tai Vo, Minh-Duy Vo, Van-Dat Nguyen, Van-Dong-Hai Nguyen, Minh-Duc Tran, Hoai-Nghia Duong, and Thanh T. Tran	
A Study of the Influence of Steel Brushes in Rail Surface Magnetic Flux Leakage Detection Using Finite Elements Simulation	247
Gong Wendong, Muhammad Firdaus Akbar, Mimi Faisyalini Ramli, and Ghassan Nihad Jawad	
AGVs and AMRs Robots: A Brief Overview of the Differences and Navigation Principles	255
Sami Abdulla Mohsen Saleh, Shahrel Azmin Suandi, Haidi Ibrahim, Qusay Shihab Hamad, and Ibrahim Al Amoudi	
Development of Delivery Robot with Application of the TRIZ Method	261
Zulkifli Ahmad and Muhd Aslam Zulkarnain	
Enhancement of Adaptive Observer for Fault Detection in Direct Current Motor System Using Kalman Filter	269
Nur Dalilah Alias and Rosmiwati Mohd Mokhtar	
Model Reference Adaptive Control for Acoustic Levitation System Based on Standing Waves	277
Ibrahim Ismael Ibrahim Al-Nuaimi and Muhammad Nasiruddin Mahyuddin	

Telecommunication Systems and Applications

A Review of Slotted Hollow Pyramidal Absorbers for Microwave Frequency Range 287
 Ala A. Abu Sanad, Mohd Nazri Mahmud, Mohd Fadzil Ain, Mohd Azmier Bin Ahmad, Nor Zakiah Binti Yahaya, and Zulkifli Mohamad Ariff

The Investigation of Perceptual Speech 5G Wireless Communication Networks 295
 Tuan Ulfah Uthailah Tuan Mohd Azran, Ahmad Zamani Jusoh, Ani Liza Asnawi, Khaizuran Abdullah, Md. Rizal Othman, and Nur Idora Abdul Razak

Comparison of Different Data Detection Methods in Orthogonal Frequency Division Multiplexing (OFDM) System 303
 Nur Qamarina Muhammad Adnan, Aeizal Azman Abdul Wahab, Syed Sahal Nazli Alhady, and Wan Amir Fuad Wajdi

A Wideband 3 dB T-Shaped Stubs-Loaded Coupler for Millimeter-Wave (mm-Wave) Beamforming Network Towards Fifth Generation (5G) Technology 311
 Nazleen Syahira Mohd Suhaimi and Nor Muzlifah Mahyuddin

Mathematical Modelling of Artificial Magnetic Conductor Backed Antenna On-Chip Using Response Surface Method for 28 GHz Application 317
 Ahmadu Girgiri, Mohd Fadzil Bn Ain, Abdullahi S. B. Mohammed, and Muhammad Bello Abdullahi

A Method Combining Compressive Sensing-Based Method of Moment and LU Decomposition for Solving Monostatic RCS 325
 Yalan Gao, Muhammad Firdaus Akbar, Jagadheswaran Rajendran, and Ghassan Nihad Jawad

Microwave Non-destructive Testing Using K-Medoids Clustering Algorithm 333
 Tan Shin Yee, Muhammad Firdaus Akbar, Nor Azlin Ghazali, Ghassan Nihad Jawad, and Nawaf H. M. M. Shrifan

Microwave Nondestructive Evaluation Using Spiral Inductor Probe 341
 Danladi Agadi Tonga, Muhammad Firdaus Akbar, Ahmed Jamal Abdullah Al-Gburi, Imran Mohd Ibrahim, Mohammed Fauzi Packeer Mohammed, and Mohammed Mydin M. Abdul Kader

Design of a Hairpin SIR Dual-Band Bandpass Filter with Defected Ground Slots for WLAN Application 349
 Nur Irdina Rizal, Azniza Abd Aziz, and Intan Sorfina Zainal Abidin

A Comparative Analysis of BLE-Based Indoor Localization with Machine Learning Regression Techniques 357
Chia Wei Khor and Nur Syazreen Ahmad

Determination of Anechoic Chamber Set-Up Using Simulation Approach: A Review 363
Roslina Hussin, Mohd Nazri Mahmud, Mohd Fadzil Ain, and Nor Zakiah Yahaya

Hybridization of Equilibrium and Grasshopper Optimization Algorithms 371
Ebinowen Tusin Dayo and Junita Mohamad-Saleh

Vision, Image and Signal Processing

An Enhanced Double EWMA Chart for Monitoring the Process Mean Shifts 381
Peh Sang Ng, Sook Yan Goh, Sajal Saha, and Wai Chung Yeong

Face Recognition Attendance Management System (FRAMS) Algorithm Using CNN Model 391
Saw Yang Yi, Mohd Izzat Nordin, and Mohamad Tarmizi Abu Seman

A Low-Cost Vibration Measurement Using MEMS MPU9250 Accelerometer with DLPF Filtering 399
Mohd Affan Mohd Rosli, Abdul Haadi Abdul Manap, and Ahmad Zhafran Ahmad Mazlan

Sampling Methods to Balance Classes in Dermoscopic Skin Lesion Images 407
Quynh T. Nguyen, Tanja Jancic-Turner, Avneet Kaur, Raouf N. G. Naguib, and Harsa Amylia Mat Sakim

Unsupervised Clustering to Reduce Overfitting Issues in Ensemble Deep Learning Models for Skin Lesion Classifications 415
Avneet Kaur, Tanja Jancic-Turner, Quynh T. Nguyen, Satyam Vatts, and Harsa Amylia Mat Sakim

Drone Detection Using Swerling-I Model with L-Band/X-Band Radar in Free Space and Raining Scenario 421
Salman Liaquat, Nor Muzlifah Mahyuddin, and Ijaz Haider Naqvi

Enhancing Generalized Electrocardiogram Biometrics Transformer 429
Kai Jye Chee and Dzati Athiar Ramli

Performance Evaluation of Different CNN Models for Motor Fault Detection Based on Thermal Imaging 437
Lifu Xu and Soo Siang Teoh

Comparative Analysis of Deep Learning-Based Abdominal Multivisceral Segmentation 445
 Junting Zou and Mohd Rizal Arshad

A Review of ECG Biometrics: Generalization in Deep Learning with Attention Mechanisms 453
 Aini Hafizah Mohd Saod and Dzati Athiar Ramli

Detecting Sleep Disorders from NREM Using DeepSDBPLM 459
 Haifa Almutairi, Ghulam Mubashar Hassan, and Amitava Datta

Application of Fuzzy Logic in Stock Markets by Using Technical Analysis Indicators 469
 Leow Wei Kang, Mohd Izzat Nordin, Abdul Sattar Din, and Mohamad Tarmizi Abu Seman

YOLOv7-Tiny and YOLOv8n Evaluation for Face Detection 477
 Ibrahim Al Amoudi and Dzati Athiar Ramli

Optimizing Feature Selection for Industrial Casting Defect Detection Using QLESCA Optimizer 485
 Qusay Shihab Hamad, Sami Abdulla Mohsen Saleh, Shahrel Azmin Suandi, Hussein Samma, Yasameen Shihab Hamad, and Ibrahim Al Amoudi

Improving the Accuracy of Gender Classification Based on Skin Tone Using Convolutional Neural Network: Transfer Learning (CNN-TL) 493
 Muhammad Firdaus Mustapha, Nur Maisarah Mohamad, Siti Haslini Ab Hamid, and Nik Amnah Shahidah Abdul Aziz

Literature Survey on Edge Detection-Based Methods for Blood Vessel Segmentation from Retinal Fundus Images 499
 Nazish Tariq, Shadi Mahmoodi Khaniabadi, Soo Siang Teoh, Shir Li Wang, Theam Foo Ng, Rostam Affendi Hamzah, Zunaina Embong, and Haidi Ibrahim

Near Infrared Remote Sensing of Vegetation Encroachment at Power Transmission Right-of-Way 507
 Pei Yu Lim, David Bong, Kung Chuang Ting, Florence Francis Lothai, Annie Joseph, and Tengku Mohd Afendi Zulcaffle

Evaluation of Three Variants of LBP for Finger Creases Classification 515
 Nur Azma Afiqah Salihin, Imran Riaz, and Ahmad Nazri Ali

R-Peaks and Wavelet-Based Feature Extraction on K-Nearest Neighbor for ECG Arrhythmia Classification 523
 A. M. Khairuddin, K. N. F. Ku Azir, and C. B. M. Rashidi

Ground Truth from Multiple Manually Marked Images to Evaluate Blood Vessel Segmentation 531
 Nazish Tariq, Michael Chi Seng Tang, Haidi Ibrahim, Teoh Soo Siang, Zunaina Embong, Aini Ismafairus Abd Hamid, and Rafidah Zainon

A Comparative Study of Noise Reduction Techniques for Blood Vessels Image 537
 Shadi Mahmoodi Khaniabadi, Haidi Ibrahim, Ilyas Ahmad Huqqani, Harsa Amylia Mat Sakim, and Soo Siang Teoh

Survey on Blood Vessels Contrast Enhancement Algorithms for Digital Image 545
 Shadi Mahmoodi Khaniabadi, Harsa Amylia Mat Sakim, Haidi Ibrahim, Ilyas Ahmad Huqqani, Farzad Mahmoodi Khaniabadi, and Soo Siang Teoh

Face Image Authentication Scheme Based on Cohen–Daubechies–Feauveau Wavelets 553
 Muntadher H. Al-Hadaad, Rasha Thabit, Khamis A. Zidan, and Bee Ee Khoo

A Finger Knuckle Print Classification System Using SVM for Different LBP Variants 565
 Imran Riaz, Ahmad Nazri Ali, and Ilyas Ahmad Huqqani

Assessment of Real-World Fall Detection Solution Developed on Accurate Simulated-Falls 573
 Abdullah Talha Sözer, Tarik Adnan Almohamad, and Zaini Abdul Halim

Deep Learning Based Distance Estimation Method Using SSD and Deep ANN for Autonomous Braking/Steering 581
 Siti Nur Atiqah Halimi, Mohd Azizi Abdul Rahman, Mohd Hatta Mohammed Ariff, Yap Hong Yeu, Nor Aziyatul Izni, Mohd Azman Abas, and Syed Zaini Putra Syed Yusoff

Wood Defect Inspection on Dead Knots and Pinholes Using YOLOv5x Algorithm 589
 Liew Pei Yi, Muhammad Firdaus Akbar, Bakhtiar Affendi Rosdi, Muhamad Faris Che Aminudin, and Mohd 'Akashah Fauthan

Electrical Power, Energy and Industrial Applications

Finite Element Modeling of Electric Field Distribution in a Defective XLPE Cable Insulation Under Different Magnitudes of Stressing Voltage



S. H. Sulaiman, Mohamad N. K. H. Rohani, A. Abdulkarim, A. S. Abubakar, G. S. Shehu, U. Musa, A. A. Mas'ud, N. Rosle, and F. Muhammad-Sukki

Abstract Air voids in solid dielectrics affect the performance and lifespan of high voltage (HV) equipment. In this research, electric field distribution within a cross-linked polyethylene (XLPE) HV cable is analyzed using a finite element analysis (FEA) software, COMSOL Multiphysics. The study was performed in the presence of air cavity of different sizes within the insulation. The average as well as the maximum field strengths for both 2D and 3D of the healthy cable were observed to be equal under five (5) stressing voltage levels. The local field for 1 mm cavity radius in 3D was however lower than that of 2D model with an approximate percentage decrease of 9% for all the applied voltages. Further investigations on the 3D model show that average field rises with voltage and slightly decreases with increasing cavity size, while field enhancement is affected more by the cavity size than voltage stress.

Keywords Finite element analysis · XLPE · Electric field · Cable · COMSOL

S. H. Sulaiman (✉) · A. Abdulkarim · A. S. Abubakar · G. S. Shehu · U. Musa
Department of Electrical Engineering, Ahmadu Bello University, Zaria, Nigeria
e-mail: shsulaiman@abu.edu.ng

M. N. K. H. Rohani · N. Rosle
Faculty of Electrical Engineering and Technology, University Malaysia Perlis, Arau, Malaysia

A. A. Mas'ud
Department of Electrical Engineering, Jubail Industrial College, Jubail, KSA, Saudi Arabia

F. Muhammad-Sukki
School of Computing and the Built Environment, Edinburgh Napier University, Edinburgh, Scotland, UK

1 Introduction

In order to provide an uninterruptible power supply to customers, ensuring the reliability of power system networks becomes a necessity. High voltage (HV) installations like power cables and transformers are integral parts of modern power networks, and proper monitoring of their conditions defines the overall efficiency and reliability of the system [1]. Operation of HV equipment relies on quality of insulation [2] which must cope with the varying operating stresses to avoid fast deterioration of the insulation systems and ensure satisfactory operation.

Insulation deterioration and breakdown due to factors like partial discharge (PD) and water treeing are major concerns in HV equipment [1, 3]. Air voids, protrusions or cracks in or/and on the dielectric serve as PD initiation regions [4]. A sustained PD may lead to total breakdown [5]. PD measurement is therefore employed for diagnostics in HV systems. Accurate modeling and estimation of the field strength and distribution within PD initiating sources aids the understanding of HV installations behavior of [6].

Most researches have utilized simple 2D models to study the impact of electric field in voids of different geometries and sizes [7–9]. Although appreciable results were obtained from these works, the 2D model in the literature is still inadequate as it does not depict a practical cable. In this paper, electric field distribution within an XLPE cable is analyzed in 2D and 3D models considering two scenarios; Case 1: healthy cable and Case 2: with a single spherical void. The field distribution as a function of void radius and applied voltage was investigated, and the two models were compared. Capabilities of COMSOL were utilized in the development and simulation of the model. Further analyses were then carried out on the 3D to examine the relationship between voltage magnitude, cavity size and field strength.

2 Materials and Methods

The material properties of the cable model used in this work are adopted from [7] as presented in Table 1, while its geometrical specifications are given in Table 2.

Under applied voltage, U_0 , the void is exposed to a local field, E_0 , given by [8]:

Table 1 Cable material properties

Material	Conductivity (S/m)	Relative permittivity
Copper	5.85×10^7	1
Aluminum	3.57×10^7	2.2
XLPE	1.0×10^{-15}	2.3
Graphite	3.0×10^{-3}	500
PVC	1.0×10^{-15}	2.9
Air	1.0×10^{-100}	1

Table 2 Geometry of the XPLE cable model

Layer	Component	Material	Value (mm)
1	Conductor radius	Copper	9.2
2	Inner sheath thickness	Graphite	1.8
3	Insulation thickness	XLPE	7.6
4	Insulation screen	Graphite	2.5
5	Earthing screen thickness	Copper	0.8
6	Bedding	PVC	1.3
7	Aarmor wire thickness	Aluminum	1.3
8	Outer sheath thickness	PVC	1.3

$$E_0 = -\nabla U_0 \quad (1)$$

Involving the free charge, the electric field displacement is represented as [3]:

$$\nabla \cdot D = \rho_j \quad (2)$$

where ρ_j is free charge density and $D = \varepsilon_0 E$. The material property, ε , and the charge density, ρ , are related by the Poisson's equation [4]:

$$\nabla^2 U_0 = -\frac{\rho}{\varepsilon} \quad (3)$$

Equation 3 can be solved through finite element approach.

Insulation charge density may be neglected. For voids with extremely small size, void charge density is negligible [4, 5]. In that case, Eq. 3 reduces to [6]:

$$\nabla^2 U_0 = 0 \quad (4)$$

The sinusoidal voltage U_0 is given as [6]:

$$U_0 = U_m \sin 2\pi f t \quad (5)$$

where f is the frequency and U_m is the peak value of the applied voltage. The boundary condition between two media is given by [4]:

$$n \cdot (\vec{D}_1 - \vec{D}_2) = \rho_s \quad (6)$$

where, ρ_s is surface charge, while $n \cdot \vec{D}_1$ and $n \cdot \vec{D}_2$ are normal components of electric displacement of any two different mediums.

2.1 Field (Electric and Potential) Equations

Neglecting the surface charges of the insulation material, Eq. (6) becomes [4]:

$$n \cdot (\vec{D}_1 - \vec{D}_2) = 0 \quad (7)$$

Equation 8 describes the field enhancement factor, η , [6].

$$\eta = \frac{E_0 - E_1}{E_1} \times 100\% \quad (8)$$

E_0 is the average field at the center of the void, E_1 is the field in a healthy cable at the same coordinate with E_0 . The center of the void in this work is at $x = 13.7$ mm, $y = 0.0$ mm for 2D and $x = 13.7$ mm, $y = 0.0$ mm, $z = 25$ mm for 3D model.

2.2 Cable Model

The model geometry in [7] was used for 2D for field strengths evaluation along a cutline that dissects the insulation through the void diameter and emanates from the cable center at P_1 ($x_1 = 0.0$ mm, $y_1 = 0.0$ mm) to the outer sheath at P_2 ($x_2 = 25.8$ mm, $y_2 = 0.0$ mm). A 3D model with the same properties was developed and meshed as shown in Fig. 1a and b respectively. The 3D cutline starts at P_1 ($x_1 = 0.0$ mm, $y_1 = 0.0$ mm, $z_1 = 25$ mm) to P_2 ($x_2 = 25.8$ mm, $y_2 = 0.0$ mm, $z_2 = 25$ mm) along work plane shown in Fig. 1c. The radius of the void is 1.0 mm for both 3D and 2D results.

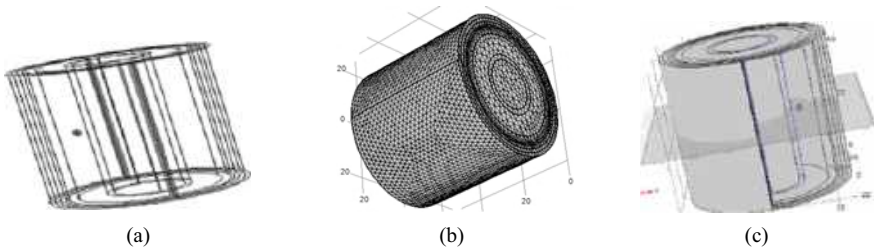


Fig. 1 3D cable model **a** wireframe showing void **b** meshed **c** work plane

3 Results

3.1 Electric Field Distributions in 2D and 3D Models

Field distribution for healthy cable in both 2D and 3D models are shown in Fig. 2a and b respectively. The simulation was performed under 18 kV, 50 Hz AC supply using a time dependent study, and the results were recorded at 0.005 s. Maximum field around the HV electrode is 3.12 kV/mm in both cases. The model was then simulated with a void of 1 mm radius, and field behavior was observed. Maximum field in 2D and 3D were 3.65 kV/mm and 3.32 kV/mm respectively as shown in Fig. 2c and d.

Field distribution along the cutlines is shown in Fig. 3a, b and c. Healthy cable in both 2D and 3D is shown in (a), while (b) and (c) show defective cable in 2D and 3D respectively. In the case of defective cable, maximum field occurs inside the void at the point closest to the HV electrode. The whole void has higher field than the insulation.

Figure 4 shows a comparison of the maximum fields in the 2D and 3D as a function of voltage for a 1 mm void radius. In a healthy cable, the two models produced similar

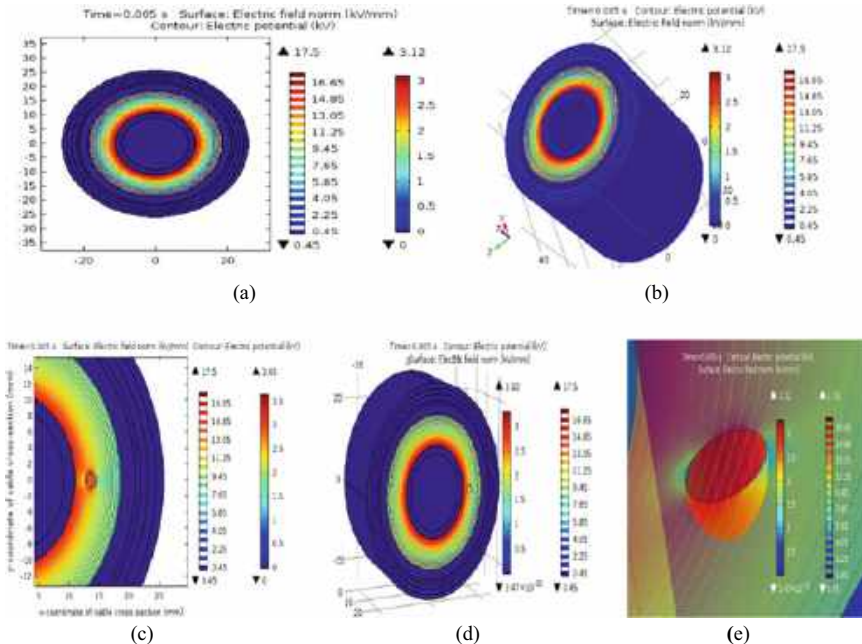


Fig. 2 Electric field and potential distributions **a** 2D healthy **b** healthy 3D **c** 2D defective **d** 3D defective **e** 3D closeup view along work plane

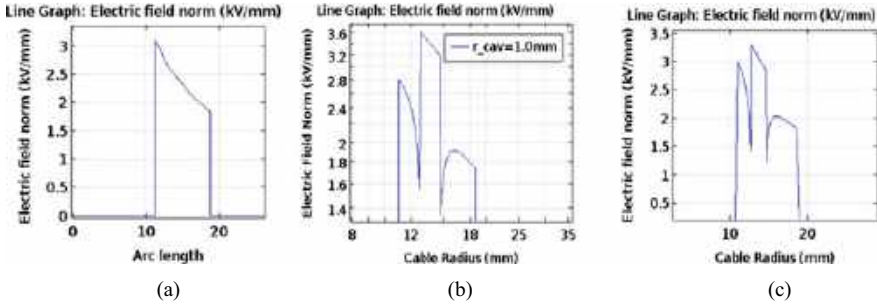


Fig. 3 Field distribution along cutlines a healthy 2D/3D b 2D defective c 3D defective

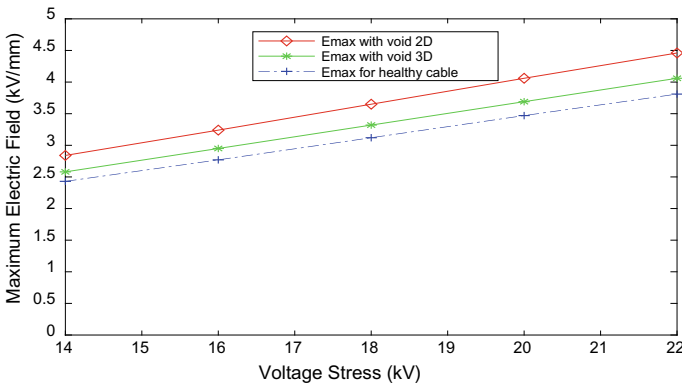


Fig. 4 Maximum field comparison between cable models

results regardless of the voltage. However, for a defective cable, the 2D model has higher field magnitudes than the 3D by around 9% under all voltages.

3.2 Effect of Void Size on Local Field Under Different Voltages

Effect of void size on the local field magnitudes under different values of the stress voltage and fixed void center is examined on the 3D model. Average fields were recorded at the void centers for cavity radii of 0.25, 0.5, 1.0 and 1.5 mm.

Figure 5a shows variation of average fields with void size at different voltages. In all cases, the field significantly increases with increase in voltage and slightly with decrease in void sizes. In Fig. 5b, smaller voids have more impact on field enhancement.

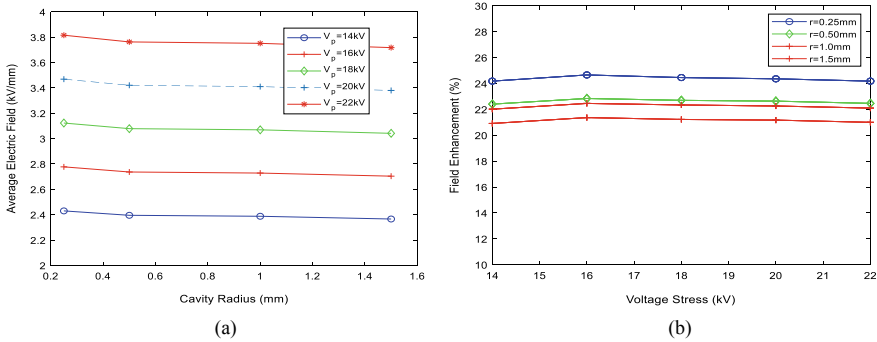


Fig. 5 Impact of void radius and voltage on a average field b field enhancement

4 Conclusion

In this paper, 3D and 2D models of an XLPE cable were developed for electric field analysis. This research has established the closeness in field results between 2 and 3D models of healthy cable. However, significant difference was observed when a defect exists within the insulation bulk. For a cavity of 1 mm radius, maximum field strengths in 3D model were always less than those in 2D model by about 9% for all stress voltages considered in this work. Finally, average field was observed to rise with voltage stress but decreases slightly with increase in void sizes, while field enhancement is more affected by cavity size than voltage stress.

Acknowledgements This work is supported by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS/1/2019/TK04/UNIMAP/03/8)

References

1. Negm TS, Refaey M, Hossam-Eldin AA (2017) Modeling and simulation of internal partial discharges in solid dielectrics under variable applied frequencies. In: 18th International middle east power systems conference MEPCON, pp 639–644
2. Mas’ud AA, Musa U, Sulaiman SH, Mahmud I, Shehu IA, Aliyu BD (2022) Partial discharge detection and classification : implications on the power industry. In: 18th International conference and exhibition on power and telecommunications (ICEPT), pp 146–152
3. Hore S, Basak S, Haque N, Dalai S, Mukherjee M (2018) Studies on the effect of void geometry and location on electric field distribution and partial discharge in XLPE insulated power cable by finite element analysis using COMSOL multiphysics simulation. In: 6th International conference on computer applications in electrical engineering-recent advances (CERA), pp 220–225
4. Joseph J (2018) Classification of PD sources in XLPE cables by artificial neural networks and support vector machine. In: IEEE electrical insulation conference (EIC), pp 407–411
5. Musa U, Mati AA, Mas’ ud AA, Shehu GS, Albarraçin-Sanchez R, Rodriguez-Serna JM (2021) Modeling and analysis of electric field variation across insulation system of a MV power cable.

- In: International conference on electrical communication & computer engineering (ICECCE), pp 1–5
6. He M, Chen G, Lewin PL (2016) Field distortion by a single cavity in HVDC XLPE cable under steady state. *Inst Eng Technol* 1(3):107–114
 7. Emna K, Rabah A, Nejib C (2016) Numerical modeling of the electric field and the potential distributions in heterogeneous cavities inside XLPE power cable insulation. *J Electr Electron Eng* 9(2):37–42
 8. Musa U, Mati AA, Mas'ud AA, Shehu GS (2021) Finite element modeling of fields distributions in a three-core XLPE cable with multiple cavities. In: 2021 IEEE PES/IAS power Africa. IEEE, pp 1–4
 9. Alsharif M, Wallace PA, Hepburn DM, Zhou C (2012) FEM modeling of electric field and potential distributions of MV XLPE cables containing void defect. In: Excerpt from the proceedings of the COMSOL conference in Milan, pp 1–4

Optimal Switching Sequence of Urban Power System Based on Dynamic Thermal Rating Parameter Adjustment



Yi Su  and Jiashen Teh 

Abstract With the emergence of flexible topology technology in UPSs, the number of breaker action has increased, making the sequence more important. In addition, during the switching process, In addition, short-term violations may occur, such as line overloading, which can necessitate additional corrective operations to maintain the stability and reliability of UPS. In response, this article proposes the use of dynamic thermal rating (DTR) technology to replace the existing overload judgment conditions and accurately evaluate the risk of the changing process. To improve the accuracy of the key parameters in DTR model, this paper also proposes a real-time feedback correction framework. Finally, a comprehensive evaluation model is constructed that takes into account node voltage to assess the operation sequence of dynamic line overload, which provides assistance for the switching of the power grid state. Simulation results show that the proposed feedback correction scheme improves the accuracy of the DTR model, and the comprehensive evaluation model effectively solves the optimal sequence problem of long-time, multi-switch stepping operations, improving system stability.

Keywords Urban power system · Switching sequence · Dynamic thermal rating

1 Introduction

The trend of optimizing operation through large-scale topology switching has become a current hot spot in response to the rapid development of urban power grids [1]. However, the application of flexible topology technology is constrained by capacity, and may cause short-term congestion, which need more operations [2]. To effectively solve this problem, capacity expansion of overhead conductors can be utilized.

Y. Su · J. Teh (✉)

School of Electrical and Electronic Engineering, Universiti Sains Malaysia (USM), Ndong Tebal, 14300 Penang, Malaysia

e-mail: jiashenteh@usm.my

Using the real-time monitoring of conductor operations and the use of the DTR model to calculate current carrying capacity, enabling delicate management of transmission capacity and dynamic capacity increase of overhead lines under unfavorable weather conditions [3]. As meteorological information collection improves and costs decrease, it is expected that dynamic capacity increase technology will be more widely applied in engineering, with increased research focus on improving model accuracy.

In addition, large-scale topology switching needs to optimize the switching sequence to ensure the minimum impact on the power grid during the switching process [1]. Therefore, we build a dynamic capacity expansion model based on DTR, and use the feedback adjustment to adjust two important parameters of DTR. Then, build a comprehensive evaluation model to choose the optimal switching sequence. The simulation shows the proposed method is useful and meet engineering needs.

2 Dynamic Capacity Expansion Using Dynamic Thermal Rating

In this chapter, the heat balance equation of the IEEE std 738-2012 [4] standard is introduced and the various parameters and variables involved are analyzed. It is pointed out that the solar radiation absorption coefficient and conductor heat dissipation coefficient, which are easily overlooked, have a significant impact on the calculation of the current carrying capacity. Propose a quasi-real-time micro-meteorological data to predict the real-time ampacity to realize the dynamic capacity expansion of overhead lines.

2.1 Parameters Adjustment of Dynamic Thermal Rating Model

The capacity of overhead lines can be expressed as follows [4].

$$I(t) = \sqrt{\left\{ Q_c(t) + Q_r(t) + mC_p \cdot \frac{dT_s(t)}{dt} - Q_s(t) \right\} / R[T_s(t)]} \quad (1)$$

$$R[T_s(t)] = \left[\frac{R(T_{high}) - R(T_{low})}{T_{high} - T_{low}} \right] \cdot (T_s - T_{low}) + R(T_{low}) \quad (2)$$

$$Q_c(t) = f(K_{angle}, V_w) \cdot [T_s(t) - T_a(t)] \quad (3)$$

$$Q_r(t) = 17.8 \cdot D_0 \cdot \varepsilon \cdot \left[\left(\frac{T_s(t) + 273}{100} \right)^4 - \left(\frac{T_a(t) + 273}{100} \right)^4 \right] \tag{4}$$

$$Q_s(t) = \alpha \cdot Q_{se} \cdot \sin(\theta) \cdot A' \tag{5}$$

where V_w , K_{angle} , $T_a(t)$, $T_s(t)$, Q_{se} , and θ are the input data, represent wind speed at the conductor [m/s], wind direction [deg], ambient air temperature at time t [°C], conductor temperature at time t [°C], intensity of solar radiation [W/m^2], and sun incidence angle, respectively. The output is the overhead line capacity [A], i.e., $I(t)$. The intermediate variables include: $Q_c(t)$, $Q_r(t)$, and $Q_s(t)$, which represent overhead line convective exchange heat [W/m], radiation heat dissipation [W/m], solar radiation absorption heat [W/m] respectively. The constant parameters include: D_0 , $R(T_{high})$, $R(T_{low})$, and mC_p , representing outside diameter of conductor [m], AC resistance corresponding resistance at high temperature and low temperature [Ω], and total heat capacity of conductor [$J/(m \cdot ^\circ C)$].

The constant parameters can be found in the overhead line manufacturer’s instructions; the input data can be found in the weather system or DTR system. But there are two parameters, i.e., α and ε is a constant [4] in the range of [0.15, 0.95], which is closely related to the reality of conductor smoothness and aging.

The α and ε are usually determined by empirical values and are not valued. In fact, they are closely related to the accuracy of the model. * MERGEFORMAT Fig. 1. shows the influence of α and ε changing to current under the same meteorological conditions.

It can be seen from * MERGEFORMAT Fig. 1 that compared with 0.8 for α and ε , the change of parameters makes the current carrying capacity of the conductor fluctuate between [- 20.3%, 12.2%], which cannot be ignored. Therefore, this paper

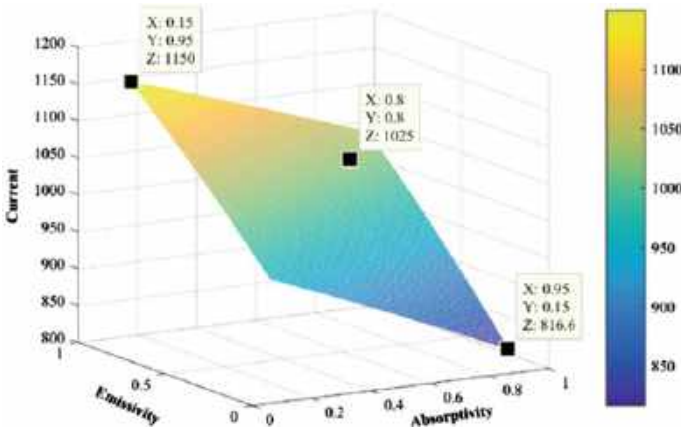
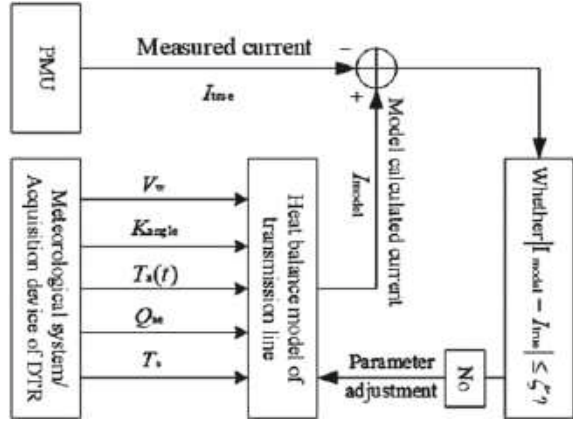


Fig. 1 Solar radiation absorption coefficient and conductor heat dissipation coefficient on the current effect (IEEE Std 738-2012 Standard calculation case conditions)

Fig. 2 Schematic diagram of the calibration process of solar radiation absorption coefficient and conductor heat dissipation coefficient



constructs α and ε adjustment systems to further improve the accuracy of DTR model. The process is shown in * MERGEFORMAT Fig. 2.

Considering that there are two parameters to be determined in the parameter adjustment process, and the iterative convergence method needs to clarify the convergence direction, the adjustment rules are formulated as follows: (a) For high temperature conductors (conductor temperature greater than 150°C), ε has a great influence on the model calculation current. That is, by maintaining α constant, the adjustment ε causes the model calculating the current I_{model} to approximate the actual current I_{true} . (b) For low-temperature conductors (temperature less than 75°C), α affects the model calculation current greatly. That is, by maintaining ε constant, the adjustment α causes the model calculating the current I_{model} to approximate the actual current I_{true} . It should be pointed out that the real-time change rate of α and ε is small, so the parameters can be used as the fixed value of model parameters for a long time after adjustment until they exceed the threshold. In this way a more accurate model is proposed.

2.2 Line Capacity Constraint Relaxation Based on Dynamic Thermal Rating

The capacity of transmission line is mainly affected by external micro meteorological conditions. Thus, use the DTR technology to dynamically calculate the capacity, to ensure no heavy load tripping during the relaxation as follows.

$$-I_l^{DRT} \leq I_l \leq I_l^{DRT} \quad (6)$$

The short-time-scale micro-meteorological data can be regarded as unchanged [5], and use it as an input to calculate the quasi-dynamic capacity. Namely, the α and

ε are adjustment at day-ahead dispatch and the dynamic capacity are calculated at intra-day dispatch, using the input data at time $t - 1$ to forecast the capacity at time t .

3 Optimal Switching Sequence Selection Model

This section built a comprehensive model considering the overvoltage and overload to choose the optimal switching sequence. Assuming the transition from \mathcal{A} to \mathcal{B} requires N number of switching, the model of the switching sequence is as follows.

$$\Theta = \min \sum_{\mathcal{R}=1}^N \left[\frac{\tau_1}{H} \left(\sum_{a=1}^H f_U^{\mathcal{R}}(a) \right) + \frac{\tau_2}{L} \left(\sum_{l=1}^L f_I^{\mathcal{R}}(l) \right) \right] \quad (9)$$

$$f_U^{\mathcal{R}}(a) = \begin{cases} \frac{|u_a^{\mathcal{R}} - 1|}{0.07}, & 0.93 \leq u_a^{\mathcal{R}} \leq 1.07 \\ 1, & \text{others} \end{cases} \quad (10)$$

$$f_I^{\mathcal{R}}(l) = \begin{cases} \frac{I_l^{\mathcal{R}}}{I_{max,l}}, & I_l^{\mathcal{R}} \leq I_{max,l} \\ 1, & I_l^{\mathcal{R}} \geq I_{max,l} \end{cases} \quad (11)$$

where the $f_U^{\mathcal{R}}(a)$ is index of nodal voltage deviation from the ideal rated voltage (1 p.u.); $f_I^{\mathcal{R}}(l)$ is the index of line loading deviation from the maximum ratings of the line. τ_1 and τ_2 are the weightages that indicate the emphasis level of the objective functions, which are set based on the priority of the user.

4 Results and Discussion

A modified 56-node test system is used here as * MERGEFORMAT Fig. 3. shows, which is formed by 1 IEEE 14-node transmission network [6] and 3 IEEE 14-node distribution networks [7]. The parameter adjustment of DTR is based on the feedback conditions of the actual system. This article performs feedback adjustment on several DTR application lines in Guangdong, China. The results are as shown in * MERGEFORMAT Table 1.

Table 1 shows that the value of a line operating for 1 year is below to 0.3, while lines operating over 5 years is over 0.7. For lines operating over 10 years is about 0.9.

Then, this paper uses the $\alpha = 0.73$, $\varepsilon = 0.82$ as the line parameters of the 56-node test system. The meteorological parameters are taken the data of Haiyang, Shandong, China in 2019. Assume that the changes between \mathcal{A} and \mathcal{B} in the test system are: $S_3 - 114$, $104-113$, $S_6 - 212$, $S_{12} - 201$, $201-202$, $201-204$, $S_9 - 301$, $301-302$,

Fig.3 Schematic diagram of the calibration process of solar radiation absorption coefficient and conductor heat dissipation coefficient

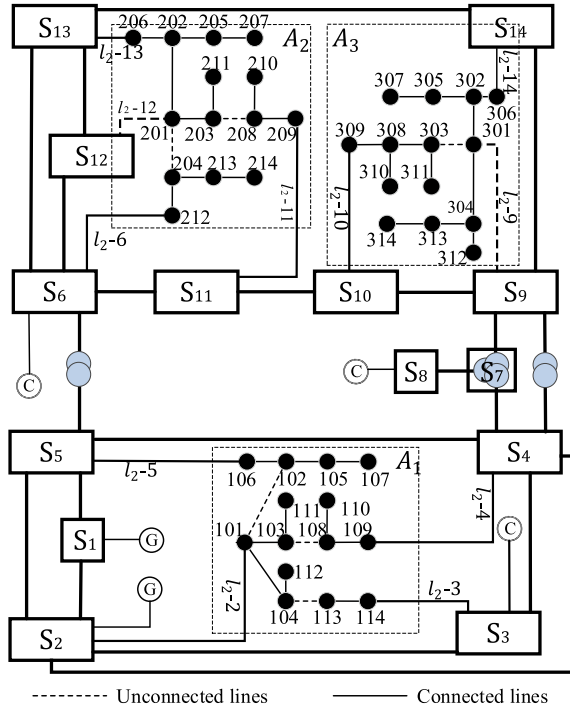


Table 1 Parameters adjustment of DTR based on real-time current feedback

Operation years	α	ε	ξ
1 year	0.23	0.31	0.042
5 years	0.73	0.82	0.047
10 years	0.86	0.92	0.041

303–308, 301–303. * MERGEFORMAT Fig. 4. shows the current changes in the transition process.

As * MERGEFORMAT Fig. 4 shows, if the dispatcher closes branch 201–204, line $S_6 - 212$ grows to 240.1 A, which is under the tolerable current based on relaxation coefficient. But it may increase the line failure probability greatly when in 5:00–10:00, because of influence of weather on the line (tolerable current calculated by DTR is only about 170–200 A). And it increases to 249 A at 18:00. Therefore, quantifying relaxation conditions based on DTR rather than relaxation coefficient is useful to dispatcher control operational risk. The current of line $S_3 - 114$ is about 290 A until the branch 301–303 closed. So, closing that branch early or not affects the stability of power grid.

So using the switching sequence model proposed here can well evaluate the risks in the process of flexible topology and improve system stability.

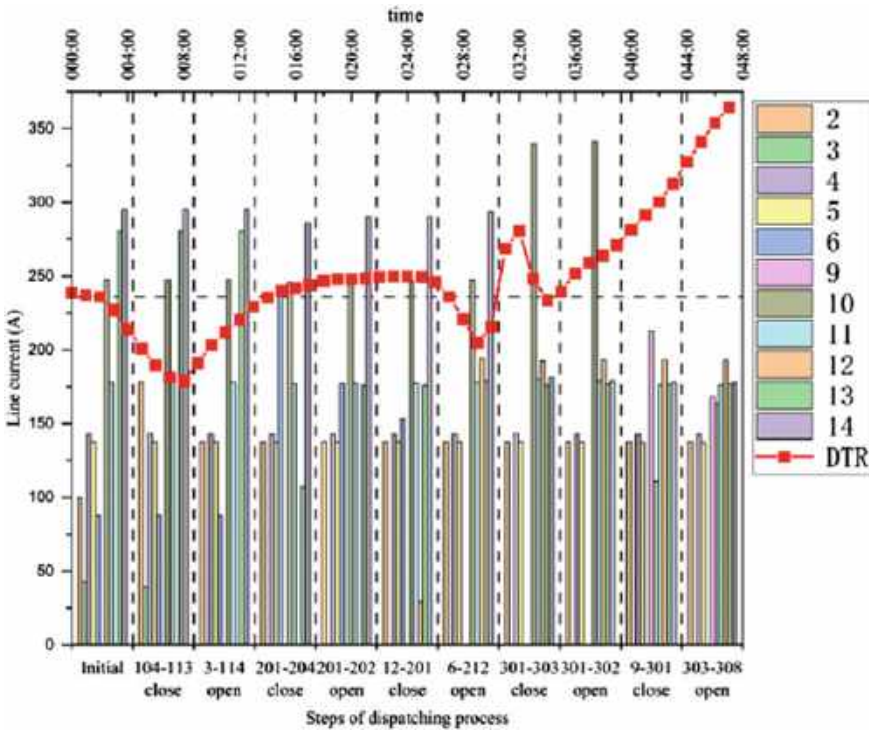


Fig.4 Current changes in the transition process

5 Conclusion

This paper proposes a feedback framework to enhance the accuracy of the dynamic thermal rating (DTR) model and utilizes DTR technology for risk assessment of line overload in urban power grids. By constructing an optimal sequence for the operation process, the paper achieves optimal control of large-scale operations in urban power grids. The experimental results demonstrate that the method described in the article improves the accuracy of the DTR model, especially for lines with different operating durations. Furthermore, the optimal sequence model effectively addresses stability concerns in multi-step continuous operation processes.

References

1. Jimada-Ojuolape B, Teh J (2020) Surveys on the reliability impacts of power system cyber-physical layers. *Sustain Cities Soc* 62:102384
2. Su Y, Teh J (2023) Two-stage optimal dispatching of AC/DC hybrid active distribution systems considering network flexibility. *J Modern Power Syst Clean Energy* 11(1):52–65

3. Lawal OA, Teh J (2023) Dynamic line rating forecasting algorithm for a secure power system network. *Expert Syst Appl* 219:119635
4. Lai CM, Teh J (2022) Network topology optimisation based on dynamic thermal rating and battery storage systems for improved wind penetration and reliability. *Appl Energy* 305:117837
5. Cherukupalli S, Adapa R, Bascom EC (2018) Implementation of quasi-real-time rating software to monitor 525 kV cable systems. *IEEE Trans Power Deliv* 34(4):1309–1316
6. Teh J, Lai CM, Cheng YH (2017) Impact of the real-time thermal loading on the bulk electric system reliability. *IEEE Trans Reliab* 66(4):1110–1119
7. Teh J, Lai CM (2019) Reliability impacts of the dynamic thermal rating and battery energy storage systems on wind-integrated power networks. *Sustain Energy, Grids Netw* 20:100268

Improved DC-Link Voltage Control Scheme for Standalone PV Renewable Energy System



Muhammad Najwan Hamidi

Abstract In this paper, an improved DC-Link voltage control algorithm is proposed for the application on a standalone photovoltaic (PV) renewable energy system. The algorithm is based on the perturb & observed based voltage regulator (POVR) algorithm. By deploying the proposed algorithm with a buck-boost converter together with the inclusion of dynamic step size, ΔD changing technique, the suggested algorithm managed to have a very good balance between tracking speed and output voltage ripple. This is proven through the conducted simulation where in overall, the proposed algorithm produced the highest and lowest output voltage ripple of 0.1% and 0.3%, respectively. In terms of tracking time, the highest and lowest recorded times are 0.05 s and 0.9 s, respectively. The tracking times are only marginally higher than the POVR with large ΔD and much lower than the POVR with small ΔD . On the other hand, the output voltage ripples using the proposed algorithm are found to be considerably lower than the POVR with large ΔD and only slightly larger than the POVR with small ΔD .

Keywords Standalone · PV system · DC-Link Control

1 Introduction

A standalone PV system typically includes a PV source, DC-DC converters, and optionally an inverter and loads. It is an off-grid system that can independently power standalone loads, making it a cost-effective solution, especially in remote locations with geographical challenges. Additionally, the continuous costs for electricity and maintenance requirements are typically lower [1].

A standalone PV system, with or without a battery storage system, operates differently from a grid-connected system. The load determines the power output, and the

M. N. Hamidi (✉)

School of Electrical & Electronic Engineering, Universiti Sains Malaysia, 14300 Nibong Tebal, Malaysia

e-mail: najwan@usm.my

system must meet the demand, even at its highest. The DC–DC converter is responsible for providing accurate voltage levels to the loads, which have specific voltage requirements. A standard pulse-width modulation (PWM) controller can adjust the duty cycle, D to achieve the desired voltage, but it has a narrow regulating range since it can only regulate based on the instantaneous PV operating point. To address this limitation, a hybrid algorithm called POVR is introduced, which has active power tracking and a wider regulating range [2].

The POVR, despite its ability to regulate DC-Link voltage over a wide range, has drawbacks. Steady-state oscillation is one of these drawbacks, caused by the P&O MPPT algorithm on which the POVR is based [3]. This oscillation is worse with POVR because the PV operating point falls between 0 V and V_{mpp} , where the PV voltage is lower. Achieving the reference voltage, V_{ref} , requires a higher D , resulting in higher ripples due to the boost converter relationships [4].

This paper proposes an improved DC-Link voltage control technique for a standalone PV renewable energy system to overcome the aforementioned drawbacks. A buck-boost converter is used with POVR instead of a boost converter, which can operate the PV panel in the higher voltage region, reducing the required D to achieve target voltage levels. The system is simulated and demonstrated using MATLAB-Simulink software.

2 System Modelling

2.1 Overview

The suggested standalone PV renewable energy system consists of a PV module, a standard buck-boost DC–DC converter and a standalone DC load as in Fig. 1. The proposed algorithm is implemented to control the duty cycle of the switch.

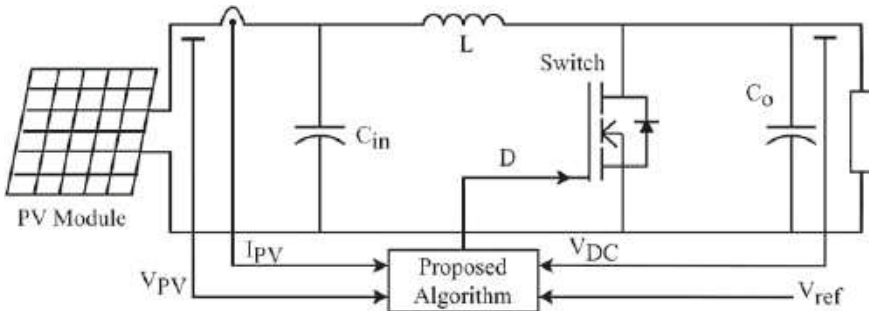


Fig. 1 The implemented boost DC–DC converter

2.2 The Proposed Control Algorithm

The proposed control scheme in this paper is based on the POVR algorithm with some modifications. The original POVR is designed for a boost DC–DC converter, with the PV operating point related to the step-size, ΔD used. Large ΔD places the operating point between 0 V and V_{mpp} , while small ΔD places it between V_{mpp} and the open circuit voltage, V_{OC} , which is more desirable as it has lower PV current (I_{PV}). Trade-offs exist between ΔD , response time, and output voltage ripple, with large ΔD having faster response time but higher ripple voltage, and vice versa for smaller ΔD .

The proposed algorithm in this paper is designed for a buck-boost converter and incorporates a dynamic ΔD changing technique to minimize steady-state oscillation while maintaining fast response time. Once V_{ref} is achieved, ΔD is reduced using a simple technique. Figure 2 shows the flowchart of the control algorithm.

As can be seen, the ΔD is set at a larger value in the beginning, and the value is reduced once the output voltage falls within 99 and 100.03% of V_{ref} . As such, the tracking time will be shortened due to the larger initial ΔD , and the oscillation can be reduced due to the reduced ΔD after V_{ref} is achieved. Both quick-response time and low output voltage ripple can be achieved. Additionally, from the deployment of the buck-boost converter, the range of D can theoretically be anywhere between 0 and 1 as compared to POVR operation with a boost converter.

Technically, the proposed algorithm operates by using the standard P&O MPPT concept. But, instead of tracking for maximum power point (MPP), the algorithm tracks for reference power, P_{ref} . P_{ref} is directly related to V_{ref} according to the

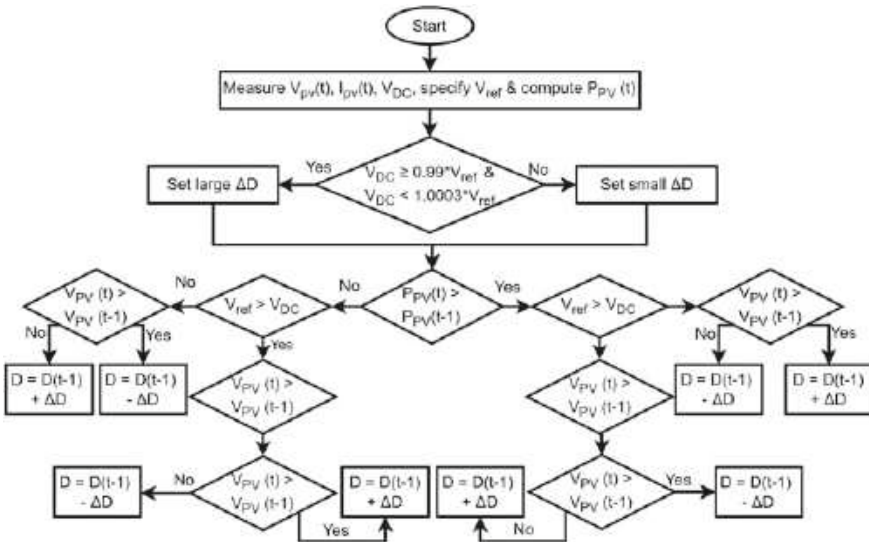


Fig. 2 Flowchart of the proposed control algorithm

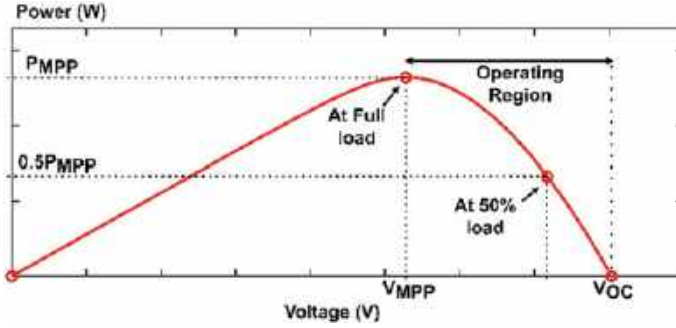


Fig. 3 Operating region depiction of the proposed algorithm

following relationship. Once P_{ref} is tracked, D will be adjusted based on the instantaneous PV voltage, V_{PV} , to produce the desired output voltage, V_{DC} . The operation can be simplified as shown in Fig. 3. With such an operation, the algorithm can also be deployed as a standard MPPT algorithm by setting the V_{ref} to a very large value.

The proposed algorithm is advantageous compared to the normal POVR since its operating region is in between V_{MPP} and V_{oc} . In this region, V_{PV} is higher and I_{PV} is lower. With higher V_{PV} , lower D is required to produce the desired output voltage. This, in turn will reduce ripple.

3 Results and Discussion

Simulation works are conducted using MATLAB/Simulink software to demonstrate the operation and performance of the proposed DC-Link voltage control scheme in a standalone PV renewable energy system. The tests conducted aimed to analyse the characteristics of V_{DC} at different V_{ref} . Comparative analysis is also done to compare the V_{DC} obtained using the proposed algorithm and the original POVR at different ΔD . Throughout the tests, the load, irradiance and temperature are set at 100Ω , 1000 W/m^2 and $25 \text{ }^\circ\text{C}$, respectively. Figure 4 shows the V_{DC} obtained using POVR and the proposed algorithm at V_{ref} of 40 V. It can be seen that V_{DC} reached the V_{ref} point in about 0.09 s. The average value of V_{DC} measured is 40.04 V, with the ripple voltage at only 0.3%. Using POVR, the tracking time is marginally lower which is at 0.06 s with ΔD of 0.1. However, the average V_{DC} value obtained is 59.94 V, and the voltage ripple is at 0.5%. If a smaller ΔD of $1e-6$ is used instead, the POVR produces an average V_{DC} value of 60.00 V with a slightly lower ripple voltage of 0.2% compared to the proposed algorithm. However, the tracking time is much slower, which is at 0.64 s. The full output characteristics analysis of V_{DC} at the tested V_{ref} and several other V_{ref} are presented in Table 1. In brief, the proposed algorithm has the best balance between tracking time and output ripple where the tracking times measured are only marginally higher than the tracking time obtained using POVR

with large ΔD and much lower than the POVR with small ΔD . The output ripples measured are also only slightly higher than the ripple observed when deploying POVR with small ΔD . and much higher than POVR with large ΔD . In terms of output voltage accuracy, all methods can be considered good. It can be concluded that the proposed technique reduced output ripple by 55% on average compared to POVR with large ΔD . It also increases tracking time by 57% on average compared to POVRs with small ΔD .

When V_{ref} is set at a very high value, both the proposed and the POVR algorithms will operate as a normal maximum power point tracking (MPPT) algorithm. Under this scenario, all techniques produced almost similar results since the operations are at MPP. So, there is theoretically no difference whether the tracking is from the left or right side of the power-voltage (P-V) curve. From Table 1 as well, it can be seen that when using POVR with small ΔD , the algorithm will fail to track for V_{ref} since the PV operating region has a voltage value higher than V_{ref} . Thus, a boost converter will not be able to track for the targeted voltage.

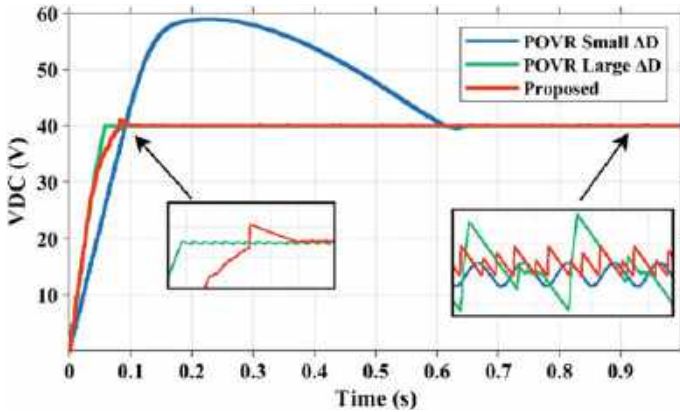


Fig. 4 V_{DC} measurements using different techniques at $V_{ref} = 40$ V

Table 1 Output characteristics using different techniques

V_{ref} (V)	V_{DC} (V)			V_{DC} ripple (%)			Tracking time (s)			Duty cycle, D		
	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3
20	20.03	19.99	-	0.30	0.40	-	0.05	0.03	-	0.33	0.93	-
40	40.04	39.98	40.00	0.30	0.90	0.20	0.09	0.06	0.67	0.51	0.90	0.02
60	60.04	59.94	60.00	0.20	0.50	0.02	0.24	0.14	1.4	0.61	0.87	0.37
80	80.05	80.04	80.00	0.20	0.50	0.10	0.35	0.25	0.80	0.69	0.83	0.55
100	100.0	99.83	100.0	0.10	0.60	0.10	0.90	0.50	2.6	0.76	0.79	0.68
MPP	111.6	112.8	111.3	0.02	0.03	0.02	2.2	2.2	2.1	0.81	0.76	0.76

Note A1 refers to the proposed algorithm, A2 refers to the POVR algorithm using large ΔD , and A3 refers to the POVR algorithm using small ΔD

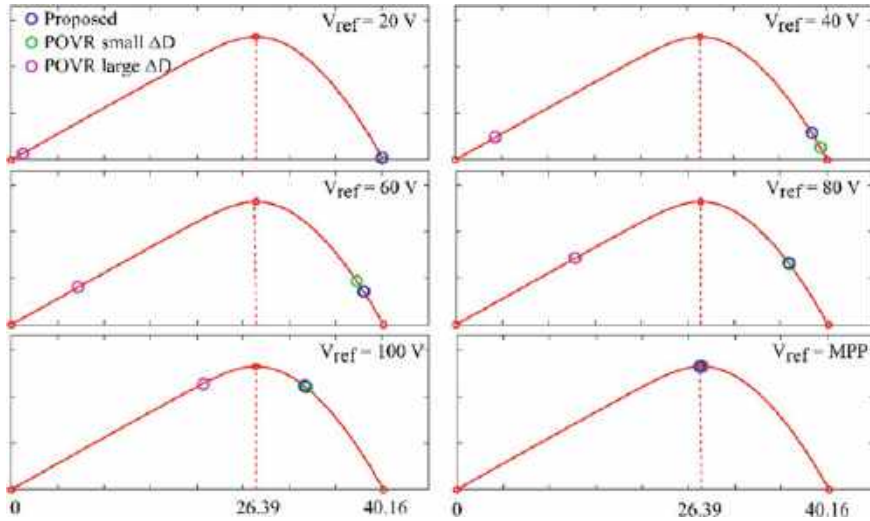


Fig. 5 PV operating points using different techniques

Figure 5 shows input characteristics on a P–V curve. The proposed algorithm and POVR with small ΔD have a power tracking region from V_{OC} of 40.16 V to V_{MPP} at 26.39 V. POVR with large ΔD tracks from 0 V to V_{MPP} . The proposed algorithm uses a buck-boost converter, allowing for a higher tracking range. At V_{ref} of 20 V, the proposed algorithm can still step down PV voltage with buck operation, which POVR with a boost converter cannot achieve.

The duty cycle is an important factor to consider. Table 1 shows that the proposed algorithm with a buck-boost converter can operate with a theoretical D range of 0–1. POVR can also achieve this with a small ΔD . However, if a boost converter is used with POVR, it will not be able to track V_{ref} if it is set below V_{MPP} . Using a large ΔD with POVR produces duty cycle values that are all greater than 0.7 and inversely proportional to V_{ref} , which results in higher voltage ripple. These findings demonstrate that the proposed algorithm has the widest regulating range.

4 Conclusion

In this research, a new DC-Link voltage control scheme is proposed for standalone PV renewable energy systems. The proposed algorithm has a good balance between tracking speed and output voltage ripple compared to the POVR algorithm. It also has the highest voltage regulating range, operating anywhere between 0 and 1. The proposed algorithm has tracking speeds of 0.05–0.9 s, only slightly higher than POVR with large ΔD and much lower than POVR with small ΔD . Output voltage ripple

using the proposed algorithm ranges from 0.1 to 0.3%, slightly higher than POVR with small ΔD and considerably lower than POVR using large ΔD .

Acknowledgements The research work is financially supported by Universiti Sains Malaysia Short Term Grant 304/PELECT/6315747.

References

1. Bonkile MP, Ramadesigan V (2019) Power management control strategy using physics-based battery models in standalone PV-battery hybrid systems. *J Energy Storage* 23:258–268
2. Hamidi MN, Ishak D, Zainuri MAAM, Ooi CA, Tarmizi T (2021) Asymmetrical Multi-level DC-link inverter for PV energy system with perturb and observe based voltage regulator and capacitor compensator. *J Modern Power Syst Clean Energy* 9:199–209
3. Manna S, Singh DK, Akella AK, Kotb H, AboRas KM, Zawbaa HM, Kamel S (2023) Design and implementation of a new adaptive MPPT controller for solar PV systems. *Energy Rep* 9:1818–1829
4. Marahatta A, Rajbhandari Y, Shrestha A, Phuyal S, Thapa A, Korba P (2022) Model predictive control of DC/DC boost converter with reinforcement learning. *Heliyon* 8:e11416

Wind Energy Distributions for Integration with Dynamic Line Rating in Grid Network Reliability Assessment



Olatunji Ahmed Lawal and Jiashen Teh

Abstract Wind power presents a promising form of sustainable energy readily available with negligible greenhouse gas emissions. However, the variability in wind speed and power presents significant challenges for modelling and forecasting for grid reliability studies. This article examines how wind varies using four statistical distributions: Weibull, gamma, Rayleigh, and lognormal. It systematically reviews the prospects of integrating wind power into power systems with dynamic line rating (DLR). It highlights each distribution's crucial considerations and limitations when selecting a wind distribution for analysis. The study shows that integrating wind power into power systems using DLR can address reliability. This approach increases infrastructure utilization and facilitate the integration of wind energy. Grid operators, and energy stakeholders seeking effective strategies for renewable energy integration that balance reliability and cost-effectiveness will find this study relevant.

1 Introduction

Dynamic line rating (DLR) is a technology that adjusts the ampacity of overhead transmission lines in real-time by considering the changes in physical and environmental conditions. Its significance lies in maintaining optimal power flow, preventing system failure caused by overloading, reducing congestion on power lines, increasing efficiency, enabling the integration of renewable energy sources (RES), and improving cost-effectiveness in power generation and dispatch. DLR data is obtained through direct sensors or weather-based measurements and modelling of atmospheric parameters such as temperature, wind velocity, and solar irradiance, which is essential for accurately assessing the transmission line's thermal capacity.

O. A. Lawal · J. Teh (✉)

School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Nibong Tebal 14300, Pulau Pinang, Malaysia
e-mail: jiashenteh@usm.my

O. A. Lawal

Kwara State Polytechnic, Ilorin, Nigeria

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
N. S. Ahmad et al. (eds.), *Proceedings of the 12th International Conference on Robotics, Vision, Signal Processing and Power Applications*, Lecture Notes in Electrical Engineering 1123, https://doi.org/10.1007/978-981-99-9005-4_4

A recent comprehensive overview [1] highlighted the significance of weather and climate in energy-related problems. It showed that DLR technology is essential in ensuring an efficient power flow and preventing system collapse due to overload.

Additionally, deploying DLRs reduces renewables' curtailment, providing a means to mitigate the challenges of climate change caused by increasing global warming when well forecasted and deployed. It is worth noting that the emission of greenhouse gases can have severe health, safety, and environmental impacts [2]. DLRs play a crucial role in reducing greenhouse gas emissions by integrating renewable energy sources, ensuring efficient and cost-effective power dispatch, and mitigating climate change challenges caused by global warming. By preventing system collapse due to overload, reducing power line congestion, and improving efficiency, DLRs can reduce curtailment of renewables. The growing deployment of renewable energy generation capacities in China, the US, and Europe highlights the increasing dependence on meteorological forecasting for power system operations. Authors in [3–6] researched the correlation between the potential power output of wind farms and the cooling of overhead line conductors. Results confirm a positive correlation between wind generation and line rating.

Consequently, implementing DLR on relevant transmission lines could mitigate wind farm curtailment. The importance of DLR forecasting in short and medium terms has been stressed in [7, 8]. The studies used stochastic and deep learning methods in the forecasting of DLR to be used in system operations and planning. At higher wind speeds, the power generation of wind farms increases. Forecasting DLR for a reliable power system network has also been paramount for researchers. With this, a novel algorithm for DLR forecasting has been proposed in [9]. This produced an accurate forecasting model that will allow the full potential of the line capacity thereby allowing wind integration. Load curtailment cost, generator dispatch cost and wind curtailment costs have also been found to reduce in a study optimizing network topology for improved wind penetration [10].

A study in [11] provided a comprehensive review of DLR technologies, methodologies and equipment, communication and reliability challenges with deployment [12–14] and real-world application, and future DLR implementation approaches. They proposed further research into DLR usage on Power system components such as transformers, circuit breakers, and protective relays. In a bid to present a probabilistic line rating forecasting model in a two-stage stochastic optimisation model in [15], a methodology to optimise reserve holding levels and energy production in the scheduling and re-dispatch actions in real-time operation was imminent. Efficient filtering approaches can handle the computational burden of DLR, balancing the benefits of higher line capacity utilisation against increased holding and reserve services costs to manage forecasting errors. The DLR system supports the cost-effective integration of high wind penetration, and multiple sources of uncertainty related to DLR must be considered to achieve optimal results. The reliability impacts of DLR have also been studied with the integration of wind energy [16] and battery energy storage systems (BESS) [17]. The study showed how the effects of increased generation and storage capacity could be enhanced on the transmission lines.

The prospects of DLR in ensuring power transmission reliability have been surveyed in [18]. The study saw smart grids incorporating ICT to better the traditional grid. It also observed that as much as the power generated does not match load demands, wheeling out power from the generating stations with the weak, aged transmission infrastructure will continue to be challenging in many parts of the world.

Network topology optimisation, DLR and BESS were combined to improve congestion and increase network flexibility [19]. DLR was used in the study to enhance the line rating, while BESS was used to time-shift wind power to avoid spillage. The combination of the three methods reduced wind and load curtailment costs.

Smart grids involve an increased penetration of sensors and ICT in power delivery; the reliability of the ICT components termed the cyber components, is pertinent in developing a complete framework for cyber-physical interdependencies of the power system components. The cyber components involve the integration of data communication gadgets, control, protection and monitoring devices, and their reliabilities as proposed in [20].

2 Wind Energy Distributions

Wind energy is a sustainable and promising renewable energy source, but its unpredictability and variability challenge integration into power systems. Statistical distributions such as Weibull, Rayleigh, Gamma, and Lognormal have been developed to model wind energy output based on wind speed, allowing for effective planning, design, and management of wind energy systems. These distributions enable wind energy integration into power systems while accounting for variability and uncertainty (Table 1).

3 Method

The four major wind distribution models will be simulated, their outputs will be compared, and they will be integrated with a reliability test system to assess the system's reliability.

1. Measure the wind speed in the target area.
2. Calculate the wind power associated with each measured wind speed.
3. Define the parameters for each distribution to be used.
4. Generate wind speed data for each distribution based on the defined parameters.
5. Use the wind turbine power curve to compute the power output for each wind speed.
6. Create a plot displaying the wind speed and power output data.

Table 1 Statistical distributions for wind models

Distribution	Description	Formulae	Merits	Assumptions
Gamma	It is a continuous probability distribution that is often used to model wind speed frequency distributions in areas with complex terrain	$f(x) = \frac{\left(x^{(k-1)} * e^{\left(\frac{-x}{\theta}\right)}\right)}{\left(\theta^k * \Gamma(k)\right)}$ k is the shape parameter, θ is the scale parameter and x is the wind speed	It can capture the influence of terrain and other local effects	Wind speed is independent and identically distributed
Log-normal	It is often used to model wind speed in areas with atmospheric stability effects, such as thermal turbulence or wind shear	$f(x) = \frac{1}{(x\sigma\sqrt{2\pi})} * \frac{(\ln(x)-\mu)^2}{2\sigma^2}$ σ is the standard deviation of the logarithm of wind speed, x is the wind speed, and μ is the mean of the logarithm of wind speed	It models wind speed frequency distributions in areas with atmospheric stability effects, and it can account for the skewness and kurtosis of the distribution	It assumes the same independence and identical distribution of wind speeds
Rayleigh	It is often used to model wind speed in areas with no dominant wind direction	$f(x) = \frac{x}{\sigma^2} * e^{-\left(\frac{x^2}{2\sigma^2}\right)}$ x is the wind speed, and σ is the scale parameter	It is a simple distribution that can be easily applied in practice	Wind speed and distribution assumption
Weibull	It is a popular method used to model wind speed	$f(x) = \frac{k}{\lambda} * \left(\frac{x}{\lambda}\right)^{(k-1)} * e^{-\left(\frac{x}{\lambda}\right)^k}$ λ is the scale parameter x is the wind speed, and k is the shape parameter	It is a flexible distribution that can take a wide range of shapes	Wind speed and distribution assumption

4 Results and Discussion

Wind speed data and corresponding power output were generated using four distinct probability distributions. A wind turbine power curve was applied to calculate the power output for each wind speed. Finally, histograms were plotted to visualize the wind speed and power output data for each distribution.

Although the histograms of wind speed and power output data from different distributions may look similar, the underlying distributions can have distinct characteristics and implications for analysis and modelling as shown in Fig. 1a and b.

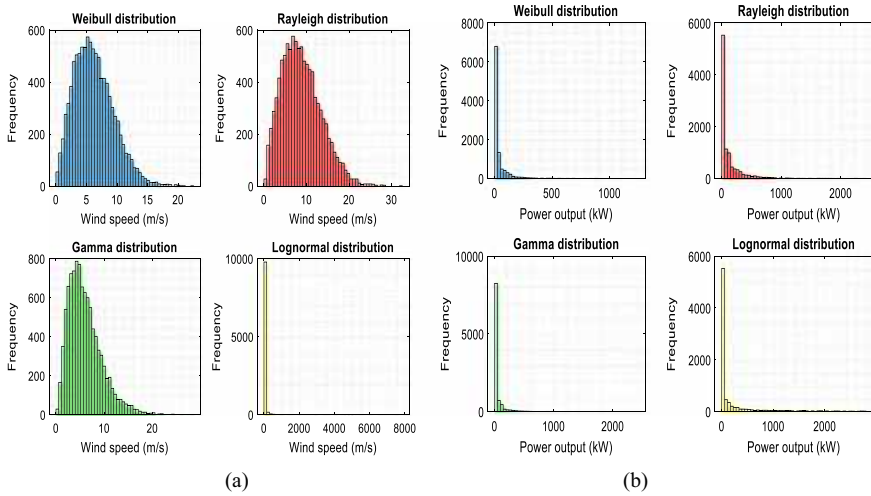


Fig. 1 a, b: Wind speed and power output of different distributions

In the context of wind energy, it is crucial to understand not only the average wind speed but also how it varies over time. The Weibull distribution offers a flexible shape that can capture different wind speed characteristics, making it a suitable choice for modeling wind speed data. In analyzing wind speed and power output data, the Weibull distribution is a suitable choice for capturing both the average and variability of wind speed, while the Rayleigh distribution is a simplified version of the Weibull. The Gamma distribution is a versatile option that can model various phenomena, whereas the Lognormal distribution is well-suited for analyzing skewed or asymmetric data. By utilizing the Weibull distribution, analysts can gain insights into the probability distribution of wind speeds at a particular location. This information is valuable in various applications related to wind energy, such as estimating the energy production potential of wind turbines, designing wind farms, and assessing the reliability and performance of wind power systems. However, it is important to carefully consider the underlying assumptions and limitations of each distribution when selecting one for modeling or analyzing wind speed and power output data.

5 Conclusion and Further Works

Wind speed data generated from various probability distributions can be utilized to model wind energy production and predict the expected output of wind turbines. This information, along with DLR technology, can be used to dynamically adjust the grid's transmission capacity based on wind power forecasts. For instance, if the forecast predicts high wind speeds and high-power output, the DLR system can increase the transmission capacity to handle the additional power. If the forecast predicts low

wind speeds and power output, the DLR system can reduce the transmission capacity to avoid grid overloading.

In a future study, the information generated from the probability distributions and wind turbine power curve will be useful for developing and optimizing models used in forecasting wind energy production. This will ensure the integration of wind energy into the grid with DLR technology to increase transmission capacity and ensure grid reliability.

References

1. Troccoli A (2014) Climatic changes: looking back, looking forward. In: Troccoli A, Dubus L, Haupt SE (eds) *Weather matters energy*. New York, NY: Springer, pp 65–89. https://doi.org/10.1007/978-1-4614-9221-4_3
2. Lawal OA, Akinyemi TO, Ramonu OJ (2018) Mitigating the challenges of global warming by harnessing the electric power generation potential of gas flaring in Nigeria 9(3):7
3. Talpur S, Wallnerstrom C, Flood C, Hilber P (2015) Implementation of dynamic line rating in a sub-transmission system for wind power integration. *Smart Grid Renew Energy* 6(8):233–249
4. Teh J, Cotton I (2015) Risk informed design modification of dynamic thermal rating system. *IET Gener Transm Distrib* 9(16):2697–2704. <https://doi.org/10.1049/iet-gtd.2015.0351>
5. Fernandez E, Albizu I, Bedialauneta MT, Mazon AJ, Leite PT (2016) Review of dynamic line rating systems for wind power integration. *Renew Sustain Energy Rev* 53:80–92. <https://doi.org/10.1016/j.rser.2015.07.149>
6. Cao J, Du W, Wang HF (2016) Weather-based optimal power flow with wind farms integration. *IEEE Trans Power Syst* 31(4):3073–3081. <https://doi.org/10.1109/TPWRS.2015.2488662>
7. Lawal OA, Teh J (2022) Dynamic thermal rating forecasting methods: a systematic survey. *IEEE Access* 10:65193–65205. <https://doi.org/10.1109/ACCESS.2022.3183606>
8. Lawal OA, Teh J (2023) Assessment of dynamic line rating forecasting methods. *Electr Power Syst Res* 214:108807. <https://doi.org/10.1016/j.epsr.2022.108807>
9. Lawal OA, Teh J (2023) Dynamic line rating forecasting algorithm for a secure power system network. *Expert Syst Appl* 219. <https://doi.org/10.1016/j.eswa.2023.119635>
10. Lawal OA (2023) Network topology optimisation for improved wind penetration and reliability. *AIP Conf Proc* 2795(1):020007. <https://doi.org/10.1063/5.0121019>
11. Lai C-M, Teh J (2022) Comprehensive review of the dynamic thermal rating system for sustainable electrical power systems. *Energy Rep* 8:3263–3288. <https://doi.org/10.1016/j.egy.2022.02.085>
12. Teh J, Lai C-M, Cheng Y-H (2017) Impact of the real-time thermal loading on the bulk electric system reliability. *IEEE Trans Reliab* 66(4):1110–1119. <https://doi.org/10.1109/TR.2017.2740158>
13. Jimada-Ojuolape B, Teh J (2022) Composite reliability impacts of synchrophasor-based DTR and SIPS cyber-physical systems. *IEEE Syst J* 16(3):3927–3938. <https://doi.org/10.1109/JSYST.2021.3132657>
14. Jimada-Ojuolape B, Teh J (2022) Impacts of communication network availability on synchrophasor-based DTR and SIPS reliability. *IEEE Syst J* 16(4):6231–6242. <https://doi.org/10.1109/JSYST.2021.3122022>
15. Teh J (2018) Uncertainty analysis of transmission line end-of-life failure model for bulk electric system reliability studies. *IEEE Trans Reliab* 67(3):1261–1268. <https://doi.org/10.1109/TR.2018.2837114>
16. Teh J, Lai C (2019) Reliability impacts of the dynamic thermal rating and battery energy storage systems on wind-integrated power networks. *Sustain Energy Grids Netw* 20:100268

17. Mohamad F, Teh J, Lai C-M (2021) Optimum allocation of battery energy storage systems for power grid enhanced with solar energy. *Energy* 223:120105. <https://doi.org/10.1016/j.energy.2021.120105>
18. Lawal OA (2021) Prospects of using dynamic thermal rating for a reliable power system network: a review. In: 2021 IEEE international future energy electronics conference (IFEEEC), pp 1–7. <https://doi.org/10.1109/IFEEEC53238.2021.9661878>
19. Lai C-M, Teh J (2022) Network topology optimisation based on dynamic thermal rating and battery storage systems for improved wind penetration and reliability. *Appl Energy* 305. <https://doi.org/10.1016/j.apenergy.2021.117837>
20. Jimada-Ojuolape B, Teh J (2020) Surveys on the reliability impacts of power system cyber-physical layers. *Sustain Cities Soc* 62:102384. <https://doi.org/10.1016/j.scs.2020.102384>

Modeling of Multijunction Solar Cells InGaAs/InGaP/GaAs/GeSi for Improving the Efficiency of PV Modules by 43%



Muhammad Shehram, Muhammad Najwan Hamidi,
Aeizaal Azman A.Wahab, and M. K. Mat Desa

Abstract In the current era, a lot of work is done on solar efficiency improvement because the world is turning toward clean energy, which is environmentally friendly and generates less global warming. More than one-junction solar cells enhance the performance of the PV modules; in this paper, four junction solar cells are exploited to convert solar power into fruitful electrical power. The InGaP/ InGaAs/GaAs/GeSi-based four-junction solar cells are produced on a micro-scale to improve the efficiency of the system by up to 43%, on the irradiated radiation intensity of the sun is near about 1000 w/m^2 , which is double the efficiency in comparison to the one junction solar cells. The concentration of the sun increases the output efficiency of the module decreases due to thermal heat losses. Silicon is used as a substrate for the formation of the four-junction solar cells to improve the performance of the solar module; the wafer-based bonding is used in this system, and the simulation has been done in Matlab Simulink to get a better result.

Keywords Four junction PV cells · Vapor phase with metal organic epitaxy (MOVPE) · Si substrate · Wafer-bonds

1 Introduction

More than one type of semiconductor material uses to produce more than one junction solar PV cells for the efficiency improvement of the solar panels to get the better utilization of the solar energy into consumable electrical power. III-V multijunction solar PV cells are adapted to enhancing the performance of the PV module on the 796 w/m^2 irradiated radiation intensity; and concentration of the sun; the junction is electrically and mechanically connected the efficiency of the module decrease when the obsession is much high 1378 w/m^2 [1]. Mechanically connected more than

M. Shehram · M. N. Hamidi (✉) · A. A. A.Wahab · M. K. M. Desa
School of Electrical and Electronics Engineering, Universiti Sains Malaysia, Nibong Tebal,
Malaysia
e-mail: najwan@usm.my

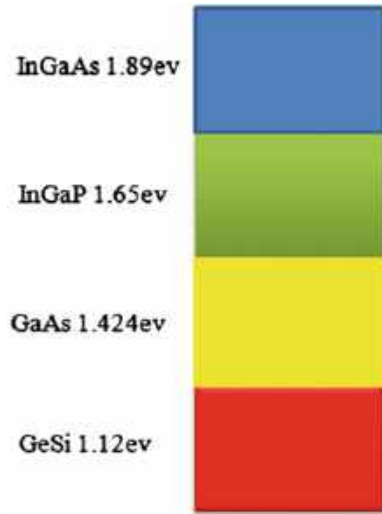
one junction solar cells contain optimistic performance; triple-junction solar cells are produced by this method for space use because it has good resistance against radiation [2]. To achieve the high performance of solar modules, triple junction solar cells are constructed with a different material by using the wafer bond, and double junction solar cells are fabricated by using the MOVPE technique in which silicon uses as a substrate for further improvement of the performance of the module Cells are produced by direct wafer bonding for highly doping materials of GaAs-n and Si-n; these bonding are performers at 120 temperature, and the maximum efficiency of the system finds 30% [3]. Multijunction solar cells have good performance; GaAs-based devices have a far efficiency of 70 mv for each junction; the dislocation and lattice mismatch present in silicon and other layers, which affect the efficiency of the PV cells the germanium layer is a mismatch with silicon introduces to produce the active layer in the system [4]. A lot of MOVPE issues occur while producing multijunction solar cells; for this purpose, high-caliber buffers use between the layers for better performance; many growth issues also occur during the fabrication process, and good quality material used to overcome the MOVPE issue like phosphine [5].

The tandem solar cell is an excellent solution for the improvement of the net power of the PV cells; the wafer bonding-based three-junction PV cells developed previously; raised the efficiency of the solar module to 34% to more improve the performance of the module 4-junction PV cells are introduced which boost the performance [6]. The simulation of three-junction PV cells conduct by using the different semiconductor materials on the specified intensity of radiation and temperature simulation done in Matlab Simulink and IV, and PV characteristic curve draw for triple junction solar cells, all fabric for the preparation of triple junction solar cells are lattice matching [7]. The five layers four junction solar cells design to utilize the full spectrum of the sun; different parameters are drawn, and compare the results as an IV and PV curve; this is an outstanding solution for the improvement of the efficiency of the PV cells and the output performance of the cells is 76.198%, and efficiency of the fill factor of the PV cell is 43.61% [8]. Four junction solar cells InGaP/InGaAs/GaAs/InGaSb with double diode used to improve the performance [9].

2 Problem Statement

The existing multijunction solar cells contain less efficiency. Add one more band gap of nearly 1.12 eV in triple junction solar cells in which silicon is grown on germanium in the bottom cell to raise the performance of the cells to 43% keen scale cells are produced, which occupy less area on a small scale.

Fig. 1 Four-junction PV cells



3 Methodology

3.1 Four-Junction PV Cells

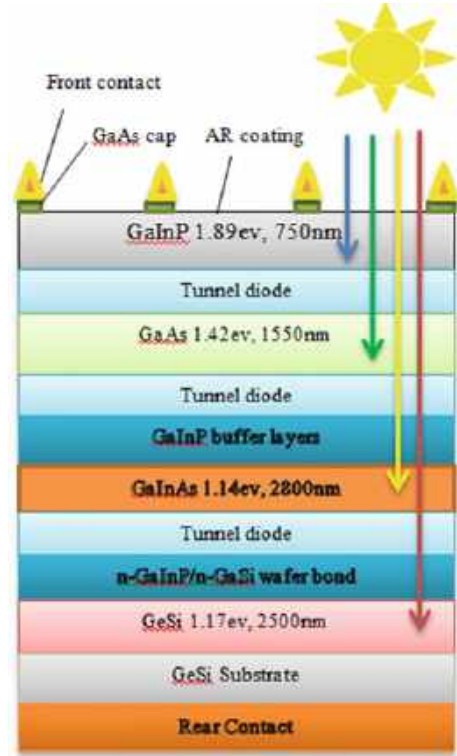
The performance of multijunction PV cells improves by adding one more junction of semiconductor materials to become the four-junction solar cells; each alliance contains its different band gap energy and material. The top cell encloses blue color and 1.89 eV energy; the middle junction is green in color and represses 1.65 eV energy; the second medial cell is yellow and holds 1.42 eV power; the bottom cell is red and contains 1.12 eV energy (Fig. 1).

3.2 Four-Junction PV Cells Fabrication

The four-junction PV cells design in this paper further improves the efficiency of the multijunction solar cells by up to 43%; each material contains its band gap energy in electron volts and consumes the full spectrum of the sunlight to improve the conversion energy of the module the below diagram shows the fabrication process of four junction solar cells (Fig. 2).

The top cell is n-InGaP which consists of 1.89 eV band gap energy and contains a 750 nm wavelength of the irradiated radiation intensity of the sunlight; this is growth on the top of the GaAs junction by using the tunnel diode in between them. The n-InGaP is grown top the GaAs; this is the second junction of the cell, and its band gap energy is 1.65 eV; and its wavelength is 1550 nm; this junction is pretending as a buffer layer below this junction third junction exists, which is the construct of

Fig. 2 MPV fabrication process



GaAs, and its band gap energy is 1.42 eV wavelength is 2800 nm, both junctions are connected through the tunnel diode the last junction is GeSi its wavelength is 2500 nm, and its band gap energy is 1.12 eV this is the fabrication process of four junction solar cells which is grown for enhance the performance of the PV modules by using the full spectrum of the sunlight.

4 Results

Four junction solar cells are simulated in Matlab Simulink the Simulink model is given in Fig. 3 the four types of different materials are used to generate this type of module, and each material contains its band gap energy and absorbs the sunlight according to its color and properties the overall system enhanced the performance of the PV module up to 43%, the size of the PV module is 2 m².

Four different solar cells with unique band gap energy are model in this paper; the top cell is InGaAs which contains 1.89 eV energy and absorbs blue color light, and the second band gap material is InGaP, its power is 1.65 eV and absorbs green color light the third junction is GaAs whose band gap energy is 1.424 eV and yellow the

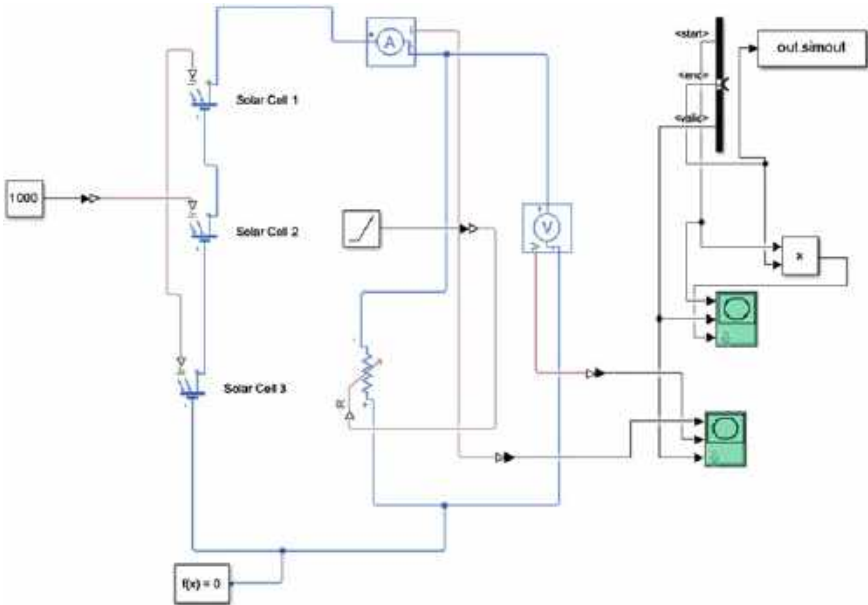


Fig. 3 Modeling of four junction solar cells in matlab simulink

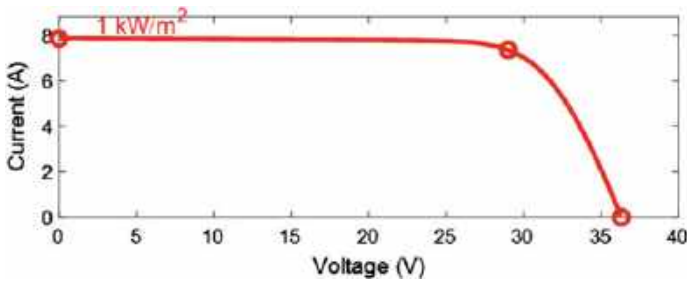


Fig. 4 Current versus voltage characteristic curve of four junction solar cells

fourth band gap material is GeSi which has 1.12 eV power and red color all materials are model in Matlab Simulink and output results is shown in Fig. 4.

The characteristic curve of four junction solar cells is obtained in Matlab Simulink for one module, and the value of irradiated radiation intensity is specified, which is 1000 W/m^2 ; the current and voltage value measured on this intensity value the current value shown on the x-axis and voltage shows on the y-axis in Fig. 4 at 1000 W/m^2 , the current value of four junction module is 7.9 A, and the voltage value is 37 v for one module of the 4-junction PV cells.

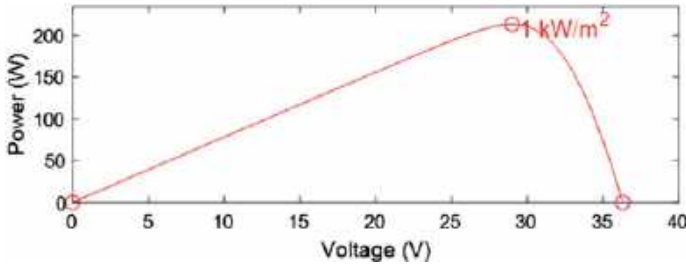


Fig. 5 Power versus voltage characteristic curve of MPV

The power versus voltage characteristic curve of the four-junction PV cells is drawn and shown in Fig. 5; the results are generated on 1000 W/m^2 , radiation intensity of the sun showing that the power of the panel is 205w and the voltage is 37 V.

5 Conclusion

The present solar modules have less efficiency for improving these modules' efficiency; different materials of different band gap energy are used in this paper. More than one junction solar cells enhance the efficiency of the overall system; the four junction solar cells are exploited to convert the solar energy into electrical energy. The InGaP/ InGaAs/GaAs/GaSi-based four-junction solar cells are produced on a micro-scale to improve the efficiency of the system by up to 43%, on the irradiated radiation intensity of the sun is near about 1000 w/m^2 , which is double the efficiency in comparison to the one-junction one band gap PV cells. The concentration of the sun increases the output efficiency of the module decreases due to thermal heat losses. The output result is generated in Matlab Simulink and proves that four-junction solar cells are admirable performance compared to the simple junction solar cells.

Acknowledgements The research work is financially endorsed by Universiti Sains Malaysia Short Term Grant 304/PELECT/6315747.

References

1. Predan F, Franke A, Hoehn O, Lackner D, Helmers H, Siefer G, Bett AW, Dimroth F (2020) Wafer-bonded GaInP/GaAs/GaInAs//GaSb four-junction solar cells with 43.8% efficiency under concentration. In: 2020 47th IEEE photovoltaic specialists conference (PVSC). IEEE, pp 0250–0252
2. Kawakita S, Imaizumi M, Makita K, Nishinaga J, Sugaya T, Shibata H, Sato S, Ohshima T (2016) High efficiency and radiation resistant InGaP/GaAs//CIGS stacked solar cells for space

- applications. In: 2016 IEEE 43rd photovoltaic specialists conference (PVSC). IEEE, pp 2574–2577
3. Essig S, Benick J, Schachtner M, Wekkeli A, Hermle M, Dimroth F (2015) Wafer-bonded GaInP/GaAs//Si solar cells with 30% efficiency under concentrated sunlight. *IEEE J Photovoltaics* 5(3):977–981
 4. Hinzer K, Beattie MN, Wilkins MM, Valdivia CE (2018) Modeling techniques for multijunction solar cells. In: 2018 International conference on numerical simulation of optoelectronic devices (NUSOD). IEEE, pp 133–134
 5. Hinojosa M, García I, Algora C, Martínez O (2017) MOVPE issues in the development of ordered GaInP metamorphic buffers for multijunction solar cells. In: 2017 Spanish conference on electron devices (CDE). IEEE, pp 1–4
 6. Dimroth F, Müller R, Predan F, Siefert G, Schygulla P, Benick J, Höhn O et al (2020) 34.1% Efficient GaInP/AlGaAs//Si tandem cell. In: 2020 47th IEEE photovoltaic specialists conference (PVSC). IEEE, pp 1543–1546
 7. Tripathi SK, Chakraborty S, Singh AK (2018) Modeling and simulation of multijunction solar or sun based cell. In: 2018 International conference on recent innovations in electrical, electronics & communication engineering (ICRIEECE). IEEE, pp 196–199
 8. Chaudhary JK, Kanth R, Skön JP, Heikkonen J (2019) Analysis and enhancement of quantum efficiency for multi-junction solar cell. In: 2019 IEEE 46th Photovoltaic specialists conference (PVSC). IEEE, pp 0210–0214
 9. Gungi DV, Das N, Elavarasan RM (2021) Modelling of 4-junction PV cell considering MPPT for conversion efficiency enhancement. In: 2021 31st Australasian universities power engineering conference (AUPEC). IEEE, pp 1–6

Control Method of Two-Stage Grid-Connected PV Inverter System



Wang Zhe, Dahaman Ishak, and Muhammad Najwan Hamidi

Abstract A two-stage, grid-connected PV inverter, and its control method are proposed in this paper. By controlling the DC link voltage at the front stage and the PWM of the inverter circuit at backstage, an LCL-type PV three-phase grid-tied inverter system is established. The paper analyzes the inverter output by this control system and compares it with the PV output power on the DC link voltage and current for maximum power tracking. The results indicate the proposed method has an excellent performance from the PV modules, the output power is always very stable, and the total harmonic distortion is very small. The current and voltage generated at the output end of the inverter have fast dynamic response, smooth waveform and strong stability.

Keywords Two-stage · Grid-connected · PV inverter

1 Introduction

As energy demands continue to increase, and a desire to utilize renewable energy sources to reduce pollution and combat the effect of climate change, PV renewable energy has become a key focus for human use. Therefore, a cutting-edge research direction is now frequently campaigned in utilizing PV energy and transmitting it more efficiently into the power grid. This paper focuses on a two-stage PV inverter and its control method for grid connection. The two-stage PV grid-connected inverter mainly controls the DC link voltage (front stage) and the inverter drive signal (backstage). Meanwhile, there is closed-loop control between the front and back stages.

W. Zhe (✉)

Changchun Automotive Industry College, Changchun of Jilin Province, Beijing 13000, China
e-mail: Wangzhe9@student.usm.my

W. Zhe · D. Ishak · M. N. Hamidi

School of Electrical and Electronic Engineering, Universiti Sains Malaysia, 14300 Nibong Tebal, Penang, Malaysia

The front stage DC link voltage can change the reference current value of the back-stage through the control loop, achieving efficient control of the input of electrical energy to the grid, improving the quality of electrical energy, and achieving a good power factor. Additionally, this method can reduce the cost of micro PV inverters [1–3]. This paper mainly introduces the structure and control strategy of an LCL-type PV three-phase, grid-connected inverter and the control method of the two-stage LCL-type PV three-phase grid-connected inverter system. The DC link voltage, output power, output voltage and current, and grid-connected power quality of the model are established and analyzed in detail. The paper also compares the effects of PV output power on the DC link voltage, output current, and harmonics.

2 Tow-Stage LCL Type PV Three Grid-Connected Inverter System

The system includes PV arrays, MPPT system, boost circuit, DC link voltage, three-phase inverter circuits, LCL filtering circuit, grid connection, and PWM control for inverter driving signal, as shown in Fig. 1.

The working process of the system is as follows: The PV arrays convert solar energy into DC power, and the MPPT system achieves maximum power tracking by matching the output voltage and current to the optimal point [4]. Then, the boost circuit boosts the DC voltage to the level required by the inverter and sends the DC power to the DC link.

Subsequently, the three-phase inverter circuits convert the DC power into AC power and filter it through the LCL filtering circuit to eliminate high-frequency noise and harmonics. Finally, the AC power is sent into the grid [5]. At the same time, the control loop is responsible for monitoring the operation status of the inverter and adjusting the PWM driving signal to achieve output control and power regulation of the inverter.

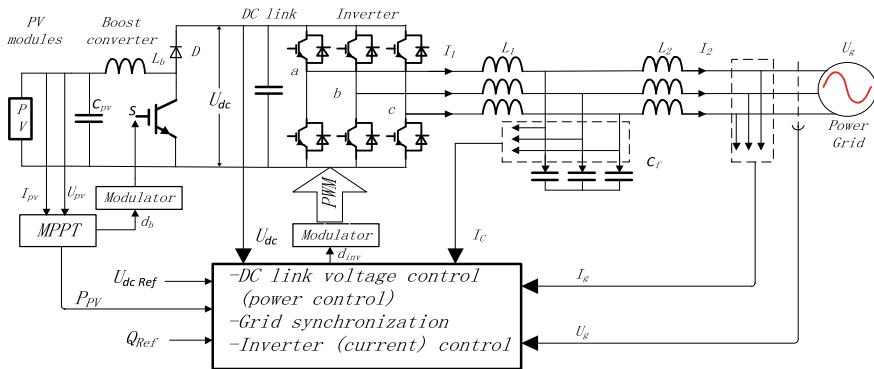


Fig. 1 LCL type photovoltaic three grid-connected inverter system

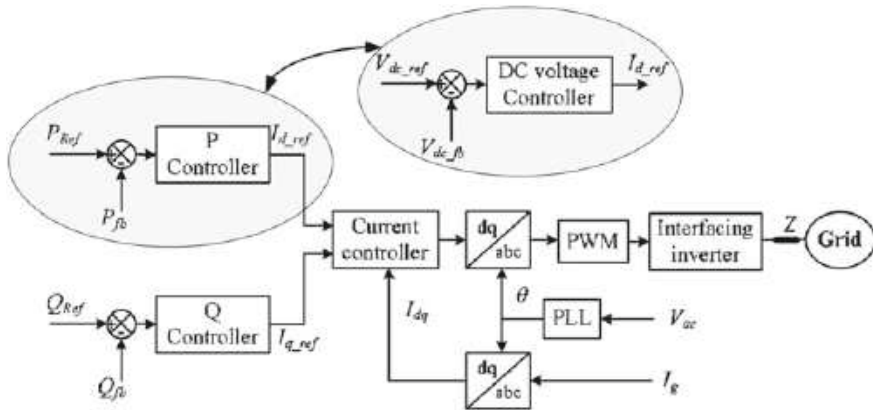


Fig. 2 Control strategy of two—stage photovoltaic inverter system

2.1 Control Strategy

The three-phase grid-connected inverter system consists of a front-end DC link voltage control stage, which uses a PI controller to regulate the DC link voltage at a given reference voltage value [6]. The second stage controls the PWM signal to adjust the values on the DQ rotating coordinate system within a given reference value range [7].

By setting up a control loop and using PI control, we can regulate the DC link voltage to a stable value of 400 V. Specifically, we can adjust the input active power P by comparing controls to change the reference current I_d in the PQ control, to achieve the goal of controlling the DC bus voltage [8]. For example, when the DC link voltage is higher than the set value, we increase the reference current I_d to increase the power input to the inverter and lower the DC link voltage. In this way, we can achieve stable control of the DC link voltage, ensuring the normal operation of power electronic devices and improving their efficiency and reliability [9], as shown in Fig. 2.

2.2 Model Building and Analysis

The current loop control model of the LCL-type PV grid-connected inverter with grid current feedback is shown in Fig. 3. The grid current I_g and reference current I_{g_ref} is adjusted and modulated with a high-frequency triangular carrier to obtain a drive signal to the inverter bridge and achieves current tracking [10].

$G_{PI}(S)$ is the transfer function of the inverter bridge controlled by PWM.

$$G_{PI}(S) = k_p[1 + 1/(k_i s)] \quad (1)$$

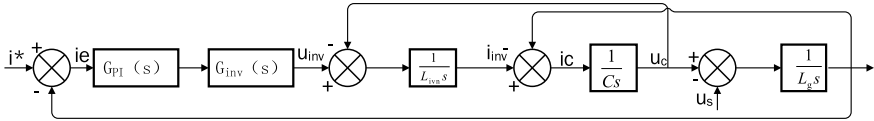
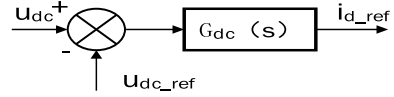


Fig. 3 Current loop control model

Fig. 4 Voltage loop control model



$G_{inv}(S)$ is the gain of the PWM inverter.

$$G_{inv}(S) = U_{dc}/(1 + T_S S) \quad (2)$$

The transfer function of the grid current I_g relative to the inverter AC voltage for phase A is

$$G_g(S) = \frac{I_g(S)}{U_{inv}(S)} = \frac{1}{L_{inv}L_gCs^3 + (L_{inv} + L_g)s} \quad (3)$$

When the voltage disturbance from the grid is not considered, due to the high switching frequency of the inverter, as T_S approaches zero, from Eqs. (1), (2) and (3), the open-loop transfer function of the inner loop of the current can be calculated as

$$G_o(S) = \frac{k_p U_{dc} [s + (1/k_i)]}{L_{inv}L_gCs^4 + (L_{inv} + L_g)s^2} \quad (4)$$

Control of DC link voltage, as shown in Fig. 4

G_{dc} Is bus voltage PI link transfer function

$$G_{dc}(S) = \frac{k_p k_{is} + k_p}{k_{is}(1 + U_{dc_ref}e)} \quad (5)$$

3 Results and Discussion

To verify the effectiveness of the proposed control method, a simulation model of an LCL-type PV grid-connected inverter was built in the MATLAB simulation environment, and the control strategy shown in Figs. 3 and 4 was implemented. The PV

array can produce output power of 4.2 kW, and the MPPT algorithm adopts the P&O method. The parameters are shown in Table 1.

After the system reaches a steady state, the simulated grid-connected PV system delivers output power of around 4 kW as shown in Fig. 5, and the system can operate efficiently and stably with a good power factor. Figure 6 shows the grid-connected output voltage, with two cycles of waveform displayed, and the waveform is stable and normal. Figure 7 shows the grid-connected output current, which is stable but with small disturbances in the waveform. Figure 8 shows the waveform of the input voltage to the DC link, which is stable at around 400 V after the system reaches steady state, consistent with the given reference voltage, but there is still a certain ripple, with fluctuations of 3–5 V. Figure 9 shows the THD analysis of the grid current, with a THD value of 1.40%. The overall control effect is good, but there are some low-frequency harmonics in the figure, with peaks at 75 Hz and 175 Hz, respectively.

To further verify the stability of the experimental system, the effect of changing the solar irradiance was investigated. The light intensity drops from 1000–800 W/m², as shown in Fig. 10. So the output power of the solar panel was reduced from 4.2–3.8 kW by changing the light intensity, as shown in Fig. 11. The grid current before and after the change is shown in Fig. 12. It can be seen that when the PV input power changes, the grid current changes smoothly without a large overshoot, and then stabilizes. At the same time, the DC link voltage is stable without any significant changes. Figure 13 shows the DC link voltage before and after the change

Tab.1 Simulation parameters

Pa Parameters and units	Numerical values
Energy storage inductance L_0/mH	2.5
DC side capacitance $C_{dc}/\mu F$	5000
Inverter AC side inductance L_2/mH	0.3
Filter capacitor $C/\mu F$	15
Peak value of power grid voltage U_M/V , frequency/Hz	155.5 50
Switching frequency f/kHz	10
Reference voltage of the DC side U_{dcRef}/V	400

Fig. 5 Output active power

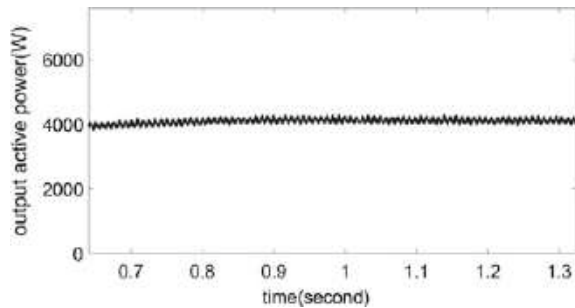


Fig. 6 Grid voltage output

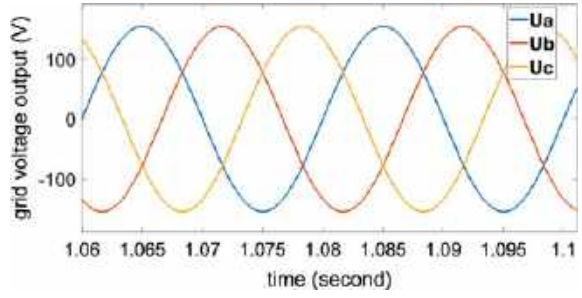


Fig. 7 Grid current output

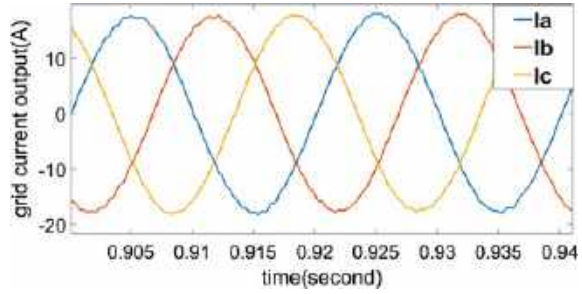
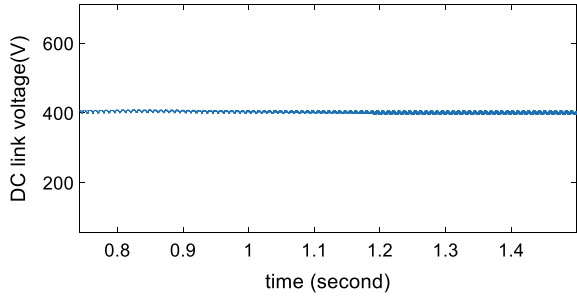


Fig. 8 DC link input voltage



in output power, which decreases after the change and stabilizes around 400 V, with smaller voltage ripple compared to before.

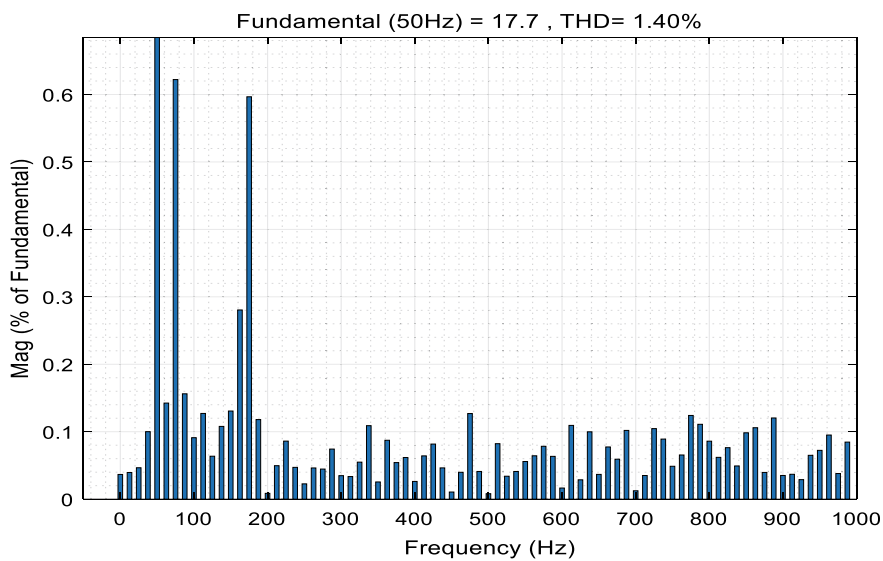


Fig. 9 THD of grid current

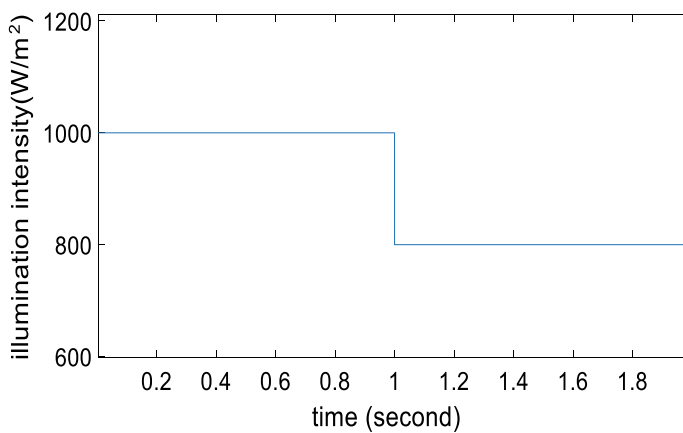


Fig. 10 Light intensity

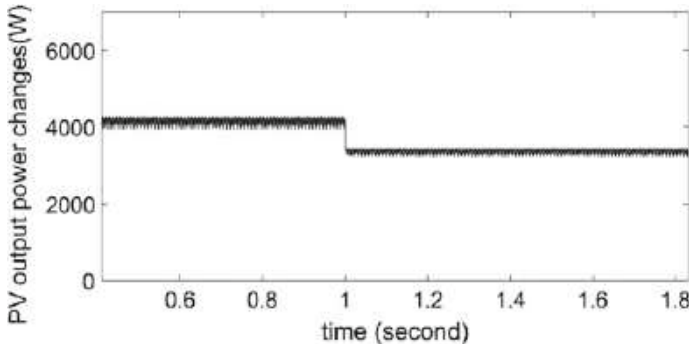


Fig. 11 PV output power changes

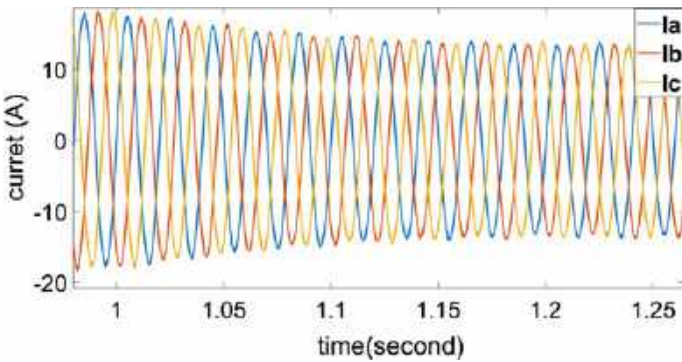


Fig. 12 Gridt current

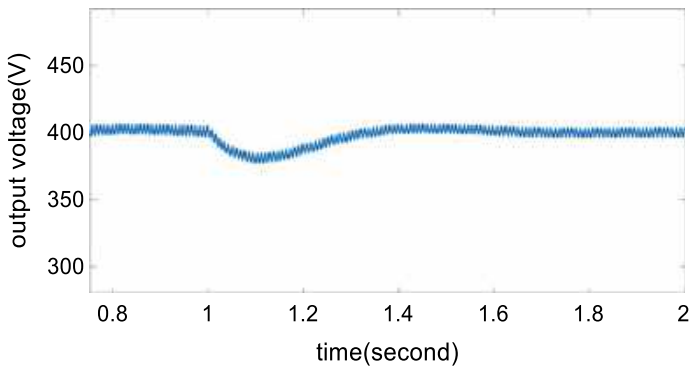


Fig. 13 DC link voltage

4 Conclusion

The two-stage, grid-connected PV inverter controls the DC link voltage (front stage) and the inverter circuit PMW (backstage), and adds a control loop for the bus voltage in the front stage. The reference current of the backstage is also changed, and the DQ coordinate system is used to control the PMW drive signal. This effectively transfers the PV energy to the grid, the power factor $\cos(\phi) = 0.95$, good grid-connected energy quality. Additionally, when the output power of the PV system changes, it has a good dynamic response and high stability. This type of control can greatly reduce costs and eliminate some components such as batteries in micro PV systems. The optimization of the two-stage control strategy is a key research focus for further improving control accuracy and efficiency in the later stage.

References

1. Li T, Fu L (2023) Research on the control strategy of LCL-type PV grid-connected inverter. *J Phys Conf Ser*
2. Lee J-S, Lee KBJE (2013) Variable DC-link voltage algorithm with a wide range of maximum power point tracking for a two-string PV system 6(1):58–78
3. Abd Rahim N, Selvaraj J (2007) Hysteresis current control and sensorless MPPT for grid-connected photovoltaic systems. In: 2007 IEEE international symposium on industrial electronics. IEEE
4. Li J et al (2022) Analysis of photovoltaic array maximum power point tracking under uniform environment and partial shading condition: a review 8:13235–13252
5. Qin D et al (2020) Adaptive bidirectional droop control for electric vehicles parking with vehicle-to-grid service in microgrid 6(4):793–805
6. Trujillo C, Santamaría F, Gaona EE (2016) Modeling and testing of two-stage grid-connected photovoltaic micro-inverters 99:533–542
7. Fahad S et al (2018) Particle swarm optimization based DC-link voltage control for two stage grid connected PV inverter. In: 2018 International conference on power system technology (POWERCON). IEEE
8. Tan AST et al (2022) model predictive control of single-phase simplified split-source inverter. In: Proceedings of the 11th international conference on robotics, vision, signal processing and power applications: enhancing research and innovation through the fourth industrial revolution. Springer
9. Wang Y, Du J, Yang K (2022) Supply and demand balance control of electrical submersible linear motor for offshore oil exploitation (2022)
10. Aroudi AE et al (2020) Multiple-loop control design for a single-stage PV-fed grid-tied differential boost inverter 10(14):4808

Methodological Comparison and Analysis for Six-Switching PMBLDC Motor Control



Musa Mohammed Gujja, Dahaman Ishak, and Muhammad Najwan Hamidi

Abstract Permanent Magnet Brushless Direct Current (PMBLDC) motor plays a significant role in our daily activities and its area of application is enormous. It can be seen in the areas of robotics, military, aerospace, domestic, and industrial machine among others. Hence, a precise approach to PMBLDC motor control is imperative. This study compares and evaluate three control strategy employing Machine Learning (ML), Response Optimizer (RO), and PID controller to attain the most suitable and effective PMBLDC motor control. A six-switching driver circuit was designed to drive the motor. The simulated result was analyzed both graphically and analytically. Some differences were observed in terms of overshoot, undershoot, rise, and settling time, however, all three-control approaches RO, ML, and PID follow the reference tracking.

Keywords Machine Learning · Optimization · PID · Driver circuit · PMBLDC motor

1 Introduction

PMBLDC motor has played a tremendous role in the modern industrial era. It is highly regarded if compared to conventional motors, especially when size, noise, speed, and maintenance implications are considered. So, the need for strategizing different control methods to achieve better speed, stability, and efficiency is required. Different control techniques were recently employed [1, 2]. These control techniques are used by different researchers for effective speed control [3]. These control approaches are compared to determine the best control option [4, 5]. The terminal voltage V , current i , and back-emf e in the PMBLDC motor windings can be expressed mathematically as:

M. M. Gujja (✉) · D. Ishak · M. N. Hamidi
School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Nibong Tebal,
Penang, Malaysia
e-mail: musa4gujja@student.usm.my

$$\begin{bmatrix} V_a \\ V_b \\ V_c \end{bmatrix} = \begin{bmatrix} R_s & 0 & 0 \\ 0 & R_s & 0 \\ 0 & 0 & R_s \end{bmatrix} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} + \frac{d}{dt} \begin{bmatrix} L - M & 0 & 0 \\ 0 & L - M & 0 \\ 0 & 0 & L - M \end{bmatrix} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} + \begin{bmatrix} e_a \\ e_b \\ e_c \end{bmatrix} \tag{1}$$

where R_s is the phase winding resistance, L and M are the phase inductance and mutual inductance respectively.

2 PMBLDC Motor Modelling

A three-phase, six-switching driving approach was used to drive the motor. Here, six IGBTs S_1, S_2, S_3, S_4, S_5 and S_6 were utilized as a switch for their high power and fast switching capabilities. The driver circuit was connected to a three-phase PMBLDC motor for effective control. The scheme works in 120° conduction with back-emf detection abilities with the help of a hall-effect sensor. In this topology, only two phases will be active at a time as shown in Fig. 1 and Table 1.

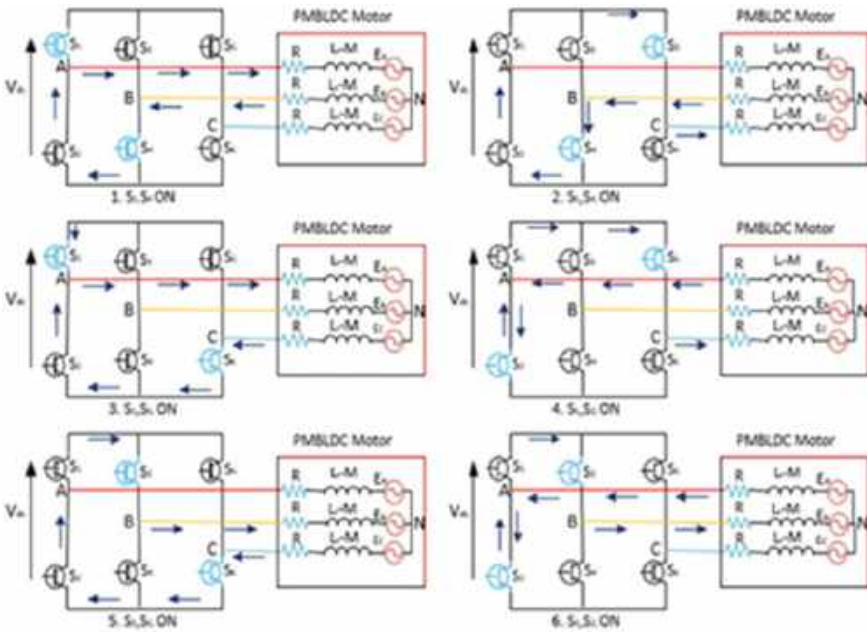


Fig. 1 Switching technique

Table 1 Switching sequence

Steps	Switch	Silent phase	Active phase
1	S ₁ ,S ₄	C	A,B
2	S ₅ ,S ₄	A	C,B
3	S ₁ ,S ₆	B	A,C
4	S ₅ ,S ₂	B	C,A
5	S ₃ ,S ₆	A	B,C
6	S ₃ ,S ₂	C	B,A

3 Control Strategy

Three control approaches i.e. PID controller, Machine Learning (ML), and Response Optimizer (RO) were employed to improve the system performance for stability, efficiency, and speed tracking. Optimization technique, Back-emf, and Pulse Width Modulation (PWM) control method are used in the model design.

3.1 Optimization

Optimization is a process of busting a system's performance to attain the best result. For this work, RO and ML were applied to improve the system performance by optimizing and training the model data to achieve a better output. More so, the PID controller was tuned, and the result was observed and compared with ML and RO to evaluate and ascertain the best control strategy.

3.2 PID Controller

The PID controller is the conventional controller used by most industries for speed, pressure, flow, and temperature control. This study used the system identification toolbox, shown in Fig. 3, to generate the model transfer function. The system model transfer function was tuned using the PID control tuner. The PID control result was compared with the other two control techniques.

3.3 Machine Learning

The ML toolbox was equally utilized to optimize the same system model. This was done by capturing the initial system set-up signal, the output of the signal was then

converted to data which was used to train the model, and the resultant output was also matched with the other control technique.

3.4 Response Optimization

The response optimizer toolbox is available in Simulink, and the model output was run at the initial PID setup. The simulation result was captured and equally optimized to improve the system performance. A new PID-optimized parameter was generated automatically, showing the model's best fit, and ensuring system stability and efficiency. The resultant output was also matched.

4 Simulation Results

MATLAB Simulink software was employed to simulate the system model. The model was tested at a speed of 1000, 3000 rpm, and at a random speed ranging from 50 to 1000 rpm consecutively for the ML, RO and PID control approaches. The resultant output was matched in terms of its settling time, rise time, undershoot, overshoot, system performance, and speed tracking. Figure 2 shows the complete model while Fig. 3 shows the input and output signal from system identification toolbox.

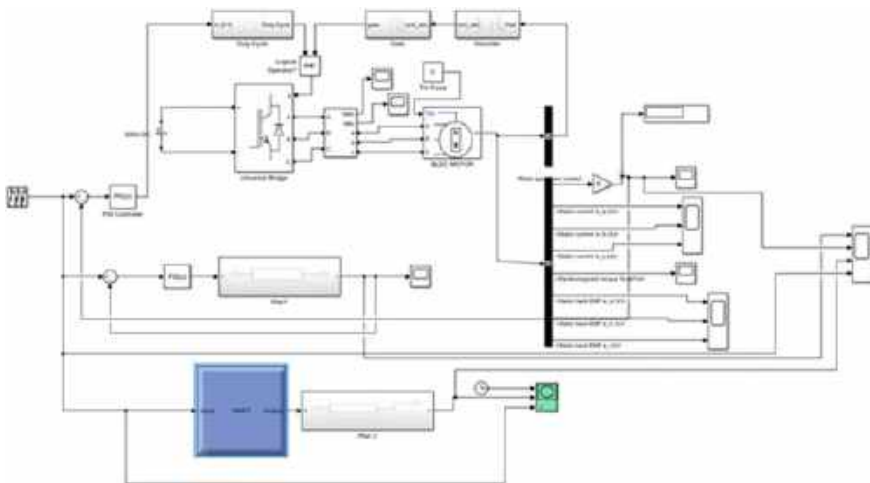


Fig. 2 Complete model

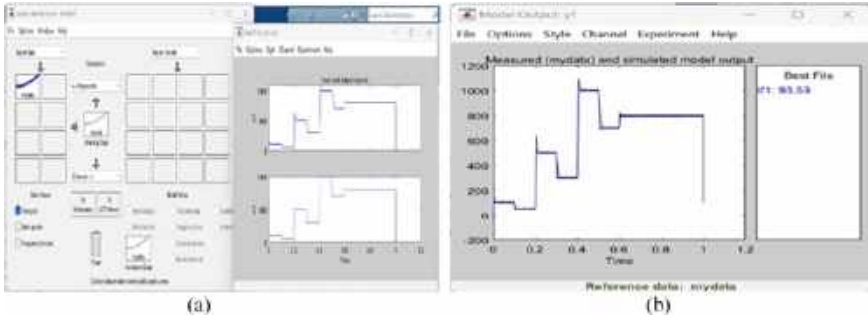


Fig. 3 a System identification toolbox and b stability and fitness of the model

4.1 Result Evaluation

The system output at 1000 and 3000 rpm indicates good reference tracking and stability. The motor speed in Fig. 4, stator current in Fig. 6, back-emf in Fig. 7, and electromagnetic torque in Fig. 8 were captured and observed graphically, more so, the resultant output was determined analytically for a good comparison as shown in Table 2. Considering the rotor speed, the ML, RO, and PID control techniques perfectly followed the reference speed, however, some differences in overshoot, undershoot, rise, and settling time were observed. The RO technique proved good when matched with the rest. The stator current also looks better with the RO compared to the ML where switching ripples are high. Also, with electromagnetic torque, some ripple, overshoot, and undershoot were highly observed in the ML when matched with the RO which looks better. Analytically, it was seen that at 1000 rpm, the RO rises at 3.562 ms compared to PID and ML with 3.655 ms and 4.246 ms respectively. Equally, considering the settling time, the RO settles at 4.899 ms which is also better compared to that of PID and ML with 15.786 and 10.618 ms correspondingly. Looking at the overshoot, the RO control strategy is equally better with 0.500% when matched with -0.028% and 10.556% for PID and ML respectively. The undershoot looks similar for all control techniques with 2% each as shown in Fig. 5.

At 3000 rpm, the PID and ML controls maintain the same parameter as that of 1000 rpm which indicates the parameter balance of the system at 1000 and 3000 rpm respectively, while that of the RO control strategy shows some fluctuation of 6.044 ms rise time, 5.862 ms settling time and 0.504% overshoot but still better.

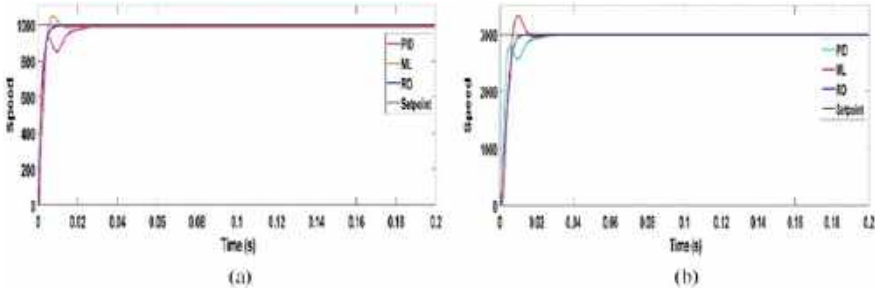


Fig. 4 a rotor speed at 1000 rpm and b rotor speed at 3000 rpm

Table 2 Analytical comparisons

Parameters	PID		ML		RO	
	1000	3000	1000	3000	1000	3000
Speed (rpm)	1000	3000	1000	3000	1000	3000
Rise time (ms)	3.655	3.655	4.246	4.246	3.562	6.044
Settling time (ms)	15.786	15.786	10.618	10.618	4.899	5.862
Overshoot (%)	- 0.028	- 0.028	10.556	10.556	0.500	0.504
Undershoot (%)	2.000	2.000	1.999	1.999	2.000	2.000

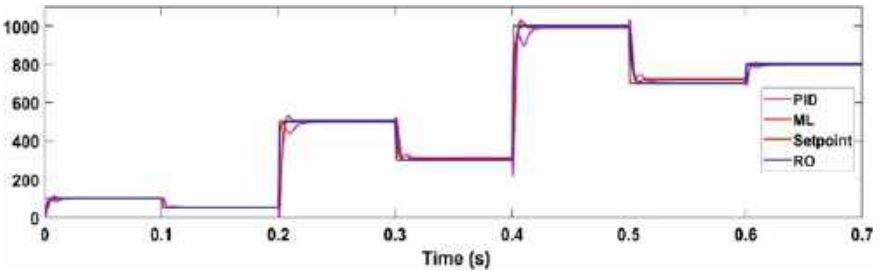


Fig. 5 Random rotor speed combine

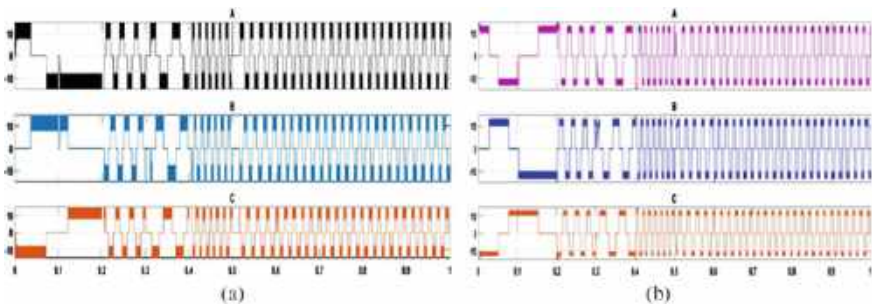


Fig. 6 Stator current at random speed a ML and b RO

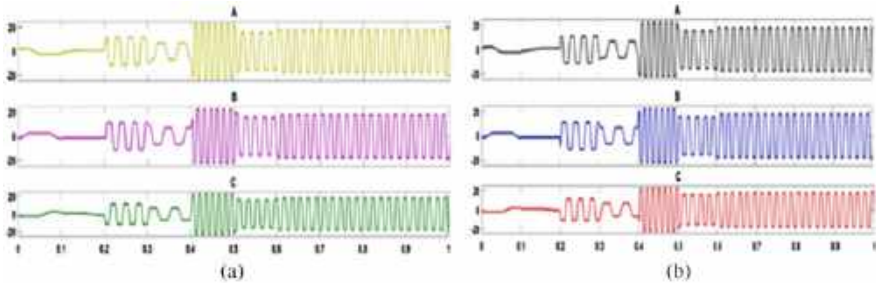


Fig. 7 Back emf at random speed **a** ML and **b** RO

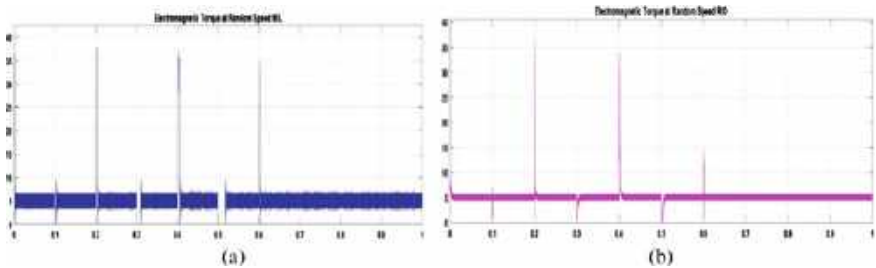


Fig. 8 **a** Electromagnetic torque at random speed ML and **b** RO

5 Conclusion

Graphically and analytically, it was observed that the RO control approach is better than the other two control strategies in terms of robustness, overshoot, settling, and rise times, but ML and PID maintain similar system values even at high speeds. Nevertheless, all three control approaches are good in reference tracking, efficiency, and good performance.

References

1. Gülbaş Ö (2020) Comparison of PI and super-twisting controller optimized with SCA and PSO for speed control of BLDC motor, pp 1–7. <https://doi.org/10.1109/HORA49412.2020.9152853>
2. Gujja MM, Ishak D (2023) Comparative evaluation and analysis of six- and four-switch drivers for PMBLDC motor control. In: 2023 (ICEPECC), pp 1–6. <https://doi.org/10.1109/ICEPECC57281.2023.10209482>
3. Auliansyah F, Qudsi OA, Ferdiansyah I (2020) controlling speed of brushless DC motor by using fuzzy logic controller. In: 2020 (iSemantic). IEEE, pp 298–304. <https://doi.org/10.1109/iSemantic50169.2020.9234290>
4. Zaid M et al (2018) Speed control of PMBLDC motor using fuzzy logic controller in sensorless mode with back-EMF detection. Springer, pp 439–447

5. Pudur R et al (2021) Speed control of PMBLDC motor using rotor position rotor speed PWM technique. Springer, pp 613–622

Energy Efficiency Performance Optimization and Surge Prediction of Centrifugal Gas Compressor



Mukhtiar Ali Shar, Masdi B Muhammad, Ainul Akmar B Mokhtar, and Mahnoor Soomro

Abstract In the modern day, oil and gas companies must meet fresh problems to effectively maximize their performance, not only about traditional performance such as reliability or productivity, but also emerging ones, related to sustainability issues. The authors present the development of a novel and robust surge prediction and energy performance control technique with the use of precise anti surge control, from the measurement to the control algorithm to the anti-surge valve and load variation with the support of speed control to ensure reliability and mitigate power loss caused by anti-surge valve opening by operating the centrifugal gas compressor adjacent surge control line. In the methodology the design and original equipment manufacturer (OEM) data for the compressor were reviewed, and then the off-design operating envelope was examined. Aspen HYSYS version 12.1 is used for dynamic simulation modelling. According to the first stage findings with 10% surge safety, the load variation control approach is more energy-efficient than compressor operation at maximum load, which consumes 13596 kWh with 99–94 MMSCFD during peak and off-peak demand and uses less energy 5183 kWh for 99–67.59 MMSCFD compression. Surging occurred when the flow rate decreased from 7022 to 4355 m³/h for variable speed and from 7022 to 5990 m³/h for fixed speed. When a surge was likely to occur, ASV opened aggressively at 22.29% with variable speed and 11.77%

M. A. Shar (✉) · M. B. Muhammad · A. A. B. Mokhtar · M. Soomro
Mechanical Engineering Department, Universiti Teknologi PETRONAS, Bandar Seri Iskandar,
32610 Perak, Malaysia

e-mail: mukhtiar_20000077@utp.edu.my; mukhtiar.shar@quest.edu.pk

M. B. Muhammad

e-mail: masdimuhammad@utp.edu.my

A. A. B. Mokhtar

e-mail: ainulakmar_mokhtar@utp.edu.my

M. Soomro

e-mail: mahnoor_22008342@utp.edu.my

M. A. Shar

Department of Mechanical Engineering, Quaid-E-Awam University of Engineering, Science and Technology (QUEST), Nawabshah, Sindh 67480, Pakistan

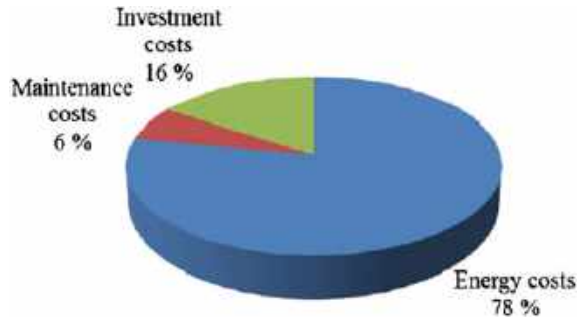
with fixed speed. Based on the results of the second stage with 10% safety, the load variation control method is more energy efficient, consuming between 23,288 and 9892 kWh with a 217 and 178MMSCFD compressor capacity, as opposed to the compressor running at maximum load, which consumes 23288 kWh with a 217–234 MMSCFD during peak and off-peak demand. It resulted in surging when the flow rate fell from 4167 to 2967 m³/h for variable speed and 4167–3906 m³/h for fixed speed. When a surge was about to occur, ASV opened aggressively at 14.79% with variable speed and 10.08% with fixed speed.

Keywords Predictive Maintenance (PdM) · Energy Efficiency (EE) · Condition Monitoring (CM) · Surge control line (SCL) · Surge limit line or backup line (SSL) · Aspen HYSYS · Anti surge control (ASC) · Reliability and Maintainability (R&M)

1 Introduction

The performance deviation of the compressor is typically limited by two phenomena on the performance curve. These are the Stonewall and Surge regions. Stonewall is the upper limit of the compressor flow, while surge is the lower boundary of a steady flow. There are different factors which resulted in a surge such as power failure, sudden load or speed variation, emergency shutdown, plant trips, suction filter chock, inter cooler leakage and discharge valve failure. Additionally, when compressor surges, it will produce a strong that shortens the compressor's lifespan and destroys shafts, bearings, and seals. Further, it reduces energy efficiency and results in unsteady flow and pressure [1]. For any industrial sector, saving energy is the most important matter, specifically for the oil and gas industry, and it assists to cut off both energy costs and environmental impacts. Monitoring techniques offer the possibility to raise energy efficiency, reduce running costs and decrease emissions when properly developed [2]. Typically, energy price dominates the life cost of centrifugal gas compressors operated in different municipal and industrial processes, developing the compressor smart objective for improving efficiency of energy [3]. A centrifugal compressor's typical function is to compress gas or air at a higher pressure. Centrifugal compressor performance breakdown and deterioration can have a bad effect on the operation, energy and profit of an enterprise or business that depends on continuous production. Compared to axial compressors, the advantages of centrifugal compressors are mostly coupled to robustness, a broad operational scale and comparatively minimum investment and maintenance costs [4]. Typically, energy price affects the life cycle costs, while considering the total lifespan costs of compressed-air equipment (LCC). Up to 78% of the overall costs of life can be the cost of energy as shown in Fig. 1. Compressor energy efficiency assessment is a source for saving energy and reducing carbon emissions, it is also an impartial process of impact evaluation of the compressor set's actual operating efficiency. The producer or a third party may evaluate their true performance in different operating conditions

Fig. 1 Life cycle cost of dynamic compressor [5]



with different working fluids for general power and refrigeration compressors through set up test rigs for smaller compressor power, simpler operation condition adjustment and easier gas component [5, 6].

Energy savings can be achieved using energy-efficient technologies, improvement of operations and efficient maintenance. A good strategy for maintenance could improve energy efficiency, improve reliability, and decrease risk [6]. The main trial of maintenance optimization is to devise an appropriate maintenance approach that increases equipment obtainability and efficiency, controls the degree of deterioration of equipment, guarantees safe and environment friendly operations, and reduces the entire cost of operation, which means equally production and energy costs [7]. As the oil and gas industry faces growing pressure to reduce the emissions and environmental impacts of its operations worldwide, the use of AI in energy efficiency and power quality is becoming increasingly important. Analytics-driven machine learning-supported AI algorithms can discern patterns in data that can signal opportunities in the afore mentioned assets to optimize energy utilization. AI already offers powerful strategies for transforming human knowledge into software. Energy efficiency analysis (EEA) and energy quality analysis (PQA) provide operators with both savings and a competitive edge through energy optimization [8]. Maintenance optimization and operational processes, even if they are rarely specifically considered in the energy-saving measures' predictive models, may deliver margins to improve energy efficiency in the industry of oil and gas. To ensure that the plant achieves the required efficiency, it is important to track operation of maintenance and maintenance outcomes and the connection amongst the inputs and the output of the process of maintenance [9]. In this paper, the authors proposing methodology in which the design and original equipment manufacturer (OEM) data for the compressor were reviewed, and then the operating envelope was examined. Aspen HYSYS version 12.1 is used for the dynamic simulation modelling to analyze the off-design performance deviation analysis in the form of surge prediction and energy performance against designed performance curves provided by the original equipment manufacturer (OEM) and model is validated with designed data provided by OEM.

2 Methodology

This study was carried out on PETRONAS Angsi two-stage centrifugal compressor powered by a gas turbine that compresses production gas for export. The unit has an inter coolers and interstage scrubbers, K-2420/2520 which are low pressure (1st stage) and high pressure (2nd stage) designed for 100 MMSCFD, and 225 MMSCFD. Before setting up a centrifugal compressor steady-state simulation model in Aspen HYSYS, process components in the form of gas mix composition and thermodynamic conditions were incorporated with a Peng-Robinson fluid package. The processes that take place in the HYSYS simulation adhere to mass balance, energy balance, and thermodynamic rules. Data from Petronas Angsi for both stages were then incorporated into this first model to generate dynamic simulation. The monitoring tool may simulate a centrifugal compressor, accept inputs on gas compositions, and set the Peng-Robinson as the EOS for calculating gas properties. Performance curve data as flow vs head, surge flow and stonewall curve from the vendor, suction, and discharge parameters (pressure, temperature, and flows), compressor operating speed, and power are the measured parameters (actual) used as inputs.

3 Results and Discussion

The performance of the gas compression system's energy efficiency optimization and surge prediction has been examined using the Aspen HYSYS Dynamics model for the centrifugal gas compressor. The polytropic head vs volume flow rate of the two compressor stages was the sole performance map that was available. The centrifugal gas compressor operated adjacent 10%, surge control and Safety line. The off-design performance deviation and energy efficiency performance optimization, and surge prediction for fixed speed and variable speed during peak and off-peak demands were then analyzed, compared with actual real-time operating conditions. The results showed that speed variation can precise the anti-surge control line to meet the system flow-head requirements for reliable surge control to adjust the operating point on compressor map, and eventually increase the overall system efficiency and lower the energy consumption. Two dynamic simulation cases A and B have been established for both stages as a result. Which are: Case A: Energy efficiency performance optimization and surge prediction with 10% surge control/safety line for fixed speed/load and variable speed/load for low pressure region (1st stage) and Case B: Energy efficiency performance optimization and surge prediction with 10% surge control/Safety line for fixed speed/load and variable speed/Load for high pressure region (2nd stage).

Case A: Energy efficiency performance optimization and surge prediction with 10% surge control/safety line for fixed speed/load and variable speed/load for low pressure region (1st stage)

The surge prediction with 10% surge control/Safety line for fixed speed/load and variable speed/load for low pressure region (1st stage) was analyzed, compared, and actual real-time operating conditions were plotted with the designed performance curve that was modelled with Aspen HYSYS based on the results of performance deviation and energy efficiency performance optimization. Figure 2a, c display the performance deviation of fixed speed, and (c, d) shows the variable speed performance deviation, energy or power consumption, and surge prediction results for peak and off-peak demand when operating at lowest and maximum operating speeds.

Case B: Energy efficiency performance optimization and surge prediction with 10% surge control/Safety line for fixed speed/load and variable speed/Load for high pressure region (2nd stage)

The surge prediction with 10% surge control/safety line for fixed speed/load and variable speed/load for low pressure region (2nd stage) was analyzed, compared, and

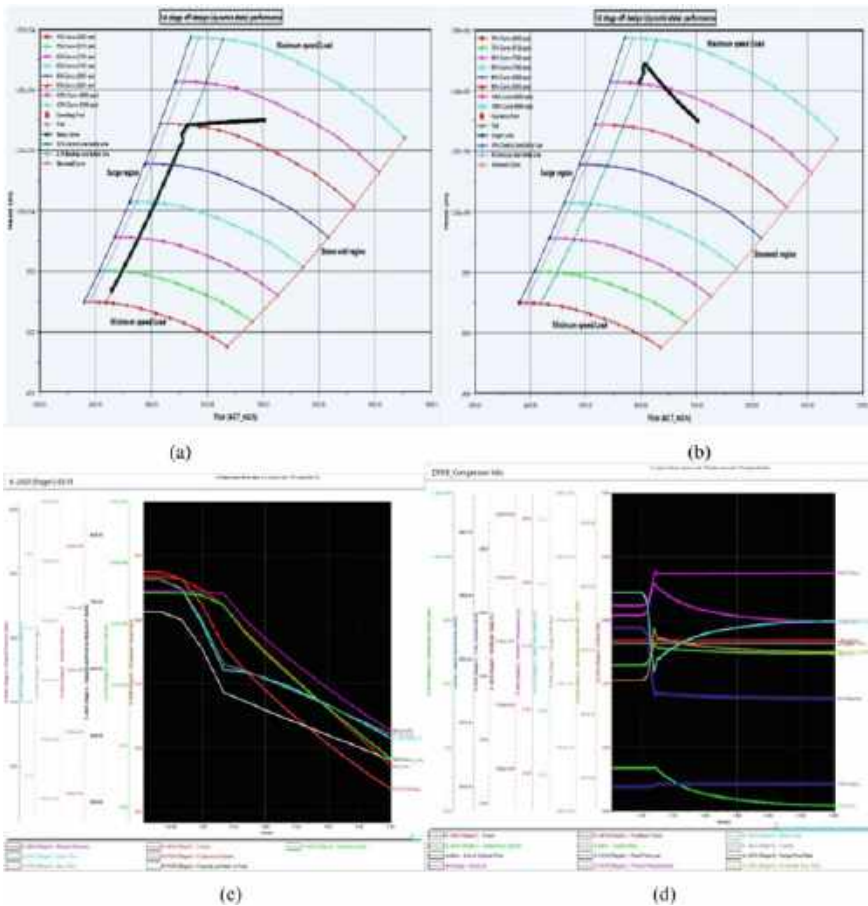


Fig. 2 1st stage a, c off design variable versus b, d fixed speed performance results

actual real-time operating conditions were plotted with the designed performance curve that was modelled with Aspen HYSYS based on the results of performance deviation and energy efficiency performance optimization. Figure 3a, c display the fixed speed, and (c, d) shows the variable speed performance deviation, energy or power consumption, and surge prediction results for peak and off-peak demand when operating at lowest and maximum operating speed.

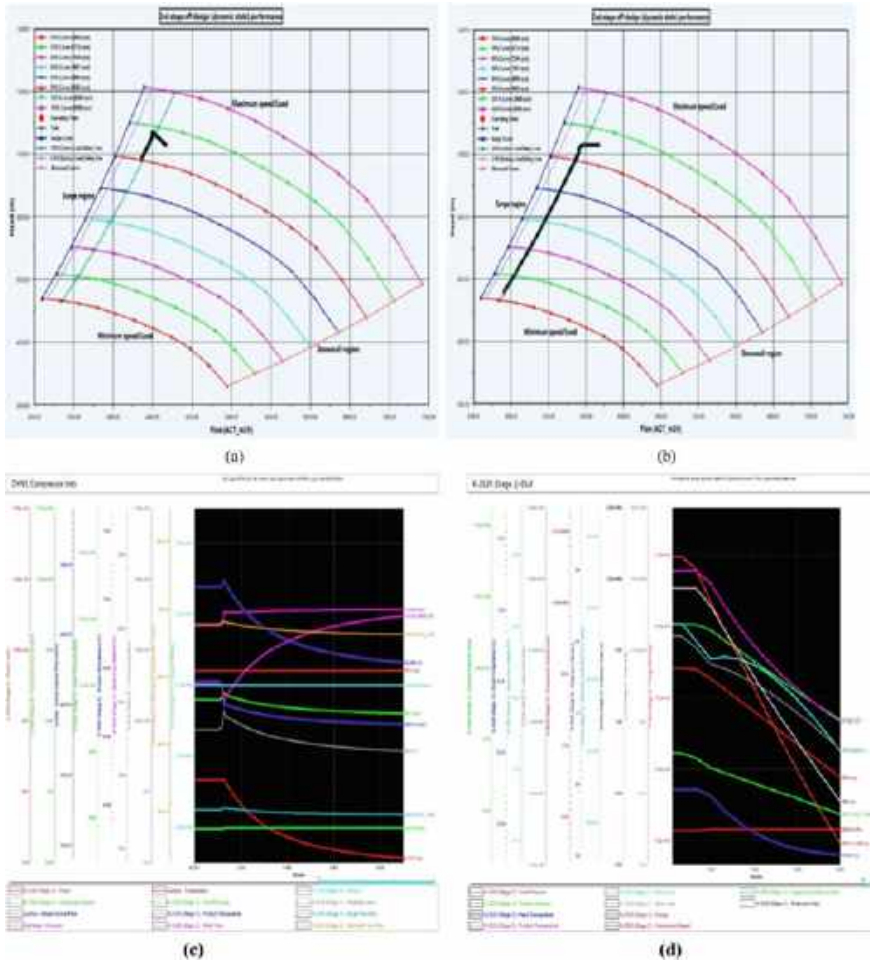


Fig. 3 2nd stage a, c off design fixed versus b, d variable speed performance

4 Conclusion

Findings showed that speed variation can precise the anti-surge control line to meet the system flow-head requirements for reliable surge control to adjust the operating point on compressor map, and eventually increase the overall system efficiency and lower the power or energy consumption. Aspen HYSYS version 12.1 is used for the dynamic simulation modelling to analyze the off-design performance deviation analysis in the form of surge prediction and energy efficiency against designed performance curves provided by the original equipment manufacturer (OEM) and model is validated with designed data provided by OEM. The model offers a dynamic baseline to forecast compressor performance under various inlet gas thermodynamic circumstances, and it can be an extremely helpful monitoring tool. The compressor power consumption, compressor speed, and compressor capacity are important indicators for monitoring the energy efficiency performance of centrifugal gas compressor. The main findings from the results are given below:

1. According to the first stage findings with 10% surge safety, the load variation control approach is more energy-efficient than compressor operation at maximum load, which consumes 13596 kWh with 99–94 MMSCFD during peak and off-peak demand and uses less energy 5183 kWh for 99–67.59 MMSCFD compression. Surging occurred when the flow rate decreased from 7022 to 4355 m³/h for variable speed and from 7022 to 5990 m³/h for fixed speed. When a surge was likely to occur, ASV opened aggressively at 22.29% with variable speed and 11.77% with fixed speed.
2. Based on the results of the second stage with 10% safety, the load variation control method is more energy efficient, consuming between 23,288 and 9892 kWh with a 217 and 178 MMSCFD compressor capacity, as opposed to the compressor running at maximum load, which consumes 23288 kWh with a 217–234 MMSCFD during peak and off-peak demand. It resulted in surging when the flow rate fell from 4167 to 2967 m³/h for variable speed and 4167 to 3906 m³/h for fixed speed. When a surge was about to occur, ASV opened aggressively at 14.79% with variable speed and 10.08% with fixed speed.

Acknowledgements The authors wishes to acknowledge his Supervisor Dr Masdi B Muhammad, Co-Supervisor Ainul Akmar Binti Mokhtar, and the University Technology Petronas for providing an opportunity to work in ideal research environment.

References

1. Hansen C Dynamic simulation of compressor control system. Thesis submitted by Aalborg University Esbjerg
2. Baldi S, Le Quang T, Holub O, Endel P (2017) Real-time monitoring energy efficiency and performance degradation of condensing boilers. *Energy Conver Manage* 136:329–339

3. Viholainen J, Grönman K, Jaatinen-Värri A, Grönman A, Ukkonen P (2015) Centrifugal compressor efficiency improvement and its environmental impact in wastewater treatment. *Energy Conver Manage* 101:336–342
4. Schiffmann J, Favrat D (2010) Design, experimental investigation, and multi-objective optimization of a small-scale radial compressor for heat pump applications. *Energy* 35(2010):436–450
5. Saidur R, Rahim NA, Hasanuzzaman M (2010) A review on compressed-air energy use and energy savings. *Renew Sustain Energy Rev* 14(2010):1135–1153
6. Li L et al (2017) Technology's present situation and the development prospects of energy efficiency monitoring as well as performance testing & analysis for process flow compressors. In: *IOP conference series: materials science and engineering*, vol 232
7. Firdaus N, Samat HA, Mohamad N (2019) Maintenance for energy efficiency: a review. In: *IOP conference series: materials science and engineering*, vol 530
8. Darabnia B, Demichela M (2013) Data field for decision making in maintenance optimization: an opportunity for energy saving. *J Chem Eng* 33
9. Blumenthal R, Siemens AG, El Naser A (2020) Siemens LLC middle east; Christian Blug, Siemens AGA, generating green value from data: applying ai-based analytics to monitor and manage energy usage across oil and gas operations. *Soc Petrol Eng*

A Novel Fuzzy PID Control Algorithm of BLDC Motor



Zhou Hongqiang  and Dahaman Ishak 

Abstract Brushless DC motors (BLDC) have many advantages, such as simple structure and high reliability, and are now widely used in various fields. However, the nonlinear characteristics of the BLDC motor make it difficult for the traditional PID controller to meet the requirements of the motor control system. In this paper, a fuzzy PID-based control method is proposed to optimize the conventional PID control by intelligent fine-tuning of the control parameters. To verify the effectiveness of the proposed method, simulation experiments are carried out under various operating conditions using MATLAB/Simulink. The results of the simulations show a significant improvement in the control performance, with a reduction in the final rise time (t_r) by approximately 0.002 S, a significant reduction in the peak time (t_p), and the current can be smoothed out much faster. The results show that the motor control system designed in this paper has good responsiveness and robustness and high application value.

Keywords BLDC · Fuzzy PID · Speed control

1 Introduction

As a new-generation motor in the field of electric motors, the BLDC motor is characterized by its simple structure and easy maintenance [1]. It has been used in various fields such as household appliances and automotive electronics and has a wide application background [2]. However, due to the nonlinearity of the system, it is difficult to describe it with an accurate system model. At the same time, as the motor speed often changes with the change of working conditions, this puts forward higher requirements

Z. Hongqiang · D. Ishak (✉)

School of Electrical and Electronic Engineering, Universiti Sains Malaysia, 14300 Nibong Tebal, Malaysia

e-mail: dahaman@usm.my

Z. Hongqiang

e-mail: zhouhongqiang@student.usm.my

on the responsiveness, speed, and robustness of motor speed control, which cannot be met by traditional PID control [3].

In this paper, a fuzzy PID-based control method is proposed to optimize the traditional PID control [4], improve the stability and tracking accuracy of the controller, and enhance the robustness of the system [5]. To study the performance of BLDC motors under the fuzzy control strategy, a simulation model of fuzzy PI improved control is built in MATLAB Simulink and experimentally verified. The speed and current are compared with the conventional PI control. The response speed and stability of the motor are analyzed, and the rationality of the control method is verified.

2 Mathematical Model of BLDC

The structure of a three-phase BLDC motor consists of two main parts, the permanent magnet rotor, and the stator winding. The armature winding can be connected in two ways, either in a star connection or in a delta connection [6]. According to the operating principle of BLDC motors, the electric potential and current of the windings need to be known to maximize the motor torque. In order to simplify the problem, the windings are set up as three symmetrical and saturation is not accounted for. They can be expressed as follows:

$$\left. \begin{aligned} u_a &= Ri_a + (L - M)\frac{di_a}{dt} + e_a = Ri_a + L_m\frac{di_a}{dt} + e_a \\ u_b &= Ri_b + (L - M)\frac{di_b}{dt} + e_b = Ri_b + L_m\frac{di_b}{dt} + e_b \\ u_c &= Ri_c + (L - M)\frac{di_c}{dt} + e_c = Ri_c + L_m\frac{di_c}{dt} + e_c \end{aligned} \right\}, \quad (1)$$

where u_x is the stator voltage; R is the stator resistance; i_x is the stator current; e_x is the back-EMF; L and M are the self and mutual-inductances respectively; L_m is equal to $L - M$. This gives the equivalent circuit for the BLDC motor, as shown in Fig. 1.

The instantaneous electromagnetic torque T_e of the BLDC motor is:

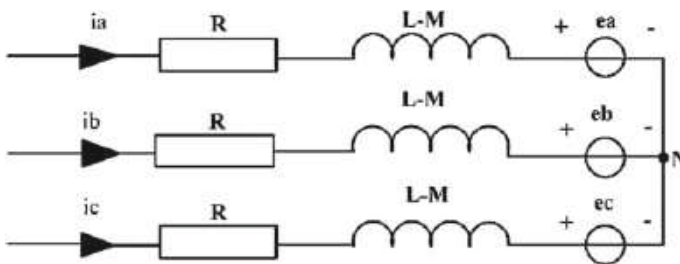


Fig. 1 Equivalent circuits

$$T_e(t) = \frac{1}{\omega_r} (e_a i_a + e_b i_b + e_c i_c), \tag{2}$$

where T_e is the electromagnetic torque; ω_r is the motor speed.

The average torque is:

$$T = \frac{1}{\omega_r t_n} \int_0^{t_n} T_e(t) dt, \tag{3}$$

where t_n is the electrical period. The equation of motion is:

$$T_e - T_L - B\omega_r = J \frac{d\omega_r}{dt}, \tag{4}$$

where T_L is the load torque; B is the damping coefficient; J is the rotational inertia.

3 The Fuzzy Controller Design

For BLDC motor speed control system, which is based on the traditional PI control, when the system is disturbed, the PI control cannot guarantee the stability of the system. Therefore, this paper proposes a Fuzzy-PI controller with a double closed-loop feedback control to improve the motor drive system i.e., the current feedback as the inner loop, and the speed feedback as the outer loop. Both loops are used in the Fuzzy adaptive PI control to achieve the targeted speed and torque under various operating conditions and disturbances. The control block diagram is shown in Fig. 2.

Both e and ec inputs are selected for the Fuzzy controller in this design, as shown in Fig. 3a, which is applied to the speed PI controller and current PI controller respectively. The Fuzzy controller is a two-dimensional input, ΔK_p and ΔK_i is taken to be the Fuzzy output variable, at which point the output parameters of the

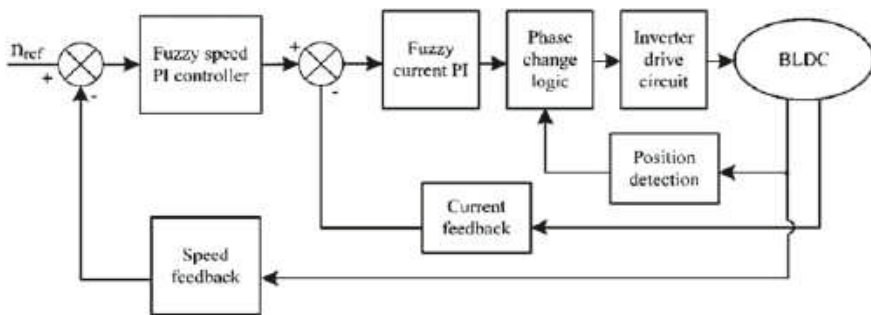


Fig. 2 BLDC double closed-loop fuzzy adaptive PI control block diagram

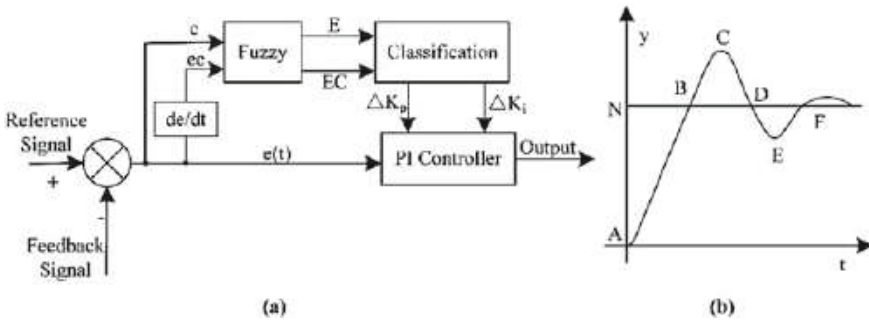


Fig. 3 **a** Fuzzy controller block diagram and **b** controller dynamic response curves

Fuzzy PI are given in Eq. 5.

$$K_p = K_{p0} + \Delta K_p, K_i = K_{i0} + \Delta K_i \tag{5}$$

where e is the input error; ec is the rate change of input error.

The dynamic response curve of the controller is shown in Fig. 3b, so that N is the given value and y is the output value of the controlled object. Then the curve is divided into five sections by the given value, the five sections are named $AB, BC, CD, DE,$ and EF in order. K_p, K_i how to choose, is divided according to these five stages, as shown in Table 1. The Fuzzy rules are then formulated by selecting a variable under each input or output quantity, which is eventually combined to form a Fuzzy rule.

A Fuzzy adaptive PI model and a simulation model were developed using MATLAB/Simulink. As applied to speed control, the actual speed is subtracted from the given speed, i.e., the reference value of the speed, to obtain a signal that is output through the Fuzzy control module with a current value that is output after limiting the reference value. Applied to the Fuzzy PI current controller, the control time is shortened, and the current is maximized. After the difference between the reference current and the actual current obtained from the speed loop, the output voltage value is controlled using the Fuzzy PI and given to the phase change logic circuit after limiting the reference voltage of the control system, as shown in Fig. 2. By

Table 1 K_p, K_i select table

Section	e	ec	K_p	K_i
AB	$e < 0$	$ec > 0$	-	+
BC	$e > 0$	$ec > 0$	+	-
CD	$e > 0$	$ec < 0$	-	+
DE	$e < 0$	$ec < 0$	+	-
EF	$e < 0$	$ec > 0$	+	-

introducing the Fuzzy PI double closed-loop control, the stability of the system is improved.

4 Simulation and Result

The design was verified by simulation in MATLAB/Simulink. The simulation parameters for the BLDC motor: rated speed $n = 2000$ rpm, rated voltage $U_d = 500$ V, stator resistance $R = 2.875 \Omega$, stator phase winding self-inductance $L = 0.0085H$, mutual inductance $M = 0.0005H$, number of pole pairs $p = 4$. The motor is running at no load, given a speed of $n = 2000$ rpm, and then applying a load of $3 Nm$ at 0.15 s.

As can be seen from Fig. 4c, the speed response of Fuzzy PID peaks in 0.0013 s with the maximum speed of 2024 rpm, and then reaches a steady state speed of 2000 rpm without any ripple. While the speed response of conventional PI has a large overshoot and many speed oscillations around 2000 rpm. The load is applied at 0.15 s where the speed oscillation appears to indicate the disturbance. The Fuzzy adaptive PI control shows robustness in achieving the targeted speed under load disturbance since the speed ripple is very small. Whereas the conventional PI control suffers a significant speed drop. The rise time t_r of the double Fuzzy PI control is shorter than traditional PI control, while the peak time t_p of the Fuzzy PI control is also shorter and the time required to reach a steady state is shorter. This shows that the dynamic performance of the Fuzzy adaptive PI control is better than the traditional PI control.

The single-phase current waveform at no load is shown in Fig. 4a. The proposed algorithm will stabilize first. Adding the load at time 0.15 s, the torque and back-emf do not differ much, with only slight differences in amplitude, as shown in Fig. 4b and d.

5 Conclusion

For a multi-variable, strongly coupled system like BLDC motor, this paper designs a new Fuzzy PID-based control method to achieve optimization on the basis of traditional PID control. The stability and tracking accuracy of the controller is improved to meet the requirements of the motor drive system for accurate speed and robustness. The control algorithm is simulated by Simulink and validated under various operating. The results prove the method's superiority and have good prospects for engineering applications.

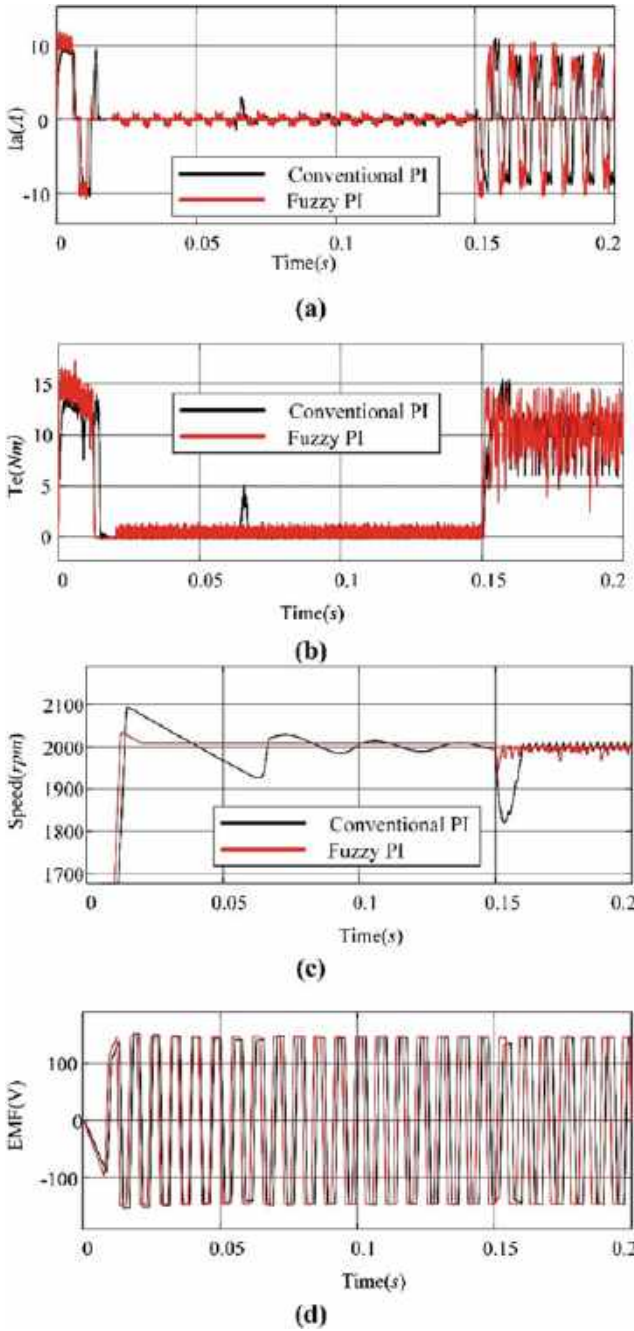


Fig. 4 The responses of the different algorithms. **a** The stator current I_a ; **b** the torque; **c** the speed response; **d** the phase back-EMF

References

1. Kim N-H, Yang O, Kim M-H (2007) BLDC motor control algorithm for industrial applications using a general purpose processor. *J Power Electr* 7(2):132–139
2. Ansari U, Alam S (2011) Modeling and control of three phase BLDC motor using PID with genetic algorithm. In: 2011 UkSim 13th international conference on computer modelling and simulation. IEEE, pp 189–194
3. Kommula BN, Kota VR (2022) Design of MFA-PSO based fractional order PID controller for effective torque controlled BLDC motor. *Sustain Energy Technol Assess* 49:101644
4. Mohanraj D, Arul david R, Verma R, Sathyasekar K, Barnawi AB, Chokkalingam B, Mihet-Popa L (2022) A review of BLDC motor: state of art, advanced control techniques, and applications. *IEEE Access*
5. Kommula BN, Kota VR (2022) An integrated converter topology for torque ripple minimization in BLDC motor using an ITSA technique. *J Amb Intell Hum Comput* 1–20
6. Hidayat N, Samman FA, Sadjad RS (2022) FPGA based controller of BLDC motor using trapezoid control. In: 2022 14th International conference on information technology and electrical engineering (ICITEE). IEEE, pp 58–63

Optimized Design Point Model of SGT500 Using GasTurb 14



Mahnour Soomro, Tamiru Alemu Lemma, Syed Ihtsham Ul-Haq Gilani, and Mukhtiar Ali Shar

Abstract This research paper presents design point simulation of three shaft industrial gas turbine i.e. SGT500 using a combination of random and gradient optimization algorithm. The modeling of design point of gas turbine is crucial initial step as it will help in simulating the complex step of off-design model. Hence, accuracy of the gas turbine model at design point will influence the rest of the model and its forthcoming results. The proposed algorithm suggest problem formulation of the multi-objective figure of merit that yields most accurate results. By using GasTurb14 software the SGT500 engine will be adopted to know all of its unknown design variables along all the stations. The design point is selected at the optimum operating point of the gas turbine which is 19.1 MW. The targeted performance variables taken are heat rate, thermal efficiency, power output, exhaust gas temperature, exhaust mass flow, and pressure ratio. The results are validated and show average accuracy of more than 99.9%.

Keywords Gas turbine · Design point · Optimization

1 Introduction

In many different sectors, gas turbines are used to drive a variety of loads, including generators, pumps, compressors, and propellers [1]. Currently, the gas turbines performance is an area of concern for many researchers as the technology is striving to get better efficiencies without compromising performance. Hence, multiple studies have been conducted to improve the gas turbines (GTs) flexibility and operability with different types of fuels to reduce emission rate and get higher performance. Also, GTs has the benefit that they provide large inertia and because GTs have a considerable rotating inertia, their frequency fluctuations are generally low [2].

M. Soomro · T. A. Lemma (✉) · S. I. U.-H. Gilani · M. A. Shar
Department of Mechanical Engineering, Universiti Teknologi Petronas, Perak, Malaysia
e-mail: tamiru.lemma@utp.edu.my

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
N. S. Ahmad et al. (eds.), *Proceedings of the 12th International Conference on Robotics, Vision, Signal Processing and Power Applications*, Lecture Notes in Electrical Engineering 1123, https://doi.org/10.1007/978-981-99-9005-4_10

Provided that, most of the research objectives require to have the specific model parameters and design values. Unfortunately, due to limited data from manufacturer [3] and many unknown design variables at hand to deal with, reaching to maximum accurate design point performance become complex.

Undoubtedly, accuracy of design point model is important, firstly, because adoption of already existing gas turbine to a new kind of modifications require full understanding of its performance behavior. Secondly, because, an inappropriate design can lead to conclusions that can cause substantial loss to environment and also it will be unable to serve the purpose of increased performance [4].

In this paper, SGT500 gas turbine is considered for modeling using well-defined objective function simulated in GasTurb14 software. The main objective is to reach the optimized design point with minimum possible error deviation.

2 Design Requirement

In this section, design point model of SGT500 will be developed using the available data as given in Table 1. The specific engine has rated power output of 19.1 MW at standard ISO conditions using natural gas with lower heating value (LHV) 46.798 MJ/kg. The SGT500 gas turbine has two stages of compressors high pressure compressor (HPC) and low pressure compressor (LPC) and two stages of turbine high pressure turbine (HPT) and low pressure turbine (LPT) with an independent power turbine (PT) with an overall pressure ratio of 13. The engine specification is shown in Fig. 1.

2.1 Optimization Problem

The target model to achieve need to be simulated in a way that it fulfills the performance characteristics. This problems can be expressed as multi-objective unconstrained function that aims to satisfy all objectives. However, objectives can be in conflict with each other and one objective may need to be minimized and other may

Table 1 Design point input [5, 6]

Parameter name	Symbol	Unit	Value
Ambient temperature	T_{amb}	K	288.15
Ambient Pressure	P_{amb}	kPa	101.325
Relative humidity	RH		60%
LPC stages	N_{stg}	–	10
HPC stages	N_{stg}	–	08
Turbine inlet temperature	TIT	K	1123
LPT stages	N_{stg}	–	2
HPT stages	N_{stg}	–	3

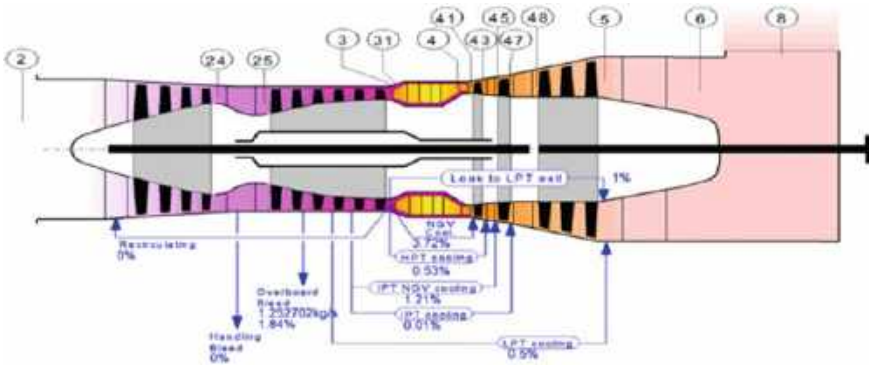


Fig. 1 Configuration of model gas turbine (SGT500)

need to be maximized, hence, a composed value is used that define all the targeted performance characteristics as one objective function in error function.

$$\text{minimize}_{x \in \mathbb{R}^n} f(x) = \sum_{i=1}^n \left(1 - \frac{\phi_i(x)}{\phi_{i,d}} \right)^2, i = 1, 2, \dots, n \tag{1}$$

ϕ_i represent simulated values and ϕ_i , represent target values, Optimize design variables x so that,

$$\text{Objective Function} = f(x) = 0 \tag{2}$$

where,

$$x = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}]^T \tag{3}$$

In normal practice of designing similar engines some parameters are given more weightage and considered as objective functions while others are considered to be constraints. In this specific case, all the performance characteristics parameters are given equal weightage in the objective function which is defined as an unconstrained optimization problem stated above. Whereas, the performance characteristics are defined in Table 2.

2.2 Optimization Strategy

In order to reach the optimized design point with minimum error value, two of the optimization strategies are used alternatively. The detail steps are as follows:

Table 2 Performance characteristics [7]

Performance variable	Unit	Target value
Thermal efficiency	%	33.8
Power output	MW	19.1
Pressure ratio	–	13
Exhaust temperature	K	642.14
Exhaust mass flow	Kg/s	97.9
Heat rate	kJ/kWh	10,664

1. From random search iteration (from 1 to N), initialize solution with random value of x_k within the initial range for each design variable, $x \in R$
2. Evaluate the function $f(x_k)$, if $f(x_k) < f(x)$, update x with x_k
3. Update the search space V_{k+1} based on current and previous points
4. Update the design variable x_{k+1}
5. Evaluate function $f(x_{k+1})$, if $f(x_{k+1}) < f(x_k)$, update x_k with x_{k+1}
6. Repeat until maximum iteration reached or when the result is same as previous.
7. After the adaptive random search stops, switch to gradient search
8. For each gradient search iteration (from 1 to M)
9. Start calculating $\nabla f(x)$ where the adaptive search ends.
10. Minimize the function by taking descent direction by subtracting the current solution from the product of learning rate and gradient:

$$x_{k+n} = x_n - \alpha \nabla f(x) \quad (4)$$

11. Repeat gradient search until the search steps becomes very small.
12. Then again switch to random search strategy.
13. This can continue until both search strategies give same values and the optimization point is reached.

3 Results

In order to analyze the optimization synthesis, two design variables are selected for parametric study. Figure 2a and b indicate that the value of x_1 and x_2 are iterated in such a way that $\phi_1(x_1, x_2)$ and $\phi_2(x_1, x_2)$ is nearly equal to 0.338 and 10664 respectively and the function $f(x)$ is approaching zero.

The OEM data from Table 1 was used as input to give the software interface to initiate the preliminary calculation. The data that is unknown in the input phase are set to default values used in the software. The main objective to reach the design point where most of the performance parameters behave around the target engine is achieved with more than 99.9% of confidence level. Table 3 shows catalog data versus GasTurb data.

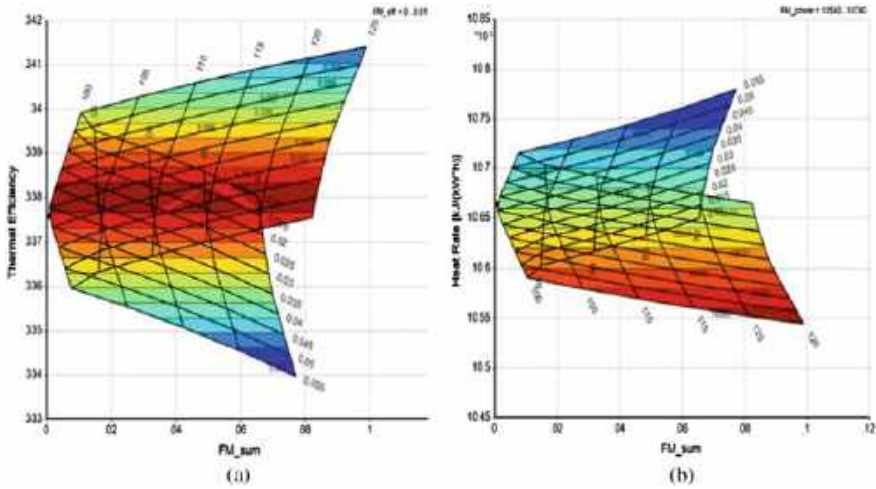


Fig. 2 Parametric study **a)** function versus thermal efficiency **b)** function versus heat rate

Table 3 Design point model validation

Parameter	Catalog data	GasTurb data	Units	Percent error (%)
Heat rate	10,664	10,664	kJ/kWh	0
Power output	19,100	19,100	kW	0
Pressure ratio	13	13.00052	–	0.004
Exhaust mass flow	97.9	97.9	kg/s	0
Exhaust gas temperature	642.15	641.21	K	0.1
Thermal efficiency	33.8	33.758	%	0.1

The model performance parameters are available with the SGT500 catalog data. There is no error in the values of heat rate, power output and exhaust mass flow. However, the other parameters have very negligible discrepancies that can be ignored. The main objective of adapting an existing gas turbine engine with minimum available data is to know the behavior of the engine at every stage. The engine performance mainly depends on pressure, temperature and mass flow at each stage. Figure 3 gives the data of each station.

The objective function in the optimization problem as defined above is minimized to 0.000455672 value which shows that on average the model has reached the target engine with 0.000455672 discrepancy overall.

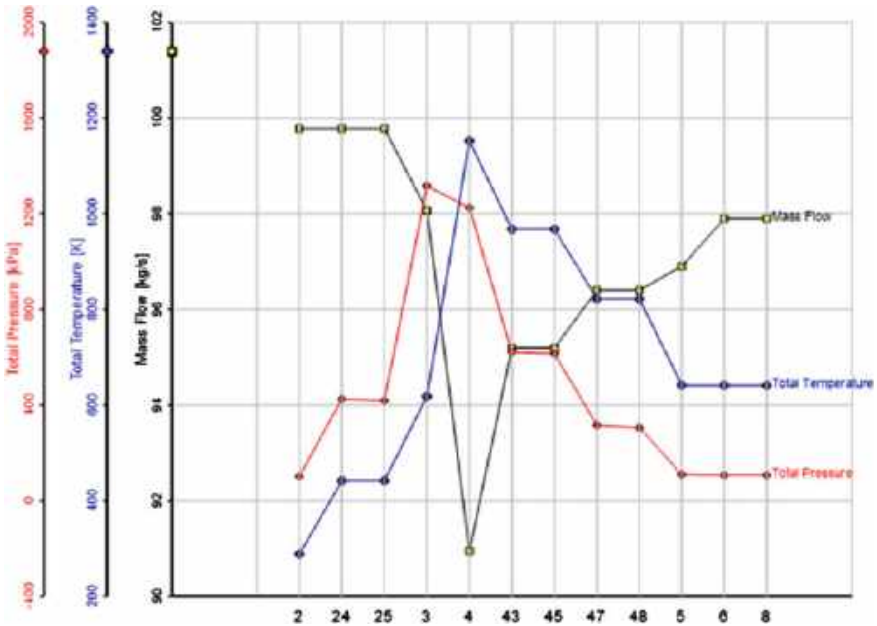


Fig. 3 Station data of mass flow rate, total pressure and total temperature

4 Conclusion

This paper discussed the method to adopt any gas turbine to its optimized design point. The method in simulation uses two alternative optimization algorithm; gradient and adaptive random search. The optimization problem is defined as a combination of target engine performance characteristics and in this way all the target parameters are given equal weightage. This method can be helpful in solving such optimization problems which have multiple conflicting objective. The method is applied to gas turbine engine SGT500 and the developed model is achieved and validated as well with very negligible difference from the target engine.

Acknowledgements Authors are grateful to Universiti Teknologi PETRONAS for providing the resources highly sought for the research.

References

1. Razak AMY (2007) Industrial gas turbines: performance and operability. Elsevier
2. Park Y-K, Moon S-W, Kim T-S (2021) Advanced control to improve the ramp-rate of a gas turbine: optimization of control schedule. Energies 14(23):8024

3. Pathirathna KAB (2013) Gas turbine thermodynamic and performance analysis methods using available data—faculty of engineering and sustainable development, department of building, energy and environmental engineering (Master's thesis, University of Gävle)
4. Schobeiri MT (2018) Blade design. In: Gas turbine design, components and system design integration. Springer, Cham, pp 249–276
5. Pierobon L, Iyengar K, Breuhaus P et al (2014) Dynamic performance of power generation systems for off-shore oil and gas platforms. In: Proceedings of ASME turbo expo 2014: turbine technical conference and exposition—GT2014. Düsseldorf, Germany
6. Benato A, Kaern MR, Pierobon L, Stoppato A, Haglind F (2015) Analysis of hot spots in boilers of organic Rankine cycle units during transient operation. *Appl Energy* 151:119–131
7. Siemens Gas Turbines (SGT) SGT-500 Industrial Gas Turbine Power Generation: (ISO) 19MW(e) baseload

Optimal RE-DG and Capacitor Placement for Cost-Benefit Maximization in Malaysia Distribution System



Maizatul Shafiqah Sharul Anuar, Mohd Nabil Muhtazaruddin,
Mohd Azizi Abdul Rahman, and Mohd Effendi Amran

Abstract Several policies have been developed to encourage the development of renewable energy sources in order to hasten the transition to a more sustainable energy system in response to the urgent environmental concerns the world faces today. Integrating renewable distributed generation (RE-DG) into the distribution network can achieve a more resilient and sustainable energy system. The system's efficiency can be increased by integrating RE-DG and capacitor into the distribution network since it can lower energy losses, enhance the voltage profile and provide economic benefit. This paper presents an optimal allocation and sizing of RE-DG and capacitor by using Artificial Bee Colony (ABC) algorithm. The proposed method has been implemented in selected public hospital distribution systems. The results show that installing RE-DG and capacitor at optimal location and capacity is important to achieve maximum power loss reduction and economic benefit in the distribution system.

Keywords Capacitor · Photovoltaic distributed generation (PV-DG) · Cost benefit

1 Introduction

Malaysia aims to use 31% renewable energy (RE) in its installed power capacity by 2025 and 40% by 2035 [1]. Hospitals are energy-intensive facilities that consume large amounts of electricity and produce significant carbon footprints due to their

M. S. S. Anuar · M. N. Muhtazaruddin (✉)

Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia

e-mail: mohdnabil.kl@utm.my

M. A. A. Rahman

Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia

M. E. Amran

Ministry of Health Malaysia, Putrajaya, Malaysia

energy consumption and reliance on fossil fuels. Malaysia's average yearly sun radiation of 1643 kWh/m² makes it ideal for producing solar PV electricity [2]. Adopting energy-efficient practices and utilizing RE sources like photovoltaic distributed generation (PV-DG) in the distribution system can aid in reducing power loss and energy costs over time. This paper will use the ABC algorithm to obtain optimal DG and capacitor location and size simultaneously. PV-DG will be considered due to Malaysia's high amount of solar irradiation throughout the year, and the economic analysis of PV-DG and capacitor placement is also considered. The selected public hospital is chosen as the test system.

2 Mathematical Modelling

2.1 Objective Function

When using optimization techniques to determine the ideal location for DG units and capacitor banks, selecting a suitable objective function is necessary. By simultaneously locating and sizing DG and capacitors in the radial distribution system (RDS), the current study's primary objective is to reduce the system's overall power losses while complying with all operational limitations. Therefore, the following definition of the goal function can be described as follows [3]:

$$P^{Tloss} = \sum_{j=0}^{N_{bus}-1} \frac{P_{j,j+1}^2 + Q_{j,j+1}^2}{|V_j|^2} R_{j,j+1} \quad (1)$$

$$f_1(x) = \min(P^{Tloss}) \quad (2)$$

P means active power loss, Q is the reactive power loss and j is the line section bus.

2.2 Constraints

The following are the problem's operational constraints [4]:

Power Balance Constraints

The size of the DG, capacitor, and overall amount of power coming from the slack bus must all be equal to the size of the load and total line losses.

DG Limit

The lowest and maximum DG sizes in this study are 0.30 MW and 3.00 MW, respectively.

Capacitor Limit

The generated reactive power should fall within 0.15 MVAR and 1.5 MVAR.
 Bus Voltage Constraints

The upper and lower limits for the bus voltages range from 0.90 p.u. to 1.05 p.u.

2.3 Cost Analysis

Based on the Malaysia electric tariff, the distribution network's total cost of power loss for the specified period is evaluated using (3).

$$C_{Loss} = E_C \times P_{T Loss} \times T \quad (3)$$

where C_{Loss} is cost of power loss in Malaysian Ringgit (RM), E_C is energy cost (RM/kWh) and T is time (8760 h). The above equation analysis (3) considers 0.365 RM/kWh energy cost for the existing loss.

3 Overview ABC Algorithm

A metaheuristic optimization technique, the Artificial Bee Colony (ABC) algorithm, was developed to solve optimization problems. It was introduced by Dervis Karaboga in 2005 [5]. The employed bee, the onlooker bee and the scout bee are the three sorts of bees that the ABC algorithm imitates. By balancing the exploitation of viable solutions and the exploration of new regions of the search space, the ABC algorithm combines exploration and exploitation tactics. In order to direct the search towards better solutions, it uses the bees' information-sharing capabilities. The mechanism of this bee can be described as follows [4]:

$$F_i = \frac{1}{(1 + OF_i)} \quad (4)$$

where OF_i is the objective function that symbolizes minimizing the overall line loss.

$$prob_i = \frac{F_i}{\sum_{i=1}^N F_i} \quad (5)$$

where N is the number of bees involved and $prob_i$ is probability.

$$x_{ij}^{new} = x_{ij}^{old} + range(0, 1) \times (x_{ij}^{old} - x_{kj}) \quad (6)$$

where the variables' new and previous values (DG and capacitor position or DG and capacitor size) are denoted by x_{ij}^{new} and x_{ij}^{old} respectively. The neighbour value, x_{kj} , was chosen at random.

4 Simulation Results and Discussion

This study was intended to examine the technical and economic impacts of RE-DG and capacitor integration on particular state-level public hospital distribution systems. The line diagram and data of the Malaysia public hospital can be obtained from [4]. The distribution system's power loss will be technically evaluated before and after integrating PV-DG and capacitors. The location and size of the PV-DG and capacitor on the networks were determined using the suggested algorithm. Four different case studies were taken into consideration in order to analyze the effects of RE-DG and capacitors. The total cost of energy (RM/year) has been determined for each situation. In each case, the total cost of energy (RM/year) has been calculated.

Case 1: Base case (without any DG or Capacitor).

Case 2: Only the capacitor is placed on the test system.

Case 3: Only PV-DG is placed on the test system.

Case 4: PV-DG and capacitor are placed simultaneously on the test system.

4.1 Distribution Network of Zone A

For case 1, the distribution system's overall line loss is measured without using a PV-DG or capacitor for comparison's sake. From the simulation result, the system's overall line loss is 349.7 kW, as shown in Table 1, or 3.063 MWh annually, corresponding to approximately 1,118,131 (RM/year).

In case 2, bus number three was chosen as the optimal location for capacitor placement with size 300 kVAR. The total line loss was reduced from 349.7 to 305.6 kW, about 12% from the initial line loss. After installing the capacitor, the overall energy loss cost dropped to 977,125 (RM/year), saving roughly 141,006 (RM/year).

In case 3, the simulation demonstrates a significant reduction in power losses to 69.7 kW, or roughly 80%, after integrating PV-DG. After using the suggested method for optimization, bus number three was chosen as the best site to install a 654 kW PV-DG. The total annual cost of energy loss has decreased to 222,859 (RM/year), representing an approximate 895,272 (RM/year) savings.

Similar to case 3, the initial line loss for case 4 was decreased to 69.7 kW, or nearly 80% of the original line loss. In this situation, the additional capacitor is likely installed in a redundant or insignificant location concerning the power loss. The system's current DG may already successfully reduce power loss. Thus, adding a capacitor offers no extra advantages. In these circumstances, the algorithm may converge to the same solution despite the insertion of a capacitor.

Table 1 Simulation outcome for all cases

Case	Zone A				Zone B			
	1	2	3	4	1	2	3	4
Total Power Loss (kW)	349.7	305.6	69.7	69.7	1466.3	1369.6	270.9	261.8
Power Loss Reduction (%)	–	12	80	80	–	7	82	82
DG Location and Size (kW)	–	–	3 (654)	3 (654)	–	–	7 (1587)	7 (1587)
Capacitor Location and Size (kVAR)	–	3 (300)	–	3 (812)	–	7 (475)	–	8 (300)
Cost of Energy Loss (RM/year)	1,118,131	977,125	222,859	222,859	4,688,348	4,379,159	866,175	837,079

Table 1 displays the outcomes of all runs. The results show that adding simply PV-DG or PV-DG and capacitor simultaneously to the test system resulted in the best power loss reduction, indirectly reducing the total yearly cost of line loss. The power loss convergence curve for case 2 to case 4 is shown in Fig. 1.

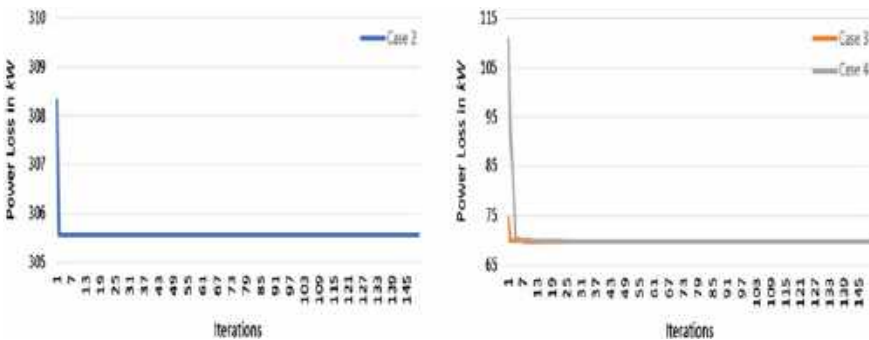


Fig. 1 Convergence curve of power loss in Zone A

4.2 Distribution Network of Zone B

For comparison's sake, in case 1, the distribution system's overall line loss is calculated without the aid of a PV-DG or capacitor. According to the simulation results, the system's overall line loss is 1466.3 kW, or 12.845 MWh per year, or approximately 1,118,131 (RM/year), as shown in Table 1.

Bus number seven was selected as the best option for installing a capacitor with a size of 475 kVAR in case 2. A 7% reduction in the overall line loss was made from 1466.3 kW to 1369.6 kW. The total energy loss cost decreased to 4,379,159 (RM/year) after installing the capacitor, saving around 309,189 (RM/year).

In case 3, the simulation shows that adding PV-DG has significantly reduced power losses to 270.9 kW, or nearly 82%. Following optimization using the provided method, bus number seven was selected as the ideal location for a 1587 kW PV-DG installation. There has been a reduction in the annual cost of energy loss to 866,175 (RM/year), saving almost 3,822,173 (RM/year).

For case 4, the initial loss was reduced to 261.8 kW or over 82% of the original loss. The yearly cost of energy loss has decreased to 837,079 (RM/year), saving nearly 3,851,269 (RM/year). Bus number seven has been selected as the best position for the installation of PV-DG, while bus number eight has been selected as the best location for the placement of capacitors with capacities of 1587 kW and 300kVAR, respectively.

The results for each case are presented in Table 1. The findings demonstrate that the test system's best power loss reduction came from simultaneously installing PV-DG and capacitors, indirectly decreasing the overall yearly cost of line loss. Figure 2 displays the power loss convergence curve for case 2 to case 4.

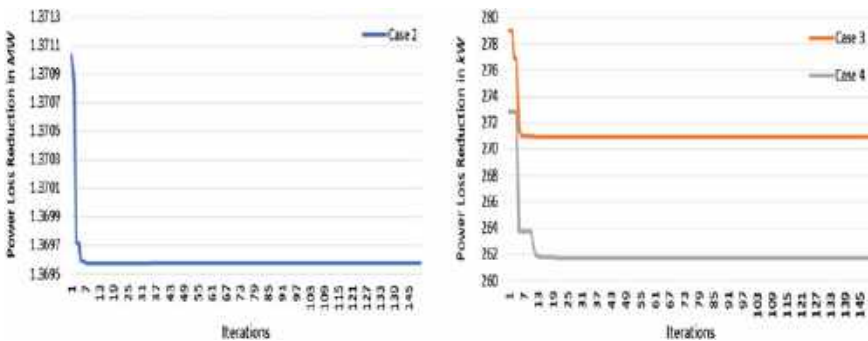


Fig. 2 Convergence curve of power loss in Zone B

5 Conclusion

The placement and size of the PV-DG and capacitor are allocated optimally in this study using the ABC method, which also lowers the cost of energy loss and minimizes power loss. The public hospital was selected as the test system in this study, which used four different scenarios to assess the impacts of PV-DG and capacitor integration on the distribution network. According to the simulation results, the simultaneous integration of PV-DG and capacitor reduces the maximum power loss, which indirectly lowers the cost of total energy loss.

Acknowledgements This work was funded by the Ministry of Higher Education under Fundamental Research Grant Schema (FRGS/1/2020/TK0/UTM/02/110).

References

1. Sustainable Energy Development Authority (SEDA) Malaysia (2021) Malaysia renewable energy roadmap pathway towards low carbon energy system. Putrajaya
2. Husain AAF, Ahmad Phesal MH, Kadir MZ, Ungku Amirulddin UA, Junaidi AHJ (2021) A decade of transitioning Malaysia toward a high-solar pv energy penetration nation. *Sustainability* 13(17)
3. Sambaiah KS, Jayabarathi T (2019) Optimal allocation of renewable distributed generation and capacitor banks in distribution systems using salp swarm algorithm. *Int J Renew Energy Res* 9(1):96–107
4. Amran ME, Muhtazaruddin MN, Bani NA, Kaidi HM, Hassan MZ, Sarip S et al (2019) Optimal distributed generation in green building assessment towards line loss reduction for Malaysian public hospital. *Bull Electr Eng Inform* 8(4):1180–1188
5. Karaboga D, Gorkemli B, Ozturk C, Karaboga N (2014) A comprehensive survey: artificial bee colony (ABC) algorithm and applications. *Artif Intell Rev* 42(1):21–57

Internal Discharge Patterns Identification of Void in High Voltage Solid Insulation Using Phase Resolved Method



N. Rosle, M. N. K. H. Rohani, N. A. Muhamad, S. A. Suandi,
and M. Kamarol

Abstract The occurrence of partial discharge in solid insulation indicates the deteriorating performance level of the high-voltage insulation system. The improvement process becomes complicated because no visual evidence of its existence in the system. Thus, one way to discover these possible insulation defects is through PD data representation. The phase-resolved PD pattern (PRPD) has become the most widely used tool to diagnose and visually represent PD data as well as discover possible insulation defects. Certain types of defects show characteristic clusters of partial discharges, which help to differentiate them. The observable parameters of PDs are crucial to relate to the characteristics of the PD defect in order to identify the type of defect and eventually ensure the reliable operation of HV equipment. This work aims to investigate the characteristics of internal discharges in solid insulation converted from the raw data to the PD patterns using PRPD. Two primary distributions which are $H_n(\varphi)$ and $H_{qn}(\varphi)$ results are presented to show the different asymmetry distribution gives different characteristics of PD defects.

Keywords Partial discharge · Solid insulation · Internal discharge · Phase-resolved

N. Rosle · N. A. Muhamad · S. A. Suandi · M. Kamarol (✉)
School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Engineering Campus,
14300 Nibong Tebal, Penang, Malaysia
e-mail: ekamarol@usm.my

N. Rosle · M. N. K. H. Rohani
Faculty of Electrical Engineering & Technology, Universiti Malaysia Perlis, 02600 Arau, Perlis,
Malaysia

N. A. Muhamad
Faculty of Engineering, Universiti Teknologi Brunei, Gadong, Brunei

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
N. S. Ahmad et al. (eds.), *Proceedings of the 12th International Conference on
Robotics, Vision, Signal Processing and Power Applications*, Lecture Notes in Electrical
Engineering 1123, https://doi.org/10.1007/978-981-99-9005-4_12

1 Introduction

Partial discharge (PD), as defined by IEC60270, refers to an electrically localized discharge that partially bridges the insulation between two conductors [1]. Generally, the focus of PDs lies in the dielectric materials employed and their role in partially connecting the electrodes across which the voltage is applied. Modern insulation systems can encompass several materials, including solids, liquids, gases, or a mix thereof. In the case of solid insulation, different sources of PDs such as internal discharge give different effects on insulation performance, especially during manufacture, installation, and operational use. In recent years, most researchers have made studies related to the extremely high failure rate of cables after only a few years in service. Different factors such as long-term operation, insufficient grounding distance, poor environmental quality for insulation aging, and insulation breakdown become the most popular reasons for the issues.

Aligned with technological progress, diverse approaches and methodologies have been suggested to define the parameters of PDs. Each form of PD defect exhibits its own distinct degradation traits [2–4]. One of the most widely used to visualize the PD activity relative to the 360° of an AC cycle is phase-resolved PD (PRPD) patterns [5]. PRPD aims to describe discharge patterns of unknown origin. Three primary parameters of PRPD patterns, which are computed within predefined time intervals across the entire 360-degree AC power cycle, include charge magnitude (q), phase angle (φ), and the total count of PD events (n) [6]. Establishing a connection between the observable characteristics of PDs and the parameters of the defect is vital for identifying the defect's type. This connection, in turn, contributes to enhancing the number of features and ensuring the dependable operation of high-voltage equipment [7]. Therefore, this research deals with the PRPD patterns in diagnosing and identifying the characteristics of the discharge in high voltage (HV) insulation systems.

2 Experimental Setup

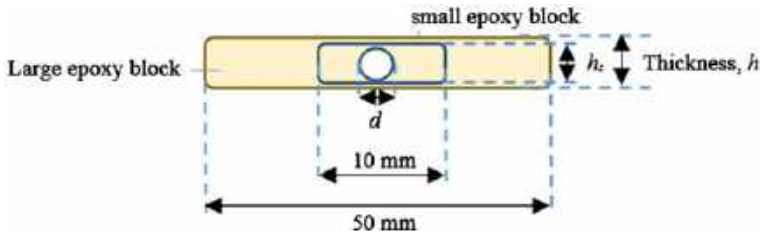
Two artificial samples with 0.5 mm sizes of the void are prepared to represent internal discharge in the solid insulator. Measurements were handled to obtain the PD signal and plotted into PRPD that represents the 1000 cycles of PD waveform. Statistical parameters are used to investigate the characteristics of internal discharges in terms of asymmetry distribution in PD patterns.

2.1 Sample Preparation

The dielectric material used in the experiment is an epoxy resin which is a highly cross-linked addition polymer. It is typically produced from a reaction between an

Table 1 Summary of sample prepared for measurements

Discharge type	Size of void, d (mm)	Thickness of small block, h_s (mm)	Thickness of sample, h (mm)
Internal discharge with air void	0.5	1.50	2.50
Internal discharge with water void	0.5	2.50	3.50

**Fig. 1** Schematic diagram of PD sample

alcohol or amine and an epoxide [8]. The PD samples were prepared as illustrated in Table 1.

The sample was prepared using a small volume of epoxy resin and hardener with a ratio of 1:1 in a 5 cm cylindrical mould. The void size has a different thickness in order to obtain the same electric field strength of 3 kV/mm with the applied voltage of 8.8 kV for air void and 12.1 kV for water void. The void defect was created separately before it cast into a large block by injecting the air bubble and water in the middle of the small epoxy block. Figure 1 shows the schematic diagram of the PD samples.

2.2 Partial Discharge Measurement

Figure 2 shows a schematic diagram of the PD measurement used in this work to measure PD activity. The setup comprised several components, including a high voltage supply, a voltage regulator, a transformer, a coupling capacitor, a limiting resistor, a specimen, a coupling device, a PD detector, and a PicoScope connected to a personal computer (PC). The Rogowski coil sensor served as a means to detect and quantify the apparent charge magnitude of the PD signal emanating from the PD activity. The signal output from the PD detector was then linked to a PicoScope, and the resulting data was transmitted to a PC for storage. To mitigate the risk of breakdown caused by surface discharges occurring along the periphery of the stainless-steel electrodes and the material under examination, the specimen was submerged in mineral oil.

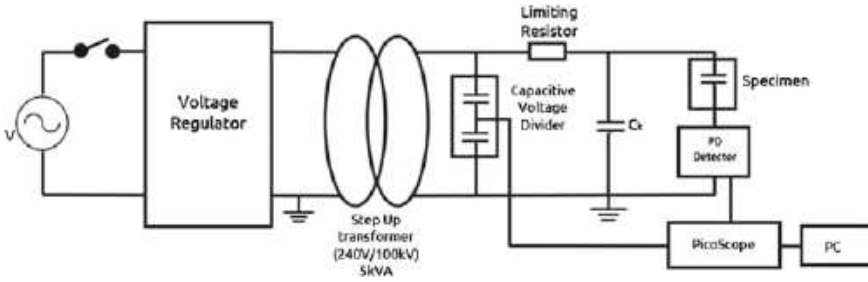


Fig. 2 Schematic diagram of PD measurement

3 Phase-Resolved Techniques

The characterization of PDs relies on three crucial parameters: phase angle (φ), PD charge magnitude (q), and the number of PD pulses (n). Each discharge type is composed of a different PD distribution pattern for PD source identification based on their PDs parameters. Due to the raw PD data is voluminous and challenging to analyze across 1000 cycles, it becomes necessary to perform feature extraction. This process aims to distill the data into a more manageable and informative representation of the PRPD, facilitating easier analysis and interpretation. The type of discharge can also be identified by analyzing the observed PD pattern, provided that these distinctions can be quantified using statistical parameters.

Initial step or known as pre-processing which extracted from the phase-resolved patterns are grouped by their phase angle with respect to $50 (\pm 5)$ Hz sine wave. Two primary distributions which are $H_n(\varphi)$ and $H_{qn}(\varphi)$ have been sorted from PRPD. $H_n(\varphi)$ represents a two-dimensional (2-D) graph showing the relationship between PD count and phase angle, whereas $H_{qn}(\varphi)$ is a 2-D graph depicting the relationship between PD charge magnitude and phase angle. The PD distributions can be split into two distinct sets for the positive and negative halves of the applied voltage cycle. Thus, during the positive half-cycle, we have distributions known as $H_{qn}^+(\varphi)$ and $H_n^+(\varphi)$, while during the negative half-cycle, they are referred to as $H_{qn}^-(\varphi)$ and $H_n^-(\varphi)$.

4 Results and Discussion

Figure 3a, b show the 2-D phase-resolved patterns (φ - q) for internal discharges with 0.5 mm air void and water void, respectively. The phase angle of PD occurrences is depicted on the x-axis, while the charge magnitude of the PDs is shown on the y-axis. Internal discharges involving air voids tend to happen during the 1st and 3rd quarter cycles, corresponding to the phase angles of the applied voltage. These discharges typically occur near the zero-crossing points of the power cycle, which

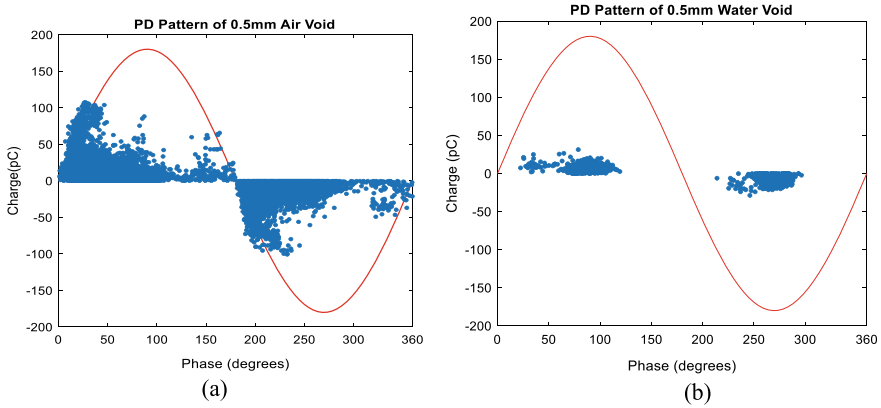


Fig. 3 2-D phase-resolved patterns (φ - q) for internal discharges with **a** 0.5 mm air void **b** 0.5 mm water void

are at 0° and 180° . In contrast, internal discharges associated with water voids take place differently. They occur in the middle of the 1st and 2nd cycles on the positive side and between the 3rd and 4th cycles on the negative side of the voltage waveform. The occurrence of PD in the void is influenced by the electric field which can occur when the electric field is higher than the inception voltage. Since the electric field within the void adheres to the sinusoidal nature of the applied voltage, the PRPD patterns consequently mirror the shape of a sinusoidal waveform.

Based on Fig. 3, the magnitude of the air void is higher than water void due to the density of the medium is different. Air density is lower than water density which can make the process of the electrons avalanche easy and fast. The electric field in this work is known due to the applied voltage in this work is fixed. However, it is foreseeable that with an escalation in the applied voltage, the magnitude of PD charge would also increase due to the heightened maximum electric field within the void. The electric field distribution on the void's surface maintains symmetry, given that the void is positioned centrally within the material. Consequently, the discharge patterns originating from the void in PRPDs during both the positive and negative cycles of the applied voltage exhibit equality.

Moreover, the PDs activity for air void is more than water void as shown in Fig. 4a compared to Fig. 4b. The highest PD count is 239 in the positive cycle and 232 in the negative cycle for air void while for water void is 39 and 30, respectively. Both air voids and water voids exhibit a higher frequency of PDs during the positive cycle due to the presence of electrons from the electrode under a positive applied voltage. This electron availability leads to a greater occurrence of electron avalanches when artificial voids are present in the samples.

The maximum PD charges captured for air void is higher than water void as shown in Fig. 5. Internal discharges with air void captured 107pC in positive cycle and 100pC in negative cycle. While for water void, the maximum charges for positive and negative cycles are 31pC and -28 pC, respectively. The maximum PD charge

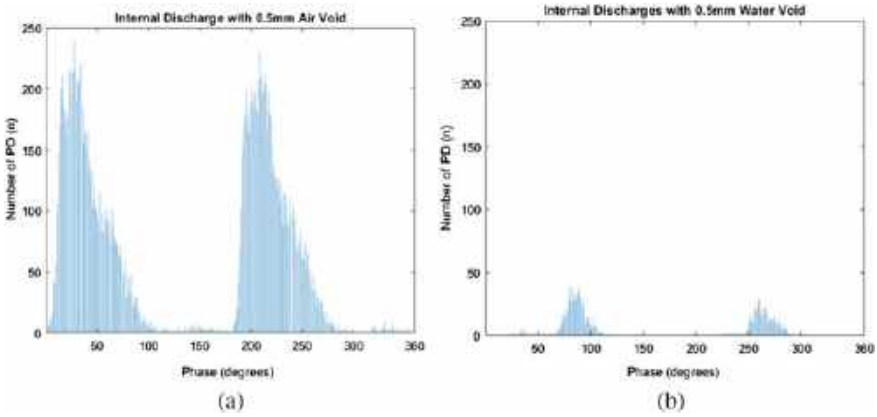


Fig. 4 $H_n(\varphi)$ distributions obtained from the measurements for: **a** air void **b** water void

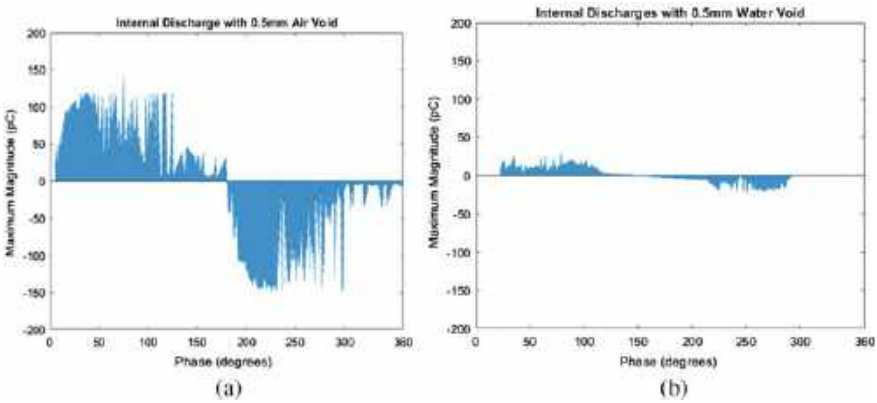


Fig. 5 $H_{qn}(\varphi)$ distributions obtained from the measurements for: **a** air void **b** water void

magnitude tends to be higher during the positive cycle because of the electric field's accumulation along the path of electron avalanches on the material surface. Consequently, certain avalanches can extend and generate a greater PD charge magnitude during this phase.

5 Conclusion

Two types of voids in internal discharges have been analyzed in this work using the phase-resolved partial discharge (PRPD) technique. Based on the PRPD, the discharge characteristics of different voids show the different asymmetry of distribution. Discharges behaviors are influenced by the applied voltage and void type.

$H_n(\varphi)$ distributions for both void types show that the number of discharges increased as the applied voltage increased. Although the number of discharges for air void is higher than for water void, due to the density of the water void being higher than air void, it would make the process of the electron avalanche in water void slower than for air void. Practically, the magnitude of discharges decreases with the insulation thickness. However, by using PRPD method, the maximum magnitude of the water void can be known in $H_{qn}(\varphi)$ distribution, which is decreased although the insulation thickness is thicker than the air void. In conclusion, the PRPD pattern can be used to diagnose and identify the characteristics of PD activity, and also acts as visual evidence of PD events, so that further action to resolve the PD issue can be made.

Acknowledgements This work was supported in part by Universiti Sains Malaysia (USM) under the Research Universiti Grant (RUI) 1001/PELECT/8014050 (UO1620/2018/0320), and in part by the Ministry of Higher Education, Malaysia, under the Fundamental Research Grant Scheme FRGS/1/2019/TK04/UNIMAP/03/8.

References

1. High-Voltage Test Techniques Partial Discharge Measurements and I. E. C. IEC 60270:2000. IEC 60270:2000—High-voltage Test Techniques—Partial Discharge Measurements (2001)
2. Dessouky S, El Faraskoury A, El-Mekawy S, El Zanaty W (2014) The optimal classification of partial discharge defects within XLPE cable by using ANN and statistical techniques. *Port-Said Eng Res J* 18(2):1–7. <https://doi.org/10.21608/pserj.2014.45254>
3. Jineeth J, Mallepally R, Sindhu TK (2018) Classification of partial discharge sources in XLPE cables by artificial neural networks and support vector machine. *IEEE Electr Insul Conf EIC* 2018:407–411. <https://doi.org/10.1109/EIC.2018.8481124>
4. Rosle N, Rohani MNKH, Muhamad NA, Kamarol M (2021) Partial discharges classification methods in XLPE cable: a review. *IEEE Access* 9:133258–133273. <https://doi.org/10.1109/ACCESS.2021.3115519>
5. Dessouky SS, Said P, El-faraskoury A, Electricity E, Company H (2014) Comparative analysis and classification of partial discharge defects within XLPE medium voltage cable by using ANN and statistical techniques. *Int J Electr Eng Technol* 5:1–16. <https://doi.org/10.1002/201>
6. Karimi M, Majidi M, Etezadi-Amoli M, Oskuoee M (2018) Partial discharge classification using deep belief networks. In: 2018 IEEE/PES transmission and distribution conference and exposition, pp 1061–1070. <https://doi.org/10.1109/TDC.2018.8440224>
7. Kothoke MP (2019) Analysis and determination of partial discharge type using statistical techniques and back propagation method of artificial neural network for phase-resolved data, vol 8, no 08, pp 430–438
8. Myers RR (1974) Epoxy resins: chemistry and technology, vol 49, no 1

Exploring the Impact of Accelerated Thermal Aging on POME-Based MWCNT Nanofluid



Sharifah Masniah Wan Masra, Yanuar Zulardiansyah Arief, Siti Kudnie Sahari, Ernieza Musa, Andrew Ragai Henry Rigit, and Md. Rezaur Rahman

Abstract This study aims to investigate the effects of accelerated thermal aging on modified refined, bleached, and deodorized (RBDPO) olein. Through a transesterification process, the RBDPO olein is converted into palm oil methyl ester (POME), which acts as the base fluid in the presence of conductive multi-walled carbon nanotube (MWCNT) at various concentrations. The accelerated aging is conducted at a temperature of 130 °C for 1000 h. Fourier transform infrared (FTIR) analysis reveals that the chemical composition of the aged nanofluids remains unchanged. The AC breakdown voltages of the aged nanofluids decrease as a result of the accelerated thermal aging over 1000 h, but they remain higher than those of the fresh POME.

Keywords Thermal aging · Palm oil methyl ester · Multi-walled carbon nanotube

S. M. W. Masra (✉) · Y. Z. Arief · S. K. Sahari · E. Musa
Department of Electrical and Electronic Engineering, Faculty of Engineering, Universiti Malaysia Sarawak (UNIMAS), 94300 Kota Samarahan, Sarawak, Malaysia
e-mail: wmmasnia@unimas.my

Y. Z. Arief
e-mail: ayzulardiansyah@unimas.my

S. K. Sahari
e-mail: sskudnie@unimas.my

E. Musa
e-mail: 22020204@siswa.unimas.my

A. R. H. Rigit
Department of Mechanical Engineering, Faculty of Engineering, Universiti Malaysia Sarawak (UNIMAS), 94300 Kota Samarahan, Sarawak, Malaysia

Md. R. Rahman
Department of Chemical Engineering and Energy Sustainability, Faculty of Engineering, Universiti Malaysia Sarawak (UNIMAS), 94300 Kota Samarahan, Sarawak, Malaysia

1 Introduction

A significant shift in research interest towards the exploration of biodegradable and renewable alternatives in the field of nanofluid research has arisen recently. Numerous studies have explored replacing mineral oils with modified vegetable oils [1–4]. Transformer insulating liquids are subjected to electrical, thermal, and chemical stresses, causing them to gradually deteriorate [5]. Aging is the mechanism responsible for resulting in gradual and irreversible changes in the properties of transformer oil [5]. An enhanced understanding of the properties changes of aged nanofluids subjected to accelerated aging conditions is thus important to quantify the variations of nanofluid properties at various concentrations and aging durations at a specified operating temperature. Therefore, the objective of this study is to expand the scope of research by modifying refined, bleached, and deodorized palm oil (RBDPO) olein into methyl ester and incorporating nanoparticles to examine the aging mechanism's effect on the oil samples. The question focuses on whether the modified nanofluids retain their dielectric strength after thermal aging.

2 Experiment

2.1 Materials

This project utilizes several key materials, including RBDPO olein, methanol sourced from Merck, potassium hydroxide (KOH) obtained from J. T. Baker, conducting multi-walled carbon nanotube (MWCNT) nanoparticles, and hexadecyltrimethylammonium bromide (CTAB) surfactant manufactured by Sigma-Aldrich.

2.2 Transesterification Reaction

The investigated methyl ester in this study was synthesized through a transesterification reaction involving RBDPO olein, methanol, and KOH as a catalyst. The reaction was conducted with a molar ratio of 6:1 (methanol:oil) at a temperature of 60 °C, while continuously stirring the mixture for 60 min. After cooling the mixture to room temperature, it was transferred to a separatory funnel and left to settle for 24 h. As a result of transesterification, the mixture separated into two distinct phases: glycerol and methyl ester. The bottom phase, consisting of crude glycerol and KOH, was discarded. The top layer, which contained fatty acid methyl ester (FAME), underwent a water-washing process to remove excess KOH. The remaining impurities were eliminated by heating and stirring the mixture at 60 °C for 30 min using a magnetic stirrer. Subsequently, the resulting FAME obtained from the transesterification of

RBDPO olein was designated as palm oil methyl ester (POME) and utilized as the base fluid for preparing the nanofluids.

2.3 Nanofluids Preparation

The nanofluids were prepared using a two-step method. The first step involved the preparation of the dried nanoparticles, specifically MWCNT, which were pre-heated in an oven to eliminate any moisture. In the second step, the nanofluids were prepared by dispersing the dried MWCNT nanoparticles into the POME base fluids through chemical and mechanical treatments. To initiate the dispersion process, an appropriate amount of the CTAB surfactant as investigated in [2], was added to 1000 ml pre-dried POME and stirred for 30 min. Then, the MWCNT nanoparticles were dispersed into the POME to achieve various concentrations required for the nanofluid samples. The considered doping concentrations of MWCNT NPs were 0.01, 0.02, 0.05, and 0.10 g/L. To ensure homogeneity and minimize the risk of agglomeration and sedimentation, the prepared nanofluid samples underwent ultrasonication at a temperature of 50 °C for 2 h. This ultrasonication process further aided in achieving well-dispersed and homogeneous nanofluids [6].

2.4 Mechanism of Nanofluids Aging

The prepared nanofluid samples were exposed to accelerated thermal aging in an oven under closed aging conditions, with the temperature at 130 °C. The analysis was performed at four different aging durations: 0, 250, 500, and 1000 h. The AC breakdown voltage (AC BDV) measurement was used to assess the aging behavior of the nanofluids. The qualitative degradation of the breakdown voltage of the modified nanofluids in response to accelerated thermal aging is studied.

3 Characterization and Measurements

3.1 FTIR Spectra Analysis

In this study, the characterization of the pure POME and the aged nanofluids was performed using a Thermo Scientific Fourier Transform Infrared (FTIR) spectrophotometer to gain insights into the aging behavior of the POME blended with MWCNT nanoparticles. The FTIR spectra were recorded with the absorbance bands of the nanofluids over a range of wavenumbers from 4000 to 600 cm^{-1} . The FTIR conditions were set with 32 scans and a resolution of 4.

3.2 AC BDV Measurement

The AC BDV measurement of the nanofluids was conducted using an HZJQ-1B transformer oil BDV tester with IEC 60156 compliance standard testing. The electrodes of brass hemispherical-shaped with a diameter of 12.5 mm and at a fixed distance of 2.5 mm were used in the testing. The voltage was automatically increased at a rate of 2 kV/s until breakdown occurred. To ensure consistency, all measurements were performed with constant stirring of the oil samples. For each nanofluid sample, we conducted six sets of six consecutive AC BDV measurements and then calculated the average value.

4 Findings and Analysis

4.1 FTIR Analysis

The FTIR study aimed to assess any changes in functional groups resulting from aging. Figure 1 displays the comparative FTIR analysis of nanofluids aged at different concentrations and durations. Visual examination of the spectra indicated that aging did not significantly alter the spectra. This finding aligns with the findings of a previous investigation conducted by [7], which examined various esters and nanoparticles. Notably, a distinct, sharp peak was observed at the absorbance wavenumber 1741 cm^{-1} , indicating the stretching vibration of C=O and confirming the presence of ester. Peaks at wavenumbers 2922 and 2852 cm^{-1} suggested the presence of carboxylic acids [8].

4.2 Average of AC BDV

A total of 36 AC BDV measurements were conducted on both fresh and aged nanofluids at different aging durations. The average AC BDV values are presented in Fig. 2. The results indicate that the BDV of the aged nanofluids after 1000 h is higher compared to the fresh nanofluids. The incorporation of MWCNT nanoparticles in the POME base fluids has a positive impact on all doping concentrations, with a concentration of 0.10 g/L nanoparticles showing the highest BDV, as depicted in Fig. 2. Overall, the analysis reveals that the dielectric strength of the nanofluids increases after 250 h of aging but subsequently decreases with prolonged aging. However, it has been observed that the AC BDVs of the aged nanofluids are always superior to the pure POME, irrespective of aging duration.

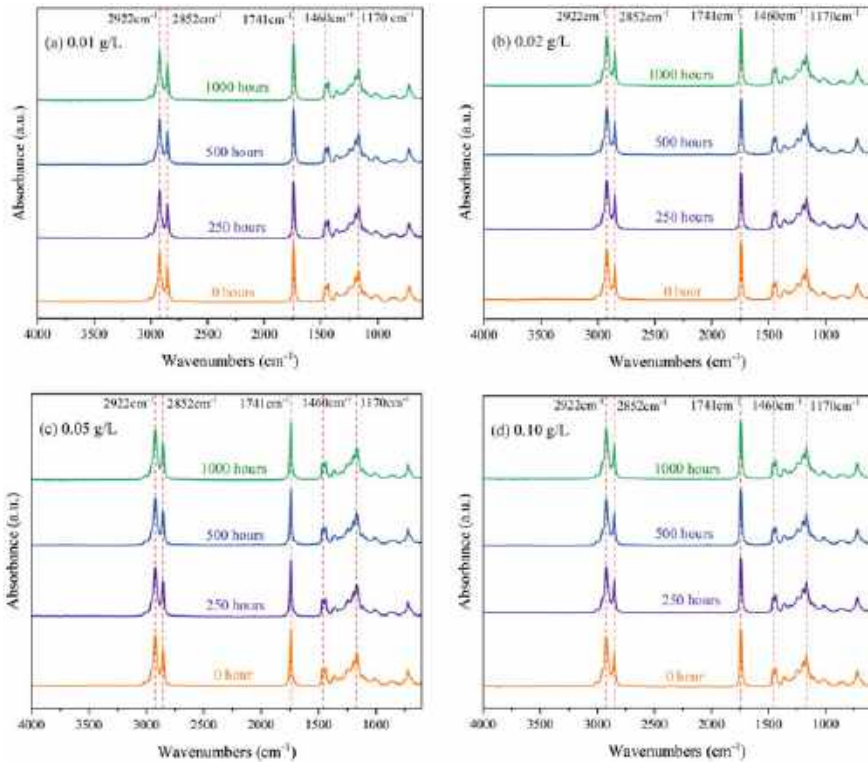


Fig. 1 FTIR spectra of fresh and aged nanofluids at various concentrations

4.3 Statistical Analysis

The Weibull distribution plots at a 95% confidence interval of nanofluids for 0- and 1000-h aging periods are shown in Fig. 3. The star, square, triangle, and circle plots indicate the doping concentrations of 0.01, 0.02, 0.05, and 0.10 g/L MWCNT nanoparticles, respectively. Table 1 shows the scale (α) and shape (β) parameters of the Weibull distributions for 1000-h-aged NFs. Then, the agreement of the BDV data following the Weibull distribution was determined by the Anderson–Darling (AD) goodness-of-fit test [9]. The ρ -value was determined and compared to the 0.05 significance level. Table 1 also summarizes the results of the AD test. The results indicate that the AC BDV at low concentrations (0.01, 0.02, and 0.05 g/L) obeys the Weibull distribution. The Weibull fit lines as shown in Fig. 3b were used to evaluate the AC BDV at 1%, 10%, and 50% probability failure, and the results are tabulated in Table 2. For $U_{10\%}$ and $U_{50\%}$, the AC BDV was optimal with concentrations at 0.02 g/L, which gave improvements of 5.2% and 50.6% higher than the pure POME.

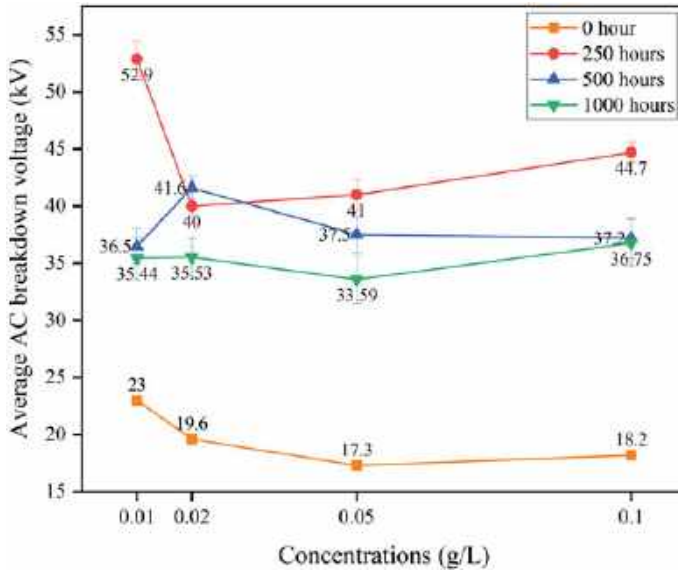


Fig. 2 Average AC BDV of fresh and aged nanofluids

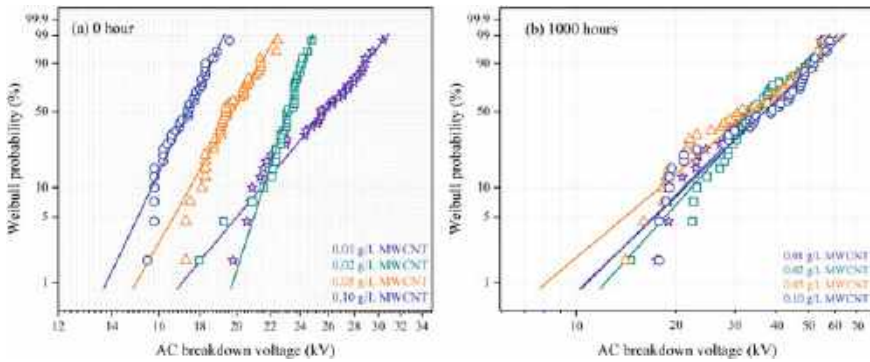


Fig. 3 Weibull distribution plots of AC BDV for a 0, and b 1000 h

Table 1 Scale, shape parameters, and AD Goodness-of-fit test for 1000-h-aged NFs

Oil samples	α (Scale)	β (Shape)	ρ -value	Decision
Pure POME	24.2	15.8	≥ 0.25	Accepted
0.01 g/L NF	39.5	3.6	0.240	Accepted
0.02 g/L NF	39.2	3.9	0.122	Accepted
0.05 g/L NF	37.8	3.0	0.054	Accepted
0.10 g/L NF	41.1	3.4	0.024	Rejected

Table 2 AC BDV at 1%, 10%, and 50% probability levels

Oil samples	U _{1%}		U _{10%}		U _{50%}	
	AC BDV (kV)	Increment (%)	AC BDV (kV)	Increment (%)	AC BDV (kV)	Increment (%)
Pure POME	18.1	–	21.1	–	23.7	–
0.01 g/L NF	10.8	– 40.3	21.0	– 0.5	35.6	50.0
0.02 g/L NF	12.3	– 32.0	22.2	5.2	35.7	50.6
0.05 g/L NF	8.3	– 54.1	17.9	– 15.2	33.5	41.4

5 Conclusions

By modifying RBDPO olein through a transesterification process and blending it with conducting MWCNT nanoparticles, this study aims to explore its potential use as an alternative insulating liquid in transformers. Two aspects of thermal aging in nanofluids were studied: FTIR spectra analysis and AC BDV measurements. The key findings can be summarized as follows: (a) The chemical structure of aged nanofluids remained unchanged, as observed in the FTIR spectra; (b) The dielectric strength of the aged nanofluids are always superior to pure POME, irrespective of aging duration; and (c) The statistical analysis demonstrated that the low concentrations of 1000-h-aged NFs results obey the Weibull distribution.

References

1. Makmud MZH et al (2018) Influence of conductive and semi-conductive nanoparticles on the dielectric response of natural ester-based nanofluid insulation. *Energies* 11(2)
2. Mohamad NA, Azis N, Jasni J (2019) Impact of Fe₃O₄, CuO and Al₂O₃ on the AC breakdown voltage of palm oil and coconut oil in the presence of CTAB. *Energies* 12
3. Oparanti SO et al (2020) Dielectric characterization of palm kernel oil ester-based insulating nanofluid, pp 225–228
4. Sitorus HBH et al (2016) Jatropha curcas methyl ester oil obtaining as vegetable insulating oil. *IEEE Trans Dielectr Electr Insul* 23(4):2021–2028
5. Abdi S et al (2011) Influence of artificial thermal aging on transformer oil properties. *Electr Power Components Syst* 39(15):1701–1711
6. Fasehullah M, Wang F, Jamil S (2022) Significantly elevated AC dielectric strength of synthetic ester oil-based nanofluids by varying morphology of CdS nano-additives. *J Mol Liq* 353:118817
7. Maharana M et al (2019) Condition assessment of aged ester-based nanofluid through physicochemical and spectroscopic measurement. *IEEE Trans Instrum Meas* 68(12):4853–4863
8. Hariram V et al (2016) Analyzing the fatty acid methyl esters profile of palm kernel biodiesel using GC/MS, NMR and FTIR techniques. *J Chem Pharm Sci* 9(4):3122–3128
9. Khelifa H, Beroual A, Vagnon E (2023) Effect of conducting and semi-conducting nanoparticles on the AC breakdown voltage and electrostatic charging tendency of synthetic ester. *IEEE Trans Dielectr Electr Insul*

Accelerating Electric Vehicle Adoption on Malaysian Islands: Lessons from Japan's Islands of the Future Initiative



M. Reyasudin Basir Khan , Jabbar Al-Fattah , Gazi Md. Nurul Islam , Ahmad Anwar Zainuddin , Chong Peng Lean , Saidatul Izyanie Kamarudin , and Miqdad Abdul Aziz 

Abstract The adoption of electric vehicles (EVs) and renewable energy solutions is imperative for achieving sustainable transportation and reducing carbon emissions. This paper explores the challenges faced by Malaysian islands in adopting EV mobility and draws lessons from Japan's Islands of the Future initiative. Through site surveys and interviews conducted on Malaysian islands and a study visit to Koshiki Island in Japan, the study identifies challenges such as limited charging infrastructure, range anxiety, high upfront costs, and lack of public awareness. By implementing strategies such as prioritizing renewable energy generation, developing a robust EV charging network, fostering public–private partnerships, offering incentives and subsidies, and conducting education campaigns, Malaysia can overcome these challenges and accelerate EV adoption on its islands. The findings of this

M. Reyasudin Basir Khan (✉) · G. Md. Nurul Islam
Tun Razak Graduate School, Universiti Tun Abdul Razak (UNIRAZAK), Kuala Lumpur, Malaysia
e-mail: reyasudin@unirazak.edu.my

J. Al-Fattah
Fakulti Kejuruteraan Elektrik Dan Elektronik, Universiti Tun Hussein Onn, Parit Raja, Johor, Malaysia

A. A. Zainuddin
Department of Computer Science, International Islamic University Malaysia (IIUM), Kuala Lumpur, Malaysia

C. P. Lean
MILA University, Putra Nilai, Negeri Sembilan, Malaysia

S. I. Kamarudin
College of Computing, Informatics and Media, Universiti Teknologi MARA (UiTM), Shah Alam, Selangor, Malaysia

M. A. Aziz
Department of Electrical Technology Section, Universiti Kuala Lumpur, British Malaysian Institute, Kuala Lumpur, Malaysia

study provide valuable insights for policymakers, industry stakeholders, and local communities involved in promoting sustainable transportation on Malaysian islands.

Keywords Malaysia · Japan · Electric vehicle · Island · Koshiki

1 Introduction

Electric vehicles (EVs) are widely regarded as a key solution for reducing greenhouse gas emissions and achieving sustainable transportation. EVs include battery electric vehicles (BEVs), plug-in hybrid electric vehicles (PHEVs), and fuel cell electric vehicles (FCEVs). EVs offer several advantages over conventional internal combustion engine (ICE) vehicles, such as lower operating costs, higher energy efficiency, and lower noise and air pollution. However, EV adoption faces several challenges, such as high upfront costs, limited driving range, insufficient charging infrastructure, and low consumer awareness and acceptance.

These challenges are particularly pronounced in island settings, where the transportation sector is heavily dependent on imported fossil fuels and contributes significantly to carbon emissions. Islands also face unique geographical and socio-economic constraints, such as small land area, limited road network, high population density, and vulnerability to natural disasters and climate change impacts. Therefore, promoting EV adoption in islands requires tailored strategies that address the specific needs and opportunities of these contexts [1–4].

This paper aims to explore the potential of EV adoption for islands in Malaysia, a Southeast Asian country with over 800 islands of various sizes and characteristics. Malaysia has set a target of achieving 10% EV penetration by 2030 [1], but its current EV market share is less than 1% [5, 6]. Moreover, most of the existing policies and initiatives for EV promotion are focused on the mainland, while the islands remain largely neglected. This paper seeks to fill this gap by examining the challenges and opportunities for EV adoption in Malaysia's islands and drawing lessons from Japan's Islands of the Future initiative, which has successfully implemented renewable energy and EV solutions in an island.

2 Japan's Island of the Future Initiative

2.1 Background

Japan has launched the Islands of the Future initiative in 2014 to promote the use of renewable energy and EVs in remote islands. One of the pilot islands is Koshiki Island in Satsumasendai City, Kagoshima Prefecture. The island has implemented various measures to support EV adoption, such as installing solar panels and charging

stations, providing EV car-sharing services, and offering discounts for EV users. The island aims to achieve 100% renewable energy and zero-emission mobility by 2030 [7, 8].

2.2 *Koshiki Island*

Koshiki Island is a group of islands that are part of the Koshikishima Quasi-National Park and have a rich natural and cultural heritage. The island is also striving to become an eco-island by promoting renewable energy, electric vehicles, and smart houses.

Koshiki Island is collaborating with Sumitomo Corporation to develop a locally generated, locally used energy model using reused electric vehicle batteries as energy storage systems. The project aims to stabilize the power supply and demand on the island, which relies on diesel generators and has a small-scale grid that is isolated from the mainland.

The project consists of a solar farm with a generation capacity of 100 kW and a power management center with a storage capacity of 800 kW, using batteries from 36 electric vehicles that have completed their service life. The project also utilizes an energy management system that monitors and controls the power generation and consumption on the island [8].

The project is expected to reduce the carbon dioxide emissions on the island by replacing some of the diesel power generation with renewable energy. It will also enhance the disaster resilience of the island by providing backup power in case of power outages caused by typhoons or other events. The project also aims to establish a new business model for providing an environment for connecting renewable energy using reused electric vehicle batteries.

Koshiki Island is one of the examples of how Japan is trying to innovate and advance in the field of electric vehicles and related infrastructure. Japan has been a leader in hybrid and fuel cell vehicles and has also introduced some policies and incentives to support electric vehicle adoption [9, 10]. Japan also has some electric charging stations available in major cities and along highways, although their number and availability are still limited compared to other countries [9, 10]. Figure 1 shows the charging station as well as the electric vehicles available on the island. Meanwhile, Fig. 2 shows the solar PV generation available.

3 Discussions

Koshiki Island is a model for the future of electric vehicles in Japan, and it has some technologies, policies and procedures that can be implemented in Malaysia Island.

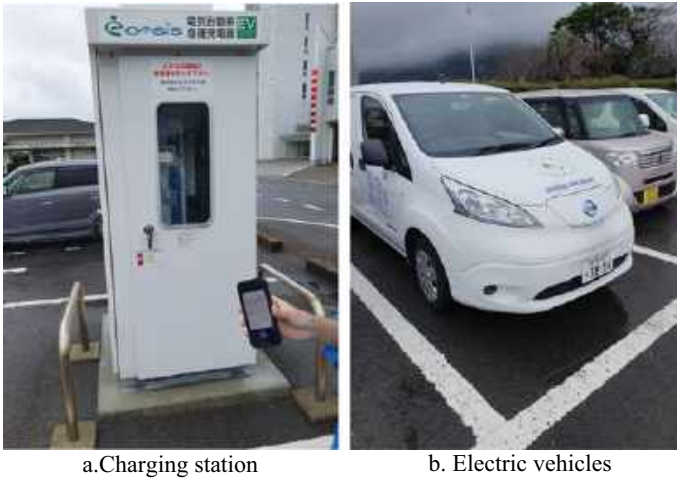


Fig. 1 Site visits to Koshiki island, **a** charging station **b** EV



Fig. 2 Solar PV generation in Koshiki Island

3.1 Technologies

Koshiki Island reuses electric vehicle batteries for energy storage, reducing waste and costs. This technology can be applied in Malaysia’s islands, where electric vehicle adoption and energy storage are needed. Koshiki Island’s energy management system improves grid efficiency and can benefit Malaysia’s islands facing challenges in renewable energy integration.

3.2 Policies and Partnerships

Koshiki Island collaborates with Sumitomo Corporation to develop a local energy model, encouraging private sector innovation in electric vehicles and renewable energy. This policy can be adopted in Malaysia's islands to support electric vehicle adoption and renewable energy development. Koshiki Island's eco-island vision promotes renewable energy and smart houses, enhancing quality of life and tourism potential, inspiring Malaysia's islands to showcase their assets.

3.3 Procedures

Koshiki Island's solar farm and power management center, powered by retired electric vehicle batteries, optimize renewable energy use and provide backup power during emergencies. Malaysia's islands can replicate these procedures to improve charging infrastructure and enhance disaster resilience for electric vehicles and renewable energy (Table 1).

Table 1 Summary of Japan EV strategies that can be implemented in Malaysia

Japan strategies	How to implement in Malaysia
Reusing electric vehicle batteries as energy storage systems	<ul style="list-style-type: none"> • Establish a system for collecting and recycling used EV batteries from the market • Develop standards and guidelines for reusing EV batteries as energy storage systems • Encourage private sector participation and innovation in providing energy storage solutions using reused EV batteries
Developing a locally generated, locally used energy model	<ul style="list-style-type: none"> • Assess the renewable energy potential and demand on each island • Invest in renewable energy projects such as solar, wind, hydro, biomass, etc. on the islands • Develop smart grids and microgrids to integrate and manage renewable energy sources and loads on the islands
Promoting renewable energy, electric vehicles, and smart houses as part of an eco-island vision	<ul style="list-style-type: none"> • Create a clear and consistent policy framework and roadmap for achieving the eco-island vision • Provide incentives and subsidies for adopting renewable energy, electric vehicles, and smart houses on the islands • Enhance the quality of life and tourism potential of the islands by showcasing their natural and cultural assets

4 Conclusion

This study has explored the challenges and opportunities of adopting electric vehicles and renewable energy solutions on Malaysian islands and has drawn lessons from Japan's Islands of the Future initiative. The study has found that Malaysia can overcome the challenges such as limited charging infrastructure, range anxiety, high upfront costs, and lack of public awareness by implementing strategies such as reusing electric vehicle batteries as energy storage systems, developing a locally generated, locally used energy model, promoting renewable energy, electric vehicles, and smart houses as part of an eco-island vision, fostering public-private partnerships, offering incentives and subsidies, and conducting education campaigns. The study has also provided valuable insights for policymakers, industry stakeholders, and local communities involved in promoting sustainable transportation on Malaysian islands. The study hopes to contribute to the advancement of electric vehicle and renewable energy sectors in Malaysia and the region, as well as to the achievement of the sustainable development goals and the Paris Agreement targets.

Acknowledgements The authors would like to express their sincere gratitude to the Sumitomo Foundation for their financial support through the Grant for Fiscal Year 2020 (Grant Number: 208422).

References

1. Zahraoui Y et al (2021) A novel approach for sizing battery storage system for enhancing resilience ability of a microgrid. *Int Trans Electrical Energy Syst* 31(12):e13142
2. Younes Z et al (2021) Blockchain applications and challenges in smart grid. In: 2021 IEEE Conference on Energy Conversion (CENCON), IEEE
3. Reyasudin Basir Khan M, Jidin R, Pasupuleti J (2016) Data from renewable energy assessments for resort islands in the South China Sea. *Data Brief* 6:117–120
4. Reyasudin Basir Khan M, Jidin R, Pasupuleti J (2016) Energy audit data for a resort island in the South China Sea. *Data Brief* 6:489–491
5. Lee J, Kim J, Kim S (2023) Capturing growth in Asia's emerging EV ecosystem. McKinsey & Company [Online]. Available: <https://www.mckinsey.com/featured-insights/future-of-asia/capturing-growth-in-asias-emerging-ev-ecosystem>. Accessed 07 Jun 2023
6. Shah KU, Awojobi M, Soomauroo Z (2022) Electric vehicle adoption in small island economies: review from a technology transition perspective. In: *WIRES energy and environment* [Online]. Available: <https://wires.onlinelibrary.wiley.com/doi/10.1002/wene.432>. Accessed 07 Jun 2023
7. Nishitatenno S, Burke PJ (2021) Current situation of electric vehicles in ASEAN. In: Promotion of electromobility in ASEAN: policy recommendations for electric vehicle adoption and development of charging infrastructure network. Economic Research Institute for ASEAN and East Asia (ERIA), Jakarta, pp 1–24 [Online]. Available: https://www.eria.org/uploads/media/Research-Project-Report/2021-03-Promotion-Electromobility-ASEAN/5_ch.1-Current-Situation-Electric-Vehicle-ASEAN-2611.pdf. Accessed 07 Jun 2023
8. Construction of a locally generated, locally used energy model on Koshiki Island. Challenge Zero [Online]. Available: <https://www.challenge-zero.jp/en/casestudy/414>. Accessed 07 Jun 2023

9. Dooley B, Ueno H (2021) Why Japan is holding back as the world rushes toward electric cars. The New York Times, 9 Mar 2021 [Online]. Available: <https://www.nytimes.com/2021/03/09/business/electric-cars-japan.html>. Accessed 07 Jun 2023
10. Capturing growth in Asia's emerging EV ecosystem. McKinsey & Company, Sept 2021 [Online]. Available: <https://www.mckinsey.com/featured-insights/future-of-asia/capturing-growth-in-asias-emerging-ev-ecosystem>. Accessed 07 Jun 2023

Effect of Incidence Angle on the Performance of a Dual Cantilever Flutter Energy Harvester



Venod Reddy Velusamy, Muhammad Izzikry Mohd Farid Suhaimi, and Faruq Muhammad Foong

Abstract This paper analyses the effect of incidence angle on the performance of a dual cantilever flutter (DCF) energy harvester. An experiment was conducted for incidence angles of 0° , 30° , 60° and 90° about the z-axis where the 0° angle correspond to the case where both cantilever beams are positioned perpendicular to the wind flow. The electromagnetic power output was then theoretically estimated from the experimental measurements. Results demonstrate that from 0° to 60° , the critical flutter speed of the device decrease with angle of incidence, which highlights a potentially larger bandwidth. However, the flutter amplitude also decreases with incident angle, recording a 54.7% decrease at 60° when compared with 0° . This corresponded to a 76.1% drop in predicted power output. At 90° , flutter was not recorded within the tested wind speeds. The experiment was then repeated by rotating the DCF 90° about the x-axis. The results obtained this time was very similar to the 0° angle along the z-axis in terms of amplitude and power. However, the critical flutter speed was reduced by 16.3%. Interestingly, the flutter frequency remains approximately constant after the critical flutter speed for all tested incidence angle. Finally, some important considerations to maximize the performance for the device were provided.

Keywords Incidence angle · Dual cantilever flutter · Wind energy harvesting · Critical flutter speed · Electromagnetic

V. R. Velusamy · M. I. M. F. Suhaimi · F. M. Foong (✉)

Faculty of Mechanical Engineering, Universiti Teknologi Malaysia, 81310 Johor, Malaysia
e-mail: faruqfoong@utm.my

F. M. Foong

UTM Aerolab, Institute for Vehicle Systems & Engineering, Universiti Teknologi Malaysia, 81310 Johor, Malaysia

1 Introduction

Interest in energy harvesting technologies quickly grew when the demand for a self-sustained wireless sensor network became necessary. A sensor network consists of many sensor nodes where for certain applications, each node can require less than 0.1 mW to operate [1, 2]. These small electronics are mainly used for monitoring purposes and at remote locations, it may be difficult to provide a direct power supply for the sensors. Even if a power source was to be provided, the installation and maintenance cost would be significant. Hence, energy harvesting comes as a viable energy source for these sensors. Among others, wind energy harvesting poses as a promising option as it is safe, ubiquitous and almost always available [3, 4]. While wind turbines are usually the first thing that comes to mind in wind power generation, some researchers have considered other methods such as harvesting wind energy from flow or vortex induced vibrations and flutter oscillations [5–7]. These methods generally produce smaller energy than wind turbines, but they cater the low wind speed range and a smaller device size [8].

Hobeck et al. [9] proposed an interesting flutter-based energy harvester named the dual cantilever flutter (DCF). The device consists of two cantilever beams placed side by side and oriented perpendicularly to the wind flow. During flutter, the device demonstrates a unique characteristic where both beams oscillate in an anti-phase motion or in the opposite direction relative to each other. In their experiment, Hobeck et al. [9] was able to achieve persistent, large amplitude vibrations and sufficient power using piezoelectric transducers. The idea of using such a simple energy harvesting device was so that they can be arranged into a grass like array consisting of multiple DCFs for a larger power generation. Nevertheless, Hobeck et al. [9] only considered the case the wind flow is perpendicular to the DCF whereas in practical applications, wind may appear from multiple directions. Hence, this study investigates the effect of different angle of incidence on the performance of the DCF energy harvester.

2 Experiment Setup

Two cantilever beams each measuring 130.0 mm × 20.0 mm × 0.5 mm (length × width × thickness) were clamped side by side inside the test section of a low-speed wind tunnel to create the DCF. The beams were separated at a distance of 4.0 mm apart from each other. Initially, the DCF were positioned at an incidence angle of 0° which is the case when the surfaces of both beams are perpendicular to the wind flow. This also corresponds to the typical DCF setup presented by Hobeck et al. [9]. The wind speed inside the tunnel was varied incrementally between 4.0 to 15.0 ms⁻¹ and two laser displacement sensors were used to capture the motion of each beam at its free end. A digital manometer was used to measure the pressure of the wind flow.

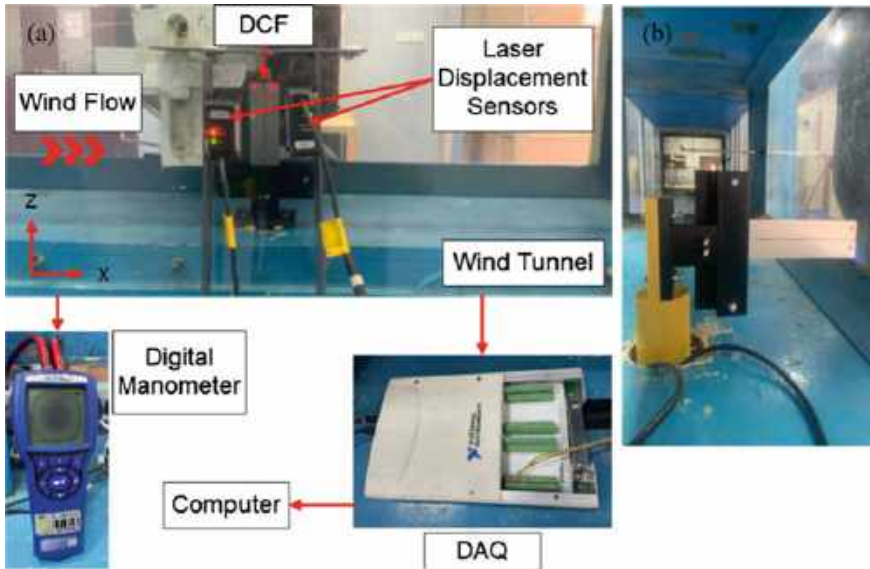


Fig. 1 **a** Experiment setup to measure effect of incidence angle on the performance of a DCF energy harvester and **b** Orientation of DCF at 90° along the x-axis

Measurements from both laser sensors were transferred to a computer for analysis via a data acquisition (DAQ) device.

The experiment was then repeated by rotating the DCF to incidence angles of 30° , 60° and 90° about the z-axis. This will simulate the scenario where the wind flow is projected onto the device from different directions. At 90° , the surfaces of the DCF beams are parallel to the wind flow as shown in Fig. 1a. Afterwards, the DCF was rotated 90° about the x-axis as shown in Fig. 1b. The experiment was again repeated for this setup under the same wind speed range. Note that under this orientation, the surfaces of the beams are also perpendicular to the wind flow, similar to the case of 0° angle about the z-axis.

3 Results and Discussion

The data obtained from the experiments was slightly off-set due to the bending of the DCF beams from the wind force. Hence, zero-centering was performed on the data to reflect the actual beam displacement due to flutter. Additionally, the amplitudes of both beams were averaged to obtain a single value.

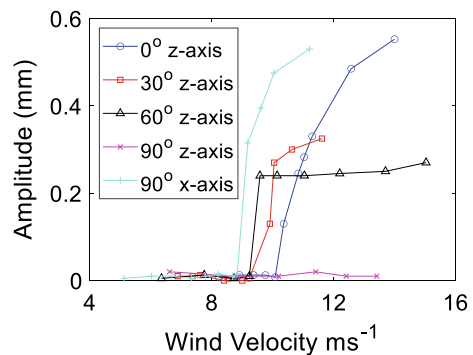
3.1 Critical Flutter Speed and Amplitude Analysis

One of the main contributors to the power output in flutter-based energy harvesters is device's flutter amplitude. In general, devices that flutter at larger amplitudes will generate a larger power output. The variation of amplitude with wind speed for all five experiments are demonstrated in Fig. 2.

Results in Fig. 2 shows that the DCF energy harvester still experience flutter when the direction of the wind flow changes between 0° to 60° about the z-axis. Additionally, a decrease in the critical flutter speed was also seen as the incidence angle increases. Although this decrease is relatively small at approximately 14.5%, it does suggest the potential increase in bandwidth of the energy harvester when subjected to wind flow from different direction. This means that the device can operate even if the wind flow is not perpendicular to the beam (0°). Nevertheless, the flutter amplitude of the device also decreases with increasing incidence angle which may reflect a lower performance at these angles despite the lower critical flutter speed. At an incidence angle of 90° , flutter was not recorded within the tested wind speeds, indicating that the device could not produce any useful output if the wind flow is parallel to the beam's surfaces.

When the device was rotated at 90° about the x-axis, experimental results show that the device exhibit a very similar trend to the 0° angle about the z-axis in terms of amplitude. The reason for this is because for both cases, the beams are oriented so that their surfaces are perpendicular to the wind flow. However, the critical flutter speed for the orientation of 90° about the x-axis is 16.3% lower than the 0° about the z-axis orientation. This shows that the former orientation is more favorable as it provides a larger operational bandwidth while not compromising the flutter amplitude. As for the flutter frequency, all experiments recorded very similar frequency of approximately 27.8 Hz, suggesting that the frequency is independent of the incidence angle.

Fig. 2 Variation in amplitude against wind speed for all tested incidence angles



3.2 Voltage and Power Output Analysis

As stated before, one of the main characteristics of a DCF energy harvester is that after the critical flutter speed, the two beams flutter in an anti-phase motion. This type of motion is favorable for electromagnetic transducers as the anti-phase motion can be used to increase the relative motion between the magnets and the conductor which in turn will multiply the voltage and power output [10]. While this have not been tested before, the theoretical analysis of electromagnetic induction for the DCF was considered in this study.

Assuming that both beams flutter with identical amplitudes and in anti-phase motion, the relation between the flutter amplitude, x_a , and the induced voltage from electromagnetic induction, E , based on Faraday's law is

$$E = 2Kx_a\omega \quad (1)$$

where K is the electromagnetic coupling factor and ω is the flutter frequency in radians. Considering that the conductor is connected in parallel to a load resistor, The voltage output at the load resistor is

$$V = E \frac{R_L}{R_L + R_c} \quad (2)$$

where R_L and R_c are the external load resistance and internal resistance of the conductor respectively. Taking the optimum load resistance case for low coupling factor where $R_L \approx R_c$ and applying's ohms law, the power output, P , of the DCF energy harvester can be estimated as

$$P = \frac{(Kx_a\omega)^2}{R_L} \quad (3)$$

To ensure the validity of Eq. (3), a low coupling factor of $K = 0.2 \text{ Tm}$ and a load resistance of 2.5Ω was considered for this analysis [11]. This also allows for the damping force induced by the electromagnetic components to be ignored. For simplicity, the effect from the mass of these components was also neglected. Under these conditions, Eqs. (1)–(3) can be applied to the amplitude results obtained in Fig. 2 to predict the electromagnetic voltage and power output for the device. Figure 3 demonstrates the estimated voltage and power output for the tested DCF at different incidence angles.

The drop in amplitude at increasing incidence angles from Fig. 2 resulted in a large reduction in power output of up to 76.1% at 60° angle. However, the 90° angle about the x-axis still display a promising output due to the similar trend in amplitude with the 0° angle about the z-axis.

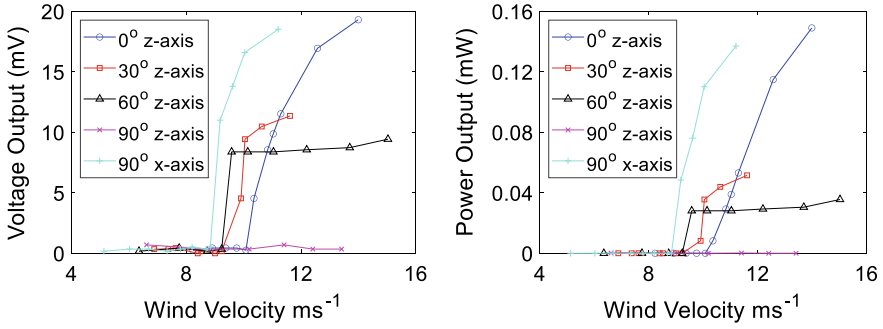


Fig. 3 Variation in voltage and power output against wind speed for all tested incidence angles

4 Conclusion

This study explores the effect of different incidence angles towards the performance of a DCF energy harvester. Experimental results show that the critical flutter speed of the device decreases when the incidence angle increased from 0° to 60° about the z-axis, but no flutter activity was recorded when the flutter beams were parallel to the wind flow (90°). Likewise, the flutter amplitude also decreases with increasing incidence angle which caused a significant drop of up to 76.1% in the power output. When the DCF was rotated 90° about the x-axis, the device demonstrated a similar trend in terms of amplitude and power output with the 0° case about the z-axis but with a 16.3% smaller critical flutter speed making it the most promising orientation. The flutter frequency, however, remained the same in all five experiments. To conclude, some important considerations for the DCF energy harvester are listed here:

1. While non-perpendicular wind direction provides a slight increase in bandwidth, the trade-off in power is too large. Hence, it is important to orientate these devices perpendicular to the direction of the majority of the wind flow in practical applications.
2. Although the 90° angle about the x-axis displayed the most promising output, it may be difficult to create arrays of the device under this orientation.
3. The power output predicted in this study is still considerably low and has very limited applications. Hence, future works must focus on optimizing the DCF energy harvester. Additionally, the variation in incidence angle about the y-axis can also be investigated.






Acknowledgements This work was supported by the Fundamental Research Grant Scheme (FRGS) from the Ministry of Higher Education (MOHE) Malaysia, Grant No: FRGS/1/2021/TK0/UTM/02/3.

References

1. Elahi H, Munir K, Eugeni M, Atek S, Gaudenzi P (2020) Energy harvesting towards self-powered IoT devices. *Energies* 13:5528
2. Hidalgo-Leon R, Urquizo J, Silva CE, Silva-Leo J, Wu J, Singh P, Soriano G (2022) Powering nodes of wireless sensor networks with energy harvesters for intelligent buildings: a review. *Energy Rep* 8:3809–3826
3. Zhao L, Yang Y (2017) Toward small-scale wind energy harvesting: design, enhancement, performance comparison, and applicability. *Shock Vib* 3595972
4. Gong Y, Yang Z, Shan X, Sun Y, Xie T, Zi Y (2019) Capturing flow energy from ocean and wind. *Energies* 12(11):2184
5. Tang B, Fan X, Wang J, Tan W (2022) Energy harvesting from flow-induced vibrations enhanced by meta-surface structure under elastic interference. *Int J Mech Scie* 236:107749
6. Lee YJ, Qi Y, Zhou G, Lua KB (2019) Vortex-induced vibration wind energy harvesting by piezoelectric MEMS device in formation. *Scie Rep* 9:20404
7. Lu Z, Wen Q, He X, Wen Z (2019) A flutter-based electromagnetic wind energy harvester: theory and experiments. *Appl Sci* 9(22):4823
8. Zhao D, Hu X, Tan T, Yan Z, Zhang W (2020) Piezoelectric galloping energy harvesting enhanced by topological equivalent aerodynamic design. *Energy Conv Manage* 222:113260
9. Hobeck JD, Inman DJ (2016) Dual cantilever flutter: experimentally validated lumped parameter modeling and numerical characterization. *J Fluid Struct* 61:324–338
10. Foong FM, Thein CK, Ooi BL, Yurchenko D (2019) Increased power output of an electromagnetic vibration energy harvester through anti-phase resonance. *Mech Syst Signal Process* 116:129–145
11. Foong, FM, Thein CK, Abdul Aziz AR (2018) Effect of electromagnetic damping on the optimum load resistance of an electromagnetic vibration energy harvester. In: 2nd international conference on smart grid and smart cities. IEEE, Kuala Lumpur, pp 127–132

Temporal Distribution of Thunderstorm Activity in Southern Region of Peninsular Malaysia



Shirley Anak Rufus , N. A. Ahmad , Z. A. Malek ,
Noradlina Abdullah , Nurul 'Izzati Hashim ,
and Noor Syazwani Mansor 

Abstract An eight-year (2011–2018) study of thunderstorm activity in the southern region of Peninsular Malaysia is presented based on the lightning data obtained from the Lightning Detection Network System (LDNS) operated by TNB-Research (TNBR). This study aims to enhance regional knowledge and analyse the temporal variation of thunderstorm activity in the southern part of Peninsular Malaysia. The main findings of this study indicated an increasing pattern of thunderstorm activity within the 8-year period of observation, and the mean annual rate was approximately 163,426 per year. About 13.2% of total CG lightning was +CG lightning, while 86.8% was dominated by -CG lightning. Even though more than 93% of annual lightning was dominated by -CG lightning, the yearly observations revealed that the number of +CG lightning has increased over the years. Therefore, further research and consideration of the -CG and +CG lightning activity is essential in this region in order to enhance and improve the current lightning protection system and predict the lightning and thunderstorm activity, particularly for directly and indirectly affected sectors and industries.

Keywords Thunderstorm · Lightning · Temporal distribution

S. A. Rufus (✉) · N. 'I. Hashim · N. S. Mansor
Department of Electrical and Electronic Engineering, Faculty of Engineering, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
e-mail: rshirley@unimas.my

S. A. Rufus · N. A. Ahmad · Z. A. Malek
School of Electrical Engineering, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 Skudai Johor, Malaysia

Institute of High Voltage and High Current (IVAT), Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 Skudai Johor, Malaysia

N. Abdullah
Lightning and Earthing Unit, TNB Research Sdn Bhd (TNBR), Selangor, Malaysia

1 Introduction

Thunderstorms and lightning vary depending on location, climate, and time. Thunderstorms are often accompanied by lightning, heavy rain, strong winds, and sometimes snow, hail, or no precipitation. Intra-cloud (IC) lightning, inter-cloud (CC) lightning, and cloud-to-ground (CG) lightning may occur at the beginning phase of a thunderstorm. Negative cloud-to-ground (−CG) lightning involves the transfer of negative charges through multiple strokes, whereas positive cloud-to-ground (+CG) lightning involves the transfer of positive charges in a single stroke, accounting for fewer than 5% of lightning strikes. The +CG lightning is more dangerous because of its greater distance, flash duration, and higher peak charges [1, 2]. Malaysia has year-round lightning and thunderstorms due to its tropical climate, monsoon variation, topography, and location near the Equatorial Belt. The Global Lightning Dataset 360 (GLD360-Vaisala) reported 17,738,435 lightning counts and 54.14 events per km² in 2021 for Malaysia, the highest in Southeast Asia. Most states in Malaysia experience more than 50 flashes per km² per year due to their mountainous nature and proximity to the Strait of Malacca (west), the South China Sea (east), and the Strait of Tebrau (south). In 2021, Johor has 3.79 million people and an area of 19,166 km². Johor ranks second in Malaysia for maximum lightning density (91.3 events per km² per year) [3].

Malaysia experiences a high rate of lightning activity, causing human fatalities, damages to electronics and machinery, financial losses, forest fires, and the destruction of agriculture and crops [4]. Lightning incidents increased to 282 cases between 2008 and 2017 [5]. Lightning incidents and fatalities were reported in southern Peninsular Malaysia [6, 7]. From 2001 to 2013, lightning and thunderstorms caused 39% to 61% of transmission line disruptions in the TNB Transmission Network, exceeding previous estimates of 36% [8]. Lightning disruptions were also reported in Brazil, Austria, Indonesia, and China as a result of significant lightning activity [9, 10]. Thus, lightning monitoring, forecasting, and understanding lightning and thunderstorm characteristics and activities in different regions are crucial for minimising lightning damage. Numerous regions, including Europe, North America, Brazil, China, Korea, India, Sri Lanka, Australia, Vietnam, and Thailand [11–19], as well as Malaysia [8, 20–23], have investigated and reported the temporal variation of lightning.

This study presents the findings of the analysis of approximately 1.3 million lightning strikes recorded by the Lightning Location System (LLS) called the Lightning Detection Network System (LDNS) and operated by TNB Research Sdn. Bhd. (TNBR) in the southern region of Peninsular Malaysia from January 1, 2011, to December 31, 2018. The temporal variation of thunderstorm activity was analysed. The main contributions of this study may provide a regional reference for lightning risk assessment and lightning protection in southern Peninsular Malaysia and its environments, benefiting diverse sectors, industries, and humans.

2 Data and Method

The study region covered an area of about a 100-km radius from the School of Engineering, Universiti Teknologi Malaysia (UTM), Johor. The lightning activity data were obtained from LDNS-TNBR, which consists of the date, time, location (longitude and latitude), and peak current (in kA) of the lightning activity. Currently, the LDNS-TNBR is empowered by five running sensors, with detection efficiency up to 95% and location accuracy improved to 250 m. The sensors utilised a combination of Time of Arrival (ToA) and Magnetic Direction Finding (MDF) techniques to detect and locate the real-time lightning activity around Peninsular Malaysia. Detailed information regarding the LDNS-TNBR can be found in the following references: [8, 24, 25]. The temporal variation of thunderstorm activity in the southern region of Peninsular Malaysia was determined using Microsoft Excel and Python 3.9.

3 Results and Discussions

The annual variation of CG lightning in the southern region of Peninsular Malaysia from 2011 to 2018 is depicted in Fig. 1. About 641,543 CG lightning activity were recorded in 2017, which is the highest compared to other years. While 8818 CG lightning activity were recorded in 2012, the mean annual rate was approximately 163,426 per year. In 2017 and 2018, the number of lightning strikes was higher (close to 123,000 flashes) because, in 2015, TNBR upgraded its LDNS, which provided a better detection efficiency of 95% and location accuracy of 250 m for recording the total lightning activity. This may have contributed to the increase in lightning activity detected after 2015.

From Fig. 2, it is found that most of the annual lightning was dominated by – CG lightning with 93% and above, and +CG lightning only accounted for less than

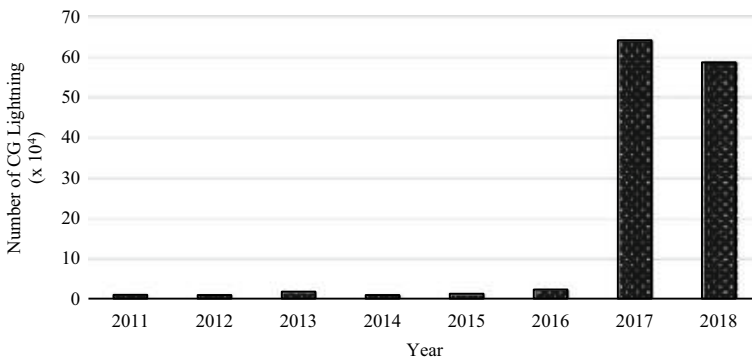


Fig. 1 Annual variation of total CG lightning from 2011 to 2018

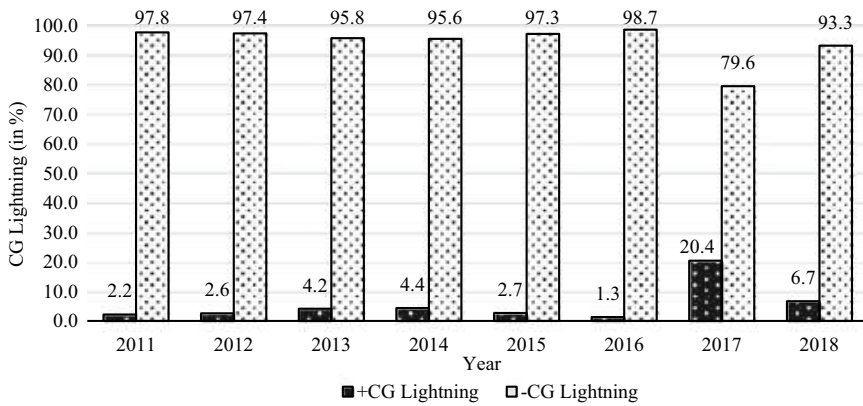


Fig. 2 Annual variations of +CG and –CG lightning from 2011 to 2018

7%. However, the yearly observations showed that the number of +CG lightning increased over the years. In 2017, the +CG lightning rate drastically increased to 20.4%, which is more than the nominal percentage of 10%. The +CG lightning was rarely observed in Malaysia, and the findings of this study are consistent with those of other studies conducted in various locations in Malaysia [8, 20–24, 26].

On the other hand, several studies in Malaysia reported that +CG lightning activity increased and occurred more than the nominal percentage of 10%, which ranged from 14 to 31%. Therefore, necessitating further investigation and study of +CG and –CG lightning in this region is crucial and unpredictable because the +CG lightning carries a greater amount of current and transfers more electrical charges than the –CG lightning.

4 Conclusion

The temporal distribution of thunderstorms in the southern region of peninsular Malaysia was analysed based on approximately 1.3 million lightning strikes recorded by TNBR-LDNS from 2011 to 2018. This study discovered an increasing pattern of lightning activity within the 8-year observation period, especially after the LDNS was upgraded in 2015. About 86.8% was –CG lightning, and only 13.2% was +CG lightning. The yearly variation showed that the +CG lightning increased over the years, with about 20.4% of +CG lightning recorded in 2017 compared to other years only that ranged from 2.2% to 6.7%. The basic characteristic and temporal variation of thunderstorm activity in the southern region may be beneficial to various sectors and industries. Overall, these findings revealed interesting trends that aligned with those obtained for tropical regions, such as the relationship between thunderstorm activity and geographical location in the southern region of Peninsular Malaysia.

Acknowledgements The authors would like to thank Faculty of Engineering, Universiti Malaysia Sarawak (UNIMAS) and Universiti Teknologi Malaysia (UTM), which provided financial and management support for this research. The cooperation from lightning technical expert Ir. Noradlina Abdullah and TNBR's technical staff is highly appreciated. This research uses data provided by the Lightning Detection System Network (LDSN), managed by the Lightning Detection System Laboratory, TNB Research Sdn. Bhd. (TNBR), for research and educational purposes.

References

1. Welcome to Vaisala's Interactive Global Lightning Density Map!. <https://interactive-lightning-map.vaisala.com/>, Jan 2022. https://interactive-lightning-map.vaisala.com/?_ga=2.251216771.448933359.1670901480-1779715020.1670901112
2. Holle RL, Dewan A, Said R, Brooks WA, Hossain MdF, Rafiuddin M (2019) Fatalities related to lightning occurrence and agriculture in Bangladesh. *Int J Disaster Risk Reduction* 41:101264. <https://doi.org/10.1016/j.ijdrr.2019.101264>
3. Riajati (2018) Malaysians unprepared for lightning strikes, 10 Nov 2018. <https://www.rj.my/2018/11/10/malaysians-unprepared-for-lightning-strikes/>
4. Bernama (2017) Indonesian man killed by lightning strike in Johor, Malaysia (Online). Available: <https://peraktoday.com.my/2017/02/indonesian-man-killed-by-lightning-strike-in-johor/>
5. Devi V (2021) Construction worker killed by lightning strike in JB, The Star, Johor Bahru, Malaysia (Online). Available: <https://www.thestar.com.my/news/nation/2021/05/30/construction-worker-killed-by-lightning-strike-in-jb>
6. Rawi IM, Abidin Ab Kadir MZ, Gomes C, Azis N (2018) A case study on 500 kV line performance related to lightning in Malaysia. *IEEE Trans Power Deliv* 33(5):2180–2186. <https://doi.org/10.1109/TPWRD.2017.2787660>
7. Zoro R (2019) Tropical lightning current parameters and protection of transmission lines. *Int J Electrical Eng Inf* 11(3):506–514. <https://doi.org/10.15676/ijeei.2019.11.3.4>
8. Diendorfer G, Pichler H, Achleitner G, Broneder M (2014) Lightning caused outages in the Austrian Power Grid transmission line network. In: 2014 international conference on lightning protection, ICLP 2014, Institute of Electrical and Electronics Engineers Inc, Dec 2014, pp 152–156. <https://doi.org/10.1109/ICLP.2014.6973112>
9. Tsenova BD, Gospodinov I (2022) Temporal and spatial distribution of lightning activity over Bulgaria during the period 2012–2021 based on ATDnet lightning data. *Climate* 10(11):184. <https://doi.org/10.3390/cli10110184>
10. Van Mai K, Laurila TK, Hoang LP, Duc Du T, Mäkelä A, Kiesiläinen S (2022) Thunderstorm activity and extremes in Vietnam for the period 2015–2019. *Climate* 10(10). <https://doi.org/10.3390/cli10100141>
11. Jayawardena IMSP, Mäkelä A (2021) Spatial and temporal variability of lightning activity in Sri Lanka. In: Multi-hazard early warning and disaster risks. Springer International Publishing, pp 573–586. https://doi.org/10.1007/978-3-030-73003-1_39
12. Makela A, Haapalainen J, Makela J (2017) Estimation of lightning hazard of an approaching thunderstorm. In: 2010 30th international conference on lightning protection, ICLP 2010. <https://doi.org/10.1109/ICLP.2010.7845831>
13. Bourscheidt V, Pinto O, Naccarato KP (2014) Improvements on lightning density estimation based on analysis of lightning location system performance parameters: Brazilian case. *IEEE Trans Geosci Remote Sens* 52(3):1648–1657. <https://doi.org/10.1109/TGRS.2013.2253109>
14. Xu M et al (2022) Lightning climatology across the Chinese continent from 2010 to 2020. *Atmos Res* 275:106251. <https://doi.org/10.1016/j.atmosres.2022.106251>

15. Moon SH, Kim YH (2020) Forecasting lightning around the Korean Peninsula by postprocessing ECMWF data using SVMs and undersampling. *Atmos Res* 243. <https://doi.org/10.1016/j.atmosres.2020.105026>
16. Mondal U, Panda SK, Das S, Sharma D (2022) Spatio-temporal variability of lightning climatology and its association with thunderstorm indices over India. *Theor Appl Climatol* 149(1–2):273–289. <https://doi.org/10.1007/s00704-022-04032-5>
17. Safronov AN (2022) Spatio-temporal assessment of thunderstorms' effects on wildfire in Australia in 2017–2020 using data from the ISS LIS and MODIS space-based observations. *Atmosphere (Basel)* 13(5):662. <https://doi.org/10.3390/atmos13050662>
18. Rufus SA, Ahmad NA, Abdul-Malek Z, Abdullah N (2019) Characteristics of lightning trends in peninsular Malaysia from 2011 to 2016. In: 2019 International Conference on Electrical Engineering and Computer Science (ICECOS), Batam, Indonesia, IEEE, Oct 2019, pp 15–18. <https://doi.org/10.1109/ICECOS47637.2019.8984514>
19. Taufik ANA et al (2022) Lightning observation around tall structures in Kuala Lumpur, Malaysia. In: 2022 IEEE international conference in power engineering application, ICPEA 2022—Proceedings, Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/ICPEA53519.2022.9744646>
20. Johari D, Aiman Misri Amir MF, Hashim N, Baharom R, Haris FA (2021) Positive cloud-to-ground lightning observed in Shah Alam, Malaysia based on SAFIR 3000 lightning location system. In: ICPEA 2021—2021 IEEE International Conference in Power Engineering Application, Institute of Electrical and Electronics Engineers Inc, Mar 2021, pp 178–182. <https://doi.org/10.1109/ICPEA51500.2021.9417761>
21. Munauwer AF, Baharudin ZA, Hanafiah MAM, Zainon M, Salim SNS, Ibrahim M (2020) Unravel the extent of the existence of positive ground flash in Malaysia. *Int J Emerging Trends Eng Res* 8(1.1):166–169. <https://doi.org/10.30534/ijeter/2020/2681.12020>
22. Abdullah N, Hatta NM (2012) Cloud-to-ground lightning occurrences in Peninsular Malaysia and its use in improvement of distribution line lightning performances. In: 2012 IEEE international conference on Power and Energy (PECon), IEEE, Dec 2012, pp 819–822. <https://doi.org/10.1109/PECon.2012.6450330>
23. Mohd Hatta N, Abdullah N, Yahaya MP, Abd Rahman NA, Reffin MS (2016) TNBR lightning detection system network: the latest generation of lightning sensors and the optimum placement for enhanced detection performance. In: MyHVnet Colloquium, Johor, p 33, Jan 2016 (Online). Available: <https://ivat.utm.my/files/2018/11/MyHVnet-Newsletter-2016.pdf>
24. Hatta NM, Abdullah N, Osman M, Abidin Ab Kadir MZ (2019) Statistical lightning study for 33kV overhead line in peninsular Malaysia. In: 2019 11th Asia-Pacific international conference on lightning, APL 2019, Institute of Electrical and Electronics Engineers Inc, Jun 2019. <https://doi.org/10.1109/APL.2019.8816020>
25. Baharudin ZA, Ahmad NA, Mäkelä JS, Fernando M, Cooray V (2014) Negative cloud-to-ground lightning flashes in Malaysia. *J Atmos Sol Terr Phys* 108:61–67. <https://doi.org/10.1016/j.jastp.2013.12.001>
26. Wooi CL, Abdul-Malek Z, Ahmad NA, Mokhtari M, Khavari AH (2016) Cloud-to-ground lightning in Malaysia: a review study. *Appl Mech Mater* 818:140–145. <https://doi.org/10.4028/www.scientific.net/amm.818.140>

Electronic Design and Applications

Enhance the AlGaIn/GaN HEMTs Device Breakdown Voltage by Implementing Field Plate: Simulation Study



Naeemul Islam, Mohamed Fauzi Packeer Mohamed,
Firdaus Akbar Jalaludin Khan, Nor Azlin Ghazali, Hiroshi Kawarada,
Mohd Syamsul, Alhan Farhanah Abd Rahim, and Asrulnizam Abd Manaf

Abstract Recently, the features of AlGaIn/GaN high electron mobility transistors (HEMTs) known as breakdown voltage (BV) have garnered a lot of interest for RF and Power applications. But due to the electric field and current collapse, the breakdown voltage of the GaN HEMTs device is reduced. Therefore, in this research, the field plate technique has been studied for enhancing the GaN HEMTs device breakdown voltage by using Silvaco TCAD software. It's observed that the dual field plate has shown a higher breakdown voltage around 1100 V, whereas the gate field plate and source field plate have illustrated a breakdown voltage of 820 and 1000 V approximately. Subsequently, GaN HEMTs presented a threshold voltage (V_{TH}) of -3.3 V and transconductance (G_M) of 16.3 mS/mm approximately.

Keywords Gallium nitride · Semiconductor devices · Wide bandgap · Breakdown voltage · Field plate

N. Islam · M. F. P. Mohamed (✉) · F. A. J. Khan · N. A. Ghazali · A. A. Manaf
School of Electrical and Electronic Engineering, Universiti Sains Malaysia, 14300 Nibong Tebal,
Pulau Pinang, Malaysia
e-mail: fauzi.packkeer@usm.my

H. Kawarada
Faculty of Science and Engineering, Waseda University, Tokyo 169-8555, Japan

M. Syamsul
Institute of Nano Optoelectronics Research and Technology (INOR), Universiti Sains Malaysia,
11900 Bayan Lepas, Pulau Pinang, Malaysia

A. F. A. Rahim
Faculty of Electrical Engineering, University Teknologi MARA, Cawangan Pulau Pinang, 13500
Pulau Pinang, Malaysia

A. A. Manaf
Collaborative Microelectronic Design Excellence Center (CEDEC), Universiti Sains Malaysia,
11900 Bayan Lepas, Pulau Pinang, Malaysia

1 Introduction

Due to the remarkable electrical characteristics of Gallium Nitride (GaN) and its associated alloys (high frequency, high carrier concentration, high density, high mobility, and high power), Gallium Nitride-based high electron mobility transistors (also known as GaN-HEMTs) have been under extensive research with many outstanding and special features like big energy band difference, low noise, high breakdown field, low thermal impedance, and high saturation velocity [1, 2]. Thus, many market-driven industries, like radio frequency, power devices, high-power conversion, high-frequency communication, photonics, and control, have reported using GaN-based devices [3]. However, GaN HEMTs face various reliability issues that may eventually cause degradation, such as high-temperature environments, high electric fields or thin films, high leakage current, current collapse, and low breakdown voltage [4].

Electrons are trapped at defects in the AlGaIn/GaN layers, and the passivation interface causes a current collapse. When a high applied voltage is used, the electric field accelerates the channel electrons, and the device captures some of the accelerated electrons. The applied voltage increases the ON resistance because, in the ON state, the 2DEG channel is depleted by trapped electrons. Consequently, the phenomenon of collapse is influenced by the electric field. In order to minimize collapse, the field plate structure could be used because it decreases the concentration of the electric field [4]. It also helps the GaN HEMTs device improve breakdown voltage where several field plate methods have been used, like gate field plate, source field plate, and drain field plate [5–7].

This study applied the field plate method to enhance the AlGaIn/GaN HEMTs device breakdown voltage. It's noticed that a dual field plate is more effective than a gate field plate or a source field plate because it's reduced the electrical field and improves the breakdown voltage ~ 1100 V, which is reliable for the device.

2 Device Design

Utilizing TCAD SILVACO software, the simulation work is carried out in this research. Figure 1 depicts the GaN HEMTs structural layout [8]. The substrate consists of a GaN buffer layer with a thickness of $0.18 \mu\text{m}$, and then an AlGaIn barrier layer with a thickness of $0.02 \mu\text{m}$ is placed on top of it. The doping concentration of the GaN buffer layer and AlGaIn barrier layer are assumed to be 10^{15}cm^{-3} and 10^{16}cm^{-3} , respectively. Si_3N_4 is used as a passivation layer above the AlGaIn layer. Moreover, the gate length of the device is $0.4 \mu\text{m}$, and with a work function of 5.23eV , the gate electrode created for the structure is assumed to be made of metal.

The gate is $0.5 \mu\text{m}$ away from the source, and from the drain is $5.1 \mu\text{m}$. The drain and source contact lengths are $0.1 \mu\text{m}$. Following that, the field plate included on the source side as well as the gate side. As for simulation, various models have been

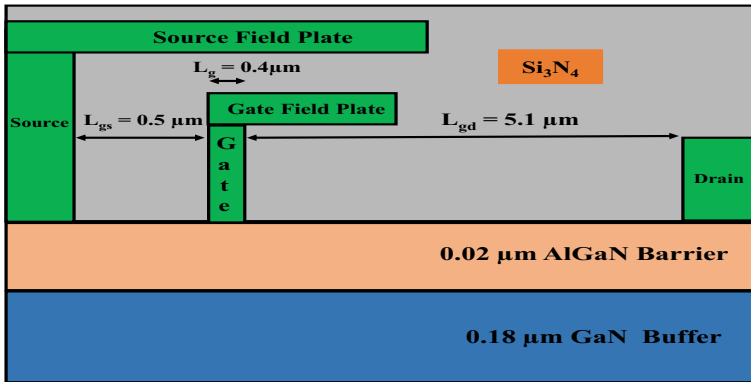


Fig. 1 AlGaIn/GaN HEMT device layout with a field plate

incorporated for it, such as for the Mobility model, the mobility depending on the field is used; The Fermi–Dirac model uses Fermi statistics; For carrier generation and recombination, the Shockley–Read–Hall model is used; domain-related mobility model also included here; the Impact Ionization model is to calculate the breakdown characteristics which mechanism is modeled as $\alpha_0 \exp(-E_C/E)$, here E_C is breakdown field of $3.4 \times 10^7 \text{ V cm}^{-1}$ and $2.9 \times 10^8 \text{ cm}^{-1}$, ionization coefficient is called α_0 [9]. Calculated strain and polarization are evoked for the epitaxial strain caused by lattice mismatch and spontaneous polarization. Subsequently, the autonr method, namely Newton-Richardson Method, is a variant of the Newton iteration; this only creates an updated coefficient matrix when slowing convergence indicates that this is required [10].

3 Result and Discussion

3.1 Characteristics Curve

The transfer and transconductance curves for a GaN HEMTs are displayed in Fig. 2a. Where the threshold voltage (V_{TH}) and transconductance (G_M) are obtained as -3.3 V and 16.3 mS/mm , respectively. In the drain current (I_{DS}) versus drain-source voltage (V_{DS}) characteristic curves, the GaN HEMTs presented a high maximum drain current of 1400 (mA/mm) at $V_{GS} = 0 \text{ V}$, as depicted in Fig. 2b.

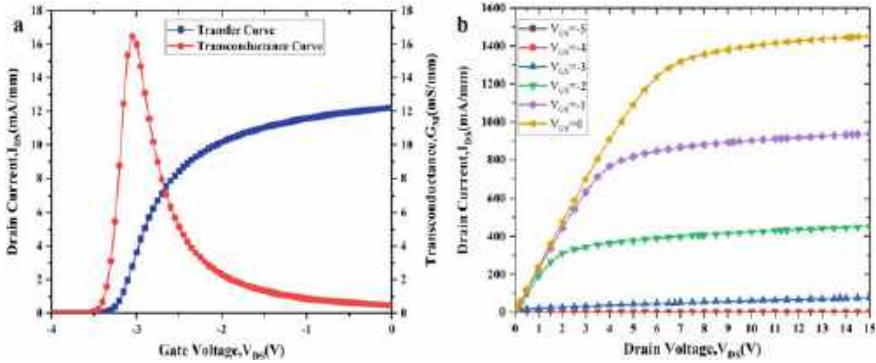


Fig. 2 AlGaIn/GaN HEMT device **a** transfer characteristics and transconductance curve **b** I versus V characteristics curve

3.2 Field Plate Simulation

In this research, three types of field plates simulated. At the gate electrode edge, the electric field impacted by the field plate electrode connection due to the gate field plate and source field plate having distinct effects on the gate edge when it comes to the gate-source voltage. The source and gate field plates are combined to create the dual field plate structure. Increasing the field plate length makes it possible to minimize the electric field at the gate edge. Meanwhile, a lengthy field plate lowers the BV because of the isolation breakdown between the drain and field plate electrodes. Despite the fact that the gate-drain offset length L_{gd} can raise the breakdown voltage, an increase has been made in the static ON resistance. Figure 3 reveals the distribution of electric fields in various field plates.

After cutting along the black line of those field plates, the measurement value is presented in Fig. 4, which displays the field plate structure's mechanism. It's observed that the gate-edge peak for the dual field plate and gate field plate is lower than the source field plate due to the relaxation of the electric field by the gate field plate electrode [4].

Earlier, by changing the gate field plate length such as 1, 1.25, 1.50, and 1.75 μm , the impact of the GaN HEMT device on the breakdown voltage was reported [8], as illustrated in Fig. 5a. It was noticed that when the gate field plate length rises from 2.5 to 3.25 μm , the BV also increases from 530 to 820 V. On the contrary, the source field plate BV displayed in Fig. 5b.

The BV grows when the source field plate length varies from 2.9 to 3.65 μm . The BV reaches the maximum of 1000 V at a source field plate length of 3.65 μm . However, for dual (gate + source) field plate structure, the BV improves highest value around 1100 V when the gate and source field plate maximum lengths are 2.75 and 3.15 μm , as indicated in Fig. 5c.

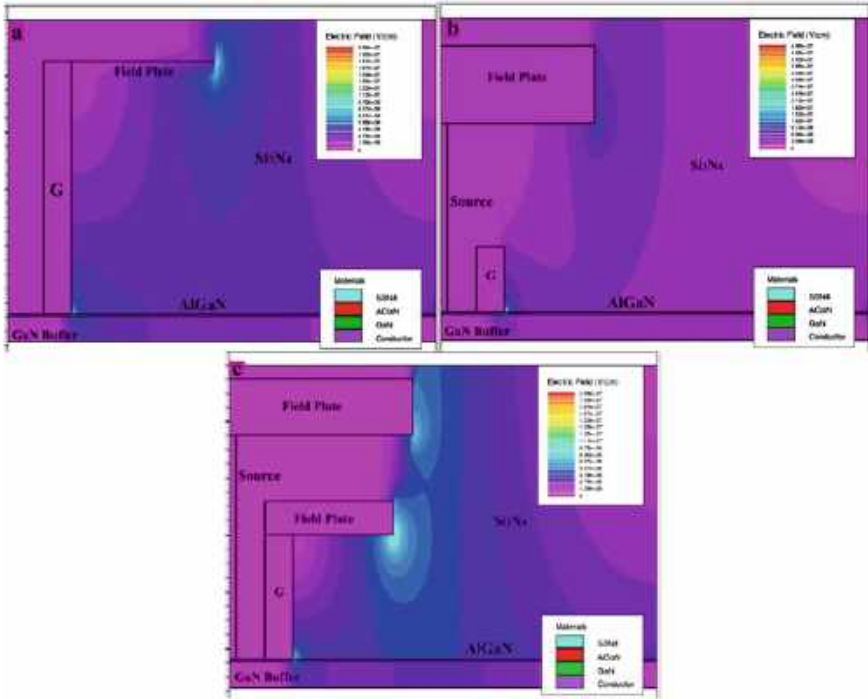
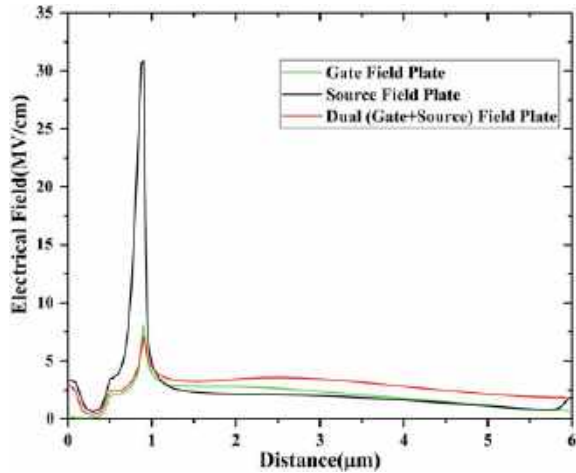


Fig. 3 Various types of AlGaIn/GaN HEMT device field plate **a** gate field plate **b** source field plate **c** gate + source (dual) field plate

Fig. 4 Electric field of various types of field plate from simulation result



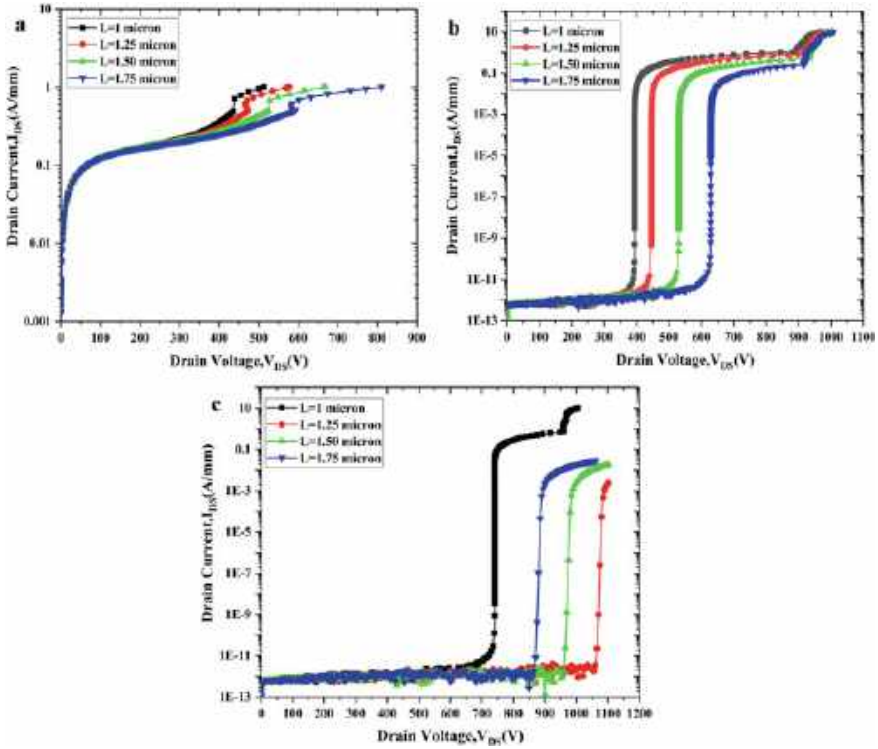


Fig. 5 Breakdown voltage of AlGaIn/GaN HEMT device with **a** gate field plate, **b** source field plate, **c** dual (gate + source) field plate

4 Conclusion

This research uses the field plate method to enhance the BV of AlGaIn/GaN HEMT devices. The gate and source field plates have shown the highest breakdown voltage, around 820 and 1000 V, respectively, at the maximum lengths of 3.25 and 3.65 μm . In contrast, the breakdown voltage of the dual field plate reached the maximum of ~ 1100 V, which increased around 34 and 10% compared to the gate and source breakdown voltage. Moreover, it reduces the electrical field, making the device more reliable.

Acknowledgements This research was funded by Universiti Sains Malaysia Research University Incentive (RUI) grant “1001/PELECT/8014134”. The authors would like to express his gratitude to USM School of Electrical and Electronic Engineering, CEDEC, and INOR for providing research facilities, also to University Technology MARA, Penang for SILVACO TCAD tool support.

References

1. Islam N et al (2022) Reliability, applications and challenges of GaN HEMT technology for modern power devices: a review. *Crystals* 12(11):1581
2. Mishra UK, Parikh P, Wu Y-F (2002) AlGaIn/GaN HEMTs-an overview of device operation and applications. *Proceed IEEE* 90(6):1022–1031
3. Meneghini M et al (2021) GaN-based power devices: physics, reliability, and perspectives. *J Appl Phys* 130(18):181101
4. Saito W et al (2010) Field-plate structure dependence of current collapse phenomena in high-voltage GaN-HEMTs. *IEEE Electr Dev Lett* 31(7):659–661
5. Akiyama S et al (2019) Analysis of breakdown voltage of field-plate AlGaIn/GaN HEMTs as affected by buffer layer's acceptor density. *Jpn J Appl Phys* 58(6):068003
6. Mao W et al (2016) Analysis of the modulation mechanisms of the electric field and breakdown performance in AlGaIn/GaN HEMT with a T-shaped field-plate*. *Chin Phys B* 25(12):127305
7. Zhao SL et al (2013) Influence of a drain field plate on the forward blocking characteristics of an AlGaIn/GaN high electron mobility transistor. *Chin Phys B* 22(11):117307
8. Karmalkar S, Mishra UK (2001) Enhancement of breakdown voltage in AlGaIn/GaN high electron mobility transistors using a field plate. *IEEE Trans Electr Dev* 48(8):1515–1521
9. Shi N et al (2022) Optimization AlGaIn/GaN HEMT with field plate structures. *Micromachines* 13(5):702
10. Silvaco IJSC (2011) ATLAS user's manual

Design of Low-Power and Area-Efficient Square Root Carry Select Adder Using Binary to Excess-1 Converter (BEC)



Poh Yui Lyn, Nor Azlin Ghazali, Mohamed Fauzi Packeer Mohamed, and Muhammad Firdaus Akbar

Abstract The Carry Select Adder (CSLA) is commonly used in VLSI design applications like data-processing processors, ALUs, and microprocessors to perform fast arithmetic operations. Compared to primitive designs like Ripple Carry Adder and Carry Look Ahead Adder, the regular CSLA offers optimized results in terms of area. However, it is still possible to reduce the area and power consumption of CSLA by implementing a simpler and more efficient gate-level modification. In this work, all the CSLA structures were designed using Verilog HDL while pre-layout simulation and synthesis were done using Quartus Prime, ModelSim and Synopsys EDA tools. The final results analysis obtained have proven that the BEC-based SQRT CSLA is better than regular square root CSLA (SQRT CSLA) as it has reduced total cell area by 19.54 (16-bit) and 19.44% (32-bit) as well as reduced total dynamic power by 8.52 (16-bit) and 8.75% (32-bit). Ultimately, the modified SQRT CSLA structure using BEC method showed significant lower dynamic power consumption and smaller cell area than the regular SQRT CSLA.

Keywords Binary to excess-1 converter (BEC) · Carry select adder · Low-power · Area efficient · Verilog

1 Introduction

In this new era of nanoelectronics, the demand for high-speed arithmetic units in micro-processors, image processing units and digital signal processing (DSP) chips is increasing rapidly due to the intensive development in digital electronics field. The development of high-speed adders is very important in the most commonly used arithmetic operation in digital signal processing applications and it acts as the basic building block for synthesis of arithmetic computations [1, 2]. The ripple carry adder

P. Y. Lyn · N. A. Ghazali (✉) · M. F. P. Mohamed · M. F. Akbar
School of Electrical and Electronic Engineering, Universiti Sains Malaysia, 14300 Nibong Tebal, Penang, Malaysia
e-mail: azlin.ghazali@usm.my

(RCA) is composed of many cascaded single-bit full-adders. The circuit architecture is simple and area-efficient, but it suffers from increased propagation delay due to the dependency on previous stage outputs. To address this, the carry select adder (CSLA) was introduced, which independently generates multiple carries and selects the appropriate one to generate the sum [3, 4]. CSLA improves computation speed by performing parallel additions in different groups [5–8]. However, the regular CSLA still has drawbacks in terms of area efficiency and power consumption due to duplicated adders [9]. To overcome this, an area-efficient CSLA can be designed by removing duplicated adder cells and utilizing multiplexers to select outputs based on previous carryout signals [2].

This work aims to reduce the cell area size and minimize dynamic power in the square root CSLA (SQRT CSLA) by manipulating the carry select adder block. SQRT CSLA, an extension of the linear CSLA, equalizes the delay and area of carry chains and block multiplexer signals [10]. It facilitates the implementation of large bit-width adders with lower delay by using cascaded CSLAs to provide a parallel path for carry propagation [11]. SQRT CSLA is chosen for modification and evaluation due to its balanced delay, lower power requirements, and improved area efficiency compared to regular CSLA [12, 13].

2 Regular SQRT Carry Select Adder (CSLA)

The carry-select adder includes two RCAs and a multiplexer for addition. It calculates the sum and carry twice, assuming carry-in values of 0 and 1, and selects the correct results based on the obtained carry [4]. The size of each carry select block can be uniform or variable, with variable sizing using the square root of the number of bits as the ideal number of full-adder elements per block to determine the delay. The 16-bit regular SQRT CSLA consists of five groups of RCAs and multiplexers of different sizes, with each RCA and its attached multiplexers performing two parallel additions assuming carry-in values of 0 and 1 [1].

The structure of the 16-bit regular SQRT CSLA is shown in Fig. 1. It consists of five groups of different size RCAs and multiplexers. Each of the RCAs starting from 2-bit RCA until 5-bit RCA and each attached multiplexers will perform two additions in parallel, one assuming a carry-in (C_{in}) of zero and another one with carry-in (C_{in}) of one.

3 Binary to Excess-1 Conversion (BEC) of SQRT CSLA

The proposed design in the regular CSLA involves replacing the conventional RCA with a BEC to achieve a smaller area and higher execution speed. The BEC logic is used with $C_{in} = 1$ [9] to restore the logic in the RCA for unlike bits. The BEC logic requires fewer logic gates compared to the n -bit full adder, making it advantageous.

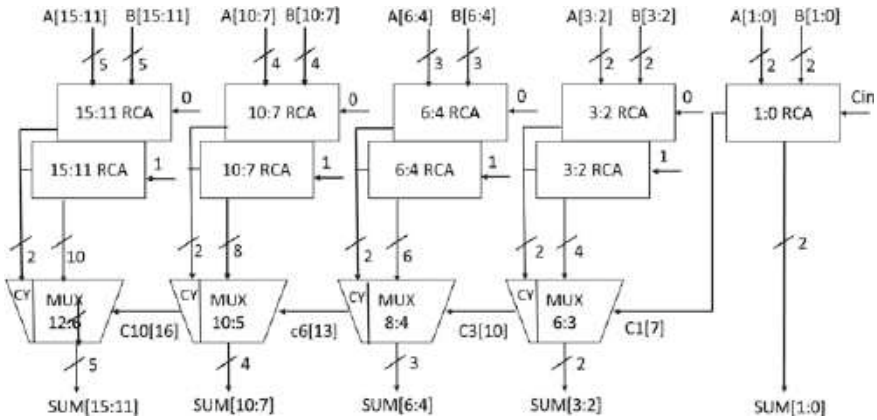


Fig. 1 Regular 16-bit SQRT CSLA [14]

However, to restore the n -bit RCA, an $n + 1$ bit BEC logic is needed. The BEC-based CSLA structure has a simpler logic structure with fewer resources than RCA, but it has a slightly higher carry propagation delay [8].

The modified structure in Fig. 2 replaces the conventional RCA with a BEC adder, simplifying the design. Verilog coding was used for ripple carry adders and multiplexers, following the procedure for regular SQRT CSLA design. The 16-bit BEC-based SQRT CSLA has four multiplexer blocks and five groups of RCAs (2-bit to 5-bit) with BEC adders (2-bit to 6-bit). The 32-bit BEC-based SQRT CSLA includes additional RCAs (2-bit to 7-bit) and BEC adders (4-bit to 8-bit) for 32-bit inputs. Both versions use six multiplexer blocks for carry selection. Another Verilog module was written for BEC adder arithmetic. The procedure was repeated as for regular SQRT CSLA.

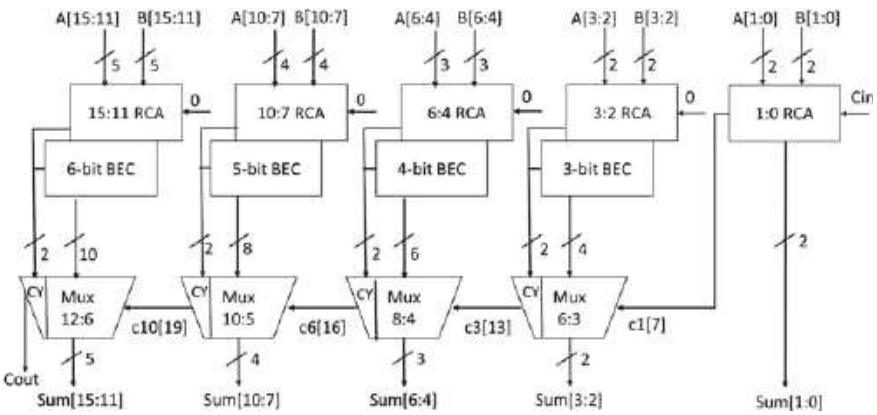


Fig. 2 Modified 16-bit SQRT CSLA; parallel RCA with $C_{in} = 1$ is replaced with BEC [14]

4 Result and Discussion

The regular and modified Sqrt CSLA designs using Verilog HDL were successfully verified with Quartus Prime and ModelSim. The testbench module was used to simulate the designs with different input stimuli. Figure 3 displays the output waveform generated by ModelSim for the 16-bit regular and BEC-based Sqrt CSLAs, while Fig. 4 shows the output waveform for the 32-bit regular and BEC-based Sqrt CSLAs.

The waveforms for the regular and BEC Sqrt CSLA will be identical as the functionality of the CSLA is unaffected by the modification using the BEC method. The modification only affects the area and power consumption of the adder. Therefore, when providing the same input stimulus to both regular and BEC-based Sqrt CSLAs using the testbench, identical outputs were obtained after the arithmetic operation. The values displayed in the output waveforms (Figs. 3 and 4) are in hexadecimal format for easier viewing and calculations.

The designs were synthesized using different libraries for both regular and modified CSLA structures. Three libraries (slow, typical, and fast) were used to observe the impact of operating conditions on area and power. Table 1 displays the synthesis results, including area and power, for the 16-bit and 32-bit regular and BEC-based Sqrt CSLA structures. The area represents the total cell area, and the total dynamic power includes internal power, net switching power, and leakage power.

Based on the synthesis results obtained shown in Table 1, it is proven that the modification of Sqrt CSLA structure using Binary to Excess-1 Conversion is able

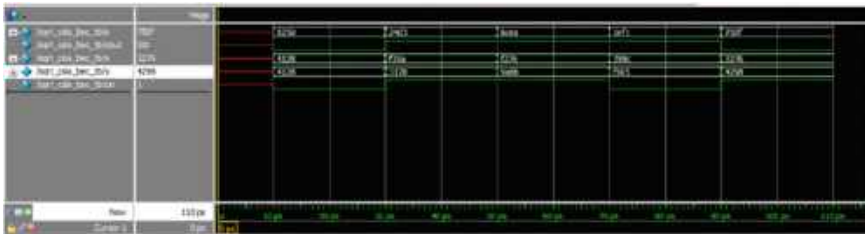


Fig. 3 16-bit regular and BEC-based Sqrt CSLA waveform generated by ModelSim



Fig. 4 32-bit regular and BEC-based Sqrt CSLA waveform generated by ModelSim

Table 1 Table of comparison of the 16-bit and 32-bit Regular and BEC-based Sqrt CSLA in terms of area and power

Type of Sqrt CSLA	16-bit regular Sqrt CSLA			16-bit BEC-cased Sqrt CSLA			32-bit regular Sqrt CSLA			32-bit BEC-cased Sqrt CSLA		
	Slow	Typical	Fast	Slow	Typical	Fast	Slow	Typical	Fast	Slow	Typical	Fast
Total cell area (μm^2)	2774217			22320144			55950049			45072720		
Cell internal power (mW)	1.6358	2.0847	2.7507	1.2933	1.6604	2.1870	3.5151	4.4494	5.8496	2.7074	3.4680	4.5566
Net switching power (mW)	0.8973	1.1517	1.4151	1.0178	1.3003	1.6046	1.6626	2.1128	2.5915	1.9782	2.5203	3.1080
Cell leakage power (nW)	155.98	10.37	23.89	124.77	8.25	18.95	325.80	21.83	50.59	256.69	17.08	39.38
Total dynamic power (mW)	2.5331	3.2363	4.1658	2.3110	2.9607	3.7917	5.1777	6.5622	8.4410	4.6855	5.9883	7.6645

to help in the reduction of area and power. The BEC-based SQR CSLAs have smaller total cell area and lower total dynamic power than the regular SQR CSLAs. The percentage reduction for total cell area of 16-bit and 32-bit proposed SQR CSLAs were 19.54 and 19.44% respectively. Moreover, it is clear that the total dynamic power of the 16-bit and 32-bit proposed SQR CSLAs that synthesized by typical synthetic libraries were reduced by 8.52 and 8.75% respectively.

The slow, fast, and typical libraries had different global operating voltages (1.62 V for slow, 1.8 V for typical, and 1.98 V for fast). The slow condition resulted in high cell leakage power despite reduced area and power consumption, while the fast condition led to increased area and power consumption due to higher current. Therefore, the analysis and comparison of the synthesized results focused on the typical library, representing normal operating conditions. The research objectives of the project were successfully achieved through the design, simulation, and synthesis processes.

5 Conclusion

The SQR CSLA structure is modified using the BEC approach in this work to make it more power and area-efficient. Because fewer logic gates are utilized in the design when adopting the BEC technique, it delivers a significant benefit in terms of reducing overall area and power. The final results obtained in this project has proven that the BEC-based SQR CSLA is better than regular SQR CSLA as it has reduced total cell area by 19.54 (16-bit) and 19.44% (32-bit) as well as reduced total dynamic power by 8.52 (16-bit) and 8.75% (32-bit). Therefore, the modified SQR CSLA using BEC method is more efficient for various VLSI hardware applications. Further work is to be extended and improvised by designing and simulating the adders with increased number of bits, such as 64-bit, 128-bit and 256-bit.

Acknowledgements This project was supported by “Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme” with project code FRGS/1/2020/TKO/USM/02/2.

References

1. Sireesha P, Kumar GR, Bhargavi CP, Sowjanya A (2016) Manga Rao P (2021) Retraction: design and analysis of 32 bit high speed carry select adder. *J Phys Confer Ser* 1:12006
2. Sreekanth G, Singh KJ, Sruthi NS (2016) Design of low power and area efficient carry select adder (CSLA) using verilog language. *Int J Adv Eng Res Sci* 3(12):78–82
3. Nayak VSP, Ramchander N, Reddy RS, Redy THSP, Reddy MS (2016) Analysis and design of low-power reversible carry select adder using D-latch. In: *IEEE international conference on recent trends in electronics, information and communication technology (RTEICT)* pp 1917–1920
4. Nagulapati Giri MD (2019) A survey on various VLSI architectures of carry select adder. *Int J Innov Technol Explor Eng* 8(5):470–474

5. Arunakumari S, Rajasekahr K, Sunithamani S, Suresh Kumar D (2023) Carry select adder using binary excess-1 converter and ripple carry adder. In: Lenka TR, Misra D, Fu L (eds) *Micro and nanoelectronics devices, circuits and systems*. Springer, Singapore, pp 289–294
6. Ramkumar B, Kittur HM (2012) Low-power and area-efficient carry select adder. *IEEE Trans Very Large Scale Integr Syst* 20(2):371–375
7. Prasad G, Nayak VSP, Sachin S, Kumar KL, Saikumar S (2016) Area and power efficient carry-select adder. In: *IEEE international conference on recent trends in electronics, information and communication technology (RTEICT)*, pp 1897–1901
8. Hebbar AR, Srivastava P, Joshi VK (2018) Design of high speed carry select adder using modified parallel prefix adder. *Proced Comput Sci* 12:317–324
9. Gayathri P, Mohan Kumar R (2019) RTL design of efficient high-speed adders using quantum-dot cellular automata. *Int J Recent Technol Eng* 8(2):2853–2857
10. Abhiram T, Ashwin T, Sivaprasad B, Aakash S, Anita JP (2017) Modified carry select adder for power and area reduction. In: *International conference on circuit, power and computing technologies (ICCPCT)*, pp 1–8
11. You H, Yuan J, Tang W, Qiao S (2019) An energy and area efficient carry select adder with dual carry adder cell. *Electronics* 8(10):1254
12. Anagha UP, Pramod P (2015) Power and area efficient carry select adder. In: *IEEE recent advances in intelligent computational systems (RAICS)*, pp 17–20
13. Srinivasareddy B, Anjularani MD (2014) Area-efficient 128-bit carry select adder architecture
14. Syed Mustafaa M, Sathish M, Nivedha S, Mohammed Magribatul Noora AK, Safrin Sifana T (2022) Design of carry select adder using BEC and common boolean logic. *Indian J VLSI Des* 2(1):5–9

Simulation of Bottom-Gate Top-Contact Pentacene Based Organic Thin-Film Transistor Using MATLAB



Law Jia Wei and Nor Azlin Ghazali

Abstract Organic transistor plays an important role in electronic applications as it provides additional benefit of flexibility and low cost compared to silicon electronic devices. Simulation and analytical model of such organic devices helps in improving and optimizing the performance of the device. There are various parameters that effect the performance of the device. In this work, analytical simulation of the organic device is performed based on the device physics to simulate the output characteristics using MATLAB for various channel length (L). Here, the impact on channel length is observed on varying the device length and its impact is observed on the drain current. Different channel length taken into consideration and the simulation result shows that the drain current increases when the channel length decreased. Thus, simulation of such analytical models helps in extracting the useful information about the performance of the organic transistors.

Keywords Organic thin film transistor (OTFT) · Pentacene · MATLAB simulation · Transfer characteristics · Output characteristics

1 Introduction

Organic electronics is gaining significance in both academic and industrial settings, as it holds great potential for diverse applications such as distributed electronics, large-scale displays, arrays of sensors, and photovoltaic devices [1–4]. This is due to the promise for large-scale, low-cost, lightweight, and flexible electronic devices made with such transistors [5]. With the advantages of the organic-based transistors, it also brings that the understanding of the device operation become more important as the manufacturer of organic electronic circuits become complex in terms of the performance of the device.

L. J. Wei · N. A. Ghazali (✉)

School of Electrical and Electronic Engineering, Universiti Sains Malaysia, 14300 Nibong Tebal, Penang, Malaysia

e-mail: azlin.ghazali@usm.my

For circuit designers and for the advancement of organic circuits, the simulation of organic circuits is particularly useful. The proper representation of organic thin film transistors (OTFTs) in commercial circuit simulation tools is essential for the design of organic electronic circuits. A precise model is required to characterize the behaviour of fundamental organic devices during simulation [6]. Additionally, modelling and circuit simulation can assist the precise extraction of transistor model parameters, which could also be a time-saving operation [7]. In this work, simulation study of a top contact pentacene-based OTFT will be conducted using MATLAB. The channel length used for the simulation of the analytical model is 5, 15, 30 and 50 μm . The result will be investigated in terms of performance parameters such as output characteristics and transfer characteristics. The effect of channel length on the output and transfer characteristic of OTFT will be analyzed.

2 Device Structure

An OTFT is a transistor made up of three electrodes, an insulator layer, and a thin film of a semiconductor that can transmit current. The main distinction between OTFTs and traditional MOSFETs is that the organic material-based device does not have a fourth terminal that is the body, making these transistors prone to the body effect [8]. Second, whereas the accumulation layer forms the conduction channel in OTFT, the inversion layer does so in inorganic-based devices [8]. Figure 1 illustrates the basic structure of the OTFT being simulated in this work, with the channel length denoted as L and its width as W . Pentacene acts as a p-type organic semiconductor (OSC). Silicon dioxide, SiO_2 is used as dielectric with the thickness d_i and dielectric permittivity is ϵ_i with an associated capacity of $C_i = \epsilon_i/d_i$. The semiconductor film has the thickness of d_s and dielectric permittivity, ϵ_s with capacity, $C_s = \epsilon_s/d_s$. The threshold voltage is simplified and set to zero. Hence, threshold voltages can be incorporated into the model by substituting V_{GS} with $V_{GS} - V_{TH}$ where V_{GS} and V_{TH} are gate and threshold voltage respectively [9].

Fig. 1 A Cross-section of TCBG OTFT device structure

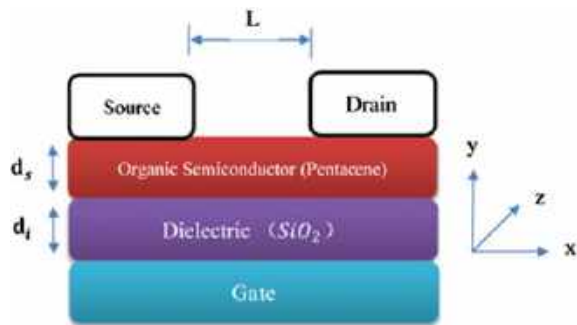


Table 1 Design specification and parameter for OTFT

Design parameter	Notation	Value
Channel length	L	5–50 μm
Channel width	W	2 mm
SiO ₂ thickness	d_i	300 nm
Dielectric permittivity	ϵ_i	$3.9 \times \epsilon_0$
Semiconductor thickness	d_s	100 nm
Semiconductor doping	N_a	10^{16} cm^{-3}
Free carrier density	n_i	$1.45^{10} \text{ cm}^{-3}$
Free carrier mobility	μ_0	$10^{-4} \text{ cm}^{-2}/\text{Vs}$
Fitting parameter	γ	$10^{-5} (\text{cm}/\text{Vs})^{-2}$
Pool–Frenkel factor	β	$4.35 \times 10^{-5} \text{ eV} (\text{cm}/\text{Vs})^{-2}$
Zero field activation energy	Δ	0.1 eV

The design specifications and parameters used in the simulation are summarized in Table 1. The simulation parameters as tabulated in Table 1 were referred from the previous published work in order to obtain material characteristic and parameters similarity thus the device performance could be compared mainly when the device structural dimension changed [10]. The architecture of OTFT used in this work is bottom-gate-top-contact (BGTC). Because organic semiconductors are delicate, it is considerably simpler to deposit them on an insulator than the reverse [11]. Therefore, most of modern OTFTs are constructed using bottom-gate architecture.

3 Analytical Model

The analytical model takes into consideration equations that describe the features and behaviour of the devices. These equations can be used to extract various device characteristics including ON current, OFF current, threshold voltage, and subthreshold slope. The analytical modelling employed standard equation of transistor for linear (Eq. 1) and saturation region (Eq. 2).

In linear region ($V_{DS} < V_{GS} - V_{TH}$), the total drain I_D can be expressed as the sum of bulk current from the free carriers in the semiconductors and the current of the carriers in the accumulation layer. Therefore, after simplification, the I_D can be express as:

$$I_D = \frac{W}{L} \mu_{lin} C_i \left\{ (V_{GS} - V_{TH}) V_{DS} - \frac{1}{2} V_{DS}^2 \right\} \quad (1)$$

where C_i is capacitance per unit area of dielectric material, μ_{lin} is field effect mobility for linear region. Whereas W and L are channel width and length of the OTFT respectively.

When drain to source voltage exceeds the gate voltage ($V_{DS} > V_{GS}$) depletion layer is formed and the device now operates in the saturation region. This depletion layer extends from the coordinate point x_0 to the point of drain contact [12]. Now the drain current is given by:

$$I_{D(sat)} = \frac{W}{2L} \mu_{sat} C_i (V_{GS} - V_{TH})^2 \quad (2)$$

where μ_{sat} is field effect mobility for saturation region. This kind on analytical modelling is useful for understanding and predicting the performance of transistor in various operating conditions.

4 Results and Discussions

Based on the Poole–Frenkel mobility model to demonstrates the channel conduction in Pentacene-based OTFT, which states mobility as:

$$\mu(E) = \mu_0 \exp \left[-\frac{\Delta}{kT} + \left(\frac{\beta}{kT} - \gamma \right) \sqrt{E} \right] \quad (3)$$

where $\mu(E)$ is the field dependent mobility, μ_0 is the zero-field mobility, Δ is the zero-field activation energy, T is the temperature, β is the electron Poole–Frenkel (PF) factor, γ is the fitting parameter, k is the Boltzman constant, and E is the electric field. In analytical modelling, a fixed value of mobility is taken but, various studies revealed that it is actually dependent on the voltage difference between the gate-source and drain-source [12]. The mobility of the OTFT is set to be $0.395 \text{cm}^2/\text{Vs}$ based on the Poole–Frenkel mobility model formula.

Output characteristics obtained for the different channel lengths 50, 30, 15 and 5 μm is shown below in the Fig. 2. From the figure, it is observed that the output characteristics looks similar to that of the conventional transistor [13].

The drain current, I_D increases linearly at low drain voltages (V_{DS}) and I_D becomes saturated at high drain voltages due to a pinch-off of the organic active channel of the fabricated devices. Figure 2a shows that maximum I_D is approximately 71.89 μA for $L = 50 \mu\text{m}$. With the channel length $L = 30 \mu\text{m}$, I_D is maximum at 186.48 μA which shown in Fig. 2b. Figure 2c and d show that the channel length with $L = 15$ and 5 μm have maximum I_D are approximately 237.29 and 711.89 μA respectively. The output current increases with decreasing the channel length at a fixed gate voltage ($V_{GS} = 1.4 \text{V}$). This behavior can be explained by the reduction in channel resistance with decreasing channel length which allows increase the density of charge carriers in the conductive channel of this top contact OTFT [12].

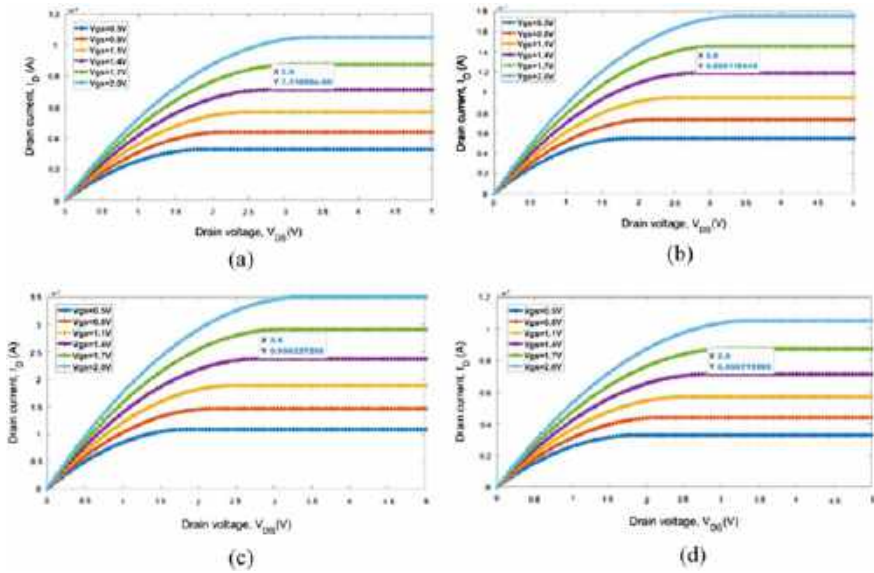


Fig. 2 Output characteristics of OTFT with channel length of **a** 50 μm **b** 30 μm **c** 15 μm and **d** 5 μm

Figure 3 shows the transfer characteristic of the devices with four different channel length plotted in logarithmic scale. The $I_{DS} - V_{GS}$ curve was obtained by sweeping gate bias from 0 to 5 V with drain bias fixed at 1 V. The device performance in terms of the electrical parameters such as threshold voltage, subthreshold slope and current on/off ratio can be extracted from Fig. 3. From Fig. 3, it shown that all the subthreshold slope value is approximately same which is 0.83 V/decade It might due to the constant mobility is used for this work so it did not bring any changes for the subthreshold slope in this case.

5 Conclusions

The performance of pentacene-based OTFT for BGTC structure through the simulation of MATLAB tools are reported in this work. It provides a better understanding of device physics and effects of the channel length to the performance characteristics of OTFT. The simulated device exhibit linear and saturation region through the output and transfer characteristics when employing varying channel lengths. The subthreshold slope remains constant despite variations in channel length. Similar simulation techniques can be used in MATLAB to better understand device behaviour and to optimize various device performances in accordance with the requirements.

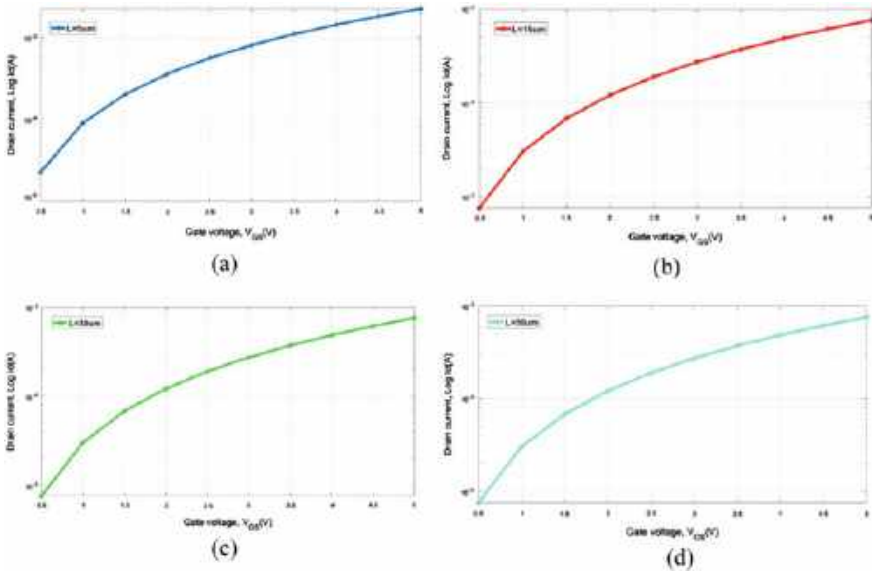


Fig. 3 Semi-logarithmic ($I_{DS} - V_{GS}$) characteristics of OTFT with channel length of **a** $5 \mu\text{m}$ **b** $15 \mu\text{m}$ **c** $30 \mu\text{m}$ and **d** $50 \mu\text{m}$

Acknowledgements This project was supported by “Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme” with project code FRGS/1/2020/TKO/USM/02/2.

References

1. Zhang G, Xie C, You P, Li S (2022) Organic field-effect transistors BT: introduction to organic electronic devices. In: Xie C, You P, Li S (eds) Zhang G. Springer Nature, Singapore, pp 107–129
2. Wang CY, Fuentes-Hernandez C, Chou WF, Kippelen B (2017) Top-gate organic field-effect transistors fabricated on paper with high operational stability. *Org Electron* 41:340–344
3. Wu X et al (2023) High-performance vertical field-effect organic photovoltaics. *Nat Commun* 14(1):1579
4. Lee MY, Lee HR, Park CH, Han SG, Oh JH (2018) Organic transistor-based chemical sensors for wearable bioelectronics. *Acc Chem Res* 51(11):2829–2838
5. Guo X et al (2017) Current status and opportunities of organic thin-film transistor technologies. *IEEE Trans Electron Dev* 64(5):1906–1921
6. Dwivedi ADD, Dwivedi RD, Zhang SD (2020) Numerical simulation and compact modeling of thin film transistors for future flexible electronics. In: Jain SK (ed) Hybrid nanomaterials. IntechOpen, Rijeka
7. Jung S, Kwon J, Tokito S, Horowitz G, Bonnassieux Y (2019) Jung S (2019) Compact modelling and SPICE simulation for three-dimensional, inkjet-printed organic transistors, inverters and ring oscillators. *J Phys D Appl Phys* 52(44):444005

8. Bansal S, Jain P (2022) Simulation of organic thin-film transistor (OTFT) having high carrier mobility. In: Proceedings of the 2022 2nd international conference on advance computing and innovative technologies in engineering (ICACITE), pp 487–490
9. Lu N, Jiang W, Wu Q, Geng D, Li L, Liu M (2018) A review for compact model of thin-film transistors (TFTs). *Micromachines* 9(11):4782
10. Yang Y, Nawrocki RA, Voyles RM, Zhang HH (2021) Modeling of the electrical characteristics of an organic field effect transistor in presence of the bending effects. *Org Electron* 88:106000
11. Nketia-Yawson B, Noh YY (2017) Organic thin film transistor with conjugated polymers for highly sensitive gas sensors. *Macromol Res* 25(6):489–495
12. Singh A, Singh MK (2020) Channel length-dependent performance study of OTFT: analytical modeling using MATLAB. In: Proceedings of the 2020 international conference on advances in computing, communication and materials (ICACCM), pp 301–305
13. Matsui H, Takeda Y, Tokito S (2019) Flexible and printed organic transistors: from materials to integrated circuits. *Org Electron* 75:105432

Acoustic Beamforming Using Machine Learning



Te Meng Ting and Nur Syazreen Ahmad

Abstract This paper shows how two microphones in an endfire array configuration was used to perform beamforming. The setup uses two condenser microphones and a sound card to allow multiple sources to be input to the computer at the same time. A cross-correlation calculation was used to determine the time shift between the two mics. Using the Delay and Sum algorithm, the time shift can be corrected, and the mic signals can be added to a superposition.

Keywords Acoustic · Delay and sum · AoA · Microphones · Cross-correlation

1 Introduction

Acoustic beamforming is a popular technique used in various applications, such as speech enhancement, noise reduction, and source localization [1]. In robotic applications, acoustic beamforming using machine learning algorithms enables robots to enhance specific sound sources of interest while suppressing background noise and interference [2, 3]. This capability can be particularly valuable in scenarios such as human–robot interaction [4, 5], where the robot needs to focus on and understand human speech in a noisy environment [6, 7]. Traditional beamforming algorithms rely on mathematical models and signal processing techniques to enhance the desired sound source while suppressing noise and interference from other directions. However, these methods often require prior information about the acoustic environment and assume stationary sources, which may limit their effectiveness in dynamic or unknown environments [8–10].

T. M. Ting (✉) · N. S. Ahmad
School of Electrical and Electronic Engineering, Universiti Sains Malaysia, 14300 Nibong Tebal,
Penang, Malaysia
e-mail: desmond_ttm@hotmail.com

N. S. Ahmad
e-mail: syazreen@usm.my

This work explores the application of omnidirectional microphones in beamforming technology. Omnidirectional microphones possess the unique ability to equally capture sounds from all directions, making them suitable for beamforming purposes. Beamforming microphone arrays can be designed to enhance sensitivity to sounds originating from specific directions while suppressing noise from other directions. This technique is based on the principle of superposition, where two or more waveforms are combined to form a resultant waveform with higher amplitude. By employing a corrective algorithm based on cross-correlation, the phase of individual microphone signals can be adjusted to achieve proper superposition. This results in a significantly improved Signal-to-Noise Ratio (SNR). The ability to dynamically change the beam direction, known as beam steering, finds applications in various devices such as smart speakers, headsets, and hands-free calling tools in cars. In this paper, we focus on elucidating the process of beamforming using two omnidirectional condenser microphones in the endfire array configuration.

2 Methodology

The concept that makes beamforming work is based on understanding the phase delay and correcting it. In order to correct the phase, a proper way of detecting the phase delay is crucial. Cross correlation works by comparing two signals, in this setup, the signals were first discrete into chunks of 200 blocks and saved into an array. If signals arrive at one mic before the other, we know that there will be a shift when comparing both arrays. By applying cross correlation between first mic (Mic 1) relative to second mic (Mic 2) and subtracting the correlation between Mic 2 and Mic 1, the distance number of chunks shifted can be calculated. The positive or negative sign of the value will indicate the direction. The sampling rate was set at 48 kHz, we can calculate the chunk time by taking the period of the sampling frequency and multiplying it with the delayed chunk. However, in this setup the signal delay was done not by adding a delay block, but by shifting the array by an amount equal to the chunk delay. The shifted cells were then populated with zeros to avoid interference when summed with the other signal. The two arrays are then summed to obtain a super position.

2.1 Polynomial Regression

In statistics, polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an n th degree polynomial in x . The goal of regression analysis is to model the expected value of a dependent variable y in terms of the value of an independent variable (or vector of independent variables) x . In general, we can model the expected value of y as an n th degree polynomial, yielding the following polynomial regression

model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_m x_i^m + \varepsilon_i (i = 1, 2, \dots, n) \quad (1)$$

2.2 Machine Learning

Data points measured are discrete values being a single point in any axis. However, accuracy issues arise when we want to obtain values in between these data points that are not previously measured. An approximation can be made by joining the two points and interpolating the data. This is only accurate for a linear regression. If the lines connecting the two dots were meant to be a polynomial expression, the interpolated value will not be accurate. The way to solve this is to figure out the best fit polynomial equation of the curve joining the lines. Machine learning uses training data to figure out the most optimal equation and uses that to either interpolate or extrapolate values needed [11, 12]. The accuracy of this relies on the number of training data and the number of degrees in the polynomial equation. Beamforming can be achieved with two or more mics. The more mics used, the better the directivity and sensitivity of the output signal. The figure below shows how beamforming is done by adding a corrective delay between different mic inputs. If the sound is coming from a different direction, the end product will not be a superpositioned waveform. By lowering the overall amplitude, and sound coming from the wrong direction will be negated. Since the signal is strongest in one direction only, we say that the beam is steered to that direction. For the signals coming from all the mics to be synchronized, cross-correlation of the different signals can be done. Cross-Correlation measures the similarity between two signals and returns the shifted copy of the second signal. By running a cross-correlation function on two signals, the delay between the two signals can be calculated. The delay correction can be done by adding a delay block in processing at a later stage.

2.3 Setup

An endfire array consists of multiple microphones arranged in line with the desired direction of sound propagation. The maximum delay occurs when signals arrive perpendicular to the direction of the mics. The distance between the two mics paired with the correct delay can form a cardioid. A cardioid pattern has no signal attenuation to the front of the array and theoretically completely cancels the sound incident to the array at 180°. The signals on the sides of a first order (2-microphone) delay-and-sum beamformer are attenuated by 6 dB.

From Fig. 1 (left), it can be observed that the mic response changes with the source's frequency. The distance between the two mics for this setup was set to form

a cardioid at 2–4 kHz which is the range of human voice frequency. The distance between the two mics was set at half wavelength of a 3 kHz source which is $d = 56.67$ mm. Two omnidirectional MEMS mics were used to capture data. Training Data was first captured in an anechoic chamber shown in Fig. 2 (right) to capture the best-case data. Once all the data points were captured, the data was fed as training data of the machine learning algorithm to generate the polynomial equation. The equation generated can be expressed as:

$$y = -657x^3 + 8.331e^4x^2 - 5.16e^5x - 1.169e \tag{2}$$

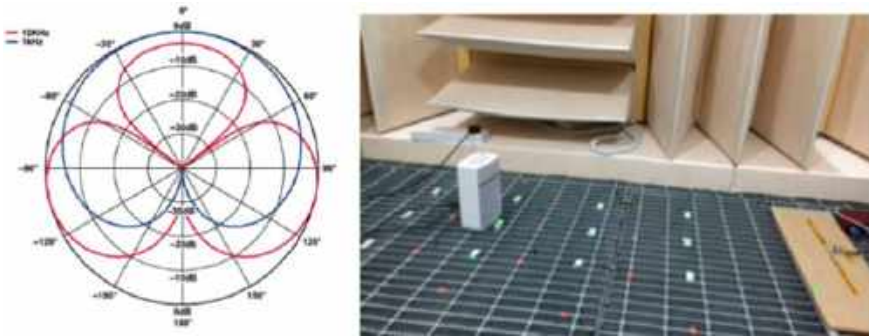


Fig. 1 Frequency aliasing in a 2 microphone endfire beamformer (left). Reference setup in anechoic chamber (right)

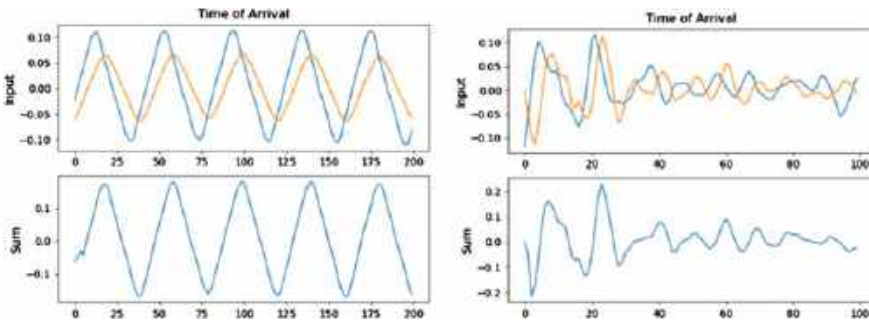


Fig. 2 Two signals shifted to form a super position—tone (left). Two signals shifted to form a super position—speech (right)

3 Results and Discussions

Figure 2 shows the results obtained when a sound source was played from different direction of the microphone array. When the sound source is played in the desired direction, the amplitude is higher than the original signal. When the sound source is played in the undesired direction, the amplitude is less than or equal to the original signal.

It can be observed that the blue waveform arrives first followed by the orange waveform indicating the sound source is coming from a direction where the signal reaches the mic with the blue wave form first. The top graph shows the raw input while the graph below shows the output after phase correction and sum. The amplitude of the output signal is almost double that of the input signal indicating that the signals have been super positioned. Figure 3 shows two waveforms from the two mics.

It can be observed that the orange waveform arrives first followed by the blue waveform indicating the sound source is coming from an angle opposite to the first one. The top graph shows the raw input while the graph below shows the sum of the two signals without any phase correction. Since the signal is coming from an undesired direction, the phase shift algorithm was not carried out. The amplitude of the output signal is about the same as the input signal indicating that the signals were not super positioned. Figure 4 shows two wave forms from the two mics.

When the signal arrives at both mics at the same time, running cross correlation will return a non-shifted value. In the case when the angle of the signal is perpendicular to the axis of both mics, the phase correcting algorithm does not need to correct any phase and allows the signal to super position naturally. Signals coming from the undesired direction can be ignored completely by comparing the time-of-arrival between the two mics. This can be achieved using software. The beamformer in this setup can beam to a resolution of 15° and completely ignores signals from any other direction. However, since only two mics were used, the mics can't distinguish between aliased signals. This, however, can be solved with more than two mics. Once the algorithm was set, a stream of data simulating the actual setup was carried out.

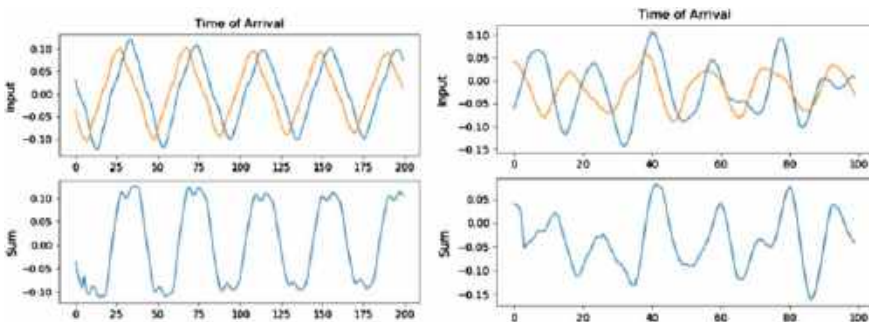


Fig. 3 Two signals not shifted—tone (left). Two signals not shifted—speech (right)

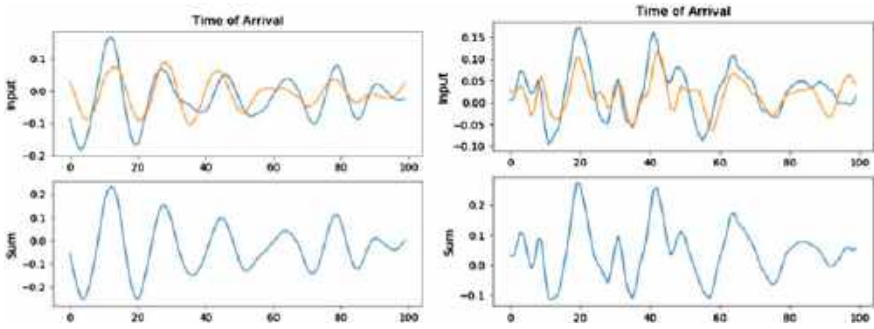


Fig. 4 Signal from equidistant of two mics—tone (left). Signal from equidistant of two mics—speech (right)

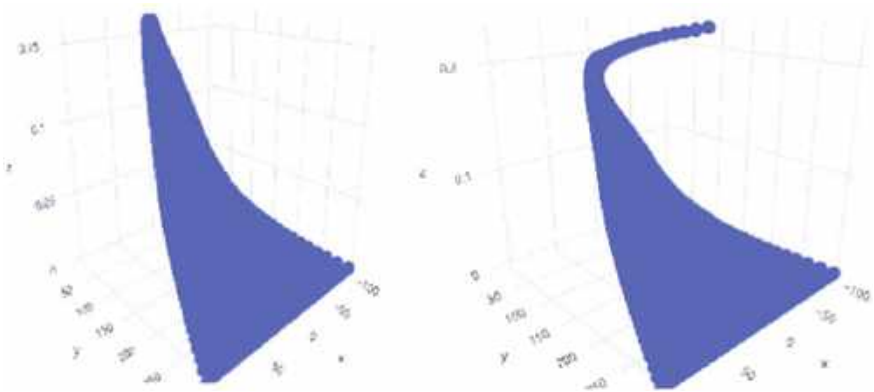


Fig. 5 3D Scatter Plot with Extrapolation (left). 3D Scatter Plot without Extrapolation (right)

The simulation generated data with an increment of 1° and an amplitude of 0.02 dBm. From this, a 3D map of the theoretical measurement was obtained.

Figure 5 shows a 3D scatter plot where the data is extrapolated beyond what was measured. The curve on the top is what the algorithm thinks the output would be given new data. The figure on the right shows another 3D scatter plot without extrapolation. Since the new data is within the range of the training data, the algorithm does not need to extrapolate excessively.

4 Conclusion

In conclusion, the setup in this document can obtain the desired results by performing the delay and sum algorithm on signals coming from the desired direction. This, however, is still a very basic form of beamforming and there are still a lot of things

that can be done to optimize it. Future work will include echo cancellation and anti-aliasing setup. The current setup was done on a computer with an abundance of computational resources. Since the concept is proven to work, the next step would be to port the method to an embedded system.

References

1. Chiariotti P, Martarelli M, Castellini P (2019) Acoustic beamforming for noise source localization: reviews, methodology and applications. *Mech Syst Sig Process* 120:422–448
2. Ting TM, Ahmad NS, Goh P, Mohamad-Saleh J (2021) Binaural modelling and spatial auditory cue analysis of 3D-printed ears. *Sensors* 21(1):227
3. Ahmad NS (2020) Robust H_∞ -fuzzy logic control for enhanced tracking performance of a wheeled mobile robot in the presence of uncertain nonlinear perturbations. *Sensors* 20(13):7673
4. Syed Mubarak Ali SAA, Ahmad NS, Goh P (2019) Flex sensor compensator via hammerstein-wiener modeling approach for improved dynamic goniometry and constrained control of a bionic hand. *Sensors* 19(18):3896
5. Ahmad NS, Boon NL, Goh P (2018) Multi-sensor obstacle detection system via model-based state-feedback control in smart cane design for the visually challenged. *IEEE Access* 6:64182–64192
6. Teo JH, Ahmad NS, Goh P (2022) Visual stimuli-based dynamic commands with intelligent control for reactive BCI applications. *IEEE Sens J* 22(2):1435–1448
7. Ahmad NS, Teo JH, Goh P (2022) Gaussian process for a single-channel EEG decoder with inconspicuous stimuli and eyeblinks. *Comput Mater Cont* 73(1):611–628
8. Loganathan A, Ahmad NS, Goh P (2019) Self-adaptive filtering approach for improved indoor localization of a mobile node with zigbee-based RSSI and odometry. *Sensors* 19:21
9. Loganathan A, Ahmad NS (2023) A systematic review on recent advances in autonomous mobile robot navigation. *Eng Sci Technol Int J* 40:101342
10. Arrouch I, Ahmad NS, Goh P, Mohamad-Saleh J (2022) Close proximity time-to-collision prediction for autonomous robot navigation: an exponential GPR approach. *Alexand Eng J* 61(12):11171–11183
11. Arrouch I, Mohamad-Saleh J, Goh P, Ahmad NS (2022) A comparative study of artificial neural network approach for autonomous robot's TTC prediction. *Int J Mech Eng Robot Res* 11(5):345–350
12. Goay CH, Goh P, Ahmad NS, Ain MF (2018) Eye-height/width prediction using artificial neural networks from S-parameters with vector fitting. *J Eng Sci Technol* 13:3

A Comparative Analysis on Electrical and Photovoltaic Performances of MIS Structures on High Resistivity Silicon with Tunneling Insulator



Nur Bashirouh binti Attaullah, Nur Zatil 'Ismah Hashim, Chong Kah Hui, Nor Muzlifah Mahyuddin, Alhan Farhanah Abd Rahim, Mohd Marzaini bin Mohd Rashid, and Mundzir Abdullah

Abstract MIS structures utilizing tunneling AlN on high resistivity silicon is superior in terms of fabrication simplicity and improved bias responses in addition to providing promising electrical and photovoltaic performances. Based on previous simulation work, tunneling behavior is absent from the AlN-based MIS photovoltaic properties, indicating inconsistencies with the experimental evidence. This work aims to highlight this inconsistency by providing a comparative analysis between AlN and other insulating materials such as SiO₂, Si₃N₄ and Al₂O₃ in terms of the dark current, photocurrent and K ratio parameters. Results show that the absence of tunneling is still prominent in AlN, whilst the other insulating materials illustrate excellent electrical and photovoltaic properties evident by the high K ratio values ranging between 10³ and 10⁵. This could be due to the misrepresentation of AlN in the simulation tool, which requires further parametric adjustments.

Keywords MIS structure · High resistivity silicon · Photodetector

N. B. Attaullah · N. Z. 'Ismah Hashim (✉) · N. M. Mahyuddin
School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Penang, Malaysia
e-mail: zatil.hashim@usm.my

C. K. Hui
Collaborative Microelectronics Design Excellence Centre, Universiti Sains Malaysia, Penang, Malaysia

A. F. A. Rahim
School of Electrical Engineering, Universiti Teknologi MARA (UiTM), Penang, Malaysia

M. M. M. Rashid
School of Physics, Universiti Sains Malaysia, Penang, Malaysia

M. Abdullah
Institute of Nano Optoelectronics Research and Technology (INOR), Universiti Sains Malaysia, Penang, Malaysia

1 Introduction

High resistivity silicon is becoming favorable to be integrated in the photovoltaic applications, especially in the metal–insulator–semiconductor (MIS) structure utilizing tunneling insulator, owing to the lower inversion voltage created by the substrate [1–3]. Incorporating thin aluminium nitride (AlN) as the insulator makes the structure superior due to the reduced fabrication complexity and promising performances [4, 5]. However, based on work conducted in [6], tunneling behavior is not observed, even though the experimental results are evident [7]. In this work, this inconsistency will be highlighted, in addition to the comparative analysis with other insulators.

2 Methodology

The structural specification utilized in this work is based from Attaullah et al. [6], with palladium (Pd) as the metal contacts and a $4200 \Omega \text{ cm}$ p -type high resistivity silicon as the semiconductor substrate. A 2 nm thick insulating material is deposited in between the metal and semiconductor structure. The insulating material is varied between AlN, SiO₂, Si₃N₄ and Al₂O₃ at that particular thickness, sufficient enough to induce tunneling mechanism in the structure as stated in [8]. These structures are built and simulated using TCAD simulation tools, and their respective dark current and photocurrent (light current) characteristics are compared and analyzed.

3 Results and Discussion

Figures 1, 2, 3 and 4 illustrate the dark current and photocurrent characteristics for structures with AlN, SiO₂, Si₃N₄ and Al₂O₃, respectively. In all accumulation cases (voltage range between -0.6 and -1.6 V), the photocurrent is higher than the dark current and showing rectifying behavior. These observations are as expected; (1) generation of electrons and holes due to photons absorptions and (2) current contribution by majority carriers in the p -type substrate. Interesting observations can be seen in the inversion region (voltage range between 0.2 and 0.8 V), where tunneling behavior is illustrated in all structures, except for AlN. Regardless, the photocurrent is still much higher than the dark current, but only by a small magnitude.

Figure 5 shows the photocurrent comparison between MIS structures on different insulating material. MIS structure with Si₃N₄ illustrates the highest photocurrent values throughout the whole simulation region, followed by a closed competition between Al₂O₃ and SiO₂, and AlN. Similar observations are seen in Fig. 6 for the dark current comparison.

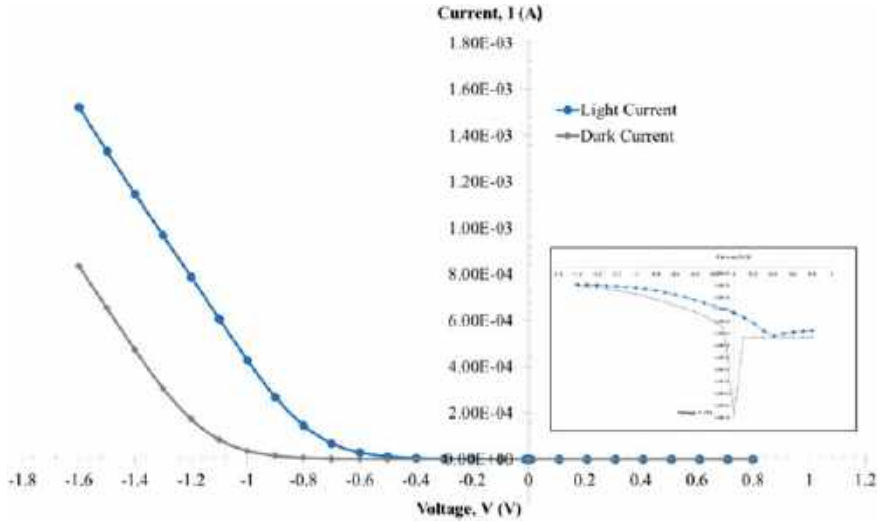


Fig. 1 Light current and dark current characteristics for MIS structure with AlN tunneling insulator. Inset plot represents the graph in log scale

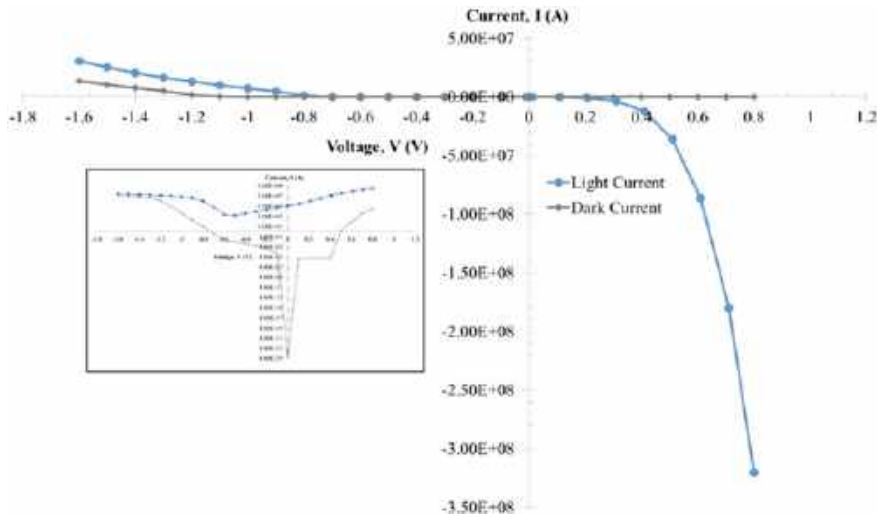


Fig. 2 Light current and dark current characteristics for MIS structure with SiO₂ tunneling insulator. Inset plot represents the graph in log scale

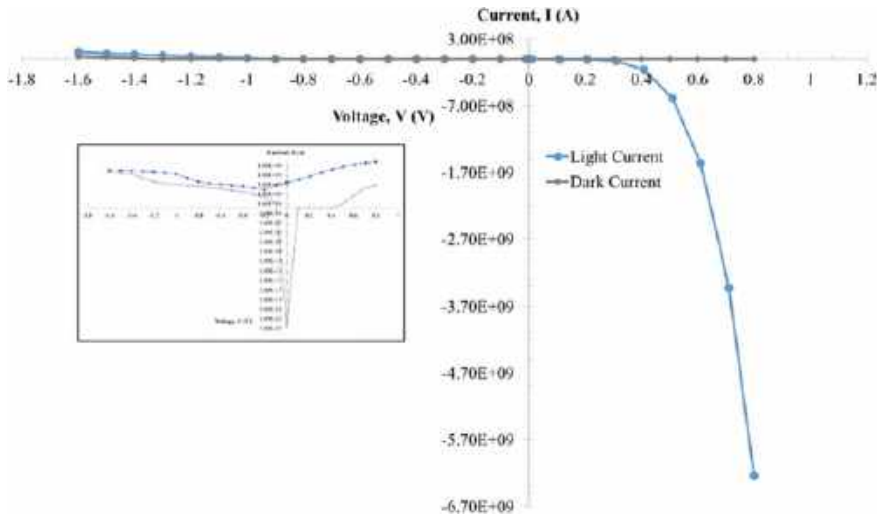


Fig. 3 Light current and dark current characteristics for MIS structure with Al₂O₃ tunneling insulator. Inset plot represents the graph in log scale

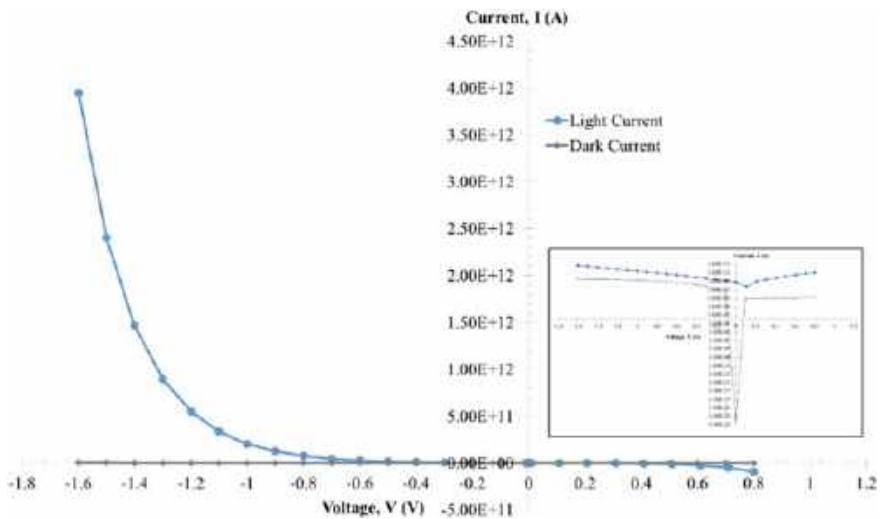


Fig. 4 Light current and dark current characteristics for MIS structure with Si₃N₄ tunneling insulator. Inset plot represents the graph in log scale

Table 1 illustrates the calculated K ratios (photocurrent/dark current) at inversion point i.e., our point of interest where tunneling is observed (or expected to observe in the case of AlN). Based on the results, structures with Si₃N₄ shows the highest

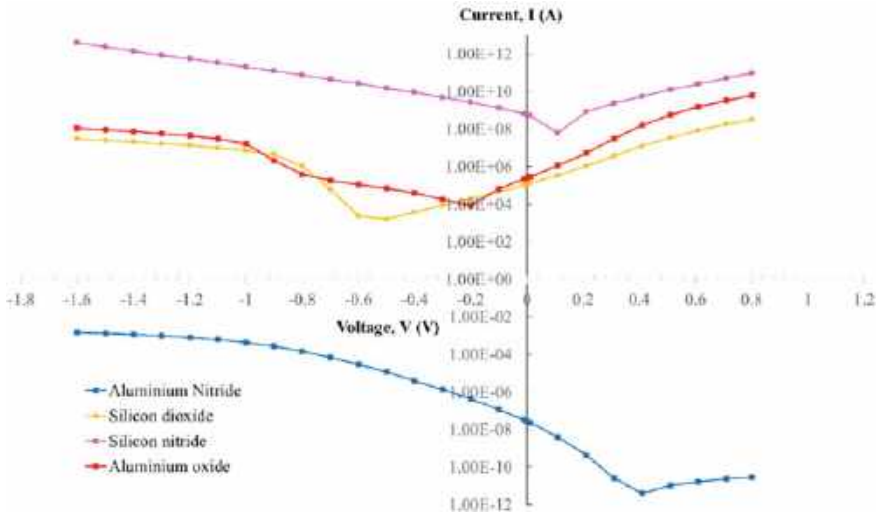


Fig. 5 Photocurrent comparison for MIS structures with varying insulating material

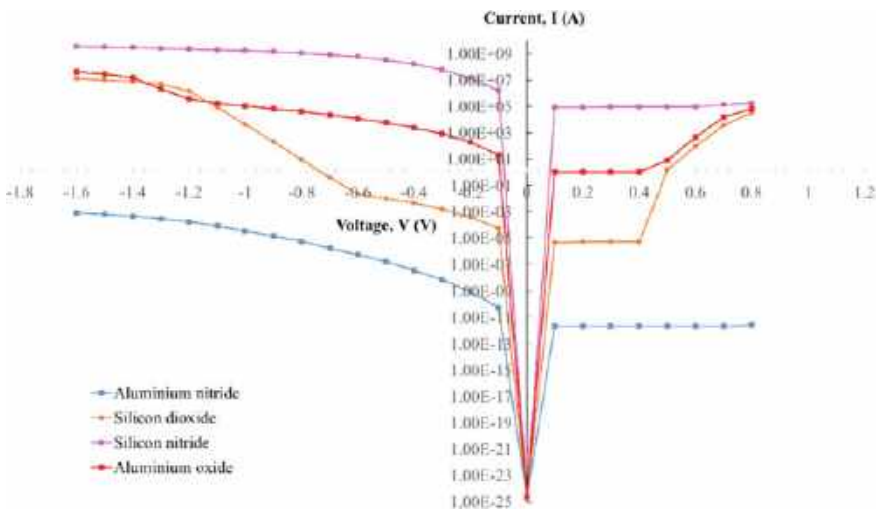


Fig. 6 Dark current comparison for MIS structures with varying insulating material

K ratio values, followed by Al_2O_3 , SiO_2 and AlN . These values corresponds to the results obtained in Figs. 5 and 6.

Despite the experimental observation in [7], the absence of tunneling behavior for structures with AlN simulated in this work is consistent with the simulation results obtained in [6], confirming the assumption made previously that the tunneling model used in the simulation work is not compatible with AlN . The main reason for this is

Table 1 K ratio for MIS structure with varying insulating material

Insulating material	K ratio
Aluminium nitride (AlN)	11.4
Silicon dioxide (SiO ₂)	9.50×10^3
Aluminium oxide (Al ₂ O ₃)	9.36×10^4
Silicon nitride (Si ₃ N ₄)	5.42×10^5

that, AlN is not, by default, defined as an insulator in the simulator as compared to the other three well-known insulating materials. Hence, parameter adjustments need to be made to ensure more accurate representation of AlN as tunneling insulator can be achieved. Regardless, the outcome of this simulation work provides better understanding on the tunneling mechanism provided by the thin insulating layer in the MIS structures on high resistivity silicon.

4 Conclusion

MIS structures utilizing tunneling insulator on high resistivity silicon are simulated and studied. The electrical and photovoltaic behavior of the structures with different insulating materials are compared and analyzed, illustrating predicted tunneling performance for structures with thin layers of Si₃N₄, Al₂O₃ and SiO₂, indicated by their respective high K ratio values. Similar tunneling properties are not portrayed in MIS structure with AlN, due to the misrepresentation of the AlN in the simulation. This requires further parameter adjustments for more accurate results.

Acknowledgements This work is supported by Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project Code: FRGS/1/2020/TK0/USM/02/6.

References

1. Dragoman M et al (2021) Multifunctionalities of 2D MoS₂ self-switching diode as memristor and photodetector. *Physica E* 126:114451
2. Dong Y et al (2019) System integration and packaging of a terahertz photodetector at W-band. *IEEE Trans Compon Packag Manuf Technol* 9(8):1486–1494
3. Liu J et al (2020) Photodiode with low dark current built in silicon-on-insulator using electrostatic doping. *Solid-State Electron* 168:107733
4. Ng DKT et al (2020) Considerations for an 8-inch wafer-level CMOS compatible AlN pyroelectric 5–14 μm wavelength IR detector towards miniature integrated photonics gas sensors. *J Microelectromech Syst* 29(5):1199–1207
5. Stewart JW et al (2020) Ultrafast pyroelectric photodetection with on-chip spectral filters. *Nat Mater* 19(2):158–162

6. Attaullah NB et al (2022) Current-voltage simulation analysis of MIS structure utilizing aluminium nitride on high resistivity silicon. In: 5th Photonics meeting 2022 (PM22), Journal of physics: conference series, vol 2411. IOP Publishing Ltd., p 012006
7. Bazlov NV et al (2020) Photoelectric properties of MIS structures on high-resistivity *p*-type silicon with aluminium nitride tunnelling insulator. In: International conference PhysicA.SPb/2020, journal of physics: conference series, vol 1697. IOP Publishing Ltd., p 012181
8. Lin C-S et al (2017) Ultraviolet emission from resonant tunnelling metal-insulator-semiconductor light emitting tunnel diodes. IEEE Photonics 9(4)

Robotics, Control, Mechatronics and Automation

An Analysis of the Performance of the SPRT Chart with Estimated Parameters Under the Weibull Distribution



Jing Wei Teoh, Wei Lin Teoh, Laila El-Ghandour, Zhi Lin Chong, and Sin Yin Teh

Abstract Most developments of the sequential probability ratio test (SPRT) control chart assume that the underlying process comes from a Normal distribution with known mean and standard deviation. Nevertheless, the true values of the process parameters are usually inaccessible in production settings, and they must be approximated from a set of Phase-I data. In certain areas, the process data can be positively skewed, which in turn affect the performance of control charts designed under the Normal distribution. In this paper, we provide a thorough analysis on the performances of the SPRT chart with estimated process parameters under the influence of Weibull distributed data. The unconditional properties of the expected value and standard deviation of the time to signal are evaluated using Monte Carlo simulation to facilitate comparisons between the Normal and Weibull distributions. Results show that both the in-control and out-of-control performances of the SPRT chart deteriorate when Weibull data are used. However, the optimal design of the SPRT chart with estimated process parameters seems to reverse the effect for large process mean shifts.

Keywords Average time to signal · Process parameter estimation · Sequential probability ratio test · Statistical process control · Weibull distribution

J. W. Teoh (✉) · W. L. Teoh
Heriot-Watt University Malaysia, Putrajaya, Malaysia
e-mail: t.jing_wei@hw.ac.uk

L. El-Ghandour
Heriot-Watt University, Edinburgh, UK

Z. L. Chong
Universiti Tunku Abdul Rahman, Kampar, Perak, Malaysia

S. Y. Teh
Universiti Sains Malaysia, Penang, Malaysia

1 Introduction

Statistical process control is essential for ensuring that the mean and variability of a manufacturing process are maintained at satisfactory levels. Popular quality control tools employed in industry include control charts. In particular, the sequential probability ratio test (SPRT) chart is highly valued due to its distinguished detection performance and sensitivity [1, 2].

In current literature, most developments related to the SPRT chart have been established assuming that the in-control process parameters are easily accessible. However, in practice, the mean and standard deviation of a production process can only be estimated from a modest number of Phase-I samples. Due to background noise, these estimates usually vary from practitioner to practitioner. Teoh et al. [3] showed that, on average, a practitioner has a 50% chance of obtaining a false alarm rate higher than the recommended rate when estimated process parameters are used. In view of the undesirable chart's performances led by parameter estimation, researchers proposed the guaranteed in-control performance (GICP) framework to combat the issue of high false alarms [4, 5]. Particularly, the GICP framework adjusts the limits of the control chart so that only a small proportion of practitioners (e.g., 5%) will obtain false alarm rates higher than the nominal rate. Many researchers have since implemented the proposed methodology to a myriad of control charts [3, 6, 7].

Many studies on control charts assume that production data come from the Normal distribution. However, this may not apply to all manufacturing processes. In fact, many manufacturing data follow a Weibull distribution, for example, the tensile strength of glass fibers and the fatigue life of a specific type of aluminum coupon [8]. It is expected that skewed distributions can have negative implications on the performances of control charts designed under the Normal distribution [9].

In this paper, we study the performances of the SPRT chart with estimated process parameters designed under the Normal distribution using the GICP method, when the data come from a Weibull distribution. We describe the SPRT chart with known and estimated process parameters in the next section. In Sect. 3, we present some statistical properties of the Weibull distribution. In Sect. 4, we tabulate the results and highlight the key interpretations. Finally, concluding remarks are given in Sect. 5.

2 The SPRT Chart Under the Normal Distribution

Suppose that the quality characteristic X of an industrial process follows the Normal distribution $N(\mu_0, \sigma_0^2)$, where μ_0 and σ_0^2 represent the in-control mean and variance, respectively. Suppose further that we want to test the hypotheses $H_0: \mu = \mu_0$ against $H_1: \mu = \mu_0 + \delta\sigma_0$, where $\delta > 0$ is the size of an upper-sided mean shift. The upper-sided SPRT chart with known process parameters has the following control statistic:

$$T_{i,j} = \sum_{\ell=1}^j \left(\frac{X_{i,\ell} - \mu_0}{\sigma_0} - \gamma \right), \tag{1}$$

for $i = 1, 2, \dots$, and $j = 1, \dots, N_i$, where $X_{i,\ell}$ is the ℓ th observation of the i th Phase-II sample, $\gamma (> 0)$ is a reference value, and N_i is the sample number of the i th SPRT. The upper-sided SPRT chart has a lower (g) and an upper (h) control limits. If $T_{i,j} < g$, the process is indicated as in-control. If $T_{i,j} > h$, then the process is signaled as out-of-control. If $g \leq T_{i,j} \leq h$, no conclusion can be drawn about the status of the process.

Since the process parameters μ_0 and σ_0 are unknown in practice, we estimate their values using m Phase-I observations (Y_1, \dots, Y_m). We use the Phase-I estimates $\hat{\mu}_0 = \sum_{\theta=1}^m Y_\theta / m$ and $\hat{\sigma}_0 = \sqrt{\sum_{\theta=1}^m (Y_\theta - \hat{\mu}_0)^2 / (m - 1)}$ in place of μ_0 and σ_0 , respectively, in Eq. (1), for the SPRT chart with estimated process parameters.

To evaluate the chart’s performances, Teoh et al. [3] derived the average time to signal (ATS), standard deviation of the time to signal (SDTS), and average sample number (ASN) as conditional functions of the random variables $V = \hat{\sigma}_0 / \sigma_0$ and $W = (\hat{\mu}_0 - \mu_0) / (\sigma_0 / \sqrt{m})$, when the process parameters are estimated. All formulae can be found in Teoh et al. [3] and are omitted in this paper due to space restrictions. To further evaluate the average performances across all parameter estimates, Teoh et al. [3] derived the average of the ATS (AATS), average of the SDTS (ASDTS), and average of the ASN (AASN) as

$$AATS = \int_0^\infty \int_{-\infty}^\infty ATS f_W(w) f_V(v) dw dv, \tag{2}$$

$$ASDTS = d \sqrt{\int_0^\infty \int_{-\infty}^\infty \frac{1 + OC(\delta) - [OC(\delta) - OC^2(\delta)] \mathbb{I}\{\delta \neq 0\}}{[1 - OC(\delta)]^2} f_W(w) f_V(v) dw dv - \frac{2}{3} \mathbb{I}\{\delta \neq 0\} - \left(\frac{AATS}{d} \right)^2}, \tag{3}$$

and

$$AASN = \int_0^\infty \int_{-\infty}^\infty ASN f_W(w) f_V(v) dw dv, \tag{4}$$

respectively. Here, $f_V(\cdot)$ and $f_W(\cdot)$ are the probability density functions of V and W , respectively. The derivations of these formulae can be found in Teoh et al. [3].

3 Statistical Properties of the Weibull Distribution

The cumulative distribution function of the two-parameter Weibull distribution is $F_U(u) = 1 - \exp\{-(\lambda u)^\alpha\}$ for $u \geq 0$, where $\alpha > 0$ and $\lambda > 0$ are the shape and scale parameters, respectively [10]. For convenience, we set $\lambda = 1$ throughout this paper. The skewness (k), in-control mean ($\mu_{U,0}$), and in-control variance ($\sigma_{U,0}^2$) of

the Weibull distribution can be calculated using the formulae [10]

$$k = \frac{\Gamma(1 + \frac{3}{\alpha}) + 2[\Gamma(1 + \frac{1}{\alpha})]^3 - 3\Gamma(1 + \frac{1}{\alpha})\Gamma(1 + \frac{2}{\alpha})}{\{\Gamma(1 + \frac{2}{\alpha}) - [\Gamma(1 + \frac{1}{\alpha})]^2\}^{1.5}}, \tag{5}$$

$$\mu_{U,0} = \Gamma\left(1 + \frac{1}{\alpha}\right), \tag{6}$$

and

$$\sigma_{U,0}^2 = \Gamma\left(1 + \frac{2}{\alpha}\right) - \left[\Gamma\left(1 + \frac{1}{\alpha}\right)\right]^2, \tag{7}$$

respectively, where $\Gamma(\cdot)$ represents the gamma function.

4 Results and Discussions

In this paper, we design the SPRT chart with estimated process parameters based on the GICP framework under the Normal distribution. The GICP constraint is as follows

$$\Pr(ATS_0 \geq (1 - \varepsilon)\tau) = 1 - p, \tag{8}$$

where ATS_0 is the in-control ATS, τ is the recommended ATS_0 , ε is the tolerance term, and p is an error probability. The GICP method is gaining popularity as it guarantees that the recommended ATS_0 is exceeded with a very high probability (e.g., 90% or 95%). The optimal design is established by minimizing the expected value of the average extra quadratic loss (AAEQL) given by [3]

$$AAEQL = \frac{1}{\delta_{\max} - \delta_{\min}} \int_0^{\infty} \int_{-\infty}^{\infty} \int_{\delta_{\min}}^{\delta_{\max}} \delta^2 ATS_1 f_W(w) f_V(v) d\delta dw dv, \tag{9}$$

where ATS_1 is the out-of-control ATS (i.e., when $\delta > 0$).

Prior to the chart’s design, we need to fix some specifications, i.e., the inspection rate $R = AASN_0/d$, minimum sampling interval d_{\min} , minimum mean shift δ_{\min} , maximum mean shift δ_{\max} , τ , m , p , and ε . Here, $AASN_0$ is the in-control AASN. The inspection rate R should be set as a suitable number that balances the ratio between the in-control average sample size and the sampling interval. d_{\min} is specified to prevent an unreasonably small d being attained as a result of the optimal design. The other design variables, i.e., δ_{\min} , δ_{\max} , τ , m , p , and ε , can be set based on practitioner’s past experience and the current in-control policy practiced by the company. In this paper, we set $(R, d_{\min}, \delta_{\min}, \delta_{\max}, \tau, p, \varepsilon) = (5, 0.25, 0.10, 2.00, 250, 0.05, 0.20)$.

Teoh et al. [3] outlined a comprehensive algorithm for developing an optimal SPRT chart with GICP-adjusted limits, given a set of design specifications. In particular, the algorithm searches the values of the charting parameters $(AASN_0, \gamma, d, g, h)$ that minimizes the AAEQL (i.e., Eq. (9)), while ensuring that the GICP constraint (i.e., Eq. (8)) is satisfied. The complete algorithm can be found in Teoh et al. [3]. Table 1 shows the optimal charting parameters, together with their corresponding AATS and ASDTS values computed using Eqs. (2) and (3), respectively, for $m \in \{250, 500, 1000\}$ and $\delta \in \{0.25, 0.50, 1.00, 1.50, 2.00\}$ under the Normal distribution. Note that when $m = +\infty$, the usual formulae for computing the ATS and SDTS values of the SPRT chart with known process parameters are used.

When the process data come from a Weibull distribution, it is expected that the optimal SPRT chart designed under the Normal distribution would perform differently, depending on the degree of skewness. In our experiment, we consider four levels of skewness, i.e., $k \in \{0.0, 1.0, 2.0, 3.0\}$, for the Weibull distribution. Using numerical root-finding methods, we solve for the values of α in Eq. (5). Tables 2 and 3 display the in-control and out-of-control (AATS, ASDTS) values, i.e., $(AATS_0, ASDTS_0)$ and $(AATS_1, ASDTS_1)$, respectively, of the SPRT chart with estimated process parameters. The process means used in Tables 2 and 3 are computed as $\mu_{U,0} + \delta\sigma_{U,0}$, where $\delta = 0$ (in-control) and $\delta > 0$ (out-of-control) are used for Tables 2 and 3, respectively. Note that $\mu_{U,0}$ and $\delta\sigma_{U,0}$ of the underlying Weibull process are computed from Eqs. (6) and (7), respectively. All the tabulated results are approximated using Monte Carlo simulations with 100,000 replications, as the sampling distributions of the Weibull random variable are not analytically tractable.

From Table 2, it can be observed that, when $k = 0$, the $(AATS_0, ASDTS_0)$ values under the Weibull distribution are reasonably close to those of the Normal distribution in Table 1. For example, when $m = 1000$, the $(AATS_0, ASDTS_0)$ values under

Table 1 Optimal charting parameters $(AASN_0, \gamma, d, g, h)$, AATS, and ASDTS values of the SPRT chart with estimated process parameters, for $m \in \{250, 500, 1000, +\infty\}$ and $\delta \in \{0.00, 0.25, 0.50, 1.00, 1.50, 2.00\}$, when the data are Normally distributed

m	250	500	1000	$+\infty$
	$(AASN_0, \gamma, d)$	$(AASN_0, \gamma, d)$	$(AASN_0, \gamma, d)$	(ASN_0, γ, d)
	(g, h)	(g, h)	(g, h)	(g, h)
δ	$(AATS, ASDTS)$	$(AATS, ASDTS)$	$(AATS, ASDTS)$	$(ATS, SDTS)$
	(2.074, 0.295, 0.415)	(2.241, 0.287, 0.448)	(2.208, 0.296, 0.442)	(2.135, 0.322, 0.427)
	(0.446, 11.228)	(0.341, 10.131)	(0.317, 9.098)	(0.259, 7.473)
0.00	(5711.79, 29,799.56)	(1251.75, 2487.63)	(597.21, 804.57)	(250.00, 249.79)
0.25	(38.79, 113.49)	(19.20, 27.79)	(15.48, 17.94)	(12.64, 12.64)
0.50	(2.52, 2.96)	(2.16, 2.28)	(2.11, 2.15)	(2.07, 2.06)
1.00	(0.61, 0.59)	(0.59, 0.56)	(0.57, 0.55)	(0.55, 0.52)
1.50	(0.35, 0.30)	(0.35, 0.30)	(0.34, 0.29)	(0.32, 0.27)
2.00	(0.26, 0.20)	(0.27, 0.20)	(0.26, 0.19)	(0.25, 0.18)

Table 2 AATS₀ and ASDTS₀ values of the SPRT chart with estimated process parameters, for $m \in \{250, 500, 1000, +\infty\}$ and $k \in \{0.0, 1.0, 2.0, 3.0\}$, when the data are Weibull distributed

		$m = 250$	$m = 500$	$m = 1000$	$m = +\infty$
α	k	(AATS ₀ , ASDTS ₀)	(AATS ₀ , ASDTS ₀)	(AATS ₀ , ASDTS ₀)	(ATS ₀ , SDTS ₀)
3.60235	0.0	(5970.64, > 20,000)	(1277.68, 2596.22)	(607.78, 822.46)	(251.70, 251.63)
1.56391	1.0	(1865.50, 8539.06)	(556.08, 1008.34)	(290.24, 371.20)	(131.07, 130.80)
1.00000	2.0	(908.99, 3922.21)	(330.05, 580.07)	(186.08, 232.86)	(90.65, 90.24)
0.76862	3.0	(579.71, 3131.25)	(241.42, 436.56)	(145.38, 182.61)	(76.03, 75.80)

the Weibull and Normal distributions are given as (607.78, 822.46) and (597.21, 804.57), respectively. When k increases, the (AATS₀, ASDTS₀) values drop for all levels of m . While a decrease in the ASDTS₀ value might suggest more consistent performances, a fall in the AATS₀ value is certainly a strong indication of deteriorated in-control performance. It means that a large proportion of practitioners is likely to suffer unacceptable levels of false alarms as a result of the positively skewed data. Referring to the out-of-control cases, from Table 3, it is noticed that, for all levels of m , the (AATS₁, ASDTS₁) values under the Weibull distribution with $k = 0$ are very close to those under Normal conditions. It is also observed that, when $0.25 \leq \delta \leq 1.00$, the (AATS₁, ASDTS₁) performances worsen as k increases. On the other hand, when $1.50 \leq \delta \leq 2.00$, the (AATS₁, ASDTS₁) values are found to decrease slightly. Though there is a slight improvement in the chart’s performance for large mean shifts, the deterioration in performance for small mean shifts due to positive skewness is much more concerning. For example, when $m = 250$, an increase in the skewness from 0 to 3 leads to a sharp increase in the AATS₁ from 38.16 to 85.77 for $\delta = 0.25$, and only a small decrease in the AATS₁ from 0.26 to 0.21 is observed for $\delta = 2.00$ (see Table 3).

5 Conclusions

This paper studies the effect of the Weibull distribution on the performances of the SPRT chart with estimated process parameters designed under the Normal distribution. It is shown that, in the in-control case, the AATS₀ values decrease as the skewness of the data increases. This leads to undesirably large false alarms being obtained across practitioners. In the out-of-control cases, as the skewness increases, both the AATS₁ and ASDTS₁ values increase for small and moderate mean shifts, and decrease for large mean shifts. The unexpected improvement in the chart’s performance for large mean shifts may be attributed to the AAELQ optimal design proposed

Table 3 AATS₁ and ASDTS₁ values of the SPRT chart with estimated process parameters, for $k \in \{0.0, 1.0, 2.0, 3.0\}$, $\delta \in \{0.25, 0.50, 1.00, 1.50, 2.00\}$ and $m \in \{250, 500, 1000, +\infty\}$, when the data are Weibull distributed

			$m = 250$	$m = 500$	$m = 1000$	$m = +\infty$
α	k	δ	(AATS ₁ , ASDTS ₁)	(AATS ₁ , ASDTS ₁)	(AATS ₁ , ASDTS ₁)	(ATS ₁ , SDTS ₁)
3.60235	0.0	0.25	(38.16, 112.80)	(18.87, 27.07)	(15.29, 17.68)	(12.41, 12.36)
		0.50	(2.50, 2.92)	(2.15, 2.26)	(2.09, 2.14)	(2.06, 2.05)
		1.00	(0.61, 0.60)	(0.59, 0.57)	(0.58, 0.55)	(0.55, 0.52)
		1.50	(0.35, 0.31)	(0.35, 0.30)	(0.34, 0.29)	(0.33, 0.28)
		2.00	(0.26, 0.20)	(0.27, 0.20)	(0.27, 0.19)	(0.25, 0.18)
1.56391	1.0	0.25	(47.65, 157.77)	(22.64, 35.32)	(17.64, 21.00)	(13.95, 13.94)
		0.50	(3.08, 4.39)	(2.55, 2.78)	(2.47, 2.56)	(2.44, 2.43)
		1.00	(0.68, 0.68)	(0.64, 0.62)	(0.63, 0.60)	(0.59, 0.57)
		1.50	(0.35, 0.31)	(0.34, 0.29)	(0.33, 0.28)	(0.31, 0.26)
		2.00	(0.23, 0.16)	(0.24, 0.15)	(0.23, 0.14)	(0.22, 0.13)
1.00000	2.0	0.25	(61.33, 272.58)	(27.56, 47.82)	(20.86, 25.90)	(15.99, 15.98)
		0.50	(3.92, 9.26)	(3.00, 3.47)	(2.87, 3.05)	(2.83, 2.82)
		1.00	(0.73, 0.74)	(0.67, 0.66)	(0.65, 0.63)	(0.60, 0.58)
		1.50	(0.32, 0.29)	(0.29, 0.23)	(0.27, 0.21)	(0.25, 0.18)
		2.00	(0.21, 0.12)	(0.22, 0.13)	(0.22, 0.13)	(0.21, 0.12)
0.76862	3.0	0.25	(85.77, 1114.55)	(33.58, 70.51)	(24.39, 31.95)	(18.33, 18.34)
		0.50	(5.69, 67.28)	(3.45, 4.72)	(3.22, 3.55)	(3.16, 3.15)
		1.00	(0.76, 0.80)	(0.67, 0.67)	(0.64, 0.62)	(0.58, 0.55)
		1.50	(0.27, 0.24)	(0.23, 0.15)	(0.22, 0.13)	(0.21, 0.12)
		2.00	(0.21, 0.12)	(0.22, 0.13)	(0.22, 0.13)	(0.21, 0.12)

by Teoh et al. [3]. In future, researchers may design the AAEQL-optimal SPRT chart with estimated process parameters under the Weibull distribution.






Acknowledgements This work was supported by the Ministry of Higher Education (MOHE) Malaysia and Heriot-Watt University Malaysia under Fundamental Research Grant Scheme (FRGS), no. FRGS/1/2021/STG06/HWUM/02/1.

References

1. Mahadik SB, Godase DG, Teoh WL (2021) A two-sided SPRT control chart for process dispersion. *J Stat Comput Simul* 91(17):3603–3614
2. Pramanik S, Johnson VE, Bhattacharya A (2021) A modified sequential probability ratio test. *J Math Psychol* 101(4):102505
3. Teoh JW, Teoh WL, Khoo MBC, Castagliola P, Moy WH (2022) On designing an optimal SPRT control chart with estimated process parameters under guaranteed in-control performance. *Comput Ind Eng* 174(12):108806
4. Gandy A, Kvaløy JT (2013) Guaranteed conditional performance of control charts via bootstrap methods. *Scandinavian J Stat* 40(4):647–668
5. Capizzi G, Masarotto G (2020) Guaranteed in-control control chart performance with cautious parameter learning. *J Qual Technol* 52(4):385–403
6. Diko MD, Chakraborti S, Does RJMM (2019) Guaranteed in-control performance of the EWMA chart for monitoring the mean. *Qual Reliab Eng Int* 35(4):1144–1160
7. Li J (2022) Adaptive CUSUM chart with cautious parameter learning. *Qual Reliab Eng Int* 38(6):3135–3156
8. Gurvich MR, Dibenedetto AT, Ranade SV (1997) A new statistical distribution for characterizing the random strength of brittle materials. *J Mater Sci* 32(5):2559–2564
9. Teoh WL, Yeong WC, Khoo MBC, Teh SY (2016) The performance of the double sampling \bar{X} chart with estimated parameters for skewed distributions. *Acad J Sci* 5(1):237–252
10. Rinne H (2008) *The Weibull distribution: a handbook*. CRC Press, Boca Raton

Semi-Automatic Liquid Filling System Using NodeMCU as an Integrated IoT Learning Tool



N. F. Adan , Z. Zainal , M. N. A. H. Sha'abani , M. F. Mohamed Nor , and M. F. Ismail 

Abstract Computer programming and IoT are the key skills required in Industrial Revolution 4.0 (IR4.0). The industry demand is very high and therefore related students in this field should grasp adequate knowledge and skill in college or university prior to employment. However, learning technology related subject without applying it to an actual hardware can pose difficulty to relate the theoretical knowledge to problems in real application. It is proven that learning through hands-on activities is more effective and promotes deeper understanding of the subject matter (He et al. in *Integrating Internet of Things (IoT) into STEM undergraduate education: Case study of a modern technology infused courseware for embedded system course*. Erie, PA, USA, pp 1–9 (2016)). Thus, to fulfill the learning requirement, an integrated learning tool that combines learning of computer programming and IoT control for an industrial liquid filling system model is developed and tested. The integrated learning tool uses NodeMCU, Blynk app and smartphone to enable the IoT application. The system set-up is pre-designed for semi-automation liquid filling process to enhance hands-on learning experience but can be easily programmed for full automation. Overall, it is a user and cost friendly learning tool that can be developed by academic staff to aid learning of IoT and computer programming in related education levels and fields.

Keywords IoT · NodeMCU · Blynk · Learning tool · TVET

1 Introduction

Internet of Things (IoT) is an essential pillar in Industrial Revolution 4.0 (IR4.0) as it gathers information from the machines via thousands of sensors to the cloud database in real time. Thus, essentially IoT has become a staple skill required by

N. F. Adan (✉) · Z. Zainal · M. N. A. H. Sha'abani · M. F. Mohamed Nor · M. F. Ismail
Universiti Tun Hussein Onn Malaysia, Pagoh Higher Education Hub, 84600 Pagoh, Johor,
Malaysia
e-mail: norfaezah@uthm.edu.my

the industry. To meet the demand for professionals in this field [1], education sector inevitably must play a role into prepping the future workforce into shape prior to employment.

Learning through hands-on activities allows students to experience and promotes deeper understanding of the subject matter [1]. It also better prepares graduates for the technology and knowledge-based jobs [2]. Therefore, through the implementation of hands-on activities using appropriate learning tool, students will be able to grasp the knowledge better and acquire the needed IoT skill [3, 4].

Apart from the IoT skill, students should be exposed to real industrial application. A simple and common application is automatic liquid filling system which is applied in food and beverages industry, chemical industry, and various others. Therefore, an integrated IoT learning tool is developed to combine the learning experience for students. The integrated learning tool is aimed to allow the students to apply their fundamental knowledge of computer programming and IoT to control a liquid filling system model.

1.1 Literature Review

Automatic liquid filling system in the industry typically uses Programmable Logic Controller (PLC) as they are more robust and can withstand harsh industrial environment. The research in [5] studied the use of PLC to automate the liquid filling process in a bottle. The developed laboratory prototype mainly consists of solenoid valve, conveyor belt and stepper motor, photo electric sensor and flow sensor. Photoelectric sensor used to detect the bottle and stop the conveyor while flow sensor is used to determine the rate of water flow and thus volume is controlled using time delay. The installation cost is expensive thus the project only implemented two sensors to reduce the cost. Hence, to reduce the cost which is mainly due to expensive controllers such as PLC, the researchers in [6, 7] implemented Arduino as the controller for their system prototype. All these prototypes are fully automated without IoT implementation.

Therefore, IoT implementation from other application can be integrated into the liquid filling system to meet the objective of this research project. The IoT controlled system in [8, 9] uses NodeMCU, smartphone and Blynk app. Smartphone is used to turn on and turn off devices such as lamp, fan, water pump etc. via internet connection. NodeMCU is similar to Arduino but with an additional Wi-Fi capability which allows easier IoT implementation. Blynk app is an IoT platform which is used to remotely control connected electronic devices and display any data transmitted from NodeMCU.

Various methods were used previously to measure and control the level of water filled in the bottle such as using time delay [5, 10], camera and image processing [6] and ultrasonic sensor [7]. Meanwhile the sensor that can be used to detect presence of bottle on the conveyor belt is photoelectric sensor [5], laser sensor [6] and infrared sensor (IR) [10]. The conveyor belt is usually operated using DC motor [5, 10] or







stepper motor [6] while solenoid valve [5, 10] is used to flow the liquid from the tank into the bottle.

2 Material and Method

Factors such as cost, level of complexity and user-friendly features are considered in the material selection process. On top of that, the model developed is meant for education purpose, hence more hands-on involvement needed to fully comprehend computer programming skills and IoT fundamentals. The material used for the integrated learning tool is summarized in Table 1.

The block diagram in Fig. 1 showed the overview of the integrated learning tool. A smartphone is used as the remote controller to manually control the solenoid valve

Table 1 List of components

No	Component		Function
1	NodeMCU		Controller—operates motor and valve, communicate with Blynk
2	12V DC motor		Runs conveyor belt
3	L298N motor driver module		Allow low-current controller to control DC motor with higher voltage and current requirement
4	Relay module		Connect solenoid valve to external higher rated power supply
5	IR sensor		Detect presence of cup
6	12V DC solenoid valve		Open and close to flow water into cup

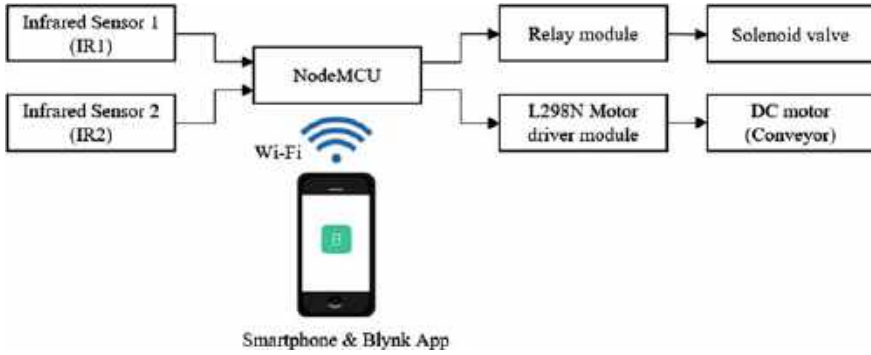


Fig. 1 Block diagram of the integrated learning tool

and starts the conveyor belt operation. IR1 is used to detect presence of cup at the filling location and IR2 is used to detect the presence of cup at the end of the conveyor belt.

Figure 2 shows the actual system together with the Blynk app interface. There are only two control buttons on the Blynk app interface which is pushbutton1 (start conveyor) and pushbutton2 (start and stop solenoid valve).

The operation of the semi-automatic liquid filling system for this integrated learning tool is represented by the flowchart shown in Fig. 3. Only one cup is allowed on the conveyor belt at a time. At the beginning, the conveyor belt is off, and the red LED is turned on to indicate conveyor belt is not in operation. When pushbutton1 in Blynk app is pressed the conveyor belt will run and green LED turned on to indicate conveyor belt is running. Next, the system will check status of IR1. If it is not triggered, the system will then check the status of IR2. If IR1 is triggered that means the cup has arrived at the filling station. But if IR2 is triggered, that means the cup has arrived at the end of the conveyor belt.

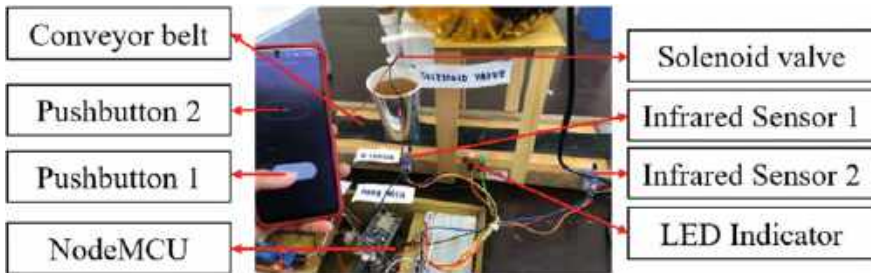


Fig. 2 Assembled integrated learning tool with Blynk App interface

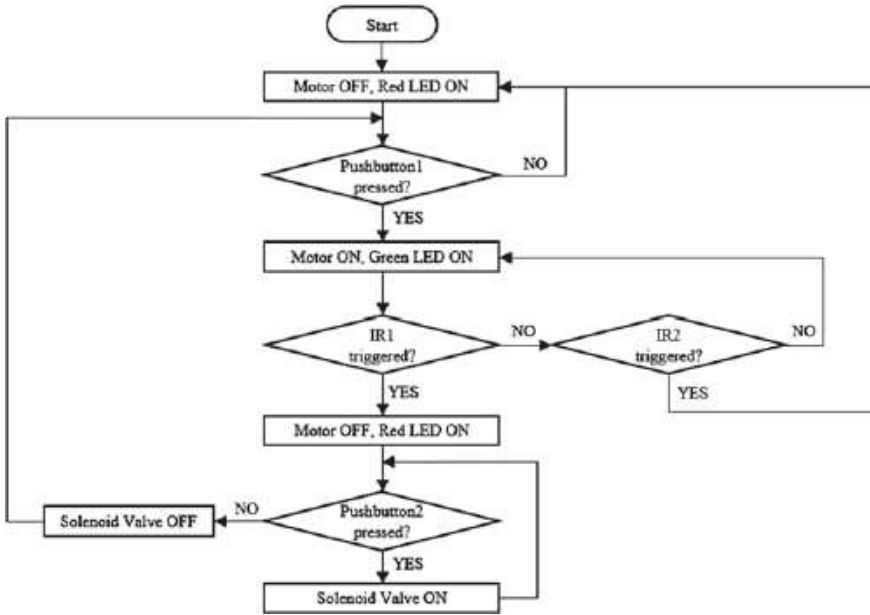
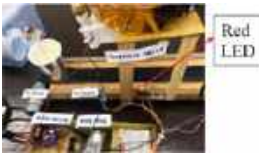
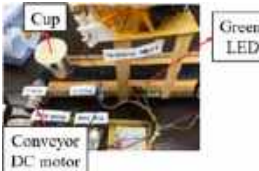
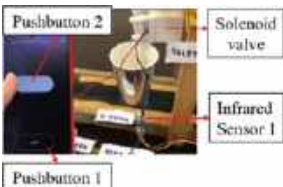

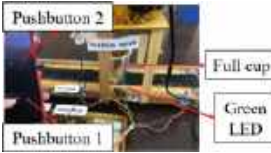



Fig. 3 System operation flowchart

3 Result

The full operation of the system has been successfully tested and the step-by-step operation is recorded as shown in Table 2. The overall system is simple to understand and allows students to engage with the operation of liquid filling process. Thus, students will be able to appreciate and understand the function of each component involved especially the IoT control via smartphone.

Table 2 Step by step operation of the semi-automatic liquid filling system

Step No	Diagram	Operation
1	 <p>Red LED</p>	<p>Power supply connected. A cup is placed on the conveyor belt. At the beginning, the conveyor belt is off and red LED is turned on</p>
2	 <p>Cup</p> <p>Conveyoe DC motor</p> <p>Green LED</p>	<p>When pushbutton 1 is pressed, the conveyor belt turned on and green LED also turned on. The cup is moving on the conveyor belt towards the filling station</p>
3	 <p>Pushbutton 2</p> <p>Solenoid valve</p> <p>Infrared Sensor 1</p> <p>Pushbutton 1</p>	<p>Cup is detected by IR1 causing the conveyor belt to turn off and red LED is turn on. Next, pushbutton 2 is pressed and solenoid valve opened to allow water flow into the cup</p>
4	 <p>Filling the cup</p> <p>Red LED</p>	<p>Student must be alert and monitor the filling process to prevent overspill as the filling operation is not automated</p>
4	 <p>Pushbutton 2</p> <p>Full cup</p> <p>Green LED</p> <p>Pushbutton 1</p>	<p>Pushbutton 2 is pressed again to close the solenoid valve and stop the filling process. The conveyor belt is turned on and green LED also turned on when pushbutton 1 is pressed. Now, the cup is moving again on the conveyor belt</p>
5	 <p>Red LED</p> <p>Infrared Sensor 2</p>	<p>Finally, the cup has reached the end of the conveyor belt and its presence is detected by IR2. The conveyor belt stopped and red LED turned on. Student must lift the cup from the belt as the system only allow one cup operation at a time</p>

4 Conclusion

The semi-automatic liquid filling system integrated with basic IoT application is a user and cost friendly learning tool that can be easily developed by academic staff to aid learning of IoT and computer programming in related education level and field. The integrated learning tool also introduces students to actual industrial system which helps to enhance their learning experience in class. The next stage of learning which is full automation and IoT monitoring can be accomplished using the same integrated learning tool. Students can modify the programs using time delay method to automate the liquid filling process into the cup and movement of the conveyor belt.


Acknowledgements This research was supported by Centre for Diploma Studies (CeDS) and Registrar Office, Universiti Tun Hussein Onn Malaysia (UTHM).

References

1. He J et al (2016) Integrating Internet of Things (IoT) into STEM undergraduate education: case study of a modern technology infused courseware for embedded system course. In: 2016 IEEE frontiers in education conference (FIE), Erie, PA, USA, pp 1–9. <https://doi.org/10.1109/FIE.2016.7757458>
2. Pusca D et al (2017) Hands-on experiences in engineering classes: the need, the implementation and the results. *World Trans Eng Technol Edu* 15(1):12–18
3. Akbar MA et al (2018) Technology based learning system in Internet of Things (IoT) education. In: 7th International conference on computer and communication engineering (ICCCE), pp 192–197. <https://doi.org/10.1109/ICCCE.2018.8539334>
4. Fidai A et al. (2019) Internet of Things (IoT) instructional devices in STEM classrooms: past, present and future directions. In: IEEE frontiers in education conference (FIE), Covington, KY, USA, pp 1–9. <https://doi.org/10.1109/FIE43999.2019.9028679>
5. Baladhandabany D et al (2015) PLC based automatic liquid filling system. *Int J Comput Sci Mob Comput* 4(3):684–692
6. Abdullah O et al (2020) Development of automated liquid filling system based on the interactive design approach. *FME Trans* 48(4):938–945
7. Mishra A et al (2020) IOT based automatic water container filler using Arduino. *Int J Innov Sci Res Technol* 5(6):370–437
8. Durani H et al (2018) Smart automated home application using IoT with Blynk App. In: 2nd International conference on inventive communication and computational technologies (ICICCT), Coimbatore, India, pp 393–397. <https://doi.org/10.1109/ICICCT.2018.8473224>
9. Joha MI et al (2021) IoT-based smart home automation using NodeMCU: a smart multi-plug with overload and over temperature protection. In: 24th international conference on computer and information technology (ICCIT), Dhaka, Bangladesh, pp 1–6. <https://doi.org/10.1109/ICCIT54785.2021.9689913>
10. Viraktamath SV et al (2020) Implementation of automated bottle filling system using PLC. In: Ranganathan G et al (eds) LNNS, vol 89. Springer, Heidelberg, pp 33–41. <https://doi.org/10.1007/978-981-15-0146-3>

Short Range Radio Frequency (RF) Data Acquisition Unit for Agricultural Product Monitoring System



S. M. N. S. Shatir , A. B. Elmi, M. N. Akhtar, M. N. Abdullah, and A. H. Ismail

Abstract This research emphasized on the experimentation performed on the 2.4 Ghz radio frequency (RF) module, utilized on Standstill Monitoring System (SMS), to discover the factors hindering the capabilities of the RF range and data stability in plantation environment, then proposed a signal enhancement technique. The architecture of SMS consists of sensor modules such as temperature, humidity, soil moisture and soil nutrition sensor to analyze as well as monitor real-time data on agriculture parameters through an IoT dashboard. The experiments were carried out, where the transmitter (Tx) was placed at the three ground stations with three different surrounding condition which is no obstruction, mild obstructions and full obstructions. The receiver (Rx) module was moved away from the ground station and transmission status was monitored. Findings showed that the transmission range without no obstruction is 70 m, with partial blockage is 50 m and total blockage is 10 m. By applying signal enhancement method which is implementing capacitor and breakout board, the range has been boosted to 350 m, 320 m and 160 m respectively. With this findings, wireless communication waypoint can be plotted out at a Harumanis plantation which allows farmers to supervise their plantation at distance.

Keywords Wireless communication · Radio module · Transmission · SMS

S. M. N. S. Shatir · A. B. Elmi (✉) · M. N. Akhtar · M. N. Abdullah
School of Aerospace Engineering, Universiti Sains Malaysia, 14300 Nibong Tebal, Penang,
Malaysia
e-mail: meelmi@usm.my

A. H. Ismail
Faculty of Electrical Engineering and Technology, Universiti Malaysia Perlis, 02600 Arau, Perlis,
Malaysia

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
N. S. Ahmad et al. (eds.), *Proceedings of the 12th International Conference on Robotics, Vision, Signal Processing and Power Applications*, Lecture Notes in Electrical Engineering 1123, https://doi.org/10.1007/978-981-99-9005-4_24

191

1 Introduction and Background

Malaysia is known with the variety of tropical fruits where one of the vast cultivated tropical fruits in Malaysia is mango (*Mangifera indica* L.), there was approximately more than 5000 hectare of mango cultivation and more than 17,000 metric tonne of mango production in Malaysia [1]. The prospect of Harumanis can be seen when the Department from Statistics Malaysia (DOSM) [2], reported that the production of Harumanis mango have rose by a whopping 57% in 2021 with 3336 metric ton with production value of RM66.7 million compared to 2121 metric ton in 2020 with production value of RM42.42 million.

Instrument testing for RF range and transmission stability is becoming increasingly important in agricultural environments due to the proliferation of wireless sensors and devices used for precision agriculture applications [3]. These devices require reliable wireless communication to collect and transmit data from remote locations. However, the agricultural environment presents several challenges that can impact RF performance, such as physical obstructions, interference from other wireless devices, and varying weather conditions. This research paper is a continuation work from the previously proposed SF system, known as Standstill Monitoring System (SMS) [4], meant to alleviate challenges faced by farmers such as difficulty in ensuring sufficient nutrients supplied to the crop along with distinguishing a diseased plant and fruit. Experimentation studies in this paper are centered on the performance of sensor modules implemented in the architecture of SMS, primarily the radio module nRF24L01+.

For the rest of the manuscript, Sect. 2 explains on the experimental setup and methodology on the testing, which is to gather insight on the radio module network coverage including location and area of experimentation, transmission band and devices model. Section 3 shows the experiment result in figures with data monitoring result seen through serial monitor. Conclusion of the proposed work is placed under Sect. 4. Meanwhile, acknowledgement is placed under end of this chapter.

2 Methodology

There are various Wi-Fi modules in the market, however the selection to use nRF24L01+ as a wireless module is due to the low-power consumption (11.3 mA) which is ideal to pair with battery powered device or power bank, supports a maximum data rate of 2 Mbps, has a range of up to 100 m (m) in direct line of sight (LOS) and able to act as both transmitter and receiver while utilizing 2.4 GHz frequency band as wireless transmission channel as well as cost efficient. This research utilizes the same radio module nRF24L01+, equipped with external antennae at both Tx and Rx board, transmitting sensor data from ground station to the receiver board.

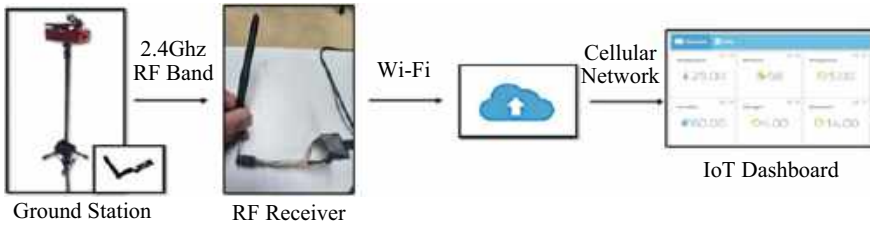


Fig. 1 Proposed data communication method for SMS

In Fig. 1, the wireless communication technology at each point, starting from ground station to end-user has been proposed. The Harumanis plantation will act as the ground station, as data are transferred to Rx board through 2.4 GHz band. The receiver board which is a waypoint, placed in an area with Wi-Fi connectivity in order to forward the received data to cloud server, where farmer is able to perform remote monitoring through IoT dashboard.

The transmission test was done to mimic the real-life application in Harumanis plantation, where several data readings are taken under different conditions such as direct Line of Sight (LOS), where the two antennas able to directly communicate without any obstacle, Near Line of Sight (nLOS) is when partial separation exists between radio signal, such as trees and bushes while Non-Line of Sight (NLOS) refers to complete obstruction between the radio modules such as buildings. For each the conditions, the Tx module is placed at the ground station while researcher moved the Rx board away from the ground station. The transmission status was observed on serial monitor at every 50 m. Transmission is considered weak when sensor readings on serial monitor were lagging, and transmission failed when new sensor readings were not registered. When failure of transmission was detected, the distance was observed and marked on google map application to obtain distance of transmission.

3 Result and Discussion

The purpose of this experiments is to find out different types of interference affecting the transmission capability that exist in outdoor condition especially plantation environment. Based on initial hypothesis, the interference can occur from the noise caused by other wireless network module such as Wi-Fi and Bluetooth, among other things such as the line of sight being obstructed by solid objects, for example walls, buildings and trees. Later, discussion on gathered results and hypothesis were made, identifying the cause of RF limitation and instability in data transmission, along with suggesting approach that can be taken to overcome the limitation of nRF24L01+.

3.1 Comparisons Before and After Signal Enhancement Technique

There are three different testing points in Universiti Sains Malaysia Engineering Campus that have been chosen for the test. Each of the point has different obstructions size. As such, point A surrounding consist of shrubs, less denser trees and no buildings nearby, as comparison to point B where the signal trajectory is not being interfered with trees or buildings as the tested location is surrounded by open sports courts. Finally, point C as testing site for NLOS, is at student’s dormitory where full radio wave obstruction should occur. The area of testing location along with travel direction (orange arrow) and ground station location (blue circle) is exhibit in Fig. 2. The initial test was done using RF24 board that has not been modified, as in Fig. 3.

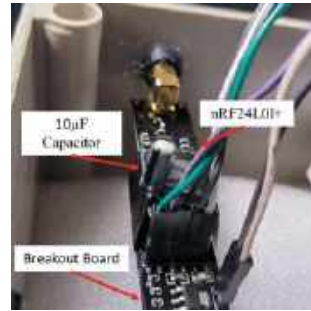


Fig. 2 Radio transmission coverage radius and testing points

Fig. 3 Nonmodified RF24 board



Fig. 4 Modified RF24 board



From the study [5] has integrated capacitor to RF24 modules as a current booster to the power supply whenever the main current supply is insufficient, due to increased current demands mainly during transmission process. The capacitance selected for this project is 10 μ F, as the suitable range of electrolytic capacitors are within 10 μ F to 100 μ f, any higher would stress the 3.3 V regulator of RF24 [6]. Additionally, breakout board is also connected to the RF24 module, which ensure a sturdy connection for wire jumpers, but also able to supply up to 1 A current, which is sufficient for data transmission process, as depicted in Fig. 4. Data transmission test was performed again with exact setup as previous experiment.

The comparison result of unmodified RF24 board and enhanced RF24 board are revealed in Table 1, where in LOS situation, the transmission of unmodified RF24 can reach up to 70 m, increased by 290 m for the enhanced radio module, with higher frequency of incoming sensor data being displayed on serial monitor. In a circumstance where the Tx and Rx are facing partial obstruction, such as spaced trees, the nLOS range, is 40–50 m for default RF board. Meanwhile, the implementation of signal enhancement technique has boosted the range up to 330 m, due to a more stable current being supplied. Finally, when the Tx board is subjected to complete obstruction, the NLOS test longest delay of incoming signal, denoting total signal loss when Rx is not within 10 m. With the addition of capacitor and breakout board, the stability of data transmission is maintained up to 160 m.

Table 1 Transmission distance comparison for nRF24L01 +

	No breakout and capacitor	With breakout and capacitor
<i>Board setup</i>		
Obstructions	Transmission distance (meters)	
No obstructions	70	358
Spaced trees	40–50	326
Buildings	10	162

3.2 Discussion

Results show that the capability of nRF24L01+ as wireless module is heavily affected by the environment of RF module for both transmission board and receiving board. The application of SMS tripod which focused on Harumanis plantation where it is an open plantation, has averted the possibility of RF communication network faced interruption from other wireless network devices such as Wi-Fi, mobile network, Bluetooth and other wireless modules that utilize the 2.4 GHz frequency band. In addition, during NLOS test, it is possible for the signal from the transmitter radio module to be reflected off wall, ceilings and metallic objects, causing a short transmission capability. Similar research such as [7] also implement nRF24L01+ in range testing, however, the study does not gather information from indoor environment. The limitations in this experimentation are the test area is not totally free from other wireless network which is Wi-Fi, due to the location is surrounded by other research laboratory with Wi-Fi access. However, the results produced is sufficient to answer the initial hypothesis by providing information on transmission range and communication stability.

As a recommendation, to ensure a maximum transmission range and constant stability are achieved in outdoor application, signal enhancement method that can be taken, which is implementing capacitor to provide the extra current that is required during data transmission, aside from ensuring that the Tx module is directly facing the Rx board with clear LOS. This can be achieved by placing the module high into the air, avoiding the Fresnel Zone phenomenon, where radio signals are diffracted or bend due to obstruction from solid object such as trees and buildings [8]. In addition, implementing breakout board for nRF24L01+ which has 3.3 V voltage regulator as well as ensuring sufficient current is supplied during wireless communication for the use for data monitoring of SMS tripod.

4 Conclusion

Instrument testing for RF range and transmission stability is crucial in agricultural environments to ensure a reliable wireless communication for precision agriculture applications. With the obtained experimentation result, further analysis will be performed from the researcher side, in order to improve the current SF monitoring system by focusing on enhancing the wireless communication range and ensuring a stable data transmission. Through implementation of suggested recommendation, it is possible to boost the RF range beyond the radius obtained from experiment result. Nonetheless, as a future work perspective, a graphical user interface transmission medium for surveillance purpose [9] with optimized scheduling can be integrated [10]. Finally, the expected outcome from this study is to provide a better SF monitoring system that is cost-effective with efficient analysis, that can benefit local farmers in managing their crop plantation, specifically Harumanis plantation.

Acknowledgements This work was supported by a **Universiti Sains Malaysia, Special (Matching) Short-Term Grant with Project No: 304/PAERO/6315715**. The authors would like to acknowledge School of Aerospace Engineering, Universiti Sains Malaysia and Universiti Malaysia Perlis for providing research facilities to conduct this work.

References

1. Sani MA, Abbas H, Jaafar MN, Abd Ghaffar MB (2018) Morphological Characterisation of Harumanis Mango (*Mangifera indica* Linn.) in Malaysia. *Int J Environ Agric Res* 4(1):36–42
2. Perlis JPN, Perlis JPMN (2021) Pertanian: Tanaman Buahhan Terpilih Negeri Perlis. *DOSM/DOSM.PERLIS/1.2021/Siri 31*, pp 1–4
3. Sishodia RP, Ray RL, Singh SK (2020) Applications of remote sensing in precision agriculture: a review (indices vegetativos utilizados na agricultura). *Remote Sens* 12(19):1–31
4. Shatir SMNBS, Ismail AH, Akhtar MN, Abdullah MN, Bakar EA (2022) Development of standstill monitoring system (SMS) for inhouse agriculture product application. *Lect Notes Electr Eng* 829:852–857. https://doi.org/10.1007/978-981-16-8129-5_130
5. Nieminen H, Ermolov V, Nybergh K, Silanto S, Ryhänen T (2002) Microelectromechanical capacitors for RF applications. *J Micromech Microeng* 12(2):177–186. <https://doi.org/10.1088/0960-1317/12/2/312>
6. HowToMechatronics.com (2019) nRF24L01—How It Works, Arduino Interface, Circuits, Codes. HowToMechatronics.com. p 40 (Online). Available: <https://howtomechatronics.com/tutorials/arduino/arduino-wireless-communication-nrf24l01-tutorial/>
7. Akinyemi LA, Shoewu OO, Ajasa AA, Alao WA (2017) Design and development of a 2.4 GHz slot antenna. *Pac J Sci Technol* 18(18):67–77
8. L-Com (2013) Wireless LOS terminology (Online). Available: <https://www.l-com.com/resource/blog/wireless-los-terminology>
9. Abdullah A, Abu Bakar E, Mohamed Pauzi MZ (2015) Monitoring of traffic using unmanned aerial vehicle in Malaysia landscape perspective. *J Teknol* 76(1). <https://doi.org/10.11113/jt.v76.4043>
10. Akhtar MN, Saleh JM, Awais H, Bakar EA (2020) Map-reduce based tipping point scheduler for parallel image processing. *Expert Syst Appl* 139:112848. <https://doi.org/10.1016/j.eswa.2019.112848>

Enhanced Parameter Estimation of Solar Photovoltaic Models Using QLESCA Algorithm



Qusay Shihab Hamad , Sami Abdulla Mohsen Saleh ,
Shahrel Azmin Suandi , Hussein Samma , Yasameen Shihab Hamad ,
and Imran Riaz

Abstract Photovoltaic (PV) systems are recognized as an important section in the utilization of solar power, and the optimisation, control, and mockup of these systems are of great significance. However, the performance of PV systems is mainly motivated by model constraints that are varying and often absent, making their accurate and robust estimation a challenge for existing methods. In this study, the effect of using the Q-learning embedded sine cosine algorithm (QLESCA) in the selection of optimal PV model parameters is investigated. The performance of QLESCA is evaluated and compared with other optimizers. The results show that QLESCA achieves higher efficiency in accurately estimating PV model parameters. This research provides an efficient and effective method for identifying optimal PV model parameters and contributes to the field of PV system optimization, control, and simulation.

Keywords QLESCA · Sine cosine algorithm · Photovoltaic models · Solar photovoltaic system · Parameter Estimation

Q. S. Hamad · S. A. M. Saleh · S. A. Suandi (✉) · I. Riaz
Intelligent Biometric Group, School of Electrical and Electronic Engineering, Universiti Sains
Malaysia, Engineering Campus, 14300 Nibong Tebal, Penang, Malaysia
e-mail: shahrel@usm.my

Q. S. Hamad
University of Information Technology and Communications (UOITC), Baghdad, Iraq

H. Samma
SDAIA-KFUPM Joint Research Center for Artificial Intelligence (JRC-AI), King Fahd University
of Petroleum and Minerals, Dhahran, Saudi Arabia

Y. S. Hamad
Ministry of Education, Baghdad, Iraq

1 Introduction

The optimization of photovoltaic (PV) models is a crucial step in obtaining accurate and efficient predictions of PV system performance. PV models can be described by different mathematical equations, with the sole diode (SD) and dual diode (DD) models being the most commonly used. Accurate parameter values are necessary for obtaining high performance of solar system. Therefore, searching for optimal parameters of solar system (PV models) can be seen as an optimization challenge, which can be solved by a robust optimization approach. Recently, population-based search engines, such as the metaheuristic technique, have made significant achievements in selecting the best values for the constraints of PV models.

In recent years, numerous studies have focused on the parameter extraction and optimization of solar photovoltaic models. Khani et al. [1] developed a computational finite volume code to accurately obtain the temperature-dependent performance of a photovoltaic-thermal solar collector by using Genetic Algorithm to maximize the electrical and thermal efficiencies. Wang et al. [2] proposed a hybrid Optimizer for parameter estimation of both static and dynamic photovoltaic models, which performs well under environmental fluctuations. Meanwhile, Abbassi et al. [3] proposed an Improved Arithmetic Optimization Algorithm (IAOA) to obtain the parameters of solar cells from experimental databases or manufacturer's datasheets. These approaches highlight the importance of optimization algorithms in accurately parameterizing solar photovoltaic models and demonstrate their potential in improving the efficiency and performance of solar photovoltaic systems.

However, some challenges are associated with the optimization of PV models, including the presence of multiple local optima and complex relationships between the model parameters. To address these challenges, various optimization algorithms have been proposed and applied to solve the problem of parameter extraction. One promising algorithm is QLESCA [4], which has shown great performance in solving optimization problems in various fields.

In this study, we propose the use of QLESCA for optimizing the restrictions of the SD and DD PV models. As far as our knowledge extends, QLESCA has not been previously employed for addressing this particular problem, and this study aims to evaluate its effectiveness in handling the challenges of optimizing PV models. The motivation for using QLESCA is based on its exceptional performance in other applications, such as [5]. The results of this study will provide insights into the potential of QLESCA for optimizing PV models and contribute to the development of more accurate and efficient PV models.

The subsequent sections of this paper are structured as follows. Section 2 presents the problem formulation of the PV system, where the mathematical models and objective functions used in the optimization process are described. In Sect. 3, we provide an in-depth analysis of the experimental results and compare the performance of the optimization algorithms. Finally, in Sect. 4, we draw conclusions and discuss future research directions.

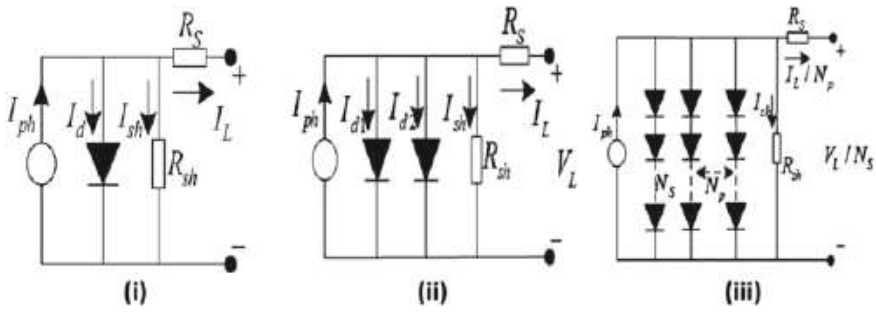


Fig. 1 The illustration (i) SD model, (ii) DD model, and (iii) PV module model

2 Problem Formulation of the PV System

The development of a mathematical structure is a crucial step in describing the behavior of PV systems. Two widely used models are SD and DD, in addition to the photovoltaic module model. A key aspect of using optimization algorithms is the design of an appropriate cost function, where high-performing individuals are selected based on the values of the objective function. In constraint estimation for PV models, the aim is to find parameter values that accurately approximate experimental data, thereby minimizing the error between them. A detailed description of these models and their associated objective functions is presented in [6] for further reference.

Based on the circuits of the three models shown in Fig. 1, the SD model contains five unknown parameters that require accurate estimation. These parameters include photo-generated current (I_{ph}), reverse saturation current of diode (I_{sd}), series resistance (R_s), shunt resistance (R_{sh}), and diode ideality factors (n).

In comparison to the SD model, the DD model needs recognition of more parameters (I_{ph} , I_{sd1} , I_{sd2} , R_s , R_{sh} , n_1 , n_2). To be precise, I_{sd1} and I_{sd2} are employed for representing the diffusion and saturation currents, respectively, while n_1 and n_2 correspond to the ideal parameters for the diffusion diode and the recombination diode, respectively. The Photovoltaic module model, on the other hand, also contains five unknown parameters (I_{ph} , I_{sd} , R_s , R_{sh} , n) similar to the SD. The parameter ranges for different PV modules are clarified in Table 1.

3 Experimental Analysis

The performance of QLESCA on solving this problem has been compared against two recent optimizers Sine Cosine Algorithm (SCA) [7, 8] and Arithmetic Optimization Algorithm (AOA) [9]. All algorithms configurations employed in the comparative analysis were maintained in accordance with their original specifications. Notably, each experiment was meticulously repeated 30 times to mitigate the influence of

Table 1 Parameter ranges for various photovoltaic models

Parameter	SD/DD	PV module		
	Smallest value	Highest value	Smallest value	Highest value
I_{ph} (A)	0	1	0	2
I_{sd}, I_{sd1}, I_{sd2} (μ A)	0	1	0	50
R_s (Ω)	0	0.5	0	2
R_{sh} (Ω)	0	100	0	2000
n, n_1, n_2	1	2	1	50

stochastic variations on the results. It is important to emphasize that the maximum iteration count for this study was set at ten thousand.

Table 2 provides a comprehensive summary of performance metrics, encompassing mean cost and standard deviation values, across all algorithms subjected to scrutiny. The findings unequivocally highlight the superior performance of the QLESCA algorithm when juxtaposed with its counterparts. Specifically, QLESCA consistently outperforms alternative algorithms.

The data presented in Table 2 indicates that QLESCA achieved the most favorable performance with the lowest mean cost and standard deviation across all three problem domains, underscoring its efficacy in the realm of Photovoltaic Models design. SCA claimed the second position in performance, with AOA following suit. The optimal parameters for all three models were successfully estimated through the optimization techniques detailed in Tables 3, 4, and 5.

To further elucidate this comparison, a comprehensive analysis was conducted by plotting the convergence curves of QLESCA in comparison to the other algorithms under evaluation. The convergence curves of QLESCA in contrast to SCA and AOA

Table 2 The average cost and standard deviation of QLESCA against other optimizers

Algorithms	QLESCA		SCA		AOA	
	Avg	Std	Avg	Std	Avg	Std
Sole diode	0.00485	0.00257	0.06151	0.01404	0.14658	0.04842
Dual diode	0.00632	0.00349	0.07203	0.04323	0.13168	0.05676
Photovoltaic module	0.00417	0.00250	0.20645	0.09196	0.42836	0.02934

Bold significance the best value corresponds to the smallest one

Table 3 Selected parameters for SD model

Algorithm	I_{ph} (A)	I_{sd} (μ A)	R_s (Ω)	R_{sh} (Ω)	n
QLESCA	0.7612	0	0.0338	62.9387	1.5454
SCA	0.7495	0	0.0127	48.5540	1.5311
AOA	0.7231	0	0.0095	69.9970	1.5509

Table 4 Selected parameters for DD model

Algorithm	I_{ph} (A)	I_{sd1} (μ A)	R_s (Ω)	R_{sh} (Ω)	n_1	I_{sd2} (μ A)	n_2
QLESCA	0.7639	0	0.0346	58.2529	1.6304	0	1.6908
SCA	0.7471	0	0.0042	54.0233	1.5632	0	1.5074
AOA	0.7113	0	0.0063	67.6931	1.4098	0	1.5798

Table 5 Selected parameters for PV module

Algorithm	I_{ph} (A)	I_{sd} (μ A)	R_s (Ω)	R_{sh} (Ω)	n
QLESCA	0.0010	0	0.0012	1.3378	0.0494
SCA	1.2721	0	0.1401	402.2723	23.6314
AOA	0.0008	0	0.0009	1.1459	0.0396

for the parameters related to the solar diode (SD), diffusion diode (DD), and module diode are presented in Figs. 2, 3, and 4, respectively.

Fig. 2 The curve of convergence for QLESCA, SCA, and AOA in SD model

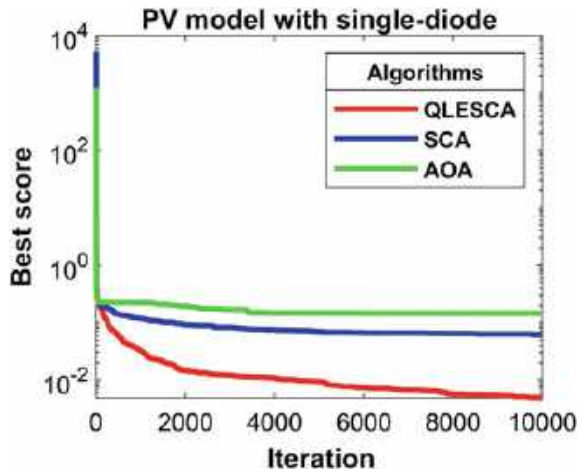


Fig. 3 The curve of convergence for QLESCA, SCA, and AOA in DD model

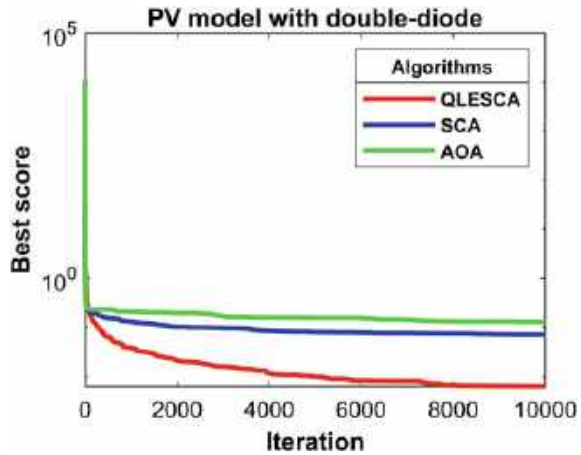
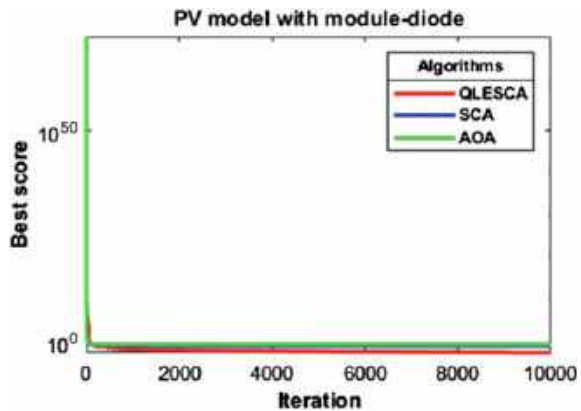


Fig. 4 The curve of convergence for QLESCA, SCA, and AOA in module diode



4 Conclusion

In this paper, the QLESCA optimizer was utilized to guess the constraints of various solar models and to solve these design problems in an efficient manner. Comprehensive experiments were conducted on different solar systems (PV models), and the results demonstrate that QLESCA outperforms other optimizers. In future studies, we will investigate the QLESCA’s performance on more complex engineering problems and explore recent design problems related to PV models to provide comprehensive solutions for renewable resources.

Acknowledgements We extend our sincere appreciation to the Malaysia Ministry of Higher Education for their invaluable support through the Fundamental Research Grant Scheme (FRGS), under grant no. FRGS/1/2019/ICT02/USM/03/3.

References

1. Khani MS, Baneshi M, Eslami M (2019) Bi-objective optimization of photovoltaic-thermal (PV/T) solar collectors according to various weather conditions using genetic algorithm: a numerical modeling. *Energy* 189:116223. <https://doi.org/10.1016/j.energy.2019.116223>
2. Wang S, Yu Y, Hu W (2021) Static and dynamic solar photovoltaic models' parameters estimation using hybrid Rao optimization algorithm. *J Clean Prod* 315:128080. <https://doi.org/10.1016/j.jclepro.2021.128080>
3. Abbassi A, Ben Mehrez R, Touaiti B, Abualigah L, Touti E (2022) Parameterization of photovoltaic solar cell double-diode model based on improved arithmetic optimization algorithm. *Optik (Stuttg)* 253:168600. <https://doi.org/10.1016/j.ijleo.2022.168600>
4. Hamad QS, Samma H, Suandi SA, Mohamad-Saleh J (2022) Q-learning embedded sine cosine algorithm (QLESCA). *Expert Syst Appl* 193:116417. <https://doi.org/10.1016/j.eswa.2021.116417>
5. Hamad QS, Samma H, Suandi SA (2023) Feature selection of pre-trained shallow CNN using the QLESCA optimizer: COVID-19 detection as a case study. *Appl Intell*. <https://doi.org/10.1007/s10489-022-04446-8>
6. Liang J et al (2020) Parameters estimation of solar photovoltaic models via a self-adaptive ensemble-based differential evolution. *Sol Energy* 207:336–346. <https://doi.org/10.1016/j.solener.2020.06.100>
7. Mirjalili S (2016) SCA: A Sine Cosine Algorithm for solving optimization problems. *Knowledge-Based Syst* 96:120–133. <https://doi.org/10.1016/j.knosys.2015.12.022>
8. Hamad QS, Samma H, Suandi SA, Saleh JM (2022) A comparative study of sine cosine optimizer and its variants for engineering design problems, pp 1083–1089
9. Abualigah L, Diabat A, Mirjalili S, Abd Elaziz M, Gandomi AH (2021) The arithmetic optimization algorithm. *Comput Methods Appl Mech Eng* 376:113609. <https://doi.org/10.1016/j.cma.2020.113609>

Active Disturbance Rejection Control of Flexible Joint System



Li Qiang and Nur Syazreen Ahmad

Abstract This paper presents an ADRC controller designed for a single link manipulator model, considering torsional vibrations in industrial robots. A fourth model is introduced to capture the complex dynamics. The ADRC successfully tracks desired trajectories despite non-linearity, uncertainties, and disturbances. Linear and nonlinear extended-state observers (ESOs) estimate and compensate for uncertainties. The control law is derived from the ESO by configuring the observer gain. Simulation results demonstrate the superior control performance of ADRC compared to PID, analyzing step and sinusoidal signals on the flexible joint system.

Keywords Active disturbance rejection control · Extended state observer · Feedback control · Flexible joint system

1 Introduction

Robots' technology has been widely used in manufacture, transportation and utilities [1, 2]. However, the complex behavior and high precision requirements of robotic applications pose significant challenges for researchers [3–8]. Robust control techniques are necessary to meet the demands of robotic tasks [9]. In industrial settings, trajectory tracking control of robotic manipulators with joint flexibility is complex, particularly when fast response is needed. The mechanical structure of flexible joints, including gears, belts, tendons, bearings, and hydraulic lines, affects joint elasticity and adds complexity to the control dynamics. A detailed description of joint elasticity is achieved by considering the link connected to a motor through a torsional

L. Qiang · N. S. Ahmad (✉)

School of Electrical and Electronic Engineering, Universiti Sains Malaysia, 14300 Nibong Tebal, Penang, Malaysia

e-mail: syazreen@usm.my

L. Qiang

e-mail: liqiang@student.usm.my

spring, resulting in a fourth model of joint flexibility. Compared to rigid joints, the plant order doubles, making precise control challenging [10].

In the literature, two main control approaches are commonly used for joint flexibility: adaptive control and robust control [9, 11]. Adaptive control requires a known mathematical model with disturbance and time-varying parameters, allowing for controller adaptation through parameter estimation and updating. Robust controllers explicitly handle uncertainties without continuous adjustment of controller parameters. Ghorbel et al. [12] present adaptive control results for flexible joint manipulators, showing that adaptive control for rigid robots can be extended to flexible joints by adding a correction term to dampen elastic oscillations. In [13], the robustness of closed-loop systems is established under specific conditions, solving the dynamic trajectory tracking control problem of manipulators with flexible joints using only position measurements on the link and motor. Sliding mode control design in [14], requires information on uncertainty bounds and complete state vectors for implementation.

While previous methods focus on mathematical models and control theories, developing a desired control system for practical applications can be challenging. Active disturbance rejection control (ADRC) has emerged as a paradigm shift in feedback control system design, combining experimental and systematic approaches to achieve a balance between experience and reasoning, particularly in control systems with uncertainties. Linear modularization ADRC has been applied to a 1-degree-of-freedom manipulator to track desired motion, demonstrating significant improvements by adding a module applicable to all observer-based controllers [15]. Linear ADRC has also been employed for trajectory tracking in a 2-degree-of-freedom rigid link manipulator, showing advantages over traditional PID controllers [16].

2 Methodology

2.1 Models for Flexible Joint System

The single-link manipulator with flexible joint operation will be considered as controlled plant. The FJS model consists of a dc motor, a linear torsional spring and joint. The motion equations from this system, can be written as

$$I\ddot{q}_1 + MgL \sin(q_1) + K(q_1 - q_2) = 0; J\ddot{q}_2 - K(q_1 - q_2) = u$$

q_1 and q_2 in equations show the angles of link and motor, respectively, I is the link inertia, J is the inertia of motor, K is the spring stiffness. The mass and length of link are symbolized as M and L accordingly. The system is a single input and single output system. Input torque u is input signal, q_1 is output signal for tracking. If the state variables $[x_1 x_2 x_3 x_4] = [q_1 \dot{q}_1 q_2 \dot{q}_2]$ are chosen from the equations, then the nonlinear dynamics can be written in state space. The output of the system is $y = x_1$.

$$\dot{x}_1 = x_2; \dot{x}_2 = -\frac{MgL}{I} \sin(x_1) - \frac{K}{I}(x_1 - x_3); \dot{x}_3 = x_4; \dot{x}_4 = \frac{K}{I}(x_1 - x_3) + \frac{1}{J}u$$

Theoretically, with the transformation of $z = T(x)$ and nonlinear feedback $u = (-a(z) + v)/b(z)$ with $b(z) \neq 0$, the new state of system z and new input v satisfy a linear time invariant relation $\dot{z} = Az + Bv$. $z = T(x)$ is selected as

$$z_1 = x_1; z_2 = x_2; z_3 = -\frac{MgL}{I} \sin(x_1) - \frac{K}{I}(x_1 - x_3);$$

$$z_4 = -\frac{MgL}{I} \cos(x_1)x_2 - \frac{K}{I}(x_2 - x_4)$$

State-space equations $\dot{z} = Az + Bv$ can be obtained by operation of $z = T(x)$.

$$\dot{z}_1 = z_2; \dot{z}_2 = z_3; \dot{z}_3 = z_4; \dot{z}_4 = a(z) + bu; y = z_1$$

$$a(z) = \frac{MgL}{I} \sin(z_1) \left(z_2^2 - \frac{K}{J} \right) - \left(\frac{MgL}{I} \cos(z_1) \right) + \frac{K}{J} + \frac{K}{I} z_3; b = \frac{K}{IJ}$$

$$A = [0 \ 1 \ 0 \ 0; 0 \ 0 \ 1 \ 0; 0 \ 0 \ 0 \ 1; 0 \ 0 \ 0 \ 0]; \quad B = [0; 0; 0; 1]$$

The control law $u = \frac{1}{b}(-a(z) + v)$ is deduced by state feedback, which establish the relationship between the states and new reference v . It's $v = \dot{z}_4^* + k_1(z_1^* - z_1) + k_2(z_2^* - z_2) + k_3(z_3^* - z_3) + k_4(z_4^* - z_4)$. where the link position, velocity, acceleration, and jerk respectively can be symbolized by z_1, z_2, z_3, z_4 , the asterisk variables represent the reference value of z_1, z_2, z_3, z_4 .

2.2 ADRC

The ADRC methodology was firstly proposed and developed by Han in 1980s [17]. The control strategy is designed to reject the effects of disturbances in a control system. Generally, there are three components in ADRC. They are tracking differentiator (TD), the extended state observer (ESO), and ESO-based feedback control.

The object of control allows the output of the system $y(t)$ to track very closely a given reference signal $r(t)$. Thus, to design a control law is essential. In consideration of the flexible joint system, a 4th order nonlinear system is introduced as follows

$$\dot{z}_1 = z_2; \dot{z}_2 = z_3; \dot{z}_3 = z_4; \dot{z}_4 = a(t, z_1, z_2, z_3, w) + bu; y = z_1$$

where $z_i (i = 1 \dots 4)$ are four states of flexible joint system, $w(t)$ is the disturbances and b is input gain. Let $b = b_0 + \Delta b$, where b_0 is the best available estimates of b

and Δb is its associated uncertainty to be determined as $d \triangleq a + \Delta bu$ and designating it as an extended virtual state $z_5 = d$, and form an augmented state-space model, which is ultimately used to estimate and compensate for the lumped disturbance in the disturbance rejection process. The dynamics aforementioned can be rewritten in a state-space form as following, in which, h is the rate of change of uncertainty which is assumed to be an unknown but bounded function.

$$\dot{z}_1 = z_2; \dot{z}_2 = z_3; \dot{z}_3 = z_4; \dot{z}_4 = z_5 + b_0u; \dot{z}_5 = h; y = z_1$$

The ESO are a series of functions which input signal are u and e_1 . It is designed as

$$\begin{aligned} \dot{\hat{z}}_1 &= \hat{z}_2 - \beta_1g_1(e_1); \dot{\hat{z}}_2 = \hat{z}_3 - \beta_2g_2(e_1); \dot{\hat{z}}_3 = \hat{z}_4 - \beta_3g_3(e_1); \\ \dot{\hat{z}}_4 &= \hat{z}_5 - \beta_4g_4(e_1) + b_0u(t); \dot{\hat{z}}_5 = -\beta_5g_5(e_1); \end{aligned}$$

where $e_1 = \hat{z}_1 - z_1$, $g_i (i = 1 \dots 5)$ are different functions. By selecting the appropriate function g_i and observer gains β_i , and the states \hat{z}_i in observer are expected to approximately match the states $z_j (j = 1 \dots 4)$ and the total disturbance a in physical plant. The extend state z_5 is used for the estimating the total disturbance.

3 Results and Discussion

The parameters of the pant are listed as $MgL = 10 \text{ N m}$, $K = 100 \text{ N m/rad}$, $I = 1 \text{ kg m}^2$, $J = 1 \text{ kg m}^2$. The plant model can be described with the extended order system in state-space as $\dot{z} = Az + Bu + Eh$, $y = Cz$. $z = [z_1z_2z_3z_4z_5]$ is the state vector of the extended -order system.

The state-space model of LESO dynamics can be written as $\dot{\hat{z}} = A\hat{z} + Bu + LC(z - \hat{z})$ where $L = [\beta_1\beta_2\beta_3\beta_4\beta_5]^T$ is the observer gain vector. The poles of observer are configured at $s_1 = -43.2, s_2 = -42, s_3 = -42.5, s_4 = -42.7, s_5 = -41.1$. The ESO and real plant have the same input signal u , which has a close relationship connection with reference signals r , the relationship between u and r constitutes the control law. The observer gain obtained is $L = [2121.8e47.6e51.6e71.4e8]$. That is $u = K_s(R - \hat{Z}) + \frac{1}{b_0}(\dot{\hat{z}}_4^* - \hat{z}_5)$. By placing the poles at $(1 + \tau_c/4)^4$ with a time constant of $\tau_c = 0.35$, we obtain ADRC controller feedback gains $K_S = [171.2 \ 59.8 \ 7.85 \ 0.46]$. The performance of ADRC is compared against PID control which is optimized using MATLAB.

Simulations are conducted to evaluate two aspects of ADRC performance. Firstly, its ability to reject internal disturbances will be examined, and secondly, its superior control effectiveness will be compared to the traditional PID controller. To gain visual insights into control behavior under varying process parameters, a simulation was

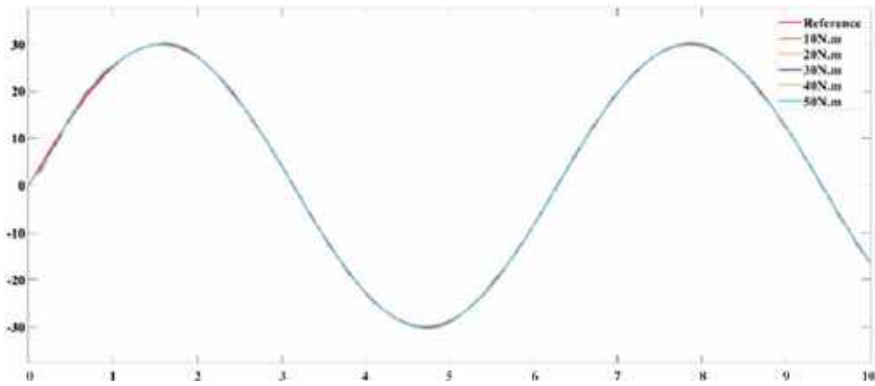


Fig. 1 Fixed ADRC controlling joint flexible process with varying load

conducted with fixed ADRC parameters while adjusting the process model parameter MgL . The values ranged from the nominal value of 10–50 N m, representing a tenfold increase. The simulation results demonstrate that regardless of the variations in load torque, the joint's output perfectly tracks the reference signal. Figure 1 illustrates the obtained results and the corresponding process parameter settings.

To evaluate the performance of the ADRC controller, another simulation is conducted to compare its speed and accuracy with the PID controller. The simulation involves applying step and sinusoidal signals, selected manually using a switch in Simulink, with a duration of 50 s. Figure 2 illustrates the step responses of the flexible joint system, comparing the ADRC and PID controllers. It is evident that ADRC outperforms PID in terms of static characteristics, reaching a stable state within 5 s, whereas PID requires nearly 40 s. Furthermore, minor variations are observed under PID control after the plant reaches stability. Figure 3 displays the sinusoidal responses with a frequency of 1 Hz and amplitude of 30° . The ADRC control strategy effectively tracks the joint trajectory in real-time, while the PID control exhibits overshoot at the peak angle and lags behind the reference trajectory. In summary, the ADRC controller demonstrates superiority over the classical PID controller for the complex flexible joint system in terms of both static and dynamic characteristics.

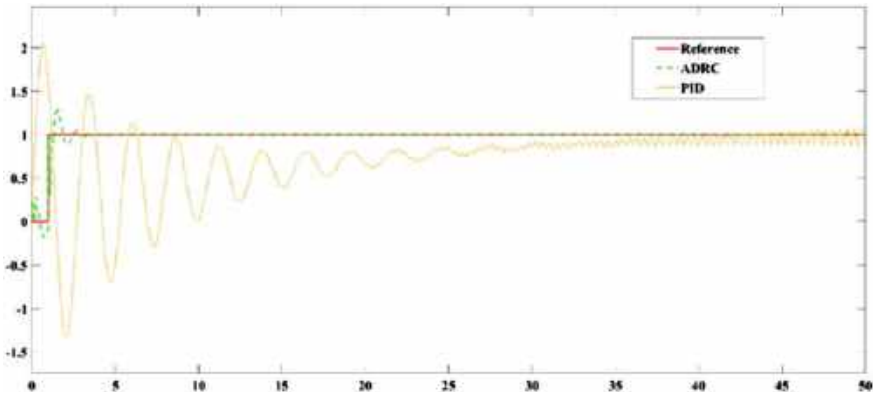


Fig. 2 Comparison of ADRC and PID under step responses

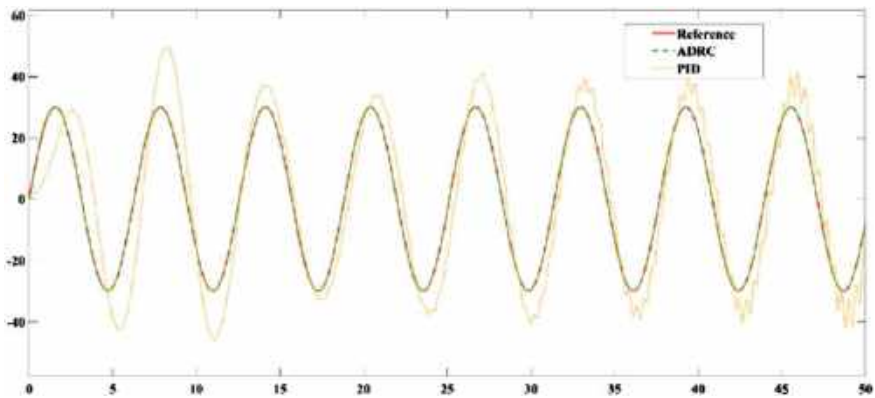


Fig. 3 Comparison of ADRC and PID under sinusoidal responses

4 Conclusion

In this paper, the canonical form of the Flexible Joint System (FJS) is derived using a diffeomorphic coordinate transformation. The Active Disturbance Rejection Control (ADRC) technology is then applied to the model. One simulation demonstrates that ADRC can effectively reject disturbances even when the process parameters change, making it suitable for time-varying systems. Comparing it to the traditional PID controller, the simulation shows that ADRC achieves satisfactory tracking performance in both step and sinusoidal responses. Therefore, ADRC proves to be robust and practical for applying to complex systems with high non-linearity, such as FJS.

References

1. Loganathan A, Ahmad NS (2023) A systematic review on recent advances in autonomous mobile robot navigation. *Eng Sci Technol Int J* 40:101342
2. Ting TM, Ahmad NS, Goh P, Mohamad-Saleh J (2021) Binaural modelling and spatial auditory cue analysis of 3D-printed ears. *Sensors* 21(1):227
3. Chan SY, Ahmad NS, Ismail W (2017) Anti-windup compensator for improved tracking performance of differential drive mobile robot. In: 2017 IEEE international systems engineering symposium (ISSE), 2017, pp 1–5
4. Teo JH, Loganathan A, Goh P, Ahmad NS (2020) Autonomous mobile robot navigation via RFID signal strength sensing. *Int J Mech Eng Robot Res* 9(8):1140–1144
5. Arrouch I, Ahmad NS, Goh P, Mohamad-Saleh J (2022) Close proximity time-to-collision prediction for autonomous robot navigation: an exponential GPR approach. *Alex Eng J* 61(12):11171–11183
6. Arrouch I, Mohamad-Saleh J, Goh P, Ahmad NS (2022) A comparative study of artificial neural network approach for autonomous robot's TTC prediction. *Int J Mech Eng Robot Res* 11(5):345–350
7. Ahmad NS, Teo JH, Goh P (2022) Gaussian process for a single-channel EEG decoder with inconspicuous stimuli and eyeblinks. *Comput Mater Continua* 73(1):611–628
8. Bin Lu L, Ahmad NS, Goh P (2023) Self-balancing robot: modeling and comparative analysis between PID and linear quadratic regulator. *Int J Reconfigurable Embedded Syst* 12(3):351–359
9. Ahmad NS (2020) Robust H_{∞} -fuzzy logic control for enhanced tracking performance of a wheeled mobile robot in the presence of uncertain nonlinear perturbations. *Sensors* 20(13):7673
10. Syed Mubarak Ali SAA, Ahmad NS, Goh P (2019) Flex sensor compensator via Hammerstein-wiener modeling approach for improved dynamic goniometry and constrained control of a bionic hand. *Sensors (Switzerland)* 19(18)
11. Ahmad NS (2017) Robust stability analysis and improved design of phase-locked loops with non-monotonic nonlinearities: LMI-based approach. *Int J Circ Theory Appl* 45(12)
12. Ghorbel F, Hung JY, Spong MW (1989) Adaptive control of flexible-joint manipulators. *IEEE Control Syst Mag* 9(7)
13. Lee HS, Bien Z (1996) Tracking control of flexible joint manipulators using only position measurements. *Int J Control* 64(3)
14. Spurgeon SK, Yao L, Lu XY (2001) Robust tracking via sliding mode control for elastic joint manipulators. *Proc Inst Mech Eng Part I J Syst Control Eng* 215(4)
15. Xue W, Madonski R, Lakomy K, Gao Z, Huang Y (2017) Add-on module of active disturbance rejection for set-point tracking of motion control systems. *IEEE Trans Ind Appl* 53(4)
16. Hu H, Xiao S, Shen H (2021) Modified linear active disturbance rejection control for uncertain robot manipulator trajectory tracking. *Math Probl Eng* 2021
17. Han J (2009) From PID to active disturbance rejection control. *IEEE Trans Ind Electron*

Research on Synchronous Control of Double-Cylinder Electro-Hydraulic Position Servo System Based on Active Disturbance Rejection Control



Liu Lizhen, Li Qiang, and Nur Syazreen Ahmad

Abstract The electro-hydraulic servo position synchronization control system of hydraulic cylinders is prone to internal and external interference, which results in the inconsistent positioning of the cylinders. To address this issue, a mathematical model of the hydraulic cylinder's electro-hydraulic servo control system was developed based on a test platform that synchronizes both cylinders' positions. The "equivalent + master-slave" control mode was employed due to the drawbacks of "equivalent mode" and "master-slave mode" in engineering applications. An active disturbance rejection controller (ADRC) was designed, and a simulation model of synchronous active disturbance rejection control of both cylinders was created. The simulation results were compared to those of traditional PID control, revealing that the active disturbance rejection controller can more effectively enhance the system's tracking performance and anti-interference capability while reducing the synchronization error.

Keywords Electro-hydraulic servo · Position synchronization control · Equal + master-slave · Active disturbance rejection controller (ADRC)

1 Introduction

Electro-hydraulic servo control technology offers the advantages of simple structure, strong reliability, and high control accuracy, making it a crucial bridge connecting modern control technology and high-power engineering equipment [1]. Hydraulic synchronous drive technology based on electro-hydraulic servo control

L. Lizhen · L. Qiang · N. S. Ahmad (✉)

School of Electrical and Electronic Engineering, Universiti Sains Malaysia, 14300 Nibong Tebal, Penang, Malaysia
e-mail: syazreen@usm.my

L. Lizhen

Department of Mechanical and Electrical Engineering, Cangzhou Normal University, Cangzhou Hebei 061001, China

has been widely applied in heavy equipment, such as machine tools, construction machinery, and hydraulics. As aviation technology, automobile manufacturing, precision machine tools, and other advanced manufacturing technologies continue to evolve, higher requirements have been set for hydraulic cylinder position synchronous control. The hydraulic cylinder position synchronization control system is affected by factors such as component installation precision, pipe fitting seal, friction damping, and load fluctuation, resulting in significant synchronization errors that reduce production efficiency. Therefore, it is of great practical significance to study and address this problem.

The precision of hydraulic cylinder synchronization control is influenced by control strategies and algorithms, considering a specific hardware setup. Researchers have explored various approaches to improve position synchronization control [2, 3]. For instance, one study enhanced the equivalent control mode by using a deviation coupling control mode and a cloud model controller as a deviation compensator, effectively reducing synchronization errors. Another study in [4] proposed a master–slave PID control strategy that improved the precision of multi-cylinder synchronization control by combining a cross-coupled fuzzy PID controller with a full decoupling compensation algorithm. However, the electro-hydraulic proportional position synchronous control system of hydraulic cylinders is a complex nonlinear and time-varying system. Many existing algorithms require precise mathematical models and can be challenging to apply in industrial control settings [5, 6]. The traditional PID control algorithm, although simple, is based on a linear model and lacks robustness against interference, making it insufficient for high-performance requirements [7–9]. On the other hand, the active disturbance rejection controller (ADRC) is a nonlinear controller with strong resistance to interference and robustness, offering a straightforward implementation [10]. Therefore, it has gained popularity in modern engineering applications. This paper adopts the “equal + master–slave” control strategy and designs an automatic disturbance rejection controller to simulate the hydraulic cylinder position synchronization control system. Compared to the traditional PID controller, this approach greatly improves synchronization precision, response speed, and anti-interference capabilities.

2 Methodology

Based on the theoretical analysis of the Hydraulic Cylinder Position Control System, the mathematical model of the system is derived according to the system structure and parameters. In this work, the system is modeled based on the test platform of electro-hydraulic servo position synchronous control of two cylinders. The hydraulic control components are servo valves. Figure 1 shows the structure of the electro-hydraulic servo control system of two cylinders.

The system is composed of two identical valve control cylinder hydraulic circuits, so one of the circuits is taken as the research object in the modeling and analysis of the system.

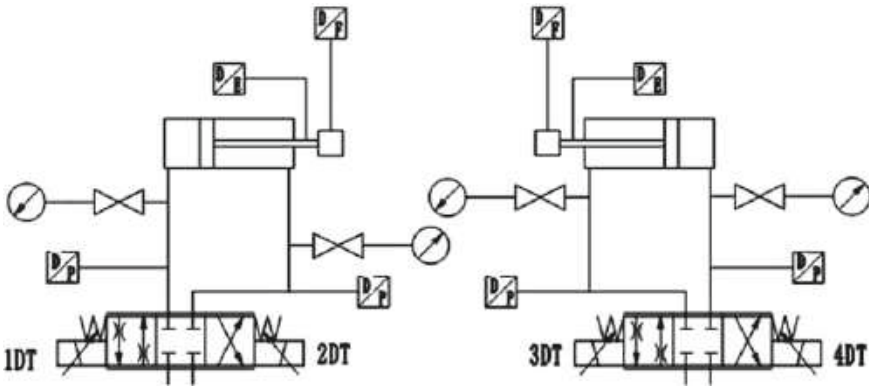


Fig. 1 Two side cylinder electro-hydraulic servo control system structure schematic

(1) Load balance equation of hydraulic cylinder

$$m\ddot{y} = P_L A - B\dot{y} - F_L \tag{1}$$

(2) Flow equation of hydraulic cylinder working chamber

$$\dot{P}_L = \frac{4\beta_c}{V_t} (Q_L - A\dot{y} - C_i P_L) \tag{2}$$

(3) Linear flow equation of servo valve

$$Q_L = K_q x_v - K_c P_L \tag{3}$$

According to Eqs. (1)–(3), it can be obtained as below.

$$y = \frac{\frac{K_q}{A} x_v - \frac{K_{ce}}{A^2} \left(1 + \frac{V_t}{4\beta_c K_{ce}} s \right) F_L}{s \left[\frac{V_t m}{4\beta_c A^2} s^2 + \left(\frac{K_{cem}}{A^2} + \frac{B V_t}{4\beta_c A^2} \right) s + \left(1 + \frac{B K_{ce}}{A^2} \right) \right]} \tag{4}$$

The simplified transfer function is as follows.

$$\frac{y}{x_v} = \frac{\frac{K_Q}{A}}{s \left[\frac{1}{\omega_n^2} s^2 + \frac{2\xi_n}{\omega_n} s + 1 \right]} \tag{5}$$

Servo valves can be equivalent to proportional links.

$$x_v = K_{sv} i = K_{sv} K_e u \tag{6}$$

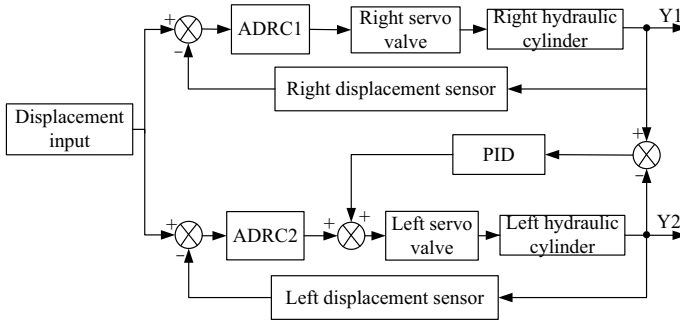


Fig. 2 Schematic diagram of “Equal + Master–slave” control mode

2.1 Synchronous Control Strategy

This paper proposes the control strategy of “equal + master–slave”, and the schematic diagram is shown in Fig. 2. “Equal” means that the given signal is added to both hydraulic cylinders at the same time. “Master–slave” is based on the right hydraulic cylinder, and the difference of displacement of hydraulic cylinders on both sides acts on the left hydraulic cylinder as the input signal to make it follow the rapid change of displacement of the right hydraulic cylinder, so as to ensure that the system can obtain better dynamic and static quality in the process of position synchronous control.

2.2 Active Disturbance Rejection Controller (ADRC)

The active disturbance rejection controller consists of tracking differentiator (TD), extended state observer (ESO) and error nonlinear feedback control law (NLSEF) [6]. In this paper, a second-order active disturbance rejection controller is designed to control the electro-hydraulic servo position synchronous control system of hydraulic cylinder. Its structure is shown in Fig. 3.

- (1) Tracking differentiator (TD)

$$\begin{cases} \dot{v}_1 = v_2 \\ \dot{v}_2 = fhan(v_1 - v, v_2, r, h) \end{cases}$$

- (2) Extended state observer (ESO)

$$\begin{cases} e = z - y \\ \dot{z}_1 = z_2 - b_{01}e \\ \dot{z}_2 = f_1(t, z_1, z_2, w) + z_3 - b_{02}fal(e, a_1, d_2) + b_0u \\ \dot{z}_3 = -b_{03}fal(e, a_2, d_3) \end{cases}$$

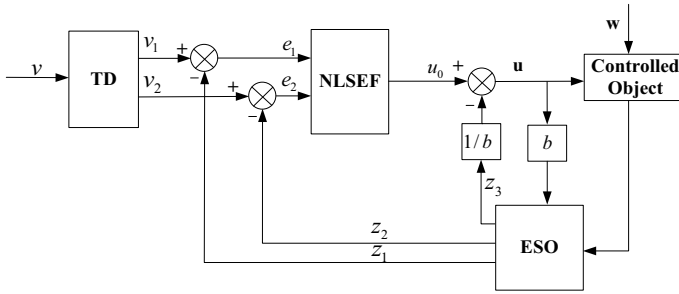


Fig. 3 Active disturbance rejection controller structure diagram

(3) Nonlinear feedback control law (NLSEF)

$$\begin{cases} e_1 = v_1 - z_1 \\ e_2 = v_2 - z_2 \\ u_0 = k_1 fal(e_1, a_1, d_1) + k_2 fal(e_2, a_2, d_2) \end{cases}$$

3 Results and Discussions

Based on the mathematical model of single-side cylinder position control system and the technical manual of the servo valve, the relevant parameters were determined, and the two-side cylinder position synchronization control system model was established in MATLAB, respectively using ADRC and PID for simulation research.

When the displacement is a step signal, the step disturbance signal is added to the right cylinder at 15 s. When the ramp signal is given, the step disturbance signal is added to the right cylinder at 20 s. The simulation results of step input and ramp input are shown in Fig. 4. According to the simulation curve, tracking performance parameters and disturbance rejection performance parameters are obtained, as shown in Tables 1 and 2.

Upon comparing the simulation curves of the active disturbance rejection controller and PID controller, it is evident that the former exhibits superior performance in the electro-hydraulic servo position synchronous control system of two cylinders. The active disturbance rejection controller achieves faster tracking of the given signal, attains stable values with reduced overshoot and shorter adjustment time, compared to the PID controller. Furthermore, the active disturbance rejection controller generates a smaller dynamic drop upon exposure to disturbance signals, which is accompanied by a shorter recovery time, thereby facilitating better interference suppression and reduction of the impact of interference signals on the synchronization control accuracy of both cylinder positions. Hence, it can be inferred that

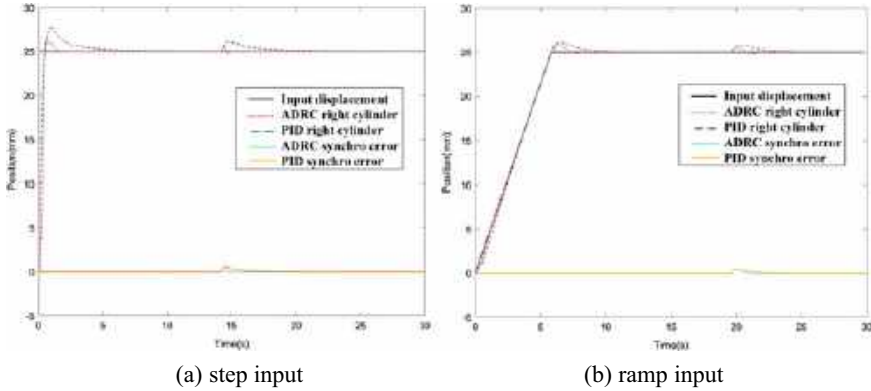


Fig. 4 Simulation curves of ADRC and PID

Table 1 Simulation curve parameters of step input

Controller	t_r/s	$M_p/\%$	t_s/s	$\Delta C_{max}/\%$	t_V/s	Maximum synchro error/mm
ADRC PID	0.53	4.98	2.04	2.16	1.72	0.4322
	0.53	10.7	8.47	4.45	6.31	0.5836

Table 2 Simulation curve parameters of ramp input

Controller	Tracking error/mm	$M_p/\%$	t_s/s	$\Delta C_{max}/\%$	t_V/s	Maximum synchro error/mm
ADRC PID	0.80–0.82	3.39	8.33	2.30	2.1	0.4325
	1.06–1.07	4.25	9.98	3.22	3.89	0.4328

the active disturbance rejection controller improves the synchronization control accuracy of the system with its stronger anti-interference ability and superior tracking performance.

4 Conclusion

Based on an experimental platform for electro-hydraulic servo position synchronization control of two cylinders, this paper establishes a mathematical model of the position synchronization control system. Adopting the “equal + master–slave” control strategy, an active disturbance rejection controller is designed and implemented in MATLAB simulations. Comparing the results with those of a PID controller, it is evident that the active disturbance rejection control algorithm provides numerous advantages in hydraulic cylinder electro-hydraulic servo position synchronization control. The overshoot is smaller, the adjustment time is shorter, synchronization

control precision is higher, and anti-interference ability is stronger, providing a theoretical foundation for the practical application of active disturbance rejection control technology in hydraulic cylinder electro-hydraulic servo position synchronization control systems in the future.

Funding Science and Technology Research Project of Colleges and Universities in Hebei Province, China (ZC2021007).

References

1. Xu B, Shen J, Liu S, Su Q, Zhang J (2020) Research and development of electro-hydraulic control valves oriented to industry 4.0: a review. *Chin J Mech Eng* 33(1):29
2. Bakar SJA, J-Shenn K, Goh P, Ahmad NS (2023) Development of magnetic levitation system with position and orientation control. *Int J Reconfigurable Embedded Syst* 12(2):287–296
3. Chan SY, Ahmad NS, Ismail W (2017) Anti-windup compensator for improved tracking performance of differential drive mobile robot. In: *Proceedings on 2017 IEEE international symposium on systems engineering*. ISSE
4. Huayong Y, Hu S, Guofang G, Guoliang H (2009) Electro-hydraulic proportional control of thrust system for shield tunneling machine. *Autom Constr* 18(7):950–956
5. Ahmad NS (2017) Robust stability analysis and improved design of phase-locked loops with non-monotonic nonlinearities: LMI-based approach. *Int J Circ Theory Appl* 45(12)
6. Loganathan A, Ahmad NS (2023) A systematic review on recent advances in autonomous mobile robot navigation. *Eng Sci Technol Int J* 40:101342
7. Ahmad NS (2020) Robust H_{∞} -fuzzy logic control for enhanced tracking performance of a wheeled mobile robot in the presence of uncertain nonlinear perturbations. *Sensors* 20(13):7673
8. Lau LB, Ahmad NS, Goh P (2023) Self-balancing robot: modeling and comparative analysis between PID and linear quadratic regulator. *Int J Reconfigurable Embedded Syst* 12(3):351–359
9. Ahmad NS (2023) Modeling and hybrid pso-woa-based intelligent pid and state-feedback control for ball and beam systems. *IEEE Access* 11:137866–137880
10. Han J (2009) From pid to active disturbance rejection control. *IEEE Trans Ind Electron* 56:900–906

Passivity-Based Control of Underactuated Rotary Inverted Pendulum System



Minh-Tai Vo, Van-Dong-Hai Nguyen, Hoai-Nghia Duong,
and Vinh-Hao Nguyen

Abstract This paper presents a feedback passivation control scheme, which named passivity-based controller (for short, PBC) for rotational inverted pendulum (for short, RIP), an under-actuated mechanical system which is nonlinear and unstable. Previous studies developed PBC by using one of two existing methods, which include choice of output and feedback passivation. The motivation of this paper is to develop PBC by combining choice of output and feedback passivation. The plant and PBC method are modeled and analyzed in MATLAB/Simulink environment. The end of this work presents a number of simulations regarding stabilization control of RIP by PBC. Theoretical analyses and validation results is provided to demonstrate the effectiveness and robustness of PBC.

Keywords Passivity-based control · Feedback passivation control scheme · Rotational inverted pendulum · PBC · Stabilization

1 Introduction

RIP plays a vital class of nonlinear systems. Many research papers have applied to this system to evaluate control algorithms such as linear quadratic regulator (LQR) controller combined with Neural Network (NN) for fast stabilizing [1], design of

M.-T. Vo (✉) · V.-H. Nguyen

Ho Chi Minh City University of Technology, VNU-HCM, Ho Chi Minh City, Vietnam
e-mail: vmtai.sdh212@hcmut.edu.vn

V.-H. Nguyen

e-mail: vinhhao@hcmut.edu.vn

V.-D.-H. Nguyen

Ho Chi Minh City University of Technology and Education, Ho Chi Minh City, Vietnam
e-mail: hainvd@hcmute.edu.vn

H.-N. Duong

Eastern International University, Thu Dau Mot City, Binh Duong Province, Vietnam
e-mail: nghia.duong@eiu.edu.vn

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
N. S. Ahmad et al. (eds.), *Proceedings of the 12th International Conference on Robotics, Vision, Signal Processing and Power Applications*, Lecture Notes in Electrical Engineering 1123, https://doi.org/10.1007/978-981-99-9005-4_28

223

robust control scheme based on optimal algorithm for RIP with unmatched uncertainty [2], etc. PBC is a methodology that casts the control problems as a search for an interconnection pattern among subsystems such that the overall dynamics exhibits passivity properties, which help infer stability [3]. Accordingly, an application of PBC method is studied in many scientific articles such as research on PBC for under-actuated system with Coulomb friction and applying that to earthquake prevention [4], and so on.

In particular, the objective of this manuscript is summarized as follows: designing of PBC scheme by combining two ways including choice of output and feedback passivation for the under-actuated RIP system. This is main contribution of this paper.

2 Rotary Inverted Pendulum Model

2.1 Dynamic Model

Schematic of representation of RIP is given in Fig. 1

The motion equations of above system can be obtained by applying the Euler-Lagrange formulation in Eq. (1): [5]

$$\begin{aligned}
 M(\theta)\ddot{\theta} + C(\theta, \dot{\theta}) + G(\theta) &= \begin{bmatrix} \tau \\ 0 \end{bmatrix} \\
 M(\theta) &= \begin{bmatrix} J_r + mL_r^2 + mL_p^2 \sin^2 \theta_2 & -mL_r L_p \cos \theta_2 \\ -mL_r L_p \cos \theta_2 & J_p + mL_p^2 \end{bmatrix} \\
 C(\theta, \dot{\theta}) &= \begin{bmatrix} B_r + \frac{1}{2}mL_p^2 \dot{\theta}_2 \sin 2\theta_2 & mL_r L_p \dot{\theta}_2 \sin \theta_2 + \frac{1}{2}mL_p^2 \dot{\theta}_1 \sin 2\theta_2 \\ -\frac{1}{2}mL_p^2 \dot{\theta}_1 \sin 2\theta_2 & B_p \end{bmatrix} \\
 G(\theta) &= \begin{bmatrix} 0 \\ mgL_p \sin \theta_2 \end{bmatrix} \tag{1}
 \end{aligned}$$

where $M(\theta)$ is inertia matrix, $C(\theta, \dot{\theta})$ contains centrifugal/Coriolis terms and $G(\theta)$ is the vector of gravitational forces.

Fig. 1 RIP model

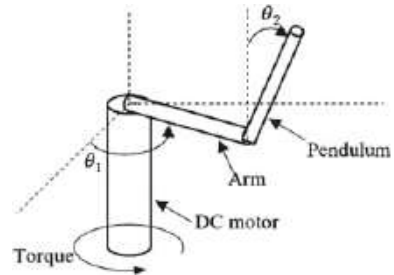


Table 1 Modeling parameters

Symbol	Description
m	A weight for pendulum (kg)
L_p	A length for pendulum (m)
J_p	Inertial moment of pendulum (kg m^2)
L_r	Length of arm (m)
J_r	Inertial moment of arm (kg m^2)
g	Gravitation acceleration (m/s^2)
B_r	Friction of arm ($\text{kg m}^2/\text{s}$)
B_p	Friction of pendulum ($\text{kg m}^2/\text{s}$)

Note that $M(\theta)$ is symmetric matrix and $\det(M(\theta)) > 0$. Therefore, Property 1 is given as follows:

Property 1 $M(\theta)$ is symmetric matrix, positive definition $\forall \theta$:

$$M(\theta) = M(\theta)^T \quad (2)$$

From Eq. (1), we have Property 2 as follows:

Property 2 $N = \dot{M}(\theta) - 2C(\theta, \dot{\theta})$ is skew symmetric matrix

$$\dot{\theta}^T N(\theta, \dot{\theta}) \theta = 0. \quad (3)$$

for any $(n \times 1)$ vector with $N^T = -N$ or $N^T + N = 0$. Property of skew-symmetric matrix is used in establishing the passivity property of RIP.

Modeling parameters of RIP system is provided in Table 1.

2.2 Passivity of RIP System

In this subsection, passivity of RIP is delivered. Storage function of RIP is defined by:

$$E(\theta, \dot{\theta}) = \frac{1}{2} \dot{\theta}^T M(\theta) \dot{\theta} + \frac{1}{2} mg L_2 (1 + \cos \theta_2) \quad (4)$$

where $\frac{1}{2} mg L_2 (1 + \cos \theta_2)$ is potential energy.

Taking derivative of $E(\theta, \dot{\theta})$ with respect to t , we obtain

$$\dot{E}(\theta, \dot{\theta}) = \dot{\theta}^T \tau \quad (5)$$

Integrating both sides of Eq. (5), we obtain (18):

$$\int_0^t \dot{\theta}^T \tau dt = E(t) - E(0) \quad (6)$$

Accordingly, RIP is passive between control input τ and output $\dot{\theta}$. When $\tau = 0$, RIP in Eq. (1) has two operation points, include $(\theta_1, \dot{\theta}_1, \theta_2, \dot{\theta}_2) = (0, 0, \pi, 0)$ corresponding to the downward vertical pendulum position, is stable equilibrium position. For the left position $(\theta_1, \dot{\theta}_1, \theta_2, \dot{\theta}_2) = (0, 0, 0, 0)$ corresponding to the upright vertical pendulum position, is unstable equilibrium position. Total energy of RIP is different between two operation points. At $(\theta_1, \dot{\theta}_1, \theta_2, \dot{\theta}_2) = (0, 0, \pi, 0)$, total energy of RIP is $E(\theta) = 0$, and $E(\theta) = 2\frac{1}{2}mgL_2$ for position $(\theta_1, \dot{\theta}_1, \theta_2, \dot{\theta}_2) = (0, 0, 0, 0)$. The control objective is to stabilize the system at vertical upright position.

3 Controller Implementation

3.1 Feedback Passivation [6]

Stabilizing RIP at $\theta = \theta_d$, we set

$$q = \theta - \theta_d; \quad \dot{q} = \dot{\theta} \quad (7)$$

The dynamics equations (1) become

$$M(\theta)\ddot{q} + C(\theta, \dot{\theta})\dot{q} + G(\theta) = \begin{bmatrix} \tau \\ 0 \end{bmatrix} \quad (8)$$

Feedback passivation control law is defined by

$$\tau = u_1 = G(\theta) - \phi_p(q) + e \quad (9)$$

where $\phi_p(0) = 0$, $q^T \phi_p(q) > \forall q \neq 0$

Based on Eqs. (7), (8) and (9), equations of motion obtained are expressed in Eq. (10)

$$M(\theta)\ddot{q} + C(\theta, \dot{\theta})\dot{q} + \phi_p(q) = \begin{bmatrix} e \\ 0 \end{bmatrix} \quad (10)$$

Storage function E_1 is selected as follows:

$$E_1 = \frac{1}{2}\dot{q}^T M(\theta)\dot{q} + \int_0^q \phi_p(\psi)d\psi \quad (11)$$

Time derivation of selected storage function E_1 in Eq. (11) becomes

$$\dot{E}_1 = \frac{1}{2}\dot{q}^T N \dot{q} - \dot{q}^T \phi_p(q) + \dot{q}^T e + \phi_p^T(q) \dot{q} \leq \dot{q}^T e \quad (12)$$

RIP system is output strictly passive with $y = \dot{q}$. In order to guarantee the system is zero-state observable, the new control input must be zero, i.e. $e = 0$

$$\dot{q}(t) \equiv 0 \Rightarrow \ddot{q}(t) \equiv 0 \Rightarrow \phi_p(q(t)) \equiv 0 \Rightarrow q(t) \equiv 0 \quad (13)$$

The passive control law is defined by

$$e = -\phi_d(\dot{q}) \quad (14)$$

where $\phi_d(0) = 0$, $\dot{q}^T \phi_d(\dot{q}) > 0 \forall \dot{q} \neq 0$

Control law of feedback control based on PBC is given by

$$u_1 = G(\theta) - \phi_p(q) - \phi_d(\dot{q}) \quad (15)$$

We select $\phi_p(q) = k_p q$ and $\phi_d(\dot{q}) = k_d \dot{q}$ where k_p , k_d are symmetric matrices and positive definition, $k_p^T > 0$, $k_d^T > 0$

3.2 Choice of Output [6]

Storage function E_2 is selected as follows:

$$E_2 = \frac{1}{2}(\theta_1^2 + \dot{\theta}_1^2 + \theta_2^2 + \dot{\theta}_2^2) \quad (16)$$

is positive definite on R^n and radially unbounded. The following conditions are satisfied: Lie derivation of selected storage function E_2 in Eq. (16) becomes

$$L_f E_2 \leq 0, \quad L_g E_2 = h^T(x) \quad (17)$$

where $L_f E_2 = (\partial E_2 / \partial x) f(x)$, $L_g E_2 = (\partial E_2 / \partial x) g(x)$.

The defined outputs of the system are chosen as follows:

$$y = h(x) = L_g E_2 \quad (18)$$

Control law is defined as follows:

$$u_2 = -\phi(y), \quad \forall \phi \quad (19)$$

such that $\phi(0) = 0$ and $y\phi(y) > 0$ for $y \neq 0$. Therefore $\phi(y)$ can be expressed by

$$\phi(y) = \kappa L_g E_2 \quad (20)$$

where $\kappa > 0$

Substituting Eq. (20) into Eq. (19) results in

$$u_2 = -\kappa L_g E_2 \quad (21)$$

3.3 Passivity-Based Control Scheme

The control law is defined by combining control law of choice of output and feedback passivation from Eqs. (15) and (21)

$$\tau = u_1 + u_2 = G(\theta) - k_p q - k_d \dot{q} - \kappa L_g E_2 \quad (22)$$

The value of controller parameters are: $k_{p1} = 29.73$, $k_{d1} = 19.74$, $k_{p2} = 99.54$, $k_{d2} = 0.08$, $\kappa = 1.9$, $\theta_{d1} = 0$, $\theta_{d2} = 0$.

4 Validation

In Fig. 2, the performance of RIP with PBC in period time of simulation 5 s. Value of system parameters: $m = 0.027$ (kg), $L_p = 0.328$ (m), $J_p = 0.0046617$ (kg m²), $L_r = 0.205$ (m), $J_r = 0.0019$ (kg m²), $B_r = 0.025$ (kg m²/s), $B_p = 0.0017$ (kg m²/s), $g = 9.81$ (m/s²). Beginning at the original point 0 (rad), after 4 s, angular position of arm θ_1 is back to the 0 rad. Angular position of pendulum θ_2 is back to equilibrium point after 4 s. Angular velocities of arm $\dot{\theta}_1$ and pendulum $\dot{\theta}_2$ are depicted in the second and fourth graph of this figure. Control input τ is also drawn in the last graph.

5 Conclusion

In this manuscript, PBC law for the RIP was designed, tested on simulation. Through simulation, PBC method guarantees close-loop system to be asymptotically stable at upright position when arm is at zero position and other constant value positions. PBC is demonstrated to be successful and effectiveness. Result of controller has been plotted and discussed. The future research for this topic is passivity-based sliding mode control scheme.

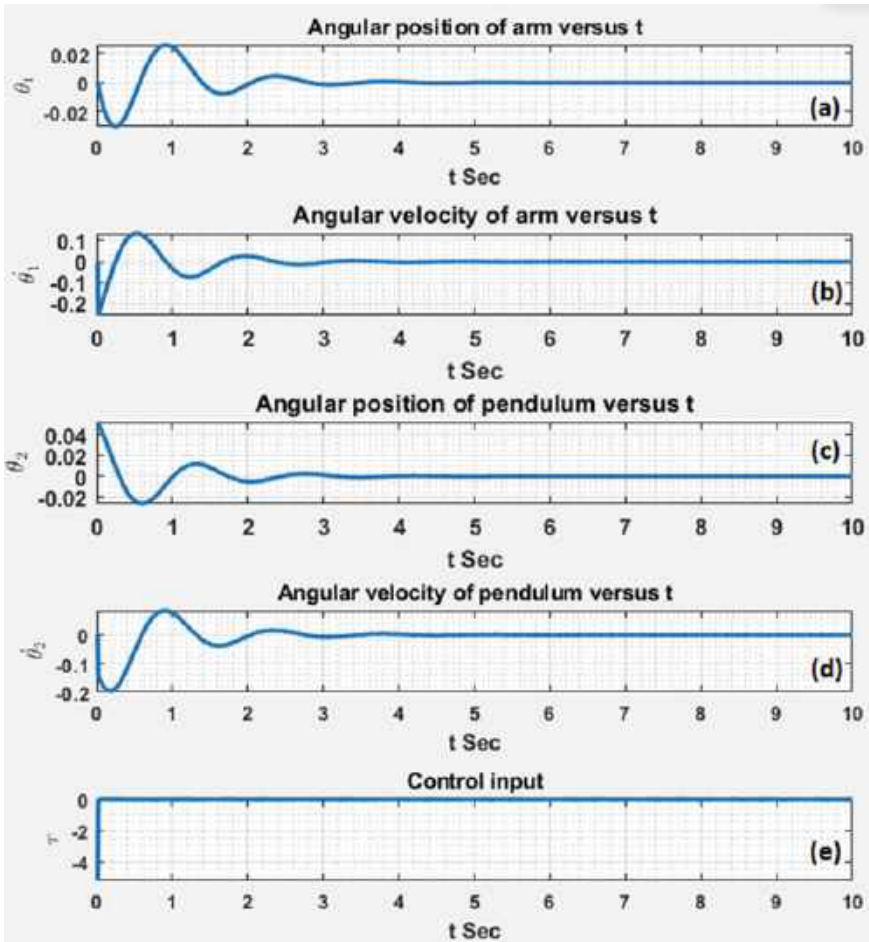


Fig. 2 State responses and control input of RIP with PBC

Acknowledgements We acknowledge the support of time and facilities from Ho Chi Minh City University of Technology (HCMUT), Vietnam National University, Ho Chi Minh City (VNU-HCM) for this study.

References

1. Nghi HV, Nhien DP, Nguyet NTM, Duc NT, Luu NP, Thanh PS, Ba DX (2021) A LQR-based neural-network controller for fast stabilizing rotary inverted pendulum. In: 2021 international conference on system science and engineering (ICSSE), Aug 2021, pp 19-22. <https://doi.org/10.1109/ICSSE52999.2021.9537940>

2. Pandey A, Adhyaru DM (2023) Robust control design for rotary inverted pendulum with unmatched uncertainty. *Int J Dynam Control* 11:1166–1177. <https://doi.org/10.1007/s40435-022-01047-8>
3. Ashenafi NA, Sirichotiyakul W, Satici AC (2022) Robust passivity-based control of underactuated systems via neural approximators and Bayesian inference. *IEEE Control Syst Lett* 6:3457–3462. <https://doi.org/10.1109/LCSYS.2022.3184756>
4. Gutierrez-Oribio D, Stefanou I, Plestan F (2022) Passivity-based control of underactuated mechanical systems with coulomb friction: application to earthquake prevention. *ArXiv preprint arXiv:2207.07181*. <https://doi.org/10.48550/arXiv.2207.07181>
5. Yang X, Zheng X (2018) Swing-up and stabilization control design for an underactuated rotary inverted pendulum system: theory and experiments. *IEEE Trans Ind Electron* 65(9):7229–7238. <https://doi.org/10.1109/TIE.2018.2793214>
6. Ortega R, Perez JAL, Nicklasson PJ, Sira-Ramirez HJ (2013) *Passivity-based control of Euler-Lagrange systems: mechanical, electrical and electromechanical applications*. Springer Science and Business Media

Comparative Analysis of Data-Driven Models for DC Motors with Varying Payloads



Helen Shin Huey Wee and Nur Syazreen Ahmad

Abstract DC motors are widely used in various industrial applications, and accurately predicting their performance under different payload conditions is crucial for optimal control and efficient system design. In recent years, data-driven modeling techniques have gained significant attention as effective tools for capturing the complex dynamics of DC motors. This study conducts a comparative assessment of diverse data-driven models applied to DC motors operating under different load conditions. A number of model structures have been tested which include continuous-time transfer function, discrete-time transfer function, and Auto-Regressive with eXogenous input (ARX). Results reveal that the ARX and second-order continuous time transfer function models outperform the rest with accuracies of at least 94%.

Keywords DC motors · Payloads · System identification · Transfer function model · ARX

1 Introduction

DC motors are widely used in various industries for their simplicity, reliability, and controllability [1]. These motors are employed in numerous applications, ranging from robotics [2–6] and industrial automation to electric vehicles and aerospace systems [7]. Numerous factors can impact the operational efficiency of a DC motor, with one such factor being the payload or load that it is tasked with transporting [8]. Understanding the behavior of DC motors under varying payload conditions is crucial for optimizing their operation and ensuring efficient performance.

H. S. H. Wee · N. S. Ahmad (✉)

School of Electrical and Electronic Engineering, Universiti Sains Malaysia, 14300 Nibong TebalPenang, Malaysia

e-mail: syazreen@usm.my

H. S. H. Wee

e-mail: helenwee@student.usm.my

In recent years, there has been a substantial surge in interest towards data-driven models, particularly in the realm of analyzing and predicting the behavior of intricate systems, such as DC motors [9]. These models utilize data collected from sensors and other sources to build mathematical representations of the system's dynamics [10, 11]. By training on historical data, data-driven models can capture the relationships between input variables, such as motor voltage and current, and output variables, such as motor speed and torque [12]. Consequently, these models can provide valuable insights into the performance of DC motors under different operating conditions.

Several studies have investigated the comparative analysis of data-driven models for DC motors with varying payloads. In [13], the Simscape electronic systems in MATLAB/SIMULINK is utilized to simulate the behavior of a real DC motor and obtain input voltage and output speed data. In their work, a nonlinear autoregressive with exogenous input (NARX) neural network is proposed to model the DC motor. While the simulation results demonstrate the effectiveness of the proposed technique, constructing a suitable controller for such a model can be challenging due to the presence of nonlinearity [14–16].

In this study, the focus is on modeling DC motors where the data is obtained from real-time experiments. A comparative analysis is conducted to evaluate various models for DC motors under different payload conditions. The models considered include continuous-time transfer function, discrete-time transfer function, and Auto-Regressive with eXogenous input (ARX). Results indicate that the ARX model and the second-order continuous-time transfer function model outperform the other models, achieving accuracies of at least 94%.

2 Methodology

2.1 Model Structure

The ARX configuration demonstrates how the input effects $u(t)$ influence the output $y(t)$. The representation of the ARX model is as follows:

$$y(t) = a_1y(t-1) - \dots - a_{n_a}y(t-n_a) + b_1u(t-1-n_k) + \dots + b_{n_b}u(t-n_b-n_k) + e(t) \quad (1)$$

where $e(t)$ denotes the assumed Gaussian noise, while a_{n_a} and b_{n_b} represent the parameters of the model. Here, n_a and n_b specify the degrees of the polynomials for the output $A(q)$ and the input $B(q)$ respectively. Additionally, n_k signifies the time delay between $u(t)$ and $y(t)$. The polynomial portrayal of Eq. (1) can be formulated as follows:

$$A(q)y(t) = B(q)u(t-n_k) + e(t) \quad (2)$$

where $A(q)$ and $B(q)$ are given by:

$$A(q) = 1 + a_1q^{-1} + \dots + a_{n_a}q^{-n_a}, \quad B(q) = b_1q^{-1-n_k} + \dots + b_{n_b}q^{-n_b-n_k} \quad (3)$$

The term q^{-1} denotes the delay operator, i.e. $u(t-1) = q^{-1}u(t)$. In continuous time setting, the first-order and second-order transfer functions that are used to model the DC motor are described as follows:

$$G_{1st-order}(s) = \frac{b_0}{s + a_0}, \quad G_{2nd-order}(s) = \frac{b_0}{s^2 + a_1s + a_0} \quad (4)$$

where $a_0, a_1, b_0 > 0$. In discrete-time setting, the structures can be simply written as

$$G_{1st-order}(z) = \frac{b_0 + b_1z^{-1}}{1 + a_1z^{-1}}, \quad G_{2nd-order}(z) = \frac{b_0 + b_1z^{-1} + b_2z^{-2}}{1 + a_1z^{-1} + a_2z^{-2}} \quad (5)$$

2.2 Design of Experiment

To collect data, a DC motor is connected to a mobile robot platform that carries a load. The load is measured by a load sensor, and its range is from 0.001 to 1.1 kg. The motor's PWM is adjusted for each trial, and the motor speed and load measurements are transmitted wirelessly to a PC through Bluetooth for later analysis. After collecting the data, it is partitioned into two subsets: an estimation set and a validation set. The estimation set is employed for the optimization of model parameters, whereas the validation set is used to objectively assess the practical performance. Before the estimation stage, a moving average filtering technique is employed. This filtering process reduced the impact of high-frequency measurement noise, allowing the identification process to focus on the desired frequency range.

3 Results and Discussion

The percentage fitness to estimation data was obtained from MATLAB simulation. Table 1 presents the percentage fitness results for first-order and second-order models with both discrete and continuous transfer functions, under various load conditions. Similarly, Table 2 displays the results for the ARX model. The load conditions include four different load weights: 0.0001 kg (L1), 0.4 kg (L2), 0.8 kg (L3), and 1.1 kg (L4).

Upon observation, it is evident that the ARX model exhibits superior performance in terms of percentage fitness to the estimation data. Specifically, for load weights L1 (0.0001 kg) and L2 (0.4 kg), the best fitness is achieved with an order of [2 1

Table 1 Percentage fit to estimation data of first and second order transfer function

Percentage fitness to estimation data (%)				
Load weight (kg)	Discrete transfer function		Continuous transfer function	
	First order	Second order	First order	Second order
L1 (0.0001)	83.01	66.6	82.28	95.69
L2 (0.4)	64.96	65.98	65.31	95.42
L3 (0.8)	69.82	70.17	69.93	79.81
L4 (1.1)	70.71	62.09	69.25	94.32

Table 2 Percentage fit to estimation data of ARX model

Percentage fitness to estimation data (%)				
ARX model order	Load weight (kg)			
	L1 (0.0001)	L2 (0.4)	L3 (0.8)	L4 (1.1)
[1 0 0]	92.17	91.35	91.03	91.41
[1 0 1]	92.17	91.35	97.89	91.41
[1 1 0]	94.21	92.56	98.59	92.74
[1 1 1]	92.82	91.79	98.63	91.77
[2 0 0]	98.31	97.82	98.01	97.73
[2 0 1]	98.31	97.82	98.01	97.73
[2 1 0]	99.59	97.84	98.02	97.73
[2 1 1]	98.34	97.82	98.01	97.75

0], reaching 99.59% and 97.84%, respectively. For L3 (0.8 kg), the best fitness is achieved with an order of [1 1 0], resulting in a fitness of 98.59%, while for L4 (1.1 kg), the best fitness is achieved with an order of [2 1 1], which amounts to 97.75%. In terms of the transfer function model, the second-order continuous-time transfer function model demonstrates good fitness above 90% for L1, L2, and L3, with fitness values of 95.69%, 95.42%, and 94.32%, respectively, except for L3, which achieves a fitness of 79.81%. On the other hand, the ARX model with an order of [1 0 0] exhibits the poorest fitness to the estimation data across all four load conditions. Figure 1 illustrates the model output of the best fit in comparison to the validation data under various load conditions.

The validation data is represented by a black line, while the colored line represents the best fit model for each load condition. For load weights L1 (0.001 kg) and L2 (0.4 kg), the ARX model with an order of [2 1 0] exhibits the closest match to the validation data. Similarly, for L3 (0.8 kg), the best fit model in comparison to the validation data is the ARX model with an order of [1 1 0]. As for L4 (1.1 kg), the best fit is achieved with the ARX model at order [2 1 1]. Equations (10), (11), (12), and (13) show the model parameters that yield the highest percentage fitness to the estimation data for the different load weights of L1, L2, L3, and L4, respectively.

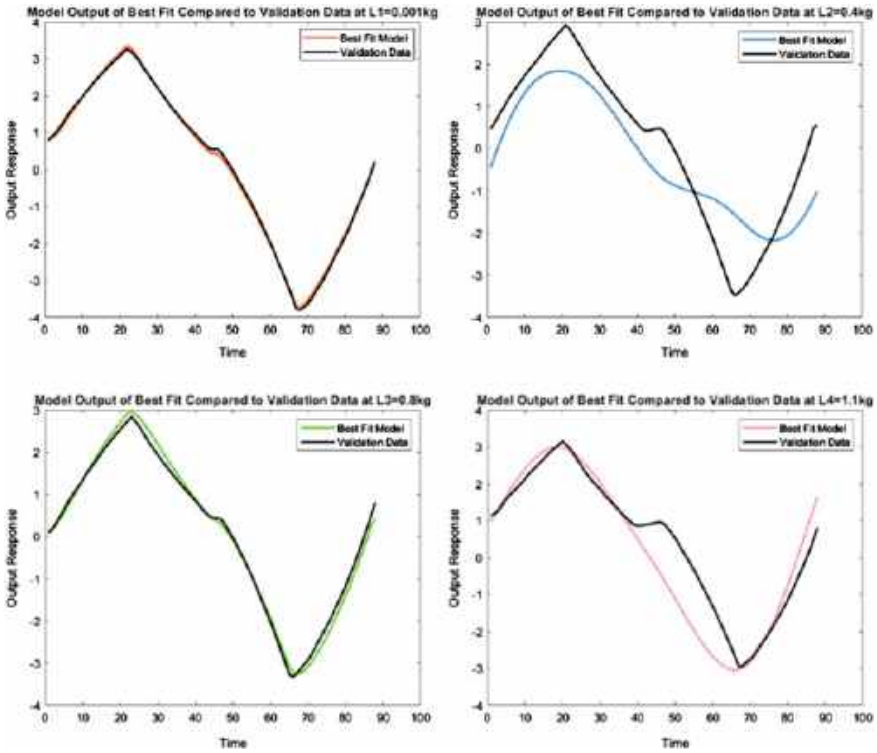


Fig. 1 Model output of best fit compared to validation data at various load conditions

The ARX model demonstrated effectiveness in estimation due to its ability to solve linear regression equations in analytical form, thereby minimizing the loss function. On the other hand, the transfer function model, which was simulated in both continuous time and discrete time for first and second-order systems, provided a mathematical representation of the overall system and its components. However, the results indicated that the transfer function model did not align well with the estimation data compared to the ARX model. The advantage of the transfer function model lies in transforming complex differential and integral equations into simpler algebraic equations. However, it is limited to linear systems, which represents a fundamental drawback of transfer functions.

$$(1 - 0.3979z^{-1} + 0.02748z^{-2})y(t) = 0.1041u(t) + \frac{1}{(1 - z^{-1})}e(t) \quad (6)$$

$$(1 - 1.93z^{-1} + 0.9448z^{-2})y(t) = 0.001475u(t) + e(t) \quad (7)$$

$$(1 - 0.6843z^{-1})y(t) = 0.0573u(t) + \frac{1}{(1 - z^{-1})}e(t) \quad (8)$$

$$(1 - 1.997z^{-1} + 0.9939z^{-2})y(t) = (-0.001837z^{-1})u(t) + e(t)$$

4 Conclusion

In conclusion, the ARX and second-order continuous time transfer function structures can accurately describe the DC motor model when subject to varying payloads. These findings emphasize the potential of data-driven modeling techniques in accurately characterizing the behavior of DC motors and can contribute to optimal control strategies and efficient system design. Further research and development in this area could lead to improved performance prediction and control of DC motors in industrial applications.

References

1. Lu Y (2022) DC motor control technology based on multisensor information fusion. *Comput Intell Neurosci* 2022
2. Ahmad NS, Teo JH, Goh P (2022) Gaussian process for a single-channel EEG decoder with inconspicuous stimuli and eyeblinks. *Comput Mater Continua* 73(1):611–628
3. Teo JH, Ahmad NS, Goh P (2022) Visual stimuli-based dynamic commands with intelligent control for reactive BCI applications. *IEEE Sens J* 22(2):1435–1448
4. Teo JH, Loganathan A, Goh P, Ahmad NS (2020) Autonomous mobile robot navigation via RFID signal strength sensing. *Int J Mech Eng Robot Res* 9(8):1140–1144
5. Bakar SJA, J-Shenn K, Goh P, Ahmad NS (2023) Development of magnetic levitation system with position and orientation control. *Int J Reconfigurable Embedded Syst* 12(2):287–296
6. Bin Lu L, Ahmad NS, Goh P (2023) Self-balancing robot: modeling and comparative analysis between PID and linear quadratic regulator. *Int J Reconfigurable Embedded Syst* 12(3):351–359
7. Loganathan A, Ahmad NS, Goh P (2019) Self-adaptive filtering approach for improved indoor localization of a mobile node with ZigBee-based RSSI and Odometry. *Sensors* 19(21):4748
8. Loganathan A, Ahmad NS (2023) A systematic review on recent advances in autonomous mobile robot navigation. *Eng Sci Technol Int J* 40:101342
9. Saucedo-Dorantes JJ, Delgado-Prieto M, Osornio-Rios RA, Romero-Troncoso RDJ (2020) Industrial data-driven monitoring based on incremental learning applied to the detection of novel faults. *IEEE Trans Industr Inform* 16(9)
10. Arrouch I, Ahmad NS, Goh P, Mohamad-Saleh J (2022) Close proximity time-to-collision prediction for autonomous robot navigation: an exponential GPR approach. *Alex Eng J* 61(12):11171–11183
11. Arrouch I, Mohamad-Saleh J, Goh P, Ahmad NS (2022) A comparative study of artificial neural network approach for autonomous robot's TTC prediction. *Int J Mech Eng Robot Res* 11(5):345–350
12. Ahmad NS (2020) Robust H_{∞} -fuzzy logic control for enhanced tracking performance of a wheeled mobile robot in the presence of uncertain nonlinear perturbations. *Sensors* 20(13):7673
13. Naung Y, Schagin A, Oo HL, Ye KZ, Khaing ZM (2018) Implementation of data driven control system of DC motor by using system identification process. In: *Proceedings of the 2018 IEEE conference of Russian young researchers in electrical and electronic engineering, ElConRus 2018*

14. Chan SY, Ahmad NS, Ismail W (2017) Anti-windup compensator for improved tracking performance of differential drive mobile robot. In: Proceedings of 2017 IEEE international symposium on systems engineering, ISSE
15. Syed Mubarak Ali SAA, Ahmad NS, Goh P (2019) Flex sensor compensator via hammerstein-wiener modeling approach for improved dynamic goniometry and constrained control of a bionic hand. *Sensors (Switzerland)* 19(18)
16. Ahmad NS, Carrasco J, Heath WP (2012) Revisited Jury-Lee criterion for multivariable discrete-time Lur'e systems: convex LMI search. In: 2012 IEEE 51st IEEE conference on decision and control (CDC), pp 2268–2273

Development and Control of Underactuated Parallel Rotary Double Inverted Pendulum System



Minh-Tai Vo, Minh-Duy Vo, Van-Dat Nguyen, Van-Dong-Hai Nguyen, Minh-Duc Tran, Hoai-Nghia Duong, and Thanh T. Tran

Abstract The Parallel Rotational Double Inverted Pendulum System (PRDIP) is a new model in the pendulum family. And there are not many articles that propose and research this model at the moment. One of the interesting features of the PRDIP system is the presence of an underactuated link, which makes it challenging to control the system. This type of system has received increasing attention from researchers in recent years due to its applicability in many areas, such as robotics and control. This model is built by adding one more link parallel to the former link, which is placed at the ends of the arm. Because of the complexity of PRDIP, this model has not been widely proposed and investigated in both simulation and experiments. Therefore, this research is going to concentrate on investigating the PRDIP system and deploying Linear Quadratic Regular (LQR) algorithm in simulation and experiments for this system. The LQR controller is responsible for controlling and maintaining an arm and two pendulums, which work around the work points. Particularly, the working points

M.-T. Vo (✉) · T. T. Tran
RMIT University, Ho Chi Minh City, Vietnam
e-mail: tai.vo3@rmit.edu.vn

T. T. Tran
e-mail: thanh.tran37@rmit.edu.vn

M.-D. Vo · V.-D. Nguyen · V.-D.-H. Nguyen
Ho Chi Minh City University of Technology and Education, Ho Chi Minh City, Vietnam
e-mail: 19151108@student.hcmute.edu.vn

V.-D. Nguyen
e-mail: 19151114@student.hcmute.edu.vn

V.-D.-H. Nguyen
e-mail: hainvd@hcmute.edu.vn

M.-D. Tran
Ho Chi Minh City University of Technology, VNU-HCM, Ho Chi Minh City, Vietnam
e-mail: tmduc.sdh222@hcmut.edu.vn

H.-N. Duong
Eastern International University, Binh Duong Province, Vietnam
e-mail: nghia.duong@eiu.edu.vn

of the two pendulums are totally different. The long pendulum will be controlled around the upward upright position, the short pendulum will be controlled around the downward upright position, and the arm will be controlled in a particular position. The result of this research shows that the LQR algorithm has been successfully deployed for the PRDIP system.

Keywords Parallel rotational double inverted pendulum · PRDIP · Hardware · MATLAB® · Solidworks

1 Introduction

The Parallel Rotational Double Inverted Pendulum (PRDIP) is proposed, which is not a popular model in teaching and research due to its flexibility. PRDIP is a type of system known as a single input-multiple output and is characterized by nonlinearity and instability. This model is considered to be the most flexible in the pendulum family. As a result, there are not many articles available on PRDIP currently. To construct PRDIP, an additional parallel link is attached to the end of the arm, and both pendulums are opposite each other. This approach contrasts with building the Serial Rotational Double Inverted Pendulum (SRDIP), which is implemented by adding another serial link to an existing link [1–3]. Due to the difference in length between the two parallel pendulums in PRDIP, the rotation of the arm has a varying effect on each pendulum. The shorter pendulum is particularly reactive and prone to dropping quickly. As a result, controlling and simulating this system is a big challenge for every researcher. So far, only a few studies have been conducted on PRDIP in order to identify its dynamic equation and investigate its stability [4, 5]. And many researchers just investigate this model by examining control theory algorithms in simulations. Some algorithms are implemented on PRDIP such as, LQR controller [6, 7], etc. In this research, the authors will focus on researching both simulation and experiment. The LQR controller is used to deploy on this system, and this controller is responsible for controlling and maintaining an arm and two pendulums, which work around the working point. The longer pendulum will be controlled around the upward upright position, while the shorter pendulum will be controlled around the downward upright position, and the arm will be controlled in a particular position.

2 Mathematical Equation of Rotary Double Parallel Inverted Pendulum

2.1 Mathematical Equation

The physical structure of PRDIP consists of a metal frame, an arm of length L (m), and two pendulums of length $lg1$ (m) and $lg2$ (m). Mass of pendulum one is denoted as m_1 (kg), and mass of pendulum two is denoted as m_2 (kg). Inertia of arm (kg m^2), pendulums one (kg m^2), and two (kg m^2) are denoted as J_0, J_1, J_2 , respectively. Symbol g represents for gravity constant (m/s^2). The two pendulums are connected to the arm via two encoders. This system is operated by DC servo motor through the arm. The operating direction of the system is shown in Fig. 1. In, the angle of arm is denoted by ϕ (rad) and the angle of two pendulums ($lg1, lg2$) are respectively denoted by θ_1, θ_2 (rad) (Table 1; Fig. 2).

As stated in reference [1], the dynamical equation of this system is computed and written in the form of an equation of state as follows:

$$\begin{bmatrix} Z_1 & X_1 & V_1 \\ Z_2 & X_2 & V_2 \\ Z_3 & X_3 & V_3 \end{bmatrix} \begin{bmatrix} \ddot{\phi} \\ \ddot{\theta}_1 \\ \ddot{\theta}_{r2} \end{bmatrix} + \begin{bmatrix} K_1 \\ K_2 \\ K_3 \end{bmatrix} = \frac{K_t}{R_m} \begin{bmatrix} V_{in} \\ 0 \\ 0 \end{bmatrix} \tag{1}$$

Fig. 1 PRDIP model

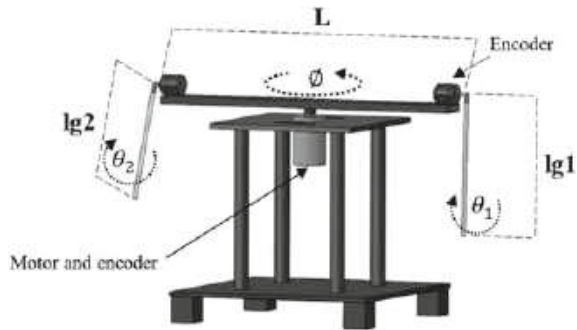


Table 1 The parameters of system

Parameter	Pendulum 1	Pendulum 2	Arm
m_i	0.059 (Kg)	0.127 (Kg)	na
l_{gi}	0.038 (m)	0.082 (m)	na
L	na	na	0.51 (m)
J_i	0.00267(kg m^2)	0.00137(kg m^2)	na
J_0	na	na	0.75(kg m^2)
C_i	1.526×10^{-4}	4.693×10^{-4}	na
C_0	na	na	4.978

Fig. 2 The real-time experimental setup



where

$$Z_1 = J_0 + m_1 l_{g1}^2 \sin^2 \theta_1 + m_1 L^2 + m_2 l_{g2}^2 \sin^2 \theta_2 + m_2 L^2 \quad (2)$$

$$X_1 = -m_1 L l_{g1} \cos \theta_1; \quad V_1 = -m_2 L l_{g2} \cos \theta_2 \quad (3)$$

$$K_1 = m_1 l_{g1}^2 \dot{\theta}_1 \dot{\phi} \sin(2\theta_1) + m_1 l_{g1} L \dot{\theta}_1^2 \sin \theta_1 + c_0 \dot{\phi} \\ + m_2 l_{g2}^2 \dot{\theta}_2 \dot{\phi} \sin(2\theta_2) + (c_0 + \frac{K_t K_b}{R_m}) \dot{\phi} \quad (4)$$

$$Z_2 = -m_1 L l_{g1} \cos \theta_1; \quad X_2 = J_1 + m_1 l_{g1}^2; \quad V_2 = 0 \quad (5)$$

$$K_2 = -m_1 l_{g1}^2 \dot{\phi}^2 \sin \theta_1 \cos \theta_1 - m_1 g l_{g1} \sin \theta_1 + c_1 \dot{\theta}_1 \quad (6)$$

$$Z_3 = -m_2 L l_{g2} \cos \theta_2; \quad X_3 = 0; \quad V_3 = J_2 + m_2 l_{g2}^2 \quad (7)$$

$$K_3 = -m_2 l_{g2}^2 \dot{\phi}^2 \sin \theta_2 \cos \theta_2 - m_2 g l_{g2} \sin \theta_2 + c_2 \dot{\theta}_2 \quad (8)$$

2.2 Linear Model

This system has to be linearized around the operation point, which is given below.

$$\phi \approx 0; \theta_1 \approx 0; \theta_2 \approx \pi; \dot{\phi} \approx 0; \dot{\theta}_1 \approx 0; \dot{\theta}_2 \approx 0 \quad (9)$$

$$x = [\phi \ \theta_1 \ \theta_2 \ \dot{\phi} \ \dot{\theta}_1 \ \dot{\theta}_2]^T \quad (10)$$

The linearized state equations of the PRDIP are as follows:

$$\begin{cases} \dot{x} = Ax + Bu \\ y = Cx \end{cases} \quad (11)$$

And matrices A and B are calculated and given as follows:

$$A = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0.1008 & 0.1108 & -3.7526 & -0.0002 & 0 \\ 0 & 20.4029 & 0.1169 & -3.9596 & -0.0424 & 0 \\ 0 & -0.1452 & -27.8634 & 5.4047 & 0.0003 & 0 \end{bmatrix}; B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0.0124 \\ 0.0130 \\ -0.0178 \end{bmatrix} \quad (12)$$

3 Results and Discussions

3.1 Simulation Results

Results show in Fig. 3 obtained by simulation on MATLAB/Simulink with feedback gain matrix as follows $K = 10^5[-0.1 \ 1.6478 \ 0.0173 \ -0.1678 \ 0.2906 \ 0.0248]$. The control signal is given by $u(t) = -Kx(t)$.

As of the start of control of this system, the first pendulum tends to oscillate less than the second pendulum. The first pendulum takes about 10 s to stabilize and maintain around 0 rad. In contrast, the second pendulum takes about 12 s to stabilize around 3.14 rad. In the processing control of the PRDIP system, the swing arm will oscillate to stabilize the two pendulums at the work point. After successfully controlling the two pendulums, the arm stabilizes around the work point, which is 0.55 rad. So, we can observe that system totally stabilize after 12 s from the start of control.

3.2 Experimental Results

To evaluate the control performance with real-time PRDIP system, LQR control scheme is tested to stabilize system at the same operation points with simulation. Figure 4 shows results obtained by experiments with feedback gain matrix as follows:

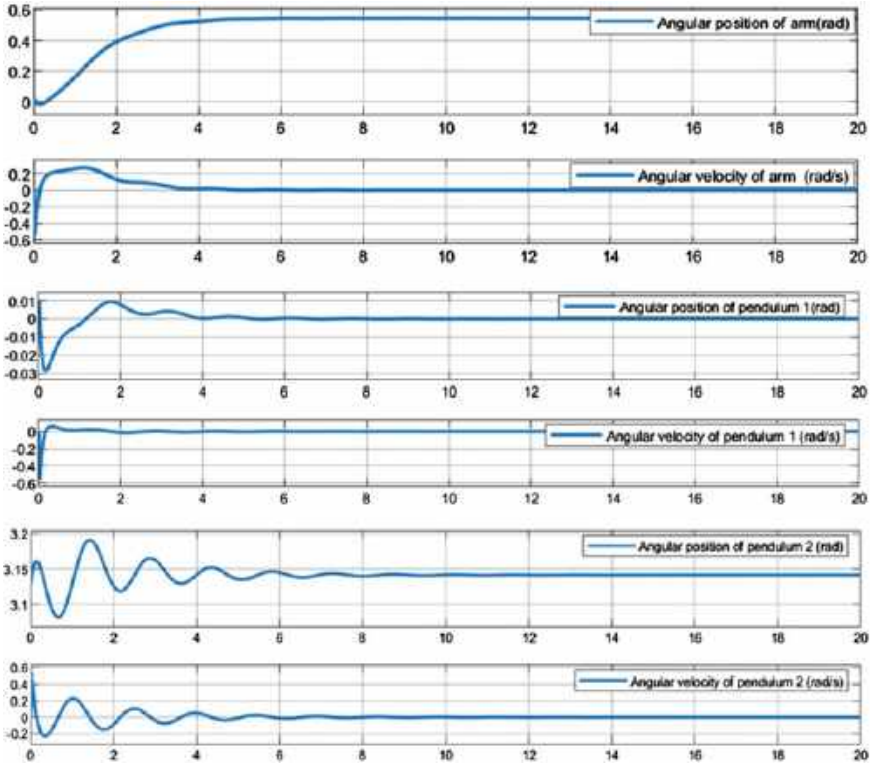


Fig. 3 The response of the PRDIP system

$$K = 10^4[-0.3162 \ 9.2841 \ 0.0017 \ -1.1494 \ 2.0510 \ 0.0039].$$

The system is in an uncontrollable state, two pendulums are in a downward upright position, and the arm is at 0 rad. When the authors supply power and put the first pendulum upright, this system will operate with the LQR controller. From about 0.8–3.1 s, this system will be controlled around the work points. When the second pendulum is suddenly impacted, it will oscillate with an amplitude of 2.3 rad. Simultaneously, the first pendulum and an arm are also affected and deviated from the work point. At this point, the controller will control and help maintain the system so that it does not destabilize. And after about 0.5 s, the system will be stable again. A video of the experiments can be watched at: <https://www.youtube.com/watch?v=TXyixHshDsQ>.

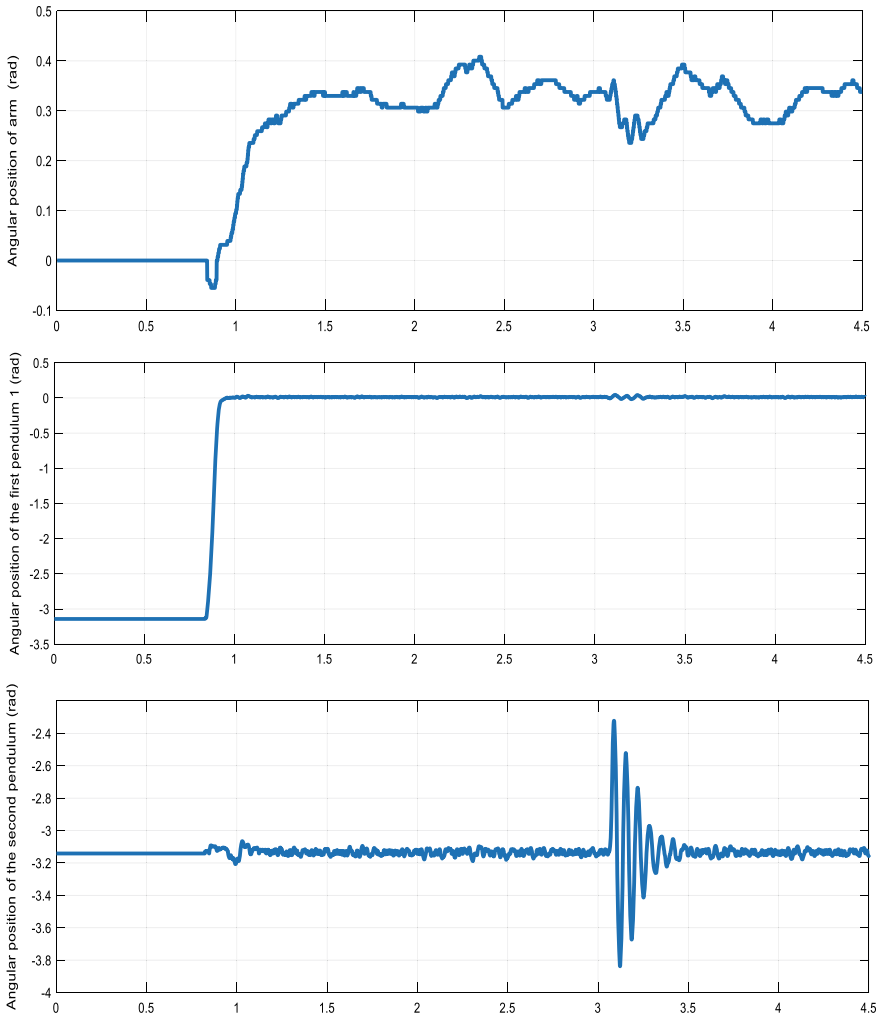


Fig. 4 The response of the real time experimental setup

4 Conclusion

In this research, PRDIP system is proposed and validated in both simulation and experiment. PRDIP is an uncommon model and has not been studied much in control engineering. In order to have a breakthrough development, the authors have constructed physical PRDIP. The authors implemented LQR controller for the PRDIP to evaluate the controllability of this controller. The results show that the LQR controller can successfully control PRDIP in both simulation and experiment.

In the future, the authors focus on developing the research by using metaheuristic approaches to find matrix Q to improve quality control.

Conflict of Interest The authors have no conflicts of interest to declare.

References

1. Sanjeeva SD, Parnichkun M (2021) Control of rotary double inverted pendulum. *J Control Decis* 89–101
2. Ananthan SS (2022) Advanced control strategies for rotary double inverted pendulum. *Int J Eng Manag Res*
3. Singh S, Swarup A (2021) Control of rotary double inverted pendulum using. In: 2021 International conference on intelligent technologies (CONIT), Karnataka
4. Aribowo AG, Nazaruddin YY, Joelianto E, Sutarto HY (2007) Stabilization of rotary double inverted pendulum using robust gain-scheduling control. In: SICE annual conference 2007, Kagawa
5. Zhao X, Zhang Z, Huang J (2016) Energy-based swing up control of rotary parallel inverted pendulum. In: 2016 12th World Congress on Intelligent Control and Automation (WCICA), Guilin, China. <https://doi.org/10.1109/WCICA.2016.7578567>
6. Dinh KH (2013) Using LQR algorithm to control circular two stage parallel inverted pendulum system. *Bulletin of College of Engineering National Ilan University, Taiwan*
7. Sugie T, Okada M (1993) H_∞ control of parallel inverted pendulum systems. *J Soc Instrum Control Eng* 6:543–551

A Study of the Influence of Steel Brushes in Rail Surface Magnetic Flux Leakage Detection Using Finite Elements Simulation



Gong Wendong, Muhammad Firdaus Akbar, Mimi Faisyalini Ramli, and Ghassan Nihad Jawad

Abstract The structure of the magnetic excitation circuit can significantly impact the Magnetic Flux Leakage (MFL) signal of rail inspection. This research proposes a unique approach for analyzing the influence of steel brushes in MFL detection equipment based on finite element simulation results. The magnetic circuit equation is established to investigate the magnetic resistance. Next, the MFL detection with steel brushes of different thicknesses and without steel brushes is compared and analyzed in the finite simulation model. The finite element situation results show that the MFL signal does not change linearly as the parameters are changed. This method demonstrates that the structure of the detection equipment with steel brush can generate the MFL signals with a higher signal-to-noise ratio.

Keywords MFL · Steel brush · Finite element simulation · Rail inspection

1 Introduction

Rails bear high mechanical loads from the wheel during the high-speed running process. Meanwhile, rails are usually exposed directly to the natural environment. Therefore, there are kinds of rail surface defects, including stripping, corrosion, abrasion, cracks, and so on [1]. MFL testing is based on the fact that when a magnetic

G. Wendong · M. F. Akbar (✉)
School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Engineering Campus,
14300 Nibong Tebal, Penang, Malaysia
e-mail: firdaus.akbar@usm.my

M. F. Ramli
Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, 86400
Batu Pahat, Johor, Malaysia

G. N. Jawad
Department of Electronics and Communication Engineering, University of Baghdad,
Baghdad 10071, Iraq

field is applied to a ferromagnetic material, the defect on the surface will cause a significant change in magnetic permeability in the test material. Therefore, the field will leak out of the material into the air beside the defects. Hence, the flux leakage can be measured by a magnetic field sensor and used to estimate the dimensions of the defect. MFL detection method can overcome the influence of the external environment, and it has high accuracy in high-speed detection. Therefore, the MFL detection technique has received wide attention for rail surface inspection, especially for cracks [2–4].

The finite element calculation is based on the vector partial differential equation of the electromagnetic field, combined with the material properties of the magnetization device and the object to be detected. Then the finite element model of the leakage field can be established, and the numerical solution of the leakage field is derived by meshing the cells and solving the nonlinear equation [5]. Kinds of finite element simulation software appeared with the development of computer technology, such as COMSOL multiphysics, Ansis Maxwell, and CST studio. In previous research, finite element simulations have been done for different shapes of defects, and the effects of different defects on the magnetic leakage field distribution were analysed [6]. Meanwhile, the magnetic leakage field distribution principle was analyzed when the size of defects, such as length, width, depth, and angle, changed [7, 8]. However, the magnetic leakage signal from the cracks is usually weak, so the magnetic excitation and detection parameters need to be optimized to get a more significant signal. In this work, the magnetic circuit equation was established to analyze magnetic resistance, and the thickness parameters of the steel brushes were compared and analyzed by finite simulation software.

The paper is organized as follows: Sect. 2 established the magnetic circuit equation and the finite element simulation model to analyze magnetic resistance. Section 3 calculated and discussed the simulation result by the COMSOL multiphysics software. The simulation results show that optimizing the parameters can enhance the MFL signals in a higher signal-to-noise ratio.

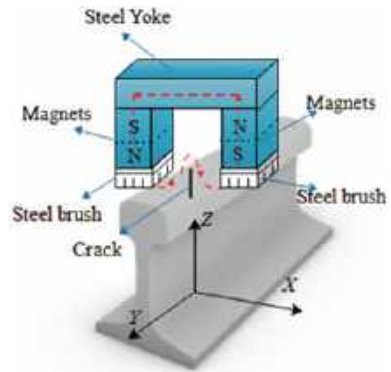
2 Theoretical Approach

2.1 Structure of Magnetization Device

The permanent magnet has been widely used in the MFL detection technique because it has a simple structure, is lightweight, and has strong residual magnetic flux density. Meanwhile, providing a continuous power supply in the field inspection environment of steel rails is difficult, and the permanent magnet magnetization method can eliminate the power supply constraint. Therefore, the permanent magnets were chosen as the excitation source.

The air gap between the permanent magnet and the measured rail affects the efficiency of magnetization. In order to reduce the magnetic resistance, one side of

Fig. 1 The structure of the magnetization device

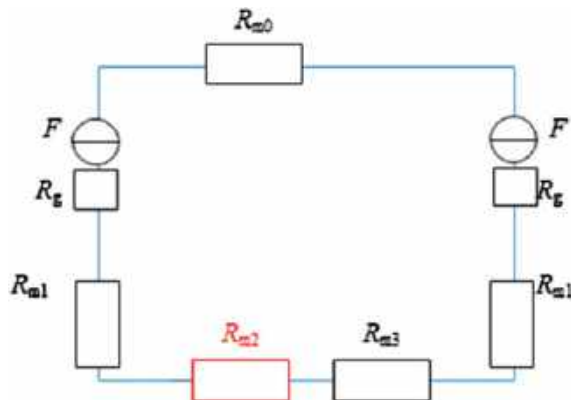


the steel brush was fixed to the permanent magnet, and the other side was always in close contact with the measured rail during the testing process. The two permanent magnets with opposite excitation directions were finally connected with the yoke to form an excitation circuit finally. The structure of the magnetization device is shown in Fig. 1.

2.2 Magnetic Circuit Analysis

From Fig. 1, it can be seen that the circuit of the magnetization device includes the yoke magnetic resistance R_{m0} , the steel brush magnetic resistance R_{m1} , the air gap magnetic resistance R_{m2} , the rail magnetic resistance R_{m3} , the permanent magnet internal resistance R_g , and the magnetic excitation source. The circuit can be shown in Fig. 2.

Fig. 2 The magnetic circuit



There are several geometric parameters of the magnetization device which can affect the magnetic resistance of the magnetic circuit, including the length and width of the cross-section, the thickness of the steel brush, the thickness of the permanent magnet, the length of the yoke, and the thickness of the yoke. However, the air has the lowest permeability. Thus, the air gap between the rail surface and detection equipment will significantly impact the total Magnetic resistance.

2.3 Finite Element Simulation Model Establishment

The three-dimensional (3D) model should be established firstly in the COMSOL multiphysics software for the finite element simulation, including the permanent magnets, the yoke, the steel brushes, the rail, and the cracks. Then the material properties need to be configured in COMSOL multiphysics software. The permanent magnets were configured to be NdFeB (neodymium-iron-boron) magnets because NdFeB has higher residual flux density than regular magnets, and the model number is N55. The yoke material was configured to be permalloy (a kind of iron-nickel alloy) because it has a high permeability to weak magnetic fields, and the model number is 1J85. The material of the brushes is confirmed to be carbon steel. Subsequently, the model meshed for calculation. The boundary of the crack was meshed in the triangular subdivision. Meanwhile, the cell size of the meshing model should be fine to ensure the accuracy of the calculation. The meshing simulation model is shown in Fig. 3.

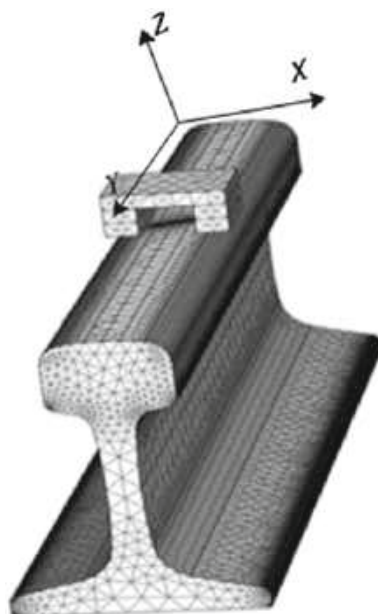
Define the origin of the cartesian coordinate system as the center of the crack on the rail surface, the X-axis direction as the direction along the rail extension, the Y-axis direction as the direction perpendicular to the rail extension, and the Z-axis direction as the direction perpendicular to the top surface of the rail. The spatial magnetic flux leakage field can be decomposed into B_x (along the X-axis direction), B_y (along the Y-axis direction), and B_z (along the Z-axis direction).

3 Results and Discussions

Steel brushes are used to connect the measured rail and the excitation device, which can effectively reduce the air magnetic resistance. When the crack is 20 mm in length, 1 mm in width, 3 mm in depth, and 60° in angle, the finite element simulation results with and without steel brushes can be shown in Fig. 4.

As shown in Fig. 4, if there is no steel brush connecting the excitation device and the tested rail, the magnetic resistance will increase significantly. Therefore, the MFL signal will become weaker and difficult to detect. Thus, the steel brush should be kept in close contact with the rail surface during the test process.

Fig. 3 The meshing simulation model



When the thickness of the steel brush is 5 mm, 10 mm, 15 mm, 20 mm, and 25 mm, respectively, the MFL signal from the finite element simulation model can be shown in Fig. 5.

It can be seen from Fig. 5 that the MFL signal in the X-direction increases substantially when the brush thickness increases from 5 to 10 mm. Nevertheless, when the brush thickness increases from 10 to 15 mm, the MFL leakage signal in the X-direction increases slowly. Especially when the brush thickness is greater than 15 mm, the thickness has almost no effect on the MFL signal in the X-direction. Meanwhile, the increase in brush thickness has almost no effect on the MFL signal in the Y and Z directions. Based on the simulation results, the thickness of the steel brush was set to be 10 mm.

4 Conclusions

This study has investigated the effects of the steel brushes on the MFL signals using the finite element simulation model. It has been shown that steel brushes can reduce the influence of the air gap between the magnetization device and the rail surface. However, simulation results have demonstrated that thickness of steel brushes do not significantly affect the MFL signal. Moreover, it has been shown that the magnetization device can be optimized to obtain MFL signals with a higher signal-to-noise

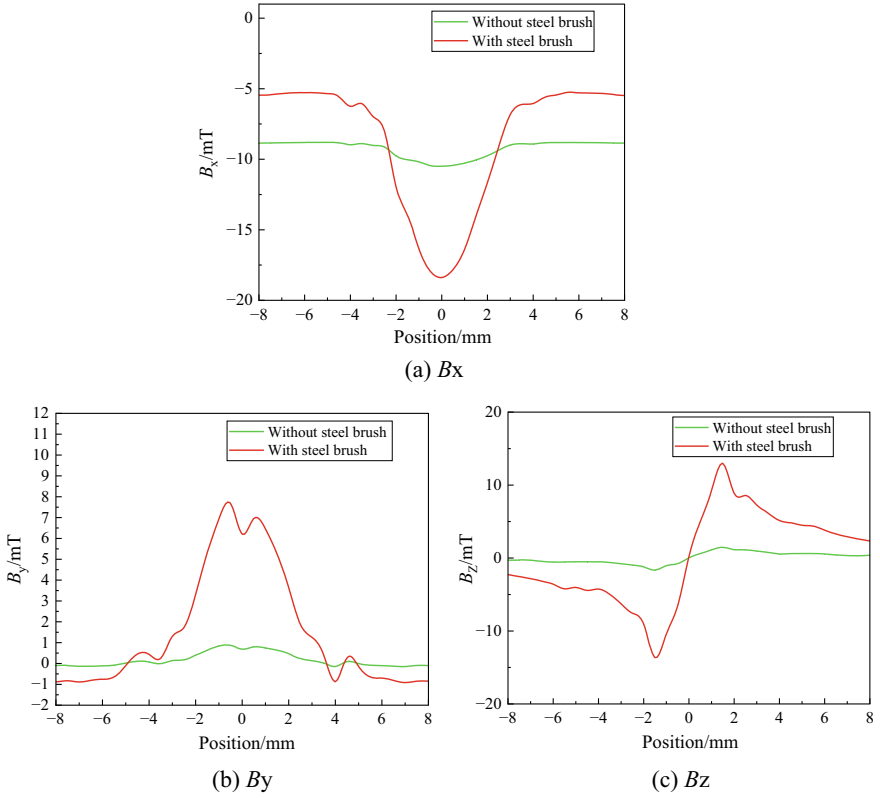


Fig. 4 MFL signal with and without steel brush

ratio. The results presented in this study help to solve the problem that some small and narrow cracks may be missed detection due to the weak MFL signals.

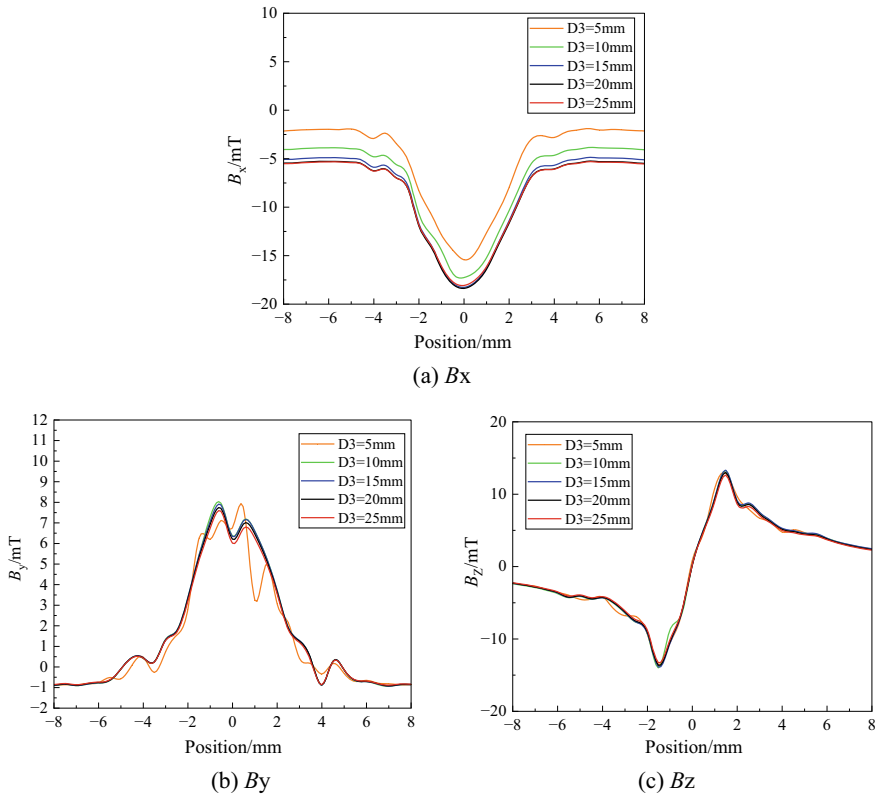


Fig. 5 MFL signal Change Rules in different steel brush thickness

Acknowledgements This work was funded and supported by a Universiti Sains Malaysia (USM), Bridging GRA Grant with Project No: 304/PELECT/6316607.

References

1. Liu G, Yang F, Wang S, Jing G, Nateghi Y (2022) Railway ballast fouling, inspection, and solutions—a review. Proc Inst Mech Eng Part F J Rail Rapid Transit. <https://doi.org/10.1177/09544097221148057>
2. Xu P, Chen Y, Liu L, Liu B (2023) Study on high-speed rail defect detection methods based on ECT, MFL testing and ACFM. Measurement 206:112213. <https://doi.org/10.1016/j.measurement.2022.112213>
3. Liu S, Chen M (2023) Wire rope defect recognition method based on MFL signal analysis and 1D-CNNs. Sensors 23:3366. <https://doi.org/10.3390/s23073366>
4. Huang S, Peng L, Sun H, Li S (2023) Deep learning for magnetic flux leakage detection and evaluation of oil & gas pipelines: a review. Energies 16:1372. <https://doi.org/10.3390/en16031372>

5. Ji K, Wang P, Jia Y, Ye Y, Ding S (2021) Adaptive filtering method of MFL signal on rail top surface defect detection. *IEEE Access* 9:87351–87359. <https://doi.org/10.1109/ACCESS.2021.3065044>
6. Antipov AG, Markov AA (2018) 3D simulation and experiment on high-speed rail MFL inspection. *NDT E Int* 98:177–185
7. Gao Y, Tian GY, Li K, Ji J, Wang P, Wang H (2015) Multiple cracks detection and visualization using magnetic flux leakage and eddy current pulsed thermography. *Sens Actuators, A* 234:269–281
8. Gong W, Akbar MF, Jawad GN, Mohamed FPM, Mohd NAW (2022) Nondestructive testing technologies for rail inspection: a review. *Coatings* 12(11):1790. <https://doi.org/10.3390/coatings12111790>

AGVs and AMRs Robots: A Brief Overview of the Differences and Navigation Principles



Sami Abdulla Mohsen Saleh , Shahrel Azmin Suandi , Haidi Ibrahim , Qusay Shihab Hamad , and Ibrahim Al Amoudi 

Abstract Automated Guided Vehicles (AGVs) and Autonomous Mobile Robots (AMRs) are swiftly revolutionizing industrial and logistics operations by automating the tasks related to material handling and transportation. While both technologies share the goal of automation, they differ in terms of control, navigation, flexibility, and interaction with humans. This research article briefly discusses the differences and the navigation principles of AGVs and AMRs. It explores the control and navigation systems, flexibility and adaptability. By understanding these differences and navigation principles, decision-makers can make informed choices to optimize material handling processes and maximize operational efficiency. This article serves as a valuable resource for industry professionals, researchers, seeking to leverage AGVs and AMRs in their operations and stay at the forefront of industrial automation.

Keywords Autonomous mobile robots · Automated guided vehicles · Navigation principles

1 Introduction

The role of material handling within a factory encompasses the control and management of materials and products throughout the manufacturing process, as well as the integration of manual, semi-automated, and automated vehicles to facilitate production logistics. Conventionally, a wide range of manufacturing industries have extensively utilized automated guided vehicles (AGVs) for material handling. These

S. A. M. Saleh · S. A. Suandi (✉) · H. Ibrahim · Q. S. Hamad · I. Al Amoudi
Intelligent Biometric Group, School of Electrical and Electronic Engineering, Universiti Sains
Malaysia, 14300 Nibong Tebal, Pulau Pinang, Malaysia
e-mail: shahrel@usm.my

H. Ibrahim
e-mail: haidi_ibrahim@ieee.org

Q. S. Hamad
University of Information Technology and Communications (UOITC), Baghdad, Iraq

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
N. S. Ahmad et al. (eds.), *Proceedings of the 12th International Conference on Robotics, Vision, Signal Processing and Power Applications*, Lecture Notes in Electrical Engineering 1123, https://doi.org/10.1007/978-981-99-9005-4_32

vehicles rely on tape, wire, or magnetic tracks to guide their movement, following pre-determined paths that determine the distance between two locations. AGVs have proven to be invaluable in high-volume manufacturing operations, although their widespread adoption has been limited due to their high initial investment costs and restricted range of applications. In recent years, there has been a notable surge in the utilization of autonomous mobile robots (AMRs) for material handling tasks. Manufacturers are increasingly opting for semi-autonomous or fully autonomous vehicles, driven by the advantages of scalability, versatility, and cost-effectiveness offered by AMRs. Unlike AGVs, AMRs possess the ability to operate autonomously, making real-time decisions and dynamically adjusting their paths based on the surrounding environment. This flexibility enables them to adapt to changing production layouts and handle various types of tasks. Moreover, AMRs are often equipped with advanced sensors and perception capabilities, allowing them to navigate complex and dynamic environments safely, without the need for physical infrastructure modifications such as tracks or wires. As a result, AMRs are gaining popularity as a viable solution for material handling in manufacturing facilities of all sizes, offering enhanced efficiency and adaptability compared to traditional AGVs [15].

The goal of this work is to present and briefly discuss the differences and the navigation principles of AGVs and AMRs. We discuss the differences of AGVs and AMRs in Sect. 2 and the navigation principles of these Robots. Finally, we draw the conclusion in Sect. 4.

2 AGVs and AMRs Robots

The utilization of AMRs for material handling is experiencing rapid growth in the manufacturing industry as manufacturers increasingly opt for these vehicles due to their scalability, versatility, and cost advantages. While AGVs have long played a significant role in production logistics, AMRs offer several distinct advantages. One key advantage is their autonomous capability, allowing them to avoid collisions and resolve conflicts between vehicles by taking alternative routes. Unlike AGVs, which are constrained by fixed routes and come to a halt when encountering obstacles, AMRs can dynamically detour around obstacles using built-in sensors such as cameras and laser scanners [5]. Figure 1 shows the guiding systems for AGVs and AMRs as described in [3]. Furthermore, AMRs can perform additional tasks during transportation, such as data collection in the era of industry 4.0, where interconnected devices and mass data are prevalent. As a result, even small- and medium-sized factories can harness the benefits and potential of AMRs.

However, the introduction of AMRs also presents unique challenges. The first challenge lies in considering the distinctive characteristics of AMRs, such as their limited payload and travel times due to their smaller size. To increase capacity and flexibility for material handling, fleets of vehicles with different specifications can be deployed. Additionally, for efficient travel time, the calculation of shortest paths considering obstacles becomes crucial in route planning. Moreover, as factories strive

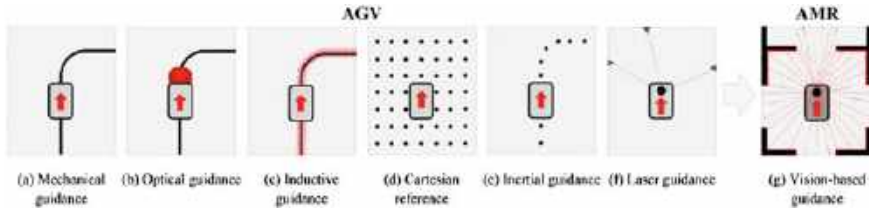


Fig. 1 Top view of guiding systems for AGVs and AMRs as described in [3]

for mass customization with small production lot sizes and low work-in-process inventory, minimizing delays in transportation requests is essential. Therefore, identifying and addressing these new characteristics, including pathfinding, maximum payloads, battery levels, and customized order fulfillment, is of paramount importance. The second challenge lies in the computational complexity associated with optimizing AMR deployment. Although mathematical models like mixed-integer linear programming can provide optimal solutions for small-scale problems, they become impractical for larger-scale ones. The inherent NP-hard nature of these problems leads to excessively long computational times. Hence, there is a pressing need for efficient algorithms capable of finding good solutions within reasonable timeframes across various environments.

As per the specifications of commercially accessible AGVs and AMRs utilized in material handling [5], Table 1 provides an overview of the common traits of these two vehicle categories.

Table 1 AGVs versus AMRs comparison [3, 5]

	AGVs	AMRs
Navigation	Wire, magnet, reflective markers, radio-frequency identification (RFID), quick response (QR) code	Pre-loaded maps with laser and depth sensors for localization
Speed	3 km/h	5–10 km/h
Payload	High (> 1500 kg)	Low (100–500 kg)
Collision avoidance	AGVs can only be routed while their path is free of obstacles	AMRs can detour around obstacles and other AMRs dynamically
Scalability	Low (new infrastructure and tracks must be installed)	High (each AMR can be controlled by a control system individually)
Power source	Large battery or inductive power transfer	Compact battery

3 Navigation Principles of AGVs and AMRs

3.1 Fixed Path Navigation

Fixed path navigation utilizes sensors on the AGV to provide guidance along a predetermined path. While relatively mature, it poses challenges in construction and altering trolley routes. Electromagnetic navigation involves burying metal wires, generating a magnetic field with low-frequency currents [2]. AGVs use induction coils to track the magnetic field for navigation. Optical navigation [10] employs luminescent belts or paint and infrared sensors to control deviation. The AGV's driving and steering motors adjust its forward direction. Tape navigation relies on a magnetic guide belt and sensor to obtain relative coordinate signals, which are transmitted to the controller for the AGV to follow the guide belt [7].

3.2 Free Route Navigation

Free path navigation enables real-time path planning and guidance for AGVs, offering high flexibility and easy trolley route changes despite higher manufacturing costs. Inertial navigation [6, 8] utilizes a gyroscope on the AGV/AMR and ground positioning blocks to calculate position and direction. Laser navigation [13] involves installing laser reflectors along the travel path and using a laser positioning device on the AGV/AMR to determine its current position and direction. Visual navigation [9, 14] a rapidly evolving method, employs a charge-coupled device camera and sensor to dynamically acquire and compare image information with a database, determining the vehicle's position and driving state. This method offers excellent navigation flexibility without the need for manual physical path settings, benefiting from advancements in computer image acquisition, storage, and processing technology.

3.3 Visual Navigation Tracking

In recent years, visual navigation technology has undergone significant advancements, primarily driven by improvements in chip performance and digital image processing capabilities. By utilizing on-board cameras, visual navigation captures environmental images, detects navigation parameters (such as position, speed, attitude), and plans paths for autonomous control of vehicles. Visual navigation finds extensive application in tracking tasks, where autonomous robots operate independently, relying on a suite of sensors, motion systems, navigation, and positioning systems to execute tasks without human intervention. Vision-based mobile robots exhibit accurate task completion using visual information as input, processed and analyzed by controller algorithms.

One of common visual navigation tracking control methods is monocular vision system. It employs a single camera mounted on the robot to capture images of the surroundings [4, 12]. Monocular vision systems require pre-calibration to extract feature information and utilize methods like edge and color detection to identify objects or obstacles. However, monocular vision lacks precise position and distance measurements, suffers from disturbances, and may lead to misinterpretation. Various studies have proposed autonomous navigation systems and obstacle detection methods using monocular cameras. Stereo vision-based navigation systems [1], are another type that involve two identical cameras mounted on the robot to acquire images. These systems can predict object distances similar to human stereo vision. Implementing a stereo vision system requires high accuracy camera alignment, specialized skills, or high-precision machinery, leading to higher costs. Researchers have proposed fusion frameworks, stereoscopic keyframe-based navigation, and deep learning-based approaches in stereo vision systems.

Deep learning plays a crucial role in visual navigation technology for mobile robots. Various applications have emerged, including end-to-end navigation using deep reinforcement learning, CNN-based architectures for navigation and location identification, and visual servoing methods. Visual servoing, an important technique in visual navigation, combines classical and deep neural network-based approaches, enabling accurate navigation and following of crop rows [11].

4 Conclusion

In conclusion, AGVs and AMRs offer distinct advantages and challenges. Organizations must carefully assess their operational requirements, environment, and desired level of automation when choosing between AGVs and AMRs. By understanding the differences and challenges presented by these technologies, organizations can implement the most suitable solution and unlock the benefits of automation, including improved efficiency, productivity, and safety. In addition, visual navigation technology has witnessed significant progress in recent years. The use of on-board cameras, combined with advanced algorithms and deep learning, enables autonomous navigation and control of mobile robots. Monocular and stereo vision systems, as well as visual servoing methods, contribute to accurate path planning, obstacle avoidance, and precise positioning. Continued research in visual navigation will lead to further advancements and applications in the field of autonomous robotics.

Acknowledgements This research is supported by Universiti Sains Malaysia (USM), the grant no. is 1001/PELECT/8021023.

References

1. Chae H-W et al (2020) Robust and autonomous stereo visual-inertial navigation for non-holonomic mobile robots. *IEEE Trans Veh Technol* 69(9):9613–9623
2. Chen X et al (2016) Electromagnetic guided factory intelligent AGV. In: 2016 3rd international conference on mechatronics and information technology. Atlantis Press, pp 200–205
3. Fragapane G et al (2021) Planning and control of autonomous mobile robots for intralogistics: literature review and research agenda. *Eur J Oper Res* 294(2):405–426. <https://doi.org/10.1016/j.ejor.2021.01.019>
4. Hao M (2020) An autonomous navigation algorithm for monocular visual recognition. In: 2020 IEEE 4th information technology, networking, electronic and automation control conference (ITNEC). IEEE, pp 1975–1978
5. Jun S et al (2021) Pickup and delivery problem with recharging for material handling systems utilising autonomous mobile robots. *Eur J Oper Res* 289(3):1153–1168. <https://doi.org/10.1016/j.ejor.2020.07.049>
6. Kim J et al (2012) Inertial navigation system for an automatic guided vehicle with Mecanum wheels. *Int J Precis Eng Manuf* 13:379–386
7. Lynch L et al (2018) Automated ground vehicle (AGV) and sensor technologies—a review. In: 2018 12th international conference on sensing technology (ICST). IEEE, pp 347–352
8. Pivarčiová E et al (2018) Analysis of control and correction options of mobile robot trajectory by an inertial navigation system. *Int J Adv Robot Syst* 15(1):1729881418755165
9. Ran T et al (2021) Scene perception based visual navigation of mobile robot in indoor environment. *ISA Trans* 109:389–400
10. Run R-S, Xiao Z-Y (2018) Indoor autonomous vehicle navigation—a feasibility study based on infrared technology. *Appl Syst Innov* 1(1):4
11. Sadeghi Esfahlani S et al (2022) The deep convolutional neural network role in the autonomous navigation of mobile robots (SROBO). *Rem Sens (Basel)* 14(14):3324
12. Sun T et al (2020) An improved monocular visual-inertial navigation system. *IEEE Sens J* 21(10):11728–11739
13. Thanh LB et al (2013) AGV trajectory control based on laser sensor navigation. *Int J Sci Eng* 4(1):16–20
14. Yasuda YDV et al (2020) Autonomous visual navigation for mobile robots: a systematic literature review. *ACM Comput Surv (CSUR)* 53(1):1–34
15. Zhang J et al (2023) Automated guided vehicles and autonomous mobile robots for recognition and tracking in civil engineering. *Autom Constr* 146:104699

Development of Delivery Robot with Application of the TRIZ Method



Zulkifli Ahmad and Muhd Aslam Zulkarnain

Abstract Delivery robots have become increasingly popular due to their potential to improve the efficiency and safety of delivery services. This study aims to develop an aesthetically pleasing delivery robot's body for indoor applications that can be cost-effective and efficient. The development process involved the delivery robot's design, construction, and testing. The robot's body was meticulously designed, focusing on creating a vigilant and suitable robot coexisting with humans. The robot includes three cargo bays with a capacity of 10 kg each. The delivery mechanism is controlled by a microcontroller and integrated with ultrasonic and infrared sensors for obstacle avoidance and line tracking. In addition, the delivery robot's design was analyzed and enhanced through the Theory of Inventive Problem Solving (TRIZ). In conclusion, the development results demonstrated that the robot could perform accessible delivery services inside buildings effectively.

Keywords Delivery robot · Line following · Industrial design · Arduino

1 Introduction

The term “robot” originates from the Czech word “robota,” meaning servant or laborer [1]. A robot is described as a mechanical device that performs automated tasks either under human supervision, following a predetermined program, or based on general rules using artificial intelligence [2]. Initially limited to industrial settings, robots have now become an integral part of our daily lives, engaging directly with humans and performing tasks that make life more convenient. The concept of robot deliveries began in late 2014 with the introduction of Relay, a room service robot by

Z. Ahmad (✉) · M. A. Zulkarnain

Faculty of Mechanical and Automotive Engineering Technology, Universiti Malaysia Pahang
Al-Sultan Abdullah, 26600 Pekan, Pahang, Malaysia

e-mail: kifli@umpsa.edu.my

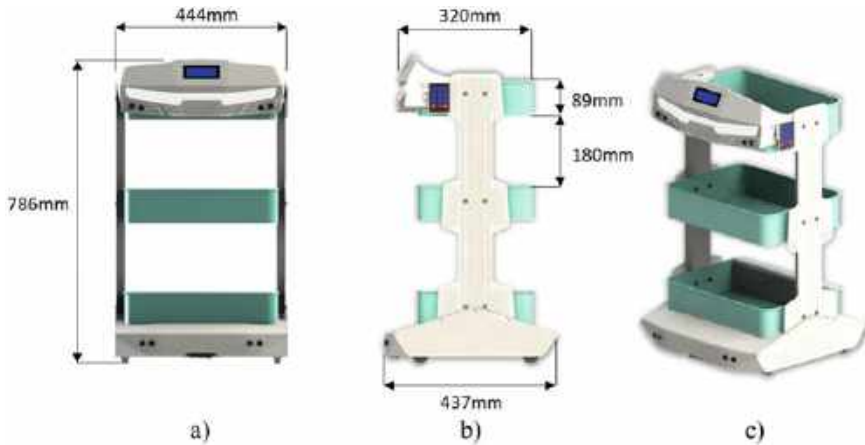


Fig. 1 a Front view, b side view, and c isometric view of the delivery robot's 3D model design

Savioké [3]. This marked the beginning of using robots in hotels for tasks like transporting ordered items to guest rooms. Delivery robot services have since expanded to include food delivery in restaurants and assistance in various settings.

2 Methodology

2.1 3D Modelling

The first step in the development process of the robot is to create 3D CAD model design of the robot's structure. The 3D model of the selected design concept is constructed from scratch utilizing SolidWorks software (Fig. 1).

2.2 Fabrication Process

The design of the delivery robot was fabricated using two primary processes: laser cutting, 3D printing, and wiring (Fig. 2).

Laser Cutting. Laser cutting uses a focused laser beam to cut materials, which melts, burns, vaporizes away by a jet of gas, leaving an edge with a high-quality surface finish. The laser cutting procedure for these components begins with the conversion of a 3D model in SLDPRT file format to the DXF file format to produce the SVG code using Adobe Illustrator software for the cutting operation.

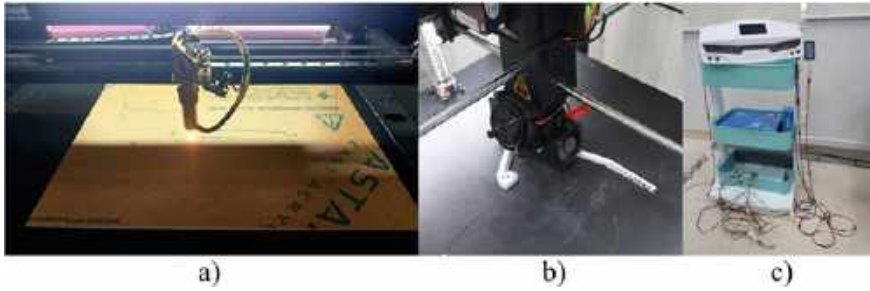


Fig. 2 a Laser cutting, b 3D printing, and c wiring process

3D Printing. 3D printing is a process of creating a physical object from a digital model where the object is built up layer by layer [4]. The 3D printing process begins with the creation of a 3D digital model of the object using SolidWorks software. This model is then sliced into thin layers, which are used as the basis for building the physical object. The 3D printer reads the digital model and begins to build the object layer by layer. This project utilizes Raise 3D Pro2 Plus, and Artillery Sidewinder x2 printer machine. The material used for the printed product in this project is polylactic acid (PLA).

Wiring. Involve interconnecting different kinds of electrical parts to produce functional and dependable systems. It includes organizing and routing wires to guarantee efficient signal transmission and reduce the risk of interference and short circuits. The robot is wired and soldered in accordance with the system schematic depicted in Fig. 3.

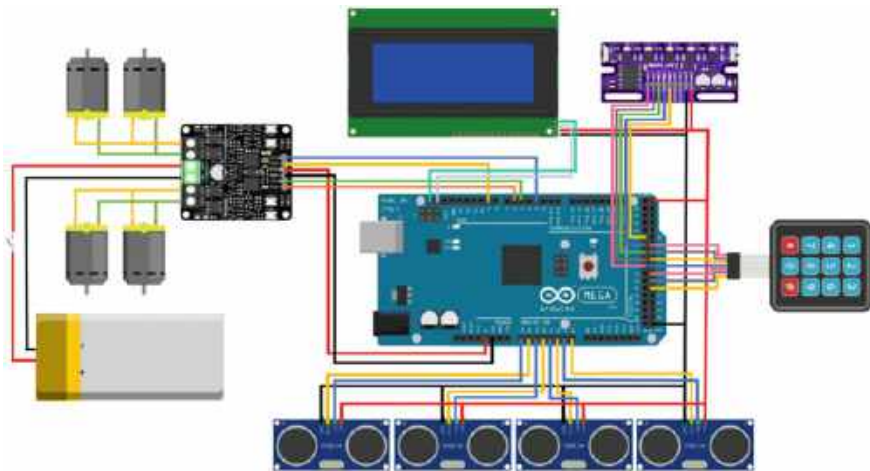


Fig. 3 System schematic diagram



Fig. 4 The delivery robot's prototype body design

3 Results and Discussion

3.1 Product Prototype

The physical prototype of the delivery robot is named as 'DezyBot' (Robot for Easy Delivery). The robot prototype demonstrates the successful integration of multiple components and subsystems, such as the chassis, navigation system, parcel compartment, and sensory modules. The delivery robot utilizes line tracking mechanism to perform easy delivery service efficiently inside of buildings (Fig. 4).

3.2 TRIZ Analysis

TRIZ is a problem-solving methodology that offers a systematic approach to innovation. Its main goal is to provide a structured framework for problem-solving by identifying inventive principles and patterns that can be applied to overcome technical contradictions and find innovative solutions.

Analysis 1. To address the problem of robot's keypad positions that automatically rotate when robot in motion (Table 1).

Based on the result, TRIZ inventive principle 24 suggests introducing an intermediate object or element between two interacting components or systems. In this case, magnets are utilized to temporarily attach the keypad to the robot's body, serving as a mechanism to secure it during movement.

Analysis 2. To solve issue regarding lengthy time consumption in assembly and disassembly process of robot's components (Table 2).

Based on the analysis, TRIZ Inventive Principle 28 is employed to resolve the issues. Mechanical substitution involves replacing mechanical components or processes with alternative options to achieve the desired functionality. In this case,

Table 1 Results of TRIZ analysis 1

Contradiction statement	If hinge is designed with big tolerance. Then less energy and friction required to rotate it. But the hinge will be unstable	
Responding variables	Positive	Negative
	User able to rotate keypad easily using less energy since hinge dimension has big tolerance	Unstable hinge due to big tolerance hinge design resulting in bigger gap around hinge
Parameter	(19) Use of energy by moving object	(13) Stability of the object composition
Inventive principle	13, 17, 19, 24	
Solution principle (24) Intermediary	Merge keypad placement temporarily with robot body by using magnet	

Note Bold text is represent the compulsory word needed in order to develop the contradiction statement in TRIZ

Table 2 Results of TRIZ analysis 2

Contradiction statement	If we use lot of fasteners. Then we able to strengthen the assembly connections. But the assembly and disassembly process will time-consuming	
Responding variables	Positive	Negative
	Studier assembly connection due to high number of fasteners in securing the connection	The assembly/disassembly process of components is time-consuming
Parameter	(14) Strength	(25) Loss of time
Inventive principle	3, 10, 28, 29	
Solution principle	(28) Mechanical substitution: change fastening methods	

Note Bold text is represent the compulsory word needed in order to develop the contradiction statement in TRIZ

excessive fasteners are replaced with more efficient and time-saving fastening techniques, such as snap-fit connections. These alternative methods ensure a secure and sturdy assembly while reducing the time required for assembly and disassembly.

Analysis 3. To eliminate the existence of joint gap in the assembly connection of the robot’s component (Table 3).

Analysis results led to the use of TRIZ principles 6 and 1. Universality involves designing cover parts and the trolley structure with greater adaptability, minimizing joint gaps caused by poor precision. Using larger tolerance holes and adjusting bolts helps close the gaps. Instead of mass-produced trolleys, self-fabricated aluminum profiles offer superior precision, and adjustable dimensions, reducing joint gaps. Segmentation involves dividing the structure into smaller segments, simplifying manipulation and addressing specific problems. Modular design of aluminum profiles

Table 3 Results of TRIZ analysis 3

Contradiction statement	If we use market-available trolley. Then we be able to save manufacturing time and cost. But it will be difficult to measure trolley dimension	
Responding variables	Positive	Negative
	Manufacturing time and cost can be saved because market-available trolley is used as robot structure	Difficult to measure the dimension of the trolley due to poor manufacturing precision of mass scale product
Parameter	(32) Ease of manufacture	(37) Difficulty of detecting and measuring
Inventive principle	1, 6, 11, 28	
Solution principle	(6) Universality: create more universal design	
	(1) Segmentation: make an object sectional	

Note Bold text is represent the compulsory word needed in order to develop the contradiction statement in TRIZ

allows easy assembly and adjustments, ensuring proper alignment and reducing joint gaps.

Analysis 4. To resolve the issue involving the robot cannot correctly reverse in a straight line if the motor speed is not set to maximum (Table 4).

The analysis suggests using inventive principle 3 to address the issues. This principle focuses on optimizing each component of an object to operate under the most favorable conditions. In this case, the robot’s system can be modified to enable it to operate in the most suitable conditions. Instead of moving in reverse, alternative actions like moving forward and making left or right turns can be programmed to allow the robot to retrace its route efficiently.

Table 4 Results of TRIZ analysis 4

Contradiction statement	If motor speed is not set at maximum. Then it can ensure smooth operations delivery process. But robot unable to reverse properly	
Responding variables	Positive	Negative
	Robot operation become smooth since robot able to avoid sudden acceleration movement	Robot unable to reverse properly and unable to perform path correction if motor speed is not set at maximum
Parameter	(33) Ease of operation	(38) Extend of automation
Inventive principle	1, 3, 12, 34	
Solution principle	(3) Local quality: change robot action and make robot to function in conditions most suitable for its operation	

Note Bold text is represent the compulsory word needed in order to develop the contradiction statement in TRIZ

4 Conclusion

In conclusion, this study has successfully developed a physical prototype of a delivery robot designed specifically for indoor applications. The robot's design was constructed, tested, and evaluated for its performance and functionality. Key features of the delivery robot include a well-designed body fabricated using laser cutting and 3D printing, four 12 V DC motors, a LIPO battery pack, and a microcontroller for precise delivery control. The integration of ultrasonic and infrared sensors enables effective obstacle avoidance and line tracking capabilities. By applying TRIZ analysis, the study effectively resolved and improved design issues, leading to an enhanced robot design that exhibited satisfactory performance. The outcomes of this research contribute to the advancement of delivery robot technology and hold practical implications for enhancing delivery services across various industries.

Acknowledgements The authors thank Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA) for funding this research under prototyping grant PDU223207 and Tabung Persidangan Dalam Negara (TPDN).

References

1. Hockstein NG, Gourin CG, Faust RA, Terris DJ (2007) A history of robots: from science fiction to surgical robotics. *J Robot Surg* 1(2):113–118. <https://doi.org/10.1007/s11701-007-0021-2>
2. Goris K (2004) Autonomous mobile robot design
3. Michael (2022) Last mile delivery robots, history, benefits and best companies. https://angarium.com/last-mile-delivery-robots/#History_of_robot_delivery. Accessed 26 Dec 2022
4. Shahrubudin N, Lee TC, Ramlan R (2019) An overview on 3D printing technology: technological, materials, and applications. *Proc Manuf* 35:1286–1296. <https://doi.org/10.1016/j.promfg.2019.06.089>

Enhancement of Adaptive Observer for Fault Detection in Direct Current Motor System Using Kalman Filter



Nur Dalilah Alias and Rosmiwati Mohd Mokhtar

Abstract Rotational machines such as direct current (dc) motors might be exposed to unexpected failures, which can cause production delays or safety problems. These failures must be detected immediately before the machines' condition worsens or fails. A fault detection strategy can be used to detect the faults in the machines. This study aims to improve the adaptive observer-based fault detection techniques by implementing the low pass and Kalman filters. The dc motor was modelled in a state-space system in the simulation, and the encoder was modelled to have faults. From results, Kalman filter can tolerate the encoder fault's effect and better estimates the actual states than the low pass adaptive observer.

Keywords Fault detection · Dc motor · Adaptive observer · Kalman filter

1 Introduction

Technological advancement in the industry causes complexity and diversity in the system. Thus, there will be a substantial possibility of fault in the system, resulting in poor efficiency in the process. A fault is an unpermitted deviation of at least one characteristic property or variable in a system from its acceptable, usual, or standard condition [1]. Blocking an actuator, losing a sensor signal, or disconnecting a system component are such faults. Dc motor drives are utilized in various speed and position control systems due to their superior performance, ease of control and high efficiency [2]. However, due to mechanical wear or ageing, the dc motor requires frequent maintenance. Plus, to discover faulty scenarios, it is essential to use well-known dynamical characterizations of the analytical model motor [3]. A rotary encoder is

N. D. Alias · R. M. Mokhtar (✉)

School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Engineering Campus, 14300 Nibong Tebal, Pulau Pinang, Malaysia

e-mail: eeerosmiwati@usm.my

N. D. Alias

e-mail: dalilahalias@student.usm.my

the most common sensor to detect the position and speed of a motor. It is made of a circular disk with marks that can be detected by mechanical, optical, or magnetic sensors to determine the rotor's position [4]. However, the information can sometimes be lost due to the malfunction in the encoder. The encoder information loss defect can be caused by various variables, including mechanical, electrical, and optical causes [4].

To ensure process efficiency, more regular monitoring, including process control and appropriate corrective actions, is required for the change in the process. Maintaining a desirable performance in industrial processes that can contain various flaws is a critical responsibility. Fault Detection and Diagnosis (FDD) is a required control method to achieve this task among numerous process supervision approaches due to most industries seeking to enhance their process performance by improving their FDD capability [5]. FDD systems have three subsystems, each integrated with fault detection, isolation, and estimation capabilities [6]. Many researchers have proposed similar fault detection projects with different types of observers. For instance, an observer-based approach was used in the fault detection for a dynamical system which is dc motor and encoder to represent the sensor has been presented in [4]. Furthermore, a neural adaptive observer that uses a neural network for fault detection while an embedded Kalman filter updates the weighting parameters has been presented in [7]. An adaptive Kalman filter for actuator malfunction diagnostic in a discrete-time stochastic time-varying system was proposed in [8]. A sliding mode controller is used in [9] to control dc motors via MATLAB/Simulink simulation and Arduino hardware implementation. Based on the research, several algorithms have been implemented for fault detection.

Thus, the motivation of this paper is to implement a quantitative model-based method by further designing an adaptive observer to be implemented in the fault detection system. The observer's algorithm will be further improved to obtain a fair value on the proposed fault, which is the encoder fault. Dc motor will represent the dynamical system, and an incremental encoder will be used as a sensor. The mathematical model of the dc motor will be determined. The fault detection mechanism will be modelled and designed using a MATLAB/Simulink environment.

2 Methodology

The dc motor was first modelled and transformed into a state space form in this project. The adaptive observer was designed based on the equations modelled in MATLAB/Simulink. Two filters, a low pass filter and a Kalman filter, were designed in the simulation to identify the effect of filters on fault detection. The encoder fault was the fault implemented in the simulations.

Table 1 Dc motor parameters

Parameter	Value
Moment of inertia, J_m	0.00025 N m/rad/s ²
Frictional coefficient, B_m	0.0001 N m/rad/s
Armature resistance, R_a	0.5 Ω
Armature inductance, L_a	1.5 mH
Motor torque constant, K_m	0.05 N m/A

2.1 Dc Motor Modelling

The dynamic equations of the dc motor were derived based on Kirchhoff's Voltage Law (KVL) and Newton's Second Law of Motion. The model was then transformed into Laplace transform and further into state space form, as shown in (1).

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned} \quad (1)$$

where $x(t)$, $u(t)$ and $y(t)$ are the state, input, and output vector, respectively. By considering the armature current, angular position and angular velocity as the state variables, armature voltage as the input and angular position as the output, the model was derived into the state space model as shown in (2). The parameters used for the dc motor model are shown in Table 1.

$$A = \begin{bmatrix} -\frac{R_a}{L_a} & 0 & -\frac{K_m}{L_a} \\ 0 & 0 & 1 \\ \frac{K_m}{J_m} & 0 & -\frac{B_m}{J_m} \end{bmatrix}; \quad B = \begin{bmatrix} \frac{1}{L_a} \\ 0 \\ 0 \end{bmatrix}; \quad C = [0 \ 1 \ 0] \quad (2)$$

2.2 Adaptive Observer with Low Pass Filter

In [10], the author proposed a method for sensor fault estimation in Multiple-Input Multiple-Output (MIMO) nonlinear system using an adaptive observer for joint estimation states and sensor faults in a state space formulation of the monitored system. By following the same approach as in [10], the algorithms for adaptive observer were performed and designed in Simulink. A low pass filter was inserted in the observer via a low pass filter block in Simulink. Consider the linear time-varying systems shown in (3) and (4),

$$\dot{x}(t) = Ax(t) + Bu(t) \quad (3)$$

$$y(t) = Cx(t) + v(t) + f(t) \quad (4)$$

where $f(t)$ is the possible sensor faults and can be expressed in linear regression shown in (5),

$$f(t) = \Psi(t)\rho(t) \quad (5)$$

where $\Psi(t) = [\Psi_1(t), \dots, \Psi_p(t)]$ are the regressors and $\rho(t) = [\rho_1(t), \dots, \rho_p(t)]^T$ are the unknown regression coefficients.

Few assumptions were applied to ensure the algorithm's convergence [10]. First, matrices A and C were assumed to be completely observable. Thus, matrix K can be designed to make the system exponentially stable, as shown in (6). Another assumption is to let the matrix of signal $\Psi(t)$ be filtered through a linear time-varying filter, as shown in (7) and (8).

$$\dot{\eta}(t) = [A - KC]\eta(t) \quad (6)$$

$$\dot{Y}(t) = [A(t) - K(t)C(t)]Y(t) - K(t)\Psi(t) \quad (7)$$

$$\Omega(t) = C(t)Y(t) + \Psi(t) \quad (8)$$

where $\dot{Y}(t)$ and $\Omega(t)$ are the state and output filters. $\Psi(t)$ is assumed to be persistently exciting. Thus, $\Omega(t)$, satisfies for some positive constant α , T and $t \geq t_0$, as shown in (9),

$$\int_t^{t+T} \Omega^T(\tau)\Omega(\tau)d\tau \geq \alpha I_p \quad (9)$$

An adaptive observer in the form of the ordinary differential equation (ODE), as shown in (10), (11) and (12), was designed in MATLAB/Simulink, and a low pass filter was implemented after \hat{x} generated in the simulation. A , B and C were previously substituted with the state space model in (2).

$$\dot{Y}(t) = [A(t) - K(t)C(t)]Y(t) - K(t)\Psi(t) \quad (10)$$

$$\dot{\hat{x}}(t) = A(t)\hat{x}(t) + B(t)u(t) + K(t)\left[y(t) - C(t)\hat{x}(t) - \Psi(t)\hat{\theta}(t)\right] + Y(t)\hat{\theta}(t) \quad (11)$$

$$\dot{\hat{\theta}}(t) = \Gamma[C(t)Y(t) + \Psi(t)]^T \cdot \left[y(t) - C(t)\hat{x}(t) - \Psi(t)\hat{\theta}(t)\right] \quad (12)$$

2.3 Kalman Filter

The Kalman filter is a recursive predictive filter that employs state space concepts and a recursive algorithm where the estimated state from the previous time step and current measurement is required to compute the estimate of the current state [11]. In the simulation, the continuous dc motor system was sampled into a discrete system before implementing the Kalman filter. Kalman filter equations applied to the dc motor are shown in (13). A , B and C were previously substituted with the state space model in (2).

$$\begin{aligned} x_{(k+1)}^- &= Ax_{(k)} + BU_k \\ P_{(k+1)}^- &= AP_{(k)}A^T + Q \end{aligned} \quad (13)$$

The measurement update equations are shown in (14), (15) and (16),

$$K_{(k)} = P_{(k)}^- C^T \left[\left(C^T P_{(k)}^- C + R \right) \right]^{-1} \quad (14)$$

$$\hat{x}_{(k)} = x_{(k)}^- + K_{(k)} \left[Z_{(k)} - (Cx_{(k)}) \right] \quad (15)$$

$$P_{(k)} = (I - K_{(k)}C) P_{(k)}^- \quad (16)$$

where $K_{(k)}$, $P_{(k)}$, $P_{(k)}^-$ are Kalman gain, posterior co-variance, and prior co-variance. Q and R are the state noise (system uncertainty) and measurement noise.

3 Result and Discussion

DC motor system observability was generated first in MATLAB to ensure that the system is observable before performing simulations on the observer. From the source code generated, the observability matrix has a rank of three, similar in dimension to the dc motor matrix A . Thus, the state space model of the dc motor system was utterly observable. The system matrices A , B and C were calculated based on the dc motor parameters in Table 1 and Eq. (2). The input signal used for the simulations was a square wave with an amplitude of -1 and frequency of 0.5 Hz. A band-limited white noise block was modelled to indicate the occurrence of the encoder's fault in the simulation. The actual and estimated output waveforms were observed on the armature current, angular position and angular velocity, and the residual on each output was generated.

Figure 1 showed the performance of the low-pass adaptive observer without the Kalman filter when the encoder fault signal was inserted in the simulation. From the simulations, this observer could not tolerate the noise signal generated from the faults. The estimated current in Fig. 1a was unable to follow the waveform of the

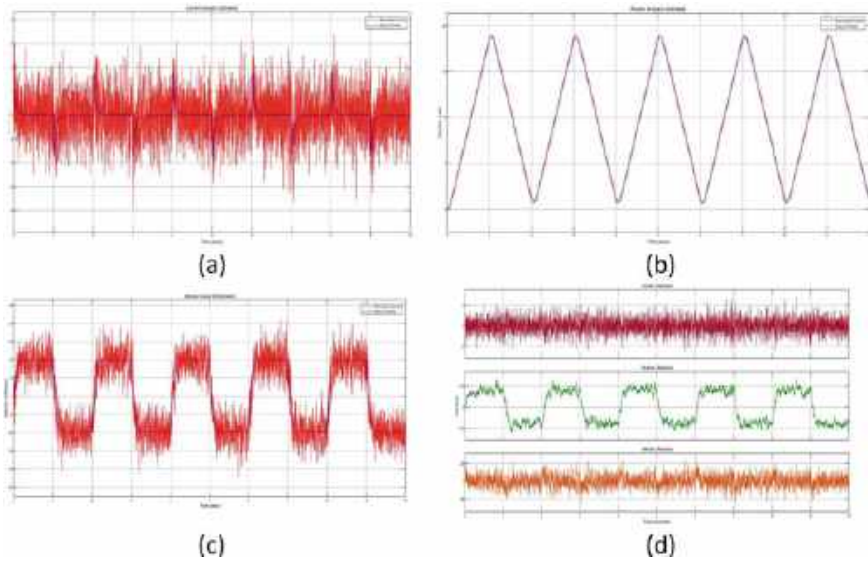


Fig. 1 The actual (blue) and estimated (red) results using low pass adaptive observer without Kalman filter on **a** current, **b** position, **c** velocity; **d** residual generated on the current, position and velocity

actual current. However, the estimated position and voltage in Fig. 1b, c could still follow their actual waveforms. However, the signal in Fig. 1c was corrupted with a noise signal. Thus, the residual shown in Fig. 1d consists of a high residual on current and velocity compared to residual on position.

Figure 2 shows the performance of the low pass adaptive observer with the Kalman filter when the encoder fault signal was inserted in the simulation. The Kalman filter adapted to the noise generated from the encoder fault. The estimated output states could follow the waveforms of their respective actual signals with very minimum distortion. The response was much better than that without the Kalman filter, in which, previously seen that the performance was poor and far from actual values. Thus, the residual shown in Fig. 2d was improved, and the values were close to zero residual.

The low pass adaptive observer and Kalman filter simulations were analyzed and compared according to the mean squares error (MSE) calculation. The results were tabulated as shown in Table 2. From the table, the observer with the Kalman filter performs better than the low-pass adaptive observer in no-fault and encoder faults implemented in the system. The MSE values on the Kalman filter were much smaller compared to the low-pass adaptive observer.

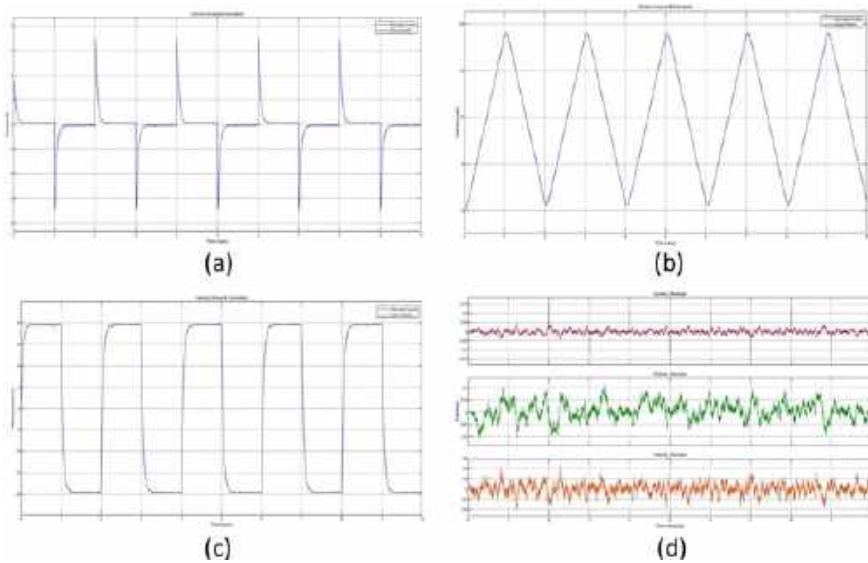


Fig. 2 The actual (blue) and estimated (red) results using low pass adaptive observer with Kalman filter on **a** current, **b** position, **c** velocity; **d** residual generated on the current, position and velocity

Table 2 MSE from simulation results

	Low pass adaptive observer			An observer with Kalman filter		
	Current	Position	Velocity	Current	Position	Velocity
No fault	0.09482	0.1379	4.2800	0	0	0
Encoder fault	3.0031	0.1444	44.6682	0.000136	0.001279	0.01219

4 Conclusion

Fault detection was studied using a low-pass adaptive observer and Kalman filter in the dc motor system. Based on the results analyzed, low-pass adaptive observers, with and without the Kalman filter, could detect faults in the dc motor system. However, the Kalman filter can tolerate or eliminate the fault’s effect and better estimates the actual states than the low pass adaptive observer. The enhancement of the standard low pass adaptive observer model with Kalman filter will maintain good performance of the dc motor system even during the influence of a faulty encoder.

Acknowledgements The Ministry of Higher Education Malaysia supports the work with FRGS Project Code: FRGS/1/2019/TK04/USM/02/12.

References

1. Gertler J (2019) Fault detection and diagnosis in engineering systems. CRC Press, New York
2. Joshi B, Shrestha R, Chaudhar R (2014) Modeling, simulation and implementation of brushed dc motor speed control using optical incremental encoder feedback. In: Proceedings of IOE graduate conference, 10–11 Oct 2014, pp 497–505
3. Adouni A, Abid A, Sbita L (2016) A dc motor fault detection, isolation and identification based on a new architecture artificial neural network. In: 2016 5th international conference on systems and control (ICSC), Marrakesh, Morocco, pp 294–299
4. Lee J, Mohd-Mokhtar R, Mahyuddin MN (2020) Adaptive observer for dc motor fault detection dynamical system. In: Zain M et al (eds) Proceedings of the 11th national technical seminar on unmanned system technology. Lecture notes in electrical engineering, vol 666, pp 285–297
5. Park YJ, Fan SKS, Hsu CY (2020) A review on fault detection and process diagnostics in industrial processes. *Processes* 8(9)
6. Ahmad M, Mohd-Mokhtar R (2021) A survey on model-based fault detection techniques for linear time-invariant systems with numerical analysis. *Pertanika J Sci Technol* 30(1):53–78
7. Abbaspour A, Aboutalebi P, Yen KK, Sargolzaei A (2017) Neural adaptive observer-based sensor and actuator fault detection in nonlinear systems: application in UAV. *ISA Trans* 67:317–329
8. Zhang Q (2018) Adaptive Kalman filter for actuator fault diagnosis. *Automatica* 93:333–342
9. Ma'arif A, Çakan A (2021) Simulation and Arduino hardware implementation of dc motor control using sliding mode controller. *J Robot Control (JRC)* 2(6):582–587
10. Ding Q, Peng X, Zhang X, Hu X, Zhong X (2017) Adaptive observer-based fault diagnosis for sensor in a class of MIMO nonlinear system. In: 36th Chinese control conference, pp 7051–7058
11. Cui W (2018) Kalman filter based fault detection and diagnosis. M.Eng. thesis. Flinders University

Model Reference Adaptive Control for Acoustic Levitation System Based on Standing Waves



Ibrahim Ismael Ibrahim Al-Nuaimi 
and Muhammad Nasiruddin Mahyuddin 

Abstract In the context of industry 4.0 and the fast development taking place nowadays all over the world towards industry 5.0, it's obvious that Robots are playing a very important role in this context alongside human in incorporating sustainability and resilience aims. so that, the robotic non-contact manipulation in the field of industry, medicine and chemical processes is considered as a main core nowadays due to its ability to achieve the desired targets with high accuracy, less time and less faults to do such a duty, and this depends on how to control the robotic operations effectively to fulfill the proper demands. This research article proposed the Model Reference Adaptive Control approach to be applied for the first time on one of the most well-known non-contact methods which is acoustic levitation based on standing waves. The control approach used in this study succeeded in accomplishing the required performance and keeping the levitated particle at the desired point within the sonic field.

Keywords Model reference adaptive control · Acoustic levitation · Standing wave

1 Introduction

Acoustic levitation is a phenomenon or technique that relies on sonic radiation from the sound waves in the medium to suspend things in air or other media, using sound moving into and out of air in order to balance the gravity force, in other word an

I. I. I. Al-Nuaimi (✉) · M. N. Mahyuddin

School of Electrical and Electronic Engineering, Universiti Sains Malaysia, 14300 Nibong Tebal, Pulau Pinang, Malaysia

e-mail: ibrahim27@student.usm.my

M. N. Mahyuddin

e-mail: nasiruddin@usm.my

I. I. I. Al-Nuaimi

Department of Electrical Power and Machine Engineering, College of Engineering, University of Diyala, Baqubah, Diyala, Iraq

object inside a sound field will be influenced by a force within the field itself. The main application of this technique is the noncontact manipulation processing, which enables the handling of all materials like solids [1], liquids [2], even small living animals [3, 4] while avoiding contact noise and pollution. Being able to levitate any material sets acoustic levitation apart from other methods like magnetic and electromagnetic levitation, which can only be used with certain types of materials like magnetic materials and conductive materials respectively. This is the most significant advantage of acoustic levitation that makes it suitable for a wide range of applications. The manufacturing of micro-electronic systems, where handling the components is very hard due to their delicate and sensitive nature [5], as well as the biological/chemical industry when handling hazardous and high purity materials [6], both demonstrate numerous advantages of non-contact processes of objects. The ideal way to prevent problems with damage, scratching, and contact pollution caused by conventional physical contact operations on very precise materials or parts is to use a non-contact manipulation approach based on acoustic levitation [7]. There are five different ways to use acoustic levitation technique based on [8] but only three of them are yet involved in the context of robotic noncontact manipulation, for more details see [9]. The most famous one is the standing wave method which is the technique used for acoustically suspending particles or objects in a sonic field. There are primarily two types of standing wave levitation. The standing wave field is created between two opposing transducers in the first type, which is known as the single axis [10]. The second form uses a closed resonant chamber to produce a standing wave field in one of the cavity's acoustic modes [11]. According to [9], there are numerous options for improvement of the control in the field of acoustic levitation, some of which can be further studied in regard to a particular kind of industrial application. This article focuses on how Model Reference Adaptive Control (MRAC) can be applied as a new control strategy for a standing wave-based acoustic levitation system. When incident and reflected waves interact to produce a stable standing wave pattern, this technique is known as the standing wave method and is frequently employed in acoustic levitation. This article lays the groundwork for additional research and advancements in acoustic levitation control methods by outlining the approaches, results, and conclusions. This encourages further development in this area and makes it possible to investigate fresh ideas and opportunities. In the following sections, the oscillatory model of the object levitated in the sound field of standing wave which can be described as a mass spring damper model is presented in Sect. 2. The control strategies used in the article, which is MRAC, are fully presented in Sect. 3. Then, the simulation results based on MATLAB/Simulink are presented in Sect. 4. In Sect. 5, a conclusion and possible future ideas are offered.

2 Theory

Standing waves are created when two waves that have the same frequency, amplitude, and phase superimpose on one another to form nodes and anti-nodes. The drag force functions as a damping force and the acoustic radiation forces act as springs when a particle is displaced from a node [12]. and because of this, the dynamic model that most researchers use to describe a particle vibrating vertically within a standing acoustic wave is:

$$m \frac{dz^2}{dt} + R \frac{dz}{dt} + k(z) = 0 \quad (1)$$

where m is the mass of the particle, R is the viscous damping coefficient, the elastic spring constant of the acoustic radiation force is K , and the vertical displacement of the object is z [12]. The equation for the viscous damping coefficient is:

$$R = 6\pi \mu r \quad (2)$$

where μ is the dynamic viscosity and r is the radius of the particle [13].

It is possible to get the elastic constant K by first obtaining the Gor'kov Potential equation, which is expressed as follows:

$$U = 2k_1(|p|^2) - 2k_2(|p_x|^2 + |p_y|^2 + |p_z|^2) \quad (3)$$

$$k_1 = \frac{1}{4}V \left(\frac{1}{c_o^2 \rho_o} - \frac{1}{c_p^2 \rho_p} \right) \quad (4)$$

$$k_2 = \frac{1}{4}V \left(\frac{\rho_o - \rho_p}{\omega^2 \rho_o (\rho_o + 2\rho_p)} \right) \quad (5)$$

where V is the volume of the particle, ω is the frequency of the wave, ρ_o is the density of air, ρ_p is the density of the particle, c_o is the speed of sound in air, c_p is the speed of sound of the particle material, p is the complex pressure field, and p_x , p_y , and p_z are its derivative over their respective axes. The elastic constant and the acoustic radiation force can be calculated following the discovery of the Gor'kov potential equation, where the elastic constant is the second derivative of the Gor'kov potential, and the acoustic radiation force is the negative gradient of the Gor'kov potential [8].

$$K = \frac{d^2U}{dz^2} \quad (6)$$

$$F_{radiation} = -\nabla U \quad (7)$$

The partial derivative of the potential in the x, y, and z directions is what is essentially meant by “taking the gradient,” and this creates a vector field that is connected to the acoustic radiation force. The second derivative of the Gor’kov potential in the z direction is the elastic constant K in the vertical direction.

Where P_0 is the experimentally measured constant pressure amplitude of the transducer, V_{pp} is the excitation signal from peak to peak, k is the wave number, a is the radius of the ultrasonic transducer, $1/d$ is the distance from the transducer, φ is the phase of the ultrasonic transducer, and D_f is the far-field directivity function [14].

3 Control Strategy

The desired behavior of a particular process that can be characterized is taken into consideration as a starting point. By making use of a model reference, it is possible. In particular, a linear time-invariant system (LTI) is used to represent the process, and MRAC is used as the model reference. The transfer function $G_m(s)$ is related with the mentioned process, which is driven by its input reference. MRAC has an inner loop and an outer loop and was derived from continuous systems. The outer loop is only used to modify controller parameters; the inner loop includes both the process itself and traditional feedback. Successfully reducing the discrepancy between system output and reference model is a key objective. The amount of this gap, known as the error signal $e(t)$, depends on the model reference that was selected, the process $y(t)$, which must be carried out after the output signal, and the command signal. The error signal is claimed to be reduced to a null value for all command signals when a perfect model is said to be possible [15]. In the specific situation of MRAC, all parameters can be modified either by applying a stability theory or a gradient method. The reference model is selected as a second order transfer function.

$$G_m(s) = \frac{25}{s^2 + 2\omega_n\zeta s + 25} \quad (8)$$

The control law and controller parameters have been adjusted as follows:

$$u(t) = k_1 r(t) - k_2 z(t) - k_3 \frac{dz}{dt} \quad (9)$$

$$\frac{dk_1(t)}{dt} = -\gamma \left(\frac{1}{s^2 + 2\omega_n\zeta s + \omega_n^2} r(t) \right) e(t) \quad (10)$$

$$\frac{dk_2(t)}{dt} = \gamma \left(\frac{1}{s^2 + 2\omega_n\zeta s + \omega_n^2} z(t) \right) e(t) \quad (11)$$

$$\frac{dk_3(t)}{dt} = \gamma \left(\frac{1}{s^2 + 2\omega_n\zeta s + \omega_n^2} \dot{z}(t) \right) e(t) \quad (12)$$

4 Simulation Results

This section presents the simulation results based on MATLAB/Simulink scheme (Figs. 1, 2 and 3).

From the MATLAB/simulink results above, it's easily to say that the MRAC succeeded by keeping the process model output $z(t)$ following the desired trajectory of the reference input, eliminating the error between the process model and reference model to zero which means keeping the levitated particle by the standing wave at the desired point as well as maintaining the stability of the system.

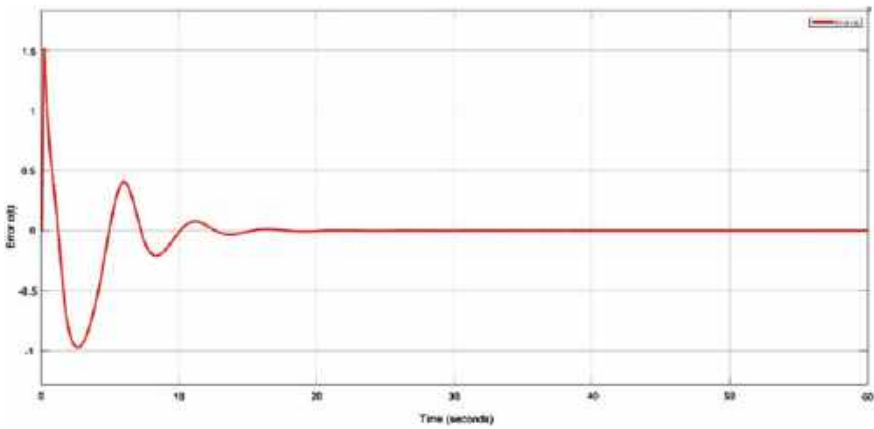


Fig. 1 Error signal $e(t)$

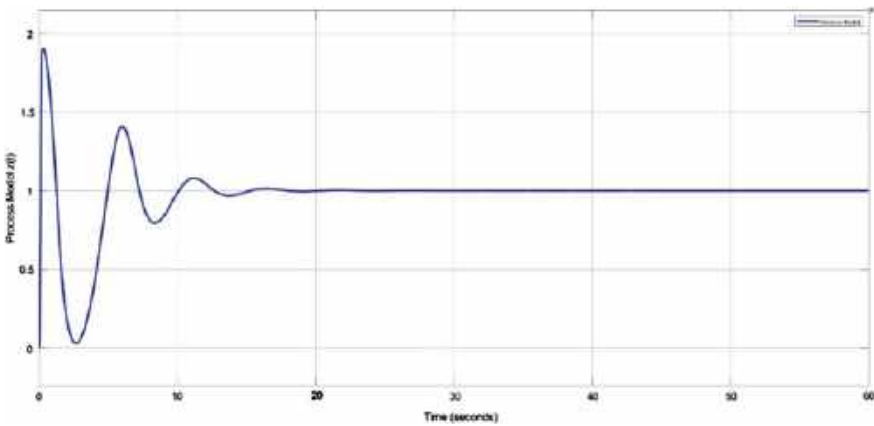


Fig. 2 Process model output $z(t)$

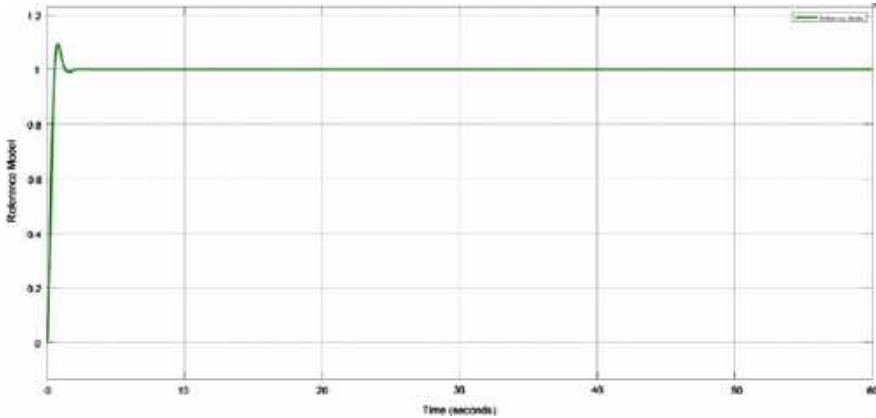


Fig. 3 Reference model output

5 Conclusion

This research article gave an introduction about robotic non-contact manipulation by one of the most effective techniques used in this field which is acoustic levitation based on standing wave. Then, described the importance of implementing and developing new control strategies to acoustic levitation so as to achieve adaptive and reliable control methods that can handle changes in the dynamics of the system and outside influences. Also, we presented MRAC as a new control strategy for controlling the acoustic levitation system based on the standing wave method and it's already achieved the desired performance of it. Hopefully this study will lead to a new advancement by focusing on applying different control approaches to acoustic levitation system in order to improve it.

Acknowledgements This work was supported in part by The fundamental Research Grant Scheme (FRGS) awarded by the Ministry of Higher Education Malaysia FRGS/1/2022/TK07/USM/02/13.

References

1. Andrade MAB, Bernassau AL, Adamowski JC (2016) Acoustic levitation of a large solid sphere. *Appl Phys Lett* 109(4). Art. no. 044101
2. Zang D, Yu Y, Chen Z, Li X, Wu H, Geng X (2017) Acoustic levitation of liquid drops: dynamics, manipulation and phase transitions. *Adv Colloid Interface Sci* 243:77–85
3. Xie WJ, Cao CD, Lü YJ, Hong ZY, Wei B (2006) Acoustic method for levitation of small living animals. *Appl Phys Lett* 89(21). Art. no. 214102
4. Sundvik M, Nieminen HJ, Salmi A, Panula P, Hæggström E (2015) Effects of acoustic levitation on the development of zebrafish, *Danio rerio*, embryos. *Sci Rep* 5(1):1–11
5. Reinhart G, Hoepfner J (2000) Non-contact handling using high-intensity ultrasonics. *CIRP Ann* 49(1):5–8

6. Santesson S, Nilsson S (2004) Airborne chemistry: acoustic levitation in chemical analysis. *Anal Bioanal Chem* 378(7):1704–1709
7. Li H, Wang Y, Li Y, Sun W, Shen Y, Zeng Q (2022) The levitation and driving performance of a contact-free manipulation device actuated by ultrasonic energy. *Int J Mech Sci* 225. Art. no. 107358
8. Andrade MAB, Pérez N, Adamowski JC (2018) Review of progress in acoustic levitation. *Braz J Phys* 48(2):190–213
9. Al-Nuaimi III, Mahyuddin MN, Bachache NK (2022) A non-contact manipulation for robotic applications: a review on acoustic levitation. *IEEE Access* 10:120823–120837. <https://doi.org/10.1109/ACCESS.2022.3222476>
10. Marzo A, Barnes A, Drinkwater BW (2017) TinyLev: a multi-emitter single-axis acoustic levitator. *Rev Sci Instrum* 88(8). Art. no. 085105
11. Wang T, Saffren M, Elleman D (1974) Acoustic chamber for weightless positioning. In: *Proceedings of the 12th aerospace sciences meeting, Washington, DC, USA, Jan 1974*, p 155
12. Andrade MAB, Polychronopoulos S, Memoli G, Marzo A (2019) Experimental investigation of the particle oscillation instability in a single-axis acoustic levitator. *AIP Adv* 9
13. Wang S, Allen JS, Ardekani AM (2017) Unsteady particle motion in an acoustic standing wave field. *Eur J Comput Mech* 26:115–130
14. Fushimi T, Hill TL, Marzo A, Drinkwater BW (2018) Nonlinear trapping stiffness of mid-air single-axis acoustic levitators. *Appl Phys Lett* 113
15. Dumont G (2013) Adaptive control. model reference adaptive control. an overview. In: *department of electrical and computer engineering. University of British Columbia*

Telecommunication Systems and Applications

A Review of Slotted Hollow Pyramidal Absorbers for Microwave Frequency Range



Ala A. Abu Sanad , Mohd Nazri Mahmud , Mohd Fadzil Ain ,
Mohd Azmier Bin Ahmad , Nor Zakiah Binti Yahaya ,
and Zulkifli Mohamad Ariff 

Abstract Electromagnetic (EM) absorbers are an essential component in the design of high-performance radar and communication systems. Hollow pyramidal absorbers are widely used among different types of EM absorbers due to their excellent absorption characteristics. However, improving their absorption performance remains challenging due to the lack of absorbent materials inside the hollow sections. In recent years, creating slots on the surfaces of hollow pyramidal absorbers has emerged as a promising approach to enhance absorption. Although various types of slots have been tried, their reviews are still lacking. This paper reviews the types of slots used to form slotted hollow pyramidal EM absorbers and elaborates on their effects on absorption

A. A. Abu Sanad (✉) · M. N. Mahmud · M. F. Ain
School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Engineering Campus,
14300 Nibong Tebal, Pulau Pinang, Malaysia
e-mail: a.abusanad@student.usm.my

M. N. Mahmud
e-mail: nazrie@usm.my

M. F. Ain
e-mail: eemfadzil@usm.my

M. A. Bin Ahmad
School of Chemical Engineering, Universiti Sains Malaysia, Engineering Campus, 14300 Nibong
Tebal, Pulau Pinang, Malaysia
e-mail: chazmier@usm.my

N. Z. Binti Yahaya
School of Distance Education, Universiti Sains Malaysia, 11800 Nibong Tebal, Pulau Pinang,
Malaysia
e-mail: norzakiah@usm.my

Z. M. Ariff
School of Materials and Mineral Resources Engineering, Universiti Sains Malaysia, Engineering
Campus, 14300 Nibong Tebal, Pulau Pinang, Malaysia
e-mail: zulariff@usm.my

performance. It also suggests new approaches that could be used to improve absorption performance. Overall, this review paper offers valuable insights into the design and development of high-performance slotted hollow pyramidal EM absorbers.

Keywords Hollow pyramidal · Microwave absorbers · Pyramidal absorbers · Type of slots

1 Introduction

The widespread use of electronics and wireless communication systems has increased Electromagnetic Interference (EMI) effects. To address this, there is a need for high-performance, broad-bandwidth microwave absorbers that are lightweight, easy to fabricate, and compatible with various applications. For anechoic chamber applications, absorbers with pyramidal geometry are commonly used due to their excellent reflectivity over a wide frequency range and ability to be optimized for broadband performance [1]. The pyramidal absorber can be solid or hollow in design. Whereas the inner side of the solid design is filled with absorbent material, the inner side of the hollow design is filled with air. Thus, improving the absorption of the hollow design is relatively more challenging. Nevertheless, the hollow pyramidal absorbers offer additional freedom for optimization partly due to the presence of the hollow section. They can be designed to absorb a range of frequencies by varying the dimensions of their hollow sections. Specifically, the height and base width of the pyramid as well as the thickness of the four surfaces can be adjusted to create a resonant cavity that is tuned to absorb a range of frequencies [2]. In addition, creating slots on their surfaces has become popular in recent years to enhance the performance and broaden the bandwidth of hollow pyramidal absorbers [3]. The induced electrical currents that pass through the slot can affect the electric field component that travels across the slot. As a result, the charge distribution around the slot can become non-uniform as more charges are accumulating on one side of the slot than on the other. This non-uniform charge distribution can lead to the formation of an electric field across the slot, which can help to enhance the absorption of the incident radiation. The exact nature of the electric field across the slot depends on various factors, including the design of the absorber and the frequency of the incident radiation [4].

This paper presents a review of the types of slots that have been used in the hollow pyramidal absorber and elaborates on how each type affects absorption performance. Section 2 reviews the slot types, Sect. 3 discusses their relative potentials for enhancing absorption, Sect. 4 suggests future work derived from this review, and Sect. 5 concludes the review.

2 Review of Slot Types in Hollow Pyramidal Absorbers

This section studies the type of slots used in hollow pyramidal absorbers and reports the studies that investigate the size, shape, and orientation of different types of slots that includes rectangular slots, fractal geometries, slot arrays, and triangle slots.

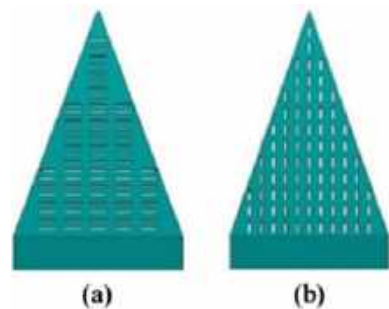
2.1 Rectangular Slots

They consist of multiple rectangular slots, each having a similar length and width, that are etched on the surfaces of the hollow pyramidal absorber. These slots may be added to all walls of the hollow pyramidal absorber or may be added to only specific walls of the absorber, depending on the design requirements and the frequency range of interest.

In [5], identical rectangular slots of size $0.3\text{ cm} \times 2.5\text{ cm}$ have been implemented on a hollow pyramidal absorber with two different orientations which are horizontal and vertical arrangement, as shown in Fig. 1a, b, respectively.

The absorption levels for both designs have been measured over the frequency range of 1–12 GHz. The results suggest that designs with different slot orientations give significantly different absorption performances. In terms of the maximum level of absorption, horizontally oriented slots give a lower value of -35.63 dB compared to the vertically oriented ones with -63.67 dB . Nevertheless, their maximum absorptions occur within the same frequency range of 4–8 GHz (C-band). However, their average levels of absorption are highest in the X-band (8–12 GHz) wherein the absorption level for the vertical orientation averages at -25.22 dB compared to the lower average of -19.26 dB for the horizontal orientation. At the lower frequency range of 1–2 GHz (L-band), however, better absorption is achieved by using horizontal rather than vertical slots. This is because in the horizontal orientation, the longer part of the rectangular opening is located horizontally, thereby giving a wider opening for more EM waves having a longer wavelength to enter the absorber.

Fig. 1 Multi-slot in hollow absorber structure, **a** horizontal slots, and **b** vertical slots [5]



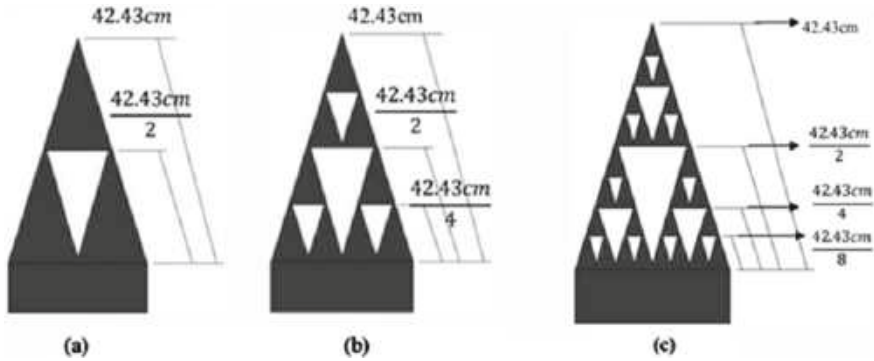


Fig. 2 Three design fractals **a** Design 1, **b** Design 2, and **c** Design 3 [6]

2.2 Fractal Geometries

A fractal slot is a type of slot that is constructed using fractal geometry. Fractals are self-similar patterns that repeat themselves at different scales, and they can be used to create intricate patterns that are highly irregular. The Sierpinski principle is employed to create the fractals for a hollow pyramidal absorber [6]. Three different fractals having different numbers and sizes of triangular slots have been designed as shown in Fig. 2.

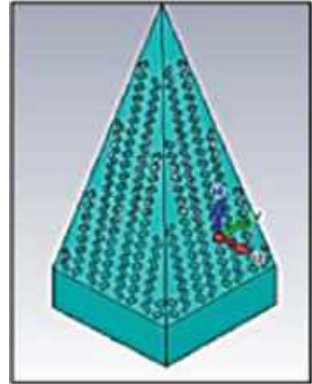
The absorption performances of the three designs are simulated in CST and measured at normal incidence in the range of 8–12 GHz. The results show that fractal design 3, which has the most number of slots and contains the smallest slot size, provides higher absorption compared to designs 1 and 2. Its impedance is more stable than the other two designs, thus giving a higher absorption of -38 dB at frequencies ranging from 11 to 12 GHz. It also achieved a wider bandwidth over the high-frequency range. This higher absorption is due to the fractal pattern of the Sierpinski triangle that creates a highly irregular surface which can increase the surface area of the absorber and improve its absorption efficiency.

2.3 Slot Array

A slot array refers to a pattern of slots that are cut into the walls of the pyramidal structure. In [7], a new slot known as slot radial array has been introduced to enhance the absorption of the hollow pyramidal absorber. A slot radial array forms a circular or radial pattern and includes other shapes, such as semi-circular, annular, and spiral. The slot radial array used in [7] is depicted in Fig. 3.

The effect of adding the slot radial array is tested in the frequency range from 8 to 12 GHz. The results show that the addition improves the reflectivity from -7.7 dB

Fig. 3 Front view of slot radial array absorber



to – 14.7. Adding the slot radial array increases the surface area of the absorber, which increases the amount of EM energy that can be absorbed.

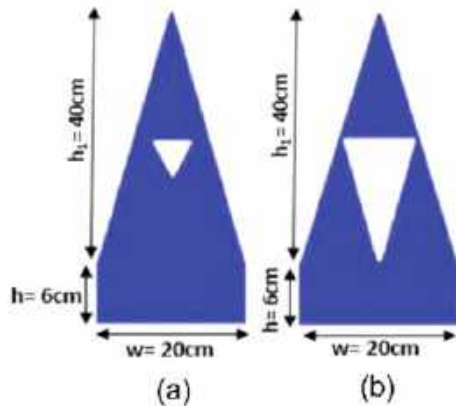
2.4 Triangle Slots

These slots are triangular in shape. They are etched on the surfaces of the hollow pyramidal absorber and are arranged in a specific pattern to achieve a desired level of absorption over a given frequency range.

Two designs with different sizes of inverted triangle slots have been presented in [8], as shown in Fig. 4.

The effect of increasing the slot size on the absorption performance has been tested at a frequency range of 1–12 GHz. It is shown that changing from the smaller to the bigger slot results in the enhancement of absorption performance from – 22.96

Fig. 4 **a** Design 1 a hollow pyramidal absorber with a small triangle slot, and **b** Design 2 a hollow pyramidal absorber with a large triangle slot [8]



to -38.12 dB under the C-band (4–8 GHz). However, in the X-band (8–12 GHz), the absorption of the bigger slot is less than that of the smaller slot. This indicates that a hollow pyramidal absorber with a smaller slot size performed better at a higher frequency. The smaller slot size is also shown to give broader absorption bandwidth.

3 Discussion

The slotted technique, which involves adding slots on the pyramidal absorber surfaces, has shown significant potential in improving the absorption performance of hollow pyramidal absorbers. The shape, size, and orientation can have a significant impact on its performance.

The shape of the slots can affect the propagation and polarization of the incident wave that is absorbed. Different slot shapes can have different polarization selectivity. For example, a circular slot can be effective for absorbing waves with both linear and circular polarization, while a rectangular slot may be better suited for absorbing waves with linear polarization [9].

Furthermore, the slot size can affect the bandwidth of the absorber. Smaller slots can provide a broader absorption bandwidth, while larger slots may provide better absorption at specific frequencies. However, larger slots can also reduce the amount of energy that is absorbed by the absorber, so it is important to find the right balance between slot size and absorption efficiency.

In addition, the orientation of the slots can affect the polarization and directionality of the absorbed waves. For example, if the slots are oriented in a radial pattern, they can effectively absorb waves coming from any direction. If the slots are oriented in a specific direction, they can be more effective at absorbing waves from the coming direction.

4 Future Work

Despite the extensive research conducted on a slotted hollow pyramidal absorber, the effect of varying the spaces between adjacent slots remains an open area for future investigation. Another promising avenue for future work involves the application of the meshing technique to create slots with very small sizes, thus facilitating the investigation into their potential for broadband absorption. Another approach is the use of linearly tapered slots which are used in antenna design. This approach can be researched for the ability to provide more uniform absorption performance across a wider frequency range. Additionally, nonlinearly tapered slots have been suggested as a new alternative, which could further enhance absorption by reducing reflection and improving impedance matching. Investigating the effectiveness of these new approaches represents an important area for future research in the design and optimization of slotted hollow pyramidal absorbers.

5 Conclusion

Several slot types to enhance the performance and increase the effective bandwidth of a hollow pyramidal absorber have been accomplished. The effect of slots' design parameters including the size, shape, and orientation has been discussed. Finally, new approaches have been suggested as future work to enhance the absorption and broaden the bandwidth of a slotted hollow pyramidal absorber.

Acknowledgements This work was funded under Fundamental Research Grant Scheme, FRGS, Ministry of Higher Education (FRGS/1/2020/TK0/USM/03/4).

References

1. Ali Z, Muneer B, Chowdhry BS, Jehangir S, Hyder G (2020) Design of microwave pyramidal absorber for semi anechoic chamber in 1 GHz~20 GHz range. *Int J Wirel Microw Technol* 10(2):22–29. <https://doi.org/10.5815/ijwmt.2020.02.03>
2. Chikhi N, Passarelli A, Andreone A, Masullo MR (2020) Pyramidal metamaterial absorber for mode damping in microwave resonant structures. *Sci Rep* 10(1):1–8. <https://doi.org/10.1038/s41598-020-76433-3>
3. Izzati M et al (2020) Effect of slot array at different angles towards the performance of hollow pyramidal microwave absorber. *Int J Emerg Trends Eng Res* 8(9):6306–6312. <https://doi.org/10.30534/ijeter/2020/224892020>
4. Lu Y, Chen J, Li J, Xu W (2022) A study on the electromagnetic-thermal coupling effect of cross-slot frequency selective surface. *Materials (Basel)* 15(2):640. <https://doi.org/10.3390/MA15020640>
5. Fazin MI et al (2022) Absorption performance of biomass hollow pyramidal microwave absorber using multi-slot array technique. *Indones J Electr Eng Comput Sci* 26(2):895–902. <https://doi.org/10.11591/ijeecs.v26.i2.pp895-902>
6. Yusof AS et al (2017) Slotted triangle on hollow pyramidal microwave absorber characteristics. In: *Proceedings of the 6th IEEE international conference control system, computing and engineering (ICCSCE)*, pp 563–568. <https://doi.org/10.1109/ICCSCE.2016.7893639>
7. Abdullah@Idris H et al (2016) Slot radial array design on hollow pyramidal microwave absorber. *Appl Mech Mater* 850:77–81. <https://doi.org/10.4028/WWW.SCIENTIFIC.NET/AMM.850.77>
8. Asmadi M, Abdulla H, Fazin MI, Razali A, Taib M, Noor N (2020) Absorption study of triangular and rectangular slotted on hollow pyramidal absorber. *ESTEEM Acad J* 16:21–30
9. Dong J, Ding C, Mo J (2020) A low-profile wideband linear-to-circular polarization conversion slot antenna using metasurface. *Materials (Basel)* 13(5):1164. <https://doi.org/10.3390/ma13051164>

The Investigation of Perceptual Speech 5G Wireless Communication Networks



**Tuan Ulfah Uthailah Tuan Mohd Azran, Ahmad Zamani Jusoh,
Ani Liza Asnawi, Khaizuran Abdullah, Md. Rizal Othman,
and Nur Idora Abdul Razak**

Abstract 5G technology has taken over the telecommunication industry and there are a lot of expectations that come with it considering the Quality of Service (QoS). However, the perceptual speech quality in a 5G wireless communication network is not yet been investigated. In communication networks, the end-user's quality has traditionally been measured based on radio link measurements such as the Signal Interference Ratio (SIR), Bit Error Rate (BER), or Frame Error Rate (FER). However, these parameters do not accurately present the perception of the end users. The ultimate measure of perceived speech quality is realized through subjective listening tests, but this is not practical for real-time day-to-day applications. In recent years, objective quality measurement algorithms have been developed to predict subjective quality with considerable accuracy. The automated end-to-end PESQ algorithm is the ideal addition to the system design to investigate the perceptual speech quality in the 5G mobile communication networks. The PESQ algorithm is applied as the objective measure to get the average of the Mean Opinion Score (MOS). MOS is the score to measure the perceptual speech quality ranging from 1 to 5, which is from poor to high quality accordingly. Therefore in this paper, the perceptual speech quality in a 5G communication network will be investigated and analysis of the investigation will be presented. MATLAB Simulink platform will be applied for 5G communication network simulation. From the result of the investigation, the parameters that adopt to improve the perceptual speech quality for the wireless 5G network can be identified and hence the perceptual speech quality can be controlled more effectively.

T. U. U. T. M. Azran · A. Z. Jusoh (✉) · A. L. Asnawi · K. Abdullah
ECE Department, Kulliyah of Engineering, International Islamic University Malaysia (IIUM),
Kuala Lumpur, Malaysia
e-mail: azamani@iium.edu.my

Md. R. Othman
Faculty of Electrical and Electronic Engineering Technology, University Malaysia Pahang (UMP),
Pekan, Malaysia

N. I. A. Razak
Faculty of Electrical and Electronic Engineering, University Technology Mara (UiTM), Shah
Alam, Malaysia

Keywords 5G · Speech quality · Perceptual · PESQ · MOS

1 Background

In general, the definition of speech quality is described as the outcome of the perceived speech assessment to achieve the suitability standard with what is expected. As the technology evolves, the current 5G communication network needs to be able to reach its full potential and capabilities as a successor to the previous mobile technologies. Consequently, the 5G services envisioned to enhance the end-users experience have been the main focus for network operators to improve the conditions concerning the Quality of Service (QoS) parameters such as latency, reliability, and availability [1]. Hence, it is needed to investigate the perceptual speech quality in a 5G wireless communication network.

The end-to-end performance measurements in communication systems especially regarding the transmission link have shown a relatively progressive stage of speech quality analysis. A few of the long-established parameters are the Bit-Error-Rate (BER), Signal to Noise Ratio (SNR), and Frame Error Rate (FER).

Even so, these metrics are non-perceptual parameters irrelevant for expressing the perceived speech quality [2] as it is lacking in terms of accuracy when compared to the perceptual evaluation method that uses MOS (Mean Opinion Score) as the quality scale. Therefore, in this research, the objective of a true speech quality measure is achieved by employing perceptual algorithms to analyze and monitor the degradation of service from the end user's perspective in a 5G wireless mobile network.

2 Background

5G is the fifth-generation mobile technology premiering first in April 2019 led by South Korea, and since then 5G networks have been deployed around the globe, covering about 58 countries based on the latest statistic up until June 2021 [3]. For economic development purposes, it has become a race against time and resources for the telecommunication industry in every country to fully operate the new global wireless standard. In order to accommodate the demand for an exceptionally ideal and sustainable 5G infrastructure, the requirement for perfect execution of 5G simulation is explored in [4]. The implementation of perceptual speech quality assessment is very useful when it comes to outlining the QoS performance of telecommunication networks. The evaluation method mentioned consists of subjective measures and objective measures. The subjective perceptual method uses human subjects to evaluate the degraded speech sample given to them through a listening test which is not practical for real-time quality monitoring [5]. On the other hand, the objective assessment method is computational of complex mathematical models following the international standard to measure the perceived quality of speech signals which can

be categorized into parameter-based and signal-based using the MOS values relative to the subjective tests [6].

There are two types of objective measure which is the intrusive and non-intrusive method. Intrusive methods in objective models comprise both the initial signal and compromised signal of the transmitted speech to estimate the perceived speech quality [7]. The formation of the model such as the deviation value is exerted to measure the MOS estimation of the compared signals [6]. The algorithms that fall under intrusive method are Perceptual Speech Quality Measurement (PSQM), Perceptual Analysis Measurement System (PAMS), Perceptual Evaluation of Speech Quality (PESQ), and the most recently introduced by ITU-T Recommendation P.863, Perceptual Objective Listening Quality (POLQA) [8].

In contrast with intrusive objective quality assessment, the non-intrusive method only uses the processed output signal and compares it with the conceptual framework of the model. Therefore, it is an acceptable method for continuous performance monitoring of telecommunication networks such as congested network traffic during data transmission and system performance degradation [9]. The most recognized objective algorithm under ITU-T recommendations for a non-intrusive approach is the P.563 [10] and the E-model defined by ITU-T Rec. G.107 [11].

PESQ is a conventional model classified as an intrusive method to measure the quality of communication systems for end-to-end narrow-band telephone networks and speech codecs [12]. The common range of the PESQ score is from -0.5 to 4.5 . High correlation (> 0.92) using the PESQ measure proves an overly parallel outcome with subjective listening tests under different scenarios encountered mostly in mobile and voice-over IP applications [5].

Consequently, PESQ is a contemporary model system that makes do in most applications such as the newly established speech coding algorithms, as well as for the analytical interpretation of the tangential quality pertaining to the speech codecs (bit rate, input levels, and channel errors). Hence PESQ is applied in this research.

3 Methodology

This section emphasizes the research methodology and design planning of the entire project to ultimately achieve the true perceptual speech quality of a 5G wireless communication network using the PESQ algorithm. For this project, the PESQ algorithm is selected due to its compatibility and efficiency to perform multiple signals testing. Figure 1 shows the detailed proposed design of the overall system in perceptual speech quality analysis over a 5G mobile network. A total of 25 speech files were used as the input signals for the 5G modeled system. To further observe the effect of the system on the MOS score, SNR values in the Additive White Gaussian Noise (AWGN) channel were varied to correlate with both the MOS and BER values.

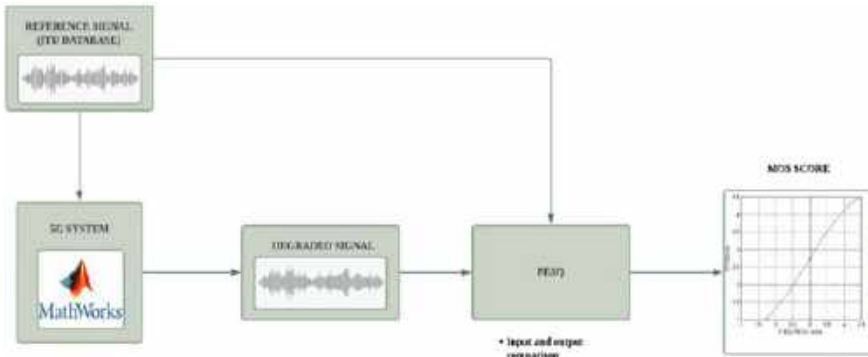


Fig. 1 Block diagram of proposed design of the perceptual speech quality analysis

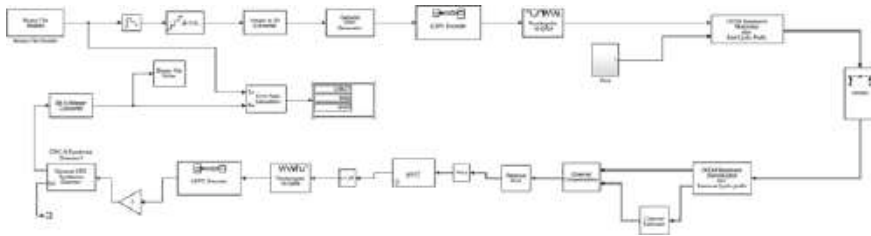


Fig. 2 The 5G digital communication system block diagram

3.1 5G Simulation Model Implementation

MATLAB Simulink was utilized in this project as the simulator to create the model for the 5G communication system. Figure 2 shows the simulation model that adheres to the 3GPP 5G standard specification's characteristics.

3.2 Simulation Parameters

The main parameters used in the simulation are shown in Table 1. It should be mentioned that during the assessment phase, the received and synthesized speech signals at the receiver and transmitter sides will be kept in separate files.

Table 1 Simulation parameters

Modulation	16 QAM
Channel model	AWGN
Power transmit	1 W
Waveform	CP-OFDM
NFFT	2048
Cyclic prefix length	144
Signal to noise ratio (Eb/No)	5–15 dB
Total bits transmitted	2025
Coding	LDPC

3.3 PESQ Simulation

Perceptual Evaluation of Speech Quality (PESQ)—ITU-T Recommendation P.862 (version 2.0—29 November 2005) is used to run the program for this section of the simulation. The PESQ algorithm will be applied to the received speech file as well as the original transmitted file for each simulation to determine the corresponding actual PESQ MOS. This analysis will remove the silent areas of the speech signal output. The perceptual speech quality will be considered good if the acquired MOS score is more than 3.5. The simulation of PESQ needs to be executed using the Microsoft Windows Command Prompt.

4 Results and Analysis

The result of the 5G model simulation and MOS score assessment with PESQ is discussed in detail in this section. The whole block diagram of the transmitter and receiver is designed and simulated in the MATLAB Simulink environment.

For this study, the PESQ MOS values for 25 carefully chosen original speech samples are analyzed. Three distinct SNRs in dB values were used to test every sample. The output’s average, highest, lowest, and standard deviation were then determined and displayed as the final result. Tables 2 and 3 contain the accumulated result of the MOS score for raw MOS and mapped MOS-LQO respectively. MOS is a mapping function to MOS-LQO as outlined in Recommendation ITU-T P.862.1 [6].

The lowest MOS score distribution from both tables is seen at a 5 dB signal-to-noise ratio. This is because the original signal is totally covered up by the high noise level, which is why such low MOS values are obtained. It is observed that the receiver would degrade severely at SNR values less than 15.

The average PESQMOS scores obtained are marginally better than the MOS-LQO in terms of evaluating quality. The voice quality is nevertheless adequate, with a top MOS of 3.361, despite the fact that the expected score of 3.5 and higher was not met.

Table 2 PESQ MOS result

SNR (dB)	Average PESQMOS	Highest score	Lowest score	Standard deviation
5	3.0272	3.271	2.454	0.315
10	3.09404	3.361	2.442	0.314
15	3.20812	3.353	2.506	0.217

Table 3 PESQ MOS-LQO (listening quality objective) result

SNR (dB)	Average MOS-LQO	Highest score	Lowest score	Standard deviation
5	2.87796	3.226	2.08	0.441636
10	2.97444	3.358	2.066	0.442400656
15	3.13632	3.346	2.143	0.307557409

5 Conclusion

In conclusion, the perceptual speech quality for 5G wireless communication network is successfully evaluated using the PESQ objective measure. Based on the result, the average of the PESQ MOS is 3.361 and is slightly less than the expected result of 3.5. This may be due to the designed system that has yet to achieve the standard and its optimum capacity to perform as a perfect model. In the future, this research could be an entry for the service providers to perform reliable network monitoring for 5G services. Based on the developed model to obtain the result, the parameters to be adopted to improve the perceptual speech quality of the 5G wireless communication systems can be identified and hence can be used to control the perceptual speech quality more effectively.

Acknowledgements This paper was part of works conducted under the IIUM-UMP-UITM Sustainable Research Collaboration 2020 grant (SRCG20-041-0041). The authors would also like to acknowledge all support given by the IIUM Research Management Centre through the grant.

References

1. Universitas Mercu Buana, Institute of Electrical and Electronics Engineers. Indonesia Section, and Institute of Electrical and Electronics Engineers. In: Program book of 2017 international conference on broadband communication, wireless sensors and powering (BCWSP 2017). Jakarta, Indonesia, 22–23 Nov 2017
2. ITU-T Recommendation P.861 (1998) Objective quality measurement of telephone band (300–34000 Hz) speech codec
3. Buchholz K (2021) Where 5G technology has been deployed. Statista, 03 Aug 2021
4. Gkonis PK, Trakadas PT, Kaklamani DI (2020) A comprehensive study on simulation techniques for 5G networks: state of the art results, analysis, and future challenges. *Electronics (Switz)* 9(3):468

5. Xu Z, Strake M, Fingscheidt T (2022) Deep noise suppression maximizing non-differentiable PESQ mediated by a non-intrusive PESQNet. *IEEE/ACM Trans Audio Speech Lang Process* 30:1572–1585
6. ITU-T Recommendation P.862 (2001) ITU-T perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs
7. Bulucea CA, WSEAS (Organization) (2009) Recent advances in circuits, systems, electronics, control and signal processing: proceedings of the 8th WSEAS international conference on circuits, systems, electronics, control and signal processing (CSECS' 09): Puerto De La Cruz, Tenerife, Canary Islands, Spain, 14–16 Dec 2009/monograph. WSEAS
8. Voznak M, Non-intrusive speech quality assessment in simplified E-model [Online]. Available: <http://voznak.eu>
9. P 863, ITU-T Rec. P.863.1 (06/2019) Application guide for recommendation ITU-T P.863 [Online]. Available: <http://handle.itu.int/11.1002/1000/11>
10. ITU-T Recommendation P.563 (2004) Single-ended method for objective speech quality assessment in narrow-band telephony applications
11. Recommendation ITU-T G.107 (2005) The E-model: a computational model for use in transmission planning
12. Alexander R (2006) Speech quality of VoIP. Wiley, England

Comparison of Different Data Detection Methods in Orthogonal Frequency Division Multiplexing (OFDM) System



Nur Qamarina Muhammad Adnan, Aeizal Azman Abdul Wahab, Syed Sahal Nazli Alhady, and Wan Amir Fuad Wajdi

Abstract Orthogonal Frequency Division Multiplexing (OFDM) is a multicarrier modulation technique that has been widely used in current technologies due to its many advantages. However, OFDM suffers from high Peak to Average Power Ratio (PAPR) that can distort OFDM's good performance. To combat this problem, Selective Mapping (SLM) was used by many researchers but later discover that SLM requires side information (SI) to be transmitted to the receiver and waste the data rate. Blind receiver was proposed so that data can be recovered without transmission of SI. This paper studies two of the most famously used blind detectors which is Maximum Likelihood (ML) and Viterbi Algorithm (VA) and compares their performance.

Keywords OFDM · PAPR · Selective mapping · Maximum likelihood · Viterbi algorithm

1 Introduction

Due to rapid innovation of technologies, a reliable and error-free system is needed to keep up with high data rates required by current high-end applications such as Internet of Things (IoT) and virtual reality (VR) [1, 2]. Orthogonal Frequency Division Multiplexing (OFDM) is seen as one of the systems that can fulfill the demand

N. Q. M. Adnan · A. A. A. Wahab (✉) · S. S. N. Alhady · W. A. F. Wajdi
School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Jln Transkrian-Bukit Panchor, 14300 Nibong Tebal, Pulau Pinang, Malaysia
e-mail: aeizal@usm.my

N. Q. M. Adnan
e-mail: qamarashvin96@student.usm.my

S. S. N. Alhady
e-mail: sahal@usm.my

W. A. F. Wajdi
e-mail: wafw_othman@usm.my

of providing high data transmission as it has been utilized by many standards like wireless local-area network (WLAN), digital audio/video broadcasting (DAB/DVB), long-term evolution (LTE), 4G and 5G [1]. OFDM is a promising multicarrier modulation technique that secures against frequency selective fading since the system divides its high data stream into many parallel low data streams with short bandwidth and specific frequencies [1–3]. These low data streams are modulated with orthogonal subcarriers and help the system to combat the issues of multipath fading and time delay [1, 3]. To avoid interferences such as inter-carrier interference (ICI) and inter-symbol interference (ISI), OFDM includes guard band between the subcarriers [2]. Orthogonality of the subcarriers allow OFDM to accomplish high data rates and better Bit Error Rate (BER) performances since they are overlapping each other and save the bandwidth used [2].

Though been used widely in many applications, OFDM has its own downsides and the biggest one is high Peak to Average Power Ratio (PAPR) [3]. This flaw could ruin the chances of OFDM to be employed in future technologies. PAPR occurs after the signals have been converted to time domain by Inverse Fast Fourier Transform (IFFT) and added up together causing peak amplitude to be much higher than the average amplitude [3, 4]. The high peak drives the power amplifier of the system into non-linear region causing non-linear distortions or out-of-band radiation and ruin the orthogonality of the subcarriers [3, 5]. PAPR can be reduced by using complex digital to analog converter (DAC) and high-power amplifier (HPA), but it will increase the system's cost and power consumption [3], so it is important to find a way to reduce effectively. PAPR reduction techniques can be divided into distortion and distortion less method [4]. In this paper, one of the distortions less methods called selective mapping (SLM) is used as PAPR reduction technique. In SLM, multiple identical signals are generated and multiplied to several different phase sequences. The signal with the lowest PAPR value is then chosen to be transmitted to the receiver [6, 7]. This method can help to reduce the occurrence probability of the peak power signal [8]. The index of the chosen signal needs to be transmitted too as side information (SI) to recover the signal at the receiver, but it causes the data rate to be wasted [9, 10]. To solve this issue, blind SLM (BSLM) has been proposed so that data can be recovered at the receiver without the transmission of SI [10–12]. In this paper, different blind detections, which is Maximum Likelihood Estimation (ML) and Viterbi Algorithm (VA) will be studied and analyzed.

2 Methodology

OFDM uses IFFT and Fast Fourier Transform (FFT) in the transmitter and receiver respectively where the transmitted signal from both algorithms can be represented by the following equations [1]

$$x_i = \frac{1}{N} \sum_{n=0}^{N-1} X_n e^{j2\pi ni/N} \quad (1)$$

$$X_n = \frac{1}{N} \sum_{i=0}^{N-1} x_i e^{-j2\pi ni/N} \quad (2)$$

PAPR happens when N subcarriers are added together in phase and cause a very high peak power with the power of N times than usual. PAPR can be calculated by the following equation [4, 5]

$$\text{PAPR}\{x(t)\} = \frac{\max_{0 \leq n \leq N-1} (|x(t)|^2)}{E(|x(t)|^2)} \quad (3)$$

It can be expressed in dB . To evaluate the performance of PAPR reduction method, complementary cumulative distribution function (CCDF) is used and defined as the probability that the PAPR exceeds the threshold value as can be seen by the following equation [5, 6]

$$\text{CCDF}(\text{PAPR}_0) = \Pr(\text{PAPR} > \text{PAPR}_0) \quad (4)$$

where $\Pr(\cdot)$ is the probability factor and PAPR_0 represents the threshold value. After signal has been mapped into the constellation diagram according to the modulation scheme chosen, elementwise multiplication will be performed between the modulated signal and U different phase sequences. In this paper, QPSK modulation scheme is used as it can transmit two bits per symbol. After IFFT has been performed on the signals, the one with the lowest PAPR value is chosen to be transmitted. At the receiver, data detection is performed once the received signal has been converted back to frequency domain by FFT. ML is based on exhaustive search [11] estimates the data received by calculating the minimum Euclidean distance of received symbols with the constellation diagram since correct de-mapping and incorrect de-mapping results in different symbol vector [10] using the following equation [12]

$$\tilde{m} = \arg \min_{\substack{u=0 \sim U-1 \\ c \in \Psi_{mod}}} \left(\epsilon = \sum_{n=0}^{N-1} |\Phi_u^*(n)d(n) - c|^2 \right) \quad (5)$$

where U is the total number of phase sequences, $\Phi_u^*(n)$ is the conjugated phase sequence, $d(n)$ is the received signal and Ψ_{mod} is the original constellation diagram. VA also utilizes Eq. (5) for blind detection. ϵ at time $t = T$, $0 \leq T \leq N - 1$ can be represented by the following equation [13]

$$\epsilon(T) = \epsilon(T - 1) + \frac{1}{N} \min_{\substack{u=0 \sim U-1 \\ c \in \Psi_{mod}}} |\Phi_u^*(n)d(n) - c|^2 \quad (6)$$

The first term of above equation represents the path metric at time t and state g , $\epsilon(g_t)$ and the second term represents the branch metric from state g' at time t to state g at time $t + 1$ and become $\zeta(g'_t \rightarrow g_{t+1})$. Path metric always begin at $t = -1$ and $\epsilon(g_{-1}) = 0$. The following equation describe how path metric entering the next state is chosen

$$\epsilon(g_{t+1}) = \min_{g'_t=0 \sim G_{max}} (\epsilon(g'_t) + \zeta(g'_t \rightarrow g_{t+1})) \tag{7}$$

Equations (6) and (7) are repeated until $t = N - 1$. To determine the surviving path metric that provide the optimal state number and optimal sequence, backward computation need to be done using the following equations

$$g_{N-1,opt} = \arg \min_{g_{N-1}=0 \sim G_{max}} \epsilon(g_{N-1}) \tag{8}$$

$$g_{t',opt} = \arg \min_{g_{t'}=0 \sim G_{max}} (\epsilon(g_{t'}) + \zeta(g_{t'} \rightarrow g_{t',opt})) \tag{9}$$

where $t' = N - 2, N - 3, \dots, 0$. Figure 1 and Table 1 show the flowchart of the system with blind detectors and the parameters used in this study respectively.

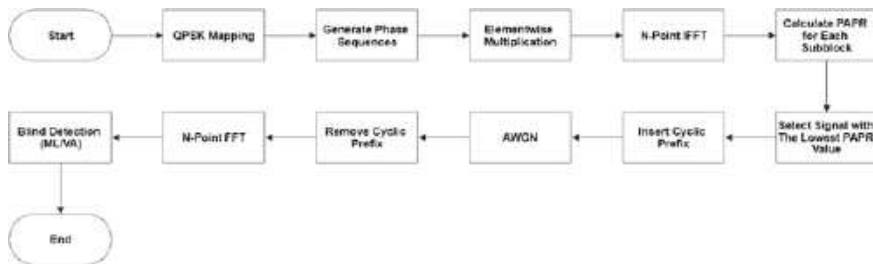


Fig. 1 Flowchart of the system with blind detectors

Table 1 Parameters used

Parameters	Value
SNR	40 dB
Frames	100
IFFT block size	1024
Channel length	10
Cyclic prefix length	9
Fading type	Frequency selective fading
Fade variance	0.5
Decoding delay	20
Phase sequences	4

3 Result and Discussion

Phase sequences are randomly generated as $U \in \{-1, 1\}$ for SLM and QPSK mapping was used for blind estimation. Figure 2 shows BER performance of ML and VA. From 0 to 20 dB, both blind detectors provide almost the same result but as SNR increases, VA starts to perform better than ML especially when SNR > 30 dB. It can be assumed that since VA considers the entire sequence of received bits rather than treating each bit in isolation, it requires much more computational than ML and results in better BER. The Viterbi algorithm explore all possible paths through the trellis diagram while ML typically involves an exhaustive search over all possible codewords, which can be impractical for long codes. Both blind receivers perform better as SNR increases. Table 2 shows the number of real multiplications and real additions needed by both ML and VA. N_b is the number of branches used in VA where the maximum is $(G_{max})^2 \times N$. When $N = 1024$, $U = 4$ and $N_b(\max) = 16,384$ with $G_{max} = 4$, ML require 36,868 real multiplication and 52,228 real additions while VA require 622,592 real multiplication and 839,685 real additions. VA requires much more computation hence it is more time consuming than ML. Computational complexity of ML and VA gets higher as the number of N and U increases. Figure 3 shows the PAPR performance for both blind detectors. Both ML and VA achieve almost the same CCDF value as OFDM's system without blind receiver proving that even by using blind detectors, the performance of SLM is not affected and eventually improves the complexity of the technique and saves the data rate from being wasted.

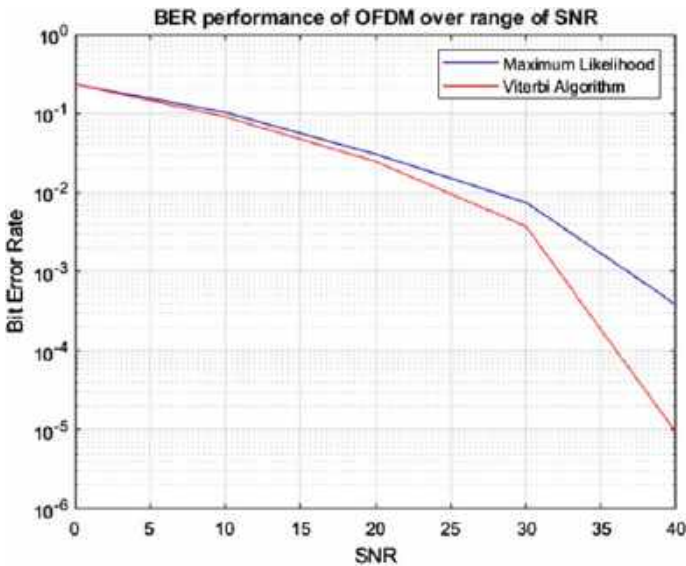


Fig. 2 BER performance of ML and VA

Table 2 Computational complexity of ML and VA

	Number of real value multiplications	Number of real value additions
Maximum likelihood	$U \times (36N + 1)$	$U \times (51N + 1)$
Viterbi algorithm	$38 \times N_b$	$(51 \times N_b) + UN$

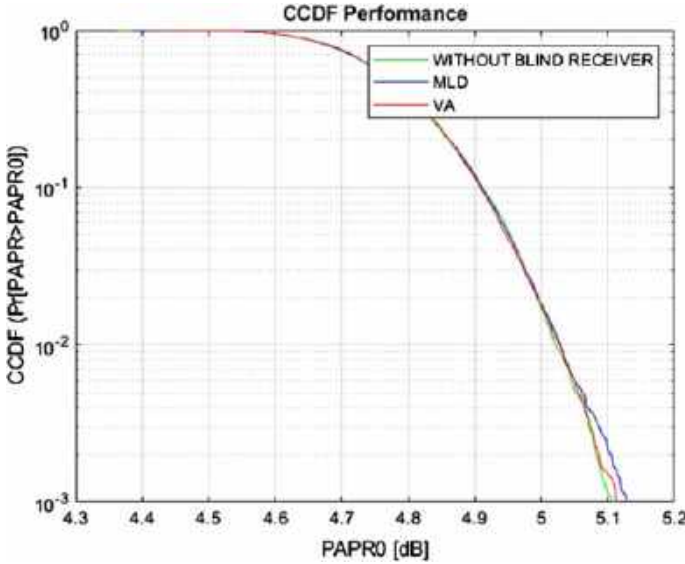


Fig. 3 PAPR performance of ML and VA

4 Conclusion

The performance of ML and VA was compared in terms of BER, PAPR, and computational complexity. When SNR is more than 30 dB, VA performs better in terms of BER and has more computing complexity than ML. Nearly identical PAPR performance is achieved by both blind detectors, ranging from 4.9 to 5 dB. Only one set of parameters was applied in this study. The performance of these characteristics on blind receivers can be compared by extending them further.

Acknowledgements This project was supported by “Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme” with project code FRGS/1/2020/TKO/USM/02/21.

References

1. Kansal L, Berra S, Mounir M, Miglani R, Dinis R, Rabie K (2022) Performance analysis of massive MIMO-OFDM system incorporated with various transforms for image communication in 5G systems. *Electronics* 11(4):621
2. Abdullahi MB (2020) Evaluation of BER performance of FFT-OFDM for wireless communication networks. *Int J Eng Technol Manag Res* 6(2):14–26
3. Lavanya P, Satyanarayana P, Ahmad A (2019) Suitability of OFDM in 5G waveform—a review. *Orient J Comput Sci Technol* 12(3):66–75
4. Zannat OR, Khatun T, Rahman M, Raza S, Huda MN (2021) PAPR reduction of OFDM signal by scrutiny of BER assessment and SPS-SLM method via AWGN channel. *IEEE Xplore*, 01 Jan 2021
5. Song X et al (2022) SCMA-OFDM PON based on chaotic-SLM-PTS algorithms with degraded PAPR for improving network security. *Opt Lett* 47(20):5293–5296
6. Niwareeba R, Cox MA, Cheng L (2021) Low complexity hybrid SLM for PAPR mitigation for ACO OFDM. *ICT Express* 8(1)
7. Agarwal P, Singh AP, Shukla MK (2020) Performance analysis of Fiedler-SLM mechanism in OFDM for PAPR reduction, 21 Feb 2020
8. Karthika J, Thenmozhi G, Rajkumar M (2021) PAPR reduction of MIMO-OFDM system with reduced computational complexity SLM scheme. *Mater Today Proc* 37(2):2563–2566
9. Gupta P, Thethi HP (2020) Performance investigations and PAPR reduction analysis using very efficient and optimized amended SLM algorithm for wireless communication OFDM system. *Wireless Pers Commun* 115(1):103–128
10. Sa'd AHY, Saad HHY, Abd AA (2021) Maximal minimum hamming distance codes for embedding SI in a data based BSLM scheme for PAPR reduction in OFDM. *IJTech Int J Technol* 19
11. Arbi T, Geller B (2019) Joint BER optimization and blind PAPR reduction of OFDM systems with signal space diversity. *IEEE Commun Lett* 23(10):1866–1870
12. Boonkajay A, Adachi F (2018) Modified blind selected mapping for OFDM/single-carrier signal transmission. *IEEE Xplore*, 01 Nov 2018
13. Boonkajay A, Adachi F (2017) 2-step phase rotation estimation for low-PAPR signal transmission using blind selected mapping. *IEEE Xplore*, 01 Oct 2017

A Wideband 3 dB T-Shaped Stubs-Loaded Coupler for Millimeter-Wave (mm-Wave) Beamforming Network Towards Fifth Generation (5G) Technology



Nazleen Syahira Mohd Suhaimi and Nor Muzlifah Mahyuddin

Abstract This article presents a design of T-shaped stubs-loaded coupler for beamforming networks towards the fifth generation (5G) technology. The low-cost, lightweight and compact size of the wideband 3-dB T-shaped stubs-loaded coupler is proposed at 26 GHz. The symmetrical rectangular-shaped slots and T-shaped stubs are introduced in order to achieve the desired coupling values of 3 dB and 90° phase difference between output ports. Four symmetrical circular slots are loaded to retain the flatness of the differential phase characteristics between 24.75 and 27.25 GHz.

Keywords Stubs-loaded coupler · Wideband · Millimeter wave · Beamforming network · Fifth-generation (5G)

1 Introduction

Quadrature hybrid coupler is one of the passive components in myriad applications such as microwave mixers, amplifier circuits and antenna arrays. Although the conventional 3-dB quadrature hybrid coupler has advantages in its simplicity and ease of fabrication, it suffers from narrow return loss bandwidth owing to the parasitic effects at the T- or cross-junctions and unwanted coupling among the adjacent low-impedance lines at high operating frequencies. In order to improve return losses, two arrays of via holes at the adjacent sides of a C-band directional coupler using substrate integrated waveguide (SIW) technique are proposed in [1]. However, very small spacing between the via holes need to be taken into account in this work to reduce the leakage loss. The coupled lines of the Lange coupler are folded as reported in [2]. The smooth swept bends are utilized in the folded Lange coupler to

N. S. M. Suhaimi · N. M. Mahyuddin (✉)

School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Engineering Campus, Nibong Tebal, Penang, Malaysia

e-mail: eemnmuzlifah@usm.my

reduce the parasitic capacitance. In order to connect the coupled lines of the folded Lange coupler, bonding wires are used to act as crossovers. However, the crossovers will provide additional inductance. According to another technique in [3], a pair of asymmetrical cross-slot patch is developed diagonally on the square patch and the horizontal slot is built with a stepped-impedance shape at the end of the slot. Although the proposed coupler cannot be described by using the closed-form microstrip line theory owing to the complicated field distribution of the etched slots, the length and width of the slots are varied and optimized to reroute the electric currents around the cross-slots. The topology as proposed in [4] is appropriate to be used in beamforming networks such as Butler matrix, Nolen matrix and Blass matrix, whereby four ports are placed at the center of the square patch sides.

In this article, the objective of this work is to construct a 3 dB T-shaped stubs-loaded coupler for 5G beamforming network that provides low amplitude imbalance performances towards 5G technology which requires wideband operation of mm-wave frequency. The Rogers RT/duroid 5880 substrate with dielectric constant of 2.2 and substrate thickness of 0.254 mm is implemented in the designs.

2 Implementation of the Proposed 3 dB T-Shaped Stubs-Loaded Patch Coupler

Physical layouts of the initial and proposed designs of 3 dB T-shaped stubs-loaded patch couplers are denoted in Fig. 1a, b. The T-shaped stubs-loaded coupler is proposed by constructing a pair of symmetrical cross slots on the square patch.

A pair of symmetrical cross slots with similar lengths (L_1 and L_2) and widths (W_1 and W_2) is developed diagonally on the square patch to perturb the original patch resonator [5]. Whilst, a pair of symmetrical rectangular-shaped slots with a dimension of $W_3 \times L_3$ is built at the end of the second cross-slot on the square patch coupler. The width W_3 of a pair of symmetrical rectangular-shaped slots on the square patch coupler should be greater than the width, W_2 of the second cross slot. By changing the slots' lengths (L_1 , L_2 and L_3) and slots' widths (W_1 , W_2 and W_3), the electric currents reroute around the cross-slots, enabling a wideband coupling. Meanwhile, four symmetrical circular slots are inductively loaded on the proposed square patch couplers to remain flatness of the differential phase characteristics. Each corner of the square patch couplers is chamfered to reduce the excess capacitance and unwanted reflection, while four ports are placed in the middle of the square patch sides.

The T-shaped stubs are loaded at every side of the microstrip lines nearby the square patch of the couplers in order to improve the s-parameters and differential phase characteristics performance within the desired frequency range. The optimization and parametric study on slots' lengths of L_1 , L_2 and L_3 as well as T-shaped stubs' lengths of L_4 are executed to satisfy a good agreement with the specifications of the couplers across the designated frequency range between 25.75 and 26.25 GHz.

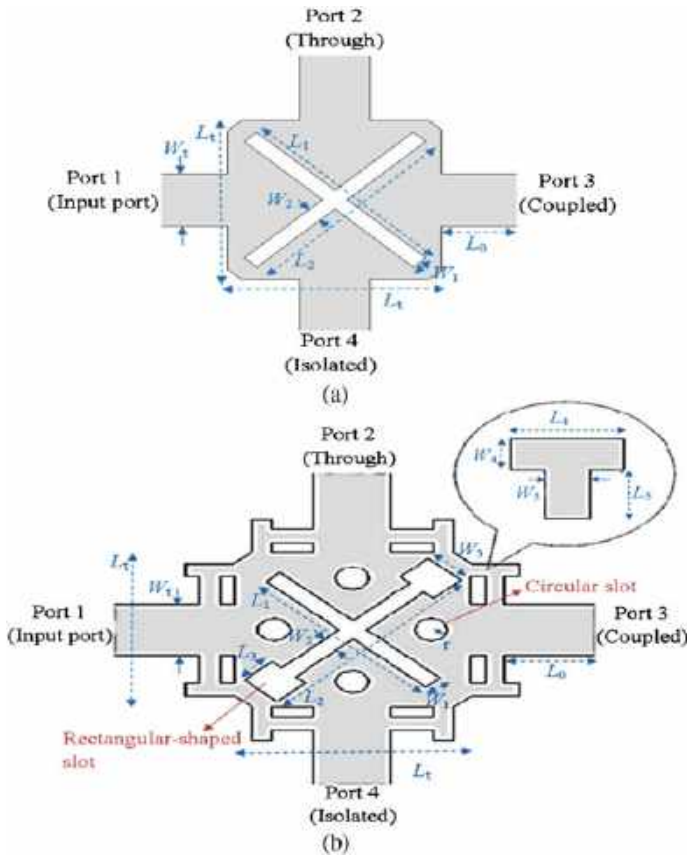


Fig. 1 Physical layout; **a** initial design; **b** proposed design of the T-shaped stubs-loaded coupler

The lengths of the microstrip feed lines, L_0 at each port of the proposed couplers are extended to $0.62 \lambda_g$ to accommodate a direct connection of receptacle connectors as well as no short circuit between center pin of the connectors and microstrip feed lines. According to a detailed parametric study, cross slots' lengths L_1 , L_2 and L_3 as well as T-shaped stubs' length of L_4 control the coupling values and the transmission phase. The optimized lengths' L_1 , L_2 , L_3 and L_4 dimensions of the proposed 3 dB T-shaped stubs-loaded couplers are listed in Table 1. The dimensions of L_0 , L_t , L_5 , r , W_3 and W_t are remained constant at 5.29 mm, 2.40 mm, 0.43 mm, 0.17 mm, 0.47 and 0.79 mm, respectively. While, the widths of W_1 , W_2 , W_4 and W_5 are fixed to be 0.20 mm due to limited specification of the fabrication.

Table 1 Length dimensions (L_1 , L_2 , L_3 and L_4) of the initial and proposed 3 dB T-shaped stubs-loaded couplers

Length	Dimension (mm)		
	$C = 1.26$ dB	$C = 1.76$ dB	$C = 3$ dB
L_1	2.46	2.47	2.30
L_2	2.46	2.56	2.65
L_3	0.52	0.40	0.40
L_4	0.54	0.50	0.50

* C = coupler coupling coefficient

3 Performance Results

The performance of the initial and proposed T-shaped stubs-loaded coupler are verified by simulation and measurement, which includes the S-parameters and phase differences. The proposed T-shaped stubs-loaded couplers as depicted in Fig. 1b is validated by conducting the measurement using a vector network analyzer (VNA). As observed in Fig. 2, the simulated return loss, S_{11} performs less than or equal to -9.15 dB between 24.75 and 27.25 GHz. Meanwhile, the simulated S_{21} and S_{41} are -3 dB ± 4.69 and less than -7.7 dB, respectively. The simulated S_{21} does not perform well on this initial design. Whilst, the simulated S_{31} is -3 ± 0.71 dB across the designated frequency range. However, the simulation result of S_{31} is acceptable, which depicts a small deviation of ± 0.71 dB between 24.75 and 27.25 GHz. The phase difference between output ports is $90^\circ \pm 26.9^\circ$ across 24.75 and 27.25 GHz. The resonant frequencies for the minimum amplitude peak of the simulated S_{11} , the maximum amplitude peak of the simulated S_{31} and phase difference are shifted to 29 GHz.

As seen in Fig. 3, the proposed 3 dB coupler with loaded T-shaped stubs has good simulated performances in terms of return loss and isolation, which are better than

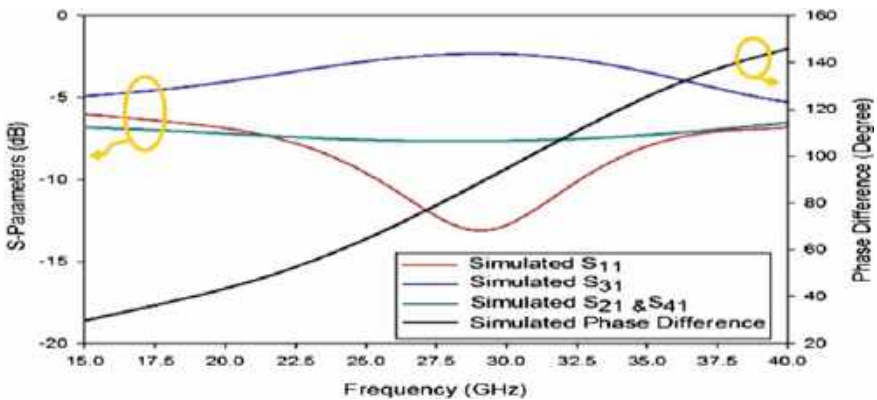


Fig. 2 Simulation results of S-parameters and phase difference for the initial design of 3 dB coupler

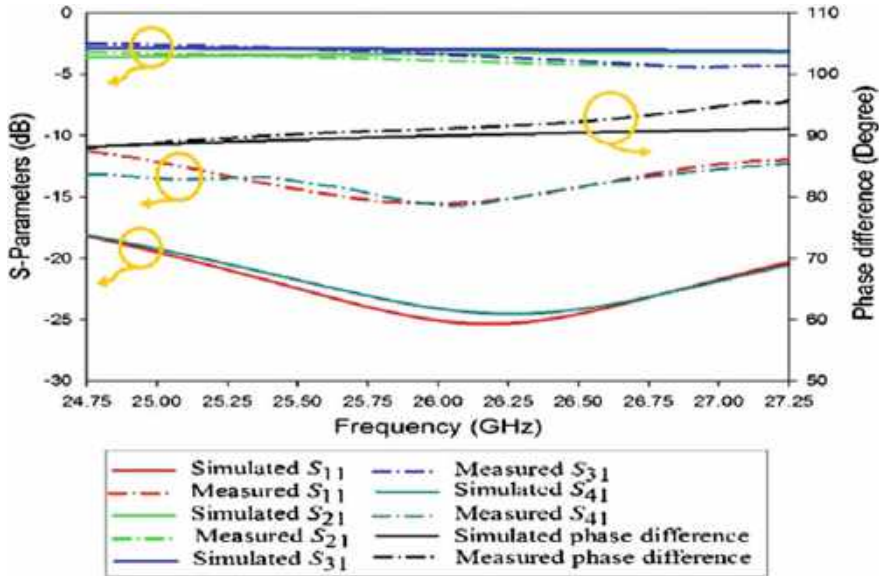


Fig. 3 Simulation and measurement results of S-parameters and phase difference for the proposed 3 dB coupler with loaded T-shaped stubs

17 dB between 24.75 and 27.25 GHz as depicted in Fig. 3. The measured return loss and isolation are better than 10 dB and 12 dB across the operated frequency interval between 24.75 GHz and 27.25 GHz, respectively. The amplitude imbalance of the simulated S_{21} is ± 0.73 dB, whereas the amplitude imbalance of the measured S_{21} is ± 1 dB, respectively. Meanwhile, the simulated and measured S_{31} are -3 dB ± -0.14 dB and -3 dB ± -1 dB, separately. The simulated and measured phase imbalances of the phase difference between port 2 and port 3 are $90^\circ \pm 2.34^\circ$ and $90^\circ \pm 5^\circ$, individually.

4 Conclusion

According to the results obtained in this work, it can be seen that the simulation and measurement results of the proposed T-shaped stub-loaded 3 dB coupler are acceptable for the mm-wave frequency band between 24.75 and 27.25 GHz with a wide frequency bandwidth of 2.5 GHz. By comparing to the initial 3 dB patch coupler design, the return losses (S_{11} and S_{41}) in the proposed coupler have been improved to be better than 10 dB, whereas the amplitude imbalance of the insertion loss (S_{21}) has been reduced by 3.96 dB within the designated frequency range. Moreover, the phase imbalance has been reduced by 21.9° in the proposed coupler compared to its initial design. Therefore, the presence of T-shaped stubs offers the better performances and

low amplitude imbalances of return loss, insertion loss and output phase difference which meets the agreement with the objective of this work.

Acknowledgements The authors would like to thank Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme (FRGS) with Project Code FRGS/1/2022/TK07/USM/02/14 for permitting them to carry out this research.

References

1. Veadesh B, Aswin S, Shambavi K (2017) Design and analysis of C-band SIW directional coupler. In: International conference on microelectronic devices, circuits and systems (ICMDCS)
2. Xu Q, Wang YE (2012) Design and realization of compact folded lange coupler. In: IEEE MTT-S international microwave symposium digest, pp 1–3
3. Sun S, Zhu L (2010) Miniaturised patch hybrid couplers using asymmetrically loaded cross slots. *IET Microw Antennas Propag* 4(9):1427
4. Fonseca NJG (2009) Printed S-band 4×4 Nolen matrix for multiple beam antenna applications. *IEEE Trans Antennas Propag* 57(6):1673–1678
5. Jing X, Sun S (2014) Design of impedance transforming 90 degree patch hybrid couplers. In: Proceedings of Asia-Pacific microwave conference, vol 1, pp 25–27

Mathematical Modelling of Artificial Magnetic Conductor Backed Antenna On-Chip Using Response Surface Method for 28 GHz Application



Ahmadu Girgiri , Mohd Fadzil Bn Ain , Abdullahi S. B. Mohammed , and Muhammad Bello Abdullahi 

Abstract In recent years, innovation in mobile and wireless communication systems has required low-profile, high speed and miniaturized antennas. One of the promising of such antennas is Antenna-on-chip (AoC) technology. Research into AoC technology is increasing as the advent of low-power, low-profile and high-speed technologies, including handheld devices, wireless sensor networks (WSNs) and internet-of-things (IoTs) increase. This study proposes a modelling of Artificial Magnetic Conductor (AMC)-backed AoC. The model offers an improved gain and radiation efficiency focusing on three independent variables: the gap between the patch (P_g), patch width (P_w), and the substrate height (h_s). The model optimization was realized using Response Surface Method (RSM) and developed two output response equations optimized operating frequency (F_0) and the AMC reflection phase (R_p) using a Reduced Quartic Model (RQM). Optimized values for the variables P_g , P_w , and h_s were 0.225 mm, 0.2 mm and 0.25 mm with a gain of 2.89 dBi, 1.83 dBi and efficiency of 59% and 46% for the design with and without DRP-AMC. Moreover, realizes a bandwidth of 1.24 GHz and 1.2 GHz, respectively.

Keywords Response surface method · Artificial magnetic conductor · Patch width · Patch gap · Substrate height · Dual patch

1 Introduction

For decades, an antenna has been a significant technology for wireless communication systems, triggering EM/RF engineers to develop low-power, miniature, and integrated antennas suitable for recent and future technologies. Most integrated antennas were designed and fabricated on high-speed, low-profile material, including CMOS

A. Girgiri (✉) · M. F. B. Ain · A. S. B. Mohammed · M. B. Abdullahi
School of Electrical and Electronic Engineering, University Sains Malaysia, 14300 Nibong Tebal, Malaysia
e-mail: ahmadu.g@student.usm.my

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
N. S. Ahmad et al. (eds.), *Proceedings of the 12th International Conference on Robotics, Vision, Signal Processing and Power Applications*, Lecture Notes in Electrical Engineering 1123, https://doi.org/10.1007/978-981-99-9005-4_40

and its derivatives. This manuscript presents the development of dual rectangular patches (DRP) AMC unit cells for enhancing the radiation efficiency and gain of Antenna-on-chip (AoC). The DRP AMC-backed AoC was modelled using Response Surface Model (RSM) for achieving an improved radiation efficiency and gain of an AoC with independent variables; reflection phase at the upper and lower frequency of the AMC surface and the resonance frequency at zero-degree reflection phase [1]. Similarly, the patch gap, P_g , patch width P_w , and substrate height h_s were considered independent variables; likewise, an AMC reflection phase and the frequency from the AMC-backed surface are the dependent variables. The independent and the response variables were approximated using a Reduced Quartic non-linear regression model with significance R-Square and p -values. An analysis of variance (ANOVA) was used to observe the model's significance with and without a DRP-AMC-backed surface.

2 Artificial Magnetic Conductor Structure

2.1 Design and Configuration

AMC is a high-impedance conducting surface designed and acts as an electromagnetic shield. It is designed to reflect incidence waves from confining into the lossy substrate and reduces the effect of surface waves [2, 3]. Besides, other techniques like localized backside etching (LBE) improve AoC gain, yet it is attributed to high cost of production [4]. However, AMC-inspired AoC realizes high gain and radiation efficiency [5–7]. In [8], a dual dipole patch AMC is proposed and performs considerably compared to the single-dipole structured AMC. In this study, a DRP-AMC-backed AoC realized an increased gain of 36.68%, an efficiency of 22.03% and BW of 3.23% over full patch AMC-backed AoC. The modelled DRP AMC-based offers better performance than full-patch AMC-backed AoC. However, substantial variations in capacitance (C) and inductance (L) were realized due to slight change with values of P_g , P_w and h_s , causing significant changes in frequency and the reflection phase. Similarly, the value of L and C can be determined from given Eqs. (1) and (2), where P_w and P_g are the patch width and the patch gap, h_s is the substrate height.

$$L = 4\pi 10^{-7} \mu h_s \quad (1)$$

$$C = 2.82 \times 10^{-12} (1 + \epsilon) \cosh \frac{(2P_w + P_g)}{P_g} \quad (2)$$

$$F = \frac{1}{2\pi \sqrt{LC}} \quad (3)$$

$$BW = \frac{1}{\eta\sqrt{\frac{L}{C}}} \tag{4}$$

3 Response Surface-Based Modeling of AMC Structure

RSM has been used for simple and multiple non-linear regression models, including quadratic, cubic, quartic, and similar regression models. This study used a reduced quartic model (RQM) to realize the response equations by eliminating the non-significant high-order variables for the experiment. Equally employed to estimate the actual and coded values [9]. Equation (5) shows the general equation for the RQM-based non-linear regression function.

$$y = \beta_0 + \beta_1z_1 + \beta_2z_2 + \beta_3z_3 + \beta_1z_1^2 + \beta_2z_2^2 + \beta_3z_3^2 + \dots + \beta_1z_1z_2^2 + \beta_3z_3^2z_1^2 + \varepsilon \tag{5}$$

where β_0 is an intercept of the polynomial equation, β_1 , β_2 , and β_3 are coefficients of 4th-order regression, and y is the model response. Thus, a general model equation is represented by:

$$y = \beta X + \varepsilon \tag{6}$$

3.1 Selection of Variables

In AoC-integrated technologies, the device’s performance is based on the significant variables liable to the change in the performance of the devices. This model constitutes three independent and two dependent variables. The experimental variables are P_w , P_g and h_s of the DRP-AMC unit cells (Table 1).

Table 1 3^k level experiment values for dd-AMC modelling

Independent factor (mm)	Real levels		
	- 1	0	1
Patch spacing, P_g	0.1	0.250	0.4
Patch width, P_w	0.05	0.225	0.4
Substrate height, h_s	0.05	0.275	0.5

3.2 Design of Experiment

As mentioned earlier, this work employs the response surface method for determining the effect of the predicting variables. The variables have a significant impact on the performance of the DRP-AMC-backed structure. The actual values were normalized and coded within a range starting from -1 to $+1$ through 0 according to Eq. (1).

$$X_c = 2X_c + \frac{(X_{max} + X_{min})}{(X_{min} - X_{max})} \tag{7}$$

The parameter X_{max} and X_{min} are the lower and upper boundaries of the actual size of the rectangular-shaped patch. X_a represents the actual size of the independent variable X_c , which defines the coded value of the independent variable. In this experiment, the actual, max, min and coded values for P_g, P_w and h_s were transformed by substituting in Eq. (7).

4 Result and Discussions

The results from the simulations indicated that a slight increase in the values of the variables, patch gap (P_g), patch width (P_w) and substrate height (h_s) of the proposed DRP-AMC unit cells, led to a significant effect on the performance of the gain and efficiency of the AoC. Table 2 contains the output developed equations obtained from the reduced-quartic model using DES. The equations are the resonance frequency at zero-degree refraction and reflection phase at upper and lower degrees.

Where the coefficients A, B and C from the equation represent; the patch gap P_g , patch-width P_w , and substrate height h_s . Moreover, the probability P -value achieved from the analysis of variance (ANOVA) is less than 0.05 , as in Table 3 for both equations, which indicates that the model is significant and equally suitable for determining the optimum performance of the DRP-AMC unit cell of an AOC at 28 GHz application.

Table 2 Modelled equations of reflection phase and frequency of modified DRP-AMC

Response	Equation	R-square
<i>Frequency at 0°</i>		
Refraction phase (F_0)	$26.28 + 0.1070A + 0.1041B - 1.65C + 0.0225AB - 0.0175AC - 0.2350BC + 1.07A^2 + 1.33B^2 + 0.5256C^2 + 0.4259A^2B + 2.64A^2C + 0.1855AB^2 - 2.11A^2B^2$	0.9192
Reflection phase (R_P)	$33.84 + 0.1100A + 0.09511B + 22.86C - 1.51AB - 2.56AC - 3.80BC - 1.47A^2 - 1.22B^2 - 0.9334C^2 + 21.60A^2B + 6.33AB^2 + AC^2 - 59.40A^2B^2$	0.9990

Table 3 ANOVA result for AMC reflection phase and frequency

Source	DF	Sum of square	Mean of square	F-value	P-value	
<i>Response 1: frequency at the reflection phase</i>						
<i>Model-reduced quartic</i>	13	34,104.19	2623.4	448.25	< 0.0001	Significant
<i>A-patch gap</i>	1	0.0684	0.0684	0.0117	0.9174	
<i>B-patch width</i>	1	0.0512	0.0512	0.0087	0.9285	
<i>C-substrate height</i>	1	0.8712	0.8712	0.1489	0.7129	
<i>AB</i>	1	18.3	18.3	3.13	0.1274	
<i>AC</i>	1	52.43	52.43	8.96	0.0242	
<i>BC</i>	1	115.52	115.52	19.74	0.0044	
<i>A²</i>	1	34.78	34.78	5.94	0.0506	
<i>B²</i>	1	25.46	25.46	4.35	0.0821	
<i>C²</i>	1	10.45	10.45	1.79	0.2298	
<i>A²B</i>	1	1546.74	1546.74	264.28	< 0.0001	
<i>A²C</i>	1	4874.83	4874.83	832.94	< 0.0001	
<i>AB²</i>	1	132.78	132.78	22.69	0.0031	
<i>A²B²</i>	1	10,989.66	10,989.66	1877.75	< 0.0001	
<i>Residual</i>	6	35.12	5.85			
<i>Lack of fit</i>	1	32.4	32.4	59.69	0.0006	
<i>Pure error</i>	5	2.71	0.5428			
<i>Cor total</i>	19	34,139.31				
R²		0.9192				
<i>Response 2: reflection phase</i>						
<i>Model-reduced quartic</i>	13	56.58	2623.4	5.25	0.0259	Significant
<i>A-patch gap</i>	1	0.0648	0.0684	0.0781	0.7892	
<i>B-patch width</i>	1	0.0613	0.0512	0.0739	0.7949	
<i>C-substrate height</i>	1	15.35	0.8712	18.5	0.0051	
<i>AB</i>	1	0.004	18.3	0.0049	0.9466	
<i>AC</i>	1	0.0025	52.43	0.003	0.9584	
<i>BC</i>	1	0.4418	115.52	0.5327	0.493	
<i>A²</i>	1	13.83	34.78	16.68	0.0065	
<i>B²</i>	1	21.23	25.46	25.59	0.0023	
<i>C²</i>	1	3.32	10.45	4	0.0925	
<i>A²B</i>	1	0.6012	1546.74	0.725	0.4272	
<i>A²C</i>	1	23.13	4874.83	27.89	0.0019	

(continued)

Table 3 (continued)

Source	DF	Sum of square	Mean of square	F-value	P-value	
AB^2	1	0.114	132.78	0.1375	0.7236	
A^2B^2	1	14.25	10,989.66	17.19	< 0.006	
<i>Residual</i>	6	4.98	5.85			
<i>Lack of fit</i>	1	0.0221	32.4	0.0223	0.8872	
<i>Pure error</i>	5	4.95	0.5428			
<i>Cor total</i>	19	61.56				
R^2		0.9990				

5 Conclusion

AoC is a low-profile, low-powered, footprint-size antenna; its significance in integrated and emerging systems remains considerable. This study developed a mathematical model for evaluating the effect of spacing between a pair of patches Pg , patch width P_w and substrate height hs on resonance frequency and reflection phase of DRP- AMC backed AoC. The developed model provides two responses by optimizing the variables $Pg = 0.225$ mm, $P_w = 0.2$ mm and $hs = 0.25$ mm. A 4th-order reduced quartic model (RQM) non-linear regression function was adopted for the optimization. CST suit 3D EM software was used for the design and simulation, and a design expert software (DES) was employed to model the result from RSM. Then, model fitness was verified by the analysis of variance ANOVA in cognizant with R-Square values for the responses.

References

1. Bean D, Venkataraman J (2020) Gain enhancement of on-chip antenna at 60 GHz using an artificial magnetic conductor. In: 2020 IEEE international symposium on antennas and propagation and North American radio science meeting. IEEECONF 2020—proceedings, pp 1423–1424. <https://doi.org/10.1109/IEEECONF35879.2020.9330242>
2. Nafe M, Syed A, Shamim A (2017) Gain-enhanced on-chip folded dipole antenna utilizing artificial magnetic conductor at 94 GHz. IEEE Antennas Wirel Propag Lett 16(c):2844–2847. <https://doi.org/10.1109/LAWP.2017.2749308>
3. Ahmad WA, Kucharski M, Di Serio A, Ng HJ, Waldschmidt C, Kissinger D (2019) Planar highly efficient high-gain 165 GHz on-chip antennas for integrated radar sensors. IEEE Antennas Wirel Propag Lett 18(11):2429–2433. <https://doi.org/10.1109/LAWP.2019.2940110>
4. Barakat A, Allam A, Pokharel RK, Elsadek H, El-Sayed M, Yoshida K (2013) Compact size high gain AoC using rectangular AMC in CMOS for 60 GHz millimeter wave applications. In: IEEE MTT-S international microwave symposium digest, vol 1, pp 1–3. <https://doi.org/10.1109/MWSYM.2013.6697475>
5. Khan MS, Tahir FA, Meredov A, Shamim A, Cheema HM (2019) A W-band EBG-backed double-rhomboid bowtie-slot on-chip antenna. IEEE Antennas Wirel Propag Lett 18(5):1046–1050. <https://doi.org/10.1109/LAWP.2019.2908891>

6. Huang HT, Yuan HT, Zhang XH, Hu ZF, Luo GQ (2016) A circular ring-shape monopole on-chip antenna with artificial magnetic conductor. In: Asia-Pacific microwave conference. Proceedings, APMC, vol 2, pp 3–5. <https://doi.org/10.1109/APMC.2015.7413137>
7. Cheema HM, Khalid F, Shamim A (2021) Antenna-on-chip: design, challenges, and opportunities. Artech House London [Online]. Available: <https://ebs.pub/qdownload/antenna-on-chip-design-challenges-and-opportunities-antennas-1608078183-9781608078189.html>
8. Ta SX, Park I (2013) Design of miniaturized dual-band artificial magnetic conductor with easy control of second/first resonant frequency ratio. J Electromagn Eng Sci 13(2):104–112. <https://doi.org/10.5515/jkiees.2013.13.2.104>
9. Isafiq M, Shayfull ZS, Nasir M, Rashidi, Fathullah M, Noriman NZ (2016) Shrinkage analysis on thick plate part using response surface methodology (RSM). In: MATEC web of conferences, vol 78, pp 0–7. <https://doi.org/10.1051/mateconf/20167801084>

A Method Combining Compressive Sensing-Based Method of Moment and LU Decomposition for Solving Monostatic RCS



Yalan Gao, Muhammad Firdaus Akbar, Jagadheswaran Rajendran,
and Ghassan Nihad Jawad

Abstract The direct method for solving the matrix equations in the method of moments offers significant advantages in monostatic scattering problems. In this paper, the traditional compressive sensing-based method of moments is transformed into a direct method, and a fast solution to the monostatic RCS is achieved by constructing a low-dimensional reduced matrix equation and LU decomposition technique. Since only part of the impedance matrix and the excitation vector are involved in the calculation, the filling and solving times in the proposed method are significantly reduced. The numerical simulation is performed using the proposed method and the traditional characteristic mode basis function method, and the results demonstrate that the proposed method can provide higher efficiency.

Keywords Characteristic basis function · Characteristic mode · Compressive sensing · Monostatic scattering

Y. Gao · M. F. Akbar (✉)

School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Engineering Campus,
14300 Nibong Tebal, Penang, Malaysia
e-mail: firdaus.akbar@usm.my

Y. Gao

e-mail: gaoyalan1212@student.usm.my

Y. Gao

School of Informatics and Engineering, Suzhou University, Suzhou, China

J. Rajendran

Collaborative Microelectronics Design Excellence Centre (CEDEC), Universiti Sains Malaysia,
11900 Bayan Lepas, Malaysia

e-mail: jaga.rajendran@usm.my

G. N. Jawad

Department of Electronics and Communication Engineering, University of Baghdad,
Baghdad 10071, Iraq

e-mail: ghassan.n.jawad@ieee.org

1 Introduction

The method of moments (MoM) [1] is one of the most efficient numerical methods for solving electromagnetic scattering problems. It discretizes the electromagnetic field integral equation into a full rank matrix equation. The solution of the matrix equation is usually performed by iterative [2] or direct methods [3]. However, for monostatic scattering problems, the matrix equation needs to be solved repeatedly for excitation sources with different incidence angles. Obviously, the iterative method demands a complete iterative procedure for each excitation, which limits the efficiency. And, for electrically large objects, the iterative method may also suffer from convergence difficulty [4]. In contrast, direct methods, such as the LU decomposition, can avoid the convergence difficulty problem, but its high computational complexity is not acceptable.

As an effective direct method, the characteristic basis function method (CBFM) [5] allows the matrix equations to be controlled to an acceptable scale by blocking techniques. The low-dimensional reduced matrices it constructs can be LU-decomposed and stored in advance. The matrix after decomposition is repeatedly called for solving the induced currents under different excitations, thus substantially improving the efficiency of monostatic scattering problem. However, the CBFM must generate a set of excitation-independent characteristic basis functions (CBFs) for the construction of the reduction matrix. Singular value decomposition techniques can be utilised in CBFM to derive excitation-independent CBFs. However, the efficiency is limited by the large number of additional excitations set. Considering that the characteristic modes (CMs) [6] are independent of the incident excitation, a characteristic mode basis function method (CMBFM) [7, 8] is proposed in recent years, which takes the CMs instead of the CBFs to improve efficiency significantly.

On the other hand, the compressive sensing-based MoM proposed in recent years can significantly compress the impedance matrix and the excitation vector thus reducing the computational complexity. In this technique, both the CBFs and the CMs can be used as sparse basis to construct a compressive sensing (CS) model [9–11]. However, the reconstruction algorithms used in CS-MoM, such as generalized orthogonal matching pursuit (gOMP) algorithm [12], are iterative methods which does not facilitate solving monostatic problems.

This paper presents a new method for fast solving the monostatic scattering problem, which combines the CS-MoM and the LU decomposition method. In the proposed method, the reconstruction algorithm in CS-MoM is transformed into a direct method, that is, the gOMP is reduced to a least squares (LS) procedure. A reduced matrix equation similar to that in CMBFM with low dimensions is then constructed. Since the CMs is used as the sparse basis, the reduced matrix is constructed independent of the excitation and does not need to be repeated in the monostatic problem. The proposed method significantly reduces the calculation time compared to CMBFM.

2 Theory of the Proposed Method

In the MoM, the general form of the matrix equation generated is as follow:

$$\mathbf{Z}\mathbf{I} = \mathbf{V}(\theta) \quad (1)$$

where \mathbf{Z} is the impedance matrix of size $N \times N$, N is the number of unknowns; \mathbf{I} and \mathbf{V} denote the current coefficient vector and the excitation vector, respectively. θ is the angle of incidence. For the electrically large problem, the dimension of the matrix equation is very large and difficult to be solved by the direct method. Thus, the blocking technique is used to divide the object into acceptably smaller blocks. Assuming that the object is divided into m blocks, Eq. (1) can be rewritten as the following blocking form:

$$\begin{bmatrix} \mathbf{Z}_{11} & \mathbf{Z}_{12} & \cdots & \mathbf{Z}_{1m} \\ \mathbf{Z}_{21} & \mathbf{Z}_{22} & \cdots & \mathbf{Z}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Z}_{m1} & \mathbf{Z}_{m2} & \cdots & \mathbf{Z}_{mm} \end{bmatrix} \begin{bmatrix} \mathbf{I}_1 \\ \mathbf{I}_2 \\ \vdots \\ \mathbf{I}_m \end{bmatrix} = \begin{bmatrix} \mathbf{V}_1(\theta) \\ \mathbf{V}_2(\theta) \\ \vdots \\ \mathbf{V}_m(\theta) \end{bmatrix} \quad (2)$$

where \mathbf{Z}_{ii} is the self-impedance on block i and \mathbf{Z}_{ij} represents the mutual impedance between block i and block j ($i, j = 1, 2, 3, \dots, m; i \neq j$). \mathbf{I} and \mathbf{V} are the current coefficient and the excitation vector on block i respectively. The CMs on each block can be derived from the extended self-impedance matrix \mathbf{Z}_{ii}^e as follow:

$$\mathbf{X}_{ii}\mathbf{J}_i = \lambda\mathbf{R}_{ii}\mathbf{J}_i \quad (3)$$

where, λ and \mathbf{J}_i represent the eigenvalue and its corresponding CM, respectively. \mathbf{R}_{ii} and \mathbf{X}_{ii} are the real and imaginary parts of \mathbf{Z}_{ii}^e respectively. The CMs basis functions on each block are obtained after selecting based on the modal significance (MS) and removing the extended part. Combination of CMs on all blocks as sparse basis \mathbf{J} :

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{J}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{J}_m \end{bmatrix} \quad (4)$$

$$\mathbf{J}_i = [\mathbf{J}_i^1 \mathbf{J}_i^2 \cdots \mathbf{J}_i^{K_i}] \quad (5)$$

The current coefficient vector \mathbf{I} can be expressed by \mathbf{J} as:

$$\mathbf{I} = \mathbf{J}\mathbf{a} \quad (6)$$

where \mathbf{a} is the sparse projection. According to the CS-MoM, the M rows of \mathbf{Z} and \mathbf{V} are uniformly extracted at a fixed step s as the measurement matrix $\tilde{\mathbf{Z}}$ and the measurement value vector $\tilde{\mathbf{V}}$, thus Eq. (1) can be transformed into a CS calculation model as follow:

$$\tilde{\mathbf{Z}}\mathbf{J}\mathbf{a} = \Theta\mathbf{a} = \tilde{\mathbf{V}}(\theta) \quad (7)$$

where $\Theta = \tilde{\mathbf{Z}}\mathbf{J}$ is called the sensing matrix. In conventional CS-MoM, the gOMP is used to solve for \mathbf{a} . However, in practice, lower order CMs correspond to larger sparse coefficients, and it is not necessary to use gOMP to iteratively find larger sparse coefficients when truncating a certain lower order CMs as sparse basis. Therefore, Eq. (7) is an overdetermined system, rather than an underdetermined equation as in traditional CS-MoM. Obviously, the least squares method can be employed to efficiently solve Eq. (7). To solve the monostatic scattering problem efficiently, we divide the least squares method into two steps. First, a low-dimensional reduced matrix is constructed as follow:

$$\Theta^H\Theta\mathbf{a} = \Theta^H\mathbf{V}(\theta) \quad (8)$$

where Θ^H is the conjugate transpose matrix of Θ . Letting $\mathbf{Z}^R = \Theta^H\Theta$ and $\mathbf{V}^R(\theta) = \Theta^H\mathbf{V}(\theta)$, Eq. (8) can be rewritten as:

$$\mathbf{Z}^R\mathbf{a} = \mathbf{V}^R(\theta) \quad (9)$$

where, \mathbf{Z}^R is referred to as the reduced impedance of size $K \times K$ and $\mathbf{V}^R(\theta)$ is referred to as the reduced excitation vector, K is the number of selected CMs on all blocks. The second step is solving Eq. (9) using the LU decomposition method.

From the above theory it is clear that \mathbf{Z}^R is independent of the angle of excitation. Therefore, \mathbf{Z}^R can be LU decomposed and stored in advance. For different angles θ , only the decomposed matrix is used repeatedly in the calculation, which can significantly improve the efficiency of solving monostatic problem.

3 Numerical Results

In this section, a perfect electric conductor cylinder model with radius 0.2 m and height 1 m is used for simulation calculations.

A set of plane wave with frequency of 1.2 GHz are set as the incident excitation. The model is meshed into 9558 triangular surface elements, and 14,337 unknowns are obtained using the Rao-Wilton-Glisson functions discrete electric field integral equation. The conventional MoM, CMBFM and the proposed method are used to calculate the monostatic RCS of the cylinder, respectively, in which the object is divided into 16 small blocks and CMs with $MS > 0.001$ are selected as sparse basis,

generating a total of 2062 CMs. The size of the measurement matrix $\tilde{\mathbf{Z}}$ obtained by extracting with step $s = 3$ is 4779×4779 , and the size of the final constructed reduction matrix \mathbf{Z}^R is 2062×2062 . The results of three methods are plotted in Fig. 1. As can be seen, the results of the proposed method agree well with those of the traditional MoM and CMBFM. The proposed method can provide excellent accuracy.

The simulation time comparisons are given in Table 1. It can be seen that the proposed method reduces the time by 47% in filling the impedance matrix and the excitation vector. This is due to the fact that the proposed method does not fill the entire impedance matrix and excitation vector, but only the extracted part. Moreover, both the constructing reduced matrix and solving time are significantly reduced. The total time of the proposed method is reduced by 43%. This demonstrates that the proposed method can provide higher efficiency.

Fig. 1 The monostatic RCS of the cylinder

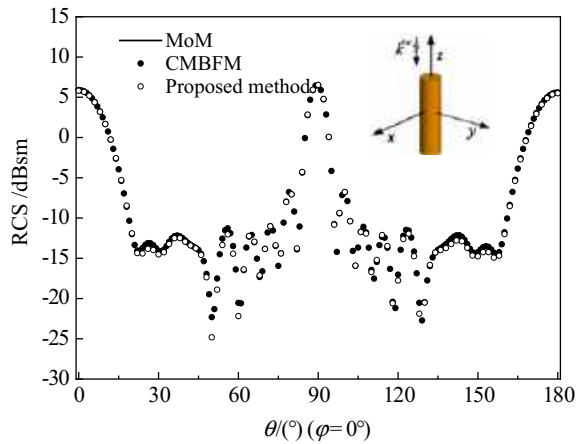


Table 1 Comparison of simulation times

Method	Filling impedance matrix and excitation time (s)	Generating CMs time (s)	Constructing reduced matrix time (s)	Solving current time (s)	Total time (s)
CMBFM	197.2	14.6	23.6	35.7	271.3
Proposed method	103.9	14.6	8.1	28.2	154.9

4 Conclusion

In this study, a novel method is proposed for fast solving monostatic RCS, which combines CS-MoM and LU decomposition. The proposed method transforms the traditional CS-MoM from an iterative solution to a direct solution. Specifically, the reconstruction algorithm gOMP is simplified to least squares, and a low-dimensional reduced matrix equation is constructed which can be solved by LU decomposition. Benefiting from the incomplete filling of the impedance matrix and the excitation vector, the proposed method can significantly improve the efficiency of solving monostatic scattering problems.

Acknowledgements This work was funded and supported by a Universiti Sains Malaysia (USM), Bridging GRA Grant with Project No: 304/PELECT/6316607.

References

1. Harrington RF (1987) The method of moments in electromagnetics. *J Electromagn Waves Appl* 1:181–200. <https://doi.org/10.1163/156939387X00018>
2. Cao X, Chen M, Qi Q, Liu X, Wang D (2022) An improved GMRES method for solving electromagnetic scattering problems by MoM. *IEEE Trans Antennas Propag* 70(11):10751–10757. <https://doi.org/10.1109/TAP.2022.3187610>
3. Gibson WC (2020) Efficient solution of electromagnetic scattering problems using multilevel adaptive cross approximation and LU factorization. *IEEE Trans Antennas Propag* 68(5):3815–3823. <https://doi.org/10.1109/TAP.2019.2963619>
4. Wang Z, Mu J, Lin H, Nie W (2019) New reduced matrix construction accelerated iterative solution of characteristic basis function method. *Acta Phys Sin* 68(17):26–32. <https://doi.org/10.7498/aps.68.20190572>
5. Lucente E, Monorchio A, Mittra R (2008) An iteration-free MoM approach based on excitation independent characteristic basis functions for solving large multiscale electromagnetic scattering problems. *IEEE Trans Antennas Propag* 56:999–1007. <https://doi.org/10.1109/TAP.2008.919166>
6. Lau BK, Capek M, Hassan AM (2022) Characteristic modes: progress, overview, and emerging topics. *IEEE Antennas Propag Mag* 64(2):14–22. <https://doi.org/10.1109/MAP.2022.3145719>
7. Wang P, Wang Z, Sun Y, Nie W (2022) A characteristic mode basis function method for solving wide-angle electromagnetic scattering problems. *J Electromagnet Waves Appl* 36:1968–1979. <https://doi.org/10.1080/09205071.2022.2051211>
8. Wang Z, Guo F, Nie W, Sun Y, Wang P (2023) Principal component analysis accelerated the iterative convergence of the characteristic mode basis function method for analyzing electromagnetic scattering problems. *Prog Electromagn Res M* 117:129–138. <https://doi.org/10.2528/PIERM23041504>
9. Wang Z, Nie W, Lin H (2020) Characteristic basis functions enhanced compressive sensing for solving the bistatic scattering problems of three-dimensional targets. *Microw Opt Technol Lett* 62:3132–3138. <https://doi.org/10.1002/mop.32432>
10. Wang Z, Wang P, Sun Y, Nie W (2022) Fast analysis of bistatic scattering problems for three-dimensional objects using compressive sensing and characteristic modes. *IEEE Antennas Wirel Propag Lett* 21:1817–1821. <https://doi.org/10.1109/LAWP.2022.3181602>

11. Wang Z, Li C, Sun Y, Nie W, Wang P, Lin H (2022) A novel method for rapidly solving wideband RCS by combining UCBFM and compressive sensing. *Prog Electromagn Res C* 124:33–42. <https://doi.org/10.2528/PIERC22072102>
12. Fu T, Zong Z, Yin X (2022) Generalized orthogonal matching pursuit with singular value decomposition. *IEEE Geosci Remote Sens Lett* 19:1–5. <https://doi.org/10.1109/LGRS.2021.3086492>

Microwave Non-destructive Testing Using K-Medoids Clustering Algorithm



Tan Shin Yee, Muhammad Firdaus Akbar, Nor Azlin Ghazali, Ghassan Nihad Jawad, and Nawaf H. M. M. Shrifan

Abstract Turbine blades' metal substrates are often coated with thermal barrier coatings made of composites, notably ceramics. Insulation defects, which might lead to a catastrophic turbine failure, must be detected by routine non-destructive testing. The microwave non-destructive testing has limited spatial imaging, complicating the defect evaluation. This research proposes a unique approach for delamination detection based on microwave non-destructive testing and k-medoids clustering. Using a double-ridged waveguide with 101 frequency points between 18 and 40 GHz, a standard ceramic coating sample is scanned. The k-medoids clustering technique reliably detects and sizes ceramic insulation delamination at each evaluated site. This finding demonstrates the k-medoids clustering method's capability of detecting delamination with 95.3% accuracy.

Keywords Insulation · Delamination · Clustering algorithm · Microwave non-destructive testing

T. S. Yee · M. F. Akbar (✉) · N. A. Ghazali
School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Engineering Campus,
14300 Nibong Tebal, Penang, Malaysia
e-mail: firdaus.akbar@usm.my

G. N. Jawad
Department of Electronics and Communication Engineering, University of Baghdad,
Baghdad 10071, Iraq

N. H. M. M. Shrifan
Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Sungai
Long Campus, Bandar Sungai Long Cheras, 43000 Kajang, Selangor, Malaysia

1 Introduction

It is a common practice to utilize ceramic coatings on turbine blades, vanes, and combustors as their resistance to corrosion and high temperatures. The lifespan of the ceramic is lowered by debonding and delamination. For this reason, non-destructive testing (NDT) must be precise and efficient to boost the system's stability and reliability [1].

Owing to field penetration constraints when analyzing dielectric materials, NDT methods such as eddy current and ultrasonic methods struggle to discover defects beneath ceramic coatings [1].

The microwave NDT approach is a potentially useful technique for analyzing the defects under ceramic coatings [2]. Microwave signals can interact with the interior structure of a dielectric. Microwave NDT using Open-Ended Rectangular Waveguide (OERW) is a technique that is effective for the under-coating examination of delamination [3]. Using a Vector Network Analyzer (VNA), changes in resonant frequency, microwave reflection coefficient in magnitude, and phase are utilized to image defects.

However, there are constraints on OERW microwave NDT applications for ceramic coatings, such as the inability to provide high-quality spatial imaging and erroneous prediction of defect size. Improved defect detection and the ability to discriminate the areas with and without defects are both achieved by post-processing methods [4]. Furthermore, double-ridged waveguides can be utilized to improve the detection process due to their wide bandwidth and small aperture.

A double-ridged waveguide based on the k-medoids clustering algorithm is proposed in this paper for analyzing delamination [5]. The proposed approach pinpoints and evaluates the defects. The ceramic sample is swept over by 101 frequency points in a double-ridged waveguide operating at 18–40 GHz. Inverse Fast Fourier Transform (IFFT) converts frequency domain signals to time domain signals. In this case, by using Principal Component Analysis (PCA), 101-time steps are condensed down to three PCA components. After that, the three PCA components are classified into delamination and delamination-free regions using the k-medoids approach for imaging underlying defects.

2 Theoretical Approach

2.1 Double-Ridged Waveguide

In this study, a double-ridged waveguide is used rather than a rectangular waveguide since it has a larger measuring bandwidth [6]. To improve the spatial resolution, it is better to use a probe with a small aperture and broad bandwidth.

2.2 *K-Medoids Clustering Algorithm*

When it comes to classifying observed data into distinct groups, the k-medoids clustering algorithm is an unsupervised machine-learning approach worth looking into. Clustering using the k-medoids algorithm identifies the representative objects and pairs them with the selected similar objects. For optimal clustering, it drastically lowers the objective function, J_m , shown in (1) [5].

$$J_m = \sum_{i=1}^k \sum_{x \in S_i} d(x, c(i))^2 \quad (1)$$

where d stands for the object that best exemplifies the certain centroid. $c(i)$ denotes the specific cluster, k is the number of clusters, and S represents the total number of data points.

3 Methodology

3.1 *Macor Sample*

A Macor sample, which is a ceramic-based coating, is used to resemble Thermal Barrier Coatings (TBCs). In this study, a Macor sample that was machine-delaminated is modeled. The test sample is a perfect electric conductor (PEC), insulated with non-porous (0% porosity) and low-loss glass ceramic (Macor) with a relative permittivity of 5.67 at 8.5 GHz.

Figure 1 and Table 1 depict the test sample with delamination depth and size. The machine-delaminated dimensions are 20×25 mm, 15×15 mm, 10×10 mm, and 5×5 mm, with depths of 1 to 2 mm. The metal sheet represents the machining defect surface, and it is used to exhibit insulation-metal delamination.

3.2 *Inspection Technique*

A standard WRD180 double-ridged waveguide is utilized as the microwave sensor. Raster scanning is carried out using an XYZ positioner with a step size of 1 mm in both x- and y-direction. The probe is set at a distance of 1 mm from the sample under evaluation. Figure 2 depicts the arrangement of the inspection system. A 3D matrix $S(m, n, f)$ is used to store the complex coefficient, position, and frequency of the sample, where m and n indicate the inspection location and f represent the operating frequency point index, which ranges from 1 to 101.

Fig. 1 The sample under test with delamination's size and location

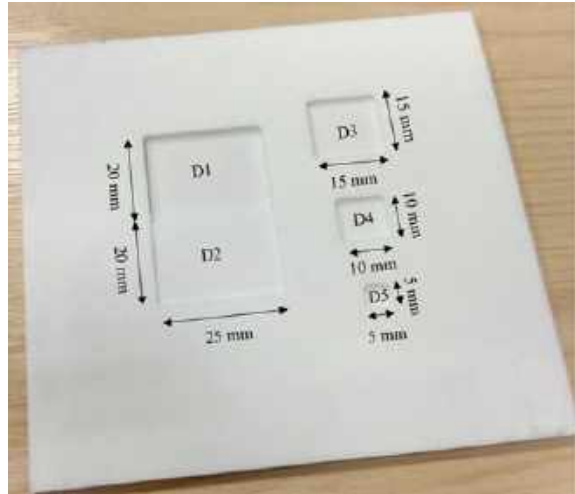
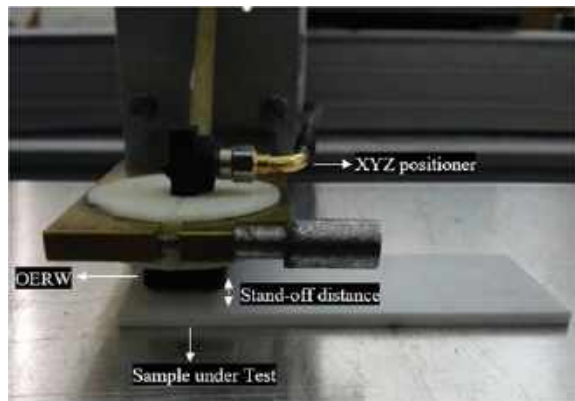


Table 1 Actual delamination size and depth in fabricated Macor sample

Macor sample	Defect 1 (D1)	Defect 2 (D2)	Defect 3 (D3)	Defect 4 (D4)	Defect 5 (D5)
Delamination size (mm ²)	20 × 25	20 × 25	15 × 15	10 × 10	5 × 5
Delamination depth (mm)	2	1	1.5	1.5	1.5

Fig. 2 Setup of the inspection system



3.3 Microwave Signal Processing

Delamination inspection and size estimation using the k-medoids clustering algorithm is shown in Fig. 3. A frequency-to-time transformation is performed by IFFT for each frequency point, $S(m, n)$.

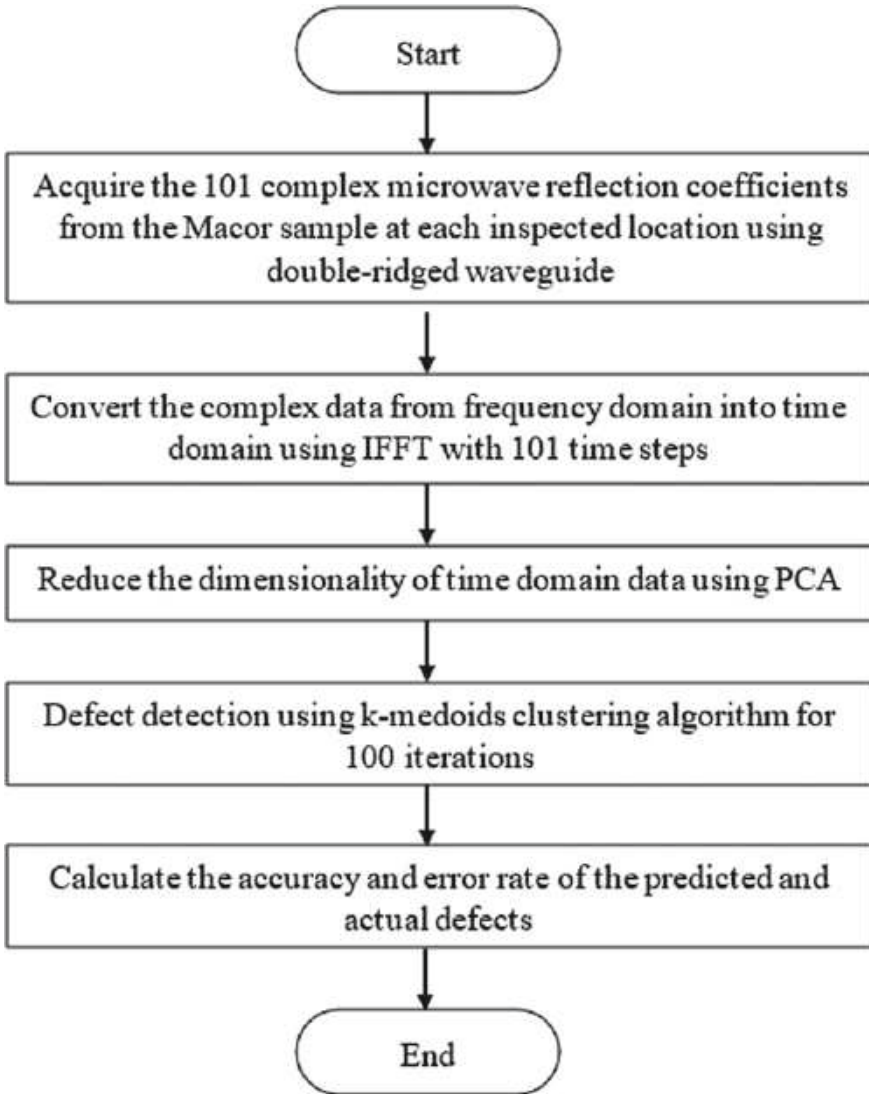
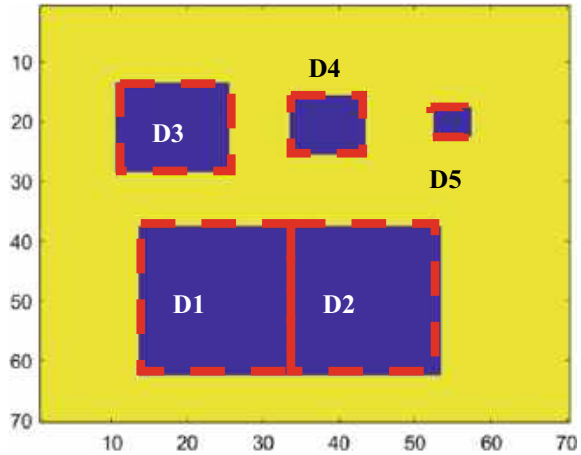


Fig. 3 Work frame of double-ridged waveguide with a k-medoids clustering algorithm

PCA broke down the 101-time steps into three uncorrelated components. After that, 100 iterations of the k-medoids are undertaken to fine-tune the clustering results. Figure 4 presents the actual delamination defects. The studied coating regions are marked with blue and yellow colour in the ground truth, corresponding to defect and non-defect zones. The distinction between actual and predicted defect sizes is reported, along with the area, error rate, and precision.

Fig. 4 Actual defect of Macor sample



4 Results and Discussions

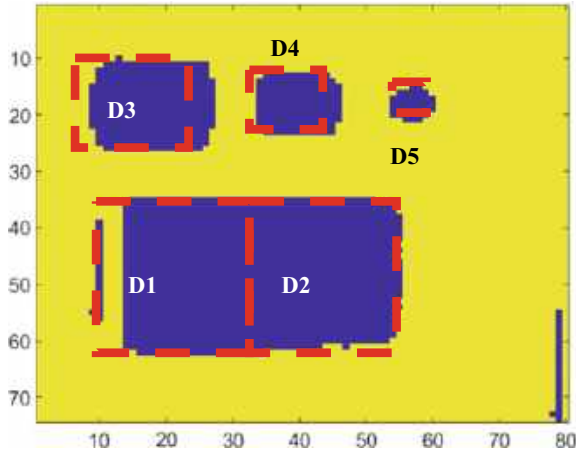
Table 2 displays the error rate for anticipated delamination under the Macor sample. The k-medoids clustering algorithm can identify all five delamination defects. The highest precision that can be attained with the k-medoids algorithm is 95.3%. In inspecting the defect with a smaller delamination depth, an error rate of 8% is achieved when predicting defect D2. This approach has better predictive accuracy for bigger delamination. The proposed method can delineate the boundaries of the defects. Yet, the projected extent of delamination is substantially greater than the delamination’s real size and area. Figure 5 depicts a spatial image utilized for k-medoids validation to determine the defect size in the Macor coating.

The electromagnetic interaction between the open end of the waveguide and the defect area causes a larger estimation error in smaller defects. The interaction typically begins when the defect’s leading edge contacts the waveguide. As a result, the variation in the reflection manifests across a region that is greater than the defect by a factor that depends on the size of the waveguide’s aperture. The projected defect

Table 2 Clustering data of the k-medoids algorithm

Actual defect size (mm × mm)	Predicted defect size (mm ²)	Actual defect area (mm × mm)	Predicted defect area (mm ²)	Error rate (%)	Accuracy (%)
20 × 25	21 × 28	500	585	17	95.3
20 × 25	21 × 27	500	540	8	
15 × 15	19 × 16	225	287	27.6	
10 × 10	13 × 11	100	136	36	
5 × 5	7 × 7	25	37	48	

Fig. 5 Spatial image using a k-medoids clustering algorithm



size becomes greater as the defect approaches the aperture. Despite these imprecisions, the suggested algorithm offers a straightforward microwave NDT method for defect identification and sizing without substantial prior knowledge, which can be employed in the industry as part of a standard maintenance schedule.

5 Conclusion

The current study introduced a double-ridged waveguide employing a k-medoids clustering algorithm for identifying delamination in ceramic coatings. The recommended approach uses microwave reflection coefficients to discover the ceramic insulating layer defects.

In this study, the k-medoids clustering method predicts the defects with 95.3% accuracy. K-medoids provide a more precise prediction of the larger delamination. Projected delamination zones are bigger than actual defects. The recommended approach provides clear demarcation of the defects' borders. It also involves fewer modifications, making it more user-friendly as an in situ microwave NDT system for defect detection. In addition to quality assurance in the manufacturing setting, it can be used for on-site service checks.

Acknowledgements This work was funded and supported by a Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project Code: FRGS/1/2020/TK0/USM/02/2.

References

1. Yee TS, Akbar MF, Ghazali NA, Mohamed MFP (2022) Defects detection using complementary split ring resonator with microstrip patch antenna. In: Proceedings of the 11th international conference on robotics, vision, signal processing and power applications, Malaysia 829, pp 625–631. https://doi.org/10.1007/978-981-16-8129-5_95
2. Yee TS, Akbar MF (2022) Under Insulation microwave non-destructive testing using dual-ridges open-ended rectangular waveguide. In: Proceedings of the 11th International conference on robotics, vision, signal processing and power applications, Malaysia, pp 684–689. https://doi.org/10.1007/978-981-16-8129-5_104
3. Shrifan NHMM, Akbar MF, Isa NAM (2019) Prospect of using artificial intelligence for microwave nondestructive testing technique: a review. *IEEE Access* 7:110628–110650. <https://doi.org/10.1109/ACCESS.2019.2934143>
4. Yee TS, Shrifan NHMM, Al-Gburi AJA, Isa NAM, Akbar MF (2022) Prospect of Using machine learning-based microwave nondestructive testing technique for corrosion under insulation: a review. *IEEE Access* 10:88191–88210. <https://doi.org/10.1109/ACCESS.2022.3197291>
5. Tan SY, Firdaus Akbar M, Shrifan NHMM, Jawad GN, Nadhir M, Wahab A (2022) Assessment of defects under insulation using K-medoids clustering algorithm-based microwave nondestructive testing. *Coatings* 12:1440. <https://doi.org/10.3390/COATINGS12101440>
6. Akbar MF, Shrifan NHMM, Jawad GN, Isa NAM (2022) Assessment of delamination under insulation using ridge waveguide. *IEEE Access* 10:36177–36187. <https://doi.org/10.1109/ACCESS.2022.3163308>

Microwave Nondestructive Evaluation Using Spiral Inductor Probe



Danladi Agadi Tonga, Muhammad Firdaus Akbar, Ahmed Jamal Abdullah Al-Gburi, Imran Mohd Ibrahim, Mohammed Fauzi Packeer Mohammed, and Mohammed Mydin M. Abdul Kader

Abstract Non-Destructive Testing (NDT) techniques are crucial in various civil and military applications for detecting and characterizing defects. The NDT methods contribute to the sustainability, operational readiness, and overall success of civil and military engineering. However, conventional NDT methods need improvement to detect delamination in carbon fibre reinforced polymer (CFRP) composite. The Microwave NDT method using a spiral inductor probe has the potential to detect a hidden or subsurface defect in CFRP. Unlike conventional eddy current NDT method that mainly detects surface and near-surface defects. Thus, the sensor can penetrate deeper into CFRP, increasing the chance of detecting concealed delamination defects. The probe is analyzed using CST Microwave Studio's computer simulation software. The designed spiral inductor exhibits a return loss of -24 dB and resonates at 419 MHz. The spatial resolution and sensitivity of the proposed probe are assessed using line scans and regression analysis. The probe demonstrates high spatial resolution and sensitivity in detecting defects at depths ranging from 0.5 to 3.0 mm. The performance of the proposed probe indicates that resonant frequencies decrease proportionally with the depth of the defect.

Keywords Spiral inductor probe · Simulation · Delamination · And microwave nondestructive evaluation

D. A. Tonga · M. F. Akbar (✉) · M. F. P. Mohammed
School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Nibong Tebal Engineering Campus, 14300 Penang, Malaysia
e-mail: firdaus.akbar@usm.my

A. J. A. Al-Gburi
Centre for Telecommunication Research and Innovation (CeTRI), Faculty of Electrical and Electronic Engineering Technology (FTKKE), 76100 Melaka, Malaysia

I. M. Ibrahim
Centre for Telecommunication Research and Innovation (CeTRI), Fakulti Kejuruteraan Elektronik Dan Kejuruteraan Komputer (FKEKK), Universiti Teknikal Malaysia Meleka, 76100 Durian Tunggal, Malaysia

M. M. M. A. Kader
Faculty of Electrical Engineering and Technology, Universiti Malaysia Perlis, Perlis, Malaysia

1 Introduction

With the introduction of modern composite material, such as carbon fibre reinforced polymer (CFRP), that shows remarkable ductility and adaptability across diverse industries, it is imperative to develop an innovative nondestructive testing (NDT) method to inspect the structural integrity of the composite [1]. Nondestructive testing (NDT) can be defined as a set of techniques utilized to assess the integrity and quality of materials [2]. The NDT methods aim to detect and evaluate the tested sample's defects, flaws, or anomalies [3]. Contrasted destructive testing (ND) refers to a set of evaluation methods which causes physical damage to the tested sample [4].

The NDT method is intensively used in various industries, such as power plants and aerospace [5]. Consistent inspection evaluation of the structural integrity in these industries is compulsory, using NDT methods to avoid severe system failure. Implementing NDT inspections to evaluate structural damage is essential to minimize the maintenance cost and significantly enhance the safety and effectiveness of the system [6].

Various regularly utilized conventional NDT techniques such as digital image correlation, eddy current and ultrasonic [7]. The selection of the methods has relied on its benefits and drawbacks.

Nevertheless, the conventional NDT methods fell short of inspecting and evaluating CFRP composite due to the restriction of field penetration. Thus, an alternate inspection method is required to overcome the restrictions of field penetration.

Compared to conventional NDT methods, microwave NDT has surfaced as an inspiring method for assessing CFRP materials [8]. Microwave NDT methods have many benefits, such as being user-friendly, less expensive and non-contact. This research proposes a microwave NDT method using a spiral inductor probe with a square shape. The probe as a resonator allows effective miniaturization and use of lower microwave frequencies.

The SI probe was designed using CST Microwave Studio. The selection of the SI probe as the microwave NDT method to evaluate the condition of the CFRP composites was based on multiple factors. These factors include low cost and ease of design compared to other microwave NDT methods, and it has the advantage of using a single port to excite the sensor [9].

2 Theoretical Approach

Theoretical approaches to spiral inductor design provide a foundation for optimizing the sensor's performance and achieving the desired characteristics for detecting delamination defects in CFRP composites. By utilizing electromagnetic modelling, magnetic field distributions and applying optimization algorithms to enhance the sensitivity and spatial resolution of the spiral inductor sensor enabling more accurate and reliable defect detection.

2.1 Spiral Inductor

This research utilizes a spiral inductor as a near-field imaging technique. The probe operates within a narrow frequency band and exhibits resonance. In contrast, non-resonant near-field imaging techniques operate across broader frequency ranges. The detection principle of the resonant spiral inductor is based on the interaction between the probe and the specific defect of interest in the near field. This interaction leads to variations in the probe's resonance frequency, enabling defect detection.

2.2 Inductance

When an alternating current (ac) flows through a conductor, it produces a magnetic field around it. The magnitude of this magnetic field depends on the current amplitude and the conductor's geometry. For spiral inductors, the shape of the coil improves the magnetic field generated.

The equation can calculate the inductance L of a spiral inductor.

$$L = \frac{\mu_0 \mu_r N^2 A}{(l + 0.5d\pi)} \quad (1)$$

where L is the inductance, μ_0 is the permeability in free space, μ_r is the relative permeability of the core material, N is the number of turns in the spiral, A is the cross-sectional area of the spiral coil, where π is a constant, and d is the mean diameter of the spiral coil.

3 Methodology

3.1 Design of the Spiral Inductor Sensor

The lower microwave range probe's design is depicted in Fig. 1 with an 8-turn square SI sensor and a 7.1 mm side length. A small loop is created around the SI to stimulate it, and both the SI and loop are built from 0.035 mm copper strips on a 1.6 mm thick FR4 4.3 substrate. The proposed SI sensor was simulated using CST Microwave Studio.

The proposed probe incorporates a loop for feeding the SI while utilizing a matching network consisting of a 12pF capacitor (C_n) and two equivalent 2pF capacitors (C_a and C_b) to ensure a 50- Ω match. This study's main objective was achieved by simulating the spiral inductor sensor and found to resonate at 419 MHz with a return loss of S_{11} of -24 dB, as shown in Figs. 1 and 2.

Fig. 1 Proposed design spiral inductor probe

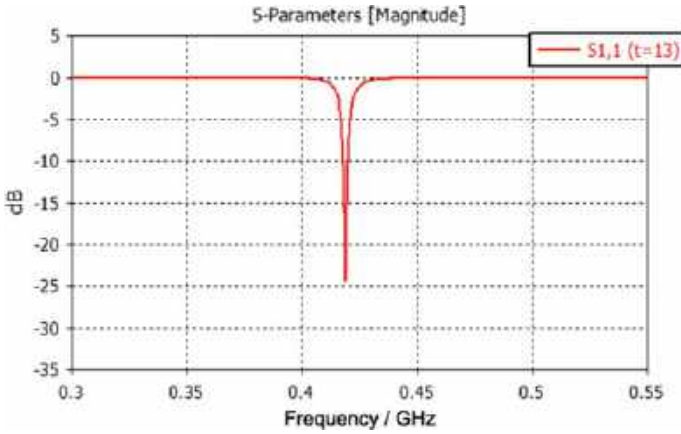
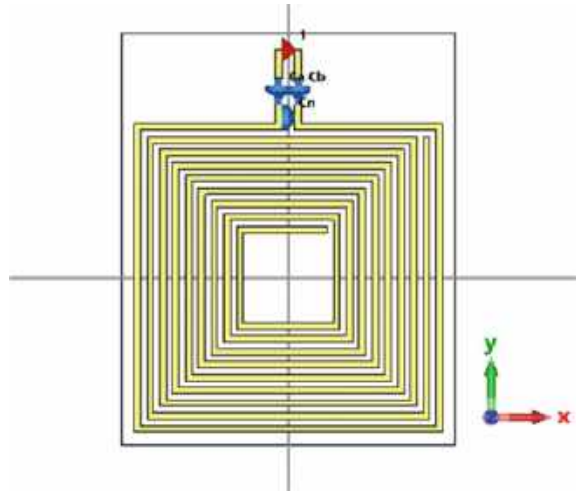


Fig. 2 Return loss (S11) in dB of the 3D simulated result of the probe

3.2 CFRP Sample

This study employed a CFRP sample in a rectangular plate with 50 mm × 50 mm × 6 mm (length, width, and thickness). Defects were created on the sample, and the proposed probe in Fig. 1 was utilized to examine the depths of the defects, which varied from 0.5 to 3.0 mm. A line scan was performed from - 15 to 15 mm in length with a step size of 2.5 mm to assess the probe’s sensitivity.

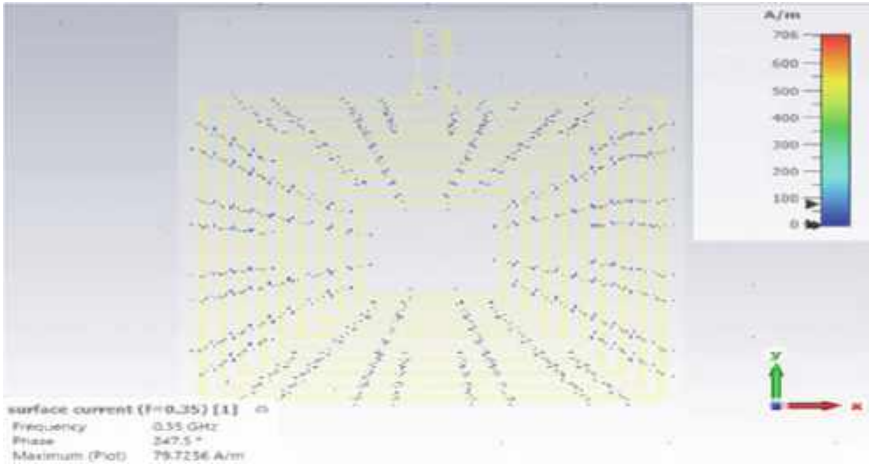


Fig. 3 Simulation of the near-field probe’s surface current distribution in the x, y plane

4 Result and Discussion

4.1 Resolution and Sensitivity of Proposed Spiral Inductor Probe

Specific features of imaging systems are determined by their ability to detect defects. The image probe’s near-field sensing region affects its spatial resolution and is influenced by the SI sensor’s geometry. The near-field region of the probe is used to assess the magnitude of the surface current and examine the sensing region. Figure 3 illustrates the surface current distribution over the x, y, and z planes in the near-field probes.

According to the findings, the imaging probe effectively detects surface current and exhibits superior spatial resolution in the near field. The surface current distribution in the probe’s plane is symmetric along the x and y axes. The maximum positions along the x, y, and z planes, as depicted in Fig. 3, are 12.900, – 2.100, and 0.035 mm, respectively. Thus, the probe’s spatial resolution performance makes it suitable for defect detection during sample scanning. Furthermore, the sensitivity of the SI probe was investigated by scanning the depth of the defect sizes, ranging from 0.5 to 3.0 mm. Six (6) observations were made, with the corresponding resonance frequency responses of 409, 404, 401, 399, 398, and 397 MHz, obtained through line scanning. A linear regression analysis technique was used to determine the relationships between the two variables. Figure 4 illustrates the regression analysis plot result.

The regression analysis produced an R-squared value of 0.9023 and an R-value equation of $- 4.571x + 409.39$. The R-squared value of 90.23% indicates that the defect’s depth can explain variability in resonance frequency. The negative slope of

Fig. 4 Regression analysis of resonant frequency versus depth of the defect

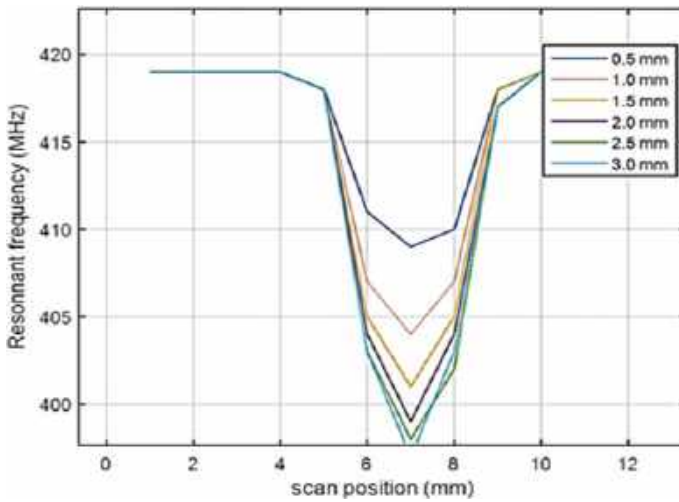
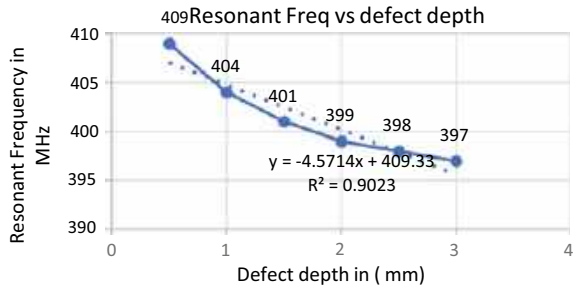


Fig. 5 Matlab plot of resonance frequency responses against the depth of the defect

the regression line ($- 4.517$) suggests that the resonance frequency decreases as the depth of the defect increases. On the other hand, the line intercept (409.39) reveals the resonance frequency when the depth of the defect is 0.5 mm. The results suggest that the SI probe used in this study demonstrates a strong sensitivity to the depth of the defect. The findings suggest that the probe can identify defects near the inspected material.

Matlab software was also used to plot the data in Table 1. The resonant frequency was plotted against the depth of the defect. As the depth of the defect increases, there is a decrease in resonant frequency. Therefore, it can be deduced that deeper flaws significantly impact the resonant frequency.

Table 1 Depth of defects and resonant frequency response

Depth in (mm)	0.5	1.0	1.5	2.0	2.5	3.0
Resonant frequency in MH	409	404	401	399	398	397

The result can be valuable in applications such as nondestructive testing, where the resonant frequency can indicate the defect's severity in the inspected material.

5 Conclusion

This research describes a novel microwave spiral inductor probe for detecting delamination in CFRP composites. The proposed probe performance shows that resonant frequencies decrease in direct proportion to the depth of the defect. This indicates that the probe is sensitive to small changes due to a material under test (MUT) defect.

References

1. Yi Q, Tian GY, Malekmohammadi H, Laureti S, Ricci M, Gao S (2021) Inverse Reconstruction of fibre orientation in multilayer CFRP using forward FEM and eddy current pulsed thermography. *NDT and E Int* 122. <https://doi.org/10.1016/j.ndteint.2021.102474>
2. Katunin A, Wronkiewicz-Katunin A, Danek W, Wyleźół M (2021) Modeling of a realistic barely visible impact damage in composite structures based on NDT techniques and numerical simulations. *Compos Struct* 267. <https://doi.org/10.1016/j.compstruct.2021.113889>
3. Chen J, Yu Z, Jin H (2022) Nondestructive testing and evaluation techniques of defects in fiber-reinforced polymer composites: a review. *Front Mater* 9
4. Segovia Ramírez I, García Márquez FP, Papaalias M (2023) Review on Additive manufacturing and non-destructive testing. *J Manuf Syst* 66:260–286
5. D'Angelo G, Palmieri F (2020) Knowledge elicitation based on genetic programming for non destructive testing of critical aerospace systems. *Futur Gener Comput Syst* 102:633–642. <https://doi.org/10.1016/j.future.2019.09.007>
6. García Márquez FP, Peco Chacón AMA (2020) Review of Non-destructive testing on wind turbines blades. *Renew Energy* 161:998–1010
7. Teng WS, Akbar MF, Jawad GN, Tan SY (2021) A past, present, and prospective review on microwave nondestructive evaluation of composite coatings, pp 1–25
8. Zhang H, Yang R, He Y, Foudazi A, Cheng L, Tian G (2017) A review of microwave thermography nondestructive testing and evaluation. *Sensors (Switzerland)* 17
9. Abou-Khousa MA, Haryono A (2020) Array of planar resonator probes for rapid near-field microwave imaging. *IEEE Trans Instrum Meas* 69:3838–3846. <https://doi.org/10.1109/TIM.2019.2937532>

Design of a Hairpin SIR Dual-Band Bandpass Filter with Defected Ground Slots for WLAN Application



Nur Irdina Rizal, Azniza Abd Aziz, and Intan Sorfina Zainal Abidin

Abstract An efficient and practical method in designing a dual-band bandpass filter operating at 2.4 and 5.2 GHz for WLAN application is presented. The proposed filter is designed by utilizing stepped impedance resonator (SIR) in hairpin-line configuration to make the structure compact and defected ground slots is introduced to attain a high coupling coefficient to enhance the performance of frequency response. The proposed design of the dual-band bandpass filter is simulated in CST Studio Suite software and fabricated using RO4003C substrate with relative permittivity of 3.55, thickness of 1.524 mm and loss tangent of 0.002. The simulation result shows low insertion loss with 1.42 dB and 1.11 dB, high return loss with 35.74 dB and 32.67 dB, fractional bandwidth (FBW) of 263 MHz (10.9%) and 349.4 MHz (6.5%) respectively at 2.4 and 5.2 GHz. The simulated and experimental results are well correlated. The proposed filter has been compared with other designs in this paper and demonstrated higher return loss compared to others.

Keywords Dual-band bandpass filter · Ground slots · SIR

1 Introduction

Nowadays, modern wireless communication systems support several standards that function under dual-frequency bands. The term dual band can be defined as microwave components that are able to operate at two separate frequencies simultaneously. For instance, Wireless Local Area Network (WLAN) operates at 2.4 and 5.2 GHz while Worldwide Interoperability for Microwave Access (WiMAX) supports 3.5 and 5.5 GHz. Due to this recent advancement in communication and

N. I. Rizal · I. S. Z. Abidin

School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Engineering Campus, 14300 Nibong Tebal, Pulau Pinang, Malaysia

A. A. Aziz (✉)

Intel Microelectronics, 11900 Penang, Malaysia

e-mail: aznizaaziz8@gmail.com

standards, the design of dual-band wireless systems that requires a filter with the capability of extracting dual-band signals, is widely encouraged.

Multiple studies have been done on the methods of realizing the dual-passband frequency response. Some of the most commonly used techniques are by connecting two separate single passband filters in parallel connection for each independent signal path [1, 2] and by separating a wideband frequency response with a notch band by using bandstop filter [3]. This approach, however, has drawbacks which required a high number of components and large circuit size. In order to significantly minimize the size of the filter and the whole system, a single filter with dual-band features, was proposed by previous studies conducted by other researchers. One of the approaches to realizing a dual-band bandpass filter is by employing a type of resonator with two tunable resonator frequencies, which is stepped impedance resonator (SIR) [4–8].

The miniaturization of the filter can be further improved by varying the standard structure of SIR, which is parallel-coupled microstrip bandpass filter into other compact structures such as hairpin-line, interdigital and combline structures. In this paper, by applying the SIR and hairpin-line technique, a dual-bandpass filter applicable for WLAN application is proposed, designed, fabricated and measured. In this research work, slots were embedded in the ground slots to increase the coupling strength to compensate the impractical requirements on the physical dimensions of the filter.

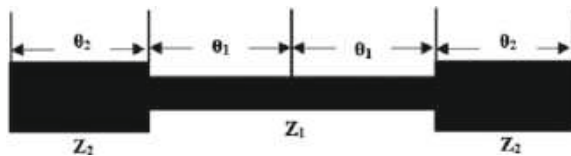
2 Theoretical Approach

SIR consists of two or more transmission lines with varying values of characteristic impedance. In this research, half-wavelength type of SIR is used, due to its controllable spurious response. As shown in Fig. 1, the basic layout of the SIR is composed of sections with high impedance Z_1 and electrical length θ_1 and another two sections, each with low impedance Z_2 and electrical length of θ_2 .

The lower passband of the dual-band filter is created by the fundamental resonance frequency which correlates to the total length of the resonator while the higher passband is generated by the first spurious response frequency that is adjusted by varying its impedance ratio, R_Z of the resonator, which is given by Eq. 1 [9],

$$R_Z = \frac{Z_1}{Z_2} = \tan \theta_1 \tan \theta_2 \tag{1}$$

Fig. 1 Basic structure of half wavelength SIR



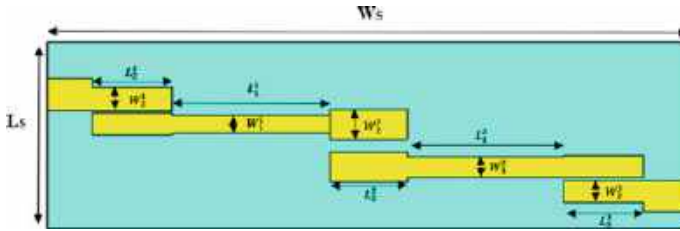


Fig. 2 Basic layout of parallel coupled SIR dual-band bandpass filter

The first spurious frequency, f_2 of the SIR is related to the fundamental frequency, f_1 as Eq. 2 [9],

$$\frac{f_2}{f_1} = \frac{\pi}{2 \tan^{-1} \sqrt{R_Z}} \tag{2}$$

The SIR dual-band bandpass filter is firstly designed in parallel-coupled configuration, as shown in Fig. 2. Element values of Chebyshev lowpass prototype of three poles ($n = 3$) with a passband ripple of 0.1 dB is utilized in this design. Once the fundamental parameters had been specified, the electrical parameters of the parallel-coupled SIR, such as inverter parameters, characteristic impedances and electrical length, are determined.

3 Methodology

3.1 Design and Simulation

In this research, the parallel coupled structure of the filter is then folded into a “U” hairpin-line configuration, since Fig. 2 filter is inefficient as it occupies larger area. To further minimize the size of the dual band bandpass filter, the layout is slightly modified by shifting the resonator so that the parallel couplings of the filter are realized by both Z_1 and Z_2 sections of the SIR. Therefore, the previous spacing values, s_n , are now invalid due to the additional coupling realized by Z_1 sections. The new value of spacing between adjacent resonators is determined by the coupling coefficient. In order to attain a high coupling coefficient, the coupling separation at the first and last stages can be reduced. However, this approach cannot be accurately achieved in practice due to the limitation in the fabrication process.

Therefore, ground slots with a dimension of length L_G and width W_G are embedded below the first and last coupled sections of the hairpin-shaped SIRs as illustrated in Fig. 3 to enhance the performance of the passband response. Parametric studies on the dimensions of the ground slots are conducted to achieve the

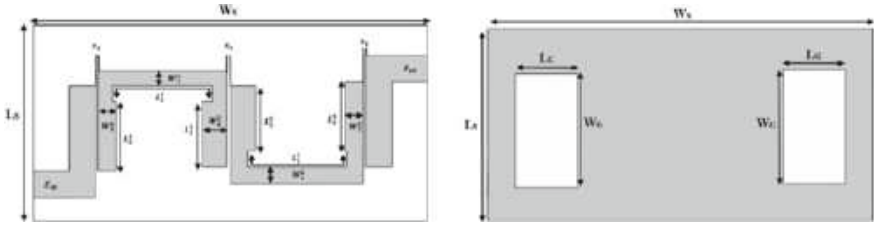


Fig. 3 **a** Layout of modified SIR dual-band bandpass filter in hairpin-line configuration. **b** Bottom view of the proposed hairpin SIR dual-band bandpass filter

optimum performance of the dual-passband response at the targeted center frequencies. Therefore, the values of L_G and W_G at 14.896 and 9.082 mm were chosen for the dimension of the ground slot. Due to the slight shifting of the dual-frequency responses when ground slots are added in the design, the values of the length of hairpin arms, L_2^1 and L_2^3 are fine-tuned until center frequencies are achieved.

3.2 Fabrication

The final layout of the proposed filter is then fabricated on Rogers RO4003C material with ϵ_r of 3.55 and thickness, h of 1.524 mm. SMA connectors are properly soldered into the feeding lines of the filter as illustrated in Fig. 4. Based on the simulation results the final structure is fabricated to compare with the simulation results, $W_2^1 = 2.368$, $L_2^1 = 8.8$, $W_1^1 = 1.871$, $W_1^1 = 3.18$, $W_1^2 = 2.142$, $W_G = 14.896$, $L_G = 9.082$ (unitmm). The fabricated filter is measured by using a vector network analyzer to compare the performance of the simulated and measured dual-passband response in terms of insertion loss and return loss.

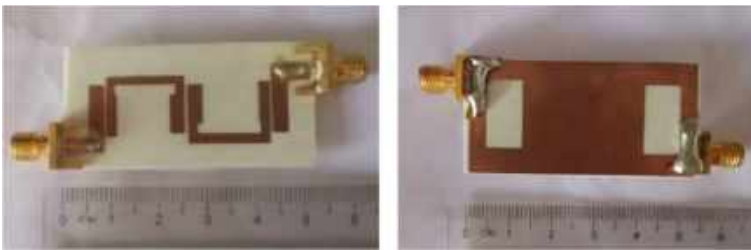


Fig. 4 **a** Top view and **b** back view of proposed hairpin SIR dual-band bandpass filter fabricated on RO4003C

4 Result and Discussion

The performance of the measured return loss and insertion loss showed well-match data between simulation and measurement as illustrated in Fig. 5. The measured return loss showed a slight decrease as compared to the simulated values, but the values still exhibited more than 25 dB. The measured center frequencies of the filter were slightly shifted from the simulated values at 2.404 and 5.3434 GHz to 2.35 and 5.41 GHz respectively. The small frequency discrepancies between the simulated and measured results may be caused by the unpredictable fabrication tolerance, imperfect cutting of the filter board and problems with soldering of the SMA connectors to the input and output feeding line. The measured 3-dB fractional bandwidth for the two passbands also showed well-correlated result with a slight deviation between the simulation values at 263 MHz (10.9%) and 349.4 MHz (6.5%) to measurement result at 166.8 MHz (7.1%) and 315.7 (5.8%). The performance characteristics of the simulated and measured filter are then summarized in Table 1.

The measured proposed filter was also compared to other previously published papers, as shown in Table 2. Based on the comparison, it can be observed that the proposed dual-band bandpass filter demonstrated the highest return loss at both passband frequency responses. The size of the proposed filter is also competitive compared to the previous works.

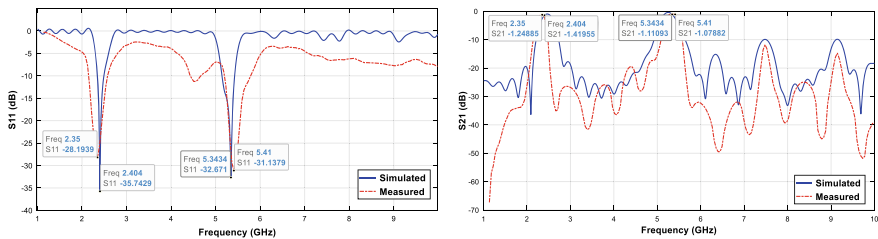


Fig. 5 a Return loss between simulated versus measured. b Insertion loss between simulated versus measured

Table 1 Comparison between simulated and measured results on proposed design

Description	Simulation results		Measurement results	
	S ₁₁	S ₂₁	S ₁₁	S ₂₁
Magnitude at f_1 and f_2 (dB)	35.74, 32.67	1.4195, 1.1109	28.193, 31.137	1.248, 1.078
f_1 (GHz) and f_2 (GHz)	2.4040, 5.3434		2.35, 5.41	
FBW (%)	10.9 and 6.5%		7.1 and 5.8%	

Table 2 Comparison between measured performance of proposed design and previous work

Ref.	Technique	f_1/f_2 (GHz)	FBW (%)	IL (dB)	RL (dB)	Size (mm ²)
[2]	Combination of two single bandpass filter	2.4 /5.2	51.9 /23.3	0.3 /0.7	22.1/20.8	24.4 × 17.5
[8]	Dual passband response synthesis	2.4/5.15	25/16.5	0.5/1	11/9	11.22/13.04
This work	SIRs in hairpin structure with ground slots	2.35/5.41	7.1/5.8	1.25/1.08	28.2/31.3	50.46 × 25

5 Conclusion

In this paper, a dual-band bandpass filter for WLAN applications at 2.4 and 5.2 GHz was proposed designed and simulated by using a new structure, consisting of a type of dual-band resonator, SIR, folded in hairpin-line configuration and two rectangular slots etched in the ground plane. The design of the bandpass filter was then fabricated on dielectric material of RO4003C with substrate permittivity of 3.55 and measured by using a vector network analyzer in order to verify the simulated results. The performance of the S-parameter of the filter prototype showed a good agreement between the simulated and measured results with passband insertion loss, S_{21} less than 1.5 dB and return loss, S_{11} more than 25 dB. The designed dual-band filter has the highest return loss compared to other designs in this paper and the size is comparatively acceptable which is suitable for WLAN.

References

- Rusdi M, Batubara FA, Anugrahwy R, Junaidi, Alifuddin S, Harianto B (2017) Design of dual-band bandpass filter for GSM 950 MHz and GSM 1850 MHz applications using lumped component. *J Phys Conf Ser* 890(1). <https://doi.org/10.1088/1742-6596/890/1/012049>
- Liang GZ, Chen FC (2020) A compact dual-wideband bandpass filter based on open-/short-circuited stubs. *IEEE Access* 8:20488–20492. <https://doi.org/10.1109/ACCESS.2020.2968518>
- Yusoff MFM, Sobri MAM, Zubir F, Johari Z (2019) Multiband hairpin-line bandpass filters by using metamaterial complimentary split ring resonator. *Indones J Electr Eng Inform* 7(2):289–294. <https://doi.org/10.11591/ijeei.v7i2.1172>
- Zhao G, Li M, Zhao R, Tu Z, Yan Y, Mo X (2021) Highly selective UWB bandpass filter with dual notch bands using stub loaded multiple-mode resonator. In: 2021 Photonics & Electromagnetics research symposium (PIERS), Hangzhou, China, pp 1299–1309. <https://doi.org/10.1109/PIERS53385.2021.9695010>
- Li Q, Wu X, Huang Z (2022) Dual-notch ultra-wideband bandpass filter based on folded dual-mode resonator. In: 2022 International applied computational electromagnetics society symposium (ACES-China), Xuzhou, China, pp 1–3. <https://doi.org/10.1109/ACES-China56081.2022.10065262>
- Khajavi M, Shakiba M (2019) Compact microstrip dual-band bandpass filter using step impedance resonators. In: 2019 27th Iranian Conference on electrical engineering (ICEE), Yazd, Iran, pp 323–325. <https://doi.org/10.1109/IranianCEE.2019.8786390>

7. Yang Y, Gu L, Dong Y (2022) Microstrip bandpass filter based on dual-mode folded SIR resonator. In: 2022 International conference on microwave and millimeter wave technology (ICMMT), Harbin, China, pp 1–3. <https://doi.org/10.1109/ICMMT55580.2022.10023430>
8. Salmani R, Bijari A, Zahiri SH (2020) Design of a microstrip dual-bandpass filter using novel loaded asymmetric two coupled lines for WLAN applications. *J Electr Comput Eng Innov* 8(2):255–262. <https://doi.org/10.22061/JECEI.2020.7250.376>
9. Khundrakpam P, Pal M, Sarkar P, Ghatak R (2016) A dual wideband bandpass filter for WLAN and 5G Wi-Fi applications. *International Conference on Microelectronics, Computing and Communications (MicroCom 2016)*, pp 4–7. <https://doi.org/10.1109/MicroCom.2016.7522575>

A Comparative Analysis of BLE-Based Indoor Localization with Machine Learning Regression Techniques



Chia Wei Khor and Nur Syazreen Ahmad

Abstract Indoor localization has become a crucial area of research due to the increasing demand for location-aware applications and services within indoor environments. Bluetooth Low Energy (BLE) is a promising technology for indoor localization due to its advantages such as low power consumption and wide compatibility with various devices. However, BLE also faces limitations, including its limited range and susceptibility to interference from other wireless devices operating in the same frequency range. This paper presents a comprehensive comparative analysis of BLE-based indoor localization techniques using machine learning regression methods. Several models, including Boosted Tree, Bagged Tree, Gaussian Process Regression (GPR), and Support Vector Machine (SVM), are trained and tested using new datasets. The results highlight the superior performance of the Boosted Tree model, with a root mean squared error (RMSE) of 0.7 m. The study's findings provide valuable insights into the strengths and limitations of different machine learning regression techniques for BLE-based indoor localization.

Keywords Bluetooth low energy · Indoor localization · Low power · RSSI

1 Introduction

In recent years, indoor localization has emerged as a vital research area, driven by the growing demand for location-aware applications and services in indoor environments [1]. Accurate and reliable indoor localization systems are essential for a wide range of applications, including asset tracking, navigation, context-aware services, and safety management [2, 3]. Among the various technologies employed for indoor localization, Bluetooth Low Energy (BLE) has gained significant attention as it provides several distinct advantages over other wireless technologies for indoor localization

C. W. Khor · N. S. Ahmad (✉)

School of Electrical and Electronic Engineering, Universiti Sains Malaysia, 14300 Nibong Tebal, Penang, Malaysia

e-mail: syazreen@usm.my

[4]. Firstly, BLE offers low power consumption, which is crucial for battery-powered devices such as smartphones and wearable devices. Compared to other wireless technologies like Wi-Fi or cellular networks, BLE consumes significantly less energy, allowing for longer battery life and extended operation in indoor localization applications. This advantage is particularly important in scenarios where devices need to be deployed over an extended period without frequent battery replacements [5, 6]. Secondly, BLE provides excellent compatibility and wide availability across various devices. BLE is natively supported by most modern smartphones, tablets, and other consumer electronic devices, ensuring a large user base and widespread adoption. This compatibility enables seamless integration with existing mobile devices, making BLE-based indoor localization solutions easily accessible and cost-effective. Furthermore, the compact form factor of BLE beacons and tags allows for convenient deployment in indoor environments without imposing significant physical or aesthetic constraints.

Despite the mentioned advantages, the BLE also has certain limitations that can impact its performance. One of the downsides of BLE is its limited range compared to other wireless technologies [7, 8]. BLE signals can be easily attenuated by physical obstacles such as walls, furniture, and human bodies, leading to signal loss or degradation [9]. This limitation can introduce inaccuracies and inconsistencies in the localization results, particularly in complex indoor environments with multiple obstructions. Furthermore, BLE-based indoor localization can be susceptible to interference from other wireless devices operating in the same frequency range. Since BLE operates in the 2.4 GHz band, which is shared with Wi-Fi, Bluetooth Classic, and other wireless technologies, signal interference can occur, leading to degraded localization accuracy. Interference from other devices can cause signal collisions, packet loss, and increased signal noise, affecting the reliability of the localization system.

To address these limitations and enhance the performance of BLE-based indoor localization, machine learning regression techniques offer significant benefits [10–14]. Machine learning regression algorithms can effectively model and predict the relationship between received BLE signal strength and the corresponding location coordinates. By training the regression models on a carefully collected dataset that includes both signal strength measurements and ground truth location data, machine learning algorithms can learn the complex patterns and variations in the BLE signals.

The goal of this study is to conduct a comprehensive comparative analysis of BLE-based indoor localization techniques using machine learning regression methods. Several best performing models, namely Boosted Tree, Bagged Tree, Support Vector Machine (SVM), and Gaussian Process Regression (GPR) have been trained and tested with new datasets obtained in this work. Results demonstrate that the Boosted Tree outperform the rest with root mean squared error (RMSE) of 0.7 m. The findings of this study are expected to provide valuable insights into the strengths and limitations of different machine learning regression techniques for BLE-based indoor localization.

2 Methodology

The flowchart of the research is presented in Fig. 1, outlining the key steps followed in this study. To conduct the experiment, a house with dimensions measuring 10.72 m by 6.3 m was selected as the indoor environment, as depicted in Fig. 2. The house comprises three rooms furnished with various items. Within the house, three BLE beacons were strategically placed at locations 1, 2, and 3, as indicated in Fig. 2. For localization purposes, a BLE tag was utilized as the target.

During the experiment, the received signal strength indicator (RSSI) values of each beacon were recorded continuously for a duration of 20 s. A total of 23 locations within the house were systematically recorded and mapped, as illustrated in Fig. 2. Among these 23 locations, 19 were marked as reference points (indicated by yellow points) and were used for training purposes. The remaining locations, represented by red crosses, served as test points to evaluate the performance of the model. These recorded locations provide a comprehensive dataset for evaluating the performance of the BLE-based indoor localization system.

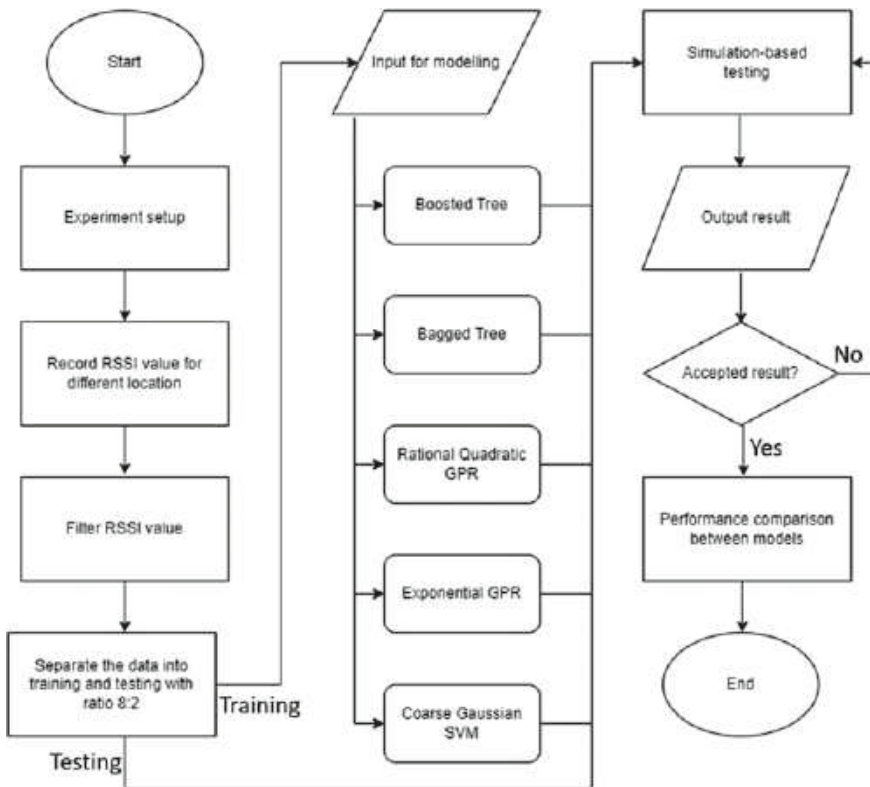


Fig. 1 The flowchart of the experiment

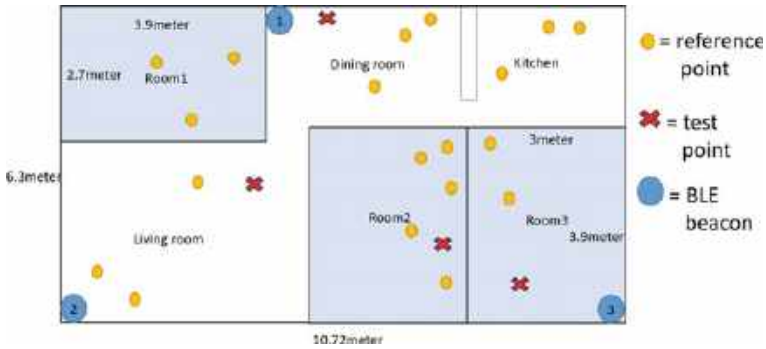


Fig. 2 The field layout of the experiment

In this work, five best performing regression models, namely Boosted Tree, Bagged Tree, Rational Quadratic GPR, Exponential GPR and Coarse Gaussian SVM, were considered. The results are presented in the next section.

3 Results and Discussions

Tables 1 and 2 compare the performance of the 5 models in terms of RMSE and R^2 values of the x- and y- coordinates based on the raw RSSI data, while Tables 3 and 4 compare the performance based on the filtered data. The overall performance where both the errors in x and y coordinates are considered is presented in Fig. 3. Based on the figure, it is evident that the Boosted Tree with filtered RSSI values outperforms the rest with RMSE of 0.70 m, and R^2 of 0.82.

In a study [15] that utilized BLE beacons for indoor localization, the best outcome was achieved using the non-linear least square method, which resulted in an average error of 1.149 m, which is slightly higher than the error via the Boosted Tree model in this work. Thus, it can be concluded that the Boosted Tree can offer a better performance for indoor localization with BLE technology.

Table 1 The result for all regression models of X coordinate with non-filtered data

Model name	RMSE validation	RMSE test	R^2 validation	R^2 test
Boosted tree	1.055	0.7337	0.85	0.82
Bagged tree	0.813	0.86143	0.91	0.75
Rational quadratic GPR	0.606	0.90632	0.95	0.73
Exponential GPR	0.611	0.89076	0.95	0.74
Coarse Gaussian SVM	1.4171	0.89177	0.74	0.74

Table 2 The result for all regression models of Y coordinate with non-filtered data

Model name	RMSE validation	RMSE test	R ² validation	R ² test
Boosted tree	0.80287	0.73877	0.69	0.77
Bagged tree	0.56112	0.72793	0.85	0.78
Rational quadratic GPR	0.43613	0.76782	0.91	0.76
Exponential GPR	0.44086	0.79208	0.91	0.74
Coarse Gaussian SVM	1.1437	0.71833	0.38	0.79

Table 3 The result for all regression models of X coordinate with filtered data

Model name	RMSE validation	RMSE test	R ² validation	R ² test
Boosted tree	1.0802	0.71707	0.85	0.83
Bagged tree	0.9156	0.70191	0.89	0.84
Rational quadratic GPR	0.85351	0.86474	0.9	0.75
Exponential GPR	0.83523	0.87876	0.91	0.74
Coarse Gaussian SVM	1.3633	0.9097	0.76	0.72

Table 4 The result for all regression models of Y coordinate with filtered data

Model name	RMSE validation	RMSE test	R ² validation	R ² test
Boosted tree	0.86092	0.68576	0.65	0.81
Bagged tree	0.65793	0.75392	0.8	0.76
Rational quadratic GPR	0.58748	0.65644	0.84	0.82
Exponential GPR	0.58851	0.68644	0.84	0.81
Coarse Gaussian SVM	1.1194	0.67161	0.41	0.81

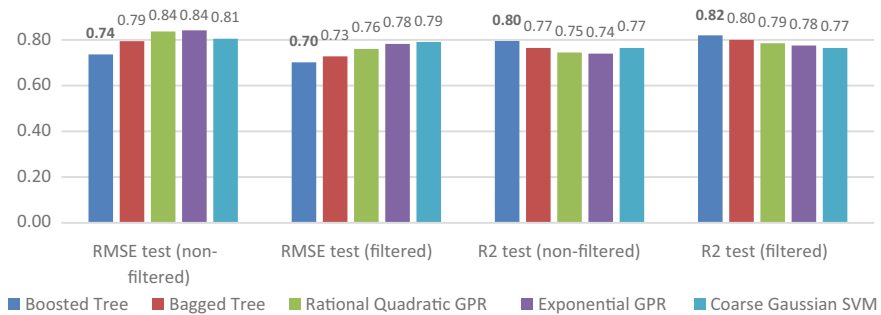


Fig. 3 The results for all regression models for both non-filtered and filtered data

4 Conclusion

In conclusion, the results obtained from this study will serve as a benchmark for future research and development efforts in the field of indoor localization. The proposed evaluation framework can be extended to explore the performance of other machine learning algorithms or alternative technologies, fostering the advancement of indoor localization solutions and contributing to the evolution of smart indoor environments.

References

1. Sadowski S, Spachos P (2018) RSSI-based indoor localization with the internet of things. *IEEE Access* 6:30149–30161
2. Obeidat H, Shuaieb W, Obeidat O, Abd-Alhameed R (2021) A review of indoor localization techniques and wireless technologies. *Wirel Pers Commun* 119(1):289–327
3. Loganathan A, Ahmad NS, Goh P (2019) Self-adaptive filtering approach for improved indoor localization of a mobile node with Zigbee-based RSSI and odometry. *Sensors* 19(21):4748
4. Teo JH, Loganathan A, Goh P, Ahmad NS (2020) Autonomous mobile robot navigation via RFID signal strength sensing. *Int J Mech Eng Robot Res* 9(8):1140–1144
5. Loganathan A, Ahmad NS (2023) A systematic review on recent advances in autonomous mobile robot navigation. *Eng Sci Technol Int J* 40:101342
6. Ahmad NS (2020) Robust H ∞ -fuzzy logic control for enhanced tracking performance of a wheeled mobile robot in the presence of uncertain nonlinear perturbations. *Sensors* 20(13):7673
7. Tiong PK, Ahmad NS, Goh P (2019) Motion detection with IoT-based home security system, vol. 998
8. Ng HT, Tham ZK, Rahim NAA, Rohim AW, Looi WW, Ahmad NS (2023) IoT-enabled system for monitoring and controlling vertical farming operations. *Int J Reconfig Embedded Syst* 12(3):453–461
9. Ahmad NS, Boon NL, Goh P (2018) Multi-Sensor obstacle detection system via model-based state-feedback control in smart cane design for the visually challenged. *IEEE Access* 6:64182–64192
10. Arrouch I, Ahmad NS, Goh P, Mohamad-Saleh J (2022) Close proximity time-to-collision prediction for autonomous robot navigation: an exponential GPR approach. *Alex Eng J* 61(12):11171–11183
11. Arrouch I, Mohamad-Saleh J, Goh P, Ahmad NS (2022) A Comparative study of artificial neural network approach for autonomous robot's TTC Prediction. *Int J Mech Eng Robot Res* 11(5):345–350
12. Ahmad NS, Teo JH, Goh P (2022) Gaussian process for a single-channel EEG decoder with inconspicuous stimuli and eyeblinks. *Comput Mater Continua* 73(1):611–628
13. Teo JH, Ahmad NS, Goh P (2022) Visual stimuli-based dynamic commands with intelligent control for reactive BCI applications. *IEEE Sens J* 22(2):1435–1448
14. Goay CH, Ahmad NS, Goh P (2021) Transient simulations of high-speed channels using CNN-LSTM with an adaptive successive halving algorithm for automated hyperparameter optimizations. *IEEE Access* 9:127644–127663
15. Sadowski S, Spachos P (2019) Optimization of BLE Beacon density for RSSI-Based indoor localization. In: 2019 IEEE international conference on communications workshops (ICC Workshops), pp 1–6

Determination of Anechoic Chamber Set-Up Using Simulation Approach: A Review



Roslina Hussin , Mohd Nazri Mahmud , Mohd Fadzil Ain ,
and Nor Zakiah Yahaya 

Abstract In this paper, we review (i) 5 full-wave numerical techniques of modeling the anechoic chamber, namely the Method of Moments (MoM), Transmission Line Matrix (TLM), Finite Element Method (FEM), Finite Differences Time Domain (FDTD) and Finite Differences Frequency Domain (FDFD), (ii) 3 asymptotic numerical techniques which are (a) Geometrical Optics (GO), (b) Physical Optics (PO) and (c) Uniform Theory of Diffraction (UTD) and (iii) Ray-tracing technique dealing with number of cells of the surface area of the enclosure. The MoM technique formulates the equivalent surface or volumetric current. The Finite Differences Time Domain (FDTD) is more stable in the computational modeling of EM for inhomogeneous materials. We conclude that a combination of a few numerical techniques in the computational Electromagnetic (CEM) simulations to the modeling AC from a basic Ray-tracing technique to the complex Finite Integral technique (FIT) is available under the CST Micro-wave Studio Suite.

Keywords Full wave anechoic chamber · Finite differences time domain (FDTD) · Ray-tracing technique

1 Introduction

Typical AC testing is radar cross-section testing, system testing, and electromagnetic compatibility (EMC) testing. To obtain a reflection-free environment inside the AC, the inner surface is installed with absorbing materials to provide an impedance

R. Hussin (✉) · M. N. Mahmud · M. F. Ain
School of Electrical and Electronic, Universiti Sains Malaysia, Engineering Campus, 14300
Nibong Tebal, Penang, Malaysia
e-mail: eeroslina@usm.my

M. N. Mahmud
e-mail: nazriee@usm.my

N. Z. Yahaya
School of Distance Education, Universiti Sains Malaysia, Main Campus, 11800 Penang, Malaysia

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024 363
N. S. Ahmad et al. (eds.), *Proceedings of the 12th International Conference on Robotics, Vision, Signal Processing and Power Applications*, Lecture Notes in Electrical Engineering 1123, https://doi.org/10.1007/978-981-99-9005-4_46

matching for reflected waves at all frequencies and angles of incidence. The ideal fully anechoic chamber design according to the standard IEEE standard 149 in AC is shown in Fig. 1 [1].

To set up the AC system, the full-wave numerical techniques are important to analyze the outcome of the factors such as the number of cell size (mesh); the frequency of the antenna measurements; and the type of source antenna in the AC. A comprehensive list for a few full-wave simulation methods used for CEM by Deslise [2] such as:—(a) Method of Moments (MOM) and Multilevel Fast Multipole Method (MLFMM) for integral-equation solvers, (b) Finite Element Method (FEM), Finite Differences Time Domain (FDTD) and Transmission Line Matrix (TLM) or differential-equation solvers. Both solvers investigate the array techniques to overcome the electrically massive problem spaces by separating the radiation boundaries. Nowadays, those methods can be hybridized to provide a less time-consuming solution because of the large matrices results. Those methods also need enormous amounts of memory and are only suitable for frequencies up to VHF. Thus, for higher frequency and reducing simulation complexity and time the ray-tracing technique known as asymptotic techniques such as:—(a) Geometrical Optics (GO), (b) Physical Optics (PO), and (c) Uniform Theory of Diffraction (UTD). These methods represent optical refraction, reflection, and propagation in 3D space.

The advantage of MoM techniques is that they can confine the surface of the scattered to a high conducting surface material, and it can derive current density from the antenna parameters namely the impedance, gain, radiation pattern, and so on. Referring to Mohammed-Aly [3] the drawbacks of the FEM compared to MoM is the discretizing of the whole objects instead of surface objects and meshes become complex for huge 3D AC structures. It is dependent on the boundaries and the material distribution of the problem. On the other hand, FDTD is one of the most popular numerical techniques for EM analysis according to Feng et al. [4]. It can generate

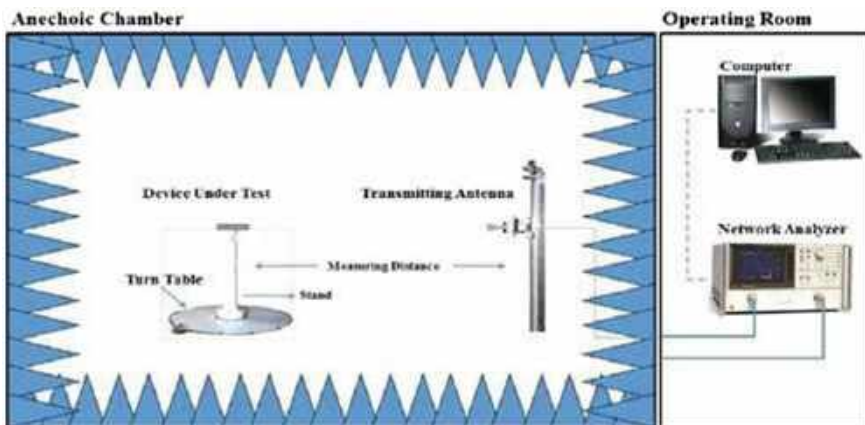


Fig. 1 Test setup between rotation DUT and the source (Tx) for antenna characteristic measurement [1]

a satisfactory result for dispersive dielectric materials applied within or around the device, however for high resonant wavelength materials FDTD method technique is more suitable. This is because FDTD will take more time to reach the frequency domain solution and be inefficient for highly resonant materials.

The purpose of this paper is to review the numerical techniques for the EM simulation solver commercially available and its application suitable for AC systems especially the run time and its approach based on the cell size, frequency, cuboid design size, and devices available.

2 Literature Review

A discrete model of compact AC shown in Fig. 2 is designed by Kawabata et al [5]. It is used to analyze the classical site attenuation (CSA) of the propagating characteristic between the receiving and transmitting antennas. The FDTD method is applied to dielectric and magnetic material used in the AC model to get the result that fulfilled Maxwell's equations. On the other hand, MoM and ray tracing methods are applied for a half-dipole antenna, a short dipole antenna, and the EM absorber. The MoM method is calculated by an analytical tool Numerical Electromagnetics Code (NEC) for 25 segments of partition of the element. Meanwhile, EM absorbers based on ferrite tiles are used to analyze the CSA, and produce the result given in Table 1 for cell size, mesh, and calculation time.

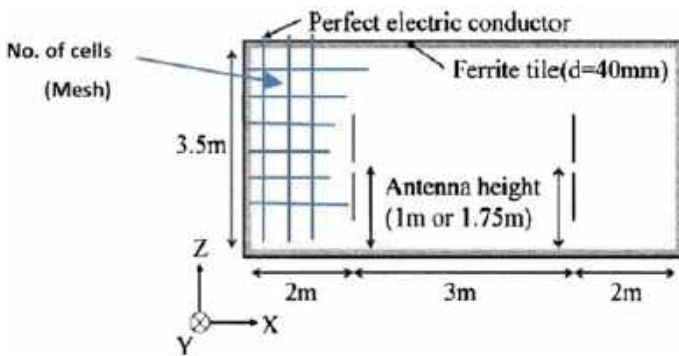


Fig. 2 Model for CSA analysis of 3.5 m H × 4.0 m W X 7.0 m L with cell size for MoM, ray tracing, and FDTD methods

Table 1 Run Time result related to cell size and mesh of the CSA model analysis using MoM, ray tracing, and FDTD methods

Cell size (mm)	Mesh	Run time (min)
40	1.54×10^6	70
20	12.25×10^6	1120
10	98.00×10^6	N/A

The result is restricted to 40 mm cell size and frequency range below 100 MHz and the computing time of 70 min is still acceptable. It is suggested to include the shield door and turntable in the modeling and simulation above 100 MHz where the ray-tracing method is applicable.

FEKO is a commercial software available for modeling and analysis used to solve full-wave AC at VHF/UHF frequencies by Campbell et al. [6]. The analysis used for design process techniques is FEM for tetrahedral element volume meshing technique and PO technique for surface meshing technique (homogeneous dielectric properties). For simulation purposes, any type of antenna source model is shown in Fig. 3 where λ is the operating wavelength with 25 current sources measuring $\lambda/15$ in length and the desired -3 dB points angle beam width in both the E-plane and H-plane.

The AC model (refer to Fig. 4.) dimensions are 17'' H \times 24'' W \times 32.5'' L with PEC as the outer shell and pyramidal absorbers patch at the inner side of the AC walls. The source antenna dimensions are 6'' H \times 6'' W \times 6'' L is located 20'' away from the receiving wall (right-hand side of Fig. 4a). Meanwhile, for the asymptotic PO technique referred to in Fig. 4b, the FEKO limit to analysis for a single material is the inner wall of AC.

The simulation results of those models using FEM and PO methods for frequencies between 150 MHz and 2 GHz are listed in Table 2. For 500 MHz frequency only, both FEM and PO methods are applied in the FEKO software indicating that FEM takes three times longer in run time than PO methods to simulate the results of full-wave AC.

Fig. 3 Source antenna model that emulated radiating pattern with a current source and PEC reflector (separated by $\lambda/4$) in the echoic chamber

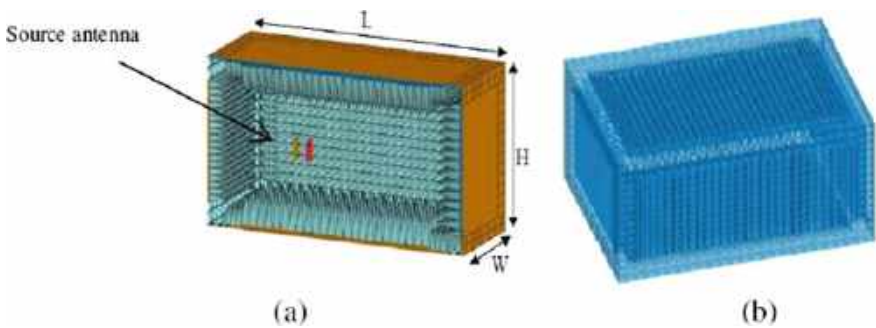
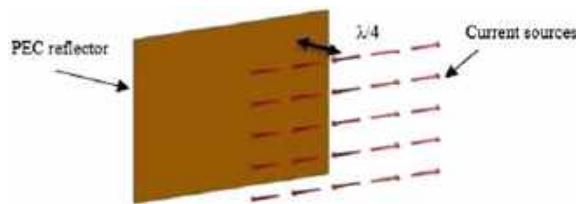


Fig. 4 The AC model for **a** full-wave FEM technique and **b** asymptotic PO technique

Table 2 Run time result related to frequency and method of numerical used

Frequency (MHz)	Method	Run time (min)
150	FEM	3
250	FEM	12
500	FEM	81
500	PO	27
1000	PO	77
2000	PO	271

Taybi et al. [7] used the high-performance software CST Microwave Studio Suite to study AC design using the FEM method for antenna measurement in the frequency domain and the FDTD method for the dielectric objects such as EM absorbers in the time domain. Furthermore, Finite Integral Technique (FIT) method is applied to a cubic elementary mesh discretization in the time domain as shown in Fig. 5 to form Eq. 1 with scalar e and b_n referred to as E-field and B-flux on the edge and the face of the mesh cell. For a large and highly resonant structure, AC with the dimensions 300 cm H × 300 cm W × 500 cm L, it was found that the FIT method becomes unstable. Thus, the EM field of the source is simulated separately with an adequate mesh with the exported file as a Huygens source (due to the spherical surface of the antenna sources). The auto-regressive (order 3) filter is introduced to calculate spectra for time signals and to overcome the equilibrium state problem. For analysis justification in AC for each of the sources implies that the error is maximum in the structure of the absorbers and 0 dB is the axial ratio of a circularly polarized antenna pattern for dipole, open-ended rectangular waveguide (OERWG), and horn sources. Table 3 shows the total number of cells (mesh) and run time of each source.

$$\hat{e}l + \hat{e}k + \hat{e}j + \hat{e}i = \frac{d}{dx}b_n \tag{1}$$

Fig. 5 Cubic elementary mesh discretization

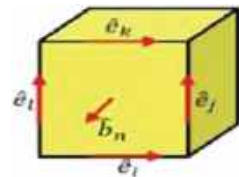


Table 3 Run Time results related to source antenna and mesh of FEM, FDTD, and FIT methods

Sources antenna	Mesh	Run time (h)
Dipole	58 × 10 ⁶	12.5
OERWG	61 × 10 ⁶	14.4
Horn	64 × 10 ⁶	4.5

Table 4 Summary of the reviewed literature

Literature [reference]	AC size (H × W × L)	AC simulations result	Gap
Kawabata et al. [5]	3.5 m × 4.0 m × 7.0 m	Run time (FDTD, MoM, and ray-tracing methods) is less time taken (in minutes) with the cell size wider and the number of cells (mesh) is less (Table 1)	Simulation consideration on the cu- boid chamber for larger AC size
Campbell et al. [6]	17'' × 24'' × 32.5'' (0.4 m × 0.6 m × 0.8 m)	Run time (in minutes) is proportional to the fre- quency. For frequency < 500 MHz (FEM method) and > 500 MHz to 2 GHz (PO method) (Table 2)	Simulation suitable for small sizes and at a lower frequency
Taybi et al. [7]	300 cm × 300 cm × 500 cm (3 m × 3 m × 5 m)	The FDTD, FEM, and FIT methods depend on the source antenna with the number of cells (mesh) being high (Table 3)	Simulation consideration on the AC system for antenna measurement at a higher frequency and larger AC size

Based on the run time of the AC simulations from the reviewed literature depends on the cell size, mesh, frequency, and source antenna, and the AC size discussed in the summary listed in the Table. 4.

3 Conclusion

From Table 4, we conclude that the [7] FDTD, FEM and FIT methods can simulate an enormous size of overall AC system compared to [5] using FDTD, MoM, and ray- tracing methods where the simulation depends on the cuboid chamber and at a low frequency (up to 100 MHz). The size for [6] is insignificant compared to [5] and [7] as well as the frequency simulated up to 2.0 GHz. Based on the summary, the gap to be investigated by the proposed research is to use of the FDTD to shorten the run time for simulating larger AC sizes at higher frequencies. It is also suggested by Xu et al. [8] that the simulation data can be reused for other methods to save computing of the run time.

Acknowledgements This work was funded under UNIVERSITI SAINS MALAYSIA Short Term Grant (304.PELECT.6315609).

References

1. Iran University of Science & Technology Homepage, <http://www.iust.ac.ir/content/55620/Antenna-Characteristics>. Last accessed 17 July 2023
2. Deslise J (2014) Know the differences between EM simulation numerical methods [White paper]. Retrieved 17 July 2023, from Microwave & RF. <https://www.mwrf.com/home/whitepaper/21847465/know-the-differences-between-emsimulation-numerical-methods-pdf-download> EM simulation-numerical-methods-pdf-download
3. Mohammed-Aly EE (2006) Analysis of electromagnetic waves in the time domain. Master's thesis, Menoufia University
4. Feng J, Santamouris M (2019) Numerical techniques for electromagnetic simulation of daytime radiative cooling: a review. *AIMS Mater Sci* 6(6):1049–1064
5. Kawabata M, Ishida Y, Shimada K (2008) FDTD method for site attenuation analysis of compact anechoic chamber using large-cell concept: electrical engineering in Japan. *IEEJ Trans Fundam Mater* 162(4):9–16
6. Campbell D, Gampala G, Reddy CJ, Winebrand M, Aubin J (2013) Modeling and analysis of anechoic chamber using CEM tools. *ACES J* 28(9):755–762
7. Taybi C, Moutaouekkil MA, Elmaground B, Ziyayat A (2021) A novel methodology for time-domain characterization of a full anechoic chamber for antennas measurements and exposure evaluation: *Int J Electr Comput Eng (IJECE)* 11(4):3285–3292
8. Xu Q, Huang Y, Zhu X, Xing L, Duxbury P, Noonan J (2017) Hybrid FEM-GO approach to simulate the NSA in an anechoic chamber. *ACES J* 32(11):1035–1041

Hybridization of Equilibrium and Grasshopper Optimization Algorithms



Ebinowen Tusin Dayo and Junita Mohamad-Saleh

Abstract This paper presents a new hybrid version which uses the swarming behaviour of the Grasshopper Optimization Algorithm (GOA) while Equilibrium Optimizer (EO) to guide the search agents towards promising regions of the search space. It makes use of the exploration and exploitation abilities of Equilibrium Optimizer to refine the search within the regions. The hybrid approach is based on the utilization of GOA to perform a global search of a wider search space during the initial phase of the algorithm. The result has shown that this process has effectively controlled the balance between exploration and exploitation during the search process and successfully hybridized the two algorithms to outperform the original algorithms.

Keywords Optimization · Grasshopper · Exploration · Exploitation · Swarm intelligence

1 Introduction

Over the past few years, optimization has become increasingly crucial in solving various engineering problems. This includes both single and multiobjective problems. Essentially, optimization refers to the process of choosing the most optimal solution among the available options for a given optimization problem [1, 2]. Driven by the hybrid trend, bio-inspired and physics-based algorithms have become an increasing research concern due to their continued prominence in engineering [3].

E. T. Dayo (✉) · J. Mohamad-Saleh
School of Electrical and Electronic Engineering, Universiti Sains Malaysia, 14300 Nibong Tebal, Malaysia
e-mail: ebitosd2000@student.usm.my

E. T. Dayo
Federal Polytechnic, Ile-Oluji, Ondo, Nigeria

The hybridization process between GOA and EO algorithms involves combining the strengths of both algorithms to improve their performance in solving optimization problems [4]. Self-adaptive quantum Equilibrium Optimizer with Artificial Bee Colony (SQEOABC) is a new variant of revised and hybridized EO that was proposed recently [5]. It incorporates quantum theory and self-adaptive mechanisms into the EO's updating rule for convergence enhancement and further uses the updating mechanism of ABC to arrive at the right solution. A new variant of revised EO termed a Normalized Mutual Information equilibrium optimizer (NMIEO) [6] incorporates the local search strategy on Normalized mutual information and chaotic maps to boost the exploitation ability and diversity of the problem as well as enhance the population initialization of the standard EO. The conceptualize potent algorithm of combining the Modified Grasshopper Algorithm (MGOA) and the Improved Harris Hawks Optimizer (IHHO) for attaining a better balance between the beginning stages of global search and the latter stages of global convergence was made possible (MGOA-IHHO) [7]. A recent study reveals the use of EO to initialize the GOA population in which the author proposed a hybrid algorithm called the Equilibrium Grasshopper Optimization Algorithm (EGOA) [8]. The authors showed that EGOA outperformed several other optimization algorithms on benchmark functions. In this paper, the "generate rate" term from EO was incorporated into the update rule of GOA which invigorates the EO's ability in exploration, exploitation, and local minima avoidance. This eventually enhances their search process and improves their performance.

1.1 The GOA Algorithm

The GOA algorithm is inspired by the behaviour of grasshoppers, where individual agents move randomly in the search space, but also take into account the position and direction of their neighbouring agents. The grasshopper swarm behaviour is mathematically modelled and used to calculate the position X_i of each solution as follows:

$$X_i = S_i + G_i + A_i \quad (1)$$

where X_i defines the position of the i th grasshopper, S_i is the social interaction, G_i is the gravity force on the i th grasshopper, and A_i shows the wind advection. Note that to provide random behaviour the equation can be written as $X_i = r_1 S_i + r_2 G_i + r_3 A_i$ where r_1 , r_2 , and r_3 are random numbers in $[0, 1]$.

1.2 The EO Algorithm

The EO algorithm was directly inspired by the solution to a simple well-mixed dynamic mass balance on a control volume V . In the EO algorithm, the individuals would update their positions with the following equation:

$$C_i = C_{eq} + (C_i - C_{eq}) \cdot F + \frac{G}{\lambda V} (1 - F) \quad (2)$$

where C_i is the concentration of i th individual, and C_{eq} is a randomly selected candidate from the equilibrium pool, which was constructed by the four best candidates and their average:

$$C_{Pool} = \{C_{eq(0)}, C_{eq(1)}, C_{eq(2)}, C_{eq(3)}, C_{ave}\} \quad (3)$$

$$C_{ave} = \frac{C_{eq(0)} + C_{eq(1)} + C_{eq(2)} + C_{eq(3)}}{4} \quad (4)$$

There are two unique control parameters involved in this algorithm, the exponential term F and the generation rate term G . They are formulated following the mathematical expressions of the solution:

$$F = a_1 \text{sign}(r_3 - 0.5) (e^{-\lambda n} - 1) \quad (5)$$

$$n = \left(1 - \frac{t}{\max Iter}\right)^{\left(a_2 \frac{t}{\max Iter}\right)} \quad (6)$$

$$G = GCP \cdot (C_{eq} - \lambda C) \cdot F \quad (7)$$

$$GCP = \begin{cases} 0.5r_4 & r_5 \geq GP \\ 0 & r_5 < GP \end{cases} \quad (8)$$

r_4 and r_5 are two other random numbers in the Gauss distribution.

Invariably, just a randomly selected candidate and the weighted distance between it and the current concentration were involved in the updating equation.

2 Methodology

The combination of the strengths of both algorithms was used in the hybridization process which effectively explores and exploits the search space to find high-quality solutions needed. This involves using the “generation rate/Exponential” term from EO to control the balance between exploration and exploitation during the local

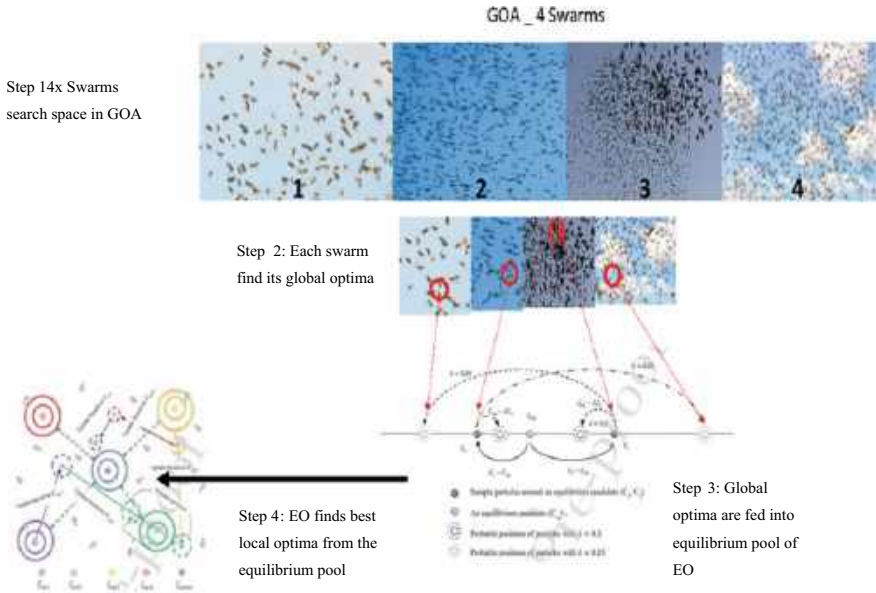


Fig. 1 Four main processes of the hybrid algorithm

search process. The hybrid approach used the GOA to perform a global search of a much wider search space (4×4 times) during the initial phase of the algorithm by using the swarming behaviour of GOA to guide the search agents towards promising regions of the search space. Once the search agents have identified promising regions, the search is switched to EO to perform a local search within those regions as shown in Fig. 1.

Pseudocode of the Hybrid Algorithm

The procedures of the proposed hybrid algorithms is given in the following pseudocode.

1. Initialization: number of agents for Swarm1, Swarm2, Swarm3 and Swarm4.
 2. Set values of Swarms: $C_{max} = [1,0.00001 \mid 1,0.00005 \mid 1,0.0001 \mid 1, 0.0005]$.
 3. Set values of Iterations and dimensions for each swarm [100], upper and lower bounds [1,0].
 4. Declare T_1, T_2, T_3 and T_4 as best search agent | set fitness value.
 5. While($1 < \text{Iterations}$) | (for Each swarm).
 6. Update C (social interaction values).
 7. For each search agent l and Each swarm.
 8. Normalize the distances between search agents.
 9. Update positions of current search agents.
 10. Bring back search agents who got outside of their respective swarms.
- End for.

11. Update T_1, T_2, T_3 and T_4 .
End while.
12. Input T_1, T_2, T_3 and T_4 as $C_{eq1}, C_{eq2}, C_{eq3}$ and C_{eq4} for equilibrium pool Set C_{avg} .
13. For $i = 1$: number of global optima-s search agents.
14. Randomly choose one candidate from the Eq-pool.
15. Generate random vectors as λ and R for constructing F and G_{CP} .
16. $F = a_1 \text{sign}(R-0.5)$ [Exponential term in EO].
17. $G_{CP} = \text{limit } 0.5R$ with [Generation term in EO].
18. Construct G_0 (initial value of G_{CP} value) and G (Control parameter of G_{CP}).
19. Update concentrations of $C = C_{eq} + (C - C_{eq}) \cdot [F + G_0/G]$.
End For.

3 Results and Discussion

A set of benchmark optimization problems (F1_CEC2005 to F25_CEC2005) that are commonly used to evaluate the performance of optimization algorithms was chosen. These problems cover a range of difficulty levels and include both unimodal and multimodal functions. The hybrid algorithm was run on each of the benchmark problems and its performance was recorded where the quality of the solutions (Large real search space) found by the algorithm, its convergence speed and robustness were measured. The performance of the hybrid algorithm to other well-known optimization algorithms, including GOA and EO to evaluate the effectiveness of the hybrid algorithm and determine advantages over the individual algorithms were compared.

Figure 2 illustrates the benchmarking tests that identify the strengths and weaknesses of the hybrid algorithm in terms of swarm size, the accuracy of finding Local optima from Global optima and the computational performance of the algorithm.

4 Conclusion and Further Works

The hybrid approach utilizes the exploration and exploitation abilities of Equilibrium Optimizer (EO) to enhance the search process provided by the Global optima of swarms in the Grasshopper Optimization Algorithm (GOA) where the “generation rate” term from EO was incorporated into the update rule of GOA. This term has been shown to invigorate EO’s ability in exploration, exploitation, and local minima avoidance. By incorporating this term into GOA, it is possible to enhance its search process was enhanced and its performance improves. New hybrids of these algorithms will be proposed with other families of swarm intelligence to solve problems in various fields.

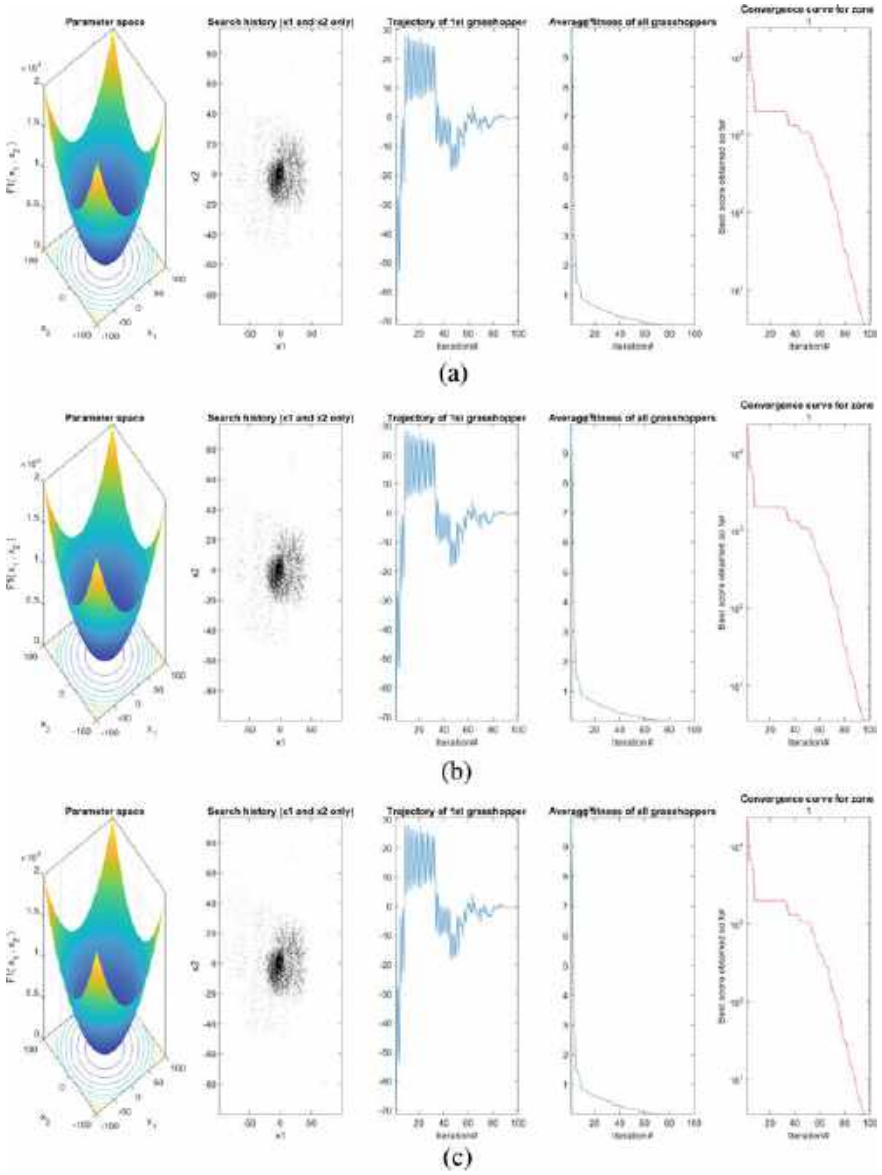


Fig. 2 a–e Zones (1)–(4); f Global optimum and best optimal value for GOA

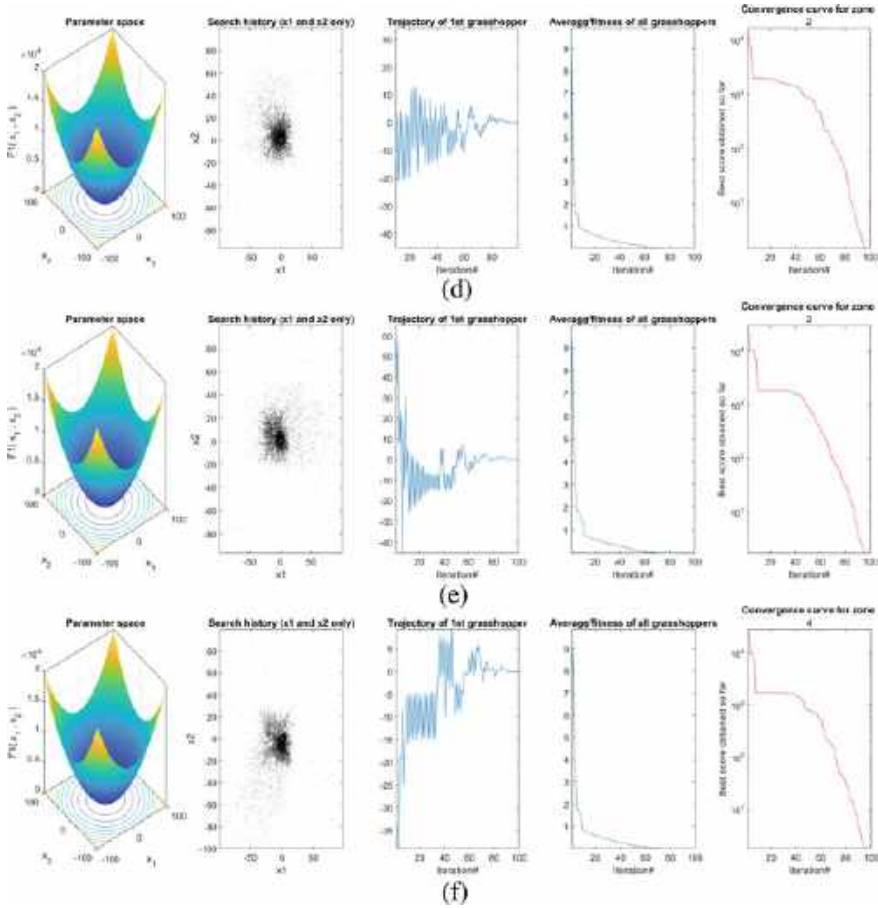


Fig. 2 (continued)

References

1. Dorigo M. Ant colony optimization: a new meta-heuristic
2. Fausto F, Reyna-Orta A, Cuevas E, Andrade ÁG, Perez-Cisneros M (2020) From ants to whales: metaheuristics for all tastes. *Artif Intell Rev* 53(1):753–810. <https://doi.org/10.1007/s10462-018-09676-2>
3. Tan WH, Mohamad-Saleh J (2023) A hybrid whale optimization algorithm based on equilibrium concept. *Alex Eng J* 68:763–786. <https://doi.org/10.1016/j.aej.2022.12.019>
4. Rai R, Dhal KG (2023) Recent Developments in equilibrium optimizer algorithm: its variants and applications. *Arch Comput Methods Engineering*. <https://doi.org/10.1007/s11831-023-09923-y>
5. Zhong C, Li G, Meng Z, Li H, He W (2023) A self-adaptive quantum equilibrium optimizer with artificial bee colony for feature selection. *Comput Biol Med* 153. <https://doi.org/10.1016/j.combiomed.2022.106520>

6. Agrawal U, Rohatgi V, Katarya R (2022) Normalized Mutual Information-based equilibrium optimizer with chaotic maps for wrapper-filter feature selection. *Expert Syst Appl* 207. <https://doi.org/10.1016/j.eswa.2022.118107>
7. Ramachandran M, Mirjalili S, Nazari-Heris M, Parvathysankar DS, Sundaram A, Charles Gnanakkan CAR (2022) A hybrid Grasshopper Optimization Algorithm and Harris Hawks Optimizer for combined heat and power economic dispatch problem. *Eng Appl Artif Intell* 111. <https://doi.org/10.1016/j.engappai.2022.104753>
8. Elmanakhly DA, Saleh MM, Rashed EA (2021) An Improved equilibrium optimizer algorithm for features selection: methods and analysis. *IEEE Access* 9:120309–120327. <https://doi.org/10.1109/ACCESS.2021.3108097>

Vision, Image and Signal Processing

An Enhanced Double EWMA Chart for Monitoring the Process Mean Shifts



Peh Sang Ng, Sook Yan Goh, Sajal Saha, and Wai Chung Yeong

Abstract Exponentially weighted moving average (EWMA) chart is one of the well-known memory-type charts that used in detecting small to moderate disturbances in the process mean. In literature, the EWMA chart with auxiliary information concept (EWMA-AIC) was shown to outperform the traditional EWMA chart. On similar lines, to further enhance the shift detection performance of the EWMA-AIC chart, we propose the double exponentially weighted moving average chart with auxiliary information concept (DEWMA-AIC) in monitoring the process mean. The Monte Carlo simulation is used to compute the run length characteristics of the DEWMA-AIC chart, which include the average run length (ARL), standard deviation of the run length (SDRL) and expected average run length (EARL). The results reveal that the DEWMA-AIC chart is more sensitive than the traditional DEWMA and EWMA-AIC charts in shift detection.

Keywords Double EWMA · Auxiliary information concept · Average run length · Standard deviation of the run length · Expected average run length

1 Introduction

Statistical process control (SPC) is a process improvement procedure that utilize statistical techniques to monitor and control the production process. Control chart is a prevalent tool in SPC toolkit. The control chart is used to ensure the stability

P. S. Ng (✉) · S. Y. Goh

Department of Physical and Mathematical Science, Faculty of Science, Universiti Tunku Abdul Rahman, 31900 Kampar, Perak, Malaysia
e-mail: psng@utar.edu.my

S. Saha

Department of Mathematics, International University of Business Agriculture and Technology, Dhaka, Bangladesh

W. C. Yeong

School of Mathematical Sciences, Sunway University, 47500 Petaling Jaya, Malaysia

of the process by identifying and removing the assignable cause throughout the process monitoring. Various applications of control charts are widely applied in manufacturing and services sectors, such as the food and chemical industries, clinical and diagnostic services and investment sector [1, 2].

The Shewhart chart is a well-known chart due to its operational simplicity as well as its ability to detect large changes in process mean quickly. However, the Shewhart chart is less sensitive in detecting small to moderate changes. This motivates the development of the exponentially weighted moving average (EWMA) and the cumulative sum charts by Roberts [3] and Page [4], respectively, which results in quick detection of small and moderate changes. To further improve the sensitivity of the EWMA chart, the double EWMA (DEWMA) chart which considers the exponential smoothing twice in the plotting statistics was proposed by Zhang and Chen [5]. They concluded that the performance of the DEWMA chart is better than the EWMA chart. For more recent work on the DEWMA chart, readers can refer to the study by Alevizakos et al. [6].

The DEWMA chart by Zhang and Chen [5] only considers the information of study variable in monitoring the process mean. For efficiently estimating the process parameter(s), Riaz [7] suggested the used of auxiliary information based estimator that considers both information of auxiliary and study variables in designing the Shewhart chart. The Shewhart chart with auxiliary information concept (AIC) was shown to surpass the traditional Shewhart chart. Subsequently, Abbas et al. [8] developed the EWMA-AIC chart and the results show that the proposed chart surpasses the traditional EWMA chart in detecting process mean changes. Ng et al. [9] proposed the run sum-AIC chart and the results reveal that the competing EWMA-AIC chart prevails over the run sum-AIC chart in detecting small changes when the correlation between the auxiliary and study variables is less than 0.75.

In this article, to further enhance the DEWMA chart in detecting small to moderate changes in the process mean, we integrate the AIC into the DEWMA chart, named as the DEWMA-AIC chart. The Monte Carlo simulation is used to compute the run length characteristics of the DEWMA-AIC chart. These run length characteristics include the average run length (ARL), standard deviation of the run length (SDRL) and expected average run length (EARL). The run length performances of the proposed chart are then compared with the existing EWMA-AIC chart.

Hereafter, we present the properties of the DEWMA-AIC chart. In Sect. 3, the run length computation, comparison and evaluation are presented. Finally, Sect. 4 concludes the main findings of the research.

2 Proposed DEWMA-AIC Chart

In this section, we extend the work of Zhang and Chen [5] by integrating the DEWMA chart with AIC to efficiently monitor the process mean. Here, the DEWMA chart with a single smoothing constant is selected as it gives better performance [5].

Suppose we have two types of characteristics in a process, say U and V , where U is the quality characteristic under study and V is the auxiliary characteristic that is correlated to U . The correlated coefficient between these two characteristics is denoted as ρ . Let assume a bivariate process (U_r, V_r) at time r follows a bivariate normal distribution, that is, $(U_r, V_r) \sim N_2(\mu_U, \mu_V, \sigma_U, \sigma_V, \rho)$ where (μ_U, μ_V) and (σ_U, σ_V) are the population means and variances, respectively, for study variable U and auxiliary variable V . Note that when the underlying process is in-control, $\mu_U = \mu_{U0}$. Otherwise, $\mu_U = \mu_{U0} + \delta\sigma_U = \mu_{U1}$, where δ is the size of the standardized mean shift.

Following Riaz [7], the regression estimator for population mean μ_U is defined as

$$Q_{U_r} = \bar{U}_r + \rho \left(\frac{\sigma_U}{\sigma_V} \right) (\mu_V - \bar{V}_r) \tag{1}$$

where \bar{U}_r and \bar{V}_r are the r th sample means of U and V , respectively, i.e., $\bar{U}_r = \frac{1}{n} \sum_{j=1}^n U_{r,j}$ and $\bar{V}_r = \frac{1}{n} \sum_{j=1}^n V_{r,j}$.

Based on the regression estimator Q_{U_r} in Eq. (1), the plotting statistic of the DEWMA-AIC chart, D_r is obtained as

$$\begin{aligned} A_r &= \lambda Q_{U_r} + (1 - \lambda)A_{r-1}, \\ D_r &= \lambda A_r + (1 - \lambda)D_{r-1}, \end{aligned} \tag{2}$$

where $\lambda \in (0, 1]$ is the smoothing constant of the DEWMA-AIC chart and Q_{U_r} has been defined in Eq. (1). The mean and variance for the D_r are then defined as

$$E(D_r) = \mu_{U0} \tag{3}$$

and

$$\begin{aligned} \text{Var}(D_r) &= \frac{\sigma_U^2 (1 - \rho^2) \lambda^4}{n(1 - (1 - \lambda)^2)^3} [1 + (1 - \lambda)^2 - (1 - \lambda)^{2r} (r^2 + 2r + 1) \\ &\quad + (1 - \lambda)^{2(r+1)} (2r^2 + 2r - 1) - r^2 (1 - \lambda)^{2(r+2)}], \end{aligned} \tag{4}$$

respectively.

The upper (UCL_r) and lower (LCL_r) control limits of the DEWMA-AIC chart are then derived as follows:

$$UCL_r = \mu_{U0} + k\sqrt{\text{Var}(D_r)}, \tag{5a}$$

and

$$LCL_r = \mu_{U0} + k\sqrt{\text{Var}(D_r)}, \tag{5b}$$

where the parameter k is the control limit coefficient.

The implementation steps for the DEWMA-AIC chart with known δ are given as:

1. Specify the desired values of n , λ , ρ , δ and in-control ARL (ARL_0).
2. Search for k that results in the desired ARL_0 using an extensive Monte Carlo simulation.
3. Determine UCL_r and LCL_r by using Eqs. (5a) and (5b), respectively.
4. Compute the DEWMA-AIC chart's plotting statistic D_r as in Eq. (2).
5. Declare the process as in-control if D_r falls between UCL_r and LCL_r . Otherwise, the process is out-of-control.

3 Computation and Comparisons of Run Lengths

Following Alevizakos et al. [6] and Abbas et al. [8], we used the Monte Carlo simulation method to compute the run length characteristics such as (i) the ARL and SDRL values when the deterministic shift size is known and (ii) the EARL values when the deterministic shift size is unknown, for both the DEWMA-AIC and the EWMA-AIC charts. Here, the ARL (EARL) is the (expected) average number of samples needed to signal an off-target process when the deterministic shift size is known (deterministic shift size is unknown). The SDRL gives us the information of the stability of the ARL.

By assuming that the underlying process follows a bivariate normal distribution, the ARL, SDRL and EARL values of the EWMA-AIC and DEWMA-AIC charts are computed where each simulation run comprises of 50 000 replications. Table 1 presents the ARL, SDRL and the corresponding parameter k value for the EWMA-AIC and DEWMA-AIC charts. We consider $n = 5$, and different values of λ , δ and ρ , where $\lambda \in \{0.05, 0.10, 0.15, 0.30, 0.50\}$, $\rho = \{0, 0.5, 0.95\}$ and $\delta \in \{0, 0.10, 0.30, 0.50, 1.00, 1.50, 2.00\}$. Here, the λ values are set to be similar to the study by Zhang and Chen [5] for the DEWMA chart, except $\lambda = 0.03, 0.08$ and 0.20 are not considered here due to space constraints. Worth to mention that the DEWMA-AIC chart reduces to the traditional DEWMA chart when $\rho = 0$.

As can be observed in Table 1, the DEWMA-AIC chart generally performs better than the DEWMA (i.e., $\rho = 0$) chart by giving smaller out-of-control ARL (ARL_1) and SDRL ($SDRL_1$) values for different shift sizes (i.e., $\delta > 0$) across different ρ values. In addition, the DEWMA-AIC chart results in smaller (ARL_1) and SDRL ($SDRL_1$) values when ρ increases. For example, the (ARL_1 , $SDRL_1$) values for the DEWMA-AIC chart are obtained as (47.91, 50.26) for $\rho = 0$, (39.37, 40.02) for $\rho = 0.50$ and (8.28, 7.43) for $\rho = 0.95$, when $\lambda = 0.05$ and $\delta = 0.10$ are considered. This shows that the AIC significantly enhances the detection ability of the DEWMA chart in process monitoring. Table 1 also shows that an increase in λ results in an increase in the ARL_1 and $SDRL_1$ values, which means larger samples are needed to detect out-of-control situation when λ increases. As expected, the ARL_1 and $SDRL_1$ values decrease when the shift size increases as a larger shift size can be detected easily and

Table 1 Run length characteristics of the EWMA-AIC and DEWMA-AIC charts when $\lambda = \{0.05, 0.10, 0.30, 0.50\}$, $\rho = \{0, 0.5, 0.95\}$ and different values of δ at $ARL_0 = 200$

	EWMA-AIC	DEWMA-AIC	EWMA-AIC	DEWMA-AIC	EWMA-AIC	DEWMA-AIC
			$\lambda = 0.05$			
$\rho = 0$			$\rho = 0.50$		$\rho = 0.95$	
δ	$k = 2.2769$	$k = 1.7073$	$k = 2.2767$	$k = 1.7085$	$k = 2.2755$	$k = 1.7091$
0	(200.0, 216.0)	(200.1, 244.8)	(200.0, 216.4)	(200.0, 245.2)	(200.0, 215.9)	(200.0, 246.4)
0.10	(57.19, 55.09)	(47.91, 50.26)	(46.83, 43.67)	(39.37, 40.02)	(9.65, 7.36)	(8.28, 7.43)
0.30	(10.75, 8.22)	(9.22, 8.27)	(8.53, 6.35)	(7.28, 6.47)	(1.84, 0.97)	(1.50, 0.86)
0.50	(4.77, 3.25)	(3.96, 3.33)	(3.81, 2.50)	(3.14, 2.53)	(1.10, 0.31)	(1.03, 0.18)
0.75	(2.58, 1.53)	(2.08, 1.47)	(2.11, 1.17)	(1.70, 1.08)	(1.00, 0.03)	(1.00, 0.01)
1.00	(1.74, 0.89)	(1.43, 0.77)	(1.47, 0.67)	(1.24, 0.54)	(1.00, 0)	(1.00, 0)
1.50	(1.15, 0.37)	(1.05, 0.23)	(1.05, 0.23)	(1.01, 0.12)	(1.00, 0)	(1.00, 0)
			$\lambda = 0.10$			
$\rho = 0$			$\rho = 0.50$		$\rho = 0.95$	
δ	$k = 2.4782$	$k = 1.9919$	$k = 2.4799$	$k = 1.9896$	$k = 2.4779$	$k = 1.9881$
0	(200.0, 206.3)	(20.0, 220.64)	(200.0, 206.6)	(200.1, 220.0)	(200.0, 206.6)	(200.0, 219.8)
0.10	(69.61, 67.15)	(57.86, 57.55)	(57.20, 54.14)	(47.12, 45.44)	(11.15, 8.24)	(9.67, 7.62)
0.30	(12.42, 9.27)	(10.74, 8.56)	(9.79, 7.07)	(8.50, 6.68)	(2.02, 1.06)	(1.69, 0.99)
0.50	(5.41, 3.52)	(4.69, 3.51)	(4.30, 2.71)	(3.70, 2.72)	(1.14, 0.36)	(1.06, 0.24)
0.75	(2.88, 1.66)	(2.43, 1.64)	(2.34, 1.27)	(1.95, 1.23)	(1.00, 0.04)	(1.00, 0.02)
1.00	(1.91, 0.97)	(1.61, 0.90)	(1.59, 0.74)	(1.35, 0.64)	(1.00, 0)	(1.00, 0)
1.50	(1.20, 0.43)	(1.09, 0.31)	(1.08, 0.28)	(1.03, 0.18)	(1.00, 0)	(1.00, 0)
			$\lambda = 0.30$			
$\rho = 0$			$\rho = 0.50$		$\rho = 0.95$	
δ	$k = 2.7174$	$k = 2.4678$	$k = 2.7184$	$k = 2.4681$	$k = 2.7193$	$k = 2.4663$
0	(200.1, 199.8)	(200.01, 201.7)	(200.0, 199.0)	(200.1, 202.7)	(200.0, 199.6)	(200.0, 202.0)
0.10	(103.7, 102.7)	(85.16, 83.47)	(88.84, 87.60)	(70.57, 68.71)	(16.2, 13.81)	(12.51, 9.93)
0.30	(18.28, 15.91)	(14.09, 11.41)	(13.82, 11.46)	(10.87, 8.36)	(2.29, 1.21)	(2.06, 1.12)
0.50	(6.80, 4.83)	(5.75, 3.85)	(5.25, 3.49)	(4.54, 2.89)	(1.20, 0.42)	(1.14, 0.36)
0.75	(3.36, 1.97)	(2.99, 1.76)	(2.68, 1.47)	(2.41, 1.35)	(1.00, 0.06)	(1.00, 0.04)
1.00	(2.15, 1.10)	(1.95, 1.03)	(1.76, 0.84)	(1.60, 0.78)	(1.00, 0)	(1.00, 0)
1.50	(1.28, 0.49)	(1.20, 0.43)	(1.13, 0.34)	(1.08, 0.28)	(1.00, 0)	(1.00, 0)
			$\lambda = 0.50$			

(continued)

Table 1 (continued)

	EWMA-AIC	DEWMA-AIC	EWMA-AIC	DEWMA-AIC	EWMA-AIC	DEWMA-AIC
	$\rho = 0$		$\rho = 0.50$		$\rho = 0.95$	
δ	$k = 2.7795$	$k = 2.6747$	$k = 2.7805$	$k = 2.6768$	$k = 2.7789$	$k = 2.6747$
0	(200.0, 198.7)	(200.0, 199.8)	(200.1, 200.9)	(200.1, 200.1)	(200.1, 198.7)	(200.0, 199.6)
0.10	(126.8, 126.1)	(107.3, 106.1)	(112.2, 112.0)	(92.39, 91.64)	(23.13, 21.46)	(16.61, 14.6)
0.30	(26.6, 25.2)	(18.99, 16.95)	(19.76, 18.04)	(14.16, 12.16)	(2.47, 1.41)	(2.23, 1.18)
0.50	(8.88, 7.33)	(6.77, 4.98)	(6.52, 5.05)	(5.18, 3.56)	(1.22, 0.45)	(1.19, 0.41)
0.75	(3.84, 2.56)	(3.29, 1.95)	(2.95, 1.80)	(2.62, 1.44)	(1.00, 0.07)	(1.00, 0.06)
1.00	(2.31, 1.28)	(2.11, 1.07)	(1.85, 0.94)	(1.73, 0.82)	(1.00, 0)	(1.00, 0)
1.50	(1.31, 0.53)	(1.27, 0.48)	(1.14, 0.36)	(1.12, 0.33)	(1.00, 0)	(1.00, 0)

thus out-of-control state can be determined quickly which results in a smaller spread in the run length distribution.

Table 2 presents the EARL values for the DEWMA-AIC and EWMA-AIC charts. The implementation steps and the parameters used in computing the EARL values are similar to that of the ARL except δ is replaced by $(\delta_{\min}, \delta_{\max}) = \{(0.1, 0.5), (0.5, 1.0), (1.0, 1.5), (1.5, 2.0)\}$. The observations found in Table 2 are similar to that of the ARL characteristic observed in Table 1. It is found that a smaller value of λ , a larger value of ρ or $(\delta_{\min}, \delta_{\max})$ results in a smaller out-of-control EARL ($EARL_1$) value.

Based on the run length comparisons in Tables 1 and 2, the DEWMA-AIC chart generally outperforms the existing EWMA-AIC chart in detecting off-target process by resulting in smaller ARL_1 and $EARL_1$ values when the exact shift size is known and unknown, respectively. For example, when $\lambda = 0.05$, $\rho = 0.5$ and $\delta = 0.1$ are considered in Table 1, the ARL_1 value for the DEWMA-AIC chart is 39.37, which is lower than 46.83 for the EWMA-AIC chart. In terms of $SDRL_1$, Table 1 shows that the DEWMA-AIC chart gives smaller $SDRL_1$ values than the EWMA-AIC chart for most of the cases, which indicates that the run lengths of the proposed chart are closed to the ARL_1 . For instance, when $\lambda = 0.10$, $\rho = 0.50$ and $\delta = 0.10$ are considered, the $(ARL_1, SDRL_1)$ values for the EWMA-AIC and DEWMA-AIC AIC charts are obtained as (57.20, 54.14) and (47.12, 45.44), respectively. As a result, based on the ARL, $SDRL$ and EARL performances, the proposed chart is shown to be more sensitive than the EWMA-AIC chart in shift detection.

Table 2 Run length characteristics of the EWMA-AIC and DEWMA-AIC charts when $\rho = \{0, 0.5, 0.95\}$ and various pairs of $(\delta_{\min}, \delta_{\max})$ at $ARL_0 = 200$

	EWMA-AIC	DEWMA-AIC	EWMA-AIC	DEWMA-AIC	EWMA-AIC	DEWMA-AIC
			$\lambda = 0.05$			
	$\rho = 0$		$\rho = 0.5$		$\rho = 0.95$	
δ_{\min}	$k = 2.2769$	$k = 1.7073$	$k = 2.2767$	$k = 1.7085$	$k = 2.2755$	$k = 1.7091$
0.1	16.14	13.66	12.96	10.99	2.73	2.30
0.5	2.80	2.28	2.28	1.86	1.01	1.00
1.0	1.38	1.19	1.21	1.09	1.00	1.00
1.5	1.06	1.02	1.02	1.00	1.00	1.00
			$\lambda = 0.10$			
	$\rho = 0$		$\rho = 0.5$		$\rho = 0.95$	
δ_{\min}	$k = 2.4782$	$k = 1.9919$	$k = 2.4799$	$k = 1.9896$	$k = 2.4779$	$k = 1.9881$
0.1	19.16	16.16	15.32	12.91	3.05	2.63
0.5	3.13	2.67	2.53	2.14	1.02	1.01
1.0	1.48	1.28	1.28	1.14	1.00	1.00

(continued)

Table 2 (continued)

		EWMA-AIC	DEWMA-AIC	EWMA-AIC	DEWMA-AIC	EWMA-AIC	DEWMA-AIC
1.5	2.0	1.09	1.03	1.03	1.01	1.00	1.00
		EWMA-AIC	DEWMA-AIC	EWMA-AIC	DEWMA-AIC	EWMA-AIC	DEWMA-AIC
				$\lambda = 0.30$			
		$\rho = 0$		$\rho = 0.5$		$\rho = 0.95$	
δ_{\min}	δ_{\max}	$k = 2.7174$	$k = 2.4678$	$k = 2.7184$	$k = 2.4681$	$k = 2.7193$	$k = 2.4663$
0.1	0.5	29.21	22.80	23.20	18.04	3.78	3.22
0.5	1.0	3.72	3.27	2.94	2.62	1.03	1.02
1.0	1.5	1.63	1.50	1.37	1.28	1.00	1.00
1.5	2.0	1.13	1.09	1.05	1.03	1.00	1.00
						$\lambda = 0.50$	
		$\rho = 0$		$\rho = 0.5$		$\rho = 0.95$	
δ_{\min}	δ_{\max}	$k = 2.7795$	$k = 2.6747$	$k = 2.7805$	$k = 2.6768$	$k = 2.7789$	$k = 2.6747$
0.1	0.5	39.74	30.30	31.86	24.06	4.69	3.76
0.5	1.0	4.40	3.65	3.34	2.88	1.04	1.03
1.0	1.5	1.71	1.60	1.41	1.35	1.00	1.00
1.5	2.0	1.15	1.12	1.05	1.04	1.00	1.00

4 Conclusion

This study improves the efficiency of the DEWMA chart in process shift detection by considering the AIC into the DEWMA chart. The run length characteristics such as the ARL, SDRL and EARL of the DEWMA-AIC chart are computed by using Monte Carlo simulations. The run length comparison reveals that the DEWMA-AIC chart is able to detect mean shift substantially quicker than the existing DEWMA and EWMA-AIC charts.

References

1. Tran KP, Castagliola P, Celano G, Khoo MB (2018) Monitoring compositional data using multivariate exponentially weighted moving average scheme. *Qual Reliab Eng Int* 34(3):391–402
2. Yeong WC, Khoo MB, Tham LK, Teoh WL, Rahim MA (2017) Monitoring the coefficient of variation using a variable sampling interval EWMA chart. *J Qual Technol* 49(4):380–401
3. Roberts SW (1959) Control chart tests based on geometric moving averages. *Technometrics* 1(3):239–250
4. Page E (1954) Continuous inspection schemes. *Biometrika* 41(1/2):100–115
5. Zhang L, Chen G (2005) An extended EWMA mean chart. *Qual Technol Quant Manage* 2(1):39–52
6. Alevizakos V, Chatterjee K, Koukouvinos C (2022) Modified EWMA and DEWMA control charts for process monitoring. *Commun Statist Theory Methods* 51(21):7390–7412
7. Riaz M (2008) Monitoring process mean level using auxiliary information. *Stat Neerl* 62(4):458–481
8. Abbas N, Riaz M, Does RJ (2014) An EWMA-type control chart for monitoring the process mean using auxiliary information. *Commun Statist Theory Methods* 43(16):3485–3498
9. Ng PS, Khoo MBC, Saha S, Teh SY (2018) Run sum chart for the mean with auxiliary information. *J Test Eval* 48(2):1554–1575

Face Recognition Attendance Management System (FRAMS) Algorithm Using CNN Model



Saw Yang Yi, Mohd Izzat Nordin, and Mohamad Tarmizi Abu Seman

Abstract This paper discusses the development of the face recognition attendance management system, FRAMS. In the development of FRAMS, there are two important stages which are face detection algorithm development and face recognition algorithm development. In face detection algorithm development, three proposed face detection methods which are Viola-Jones Haar Cascade Classifier, Local Binary Pattern, LBP and Multi-Task Cascaded Convolutional Neural Networks, MTCNN are used to evaluate their performance in terms of face detection accuracy and total detection time required by those methods. From those experiments, it can be known that MTCNN is the best method as it provides a 100% face detection accuracy compared to another two proposed methods although the processing time is longer than another two methods. In face recognition algorithm development, the pre-trained VGG-16 CNN model is used to perform transfer learning by loading the train and validation image datasets. The confusion matrix is plotted out to evaluate the performance of the trained CNN model. The trained VGG-16 CNN model achieved an accuracy of 99%. Finally, a complete FRAMS has been developed and able to recognise the face feature of students with extremely few misclassification mistakes.

Keywords FRAMS · Face detection · MTCNN · Face recognition · VGG-16 CNN model first section

1 Introduction

The attendance management system is a system that enables the administration to produce daily attendance of the lecture class while not using an inordinate amount of time. Face recognition is a biometric technique [1] that is used in a system to match the face received from an image or video against the information about the faces

S. Y. Yi · M. I. Nordin · M. T. A. Seman (✉)

School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Engineering Campus, 14300 Nibong Tebal, Pulau Pinang, Malaysia

e-mail: mohdtarmizi@usm.my

created by the system administrator. Face recognition necessitated the completion of two critical steps: face detection and face identification [2]. In the face detection process, human facial characteristics are retrieved from picture databases. The classifier will carry out the face feature comparison procedure between the humans' faces acquired by the camera and the humans' faces in the current database during the face identification stage.

Deep Learning is a subset of both machine learning and artificial intelligence which can be achieved by using a multilayer neural network to learn data representations. The deep learning model is used to precise the features and train itself to categories itself based on the exact features of the datasets acquired to generate a more accurate outcome by using the transfer learning method.

Because of the Covid-19 epidemic, all face-to-face lectures in schools and colleges have been replaced with online-based classes. Traditional methods of collecting attendance are not appropriate for use in online classrooms. These situations become a challenge to lectures, especially when considering the students' attendance during the live online class [3]. As a result, a Face Recognition Attendance Management System (FRAMS) with the integration of deep learning VGG-16 pre-trained models is proposed in this paper to overcome the problems faced by lectures especially collecting attendance during an online class. The objectives of this paper are to create a FRAMS using the Convolution Neural Network (CNN) as the face recognition algorithm and to analyse and find out the suitable face detection algorithm to pre-process the image dataset of the FRAMS and assist FRAMS to detect the face.

The rest of the paper is organised as follows. Section 3 discusses the result and discussions obtained from the research. Finally, the paper is concluded in Sect. 4.

2 Methodology

2.1 Image Datasets

Figure 1 shows the image of Person A, Person B and Person C. Image of Person A is taken from the dataset of class HW while the image of person C is taken from a dataset of class YY.

2.2 Face Detection Algorithm Development Flow

The image of person B and person C as shown in Fig. 1b, c will be used to perform a single face detection experiment. The image of different facial views will be used to perform multiple face detection experiments to evaluate the performance of each proposed face detection algorithm in terms of the accuracy of the face detection and the detection time which can be expressed in Eq. 1 and Eq. 8 respectively. This idea

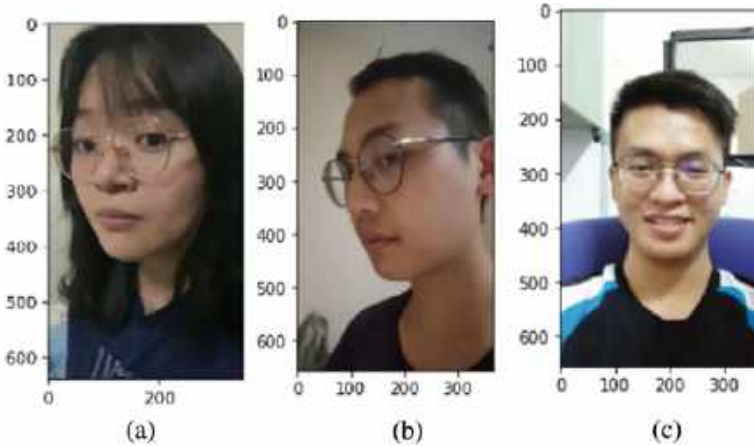


Fig. 1 a Person A, b person B, c person C

of the face detection algorithm development process is inspired by the research paper [4-6].

$$Accuracy = \frac{Number\ of\ the\ face\ detected\ in\ image}{Total\ number\ of\ face\ in\ the\ image} \times 100\% \tag{1}$$

$$Detection\ time = t_{final} - t_{initial} \tag{2}$$

where t_{final} is the final time is taken and $t_{initial}$ is the initial time taken

2.3 Face Recognition Algorithm Development

In this stage, the pre-trained VGG-16 CNN model will be used to perform transfer learning by loading the pre-processing image train and validation datasets of class HW and class YY as shown in Fig. 2 into the model. The epoch used in transfer learning is set to 10. The performance of the CNN model is evaluated based on the accuracy and the loss plot.



Fig. 2 Pre-process image datasets of person A and person C

Fig. 3 Structure of 2×2 confusion matrix

	Predicted Positive	Predicted Negative
Actual Positive	True Positive, TP	False Negative, FN
Actual Negative	False Positive, FP	True Negative, TN

The confusion matrix as shown in Fig. 2 is plotted out by loading the test datasets into the CNN model and several parameters such as Precision, Recall, F1-Score and Accuracy will be calculated by using Eqs. 3, 4, 5 and 6 as shown below (Fig. 3).

$$\text{Precision, } P = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Recall, } R = \frac{TP}{TP + FN} \quad (4)$$

$$\text{F1 Score} = \frac{2 \times P \times R}{P + R} \quad (5)$$

$$\text{Accuracy, } A = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative Once the performance is satisfied, the CNN model will be saved in.pth format and will be used in FRAMS program later.

3 Results and Discussions

3.1 Results Obtained from Face Detection Algorithm Development

i. Single Face Detection Experiment

The image of Person B in image of different facial views was used as the non-frontal face image while the image of Person C in Figure was used as the frontal face image. The result of the frontal face condition experiment and non-frontal face condition experiment were shown in Tables 1 and 2 respectively.

By referring to Table 1, it was concluded that those three face detection methods performed well when detecting the frontal face.

Table 1 Result of the frontal face condition experiment


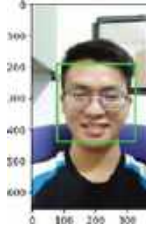



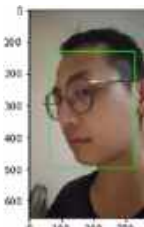
Frontal face			
Methods used			
	Viola-Jones Haar Cascade Classifier	LBP	MTCNN
			
Detection status	Face detected	Face detected	Face detected

Table 2 Result of the non-frontal face condition experiment

Non-Frontal Face			
Methods used			
	Viola-Jones Haar Cascade Classifier	LBP	MTCNN
			
Detection status	Face not detected	Face not detected	Face detected

By referring to Table 2, it was observed that MTCNN could only detect the non-frontal face compared to another 2 proposed methods.

Table 3 showed the face detection accuracy and the total detection time used by each face detection method in Fig. 8 during the multiple face detection experiment conducted.

From Table 3, it was observed that MTCNN have a high face detection accuracy of 100% and zero false case detection compared to another two proposed face detection methods. In terms of detection time, MTCNN took the longest time to perform the face detection process.

From the results obtained from single face detection and multiple face detection experiments, it could be observed that MTCNN achieved the best performance in those two experiments compared to another two proposed face detection methods. This is because MTCNN was able to detect either the frontal face or the non-frontal.

Table 3 Comparison of performance of each face detection methods

Algorithm used	Accuracy (%)			Detection time (s)
	True case detected	False case detected	Not detected	
Viola-Jones Haar Cascade Classifier	$\frac{9}{15} \times 100 = 60$	$\frac{1}{15} \times 100 = 6.67$	$\frac{5}{15} \times 100 = 33.33$	0.577
LBP	$\frac{8}{15} \times 100 = 53.33$	$\frac{0}{15} \times 100 = 0$	$\frac{7}{15} \times 100 = 46.67$	0.283
MTCNN	$\frac{15}{15} \times 100 = 100$	$\frac{0}{15} \times 100 = 0$	$\frac{0}{15} \times 100 = 0$	1.799

Viola-Jones Haar Cascade Classifier could be detected the face view of 0, 90, 180, 270, and 360 degrees during the face detection process [5]. LBP was able to detect the limited poses of the face and was not sensitive to the small changes in face localization. Hence, MTCNN was chosen as the face detection method and would be implemented in FRAMS.

4 Conclusions

In this research, there are two important stages which are face detection algorithm development and face recognition algorithm development. In face detection algorithm development, MTCNN has been chosen as the face detection algorithm which will be used in FRAMS later on due to its good performance with 100% of face detection accuracy without creating any false condition case compared to another two proposed methods which are Viola-Jones Haar Cascade Classifier and LBP. In face recognition algorithm development, the VGG-16 model is trained and able almost properly identify the two different classes of students' face features which gives the accuracy of 99%. Lastly, a complete FRAMS has been developed and able to recognise the face feature of students with extremely few misclassification mistakes. Therefore, the objectives of the research are achieved.

Acknowledgements This work was supported by Universiti Sains Malaysia Short Term Grant 304/PELECT/6315342 and FRGS KPT Grant 203.PELECT.6071535 .

References

1. Setta S, Sinha S, Mishra M, Choudhury P (2021) Real-time facial recognition using SURF-FAST. Springer, Singapore
2. Bhatti K, Mughal L, Khuhawar F, Memon S (2018) Smart attendance management system using face recognition. *EAI Endorsed Trans Creat Technol* 5(17):159713. <https://doi.org/10.4108/eai.13-7-2018.159713>
3. Archana MCP, Nitish CK, Harikumar S (2022) Real time face detection and optimal face mapping for online classes. *J Phys Conf Ser* 2161(1). <https://doi.org/10.1088/1742-6596/2161/1/012063>
4. Filali H, Riffi J, Mahraz AM, Tairi H (2018) Multiple face detection based on machine learning. In: 2018 International conference on intelligent system and computer vision (ISCV 2018), vol. 2018, pp 1–8. <https://doi.org/10.1109/ISACV.2018.8354058>
5. Malhotra S, Aggarwal V, Mangal H, Nagrath P, Jain R (2021) Comparison between attendance system implemented through Haar cascade classifier and face recognition library. *IOP Conf Ser Mater Sci Eng* 1022(1). <https://doi.org/10.1088/1757-899X/1022/1/012045>
6. Khan MZ, Harous S, Hassan SU, Ghani Khan MU, Iqbal R, Mumtaz S (2019) Deep Unified model for face recognition based on convolution neural network and edge computing. *IEEE Access* 7©:72622–72633. <https://doi.org/10.1109/ACCESS.2019.2918275>

A Low-Cost Vibration Measurement Using MEMS MPU9250 Accelerometer with DLPF Filtering



Mohd Affan Mohd Rosli, Abdul Haadi Abdul Manap,
and Ahmad Zhafran Ahmad Mazlan

Abstract Nowadays, the integration of MEMS accelerometer in vibration measurement system such as a condition-based monitoring (CBM) become necessary as it is tiny, low in weight, power consumption and cost. The MPU9250 is one of the cheapest accelerometers that capable to measure large movements and low frequencies. It also has lower noise level and consumes less energy compared to MPU6050. The DLPF enables MPU9250 to be configure by selecting the suitable noise filtration. In this study, an experiment is conducted to validate the MPU9250 accelerometer measurement with the embedded MEMS accelerometer of smart mobile device. The vibration at the end of motorcycle's handlebar is measured and the result shows that the used DLPF filtering enable the MPU9250 to achieve an equal result as embedded MEMS accelerometer of smart mobile device.

Keywords MEMS accelerometer · MPU9250 · Vibration measurement

M. A. M. Rosli · A. H. A. Manap · A. Z. A. Mazlan (✉)
School of Mechanical Engineering, Engineering Campus, Universiti Sains Malaysia, 14300
Nibong Tebal, Pulau Pinang, Malaysia
e-mail: zhafran@usm.my

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
N. S. Ahmad et al. (eds.), *Proceedings of the 12th International Conference on
Robotics, Vision, Signal Processing and Power Applications*, Lecture Notes in Electrical
Engineering 1123, https://doi.org/10.1007/978-981-99-9005-4_50

399

1 Introduction

In general, accelerometers are intended to measure the acceleration of a mechanical system and widely used in vibration measurement due to its compact size with the direct data acquisition obtained. Conventional accelerometers use piezoelectric technology in order to make the relative displacement of the mass with respect to the base proportional to acceleration. This technology uses a spring and mass system, those the natural frequency is significantly higher than the frequency of the measured system [1]. Meanwhile, due to the last decade's tremendous advancements in micro machining, it is now feasible to design mechanical elements with overall dimensions as small as a few micrometers. Micro Electromechanical Systems (MEMS), which contain microstructures, microelectronic parts, and actuators, as well as microsensors, were developed as a result of this to enable the vibration measurement. In navigation applications, gyroscopes, magnetometers, and barometers are often utilized the MEMS sensors [2]. These sensors can now be extremely lightweight, power-efficient, and inexpensively miniaturized to the size of an integrated circuit chip. Using similar approaches, an accelerometer with a total size of a few hundred micrometers have been developed. These MEMS accelerometers are highly affordable and have a very tiny physical size when compared to traditional piezoelectric accelerometers. Nevertheless, deterministic, and stochastic noise significantly affects MEMS sensors, which is regarded as a very complex issue [3]. Figure 1 shows the MEMS-based accelerometer's architecture.

Due to the MEMS accelerometers considerable as on-going technological advancements, they are now being used in a variety of industrial applications. Compared to more conventional uses, some of these accelerometers provide affordable alternatives [4]. The price of each accelerometer is based on its capability, and they may also incorporate with other devices. The cost of acceleration acquisition techniques is not limited to the price of accelerometers (such as real time controller, data acquisition software, and workforce for data analysis). Sensors with shorter acceleration ranges have lower noise densities but the low-cost MEMS usually have higher noise density compared with the traditional commercial alternatives and not offering a vast frequency range. Because MEMS accelerometers were not precise enough to compete with conventional accelerometers in low acceleration ranges, they

Fig. 1 MEMS accelerometer architecture [1]

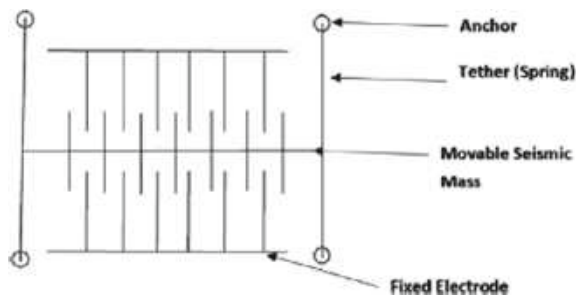


Table 1 Characteristic of MPU6050 and MPU9250 accelerometer [2]

Parameters	MPU6050	MPU9250
Full scale range	$\pm 2 \text{ g}, \pm 4 \text{ g}, \pm 8 \text{ g}, \pm 16 \text{ g}$	$\pm 2 \text{ g}, \pm 4 \text{ g}, \pm 8 \text{ g}, \pm 16 \text{ g}$
Nonlinearity	0.5%	0.5%
Cross axis sensitivity	$\pm 2\%$	$\pm 2\%$
Noise spectral power density	$400 \mu\text{g}/\sqrt{\text{Hz}}$	$300 \mu\text{g}/\sqrt{\text{Hz}}$

were mostly used in projects with large movements and low frequencies [5]. Based on the study, the MPU9250 is the most cost-effective, consumes less energy than MPU6050 and has lower noise levels. The price of MPU9250 is just €5.8 compared to the priciest MEMS accelerometer; the 3713B112G that cost €2070 [5]. Table 1 shows the characteristic of the sensors.

Three-axis accelerometer MPU9250 employs distinct proof weights for each axis. Capacitive sensors measure the displacement difference when an axis of acceleration generates deformation on the associated proof mass. The design of the MPU9250 minimizes the sensitivity of the accelerometers to manufacturing variations and thermal drift. The scale factor of the accelerometers is calibrated at the production and is mostly unaffected by supply voltage [6]. Due to this, the MPU9250 accelerometer is used in this study to measure the vibration of motorcycle's handlebar with the comparison of available MEMS accelerometer in smart mobile device. By applying the DLPF filtering, the performance of the accelerometer is investigated and compared.

2 Methodology

As shown in Fig. 2a, the MPU9250 accelerometer is connected to the Arduino Uno board to measure the acceleration data while the block diagram is shown in Fig. 2b. The VCC pin at the accelerometer is connected to the 5 V pin at the Arduino. Meanwhile the GND pin for both boards are connected. The SCL and SDA pins are attached to the A5 pins, respectively.

The Arduino Uno board shown in Fig. 2 is serially connected to the computer using USB port. Using Arduino IDE software, the file library and codes are programmed together using PYTHON programming language to enable the acceleration modules works. In this study, a single cylinder four stroke engine with 150 cc engine displacement is used. The MPU9250 accelerometer is attached to the non-moveable end of the right side of the motorcycle's handlebar [7], as shown in Fig. 3. The back of the accelerometer board is covered using the insulation tape to prevent the pins connected

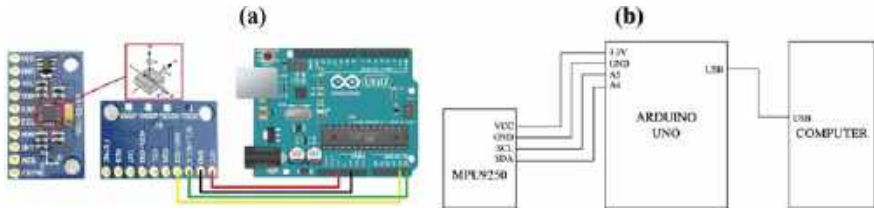


Fig. 2 The connection: **a** MPU9250 and Arduino Uno board for acceleration data measurement; **b** block diagram

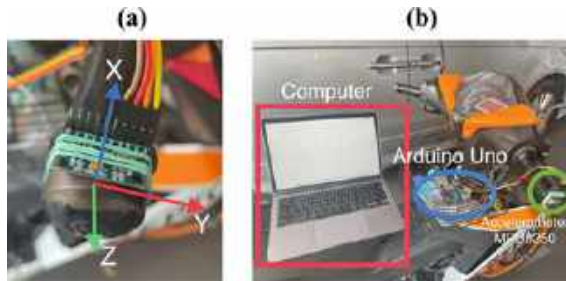


Fig. 3 Setup for handlebar vibration measurement: **a** MPU9250 accelerometer placement at the end of handlebar; **b** connection between the accelerometer, Arduino Uno and computer

directly to the handlebar metal parts, as the short circuit may be occurred. The important part is to make sure the direction of the accelerometer is correct based on x , y and z axis. Since the end of the bar is curvy, a high strength rubber band is used to securely placed the accelerometer, as show in Fig. 3a, while the complete experiment setup is shown in Fig. 3b.

The measurement is carried out at a constant idle speed of engine at 1650 RPM. The sampling rate set at 500 Hz as the MPU9250 supports the rate up to 500 Hz only [6]. All three directions of vibration (x , y , and z axis) are measured and display using the serial plotter of Arduino IDE output. As discussed earlier, the MEMS accelerometer is easily affected by noise [3]. Therefore, in this measurement, 8 states of noise filtering mode by coded Digital Low Pass Filter (DLPF) is used for measurement validation. The validation used embedded MEMS accelerometer in smart mobile device and the real-time vibration data is displayed using Seismic Vibrations Detector version 1.3.

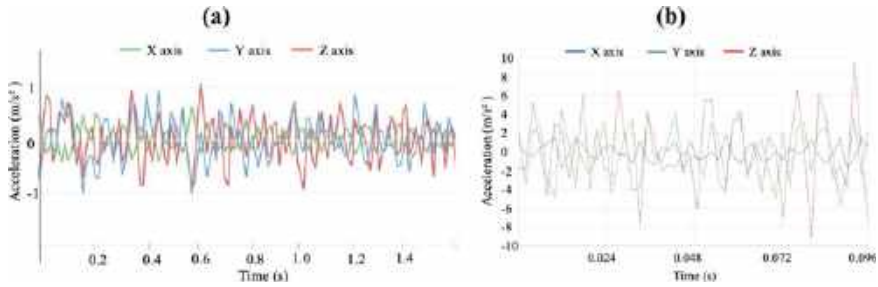


Fig. 4 The vibration of handlebar measured by: **a** MEMS accelerometer in smart mobile device; **b** MEMS accelerometer of MPU9250 without filtering

3 Results and Discussion

The vibration measurement values are recorded using the smart mobile device with the running idling speed of the motorcycle engine at 1650 RPM. From Fig. 4a, the result shows that the peak amplitude *A_{peak}* of the acceleration is about 1.1 m/s². Meanwhile, the vibration measurement using MPU9250 accelerometer without any filtration shows that, at the same motorcycle’s engine condition, the *A_{peak}* of the acceleration is about 9.6 m/s² in z-axis direction, as shown in Fig. 4(b). The large difference of *A_{peak}* values requires the MPU9250 accelerometer to enable a suitable filtering state.

Figure 5a–h show the handlebar vibration measurement results for all the 8 states of DLPF filtering that been selected and measured at each of corresponding running engine speed. In this case, the smart mobile device vibration measurement in Fig. 4a is taken as reference for data validation. Based on the figures, the most equal vibration measurement result of MPU9250 accelerometer compared to the smart mobile device accelerometer with he enabled DLPF filtering is occurred at State 5 (Fig. 5(f), with the peak acceleration amplitude *A_{peak}* of 1.26 m/s² (14.5% error). This filtering can be further tuning at more accurate decimal numbers to achieve lower error percentages between both accelerometers. Meanwhile, other DLPF filtering states differences between 45 and 445% error.

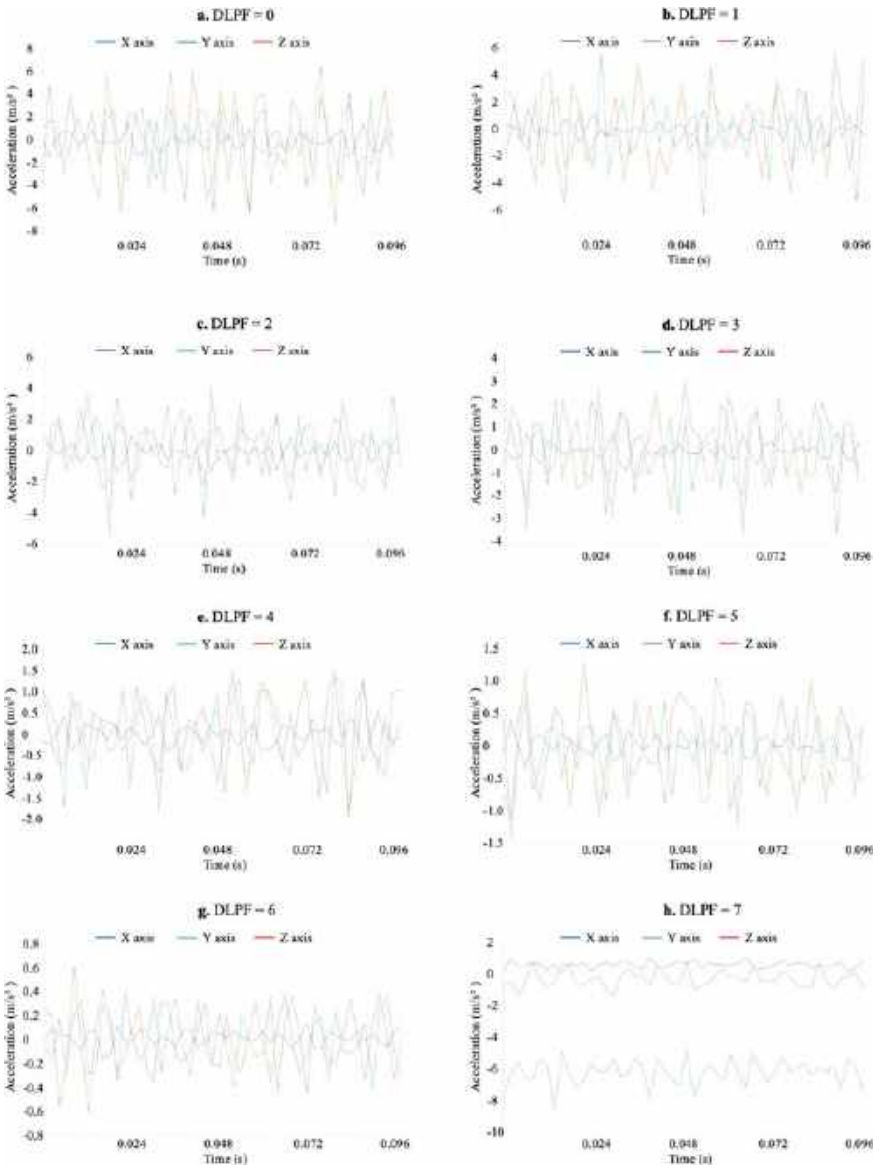


Fig. 5 DLPF filtering enable for vibration measurement using MPU9250 accelerometer: **a** DLPF = 0; **b** DLPF = 1; **c** DLPF = 2; **d** DLPF = 3; **e** DLPF = 4; **f** DLPF = 5; **g** DLPF = 6; **h** DLPF = 7

4 Conclusion

This study presents the implementation of low-cost MEMS accelerometer using MPU9250 for vibration measurement of motorcycle's handlebar with the comparison of embedded MEMS accelerometer of smart mobile device. For the acceleration validation, the best noise DLPF filtration is selected at State 5, whereby both measured peak acceleration amplitude A_{peak} is occurred at 1.1 m/s^2 and 1.26 m/s^2 , respectively. From the experiment, it is also observed that the vibration movement are actively accelerate in y -axis and z -axis acceleration for all mode of configurations.

Acknowledgements The authors would like to thank the Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme (FRGS) with Project Code: FRGS/1/2021/TK0/USM/03/6.

References

1. Guru Manikandan K, Pannirselvam K, Kenned JJ, Suresh Kumar C (2021) Investigation on suitability of MEMS based accelerometer for vibration measurements. *Mater Today Proc* 45:6183–6192
2. Tjhai C, O'Keefe K (2019) Using step size and lower limb segment orientation from multiple low-cost wearable inertial/magnetic sensors for Pedestrian navigation. *Sensors* 19:3140
3. Hayouni M, Vuong T-H, Choubani F (2022) Wireless IoT universal approach based on Allan variance method for detection of artificial vibration signatures of a DC motor's shaft and reconstruction of the reference signal. *IET Wirel Sens Syst* 12:81–92
4. Mark L (2014) An introduction to MEMS vibration monitoring. *Analog Dialogue* 48(6):1–3
5. Komarizadehasl S, Mobaraki B, Ma H, Lozano-Galant JA, Turmo J (2021) Development of a low-cost system for the accurate measurement of structural vibrations. *Sensors* 21:6191
6. MPU-9250 Product Specification Revision 1.1: InvenSense Inc. Indianapolis, CA 95110 U.S.A. (2016)
7. Usamah NM, Mazlan AZA, Ripin ZM (2022) Investigation of motorcycle handle vibration attenuation using a suspended handlebar with different rubber mount characteristics. *Int J Acoust Vibr* 27(3):276–284

Sampling Methods to Balance Classes in Dermoscopic Skin Lesion Images



Quynh T. Nguyen , Tanja Jancic-Turner, Avneet Kaur, Raouf N. G. Naguib , and Harsa Amylia Mat Sakim

Abstract Convolutional neural networks are used to classify dermoscopic skin lesion images. The high accuracy of deep learning models is well documented; however, those models do not perform very well on testing (unseen data) sets due to imbalanced classes of images. To tackle this problem, over-sampling and under-sampling methods are explored in this study. Part 1 of the study focuses on the details of these sampling techniques, while Part 2 highlights the architecture of the deep learning model and its performance when using both sampling approaches. The results of Part 1 show that through the use of unsupervised learning techniques, namely, Hierarchical Clustering, Self-Organizing Maps, and K-Means, similar images are clustered, based on the skin lesions' shape and color. Using augmentation for oversampling, 32,731 images are included for the training task in total. For undersampling, unsupervised learning techniques suggested 3 or 4 sub-groups of melanocytic nevi. Going through those clusters, the image background color also affects the way unsupervised learning techniques group similar images together.

Q. T. Nguyen (✉) · T. Jancic-Turner · A. Kaur
Mathematics and Statistics Department, Langara College, Vancouver, Canada
e-mail: qnguyen@langara.ca

T. Jancic-Turner
e-mail: tjancicturner@langara.ca

A. Kaur
e-mail: a196@mylangara.ca

Q. T. Nguyen
Department of Computer Science, Dai Nam University, Hanoi, Vietnam

R. N. G. Naguib
School of Mathematics, Computer Science and Engineering, Liverpool Hope University,
Liverpool, UK
e-mail: naguibr@hope.ac.uk; r.naguib@ieee.org

H. A. M. Sakim
University Sains Malaysia, Nibong Tebal, Pulau Pinang, Malaysia

Keywords Class imbalance · Overfitting · Oversampling · Augmentation · Undersampling · Unsupervised learning

1 Introduction

Skin cancer is a key health concern with over 10,000 newly reported cases every month around the world [1]. Melanoma, the most severe form of skin cancer, represents under 5% of all skin cancer cases but accounts for 70% of skin cancer mortality [2]. Given the cost and discomfort associated with traditional clinical screenings, significant research is focused on the automation of skin lesion classification.

One of the main labelled dermatoscopic image databases used in research for building Convolutional Neural Network (CNN) models is provided by the International Skin Imaging Collaboration (ISIC) [3]. The ISIC-2018 training dataset includes 10,015 images, with 7 classes of skin lesions (2 malignant and 5 benign). However, it is heavily imbalanced, with the largest class, melanocytic nevi (*nv*), having almost 60 times more images than the smallest class, dermatofibroma (*df*).

Ensemble Deep Learning models (EDL) of impressive performance have been built that can even surpass experienced human experts under controlled conditions, however, they underperform on unknown data, a phenomenon known as overfitting. Reasons for overfitting are (1) imbalanced data; (2) biased training data with images taken primarily from light skinned individuals; (3) images of different quality and (4) interclass overlap and intraclass diversity.

Out of those issues, imbalanced data can be tackled with reasonable resources with sampling. In this study, we propose to conduct both oversampling and undersampling to create balanced classes of skin lesions. The paper is structured in 4 sections. Section 2 provides a literature review; Sect. 3 briefly discusses the methodology; while Sect. 4 presents the rationale and results of the three clustering techniques and main findings of this first part of the study.

2 Related Work

As observed in the survey of published work in [4], skin cancer classification is mostly treated as a classical classification problem, without addressing the models' clinical constraints. Our primary interest is in how to deal with overfitting and, specifically, with imbalanced datasets which we believe is the main contributing factor to overfitting.

It is commonly agreed that imbalanced datasets adversely impact the performance of the classifiers as the learned model is biased toward the majority class to minimize the overall error rate [5]. The available methods to deal with imbalanced datasets can be classified into algorithmic (internal) or data level (external) [5]. The external interventions involve either replicating data points of the minority classes (oversampling)

or forming small samples of the majority classes (undersampling). The advantage of using external approaches is that they are easily adaptable and independent of feature selection and classification algorithms [5]. There is a controversial school of thought related to sampling techniques. For example, in [6], the authors found that over-sampling resulted in better performance than regularization and dropout in tackling overfitting, while in [7], it was concluded that undersampling performed significantly better. Since no ground truth was provided for sub-groups of nv by ISIC, it would be important to explore whether clustering techniques are able to form any meaningful group. Therefore, this study endeavored to use 3 main clustering algorithms to form different sub-groups of nv .

Data augmentation usually involves applying image transformation functions such as translation, rotation, or flipping. Decisions on augmentation strategies still require human expertise to avoid introducing bias in the training data [8]. While these balancing methods are powerful, their utility might be limited to skin lesion datasets with substantial imbalance. Repeated sampling from the minority class can increase the risk of overfitting.

One of the issues noted in this dataset is the high occurrence of intraclass variability in the largest nevi class, as well as intraclass overlap between nevi and some of the smaller classes. Our rationale for attempting various clustering techniques to select the nv sample in an under-sampling scenario was that unsupervised learning can detect certain salient patterns in the data, and group the images based on their similarities. This would result in a more generic “true” class of nv . In 2009, a study experimented with various clustering approaches to undersample and improve class distributions [9]. Through using MRI neuroimage data, it was demonstrated that undersampling using k-medoids (unsupervised learning) for binary classification can be an effective way to narrow the gap between sensitivity and specificity [5].

3 Methodology

3.1 Preprocessing Framework

Several issues related to the images were identified, including hair, vignettes, pen marks and air bubbles. For our testing, we split that dataset (70:30) after undertaking the pre-processing steps.

DullRazor package in Python [10] was applied to images manually selected for hair removal. Images were resized to the same standard, which was 128×128 pixels. Then they were centre-cropped, if required, to remove vignettes, ink, and other artifacts.

3.2 Balancing Classes Through Over- and Undersampling

Oversampling Approach: Augmentation

The purpose of augmenting images is to deal with the problem of skewed classes, overfitting, and training image scarcity. As can be seen from Table 1, the *nv* class dominates with approximately 67% of images in the dataset. Hence, augmentation aims to match the size of that largest class. Scaling, rotation, shift, horizontal/vertical flip, brightness, and contrast were applied in sequential order to all images resulting from the previous augmentation of a class (Fig. 1).

Undersampling Approach

A second approach applied to deal with unbalanced data was to undersample the images of only the *nv* class. Instead of taking random samples, unsupervised machine learning techniques were employed to generate clusters of closely-related images. Agglomerative Hierarchical Clustering, K-Means and SOM (Self-Organizing Maps), which are more advanced clustering techniques based on Neural Networks and recommended for image datasets, were applied. Each cluster was then separately fed into the ensemble deep learning (EDL) models in lieu of the *nevi* class, and the

Table 1 Number of images per class before and after augmentations (SSR: Shift Scale Rotate; HF: Horizontal Flip, VF: Vertical Flip, BC: Bright Contrast, Sc: Scaling, R: Rotate, RSS: Rotate with Sheer Shifting)

Images	<i>nv</i>	<i>mel</i>	<i>bkl</i>	<i>bcc</i>	<i>akic</i>	<i>vasc</i>	<i>df</i>	Total
Total images	6705	1113	1099	514	327	142	115	10,015
Train Before	4693	779	769	359	228	99	80	7007
Train After	4693	4693	4693	4693	4693	4641	4625	32,731
Validation	2012	334	330	153	99	43	35	3006
Type of Augmentation	N/A	VF, BC, RSS	SSR, VF, RSS	SSR, HF, VF, RSS	SSR, HF, VF, BC, Sc	SSR, HF, VF, BC, R, RSS	SSR, HF, VF, R, RRS	

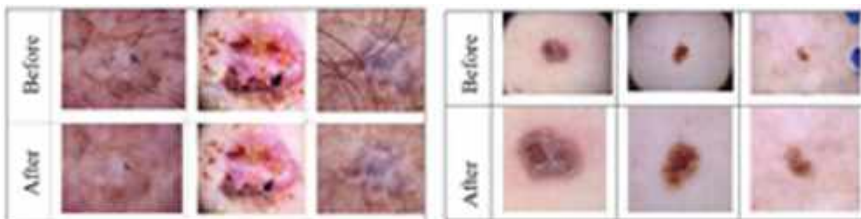


Fig. 1 Results of hair removal and center-cropping respectively (*nv*—melanocytic nevi; *mel*—melanoma; *bkl*—benign keratosis; *bcc*—basal cell carcinoma; *akic*—actinic keratosis; *vasc*—vascular; *df*—dermatofibroma)

performance of the model was assessed for accuracy, as well as the gap between training and testing accuracy, as a measure of overfitting.

Agglomerative Hierarchical Clustering

Principal Component Analysis was applied to reduce the number of dimensions and the Euclidean distance metric was used. Using Silhouette Score, the average similarity of the samples within a cluster and their distance to other objects in the other clusters was measured. Cluster forming using Ward Linkage calculation was performed.

K-Means

The K-Means clustering technique behaves differently than the Hierarchical one such that it looks for centroids of K clusters and evaluates the distance of each point to the centroids [11]. The distance method used for K-Means is also Euclidean.

Self-organizing Maps (SOM)

SOM uses processing units (neurons) to place centroids on an adjustable map [11]. Processing units maintain proximity relationships as they grow. However, this method has the advantage of providing data visualization, which helps to understand high dimensional data by reducing the dimensions of the data to a map. The parameters for this method are the number of dimensions and the radius of neighboring attraction.

4 Results and Conclusion

The ISIC-2018 dataset contains various artifacts, which were preprocessed prior to feeding images into the EDL models. Hair removal, eliminating vignettes through center-cropping, and reducing the size of the original images to 128 by 128 pixels were handled. Since the classes of distinct types of skin lesions are not balanced, augmentation techniques were applied on all classes, except nv for oversampling. Unsupervised learning algorithms were used to create sub-groups of the nv class for undersampling.

For oversampling, since the training dataset size of the nv class is 4693 (70% of 6705), the target training size of other classes should be in a similar range. The objective was to create naturally similar additional images of each class, without duplicating them. The training set was increased to 32,731 images after augmentations were completed (Table 1).

Three main clustering methods were utilized: Hierarchical Clustering, SOM, and K-Means. The main criterion prior to utilizing those clustering methods is to create balanced nv sub-groups.

The Agglomerative Hierarchical method resulted in 3 sub-classes of nv with a reasonable size of 2266 images for cluster 1, 692 for cluster 2, and 3747 for cluster 3. The K-means method also suggested 3 sub-classes of nv: cluster 1 with 3845 images, cluster 2 with 867, and cluster 3 with 1993. The results of SOM proposed 4

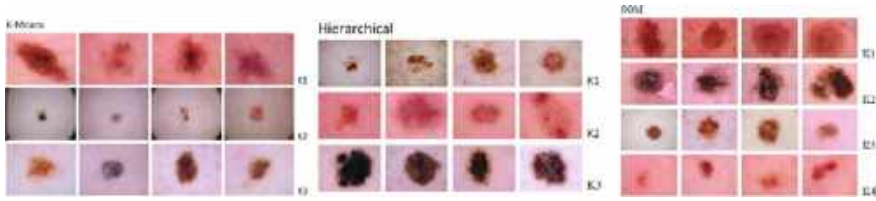


Fig. 2 Sample of images for each cluster from each clustering method

sub-classes of *nv* with 1778 images for cluster 1, 2674 for cluster 2501 for cluster 3, and 1752 for cluster 4.

Note that cluster 1 from the Agglomerative Hierarchical method does not have the same label as cluster 1 yielded from the SOM or K-Means method. While the unsupervised learning methods suggested 3 or 4 clusters, clinically, there are multiple classifications of *nevi* with 7 or 10 different classes [12]. We are uncertain that all those subclasses are represented in the ISIC-2018 dataset, therefore we elected to rely on a programmatically derived optimal number of clusters (Fig. 2).

In conclusion, unsupervised learning techniques demonstrated the ability to separate the *nv* images into several meaningful subclasses based on different features of images. Visually, we note that, across all 3 unsupervised learning methods, the *nv* images were grouped based on the color of the lesion itself, as well as the background color. This demonstrates that skin color, as well as background color (photographic technique used), have a marked impact on the unsupervised learning algorithms. Hence, the main contribution of this study is the use of unsupervised learning techniques to improve the learning tasks of CNN models.


References

1. Harangi B (2018) Skin lesion classification with ensembles of deep convolutional neural networks. *J Biomed Inform* 86:25–32
2. Skin Cancer (Including Melanoma)—Patient Version, National Institute of Health page; <https://www.cancer.gov/types/skin>. Last accessed 25 Feb 2023
3. ISIC Challenge Datasets, ISIC Challenge page, <https://challenge.isic-archive.com/data/#2018>. Last accessed 25 Feb 2023
4. Wu Y, Chen B, Zeng A, Pan D, Wang R, Zhao S (2022) Skin cancer classification with deep learning: a systematic review. *Front Oncol* 12
5. Dubey R, Zhou J, Wang Y, Thompson PM, Ye J (2014) Alzheimer’s disease neuroimaging initiative. Analysis of sampling techniques for imbalanced data: an $n = 648$ ADNI study. *NeuroImage* 87:220–241
6. Kim HC, Kang MJ (2020) A comparison of methods to reduce overfitting in neural networks. *Int J Smart Converg* 9(2):173–178
7. Jeong DH, Kim SE, Choi WH, Ahn SHA (2022) Comparative study on the influence of under-sampling and oversampling techniques for the classification of physical activities using an imbalanced accelerometer dataset. *Healthcare* 10(7):1255
8. Yang Z, Sinnott RO, Bailey J, Ke QA (2022) Survey of automated data augmentation algorithms for deep learning-based image classification tasks. [arXiv:2206.06544](https://arxiv.org/abs/2206.06544)

9. Yen S, Lee Y (2006) Cluster-based sampling approaches to imbalanced data distributions. expert systems with applications. In: Proceedings of international data warehousing and knowledge discovery conference, Krakow, Poland, vol 8, pp 427–436
10. Lee T, Ng V, Gallagher R, Coldman A, McLean D (1997) Dullrazor®: a software approach to hair removal from images. *Comput Biol Med* 27(6):533–543
11. Riveros NAM, Espitia BAC, Pico LEA (2019) Comparison between K-means and self-organizing maps algorithms used for diagnosis spinal column patients. *Inform Med Unlocked* 16:100206
12. Oakley A. Melanocytic Naevus. <https://dermnetnz.org/topics/melanocytic-naevus>. Last accessed 26 Feb 2023

Unsupervised Clustering to Reduce Overfitting Issues in Ensemble Deep Learning Models for Skin Lesion Classifications



Avneet Kaur, Tanja Jancic-Turner, Quynh T. Nguyen , Satyam Vatts, and Harsa Amylia Mat Sakim

Abstract Class imbalance in skin lesion cancer is pronounced, which in turn impacts the performance of deep learning models for classification tasks. In the case of using dermoscopic images, training a convolutional neural network-based ensemble deep learning model on 7 classes from the International Skin Imaging Collaboration (ISIC) dataset 2018 with the 2 minor classes as malignant yields decent performance in classifying skin lesions. However, the predictive ability of the model on unseen data, ISIC 2019 is unsatisfactory. To narrow the gap of overall accuracy between the training and testing tasks, Part 1 of this study conducts 2 different sampling techniques: augmentation and clustering algorithms. The results of this study—Part 2, show that 4 samples taken from Self-Organizing Maps, Hierarchical and K-Means techniques have narrower gaps of the overall accuracies between the training and testing sets. The best-balanced results of the overall accuracies for the training and testing sets are 82 and 43%, respectively, if one cluster of melanocytic nevi from hierarchical clustering is used, while augmentation gives 92 and 40%, respectively.

Keywords Deep convolutional neural networks · Ensemble learning models · Undersampling · Clustering

A. Kaur · T. Jancic-Turner · Q. T. Nguyen (✉) · S. Vatts
Mathematics and Statistics Department, Langara College, Vancouver, BC, Canada
e-mail: qnguyen@langara.ca

A. Kaur
e-mail: a196@mylangara.ca

T. Jancic-Turner
e-mail: tjancicturner@langara.ca

S. Vatts
e-mail: svatts00@mylangara.ca

Q. T. Nguyen
Department of Computer Science, Dai Nam University, Hanoi, Vietnam

H. A. M. Sakim
Univerisity Sains Malaysia, Penang, Malaysia

1 Introduction

Classification of skin lesions from digital dermoscopic images using neural network-based systems, and especially Convolutional Neural Networks (CNN) has established some significance in training and predicting melanoma. Several researchers illustrated that CNN could perform better than certified dermatologists in controlled studies [1]. Popescu et al. [2] conducted a systematic review and found that between 2015 and 2021 over 300 papers have been published investigating skin cancer detection, segmentation, and classification on various skin lesion datasets. Among these articles, the authors reviewed 134 papers from high-ranking conferences and journals which were cited frequently, and which covered recent trends in melanoma detection. The highest reported accuracy was 97.5% using a multi neural network (MNN)-based system. However, using models of such high accuracy on severely imbalanced data raises the possibility of overfitting and bias towards the majority class.

To illustrate the issue of imbalanced classes, we used the ISIC 2018 and ISIC 2019 datasets in which the malignant classes (melanoma as *mel* and basal cell carcinoma as *bcc*) are 6 times less frequent than the non-malignant ones (melanocytic nevi as *nv*, benign keratosis as *bkl*, actinic keratosis as *akic*; vascular as *vasc*, and dermatofibroma as *df*). This study employed augmentation to increase the size of minor classes, and clustering algorithms to reduce the size of one major class. The training task, along with the sampling technique, was performed on the ISIC 2018 dataset [8], and the testing task was conducted on ISIC 2019 [9]. The results of these sampling techniques are described in Part 1 [3]. Each cluster was then used in place of *nv* to find out the performance of ensemble models for further comparison. This paper is structured in 5 sections. Section 2 provides a literature review; Sect. 3 discusses the architecture of the CNN models; Sect. 4 presents the results of CNN models; and Sect. 5 describes the limitations of our study and recommendations for future work.

2 Related Work

In the early days of machine learning, both technical and domain knowledge were required for feature selection, thus slowing the development of classifiers capable of high accuracy. Deep learning neural networks (DNN) have solved the need for upfront feature selection and advanced the research of image detection and classification in clinical applications [4].

While CNN models have proven to be flexible in the process of learning, they can be sensitive to imbalanced data as a result, the prediction outputs can be different, a result known as high variance. To tackle this issue, multiple models are designed and built in the training process, instead of a single model, and subsequently the predictions of all models are combined. This process, called Ensemble Deep Learning (EDL), combines the predictions from multiple neural network models to reduce the variance of predictions and reduce generalization errors. One of the methods used

to build EDL models is called bagging or bootstrapping, which is an aggregation of models trained with subsets of data that are selected randomly from the training set to improve model variance [5].

According to [6] class imbalance in the dataset(s) can dramatically skew the performance of classifiers, introducing a prediction bias for the majority class. Nguyen et al. [3] has conducted two different sampling approaches, including oversampling and undersampling, on ISIC 2018. Augmentation of oversampling was applied on all classes except nv as the major class and unsupervised learning techniques [Hierarchical, Self-Organizing Maps (SOM), and K-Means] were applied to the nv class. The motive for using unsupervised clustering algorithms is that, in the medical field, there are sub-classes of nv and the task of clustering is to form sub-groups of similar images. As a result, the sizes of all classes become similar prior to the classification task. Besides the overall accuracy of the model, we were also interested in narrowing the gap between the training and testing sets as a measure of improvement in overfitting.

3 Ensemble Deep Learning Models

The CNN model proposed by researchers at Telkom University [7] was applied to the ISIC 2018 images. These researchers developed model for 4 different classes of images, while we tweaked the model to work for 7 classes of output. We used this model as a base reference because it was used for skin lesion classification and achieved good accuracy. An Ensemble, so called Ensemble 1, was then built from this model by creating a list of 10 deep learning algorithm of CNN Base Models (DCNN) and compiled using appropriate metrics. Each model was trained on a randomly selected subset of images with 50 epochs per training set; in terms of batch size, the model was trained on 64 batches of images for a single forward and backward pass (epoch). Those batches were run over 10 iterations. We fine-tuned the second model (Ensemble 2) by:

- Switching to Average Pooling from Max Pooling. This affected the pixel value taken from portions of the images to develop feature map. In general, Average Pooling helps in extracting overall features, such as image contrast, whereas Max Pooling is useful for edge detection.
- Adding a convolution layer of 128 filters. In CNN, an additional layer results in extraction of more features from the images. In a dataset with a large number of images with significant variations, increasing the number of convolution layers increases the accuracy by extracting more detail from the images. We did not further increase the number of layers so as to constrain the computational power and memory required for training within reasonable limits. It is current best practice to increase the number of filters in powers of 2, so we used 128 filters.

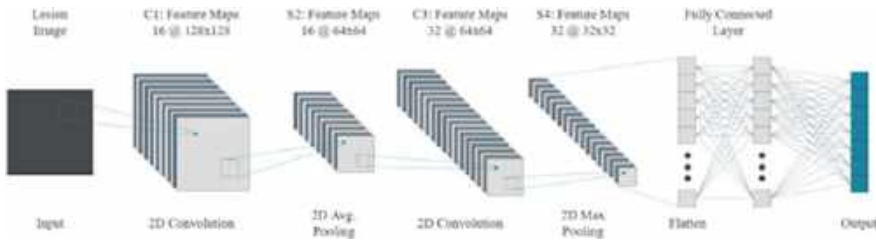


Fig. 1 Final architecture of CNN layers

- Decreasing dropout from 50 to 40% after the third CNN + Pooling Layer. The dropout is responsible for deactivating neurons in layers so that the model doesn't overfit. However, a relatively high value of 50% may result in underfitting in our case as ISIC 2018 contains a large number of images with wide variations. We found a dropout rate of 40% to be a suitable compromise between over- and underfitting.
- Decreasing the number of epochs: In the training of Ensemble 1, we observed that the model did not improve significantly after 30 epochs, but consumed resources. Thus, we reduced the number of epochs for Ensemble 2.
- The result of this tweaking was a very high accuracy on training at the expense of overfitting. Hence, in a final model, the last hyperparameter tuning step was to increase the dropout rate to 40%. Figure 1 demonstrated the final architecture of the CNN layers.

4 Results and Conclusions

4.1 Analysis

DCNN for the training task was implemented with the TensorFlow framework using 70% of the total images (both for augmentation/oversampling and unsupervised learning/undersampling). Further, a testing set was created from ISIC-2019 dataset since the main aim of the study was to determine the performance of the model on unseen data. Duplicated images were removed from ISIC-2019 before a randomly selected sample of 1300 images was created for all tests. The details of the sample size of different classes using oversampling approach were provided in [3]. During the training process of the Ensemble models, each model is trained on a different random split set of data.

The overall accuracy for the training set using the oversampling approach was 92%. The overall accuracy of the unsupervised learning techniques, including hierarchical clustering and SOM applied to both the training and testing sets, were lower than the that of the oversampling approach (Fig. 2). However, in one of the clusters derived using K-Means and another using hierarchical clustering, higher accuracies



Fig. 2 Accuracy comparisons between the training and testing sets for clusters of images under different sampling approaches

were achieved on the testing set compared to the oversampling approach. In addition, in 8 out of 10 runs with undersampled nv , the overall accuracy was still over 80%, and the difference between the training and testing sets had decreased compared to oversampling in 4 samples of different clustering techniques.

5 Conclusion

The training of DCNN was conducted on the ISIC 2018 dataset with 10,015 images and testing used a sample of 1300 images from ISIC 2019. The DCNN Ensemble model, developed using a simple base model with just 4 convolutional layers and 154,000 trainable parameters, achieved an overall accuracy of 92% on training sets when an oversampling method was applied with augmentation.

In case of model generalization, the Ensemble model did not perform well with augmentation (92% for the training set, and 40% for the testing set). Since we are interested in the performance of DCNN on unseen data, we selected ISIC 2019 as the testing set, on which we expected an overall low accuracy.

The sub-groups of nv that were clustered using unsupervised learning algorithms had noticeable resemblance to skin lesions. For example, a sub-group of nv produced by K-Means or Hierarchical clustering showed images with small brown or dark coloured lesions. Using the undersampling approach with K-Means, SOM and Hierarchical clustering techniques yielded lower overall accuracy. However, 8 out of 10 samples yielded overall accuracies of over 80% on the training sets. Moreover, in 4 out of 10 samples, the variance between the training and testing sets was smaller than that found using the augmentation approach. For Hierarchical with 3 clusters, the testing accuracy of 43% was higher than with oversampling (40%) while the training accuracy was still acceptable at 82%. Using this sample, the gap between the training and testing set is 39% while the gap for oversampling is 52%. This represents a considerable improvement in overfitting.

6 Limitations and Further Work

Although this study has demonstrated that unsupervised learning algorithms are able to separate nv class into sub-groups based on lesion features, there is a need to further investigate the number of sub classes of nv for ISIC-2018 to gain more understanding as to why sub-grouping the nv class gives a better model in terms of overfitting. We aim to assess precision and recall of the models we built to examine whether the proposed undersampling approach with clustering of nv also improves recall of melanoma class. Finally, since most of the images were taken from the light skinned individuals, it would be interesting to test our models on the images of lesions from persons of darker complexion to see how the models generalize. The clustering of nv class already demonstrated that skin and lesion colors play a role in the separation of images into clusters.

Acknowledgements We sincerely appreciate the time that Dr. Nathan Jones spent on reviewing this manuscript and the insightful feedback that Prof. Raouf Naguib provided.

References

1. Maron R, Weichenthal M, Utikal JS, Hekler A, Berking C, Hauschild A, Enk AH, Haferkamp S, Klode J, Schdendorf D, Jansen P, Holland-Lets T, Schilling B, Kalle CV, Fröhiling S, Gaiser MR, Hartmann D, Gesierich A, Kähler KC, Wehkamp U, Thiem A (2019) Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. *Eur J Cancer* 119:57–65
2. Popescu D, El-Khatib M, El-Khatib H, Ichim L (2022) New trends in melanoma detection using neural networks: a systematic review. *Sensors* 22(2):496
3. Nguyen TQ, Jancic-Turner T, Kaur A, Naguib RNG, Sakim HAM (2022) Sampling methods to balance classes in dermoscopic skin lesion images
4. Harangi B (2018) Skin lesion classification with ensembles of deep convolutional neural networks. *J Biomed Inform* 86:25–32
5. Subasi A (2020) Machine learning techniques, practical machine learning for data analysis using Python. Academic Press, New York, pp 91–202
6. Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N (2018) A survey on addressing high-class imbalance in big data. *J Big Data* 5(1):42
7. Fu'adah YN, Pratiwi NKC, Pramudito MA, Ibrahim N (2020) Convolutional neural network (CNN) for automatic skin cancer classification system. *IOP Confer Ser* 982(1):012005
8. ISIC Challenge Datasets (2018) ISIC challenge page. <https://challenge.isic-archive.com/data/#2018>. Accessed 25 Feb 2023
9. ISIC Challenge Datasets (2019) ISIC Challenge page. <https://challenge.isic-archive.com/data/#2019>. Accessed 25 Feb 2023

Drone Detection Using Swerling-I Model with L-Band/X-Band Radar in Free Space and Raining Scenario



Salman Liaquat[✉], Nor Muzlifah Mahyuddin[✉], and Ijaz Haider Naqvi[✉]

Abstract Unmanned Aerial Vehicles, also known as drones, are being utilized increasingly in numerous fields due to their multiple applications. To ensure safe operations, these drones must be detected successfully in free space and raining scenarios. However, the current radars that identify larger targets, such as aeroplanes, may not be helpful as these drones are relatively small. To successfully identify a drone, the radar designer must carefully design the system using the attributes of the target to be picked up successfully by the radar. The characteristics of an X-band radar would differ significantly from those of an L-band radar. We examine the Swerling models and their application to drones to simulate the drones' radar cross-section fluctuation using the Swerling-I model. The radar range equation uses the signal-to-noise ratio calculated from the Receiver Operating Characteristic curve to detect a drone using L-band and X-band radars in free space and rain scenarios. The analysis reveals that X-band radar experiences greater attenuation in rainy conditions than L-band radar, even though it possesses superior resolution capabilities at the same transmitted power. Nevertheless, the selection between these two radar types hinges on the particular detection scenario and prevailing environmental conditions.

Keywords Drones · Radar cross-section · Swerling model

S. Liaquat · N. M. Mahyuddin (✉)
School of Electrical and Electronic Engineering, Universiti Sains Malaysia, 14300
Penang, Malaysia
e-mail: eeammuzlifah@usm.my

S. Liaquat
e-mail: salman.liaquat@student.usm.my

I. H. Naqvi
School of Electrical Engineering, Lahore University of Management Sciences,
54792 Lahore, Pakistan
e-mail: ijaznaqvi@lums.edu.pk

1 Introduction

Drones, also known as unmanned aerial vehicles (UAVs), usage has increased in recent years due to several factors, including technological advancements, the availability of low-cost drones, and the development of new drone applications. Advancements in drone technology have made drones more reliable, easier to control, and more capable of carrying out various tasks. For example, drones can now be equipped with high-resolution cameras, sensors, and other specialized equipment to perform aerial mapping, surveying, inspection, and monitoring of infrastructure, agriculture, and wildlife [1]. The availability of low-cost drones has also contributed to their increased use. As the cost of drones has decreased, they have become more accessible to individuals and organizations that may not have been able to afford them. This availability has led to an increase in hobbyist and commercial drone use. The development of new drone applications has also contributed to their increased use. Drones are now used in various industries, including agriculture, construction, film and entertainment, transportation, and public safety. In some cases, drones perform complex or dangerous tasks for humans, such as inspecting power lines or searching for missing persons. As drones become more advanced and accessible, we can expect their use to continue to grow and evolve in the years to come [2]. Detection of drones is therefore important to ensure the safety and surveillance of these flying objects. They can be detected in numerous ways, including with radars [3], radio frequency (RF) detection [4], and infrared (IR) detection [5].

2 Drone Detection Using Radar

Radars have been used for a long time to perform surveillance and tracking in military and civilian applications. A monostatic radar (in which the transmitter and receiver are located in the same location) can be used to detect a drone. The radar transmits an RF signal toward the airspace where the drone is. When the RF signal encounters the drone, some of the signal is reflected back to the radar receiver, which can then be analyzed after suitable processing to determine the presence and location of the drone. However, some challenges are associated with using monostatic radar to detect drones. For example, drones are generally smaller and have a lower radar cross-section (RCS) than other aircraft, making them more difficult to detect. Additionally, drones flying at low altitudes and slow speeds may not be easily detected by radar systems designed to track larger, faster-moving objects. The radar system works on the radar range equation, which contains parameters that can be modified to suit the specific requirements to detect a target. It is given by [6],

$$P_r = \frac{P_t G_t^2 \lambda^2 \sigma}{(4\pi)^3 R^4}, \quad (1)$$

Table 1 Swerling Model and corresponding distribution [7]

Swerling type	RCS characteristics	Probability density function
Swerling-0	Invariant in all directions	Dirac distribution centered at the RCS value
Swerling-I	Does not vary from pulse to pulse but from scan to scan	Chi-square distribution with two degrees of freedom
Swerling-II	Varies from one pulse to another	Chi-square distribution with two degrees of freedom
Swerling-III	There exists a predominant scattering point as compared to other scatterers. RCS remains constant from pulse to pulse but varies from scan to scan	Chi-square distribution with four degrees of freedom
Swerling-IV	Varies from one pulse to another	Chi-square distribution with four degrees of freedom

where P_t is the transmitted power, G_t is the gain of the transmitting antenna, λ represents the wavelength of the signal, σ is the RCS of the target, and R is the distance from the radar to the target. Equation 1 demonstrates that the received power is directly proportional to the RCS, σ , which is a critical parameter of a target that varies with its orientation over time and must be computed correctly to detect a particular target. The RCS of a target can be estimated using simulation software or various methods, such as the computation of reflection in an anechoic chamber. The targets have been classified as non-fluctuating and fluctuating into Swerling Models as shown in Table 1.

Many researchers have carried out the RCS computation for drones using experimental methods and classified drones as Swerling-I targets, with the RCS value of drones reported to be -17 dBsm at 2.4 GHz [8], -20.9 dBsm at 8.5 GHz [9], and -15 dBsm at 9.0 GHz [10].

Radars have Receiver Operating Characteristic (ROC) curves that provide the signal-to-noise ratio (SNR) required to ensure the detection of a target with a certain probability of detection for a selected false alarm probability. The shape of a ROC curve depends on the SNR of the received signal. If the arriving signal SNR is known, then the ROC curve illustrates how well the system performs in terms of probabilities of false alarm and detection. They are a way to evaluate the performance of a radar system in terms of its ability to detect targets and reject false alarms.

Rain can affect the performance of radar detection systems. When electromagnetic waves pass through rain, some of the energy is absorbed and scattered in different directions, causing the signal to weaken and become less accurate. This effect is called attenuation, which can reduce the range and accuracy of radar detection and is dependent upon the frequency of the radar [11]. We use two frequency bands, L-band (1–2 GHz) and X-band (8–12 GHz), to demonstrate the effect of rain on drone detection at different frequencies. We use these two frequency bands as many radars commonly use these frequencies for target detection.

3 Results and Discussion

We simulate two radars using Table 2 parameters at L-band and X-band. The transmitter, antenna, and receiver subsystems have been simulated in MATLAB, and the waveform is generated and transmitted along with receiver-side processes, including matching filtering and threshold-based detection. Simulating four drones as a Swerling-I target, the SNR is computed using the ROC curve to get a 90% probability of detection for a 0.01% probability of false alarm. The first two drones are placed 500m apart to see the effect of range resolution on the radar performance. The radar detection performance for an L-band radar and an X-band radar in the free space and rain of 10mm/hr is shown in Fig. 1.

The x-axis represents the signal return time in seconds, while the y-axis depicts the received power in dBW at the radar receiver. Figure 1 (a) through (d) present the results for the free space scenario, while Fig. 1e–h illustrate the results under rainy conditions with a precipitation rate of 10 mm/h. As seen in Fig. 1a–b, both L-band and X-band radars successfully detect all four drones. However, when maintaining constant transmitted power, the received power diminishes for the X-band radar due to its smaller wavelength compared to the L-band radar. Figure 1c–d compare the radar detection performance at a range resolution of 500 m, revealing that only one of the first two targets is identified as separate drones. Although more power is transmitted as the pulse width increases with an increase in range resolution, but the resolution capabilities of both the radars are sacrificed, with the L-band radar being more affected than the X-band radar. Under rainy conditions, Fig. 1e–g demonstrate that the L-band radar remains largely unaffected, while significant performance degradation occurs for the X-band radar, as shown in Fig. 1f–h. The impact of rainfall on the X-band radar is substantial; the radar with a 50 m resolution fails to detect the last drone in this scenario, and barely identifies the third and fourth drones with a 500 m range resolution. These findings underscore the importance of selecting radar parameters based on the specific operational context and the atmospheric conditions in which they are deployed.

Table 2 Parameters used for simulation of drone detection using radar

Parameter	Value	Parameter	Value
Frequency	L-band, X-band	Range Resolution (RR)	50 m, 500 m
Max Range	10 km	RCS (σ)	- 17 dBsm
Peak Power	30 kW	Channel	Freespace, Rainfall (10 mm/h)

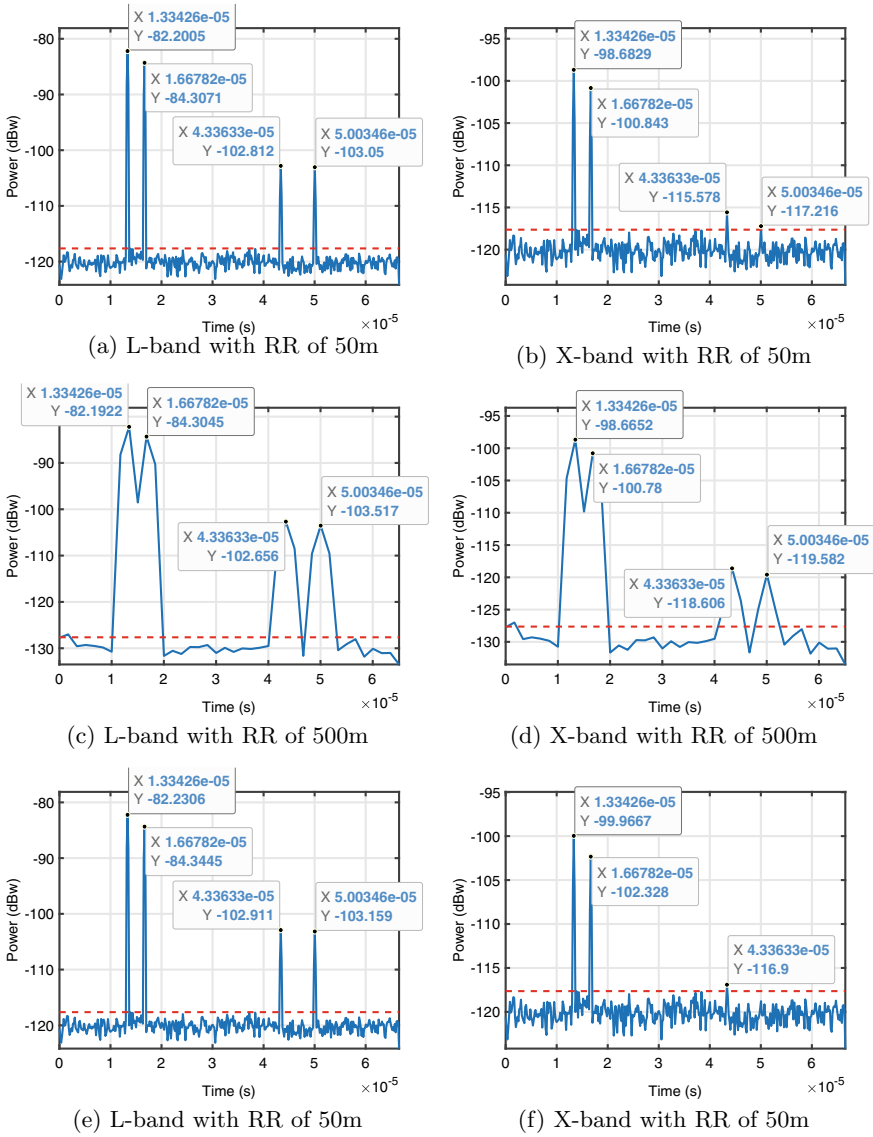


Fig. 1 Drone detection in free space (a-d) and rain (e-h) using L-band and X-band radar.

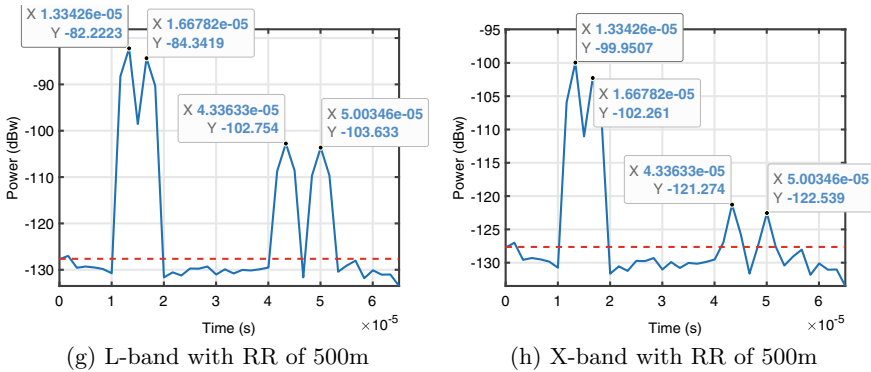


Fig. 1 (continued)

4 Conclusion

For reliable drone detection under varying atmospheric conditions, radar settings should be tailored to the specific operational context, ensuring the desired probability of detection is achieved at a given false alarm probability. The radar's ability to detect drones is contingent on the chosen frequency band, as each band influences radar performance differently. In our simulation, we compared the detection performance of L-band and X-band radars, finding that L-band radar outperforms X-band radar in rainy conditions. Due to its shorter wavelength, X-band radar offers superior resolution compared to L-band radar, enabling precise identification and tracking of small objects. However, X-band radar is more susceptible to atmospheric conditions, resulting in performance degradation during rainy situations in contrast to L-band radar.

Acknowledgements The authors would like to thank Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project Code FRGS/1/2022/TK07/USM/02/14 for permitting them to carry out this research.

References

1. Banu TP, Borlea GF, Banu C (2016) The use of drones in forestry. *J Environ Sci Eng B* 5(11):557–562
2. Michels M, von Hobe CF, Weller von Ahlefeld PJ, Musshoff O (2021) The adoption of drones in German agriculture: a structural equation model. *Precis Agric* 22(6):1728–1748
3. Coluccia A, Parisi G, Fascista A (2020) Detection and classification of multirotor drones in radar sensor networks: a review. *Sensors* 20(15):4172
4. Nguyen P, Ravindranatha M, Nguyen A, Han R, Vu T (2016) Investigating cost-effective rf-based detection of drones. In: *Proceedings of the 2nd workshop on micro aerial vehicle networks, systems, and applications for civilian use*, pp 17–22

5. Uzair M, Brinkworth RSA, Finn A (2021) A bio-inspired spatiotemporal contrast operator for small and low-heat-signature target detection in infrared imagery. *Neural Comput Appl* 33:7311–7324
6. Balanis CA (2015) *Antenna theory: analysis and design*. Wiley, London
7. Diao PS, Alves T, Poussot B, Azarian S (2022) A review of radar detection fundamentals. *IEEE Aerosp Electron Syst Mag* 1:1
8. Ritchie M, Fioranelli F, Griffiths H, Torvik B (2015) Micro-drone RCS analysis. In: 2015 IEEE radar conference. IEEE, pp 452–456
9. Guay R, Drolet G, Bray JR (2017) Measurement and modelling of the dynamic radar cross-section of an unmanned aerial vehicle. *IET Radar Sonar Navig* 11(7):1155–1160
10. Sedivy P, Nemeč O (2021) Drone RCS statistical behaviour. In: Proceedings of the MSG-SET-183 Specialists' meeting on "Drone Detectability: Modelling the Relevant Signature". North Atlantic Treaty Organization (NATO), Held Virtually (via WebEx), pp 1–4
11. Richards MA (2014) *Fundamentals of radar signal processing*. McGraw-Hill Education, New York

Enhancing Generalized Electrocardiogram Biometrics Transformer



Kai Jye Chee  and Dzati Athiar Ramli 

Abstract Using electrocardiogram (ECG) as biometrics has been explored over the years because it fits well in health monitoring applications. Most of the ECG biometrics models are specialized and can only work with very specific conditions otherwise fine-tuning, retraining or redesigning are required. Generalized ECG biometrics models are more suitable for real world applications but require large and diverse datasets to train. In this study, we introduced three databases into training the generalized ECG biometrics transformer to enhance its generalization capability. The model scored 2.93, 0.86, 2.47 and 0.29% in equal error rate for authentication task and 97.15, 99.87, 97.46 and 100.00% in identification accuracies on AFDB, NSRDB, STDB and CEBSDB respectively.

Keywords ECG · Biometrics · Transformer · BERT · Attention mechanism

1 Introduction

Proving an identity for authentication and identification is an important task to protect restricted information and to retrieve relevant information [1]. Authentication is the process of verifying whether an individual is who the individual claims to be. The goal of authentication is to ensure that only authorized individuals are granted access to protected resources or information [2]. Identification is the process of distinguishing an individual from another and then establishing the individual's unique identity. Identification is used in surveillance applications [1].

Non-biometric based methods are loosely tied to the physical identity as a result they could be lost or stolen [3]. Biometrics, by definition, is directly associated with

K. J. Chee · D. A. Ramli (✉)

School of Electrical & Electronic Engineering, USM Engineering Campus Kejuruteraan, Universiti Sains Malaysia, 14300 Nibong Tebal, Pinang, Malaysia
e-mail: dzati@usm.my

K. J. Chee

e-mail: kai_jye@student.usm.my

its identity [4]. Electrocardiogram (ECG) is an electrical signal generated by the heart which has sufficient inter-person variability to be used as biometrics [5]. Being an internal biological signal, ECG is more difficult to spoof than external biometrics like fingerprint and face [6].

ECG is already used in real-time health monitoring devices [7]. It is natural to use ECG biometrics as additional security to protect patient health records. Since ECG monitoring is continuous, it can be used for continuous authentication, which password or fingerprint cannot [8].

2 Related Works

2.1 Specialized ECG Biometrics

Most designs in the ECG biometrics literature are specialized which means a model only works on a specific set of individuals, so registering new users requires retraining or redesigning the whole model.

Jaya Prakash et al. [6] uses deep learning technique for ECG biometrics authentication by analyzing the ECG with a convolutional neural network (CNN) and a long short-term memory (LSTM) network in parallel. They trained and tested their model with ECGIDDB, PTBDB and CYBHi databases and achieved authentication accuracies of 99.94, 99.62 and 94.86% respectively.

Abd El-Rahiem and Hammad [9] transform 1D ECG signals into 2D spectrogram, use pretrained image models, namely VGG-16, VGG-19, Alex-Net, ResNet and GoogleNet, to extract features, then uses support vector machine classifier for authentication. The model scored 99.6% authentication accuracy in PTBDB.

Sun et al. [10] use blind segmentation on ECG signals, then extract features in time, frequency and energy domain using various transformations. A multiclass output CNN is used as classifier for identification. They trained and tested their model on different ECG recording sessions and achieved identification accuracies of 56.93 and 85.94% on PTBDB and ECGIDDB respectively.

Zhang et al. [11] divide ECG signals using blind segmentation. The results from wavelet transforming the ECG are used as inputs to a multiclass output CNN classifier. They achieved identification accuracies of 99, 95.1, 90.3, 93.9% on CEBSDB, NSRDB, STDB and AFDDB respectively.

Ingale et al. [12] experimented with different combinations of filters, segmentation methods, feature extraction methods and classifiers on various ECG databases, and the results showed that there is no one-size-fits-all solution that can be applied to all ECG databases. Therefore, all specialized designs are not suitable for big scale applications like centralized biometrics used in a hospital because retraining the model on every new patient is unacceptable.

2.2 Generalized ECG Biometrics

The generalized approach of designing an ECG biometrics is to train the model with a set of individuals and then test it on a completely different set of individuals without retraining and fine-tuning. This allows the training process to be separated from the enrollment process thus requires shorter ECG to enroll. Li et al. [13] proposed the generalized ECG biometrics identification. The model consists of two CNNs that are trained with subjects from FANTASIA database. It is tested on CEBSDB, NSRDB, STDB and AFDB and achieves 95, 96.1, 95.2 and 90.9% identification accuracies respectively. One training database cannot provide enough ECG variability to generalize the model hence the low accuracies.

Chee and Ramli [14] proposed the generalized ECG biometric transformer (GEBT) that can handle both authentication and identification tasks. It is trained with 10 ECG databases and tested with another 6 ECG databases. In pre-processing, the ECG recordings are resampled to 128 Hz. They are blindly segmented into 3-s segments without any fiducial points. Fifth-order Butterworth bandpass filter is applied to the ECG segments. Standard score normalization is applied to the ECG segments.

The enrolled ECG segments, G_1, G_2, \dots, G_h , and query ECG segment, G_q , are the input to ECG biometrics transformer as shown in Fig. 1. Feature space expansion transforms 1D ECG data into 2D sequence, denoted as X . Each enrolled sequence is first paired with the query sequence, then fed into ECG pair encoder for feature extraction. ECG pair encoder is adapted from Bidirectional Encoder Representations from Transformers (BERT) [15] where it analyzes two ECG segments using the transformer’s self-attention mechanism [16] to produce a representation of the two segments. The feature vectors are analyzed by authentication classifier for authentication probabilities and by ID encoder then Softmax for identification probability distribution.

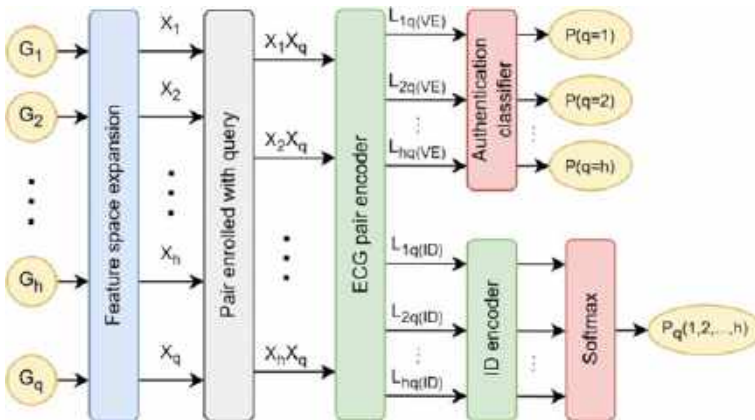


Fig. 1 Block diagram of GEBT

3 Method

This research enhances the generalization capability of GEBT by including additional ECGIDDB, PTBDB and CHYBi into the training dataset generation. Table 1 has the details of these databases, including the training-validation split ratio when generating training and validation datasets.

Table 1 Databases used in training process

Database	Name	Training-validation split	Version
APNEA-ECG	Apnea-ECG database	38:32	O
LTAfDB	Long term atrial fibrillation database	48:32	O
MITDB	MIT-BIH arrhythmia database	31:16	O
LTDB	MIT-BIH long-term ECG database	6:1	O
VFDB	MIT-BIH malignant ventricular ectopy database	14:8	O
SLPDB	MIT-BIH polysomnographic database	8:8	O
SVDB	MIT-BIH supraventricular arrhythmia database	46:32	O
INCARTDB	St. Petersburg INCART arrhythmia database	16:16	O
FANTASIA	Fantasia database	24:16	O
PTB-XL	PTB-XL, a large publicly available ECG dataset	18,853:32	O
ECGIDDB	ECG-ID database	47:32	O, P
PTBDB	PTB diagnostic ECG database	248:32	O, P
CYBHi	Check your biosignals here initiative	96:32	O, P

The “Version” column indicates whether the database is used to train the original (O) or the proposed (P) GEBT

Algorithm 1 is the single example generator used to generate a single instance of training or validation example. It runs repeatedly to obtain training and validation examples. An arbitrarily large number, 3,200,000, is chosen as the number of training examples to thoroughly exploit all ECG recordings while 32,768 validation examples, roughly 1% of the number of training examples, is used to monitor the training process. The algorithm ensures datasets will not reuse a subject in the same training example.

Table 2 Databases used to evaluate the model performance

Database	Name
AFDB	MIT-BIH atrial fibrillation database
NSRDB	MIT-BIH normal sinus rhythm database
STDB	MIT-BIH ST change database
CEBSDB	Combined measurement of ECG, breathing, and seismocardiography

Algorithm 1 Single example generator.

1	$S \leftarrow$ 32 distinct random subjects from 1 random database
2	while size of S is less than 32:
3	****db \leftarrow random database
4	**** $k \leftarrow$ random subject from db
5	**** if k is not in S :
6	*****add subject to S
7	$q \leftarrow$ random subject from S
8	$J \leftarrow$ empty set
9	for each k in S :
10	**** if k is equal to q :
11	***** $G_k, G_q \leftarrow$ 2 random ECG segments without overlapped
12	**** else :
13	***** $G_k \leftarrow$ random ECG segment
14	****add G_k to J
15	filter J and G_q
16	standardize J and G_q
17	return J, G_q, q

Table 2 lists the databases with different measuring conditions and health conditions to evaluate GEBT's generalization capability. AFDB contains unhealthy signals, STDB contains noisy signals and NSRDB and CEBSDB contains healthy signals.

4 Results

Authentication task is evaluated through equal error rate (EER) where the false positive rate equals the false negative rate. Table 3 shows that the enhanced GEBT improves in AFDB, STDB and NSRDB but remains the same for CEBSDB.

For identification task, the enhanced GEBT increases in accuracies for AFDB and STDB. It performs slightly worse for NSRDB but maintains perfect accuracy

Table 3 Comparisons of EER for authentication task

	AFDB (%)	NSRDB (%)	STDB (%)	CEBSDB (%)
Original GEBT [14]	3.44	0.87	3.00	0.29
Enhanced GEBT	2.93	0.86	2.47	0.29

Table 4 Comparisons of accuracies for identification task

	AFDB (%)	NSRDB (%)	STDB (%)	CEBSDB (%)
Li et al. [13]	90.90	96.10	95.20	95.00
Original GEBT [14]	96.20	99.91	96.09	100.00
Enhanced GEBT	97.15	99.87	97.46	100.00

for CEBSDB. It performs significantly better when comparing to Li et al. [13] generalized design as shown in Table 4.

5 Conclusion

Including more data to train the generalized ECG biometrics transformer helps the model to perform better in noisy and unhealthy conditions. However, the tradeoff is a slight performance decrease in healthy normal conditions which may be a sign of overfitting. Therefore, we suggest increasing the number of neurons in GEBT to increase its capacity to learn from even more ECG databases for future work.

Acknowledgements This paper is supported under the Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project Code: FRGS/1/2020/ICT03/USM/02/1.

References

1. Biel L, Petersson O, Philipson L, Wide P (2001) ECG analysis: a new approach in human identification. *IEEE Trans Instrum Meas* 50:808–812. <https://doi.org/10.1109/19.930458>
2. Turner D (2017) Digital authentication: the basics. <https://www.cryptomathic.com/news-events/blog/digital-authentication-the-basics>
3. Pal A, Singh YN (2018) ECG biometric recognition. In: International conference on mathematics and computing. Springer, Varanasi, pp 61–73. https://doi.org/10.1007/978-981-13-0023-3_7
4. Agrafioti F, Hatzinakos D (2008) Fusion of ECG sources for human identification. In: Proceedings of the 2008 3rd international symposium on communications, control, and signal processing, ISCCSP 2008. IEEE, Saint Julian's, Malta, pp 1542–1547. <https://doi.org/10.1109/ISCCSP.2008.4537472>
5. Arteaga-Falconi JS, Al Osman H, El Saddik A (2016) ECG authentication for mobile devices. *IEEE Trans Instrum Meas* 65:591–600. <https://doi.org/10.1109/TIM.2015.2503863>

6. Jaya Prakash A, Patro KK, Hammad M, Tadeusiewicz R, Pławiak P (2022) BAED: a secured biometric authentication system using ECG signal based on deep learning techniques. *Biocybernet Biomed Eng* 42:1081–1093. <https://doi.org/10.1016/J.BBE.2022.08.004>
7. Ullah S, Khan P, Ullah N, Saleem S, Higgins H, Sup Kwak K (2009) A review of wireless body area networks for medical applications. *Int J Commun Netw Syst Sci* 02:797–803. <https://doi.org/10.4236/ijcns.2009.28093>
8. Ekiz D, Can YS, Dardagan YC, Ersoy C (2020) Can a smartband be used for continuous implicit authentication in real life. *IEEE Access* 8:59402–59411. <https://doi.org/10.1109/ACCESS.2020.2982852>
9. Abd El-Rahiem B, Hammad M (2022) A multi-fusion IoT authentication system based on internal deep fusion of ECG signals. In: *Studies in big data*. Springer, Cham, pp 53–79. https://doi.org/10.1007/978-3-030-85428-7_4
10. Sun H, Guo Y, Chen B, Chen Y (2019) A practical cross-domain ECG biometric identification method. In: *Proceedings of the 2019 IEEE global communications conference (GLOBECOM)*. IEEE, Waikoloa, HI, pp 1–6. <https://doi.org/10.1109/GLOBECOM38437.2019.9014278>
11. Zhang Q, Zhou D, Zeng X (2017) HeartID: a multiresolution convolutional neural network for ECG-based biometric human identification in smart health applications. *IEEE Access* 5:11805–11816. <https://doi.org/10.1109/ACCESS.2017.2707460>
12. Ingale M, Cordeiro R, Thentu S, Park Y, Karimian N (2020) ECG biometric authentication: a comparative analysis. *IEEE Access* 8:117853–117866. <https://doi.org/10.1109/ACCESS.2020.3004464>
13. Li Y, Pang Y, Wang K, Li X (2020) Toward improving ECG biometric identification using cascaded convolutional neural networks. *Neurocomputing* 391:83–95. <https://doi.org/10.1016/j.neucom.2020.01.019>
14. Chee KJ, Ramli DA (2022) Electrocardiogram biometrics using transformer’s self-attention mechanism for sequence pair feature extractor and flexible enrollment scope identification. *Sensors* 22:3346. <https://doi.org/10.3390/s22093446>
15. Devlin J, Chang MW, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. [arXiv:1810.04805](https://arxiv.org/abs/1810.04805)
16. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: *Advances in neural information processing systems*. Long Beach, CA

Performance Evaluation of Different CNN Models for Motor Fault Detection Based on Thermal Imaging



Lifu Xu  and Soo Siang Teoh 

Abstract Motor faults can lead to significant operational and financial losses. In this paper, we evaluated the performance of pre-trained Convolutional Neural Network (CNN) models, including Alexnet, VGG, and ResNet, for motor fault detection using thermal imaging. We have also tested the effect of Contrast Limited Adaptive Histogram Equalization (CLAHE) on the accuracy of these models. Our approach involves first generating a new dataset using CLAHE to enhance the original dataset, followed by training AlexNet, VGG, and ResNet models using transfer learning technique. The accuracy of the models is tested using the enhanced and original datasets as inputs. The results show that using the enhanced images significantly improves the accuracy of VGG11 and VGG13 models, but it deteriorates the performance of AlexNet. For ResNet models, the improvement is minimal or even slightly decreased. The reason could be due to insufficient training or overfitting since there are limited number of images in the dataset. Overall, ResNet models achieved the best performance with the highest accuracy of 99.62%. Thus, it can be concluded that ResNet models are better suited for detecting motor fault from thermal images. Future research could focus on optimizing the CNN models or incorporating multi-sensors fusion to improve the detection performance.

Keywords Motor fault detection · Thermal imaging · CNN models · CLAHE · Pre-trained weights · ResNet · AlexNet · VGG

L. Xu · S. S. Teoh (✉)

School of Electrical and Electronic Engineering, Engineering Campus, Universiti Sains Malaysia, 14300 Nibong Tebal, Penang, Malaysia

e-mail: eteoh@usm.my

L. Xu

e-mail: xulifu2022@student.usm.my

1 Introduction

Motor faults can have serious consequences in industrial automation, making the detection of these faults a critical task. While there are existing methods for detecting motor faults, thermal imaging has emerged as an effective technique. Convolutional Neural Networks (CNNs) have been widely used in image processing tasks, including motor fault detection. In this paper, we evaluate the performance of different CNN models for motor fault detection based on thermal imaging.

The objective of this study is to compare the performance of three pre-trained CNN models, namely AlexNet [1], VGG [2], and ResNet [3], for motor fault detection based on thermal imaging. We used a public dataset obtained from [4] in the evaluation. The dataset consists of 394 thermal images of motors collected under different faults and operating conditions. To enhance the contrast of the thermal images, we apply Contrast Limited Adaptive Histogram Equalization (CLAHE) [5], which has been found to be an effective method for improving the performance of CNNs on images.

Our study aims to investigate the effect of CLAHE enhancement on the performance of pre-trained CNN models for motor fault detection. We seek to answer the following research questions: What is the impact of using different pre-trained CNN models, i.e., AlexNet, VGG, and ResNet, on the performance of motor fault detection based on thermal imaging? What is the effect of applying CLAHE enhancement on the performance of the pre-trained CNN models for motor fault detection?

To the best of our knowledge, limited studies have compared the performance of these pre-trained CNN models for motor fault detection based on thermal imaging. Moreover, the effect of CLAHE enhancement on the performance of CNN models for motor fault detection has not been extensively investigated.

In the following sections of this paper, we review the related work, explain the methodology employed, and the experimental results obtained. Finally, we discuss the implications of the findings and suggest future research directions.

2 Related Work

The use of thermal imaging for motor fault detection has gained significant attention in recent years. In the literature, several studies have proposed the use of CNN models for motor fault detection based on thermal images.

Khanjani and Ezoji [6] proposed a method for thermal imaging-based motor fault detection using AlexNet, K-means, and SVM. They evaluated the method on two datasets, achieving 100% accuracy with a two-stage classification approach. The first dataset included six motor conditions, and the second dataset included 11 situations, such as blocked rotor and different levels of Short Circuit Turns (SCTs).

In Janssens et al. [7], proposed a method to monitor the condition of a rotating machine based on infrared thermal images. They used VGG pre-trained CNN

and achieved an accuracy of 95%. Sakallı and Koyuncu [8] investigated the efficacy of various deep learning architectures, including DenseNet201, MobileNetV2, ResNet50, ShuffleNet, and Xception, for identifying 20 different faults that may occur in transformers and asynchronous motors. The authors employed an 80–20% training–testing split of the dataset as their testing method. The results showed that each model achieved 100% accuracy when tested on their dataset.

To improve the classification performance of CNN models, various image enhancement techniques have been proposed, including CLAHE. CLAHE has been shown to be able to improve the contrast of thermal images, making it easier for CNN models to perform image classification. For example, Głowacka and Rumiński [9] reported an improvement in the accuracy of face mask detection from 72.9 to 76.7% when using ResNet50 model with thermal images enhanced with CLAHE. Therefore, this study attempts to apply the CLAHE method to motor thermal images to investigate its impact on the performance of the pre-trained CNN models for fault detection.

Overall, the review shows that CNN models are effective in motor fault detection based on thermal imaging, and the potential of using CLAHE to enhance the performance of these models. Therefore, it is essential to evaluate different CNN models and image enhancement techniques to find the optimal approach for motor fault detection using thermal images.

3 Proposed Methodology

The dataset used in this study was obtained from [4], which was collected by the Electrical Machines Laboratory at the Babol Noshirvani University of Technology. The dataset consists of various performance parameters of a 1.1 kW three-phase induction motor, operating at 2800 RPM, 50 Hz, and 220/380 V. The thermal images of the motor under different operating conditions such as normal, rotor blockage, cooling fan blockage, and stator winding short-circuit fault are provided in the dataset. There is a total of 394 thermal images, which are categorized into 11 fault types. Figure 1 shows some examples of thermal images from the dataset.



Fig. 1 Examples of thermal images from the dataset. **a** Is an image with 50% short circuit in one phase of stator. **b** Is an image with 50% short circuit in 2 phases of stator. **c** Is image with blocked-rotor and **d** is image with blocked fan

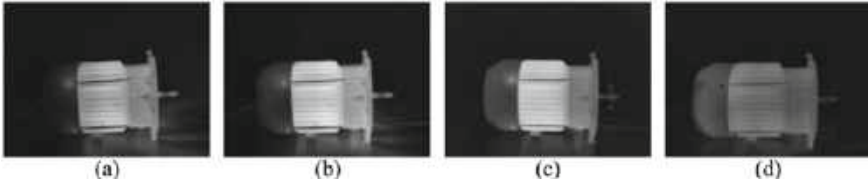


Fig. 2 Enhanced images using CLAHE are shown in **a** through **d**, corresponding to the images in Fig. 1

We utilized the CLAHE method as a pre-processing step to enhance the thermal images in the dataset. CLAHE is a widely used technique that can improve the contrast of images by redistributing the intensity values in the image. Specifically, CLAHE divides the image into small sub-regions, calculates a histogram for each sub-region, and then applies contrast enhancement to each sub-region independently. This approach can lead to more balanced contrast and better visibility of details in the image.

To evaluate the effectiveness of CLAHE, we conducted a comparative study to test the classification performance using thermal images with and without CLAHE enhancement. First, the images are converted into grayscale since CLAHE method can only be performed on a single channel image. Then CLAHE is applied to enhance the images. Figure 2 shows some examples of resulting images. It can be observed that there are some improvements in the image's contrast and detail. This is particularly obvious in images (a) and (d), where the enhanced images exhibit clearer boundaries between objects and sharper feature edges.

In the evaluation, PyTorch framework is used as the platform for implementation and experiment. We imported several pre-trained models including AlexNet, VGG11, VGG13, VGG16, ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152 in the evaluation. These models have distinct architecture designs. AlexNet introduced deeper architectures with ReLU activation and Dropout regularization, while VGG models adopt a relatively deeper structure by stacking multiple small-sized convolutional layers and pooling layers. ResNet models tackle the challenges of training deep networks by incorporating residual connections. These differences in architecture design make each model suitable for different tasks and dataset complexities.

Transfer learning [10] was utilized in the network training to improve the models' performance and accelerate convergence. This is done by importing the pre-trained weights of the models as the initial values and then retrain the models using the motor dataset. In the experiment, the models were trained on a 70–30% split of the training and testing sets using the Stochastic Gradient Descent (SGD) optimizer and the cross-entropy loss function. Each model was trained for 50 epochs using the training dataset. After training, the models were tested using the testing dataset and the accuracy was determined using Eq. 1. The average accuracy of 10 tests was calculated and used as the models' performance. All experiments were conducted on a PC with Intel Core i5-2.90 GHz CPU and an Nvidia 3060 12 GB GPU.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

where TP, TN, FP and FN are true positive, true negative, false positive and false negative respectively.

4 Experimental Result and Discussion

The experiment results are presented in Fig. 3. Overall, ResNet models show better performance compared to VGG and AlexNet. The best performing model is ResNet50 with an accuracy of 99.62%. The performance of AlexNet is the worst among the pre-trained models evaluated. On the analysis of CLAHE image enhancement, it can be observed that this method significantly improves the accuracy of VGG11 and VGG13 models, which achieves increments of 2.81 and 3.14% respectively. However, it gives an adverse effect on AlexNet with a decrease of 2.05% in accuracy. On the other hand, its impact on ResNet models is less and inconsistent. For example, in the case of Resnet101, the accuracy increases from 99.15 to 99.62% while for ResNe152, the accuracy decreases from 99.25 to 99.06%. There are three out of five ResNet models showing various degrees of deterioration in accuracy using the enhanced images.

Analysis on the relationship between the model depth and accuracy found that if the type of model is disregard, it appears that deeper models tend to exhibit better performance. However, when analyzing the individual models of ResNet, there seems to be no direct correlation between the depth of the model and its classification accuracy. The possible reasons for this are, first, the dataset itself may be relatively simple, as it consists of infrared thermographic images of motors with high similarity

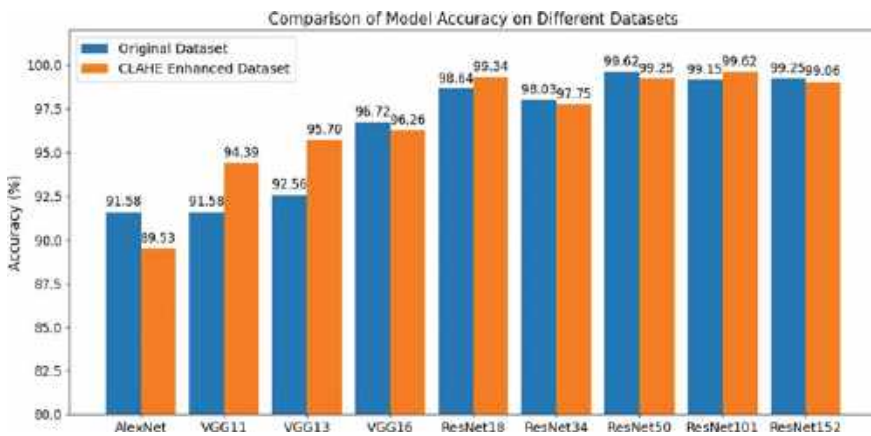


Fig. 3 Accuracy results for different models on original and CLAHE-enhanced datasets

and clear features. This makes it easier for relatively shallow models to achieve high accuracy. Second, the dataset size may be insufficient to support the increasingly complex models. Deeper models generate more parameters, which may result in overfitting and thus lowering the testing accuracy. This explains why in some cases, deeper models have even lower accuracy than that of shallower models.

5 Conclusion

This study has investigated the performance of AlexNet, VGG, and ResNet pre-trained CNN models for the detection of motor fault based on thermal imaging. The impact of applying CLAHE to enhance the input images on the models' performance have also been investigated. The results show that using enhanced images gives significant improvement for VGG11 and VGG13, but deteriorates the performance of AlexNet, while its impact on ResNet models is minimal. Overall, ResNet models show the best performance, followed by VGG and AlexNet. In conclusion, ResNet models with deeper networks and able to overcome the problem of vanishing gradients through residual connections have demonstrated excellent performance in classifying the thermal images. As a future work, ResNet models could be extended to multi-modal learning, such as incorporating the vibration signals and thermal images, to improve the accuracy and robustness of motor health monitoring.

Acknowledgements This work was financially supported by Universiti Sains Malaysia Research University Grant (RUI 1001/PELECT/8014053).

References

1. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. *Commun ACM* 60(6):84–90
2. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. In: *Proceedings of the 9th international conference on learning representations*, arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)
3. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 770–778
4. Najafi M, Baleghi Y (2017) Designing an algorithm to automatically detect and classify faults in electrical equipment using thermal images. MSc thesis, Babol Noshirvani University of Technology
5. Pizer SM, Amburn EP, Austin JD, Cromartie R, Geselowitz A, Greer T, Haar Romeny BT, Zimmerman JB, Zuiderveld K (1987) Adaptive histogram equalization and its variations. *Comput Vis Graph Image Process* 39(3):355–368
6. Khanjani M, Ezoji M (2021) Electrical fault detection in three-phase induction motor using deep network-based features of thermograms. *Measurement* 173:108622
7. Janssens O, Van de Walle R, Loccupier M, Van Hoecke S (2017) Deep learning for infrared thermal image based machine health monitoring. *IEEE/ASME Trans Mechatr* 23(1):151–159

8. Sakallı G, Koyuncu H (2023) Identification of asynchronous motor and transformer situations in thermal images by utilizing transfer learning-based deep learning architectures. *Measurement* 207:112380
9. Głowacka N, Rumiński J (2021) Face with mask detection in thermal images using deep neural networks. *Sensors* 21(19):6387
10. Ribani R, Marengoni M (2019) A survey of transfer learning for convolutional neural networks. In: *Proceedings of the 2019 32nd SIBGRAPI conference on graphics, patterns and images tutorials (SIBGRAPI-T)*. Rio de Janeiro, pp 47–57

Comparative Analysis of Deep Learning-Based Abdominal Multivisceral Segmentation



Junting Zou and Mohd Rizal Arshad

Abstract The segmentation of multiple abdominal organs is essential for medical diagnosis and treatment of various abdominal conditions, such as surgical planning, image-guided interventions and diagnosis. The main challenges are the highly heterogeneous and complex anatomy, as well as the variability in size, shape and position of abdominal organs. And in recent years, deep learning techniques have been successfully applied to various medical image segmentation tasks. Therefore combining accurate and effective deep learning based segmentation methods is essential to obtain better clinical results. In this study, we present a comparison of three deep learning architectures for abdominal multi-organ segmentation, namely the Multiscale Attention Network (MA-Net), ResNet50-U-Net and U-Net++. We evaluated the performance of these three architectures on an abdominal MRI dataset consisting of different pathological and anatomical conditions. Our results show that MA-Net equipped with a multiscale attention mechanism outperforms ResNet50-U-Net and U-Net++ in terms of Dice coefficient, Jaccard index and Hausdorff distance. By effectively capturing and integrating multi-scale contextual information, MA-Net can better depict complex organ boundaries in the dataset. Therefore, the application of MA-Net or its variants to abdominal organ segmentation has the potential to significantly enhance clinical decision-making and patient care.

Keywords Abdominal organs · Deep learning · Multiscale Attention Network · U-Net

J. Zou · M. R. Arshad (✉)

School of Electrical and Electronic Engineering, USM, Engineering Campus,
Seberang Perai Selatan, Nibong Tebal, Penang 14300, Malaysia
e-mail: erizal@usm.my

J. Zou

e-mail: junting.zou@student.usm.my

1 Introduction

Accurate diagnosis and treatment of abdominal disorders rely heavily on segmentation of multiple abdominal organs, which aids in surgical planning, image-guided operations, and diagnostic procedures [1]. The complexity and heterogeneity of abdominal anatomy, combined with the variability in organ size, shape and location, present significant challenges in achieving accurate and effective segmentation [2]. Traditional methods of segmenting medical images have a limited capacity to meet these challenges, consequently, innovative approaches need to be investigated in order to improve segmentation performance. In recent years, deep learning techniques have been increasingly applied to medical image segmentation tasks [3], demonstrating superior performance in capturing complex organ boundaries and dealing with the variability of anatomical structures. Therefore, the development and implementation of effective deep learning-based segmentation methods is critical to improving clinical outcomes.

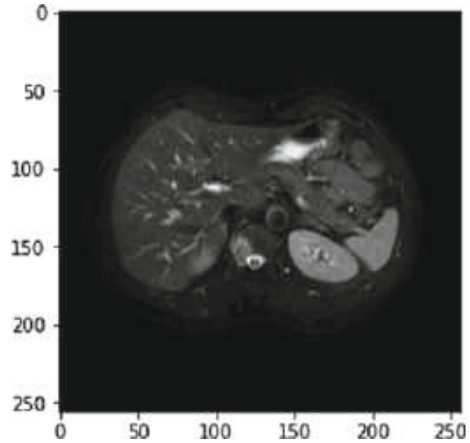
In this study, we aim to compare three advanced deep learning architectures for abdominal multi-organ segmentation, focusing on their performance and applicability in real clinical scenarios. Specifically, we investigate the Multiscale Attention Network (MA-Net) [4], ResNet50-U-Net and U-Net++ architectures [5], each one of these methods has proven successful in medical image segmentation tasks. To evaluate the performance of these architectures, we used an abdominal MRI dataset containing a variety of pathological and anatomical conditions, thus ensuring a representative evaluation. Through systematic analysis, we investigated the effectiveness of each architecture based on key segmentation metrics such as the Dice coefficient, Jaccard index and Hausdorff distance. Our findings not only help to understand the strengths and weaknesses of each architecture, but also provide recommendations for the clinical application of these approaches. By determining the optimal deep learning architecture for abdominal multi-organ segmentation, we can facilitate the development of diagnostic tools that will improve patient care and clinical decision-making.

2 Multi-organ Abdominal Segmentation Model

2.1 Dataset

We used a dataset including 120 DICOMs from two different MRI sequences [T1-DUAL in-phase (40 datasets), out-of-phase (40 datasets) and T2-SPIR (40 datasets)], each routinely scanning the abdomen using different combinations of radiofrequency pulses and gradients. There are no tumors or annotated lesions at the organ borders in this dataset (liver, kidney, spleen). These datasets were acquired from a 1.5T Philips MRI, which produces 12-bit DICOM images with a resolution of 256×256 . ISD ranges between 5.5 and 9 mm (mean 7.84 mm), X-Y spacing between 1.36 and

Fig. 1 A DICOM format image in the dataset



1.89 mm (mean 1.61 mm) and slice count between 26 and 50 (mean 36). A total of 1594 slices (532 slices per sequence) will be available for training and 1537 slices will be available for testing [6]. Figure 1 is a random image in DICOM format from the dataset.

2.2 Data Pre-processing

To normalise the input data and improve the performance of the deep learning architecture, the following steps were performed to pre-process the raw MRI images:

- a. We normalise the intensity values of the MRI images to the range $[0, 1]$ using min-max normalisation.
- b. The images were resized to 256×256 pixels using bilinear interpolation to prevent inconsistencies in the size of some images in the dataset and to ensure compatibility with the input requirements of the deep learning model.
- c. To increase the size of the training dataset and improve the model's generalization, we applied various data enhancement techniques, including random rotation, scaling and flipping.

2.3 Deep Learning Architectures

We compared three advanced deep learning architectures for multi-organ segmentation of the abdomen:

- a. Multiscale attention network (MA-Net). MA-Net is a deep learning based segmentation method that integrates multi-scale features and attention mechanisms.

It consists of an encoder-decoder architecture with a multi-scale feature fusion module and a self-attentive module. The multi-scale feature fusion module aggregates features at different scales, enhancing the model's ability to capture local and global contextual information. The self-attentive module focuses on the most relevant features and suppresses irrelevant features, thereby improving segmentation accuracy.

- b. ResNet50-U-Net. ResNet50-U-Net is a combination of the Residual Network (ResNet) and U-Net architectures; ResNet50 is a 50-layer residual network used as an encoder and the U-Net architecture used as a decoder. The residual connections in ResNet50 help alleviate the gradient disappearance problem and improve the learning capability of the network. The U-Net decoder reconstructs the segmentation map using jump connections, which combine high-resolution features from the encoder with upsampled features from the decoder.
- c. U-Net++. U-Net++ is an improved version of the U-Net architecture, which contains nested and dense skip connections. These connections help to mitigate semantic gaps between feature maps at different scales and prevent overfitting, thus improving segmentation performance. The architecture consists of an encoder-decoder structure with multiple nested decoder paths that progressively refine the segmentation results.

2.4 Model Training

All three models were implemented in Pytorch and trained using the Adam optimizer with a learning rate of 0.0001 and a batch size of 8. The loss functions used for training were a weighted sum loss of binary cross-entropy and a dice loss. The models were trained for 100 epochs with early stopping, and model weights were saved based on the best validation Dice score.

2.5 Evaluation Metrics

We used three widely used evaluation metrics to assess the performance of the three deep learning architectures: the Dice coefficient, the Jaccard index and the Hausdorff distance. The Dice coefficient and the Jaccard index were used to measure the similarity of the ground truth to the predicted segmentation, while the Hausdorff distance was used to assess the accuracy of the organ boundaries.

3 Result and Analysis

We compared the performance of the three deep learning architectures based on the Dice coefficient, Jaccard index and Hausdorff distance metrics. The means and standard deviations of these metrics are shown in Table 1.

As can be seen from Table 1, MA-Net obtained the highest Dice coefficient and Jaccard index, indicating better segmentation accuracy than the ResNet50-U-Net

Table 1 Performance comparison of deep learning architectures

Architecture	Dice coefficient	Jaccard index	Hausdorff distance
MA-Net	0.89 ± 0.04	0.81 ± 0.05	7.2 ± 2.1
ResNet50-U-Net	0.87 ± 0.05	0.78 ± 0.06	8.0 ± 2.5
U-Net++	0.88 ± 0.04	0.80 ± 0.05	7.6 ± 2.3

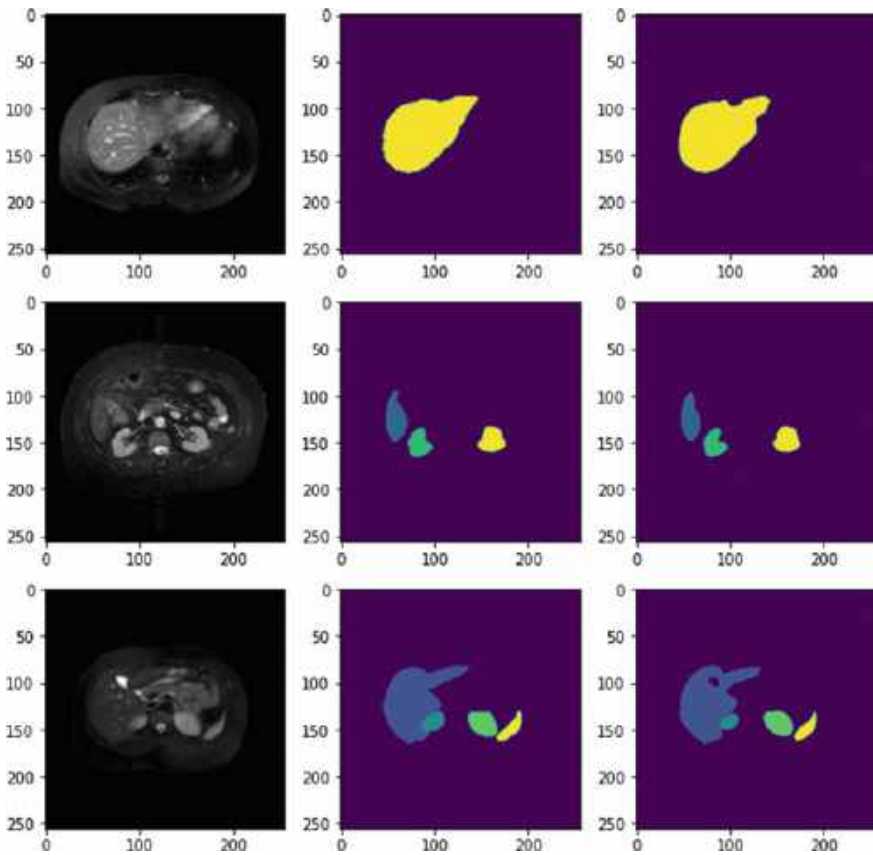


Fig. 2 MA-Net segmentation results

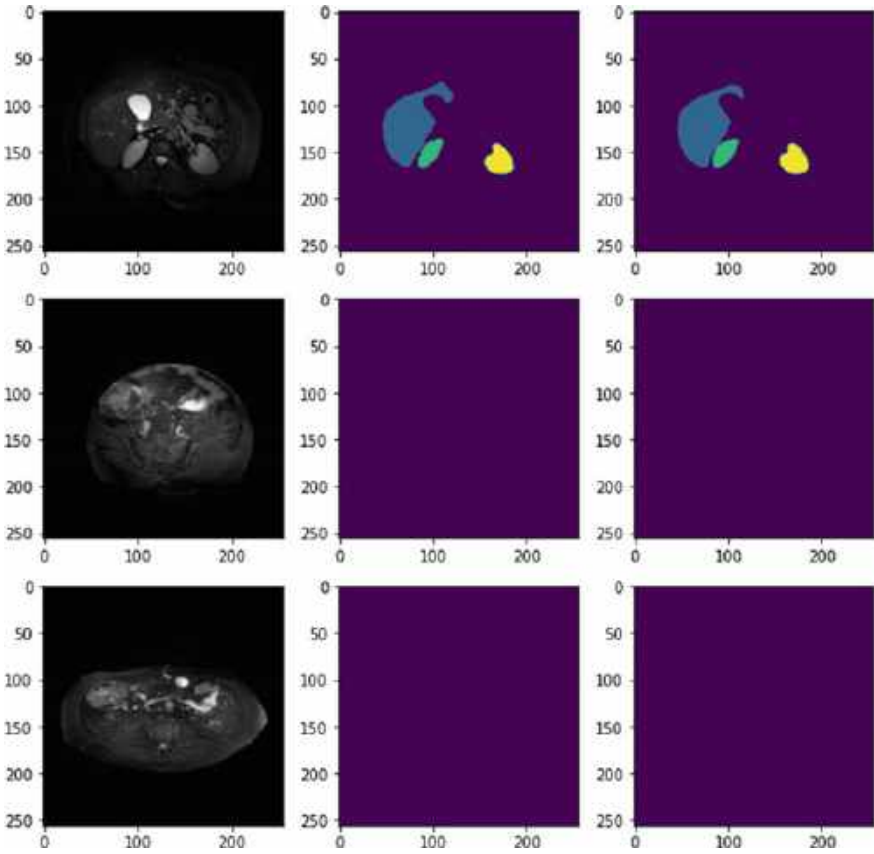


Fig. 3 ResNet50-U-Net segmentation results

and U-Net++ architectures. MA-Net also had the lowest Hausdorff distance, indicating fewer segmentation errors. This also demonstrates that the multi-scale attention mechanism can lead to better segmentation performance in terms of integrating multi-scale contextual information.

Figures 2, 3 and 4 provide a comparison of the segmentation results obtained in the test set for each of the three deep learning architectures. The left column shows the images in the test set, the middle column shows the labeled mask images, and the right column shows the segmentation results. Compared to ResNet50-U-Net and U-Net++, MA-Net produces more accurate and detailed segmentation results. In particular, MA-Net is better at handling complex organ boundaries and dealing with the inherent variability in the size, shape and location of abdominal organs.

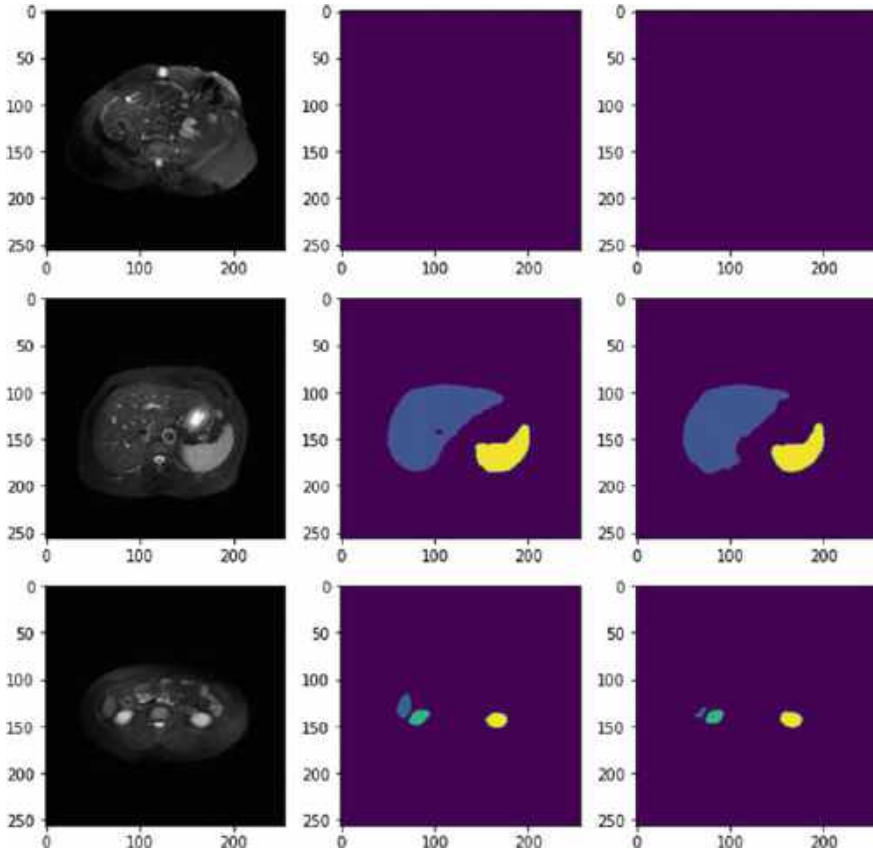


Fig. 4 U-Net++ segmentation results

4 Conclusion

In this study, we provide a comparison of three advanced deep learning architectures for abdominal multi-organ segmentation, including the Multiscale Attention Network (MA-Net), ResNet50-U-Net and U-Net++. We evaluate these architectures using abdominal MRI datasets containing a variety of pathological and anatomical conditions. Quantitative analysis showed that MA-Net consistently achieved higher Dice coefficients and Jaccard indices, as well as lower Hausdorff distances. This suggests that MA-Net is more accurate in segmenting abdominal organs and identifying their boundaries. In our qualitative evaluation, we observed that the multiscale attention mechanism embedded in MA-Net is very effective in capturing and integrating multiscale contextual information, a feature that allows the model to more accurately depict complex organ boundaries.

5 Future Works

MA-Net appears to be the most promising architecture for abdominal multi-organ segmentation tasks, it and its variants could facilitate the development of diagnostic tools that could improve patient care and clinical decision-making. Future research could combine other imaging modalities, such as CT, ultrasound, to make deep learning architectures more broadly applicable, and could also explore the integration of these architectures into clinical workflows and investigate their performance in real-life clinical situations.

References

1. Okada T et al (2012) Multi-organ segmentation in abdominal CT images. In: Annual international conference of the IEEE engineering in medicine and biology society, San Diego, CA, pp 3986–3989. <https://doi.org/10.1109/EMBC.2012.6346840>
2. Wolz R, Chu C, Misawa K, Fujiwara M, Mori K, Rueckert D (2013) Automated abdominal multi-organ segmentation with subject-specific atlas generation. *IEEE Trans Med Imaging* 32(9):1723–1730. <https://doi.org/10.1109/TMI.2013.2265805>
3. Hesamian MH, Jia W, He X et al (2019) Deep learning techniques for medical image segmentation: achievements and challenges. *J Digit Imaging* 32:582–596. <https://doi.org/10.1007/s10278-019-00227-x>
4. Fan T, Wang G, Li Y, Wang H (2020) MA-Net: a multi-scale attention network for liver and tumor segmentation. *IEEE Access* 8:179656–179665. <https://doi.org/10.1109/ACCESS.2020.3025372>
5. Siddique N, Paheding S, Elkin CP, Devabhaktuni V (2021) U-Net and its variants for medical image segmentation: a review of theory and applications. *IEEE Access* 9:82031–82057. <https://doi.org/10.1109/ACCESS.2021.3086020>
6. CHAOS—Grand Challenge (n.d.) Grand. <https://chaos.grand-challenge.org/Data/>. Accessed 13 Apr 2023

A Review of ECG Biometrics: Generalization in Deep Learning with Attention Mechanisms



Aini Hafizah Mohd Saod  and Dzati Athiar Ramli 

Abstract Common biometric systems like fingerprint and face recognition are more convenient in daily applications, but biological and behavioral characteristics of the biometrics features can be fabricated and digitally stolen. Thus, biometrics features with liveness detection such as the electrocardiogram (ECG) have been introduced as its features are hidden and difficult to forge. This study presents a review of ECG biometrics based on deep learning and generalization issues in deep learning. Based on the review, deep learning methods such as recurrent neural networks (RNN) and long short-term memory networks (LSTM) with attention mechanisms can be employed to improve the performance and generalization ability of ECG biometrics systems.

Keywords ECG biometrics · Deep learning · RNN · LSTM · Attention mechanism · Self-attention mechanism

1 Introduction

Electrocardiogram (ECG) is related to the heart's electrical activity that is recorded by placing electrodes on the human body's skin. ECG interpretation is crucial to accurately detect heart problems for further medical treatment [1]. On the other hand, ECG can be adopted in biometrics systems as an alternative to secure personal data and prevent spoofing attacks that potentially occurred in conventional static biometrics like fingerprint and face recognition systems. This paper aims to provide a comprehensive review of ECG biometrics based on deep learning and generalization

A. H. M. Saod

Centre for Electrical Engineering Studies, Universiti Teknologi MARA, Penang Branch, Bukit Mertajam Campus, 13500 Permatang Pauh, Penang, Malaysia

D. A. Ramli (✉)

School of Electrical and Electronic Engineering, USM Engineering Campus, Universiti Sains Malaysia, 14300 Nibong Tebal, Malaysia

e-mail: dzati@usm.my

issues. The paper is organized as follows; Sect. 1 presents the introduction, and the overview of ECG biometrics systems is discussed in Sect. 2. Then, Sect. 3 focuses on ECG biometrics based on deep learning. The generalization in deep learning is discussed in Sect. 4, and Sect. 5 concludes the overall review.

2 ECG Biometric Systems

Nowadays, biometrics are commercialized, covering various applications such as premise access, online banking, and border crossing by adopting fingerprint and face recognition [2]. These kinds of common biometrics have unique features of biological and behavioral traits of individuals to distinguish one person from another. Although they are easy to use and provide high accuracy, their static features are vulnerable, leading to issues of unauthorized access and personal data theft [3]. In contrast, the intrinsic and dynamic nature of ECG signals is concealed and hard to copy and forge, therefore ECG biometrics has been introduced to address the drawbacks of the static features.

Implementation of ECG biometrics typically involves enrollment and identification phases as depicted in Fig. 1. Initially, an ECG signal (known as ECG) is captured during the enrollment phase. Before the data storage, pre-processing, and feature extraction are conducted to obtain the template of ECG features. Meanwhile, in the identification phase, a new ECG signal (unknown ECG) is inputted, and the same steps, pre-processing, and feature extraction are repeated. Then, a specific classification technique is employed to identify the unknown ECG based on the templates in the database [4]. In addition, feature transformation can be performed after feature extraction to transform the extracted features into higher dimensional space (2D or 3D) [5], and feature selection can be assigned to select appropriate feature subsets for a larger number of extracted features [6].

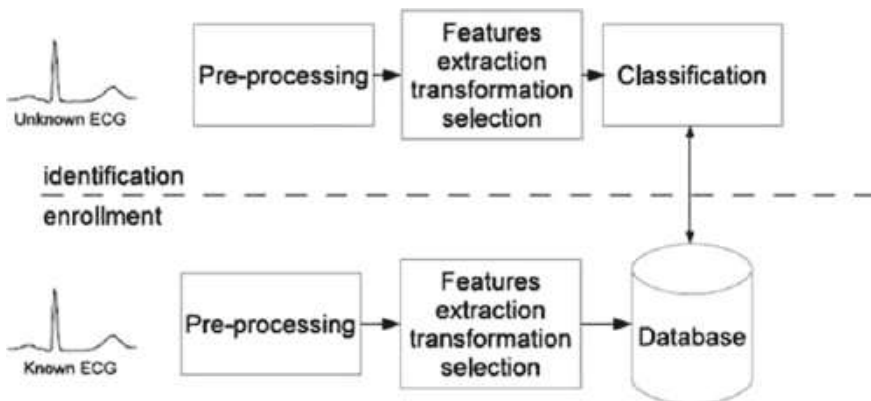


Fig. 1 Example of ECG biometrics system [4]

3 ECG Biometrics Based on Deep Learning

Deep learning is adopted in the deep neural network (DNN), which consists of an input layer, hidden layers, an output layer, and nodes that are connected between adjacent layers [7]. Based on the machine learning concept, DNN aims to automatically find appropriate representations for input data based on a given task. The “deep” in deep learning represents the depth of the network model, which is the number of successive layers or hidden layers that contribute to the representations [8]. DNNs are popular among researchers for classification and prediction purposes in various research domains as the learning approaches always produce successful results with remarkable performance.

Over the years, various DNN methods have been introduced for sequential data processing as in ECG signals. For instance, a convolutional neural network (CNN) is one of the common DNN architectures that perform convolutional operations and classification based on the learned features and can overcome the overfitting issues mostly in image processing and pattern recognition [5]. On the other hand, recurrent neural network (RNN) provides feedback loops that are suitable for operation with sequential data. However, due to the vanishing gradient issue, RNN is restricted by its short-term memory when long-term dependencies are necessary. To maintain the long-term dependencies, memory cells were added while designing RNN networks such as in long short-term memory (LSTM) and gated recurrent unit (GRU), which are the variants of RNN. LSTM and GRU use a gate structure to regulate the information flow through the sequence chain, but the difference is that GRU has one gate less than LSTM [9]. For example, Lyn et al. employed a multi-resolution bidirectional LSTM with random segmentation and auto-correlation method to perform time–frequency transformation for ECG signals. The classification results of the proposed method (> 97.3% accuracy) outperformed most RNN-based networks using selected ECG public databases [10]. Furthermore, two-hybrid network models based on LSTM and bidirectional LSTM, which are CNN-LSTM and CNN-biLSTM were proposed in [11]. Based on the experimental results, the CNN-biLSTM surpassed the CNN-LSTM with an accuracy of 94.67% since bidirectional learning was executed in two ways for fast learning results.

Besides, encoder-decoder architecture is widely used in neural machine translation, however, the performance of the encoder-decoder pair declines for long input sentences [12]. Therefore, a bidirectional RNN-based encoder and a gated RNN-based decoder were used for annotating long sequences. To reduce the burden of the encoder to encode all information, the decoder can select which parts of the source sentence to pay attention to with the implementation of an attention mechanism [13]. On the other hand, Transformer, a superior sequence transduction model, was established to generalize various tasks with large and limited training data. The model architecture of the Transformer employed stacked self-attention and pointwise, fully connected layers for both the encoder and decoder. Instead of relating the input–output sequence as in a common attention scheme, the self-attention is concentrated on a single sequence, so the sequence learns about itself [14].

4 Generalization in Deep Learning

Generalization can be related to the ability of a model architecture to respond effectively to unseen data based on prior knowledge and to attain good performance for inputs that are not the same as the training dataset [15]. Several approaches have been introduced to improve generalization in deep learning such as dropout regularization to mitigate overfitting [18], data augmentation to increase the size of the training dataset [17], and transfer learning by utilizing pre-trained models [19]. In the ECG biometrics research domain, the main goal is to accurately identify people based on their ECG signals. At the same time, the ECG biometrics systems are designed to improve the generalization ability by handling the ECG variations including extraneous noises [3]. Therefore, it is crucial to enhance or adapt models that can tackle the generalization issues in ECG biometrics based on deep learning approaches.

ECG biometrics can be generalized based on several factors such as the variability of ECG datasets and the selection of model architecture [6, 16]. The heart condition is one of the ECG variations that differ between healthy people and those with heart problems. Furthermore, body postures during data acquisition such as standing, sitting, or lying down, and emotions such as nervousness, panic, and fear can affect the shape of the heart [2], contributing to the variability of ECG signals. The placement of the ECG sensors or electrodes during data acquisition is another factor that contributes to ECG variations [3]. Thus, recent studies have been encouraged to employ large-scale ECG datasets by combining numerous databases that contain diverse and high-quality ECG data with deep learning implementation to achieve greater generalization [17].

Inspired by the biological visual attention system [20], attention mechanisms have been incorporated into deep learning architecture to generalize the employed model. Attention mechanisms are the techniques used in deep learning to make the DNN models focus on essential information while processing sequential data like text streams and time-series signals. This scheme allows the model to concentrate on the most significant parts of the input sequence and disregard the less important parts. The DNN model can be modified to include the attention mechanism that can learn to assign weights to different input features based on their significance to the output prediction. For example, an attention mechanism was proposed to combine both content and location information for the speech recognition model. An attention-based recurrent sequence generator was developed to generate an output sequence that reduced the phoneme error rate (PER) from 18.7 to 17.6% with methods of location-awareness and smooth focus [21]. Furthermore, Jyotishi and Dandapat designed a hierarchical LSTM (HLSTM) model by stacking layers of LSTM. ECG identification was generalized with an attention mechanism that provided more weight to the segments where the parts with more color saturation area have been given more attention weight compared to the parts with less color saturation area. The performance of the HLSTM model with attention mechanism was compared to other types of LSTM-based models, and the proposed model gave superior results with 98.6 and 98.7% for accuracy and F1 score, respectively [22].

On the other hand, self-attention schemes have been adapted in Transformer to concentrate on a single sequence, so that the sequence learns information about itself [14]. Specifically, self-attention mechanisms are employed to relate different parts of the same input sequence, rather than operate between two different sequences as in attention mechanisms. Since the self-attention mechanisms do not require a separate context sequence, this scheme is more computationally efficient than attention mechanisms. In [17], the Bidirectional Encoder Representations from Transformers (BERT) was used to get a dynamic representation of a pair of ECG signals. The self-attention mechanism of the Transformer was employed to draw an inter-identity relationship when performing ECG identification tasks with the outcomes of > 96.09% identification accuracy on the Physionet database. Besides, LSTM models are enhanced with statefulness and self-attention mechanisms in [23] to enable long-term inter-sequence modeling while maintaining focus on each sequence's local characteristics. The self-attention mechanisms are utilized to generalize the LSTM models by concentrating on the different parts of each sequence that can affect the classification outcome the most. Overall, based on the previous works that have been discussed, attention mechanisms can be integrated with the DNN models such as RNN and LSTM in improving generalization. This is because the attention mechanisms are computationally effective and allow the model to deal with long-term dependencies between different parts of the input sequence such as in ECG signals, where the input sequence can be very long.

5 Conclusions

ECG biometrics focuses on developing robust authentication techniques that can precisely distinguish between individuals based on their unique ECG signals. Deep learning methods such as RNN and LSTM can be implemented for processing and classification tasks to work on sequential data like ECG signals. Additionally, attention mechanisms are very helpful when incorporating deep learning models because they can enable the model to concentrate on the significant parts of the input while neglecting the less important parts. In conclusion, by employing deep learning with attention mechanisms, the generalization ability of deep learning models in ECG biometric systems can be significantly improved to achieve great performance with better generalization.

Acknowledgements This research is supported by the Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project code: FRGS/1/2020/ICT03/USM/02/1.

References

1. Hampton J, Hampton J (2019) *The ECG made easy*, 9th edn. Elsevier, Amsterdam
2. Pinto JR, Member S (2020) Evolution, current challenges, and future possibilities in ECG biometrics. *IEEE Access* 8:34746–34776
3. Ingale M, Cordeiro R, Thentu S, Park Y, Karimian N (2020) ECG biometric authentication: a comparative analysis. *IEEE Access* 8:117853–117866
4. Fratini A, Sansone M, Bifulco P, Cesarelli M (2015) Individual identification via electrocardiogram analysis. *Biomed Eng* 14(1):1–23
5. Hammad M, Zhang S, Wang K (2019) A novel two-dimensional ECG feature extraction and classification algorithm based on convolution neural network for human authentication. *Fut Gener Comput Syst* 101:180–196
6. Uwaechia AN, Ramli DA (2021) A comprehensive survey on ECG signals as new bio-metric modality for human authentication: recent advances and future challenges. *IEEE Access* 9:97760–97802
7. Mannam S (2019) Artificial intelligence, machine learning, and deep learning: are they all the same? *J Young Invest* 14:1–3
8. Chollet F (2018) *Deep learning with Python*. Manning Publications, New York
9. Yang S, Yu X (2020) LSTM and GRU neural network performance comparison study. In: *Proceedings of the international workshop electronic communications AI*. IEEE, p 98
10. Lynn HM, Kim P, Pan SB (2021) Data independent acquisition based bi-directional deep networks for biometric ECG authentication. *Appl Sci* 11(3):1125
11. Ern ESY, Ramli DA (2022) Classification of arrhythmia signals using hybrid convolutional neural network (CNN) model. *AI and ML for healthcare: image and data analytics*. Springer, Cham, pp 105–132
12. Cho K, van Merriënboer B, Bahdanau D, Bengio Y (2014) On the properties of neural machine translation: encoder–decoder approaches
13. Bahdanau D, Cho KH, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. In: *Proceedings of the 3rd international conference on learning representation*, pp 1–15. arXiv preprint [arXiv:1409.1259](https://arxiv.org/abs/1409.1259)
14. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. *Adv Neural Inf Process Syst* 30:5999–6009
15. Wang Y, Yao Q, Kwok JT, Ni LM (2020) Generalizing from a few examples: a survey on few-shot learning. *ACM Comput Surv* 53(3):1–34
16. Chen SW, Wang SL, Qi XZ, Samuri SM, Yang C (2022) Review of ECG detection and classification based on deep learning: coherent taxonomy, motivation, open challenges, and recommendations. *Biomed Sig Process Control* 74:103493
17. Chee KJ, Ramli DA (2022) Electrocardiogram biometrics using transformer’s self-attention mechanism for sequence pair feature extractor and flexible enrollment scope identification. *Sensors* 22(9):3446
18. Labati RD, Muñoz E, Piuri V, Sassi R, Scotti F (2019) Deep-ECG: convolutional neural networks for ECG biometric recognition. *Pattern Recogn Lett* 126:78–85
19. Wu SC, Wei SY, Chang CS, Swindlehurst AL, Chiu JK (2021) A scalable open-set ECG identification system based on compressed CNNs. *IEEE Trans Neural Netw Learn Syst* 14:e11107
20. Soydaner D (2022) Attention mechanism in neural networks: where it comes and where it goes. *Neural Comput Appl* 34(16):13371–13385
21. Chorowski JD, Bahdanau S, Serdyuk D, Cho K, Bengio Y (2015) Attention-based models for speech recognition. *Adv Neural Inf Process Syst* 28:577–585
22. Jyotishi D, Dandapat S (2022) An ECG biometric system using hierarchical LSTM with attention mechanism. *IEEE Sens J* 22(6):6052–6061
23. Katrompas A, Metsis V (2022) Enhancing LSTM models with self-attention and stateful training. *The 2021 intelligent systems conference*. Springer, New York, pp 217–235

Detecting Sleep Disorders from NREM Using DeepSDBPLM



Haifa Almutairi, Ghulam Mubashar Hassan, and Amitava Datta

Abstract Sleep disorders have negative effects on human health. Sleep Disorder Breathing (SDB) and Periodic Leg Movement (PLM) are common sleep disorders that happen during sleep. Early detection of SDB and PLM from Non-Rapid Eye Movement (NREM) can protect patients from hypertension and cardiovascular diseases. In this study, we propose a novel deep learning architecture DeepSDBPLM for classifying Normal, SDB and PLM from NREM using Electroencephalogram (EEG) and Electromyogram (EMG) signals. Our proposed model is tested in three different classification problems using ISRUC-Sleep database. The results show that our proposed model achieves the best result of F1 score as compared to the state-of-the-art techniques.

Keywords Sleep disorder · Convolutional neural networks · EMD · Attention

1 Introduction

Classification of sleep stages helps to identify sleep disorders. The most common sleep disorder diseases are Sleep Disorder Breathing (SDB) and Periodic Leg Movement (PLM). SDB causes stopping of the breath temporarily during the sleep. SDB is categorized into three different types: Obstructive Sleep Apnoea/Hypopnoea (OSAH), Central Sleep Apnoea/Hypopnoea (CSAH) and Mixed Sleep Apnoea/Hypopnoea (MSAH). PLM disorder happens during sleep when the patient's legs are moved periodically. The guidelines of American Academy of Sleep Medicine (AASM) categorize the sleep stages into five primary stages which are: Wake (W) stage, Non-Rapid Eye Movement (NREM) stage which includes three stages (N1, N2, N3), and Rapid Eye Movement (REM) stage. The absence of sleep stages or abnormal

H. Almutairi (✉) · G. M. Hassan · A. Datta

Department of Computer Science and Software Engineering, The University of Western Australia, Crawley, WA, Australia

e-mail: haifa.almutairi@research.uwa.edu.au

cycling between stages is associated with an increased rate of sleep disorders and related diseases.

Detection of SDB and PLM from NREM is a challenging task. Medical studies proved that when a SDB or PLM event occurs in NREM, it has an adverse effect on human health [1]. In sleep clinics, Polysomnography (PSG) is used to classify sleep stages and detect sleep disorders which records Electroencephalogram (EEG) and Electromyogram (EMG) signals. The drawbacks of this method are time-consuming and having a high rate of human errors in diagnosing.

Recently, studies proposed Deep Learning (DL) models including Convolutional Neuron Networks (CNN) to classify sleep stages [2]. Most of the studies primarily rely on the classification of sleep stages with patients who have sleep disorders. In contrast some studies proposed models to detect PLM from signals [3–7]. For detection of SDB, most of the existing works developed models to detect OSAH from ECG signals [8]. However, few studies focus on using a combination of all types of SDB segments to classify them as Normal and SDB [9–11].

Previous studies share some limitations. Firstly, they focus on developing models for the detection of PLM and SDB from signals. They do not detect SDB or PLM from a particular sleep stage. We found that measuring the total number of occurrences of SDB and PLM per hour during NREM help physicians to determine the severity of SDB and PLM. Secondly, they train their models with datasets that include segments of one type of SDB which is OSAH. To address the above mentioned limitations, this study aims to introduce DeepSDBPLM architecture to classify Normal, SDB and PLM from NREM. The proposed network was tested for three different classification problems related to sleep disorders. Classification 1 is a three-class classification of Normal, SDB and PLM; classification 2 is a binary classification of Normal and PLM; and classification 3 is a binary classification of Normal and SDB. Furthermore, we re-implemented two deep learning models from the literature [8, 12] which are used to classify signals and time series data. The purpose of this re-implementation was to compare the performance of our network with existing state of the art networks.

2 The Proposed DeepSDBPLM

Our proposed architecture DeepSDBPM is presented in Fig. 1, and its components are explained below.

2.1 Dataset and Preprocessing

We used a publicly available open dataset ISRUC-Sleep, which is collected by Sleep Medicine Centre of Hospital of Coimbra University (CHUC). According to AASM guidelines, sleep physicians segmented the recordings into 30-s segments and labelled them. Each segment is labelled as one of the five sleep stages (W, N1,

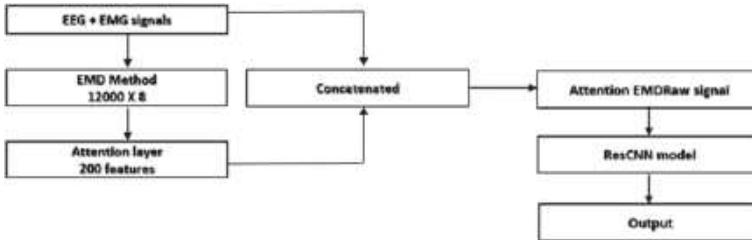


Fig. 1 The proposed DeepSDBPLM’s architecture

N2, N3, and REM) and the locations of normal, arousal, OSAH, CSAH, MSAH and PLM events are marked. We removed W and REM segments and relabelled all segments that were labelled as OSAH, CSAH and MSAH as SDB. We relabelled N1, N2 and N3 to NREM. The normal class includes normal and arousal segments, which has a high percentage in the dataset. Therefore, we reduced these segments to prevent overfitting of our deep learning architecture. The distribution of the dataset of NREM with three classes: Normal, SDB and PLM are 5639, 5639 and 1441, respectively. After selecting the data, we normalised all the segments by using Z score.

2.2 Empirical Mode Decomposition Method (EMD)

EMD is similar to Fourier Transforms and Wavelet decomposition methods. For a component to be considered an Intrinsic Mode Functions (IMF), it must satisfy two conditions as described by Karatoprak and Seker [13]:

1. Total zero crossings and the total extrema in the whole data set should be equal or vary by at most one.
2. The mean value of envelope from maxima and minima should be equal to zero at any interval of the component.

The procedure of the EMD sifting algorithm is described as follows [13]:

1. Identify all the local minima and maxima.
2. Connect all the local maxima/minima by a cubic spline to form the upper E_U / lower E_L envelope.
3. Calculate $C_1 = X[n] - M$ where the mean of these envelopes M is described as $M = \frac{(E_U + E_L)}{2}$ where E_U and E_L represent the upper and lower envelopes respectively.
4. Check if C_1 meets the two conditions to be an IMF. If not, iterate over steps 1–3 until the two conditions are met or the stopping criterion is satisfied. If these two conditions are achieved, an IMF is extracted and the same procedure is applied to the residual signal.

In this study, a total of five IMFs are produced from raw signals. We concatenated all the five IMFs to pass to an attention layer.

2.3 Attention EMDRaw Signal

The attention mechanism focuses on some of the most significant information while ignoring duplicate and unnecessary information. The mechanism of attention network gets the input EMD signal with shape 3000×5 matrix represented by the hidden state h_t and passes it to a single layer of multilayer perceptron for obtaining attention score vector $u_t = \tanh(W_s h_t + b_s)$. After that, attention weight $\alpha_t = \frac{\exp(\text{score}(u_t, h_t))}{\sum_i \exp(\text{score}(u_i, h_i))}$ is obtained by calculating the similarity between attention score vector u_t and hidden state h_t . Then, context vector $e_t = f(\sum_i \alpha_i h_i)$ is calculated by multiplying the sum of attention weight α_t and h_t , and passes to f which is a fully connected layer with \tanh activation function. The default weights W_s and bias vector b_s are randomly initialized during the training process.

We used 200 features produced by the attention network and concatenated them with the raw signals to create a new input form called *Attention EMDRaw signal*.

After that, we reshaped Attention EMDRaw signal into 3050×4 matrix to pass to ResCNN model.

2.4 ResCNN Architecture

The block diagram of ResCNN is presented in Fig. 2. We implemented three blocks of CNN with different parameters. Block 1 and 2 have three 1D-CNN layers followed by dropout layers of 0.03. Each 1D-CNN layer comprises Kernel filters and nonlinear activation function ReLU. The outputs of the first 1D-CNN layer and dropout from block 1 are added to the final output of block 1 to produce F1, which is passed as an input to block 2. The final outputs of block 2 are added to F1 from block 1 to produce F2, which is passed as an input to block 3. Block 3 has two 1D-CNN layers followed by max-pooling and dropout layers. The last two layers of block 3 are flattened and fully connected to classify the final set of extracted features.

3 Results and Discussion

Table 1 presents the performance of the selected three deep learning models with different input forms for classification 1 problem, which involved classification of Normal, SDB and PLM. From the results, it can be observed that the proposed input form Attention EMDRaw signal has the best results on all evaluation metrics

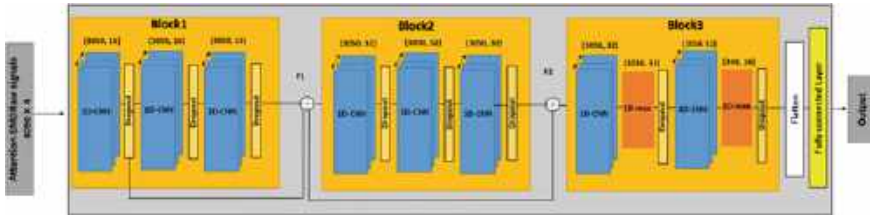


Fig. 2 The detailed architecture of the proposed ResCNN

with all the three selected deep learning models. Furthermore, our proposed model comprising of ResCNN and Attention EMDraw signal achieves the best results of sensitivity and F1 score which are 72.75 and 78.22% respectively. Whereas, Almutairi et al.'s [8] architecture comprising of CNN and LSTM with Attention EMDraw signal achieves the best results of accuracy and specificity, which are 86 and 87.88% respectively.

Table 2 presents the comparison of the results of the proposed DeepSDBPLM with state-of-the-art techniques. We compare DeepSDBPLM architecture for classification 1 problem with the two implemented models [8, 12] and Olesen et al.'s [10] proposed technique. The results show that the proposed DeepSDBPLM achieves the best F1 score of 78.22%. The accuracy obtained by Almutairi et al.'s [8] architecture (86%) is slightly higher than our proposed architecture's accuracy (85.66%). However, our proposed architecture has less complexity than Almutairi et al.'s [8] architecture. Our proposed model has 24,338 total parameters, whereas, Almutairi et al.'s [8] model has 55,194 total parameters, which makes our model computationally efficient.

For classification 2 problem of Normal and PLM segments, Table 2 shows that Umut and Centik [6] achieves the best accuracy of 91.87%, whereas our proposed model achieves the best F1 score of 86.52%. This might be due to the number of channels used in Umut and Centik [6] study, which uses a combination of EEG, EMG, EOG and SpO2 signals.

Similarly, for classification 3 problem of Normal and SDB, we compare our proposed model with studies include a combination of the three types of SDB segments in testing set. Due to the lack of literature for classification 3, we implemented Almutairi et al.'s [8] architecture and Wang et al.'s [12] architecture to classify the segments into Normal and SDB. The results in Table 2 shows that our architecture achieves an accuracy of 82.26% and an F1 score of 82.16%, which is better than other studies [8, 9, 11, 12].

Table 1 The comparison of the performance of the selected state of the art models with different input forms for the classification 1 problem with classes Normal, SDB and PLM

Type of input	Architecture	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 score (%)
Raw signal	CNN + LSTM	84.59	62.44	86.00	75.00
EMD signal	CNN + LSTM	79.78	54.31	82.00	66.48
Raw + EMD signal	CNN + LSTM	80.93	57.62	83.55	68.57
Attention RawEMD signal	CNN + LSTM	80.70	55.21	82.96	68.42
Attention EMDRaw signal	CNN + LSTM	86.00	66.10	87.88	77.89
Raw signal	ResNet	73.22	53.41	78.19	57.00
EMD signal	ResNet	75.67	53.63	78.31	61.61
Raw + EMD signal	ResNet	75.56	53.29	79.16	62.53
Attention RawEMD signal	ResNet	76.69	53.19	79.53	62.98
Attention EMDRaw signal	ResNet	82.80	58.92	84.70	71.24
Raw signal	ResCNN	84.60	69.55	86.72	76.45
EMD signal	ResCNN	77.20	56.00	80.26	64.65
Raw + EMD signal	ResCNN	82.00	66.74	84.92	72.39
Attention RawEMD signal	ResCNN	84.06	66.26	86.20	75.12
Attention EMDRaw signal	ResCNN	85.66	72.75	87.73	78.22

Bold represents the best results in the category

Table 2 Comparison of results of our proposed DeepSDBPLM with state-of-the-art techniques for all the three classification problems

Models	Details	Dataset	Classes	Accuracy (%)	F1 score (%)
Olesen et al. [10]	Raw signals + CNN and GRU	MrOS Sleep Study	Arousal, SDB and PLM	–	65.00
Wang et al. [12]	Attention EMDRaw signals + Reimplemented Resnet	ISRUC-Sleep dataset	Normal, SDB and PLM	82.80	71.24
Almutairi et al. [8]	Attention EMDRaw signals + Reimplemented CNN and LSTM	ISRUC-Sleep dataset	Normal, SDB and PLM	86.00	77.89
DeepSDBPLM	Attention EMDRaw signals + ResCNN	ISRUC-Sleep dataset	Normal, SDB and PLM	85.66	78.22
Watter et al. [3]	Rule based model	Wisconsin Sleep Cohort	Normal, PLM	47.00	63.00
Ferri et al. [4]	Rule based model	Wisconsin Sleep Cohort	Normal, PLM	62.00	72.00
Umut et al. [6]	Feature engineering + MLP	Private dataset	Normal, PLM	91.87	83.27
Carvelli et al. [7]	Raw signals + CNN, FC and LSTM	Wisconsin Sleep Cohort	Normal, PLM	81.00	85.00
DeepSDBPLM	Attention EMDRaw signals + ResCNN	ISRUC-Sleep dataset	Normal, PLM	88.20	86.52
Olsen et al. [9]	Feature engineering + GRU	SHHS, MESA	Normal, SDB	–	72.00
Sharma et al. [11]	Rawdata + CNN	SHHS	Normal, SDB	82.20	51.00
Wang et al. [12]	Attention EMDRaw signals + Reimplemented Resnet	ISRUC-Sleep dataset	Normal, SDB	76.80	77.00

(continued)

Table 2 (continued)

Models	Details	Dataset	Classes	Accuracy (%)	F1 score (%)
Almutairi et al. [8]	Attention EMDRaw signals + Reimplemented CNN and LSTM	ISRUC-Sleep dataset	Normal, SDB	81.70	81.89
DeepSDBPLM	Attention EMDRaw signals + ResCNN	ISRUC-Sleep dataset	Normal, SDB	82.26	82.16

Bold represents the best results in the category

4 Conclusion

In this paper, we proposed DeepSDBPLM model, which comprises of an introduced input form *Attention EMDRaw* signal and ResCNN architecture for classifying Normal, SDB and PLM from NREM segments. The results showed that our proposed model achieved the best F1 score for all three different classification problems. In the future, we will conduct experiments with different combined channels of EEG, EMG and EOG signals to improve the results and observe which channels play critical role.

References

1. Punjabi N, Bandeen-Roche K, Marx J, Neubauer D, Smith P, Schwartz A (2002) The association between daytime sleepiness and sleep-disordered breathing in nrem and rem sleep. *Sleep* 25(3):307–314
2. Sokolovsky M, Guerrero F, Paisarnsrisomsuk S, Ruiz C, Alvarez S (2019) Deep learning for automated feature discovery and classification of sleep stages. *IEEE/ACM Trans Comput Biol Bioinform* 17(6):1835–1845
3. Wetter TC, Dirllich G, Streit J, Trenkwalder C, Schuld A, Pollmacher T (2004) An automatic method for scoring leg movements in polygraphic sleep recordings and its validity in comparison to visual scoring. *Sleep* 27(2):324–328
4. Ferri R, Zucconi M, Manconi M, Bruni O, Miano S, Plazzi G, Ferini-Strambi L (2005) Computer-assisted detection of nocturnal leg motor activity in patients with restless legs syndrome and periodic leg movements during sleep. *Sleep* 28(8):998–1004
5. Moore H, Leary E, Lee S-Y, Carrillo O, Stubbs R, Peppard P, Young T, Widrow B, Mignot E (2014) Design and validation of a periodic leg movement detector. *PLoS ONE* 9(12):114565
6. Umut I, Centik G (2016) Detection of periodic leg movements by machine learning methods using polysomnographic parameters other than leg electromyography. *Comput Math Methods Med* 2016:1–7
7. Carvelli L, Olesen AN, Brink-Kjær A, Leary EB, Peppard P, Mignot E, Sørensen H, Jennum P (2020) Design of a deep learning model for automatic scoring of periodic and non-periodic leg movements during sleep validated against multiple human experts. *Sleep Med* 69:109–119

8. Almutairi H, Hassan G, Datta A (2021) Classification of obstructive sleep apnoea from single-lead ecg signals using convolutional neural and long short term memory networks. *Biomed Sig Process Control* 69:102906
9. Olsen M, Mignot E, Jennum P, Sorensen H (2020) Robust, ecg-based detection of sleep-disordered breathing in large population-based cohorts. *Sleep* 43(5):276
10. Olesen A, Jennum P, Mignot E, Sorensen H (2021) Msed: a multi-modal sleep event detection model for clinical sleep analysis. arXiv preprint [arXiv:2101.02530](https://arxiv.org/abs/2101.02530)
11. Sharma P, Jalali A, Majmudar M, Rajput KS, Selvaraj N (2022) Deep-learning based sleep apnea detection using spo2 and pulse rate. In: Proceedings of the 2022 44th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 2611–2614
12. Wang Z, Yan W, Oates T (2017) Time series classification from scratch with deep neural networks: a strong baseline. In: Proceedings of the 2017 international joint conference on neural networks (IJCNN). IEEE, pp 1578–1585
13. Karatoprak E, Seker S (2019) An improved empirical mode decomposition method using variable window median filter for early fault detection in electric motors. *Math Probl Eng* 2019:1–9

Application of Fuzzy Logic in Stock Markets by Using Technical Analysis Indicators



Leow Wei Kang, Mohd Izzat Nordin, Abdul Sattar Din,
and Mohamad Tarmizi Abu Seman

Abstract Investing with proper understanding on the stock is risk taking, while investing based on rumor is gambling. Increase in retail investors during the pandemic gives an opportunity of a more volatile market, therefore it leads to a need of using proper tools to screen through the stock market to find worthy assets to invest in. Previous research shows that there are possibilities of utilizing artificial intelligence techniques in assessing the market performance such as utilizing genetic algorithm with moving average convergence-divergence to generate trading signals or using deep recurrent neural network with closing data to predict the next day stock price. Even though there are good algorithms out there, it has not been utilized and made available to the public to access. Therefore, this project aims to develop an application to screen through the market using artificial intelligence techniques such as fuzzy logic to find a company which is worthy to invest in. Historical price data from 100 companies that made up the Kuala Lumpur Composite Index (KLCI) is used to assess the performance of the fuzzy logic application developed. The technical indicators used for the system is RSI, stochastic and MACD. The trading strategy using this application is to select stocks which have score lower than 0.5 for a buy signal. The results almost achieve the primary objective of generating 70% correct buy signals for short-term trading.

Keywords Fuzzy inference system · Technical indicator · RSI · MACD · Stochastic

L. W. Kang · M. I. Nordin · A. S. Din · M. T. A. Seman (✉)
School of Electrical and Electronic Engineering, Engineering Campus, Universiti Sains Malaysia,
14300 Nibong Tebal, Pulau Pinang, Malaysia
e-mail: mohdtarmizi@usm.my

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
N. S. Ahmad et al. (eds.), *Proceedings of the 12th International Conference on
Robotics, Vision, Signal Processing and Power Applications*, Lecture Notes in Electrical
Engineering 1123, https://doi.org/10.1007/978-981-99-9005-4_59

469

1 Introduction

Retail investors are increasing and are actively trading in Kuala Lumpur Stock Exchange (KLSE) during the pandemic. The ability to trade online and ease of access of business performance promotes the growth of retail investors community. The value of shares traded by retail investors was RM32.3 billion in June, while foreign investment was RM14.6 billion and local institutional fund was RM30.28 billion [1]. The increase in amount of retail investment is also due to the low interest rate of fixed deposit of Malaysian bank due to CoViD-19. The return per annum offered by most banks was $< 3\%$ [2]. Most people prefer to use their money to invest in the stock market rather than keeping their money in fixed deposits. This is because investment in the stock market will provide more return compared to fixed deposits.

When to buy in, when to exit has been a huge question for retail investors. They regret most of their decision most of the time, by selling early when the stock is on rise and not cutting loss when the stocks start to go down [3]. Some of them even chase the stock price when it is going up, believing that it will continuously go up but ended up the stock price went down.

Due to lack of analysis and judgement, retail investors tend to follow the majority when making investment decisions based on incomplete information, analyst reputation and speculative mentality [4]. This is what is known as herd mentality. Having access to social media makes things even worse. Since investment advice is heavily sought after during this pandemic, this causes the rise to some self-proclaimed investment gurus, who give investment advice without license from Security Commission [5]. This further increases the risk if retail investors do not make any effort to analyse the recommended stock and fall into the scheme of the investment gurus, resulting in loss of money due to buy high.

Contra traders or intraday traders are short-term investors who sell and buy the same shares without paying for them within the 3-day period. They will either pay the buying price for the time or paying or receiving the difference from the buying price after they sell the stock for losing and gaining respectively. This is a very risky method as it will incur a lot of losses if there is an unexpected movement of the stock price [6].

2 Methodology

A fuzzy logic inference system using technical indicators such as MACD, RSI and stochastic oscillator is proposed. The methods include the selection of data for program performance testing, methods on acquiring end of day data using Uniform Resource Identifier (URI), procedure on processing the data acquired, development of fuzzy logic system and methods on evaluating the performance of the program.

The data returned includes the opening price, closing price, high and low for the day, adjusted close price and traded volume. The data for closing price is extracted for calculation for RSI, stochastic and MACD. The calculation is done using MATLAB Financial Toolbox. The period for RSI is set to 14 days. The period for stochastic is set to 3 periods for %D and 14 days for %K, however only slow (smoothed) %K is used in the analysis. This is to prevent the stochastic oscillator from having drastic changes due to an increase in sensitivity towards price fluctuations. MACD is set to 12 EMA—26 EMA and the signal line is set to 9 EMA. The technical indicators settings are the conventional setting for short-term trading. For RSI and stochastic, the values are divided by 100 so that input to the fuzzy system is normalized between 0 and 1 as RSI and stochastic are returned as percentage. For MACD, two features are being observed. One of them is the difference between the MACD and the signal line [7]. The other one is the changes of the difference compared to the previous difference. Since the output of the results are not bounded, normalization is required. However, conventional scaling from 0 to 1 is not suitable as the value 0 is quite important in the difference between the MACD and signal line. It is a signal that a crossing occurs and it usually signifies a buy or a sell call depending on the condition of it converges or diverges.

The fuzzy inference system used is Sugeno FIS instead of Mamdani FIS. It is more computationally efficient and well suited for mathematical analysis.

Figure 1 shows the proposed Sugeno FIS. There are 4 inputs to the system, which is the RSI, stochastic, MACD_CD which is the measurement of the difference between current day and previous day MACD line and signal line difference. The final one is the MACD_zero cross which is the measurement of difference between MACD line and signal line. In the fuzzification process, AND method is the product of the input to ensure all variables in a rule are involved in the computation. The defuzzification process uses the weighted average method so that the output is between 1 and 0. All the input values had been normalized to the range of 0–1 during data processing.

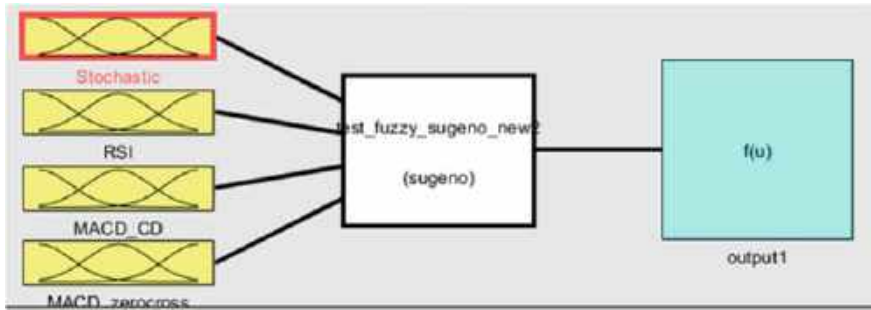


Fig. 1 Proposed Sugeno FIS

3 Results and Discussions

Table 1 shows the performance of the fuzzy logic system over 18 days of observation. There are 206 buy signals generated throughout the period. The overall accuracy is 34.47% for intraday trading (T), 52.43% for contra trading (3D), 65.05% for week trading (5D), 79.61% for 15 days observation period (15D) and finally 63.59% for generating profit larger than 1% within 15 days observation (> 1%E).

From Fig. 2, it is observed that the accuracy of the buy signals generated increases as the observation day increases. This is because when the observation period is longer, it gives more tolerance for the stock to fluctuate, hence increasing the chances of higher close within the period. An example could be seen from a buy signal generated on 25th May on YTLREIT.

Figure 3 shows the price change of YTLREIT. It is observed that even though within 5 days the stock price does not close high, there are still chances for the stock to close with higher price after longer observation period.

The fuzzy logic system developed analyze the stock generally, where it utilizes technical indicator performance for analysis. However, there are possibilities where the real stock price performance does not agree with the general strategy. A good example will be BURSA (1818).

From the orange boxes in Fig. 4, Bursa was among the most recommended stock to buy for the day. However, except for 4th May, other buy signals were off. It is also observed that the score decreases for 5th May but increases on 6th May. 4th May closes higher, indicating that the stock is going to move up, resulting in the score for Bursa in 5th May decreases. However, on 5th May, the closing price drops instead of increases. The rise of score on 6th May is caused by the correction from 5th May, signaling the buying confidence decreases.

Table 1 Performance of fuzzy logic system

Date	Samples	T	T (%)	3 days	3D (%)	5 days	5D (%)	15 days	15D (%)	> 1%E	> 1%E (%)
3-May	8	2	25.00	4	50.00	5	62.50	5	62.50	4	50.00
4-May	6	1	16.67	1	16.67	2	33.33	2	33.33	0	0.00
5-May	7	0	0.00	2	28.57	3	42.86	4	57.14	2	28.57
6-May	4	1	25.00	2	50.00	2	50.00	2	50.00	2	50.00
7-May	4	3	75.00	4	100.00	4	100.00	4	100.00	2	50.00
10-May	13	3	23.08	5	38.46	6	46.15	6	46.15	2	15.38
11-May	10	2	20.00	3	30.00	5	50.00	6	60.00	3	30.00
12-May	4	1	25.00	3	75.00	3	75.00	3	75.00	2	50.00
17-May	10	3	30.00	5	50.00	5	50.00	9	90.00	8	80.00
18-May	8	4	50.00	4	50.00	4	50.00	8	100.00	6	75.00
19-May	17	1	5.88	1	5.88	6	35.29	14	82.35	12	70.59
20-May	7	0	0.00	3	42.86	5	71.43	5	71.43	5	71.43
21-May	3	3	100.00	3	100.00	3	100.00	3	100.00	3	100.00
24-May	6	5	83.33	6	100.00	6	100.00	6	100.00	6	100.00
25-May	17	5	29.41	10	58.82	11	64.71	15	88.24	15	88.24
27-May	16	11	68.75	13	81.25	14	87.50	15	93.75	12	75.00
28-May	44	12	27.27	20	45.45	31	70.45	35	79.55	28	63.64
31-May	22	14	63.64	19	86.36	19	86.36	22	100.00	19	86.36
Overall	206	71	34.47	108	52.43	134	65.05	164	79.61	131	63.59



Fig. 2 Accuracy of buy signals generated

Fig. 3 Changes of YTLREIT (5109) Closing Price in 15 days



4 Conclusions

Overall, this project manages to get close to the objective of 70% accuracy (63.59%) in generating at least 1% return). It manages to prove the efficiency of using conventional technical indicators as a variable in generating correct buy signals. The inaccuracy of the signal generated is because technical indicators are just an analysis of past trends. It is still subject to the investors' interest in the stock to push up or pull down the price for the future. Besides, it also depends on other factors such as economic stability, sector interests, demand and supply. These factors might fluctuate wildly due to unexpected issues. The success of developing this system even though there are still improvement spaces proves that fuzzy logic is suitable to be used when dealing with uncertainties and lack of information since stock price movement is always full of uncertainties and no matter how much research is made, not all information of the company or the investors sentiments are captured.

Date	Stock Name	Stock Code	Score	TOpen	TClose	T+Close	T+2Close	T+3Close	T+4Close	T	T+1	T+2	T+3day	T+3	T+4	T+4	Days
4-May	BURSA	1818	0.143	8.46	8.52	8.51	8.45	8.46	8.45	1	1	0	1	0	0	4	1
	IOICORP	1961	0.353	4.09	4.07	4.04	4.08	4.1	4.1	0	0	0	0	0	1	1	1
	SAPNRG	5218	0.364	0.14	0.14	0.14	0.135	0.14	0.14	0	0	0	0	0	0	0	0
	YTLPOWR	6742	0.392	0.72	0.715	0.71	0.705	0.72	0.715	0	0	0	0	0	0	0	0
	GDEX	78	0.444	0.4	0.39	0.38	0.39	0.385	0.385	0	0	0	0	0	0	0	0
	DPHARMA	7148	0.466	2.98	2.92	2.85	2.82	2.84	2.87	0	0	0	0	0	0	0	0
	BURSA	1818	0.050	8.56	8.51	8.45	8.46	8.45	8.32	0	0	0	0	0	0	0	0
	GENP	2291	0.099	8.86	8.8	8.73	8.77	8.59	8.89	0	0	0	0	0	0	1	1
	SAPNRG	5218	0.408	0.14	0.14	0.14	0.135	0.14	0.135	0	0	0	0	0	0	0	0
	TENAGA	5347	0.422	9.98	9.9	9.92	9.9	9.9	9.91	0	0	0	0	0	0	0	0
5-May	IOICORP	1961	0.440	4.08	4.04	4.08	4.1	4.1	4.12	0	0	1	1	1	1	1	1
	EKOVEST	8877	0.451	0.465	0.455	0.455	0.455	0.445	0.44	0	0	0	0	0	0	0	0
	LCITIAN	5284	0.478	3.42	3.37	3.42	3.45	3.5	3.2	0	0	1	1	1	1	0	1
	BURSA	1818	0.149	8.49	8.45	8.46	8.45	8.32	8.2	0	0	0	0	0	0	0	0

Fig. 4 Snippet on buy signal generated for 4th May to 6th May

Acknowledgements This work was supported by Universiti Sains Malaysia Short Term Grant 304/PELECT/6315342.

References

1. Kana G (2020) Retail investors are back. Star
2. The ringgit plus team, best fixed deposit account in Malaysia. <https://ringgitplus.com/en/blog/fixed-deposits/best-fixed-deposit-accounts-in-malaysia.html>. Accessed 10 Dec 2020
3. Pareto C (2020) Understanding investor behaviour. <https://www.investopedia.com/articles/05/032905.asp>. Accessed 8 Dec 2020
4. Liu X, Liu B, Han X (2019) Analysis of herd effect of investor's behavior from the perspective of behavioral finance. In: Proceedings of the 2019 international conference on management, education technology and economics (ICMETE 2019)
5. Li KS (2020) Even if you're an investment guru, you can be jailed up to 10 years without a licence. Edge Malaysia
6. Tan L (2020) What is "Contra trading?". <https://www.straitstimes.com/business/invest/what-is-contra-trading>. Accessed 23 Nov 2020
7. Aguirre AAA, Medina RAR, Méndez NDD (2020) Machine learning applied in the stock market through the moving average convergence divergence (MACD) indicator. Invest Manag Finan Innov 17(4):44

YOLOv7-Tiny and YOLOv8n Evaluation for Face Detection



Ibrahim Al Amoudi and Dzati Athiar Ramli

Abstract The lightweight face detection models were developed to match the specifications of edge devices. This makes them suitable for use in real-life applications where the high GPU might not be available. This study compares the performance of two lightweight object detection models, YOLOv7-tiny and YOLOv8n, for face detection applications. Both models were trained on WIDERFACE dataset, and their performance was evaluated using mean average precision (mAP). Regarding model size, YOLOv8n is lighter than YOLOv7-tiny with only 3.01 million parameters, making it more efficient for deployment on resource-constrained devices. In terms of accuracy, the results showed that YOLOv7-tiny achieved a mAP50 of 69.8% and a mAP50-95 of 36.2%, while YOLOv8n achieved a mAP50 of 37% and an mAP50-95 of 67.6%

Keywords Face detection · Lightweight · YOLOv7 · YOLOv8

1 Introduction

Face detection is a crucial task in computer vision with numerous real-world applications. With the rise of edge computing and Internet of Things (IoT) devices, there is a growing demand for lightweight face detection models that can run on resource-constrained devices with limited computational power and memory. Traditionally, face detection models were based on classical machine learning algorithms such as Haar cascades [1, 2]. These models were lightweight and fast but struggled with challenges such as pose variations, occlusions, and low lighting conditions [3].

With the emergence of deep learning, Convolutional Neural Networks (CNNs) have become the state-of-the-art for face detection. However, most of the popular

I. Al Amoudi · D. A. Ramli (✉)

School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Engineering Campus, 14300 Nibong Tebal, Penang, Malaysia

e-mail: dzati@usm.my

CNN-based models such as MTCNN [4] or RetinaNet [5] are computationally expensive with millions of parameters and require high-end GPUs for training and inference. These models are not suitable for deployment on edge devices due to their high memory and power requirements. To address these challenges, researchers have developed a new generation of lightweight face detection models that are optimized for edge devices. These models employ various techniques such as Depthwise separable convolutions [6], Channel shuffling [7], and reparameterization [8] to reduce their size and computational complexity while maintaining a high level of accuracy.

Some examples of popular lightweight face detection models include Face SSD [9], BlazeFace [10], and SCRFD [11]. These models have achieved high accuracy while being lightweight and efficient, making them suitable for deployment on embedded systems, mobile devices, and IoT devices. Researchers have recently developed scalable object detection models that can be used in various scenarios. The recent YOLO series YOLOv7 [8], and YOLOv8 [12] are examples of scalable object detection models that can be utilized for face detection. YOLOv7-tiny and YOLOv8n are the smallest variants from their series that are suitable for deployment on edge devices.

In this paper, we trained two lightweight versions of the recent YOLO series: YOLOv7-tiny and YOLOv8n. We conducted a comparison of these models based on their size and accuracy. Furthermore, we evaluated the performance of both models on the WIDERFACE dataset.

2 Related Works

You Only Look Once (YOLO) is one of the earliest single-stage object detection models. YOLO is a real-time object detection algorithm that is fast and precise. It can predict bounding boxes and their class probabilities simultaneously using a single-stage CNN. After the success of YOLOv1, the authors modified the model and then published YOLOv2 followed by YOLOv3. In YOLOv2, the authors used anchor boxes in the head and Darknet19 as the backbone model. On YOLOv3 they replaced the backbone with a deeper CNN architecture named DarkNet-53. To solve the issue of detecting small objects, YOLOv3 uses multiscale prediction [13].

YOLOv3 was a great success yet it required more improvements to continue the competition with other object detection algorithms. Therefore Bochkovskiy et al. came up with a new version of YOLO and called it YOLOv4 in 2020. In YOLOv4 they modified the backbone architecture by using CSPDarknet53 and added SPP and PAN as neck [14]. YOLOv5 is another version of the YOLO series produced by Ultralytics in 2020. Ultralytics deployed several releases of YOLOv5 in GitHub. Similar to YOLOv4, YOLOv5 used CSPDarknet53 as a backbone, SPP and PAN as a neck, and the same head as YOLOv3. YOLOv5 has several variants, including YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, which vary in scale and level of accuracy. The smaller models, such as YOLOv5s, are designed for faster inference times and lower memory requirements, while the larger models, such as YOLOv5x, are designed for

improved accuracy and performance on complex object detection tasks. There is no change in the number of layers between YOLOv5 variants. The difference between them is in the scaling multipliers of the width and depth of the network [15]. After the release of YOLO5, Wang et al. proposed scaled-Yolov4 and developed a model scaling method by scaling up and down the number of stages in the model to utilize it for usage in different scenarios. They make YOLOv4-tiny, YOLOv4-CSP, P5, P6, and P7 variants [16]. To solve the issues associated with YOLOv4 the authors of YOLOv4 developed YOLOv7 which is another version of the YOLO series released in 2022. The authors used YOLOv4 and Scaled-YOLOv4 as bases for developing the model. The focus of YOLOv7 is to use optimization techniques that can help increase the model's efficiency and accuracy without increasing the inference cost [8]. YOLOv7 is one of the recent anchor-based object detection models. YOLOv8 is a scalable anchor-free version of the YOLO series. It was released by Ultralytics in 2022. YOLOv8 was developed to be an efficient and accurate object detection model [12].

3 Methodology

In this project, the lightweight face detection models based on object detection YOLOv7-tiny and YOLOv8n were trained and evaluated. The basic idea of a lightweight face detection model is as follows. First, the input image is fed into the face detection model. After that, the face detection model is responsible for extracting features, detecting, and locating faces within the images. Finally, the output detection results are generated, which provide information about the location and characteristics of the detected faces. Figure 1 shows the block diagram of the basic lightweight face detection model.

Object detection models can be broken down into several components, including the backbone, neck, and head. The backbone is the first component in the object detection model that takes in the input image and extracts high-level features from it. The extracted features are used to represent the image in a higher-level feature space, which helps in identifying objects in the image. The neck is the second component

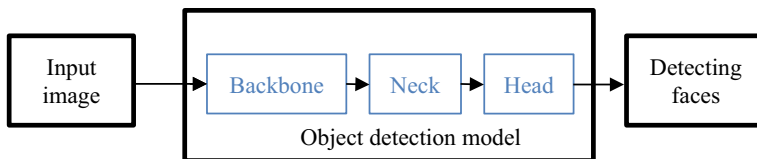


Fig. 1 The block diagram of the basic lightweight face detection model consists of three components: the input, the object detection model, and the output. The object detection model consists of three components: the backbone, neck, and head. First, input images are fed to the backbone to extract features, which are then enhanced by the neck. Finally, the head uses the results from the neck to detect and localize faces in the image

in the object detection model that takes the features extracted by the backbone and performs additional processing on them. The neck is usually a set of convolutional layers that are used to fuse and refine the features extracted by the backbone. The neck helps to improve the quality of the features and makes them more suitable for object detection. The head is the final component in the object detection model that uses the features extracted by the backbone and refined by the neck to detect and locate objects within the image. The heads can be classified into anchor-based and anchor-free heads.

YOLOv7 is an anchor-based object detection model while YOLOv8 is an anchor-free object detection model. Usually, the anchor-free models are lighter than the anchor-based models because the anchor-free models avoid the computation related to anchor boxes. Both YOLOv7 and YOLOv8 made several variants with different model sizes. Having several variants made both models suitable for deployment in various devices from a high-power GPU to a resource-constrained device. Both models try to provide the best trade-off between model accuracy and model size. Generally, larger and more complex models tend to have higher accuracy, but they also require more computing resources and storage space. On the other hand, smaller and simpler models may be less accurate, but they are more lightweight and faster to train and deploy. YOLOv7-tiny and YOLOv8n are chosen because they are the smallest variants of their series. Therefore, they are suitable for real-time face detection on re-source-constrained devices.

4 Experiment

4.1 Datasets

To verify the effectiveness of the targeted model's WIDERFACE dataset was used. WIDERFACE dataset is a widely used face detection benchmark dataset consisting of a diverse set of images with faces in various poses, scales, and occlusion levels. It includes over 32,000 images and more than 393,000 annotated face instances, making it an important resource for training and evaluating face detection models. The dataset is popular among researchers and has been used in numerous studies to develop and test new face detection algorithms [17].

4.2 Experimental Setting

Firstly, the dataset images were resized to 640×640 , and changed the annotation to fit the new size. Only the training and validation images from WIDERFACE dataset were used for both models. Then Roboflow online tool was used to change the annotation files to YOLO format. Secondly, YOLOv7-tiny was trained using the pre-trained

model. For the training, the input size was set to 640×640 , batch size 16, and the number of epochs to 100. Thirdly, the same setting was repeated with YOLOv8n. Both models were trained using Google Colab's Nvidia Tesla T4. After each epoch, the performance of two models was recorded and then evaluated using standard evaluation metrics, which included mean average precision at 50 IoU threshold (mAP50) and mean average precision over the intersection over union (IoU) thresholds ranging from 50 to 95% (mAP50-95). mAP50 represents the average precision calculated when the model detects an object with at least 50% overlap with the ground truth bounding box. In contrast, mAP50-95 represents the average precision calculated when the model detects an object with different levels of overlap with the ground truth bounding box, ranging from 50 to 95%. The model size was evaluated based on two metrics, the number of parameters used and Giga Floating Point Operations Per Second (GFLOPs). GFLOPs measures the computational complexity of a neural network model and indicates the number of floating-point operations a model can perform in one second. A higher GFLOPs value indicates a more computationally expensive model. The number of parameters in a neural network model includes the learnable parameters like weights and biases that the model learns during training to make accurate predictions. The complexity of the model increases with the number of parameters, and it requires more data to learn effectively.

5 Results and Conclusion

After the models were trained on WIDERFACE dataset, they were put into comparison based on model size and accuracy. For the model size, YOLOv7-tiny contains a 6.01 M parameter with 13.2 GFLOPs while YOLOv8n contains a 3.01 M parameter with 8.2 GFLOPs. With this, we can see that YOLOv8n is significantly smaller than the YOLOv7-tiny. For the mAP50, YOLOv7-tiny reached 69.8% and YOLOv8n reached 67.6%, and for the mAP50-95 YOLOv7-tiny achieved 36.2% and YOLOv8n achieved 37%. With mAP50 YOLOv7-tiny achieved better results and with mAP50-95 YOLOv8n was better. This result shows YOLOv8n is more efficient than YOLOv7-Tiny because the model size is significantly smaller than YOLOv7-Tiny while there are small differences in the accuracy metrics. Therefore, YOLOv8n is more suitable to be deployed in edge devices than YOLOv7-tiny (Fig. 2, Table 1).

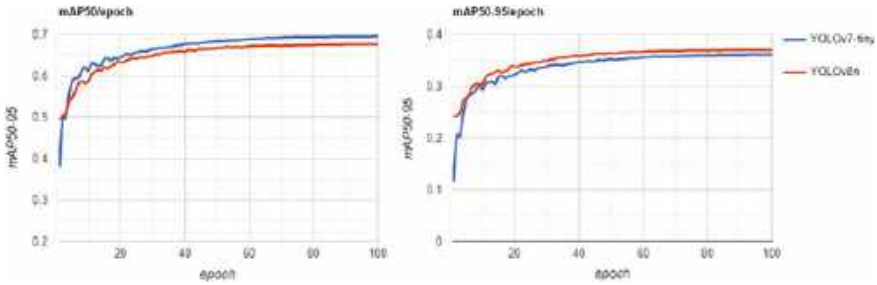


Fig. 2 The first graph shows the mAP50 per epoch result of YOLOv7-tiny and YOLOv8n. The second shows the mAP50-95 per epoch result of YOLOv7-tiny and YOLOv8n

Table 1 Comparison between YOLOv7-tiny and YOLOv8n on WIDERFACE dataset

Model	Param	GFLOPs	mAP50	mAP50-95
YOLOv7-tiny	6.01 M	13.2	0.698	0.362
YOLOv8n	3.01 M	8.2	0.676	0.37

Acknowledgements This paper is supported under the Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project Code: FRGS/1/2019/TK04/USM/02/1.

References

- Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, vol 1. IEEE
- Viola P, Jones MJ (2004) Robust real-time face detection. *Int J Comput Vis* 57:137–154
- Jadhav A, Lone S, Matey S, Madamwar T, Jakhete S (2021) Survey on face detection algorithms. *Int J Innov Sci Res Technol* 6:291–297
- Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Sig Process Lett* 23(10):1499–1503
- Lin TY, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988
- Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T et al (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint [arXiv: 1704.04861](https://arxiv.org/abs/1704.04861)
- Zhang X, Zhou X, Lin M, Sun J (2018) Shufflenet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6848–6856
- Wang CY, Bochkovskiy A, Liao HYM (2022) YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint [arXiv:2207.02696](https://arxiv.org/abs/2207.02696)
- Ye B, Shi Y, Li H, Li L, Tong S (2021) Face SSD: a real-time face detector based on SSD. In: Proceedings of the 2021 40th Chinese control conference (CCC). IEEE, pp 8445–8450
- Bazarevsky V, Kartynnik Y, Vakunov A, Raveendran K, Grundmann M (2019) Blazeface: sub-millisecond neural face detection on mobile GPUs. arXiv preprint [arXiv:1907.05047](https://arxiv.org/abs/1907.05047)

11. Zhu X, Lou Y (2022) An efficient anchor-free face detector with attention mechanisms. Sci Program
12. YOLOv8 Docs'. <https://docs.ultralytics.com/#ultralytics-yolov8>
13. Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. arXiv preprint [arXiv:1804.02767](https://arxiv.org/abs/1804.02767)
14. Bochkovskiy A, Wang CY, Liao HYM (2020) Yolov4: optimal speed and accuracy of object detection. arXiv preprint [arXiv:2004.10934](https://arxiv.org/abs/2004.10934)
15. GitHub: ultralytics/yolov5: YOLOv5 in PyTorch > ONNX > CoreML > TFLite. <https://github.com/ultralytics/yolov5>
16. Wang CY, Bochkovskiy A, Liao HYM (2021) Scaled-yolov4: scaling cross stage partial network. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, pp 13029–13038
17. Yang S, Luo P, Loy CC, Tang X (2016) Wider face: a face detection benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5525–5533

Optimizing Feature Selection for Industrial Casting Defect Detection Using QLESCA Optimizer



Qusay Shihab Hamad , Sami Abdulla Mohsen Saleh ,
Shahrel Azmin Suandi , Hussein Samma , Yasameen Shihab Hamad ,
and Ibrahim Al Amoudi

Abstract Feature selection is critical in fields like data mining and pattern classification, as it eliminates irrelevant data and enhances the quality of highly dimensional datasets. This study explores the effectiveness of the Q-learning embedded sine cosine algorithm (QLESCA) for feature selection in industrial casting defect detection using the VGG19 model. QLESCA's performance is compared to other optimization algorithms, with experimental results showing that QLESCA outperforms the other algorithms in terms of classification metrics. The best accuracy achieved by QLESCA is 97.0359%, with an average fitness value of -0.99124 . The proposed method provides a promising approach to improve the accuracy and reliability of industrial casting defect detection systems, which is essential for product quality and safety. Our findings suggest that using powerful optimization algorithms like QLESCA is crucial for obtaining the best subsets of information in feature selection and achieving optimal performance in classification tasks.

Keywords Machine learning · Support vector machine (SVM) classifier · Convolutional neural networks · Swarm intelligence · Feature engineering

Q. S. Hamad · S. A. M. Saleh · S. A. Suandi (✉) · I. Al Amoudi
Intelligent Biometric Group, School of Electrical and Electronic Engineering, Engineering
Campus, Universiti Sains Malaysia, 14300 Nibong Tebal, Penang, Malaysia
e-mail: shahrel@usm.my

Q. S. Hamad
University of Information Technology and Communications (UOITC), Baghdad, Iraq

H. Samma
SDAIA-KFUPM Joint Research Center for Artificial Intelligence (JRC-AI), King Fahd University
of Petroleum and Minerals, Dhahran, Saudi Arabia

Y. S. Hamad
Ministry of Education, Baghdad, Iraq

1 Introduction

Casting is a common manufacturing method that involves pouring liquid material into a mold with a hollow shape and allowing it to harden. However, certain casting faults may occur, which can affect the quality of the final product. Identifying internal flaws during the casting process is crucial for maintaining high product standards. Early detection of these faults can help identify defective items quickly and save both time and money. Undetected faults can lead to the failure of critical mechanical components. Thus, intelligent systems play an essential role in production lines to increase detection accuracy and maintain product quality [1]. Numerous studies have explored deep-learning-based techniques for detecting casting defects in literature. For example, Wang et al. [2] proposed a deep model to detect subtle casting defects in aluminum alloys using X-ray images. Du et al. [1] developed an X-ray-oriented deep learning-based defect detection system. Pastor-López et al. [3] proposed a quality assessment system based on machine learning using a small dataset. Wu et al. [4] introduced BX-Net, a transfer learning approach that accurately detects casting defects in X-ray images and extracts highly discriminative features from diverse categories. Ji et al. [5] proposed an artificial intelligence approach to detect and recognize faults in aerospace titanium castings using X-ray images. Jiang et al. [6] developed a weakly-supervised Convolutional Neural Network model for detecting faults in casting X-ray images. Han et al. [7] suggested a deep convolutional network approach for defect segmentation in polycrystalline silicon wafers.

Feature selection is a crucial preprocessing step in machine learning to reduce the amount of data in a set. Several algorithms have been proposed to address the combinatorial optimization problem of selecting an optimal subset of features that can maintain or enhance classification accuracy. Bacanin et al. [8] presented an enhanced version of the arithmetic optimization algorithm (AOA) that selects an optimal feature subset for a specific classification problem. In another work, Zhong et al. [9] hybridized the equilibrium optimizer with an artificial bee colony for feature selection. Houssein et al. [10] utilized the Sooty Tern Optimization Algorithm to improve classification accuracy by selecting optimal features. More recently, Hamad et al. [11] applied the Q-learning embedded sine cosine algorithm (QLESCA) to select a subset of features for COVID-19 detection.

The main aim of this work is to enhance the performance of automatic defect detection in industrial casting by selecting the optimal features extracted via a deep learning model. We use the QLESCA algorithm [12] for feature selection and evaluate its performance against other optimization algorithms. The results of this study will aid in the development of more accurate and efficient defect detection systems in the manufacturing industry.

The remaining sections of this paper are organized as follows: Sect. 2 presents the dataset used in the study. In Sect. 3, the methodology is proposed. The experimental results are presented in Sect. 4. Finally, in Sect. 5, conclusions are drawn and future research directions are discussed.

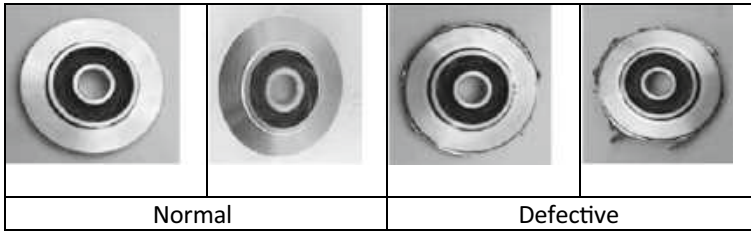


Fig. 1 Various normal and defective casting case images

2 Dataset Used in the Study

In order to authenticate the viability of our suggested framework, an assemblage of 8012 images [13] was employed as the foundation. This dataset is inherently partitioned into two distinct classifications: 4613 instances characterized as defective samples and 3399 instances cataloged as normal samples. The allocation of this dataset adhered to a bifurcation wherein 60% of the data, encompassing 2768 defective and 2039 normal samples, was reserved for the training endeavor. The remaining 40%, comprising 1845 defective and 1360 normal samples, was designated for the testing phase. Visual insight into the dataset is proffered through Fig. 1, which visually presents a selection of normal and defective casting case images.

3 Methodology

Within this investigation, we harnessed a Convolutional Neural Network (CNN)-based framework to conduct feature extraction. Specifically, the architecture of choice was VGG19. VGG19 encompasses a trio of fully connected strata (namely fc6, fc7, and fc8) boasting dimensions of 4096, 4096, and 1000 correspondingly. These strata encapsulate features gleaned from the dataset, a portion of which may prove superfluous and cast a detrimental shadow on classification precision. To counteract this concern, we deliberately focused solely on the features originating from fc8 as the designated feature subset for our model. Subsequently, we invoked the QLESCA algorithm to curate a subset of these features to be employed in classification tasks via a support vector machine (SVM). The architectural blueprint of our proposed model is visually expounded in Fig. 2, while the underpinning fitness function finds its elucidation in Eq. 1.

$$\text{Fitness, } F = K * (\alpha * A - \beta * (S_Features/T_Features)) \tag{1}$$

In Formula 1, A symbolizes the accuracy of classification, while $S_Features$ denotes the count of chosen characteristics. The variables α and β , both confined

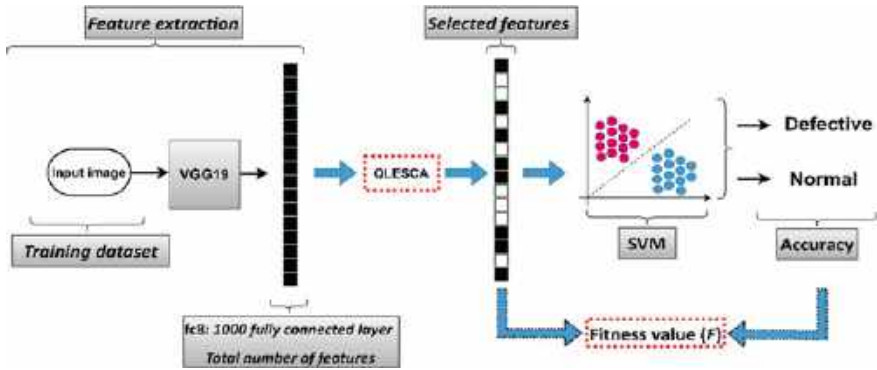


Fig. 2 The proposed model

within the $[0, 1]$ spectrum, epitomize the varying significance assigned to classification precision and the count of chosen features within the fitness function. $T_Features$ signifies the complete assortment of features within $fc8$ (amounting to 1000 features) extracted through the VGG19 model. To align with the optimization algorithm’s pursuit of minimizing the fitness value, the outcome of the fitness evaluation is scaled by a factor of $K = -1$, rendering it negative.

4 Experimental Results

The effectiveness of QLESCA in addressing the feature selection challenge was assessed by comparing it against two contemporary optimization algorithms: Sine Cosine Algorithm (SCA) [14, 15] and AOA [16]. All algorithmic setting utilized in comparison remained consistent with the original algorithms. Notably, each experiment was meticulously replicated 10 times to mitigate the influence of random variations on the outcomes. Furthermore, it’s pertinent to highlight that the maximum iteration count in this investigation was set at 200. Table 1 displays the best, mean, median, worst, and standard deviation values for all tested algorithms. The results demonstrate that QLESCA outperformed the other algorithms. As shown, QLESCA achieved better results. Further analysis was conducted by plotting the convergence curve of QLESCA compared to other evaluated algorithms, as illustrated in Fig. 3. The horizontal axis represents 200 iterations, and the vertical axis represents the best score obtained.

The experimental results presented in Table 2 demonstrate that the features selected via QLESCA outperformed the other optimization algorithms in terms of accuracy, precision, specificity, sensitivity, and F1-score. Specifically, the QLESCA algorithm attained the highest average performance across all classification metrics, indicating that it is the most effective algorithm for feature selection in industrial casting defect detection using the VGG19 model. These findings suggest that

Table 1 The average cost and standard deviation of QLESCA against other optimizers

Algorithm	Best	Mean	Median	Worst	STD
QLESCA	- 0.9926	- 0.99124	- 0.99137	- 0.9899	0.001
SCA	- 0.9898	- 0.98915	- 0.98898	- 0.9889	0.0004
AOA	- 0.9897	- 0.98917	- 0.9891	- 0.9885	0.0005

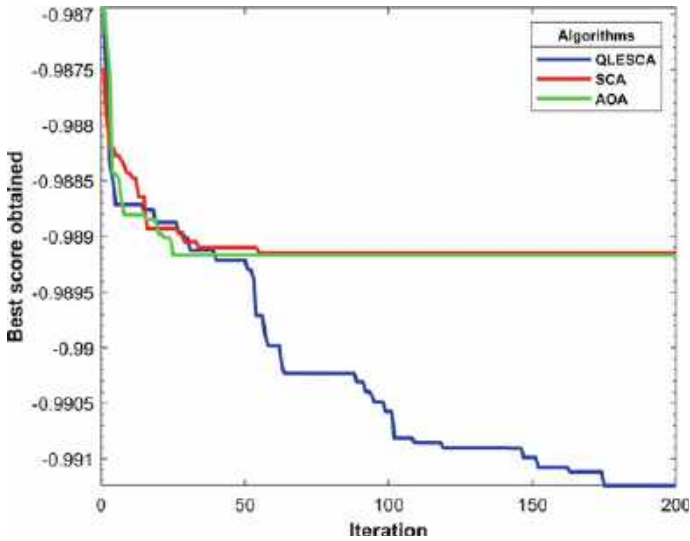


Fig. 3 The convergence curve of QLESCA, SCA, and AOA in SD model

QLESCA can be a valuable tool for improving the accuracy and reliability of industrial casting defect detection systems.

Table 2 Comparison between QLESCA and other algorithms regarding the optimal parameters on the SD model

Model	TP	TN	FP	FN	Accuracy (%)	Precision (%)	Specificity (%)	Sensitivity/recall (%)	F1-score (%)
QLESCA	1790	1320	40	55	97.0359	97.8142	97.0588	97.0190	97.4150
SCA	1782	1314	46	63	96.5991	97.4836	96.6176	96.5854	97.0324
AOA	1785	1316	44	60	96.7551	97.5943	96.7647	96.7480	97.1693

Bold indicates, the best value corresponds to the highest one

5 Conclusion

In this study, we proposed a novel feature selection algorithm, QLESCA, for industrial casting defect detection using the VGG19 model. Our results show that QLESCA outperforms other optimization algorithms in terms of accuracy, precision, specificity, sensitivity, and F1-score, demonstrating its effectiveness in selecting the most informative features for defect detection. Our findings suggest that QLESCA can be a valuable tool for improving the accuracy and reliability of industrial casting defect detection systems. Overall, our study highlights the importance of feature selection in machine learning-based defect detection systems, and the potential benefits of using QLESCA in industrial applications. Future work can investigate the scalability and generalizability of QLESCA to other datasets and models, and explore the potential of integrating it with other techniques for further improvements in defect detection performance.

Acknowledgements We extend our sincere appreciation to the Malaysia Ministry of Higher Education (MOHE) for their invaluable support through the Fundamental Research Grant Scheme (FRGS), under grant no. FRGS/1/2019/ICT02/USM/03/3.

References

1. Du W, Shen H, Fu J, Zhang G, He Q (2019) Approaches for improvement of the X-ray image defect detection of automobile casting aluminum parts based on deep learning. *NDT Eng Int* 107:102144. <https://doi.org/10.1016/j.ndteint.2019.102144>
2. Wang Y, Hu C, Chen K, Yin Z (2020) Self-attention guided model for defect detection of aluminium alloy casting on X-ray image. *Comput Electr Eng* 88:106821. <https://doi.org/10.1016/j.compeleceng.2020.106821>
3. Pastor-López I, Sanz B, Tellaache A, Psaila G, de la Puerta JG, Bringas PG (2021) Quality assessment methodology based on machine learning with small datasets: industrial castings defects. *Neurocomputing* 456:622–628. <https://doi.org/10.1016/j.neucom.2020.08.094>
4. Wu B et al (2021) An ameliorated deep dense convolutional neural network for accurate recognition of casting defects in X-ray images. *Knowl Based Syst* 226:107096. <https://doi.org/10.1016/j.knosys.2021.107096>
5. Ji X et al (2021) Filtered selective search and evenly distributed convolutional neural networks for casting defects recognition. *J Mater Process Technol* 292:117064. <https://doi.org/10.1016/j.jmatprotec.2021.117064>
6. Jiang L, Wang Y, Tang Z, Miao Y, Chen S (2021) Casting defect detection in X-ray images using convolutional neural networks and attention-guided data augmentation. *Measurement* 170:108736. <https://doi.org/10.1016/j.measurement.2020.108736>
7. Han H, Gao C, Zhao Y, Liao S, Tang L, Li X (2020) Polycrystalline silicon wafer defect segmentation based on deep convolutional neural networks. *Pattern Recognit Lett* 130:234–241. <https://doi.org/10.1016/j.patrec.2018.12.013>
8. Bacanin N et al (2023) Quasi-reflection learning arithmetic optimization algorithm firefly search for feature selection. *Heliyon* 9(4):e15378. <https://doi.org/10.1016/j.heliyon.2023.e15378>
9. Zhong C, Li G, Meng Z, Li H, He W (2023) A self-adaptive quantum equilibrium optimizer with artificial bee colony for feature selection. *Comput Biol Med* 153:106520. <https://doi.org/10.1016/j.compbimed.2022.106520>

10. Houssein EH, Oliva D, Çelik E, Emam MM, Ghoniem RM (2023) Boosted sooty tern optimization algorithm for global optimization and feature selection. *Exp Syst Appl* 213:119015. <https://doi.org/10.1016/j.eswa.2022.119015>
11. Hamad QS, Samma H, Suandi SA (2023) Feature selection of pre-trained shallow CNN using the QLESCA optimizer: COVID-19 detection as a case study. *Appl Intell* 6:1–23. <https://doi.org/10.1007/s10489-022-04446-8>
12. Hamad QS, Samma H, Suandi SA, Mohamad-Saleh J (2022) Q-learning embedded sine cosine algorithm (QLESCA). *Exp Syst Appl* 193:116417. <https://doi.org/10.1016/j.eswa.2021.116417>
13. Dabhi R (2020) Casting product image data for quality inspection. <https://www.kaggle.com/ravirajsinh45/real-life-industrial-dataset-of-casting-product>
14. Mirjalili S (2016) SCA: a Sine Cosine Algorithm for solving optimization problems. *Knowl Based Syst* 96:120–133. <https://doi.org/10.1016/j.knosys.2015.12.022>
15. Hamad QS, Samma H, Suandi SA, Saleh JM (2022) A comparative study of sine cosine optimizer and its variants for engineering design problems, pp 1083–1089. https://doi.org/10.1007/978-981-16-8129-5_166
16. Abualigah L, Diabat A, Mirjalili S, Abd-Elaziz M, Gandomi AH (2021) The arithmetic optimization algorithm. *Comput Methods Appl Mech Eng* 376:113609. <https://doi.org/10.1016/j.cma.2020.113609>

Improving the Accuracy of Gender Classification Based on Skin Tone Using Convolutional Neural Network: Transfer Learning (CNN-TL)



Muhammad Firdaus Mustapha , Nur Maisarah Mohamad ,
Siti Haslini Ab Hamid , and Nik Amnah Shahidah Abdul Aziz 

Abstract Gender classification is one of the key features in soft biometrics besides age, ethnicity, facial expression, etc. Gender classification based on skin tone has its own importance that can further improve the performance of facial recognition systems. Most Convolutional Neural Network (CNN) models require a large amount of training data to improve classification accuracy and increase processing time. Fortunately, the MobileNetV2 model overcomes this problem by running faster than the other models. However, the model's accuracy suffers when the gender classification results based on skin tone reach 50% accuracy, indicating that the model suffers from an “overfitting” problem. To address this issue, the proposed research constructed a novel face images dataset containing 6250 face images that chosen from original FaceARG dataset and divided equally into Bright (3125) and Dark (3125) skin tone. Each skin tone (Bright and Dark) is equally divided into two genders, 1563 (male) and 1563 (female). The new FaceARG dataset is then used to run two types of experiments (Bright and Dark) on the MobileNetV2 model. The Fine-Tuning method from Transfer Learning is then applied to the MobileNetV2 model, along with method gains from previous studies. The Dark experiment achieved the highest accuracy on training dataset, which is 97.4%, compared to 50% on the model without fine-tuning, and the Bright experiment achieved the highest accuracy on test dataset, which is 89.8%, compared to 50% on the model without fine-tuning. The findings of this study will determine the ability of the proposed model to accurately classify gender based on skin tone.

M. F. Mustapha · N. M. Mohamad (✉) · N. A. S. Abdul Aziz
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Cawangan
Kelantan, Bukit Ilmu, 18500 Machang, Malaysia
e-mail: 2020273478@student.uitm.edu.my

M. F. Mustapha
e-mail: mdfirdaus@uitm.edu.my

S. H. Ab Hamid
Department of Information Technology, FH Training Center, 16800 Pasir Puteh, Malaysia

Keywords Gender classification · Convolutional neural network · Fine-tuning · Transfer learning

1 Introduction

Gender classification is a well-known binary classification type of classification that involves two types of categories (male and female). Recently, the issue of uneven accuracy rates for different groups has been revealed in the context of gender classification using images of faces with different skin tones.

In [1], suggests a combination of gender and race research. They investigated the differential performance of gender classification algorithms across gender racial groups using the UTKFace [2] and FairFace [3] datasets. For all the experiments they ran, Black Race and Black Women in the dataset had the lowest accuracy rates. Middle Eastern men and Latino women obtained the highest accuracy rates, also observed in [4]. This is the case, possibly due to the reflective nature of skin color in varying lighting combined with facial morphology. Nevertheless, another issue started to arise when the classification accuracy of the MobileNetV2 model plateaued at 50%, indicating “overfitting” behavior. To solve this issue, the proposed study improves the MobileNetV2 model using the Transfer Learning with Fine Tuning approach.

The proposed research presents two key contributions. The first contribution is collected a new balanced FaceARG dataset in a small-scale size separated between Bright and Dark skin tones while the second contribution introduced a CNN-TL model that uses a novel fine tuning method combined with a head model from previous work.

2 Literature Review

2.1 Skin Tone

The topic of skin tone is the important element of the proposed research which highlights gender classification methods based on their skin tone rather than their race or ethnicity. The previous research to tackle the problem in skin tone are depicted in Table 1.

As shown in Table 1, Haar Cascade classifier [9] is widely used in face detection due to its open source nature and fast execution time. The proposed research uses research that has been conducted in [6] but applied in Hex Color Code and categorized skin color using K-Means skin segmentation. As illustrated in Table 2, the skin tone is divided into three categories.

Table 1 Approaches to tackle the problem in skin tone

Author and year	Approaches	Face detection
Azzopardi et al. [5]	Automatic procedure that combines trainable shape and color features for gender classification	Viola Jones
Molina et al. [6]	K-means method was used to categorize the color faces using clusters of RGB pixel values	Pretrained CNN
Demirezen and Erdem [7]	Signal decomposition method, nonlinear mode decomposition (NMD) in estimating the heart rate signal from face videos in the presence of subject motion	Viola Jones
Muthukumar et al. [8]	Color-theoretic, luminance mode-shift and optimal transport	Viola Jones

Table 2 Skin tone divide by color range using hex color code

Skin tone	Hex color code
Bright	[16777215–12619362]
Mild	[12619362–10300000]
Dark	Other than [8421504–0]

2.2 Convolutional Neural Network: Transfer Learning (CNN-TL)

Transfer learning is a technique used to improve learner performance in one domain by transferring data from related domains. It involves taking a network that has been trained on a dataset and applying it to a new dataset to recognize categories of untrained images or objects. It uses robust discriminatory filters learned by state-of-the-art networks on challenging datasets to identify objects for which the model has never been trained.

CNN enables more advanced functionality than traditional approaches, allowing for a more nuanced understanding of advanced architecture [10]. CNN is a popular computer vision leader, with deep convolution neural networks achieving high performance in a variety of tasks [11]. MobileNetV2 [12] was chosen as a low-power CNN architecture for applications based on mobile and embedded devices, as it performed better than other models in terms of execution time.

There are also researchers who combine both methods by performing fine-tuning methods and adding layers to Fully Connected (FC) layers. To the best of my knowledge, no researcher has specified exactly which layers they frozen or unfrozen during the fine-tuning process run against the MobileNetV2 model. The proposed research uses this opportunity to add novelty to the research by performing a fine-tuning process by determining which layers need to be frozen or unfrozen and combining them by modifying the layers on the FC layer during the MobileNetV2 model.

3 Methodology

The experiments were compiled starting with the input of the dataset, the preprocessing phase, the gender classification phase and subsequently obtaining the output i.e., gender classification based on skin tone. The experimental design is illustrated in Fig. 1. The dataset used in this experiment was the FaceARG dataset [13] which is publicly available online. The FaceARG data collection includes four racial groups, but only considers two of them. The sample was split into Afro-Americans with dark skin and Asians with bright skin. The proposed research dataset only selects clearly visible faces that have been stripped of any accessories.

Preprocessing phase is the process of skin tone classification-based ethnicity and gender in FaceARG dataset. 6256 face images are divided into four classes, Bright Male, Bright Female, Dark Male, and Dark Female. The most important idea is that the preprocessing phase produces a new FaceARG data set, which is 3.6% of the total data set that will in turn be used in the gender classification phase. Datasets are separated using the holdout method which is 80% for training and 20% for testing.

There are two separate datasets that are used in Gender Classification phase. The training data is divided into two types of data sets, which are training and validation data. The proposed gender classification is based on the method [14] of a previous study, which carried out experiments without using any face detector during the training phase and then saved the loaded model for use in the next step, on evaluation of architecture. Therefore, the proposed study takes into account what was told in [15] that a balanced dataset during the training process is important to maintain high accuracy against both classes. Face detection is not carried out in the training dataset, as it will reduce the amount of data when image datasets that cannot be detected are discarded. The model saved from the training phase is reused to classify gender using the test data set, providing gender classification results to the proposed model.

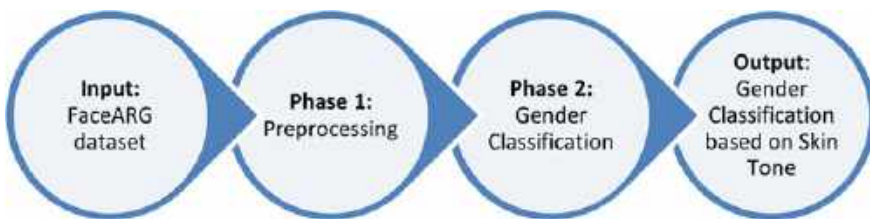


Fig. 1 Experimental design

Table 3 Result of the accuracy in training and test dataset

Model	Training dataset (%)		Test dataset (%)	
	Bright	Dark	Bright	Dark
Baseline	50.0	50.0	50.0	50.0
Freeze from bottleneck block 11 + custom head model [16]	97.2	92.0	89.6	81.5
Freeze from bottleneck block 12 + custom head model [16]	87.6	96.4	78.8	81.3
Freeze from bottleneck block 13 + custom head model [16]	95.8	97.4	89.8	84.3
Freeze from bottleneck block 11 + custom head model [17]	95.4	91.4	89.5	79.2
Freeze from bottleneck block 12 + custom head model [17]	96.4	93.6	89.0	86.1
Freeze from bottleneck block 13 + custom head model [17]	90.8	96.2	81.0	83.1

4 Result and Discussion

Table 3 stated that the overfitting conditions at the training phase, causing the model baseline to be overfitting and even worse than the training phase, due to the classification phase through the process of face detection approach before classification. In addition, the highest accuracy result for Bright skin tone was 89.8% accuracy, while for Dark skinned tone achieved 86.1% accuracy.

5 Conclusion and Future Research

The MobileNetV2 model without performing methods of transfer learning coupled with a small number of epochs, can lower accuracy, and even make the model overfitting. Overfitting is an undesirable behavior that indicates the model is only suitable for data training only and not for new data for my case validation data that causes it to be overfitting. There are many reasons for overfitting, including the lack of data training in the model, or the lack of epoch. Overfitted model on this proposed research is 50% accuracy that usually occurs in binary class.

Although the most important limitation is in hardware equipment, this research only uses CPU instead of GPU. Therefore, the future research will continue research related to this transfer learning using the GPU and then use the latest model to measure the performance of accuracy and processing time against gender classification based on this skin tone.

Acknowledgements The Fundamental Research Grant Scheme (FRGS) of the Ministry of Education (MOE) supported this research with grant number 600-IRMI/FRGS 5/3 (234/2019).

References

1. Nair AKU, Almadan A, Rattani A (2020) Understanding fairness of gender classification algorithms across gender-race groups. In: Proceedings of the 19th IEEE international conference on machine learning and applications
2. Kim YUH, Nam SEH, Park KR (2021) Enhanced cycle generative adversarial network for generating face images of untrained races and ages for age estimation, pp 6087–6112. <https://doi.org/10.1109/ACCESS.2020.3048369>
3. Karkkainen K, Joo J (2021) FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of the 2021 IEEE winter conference on applied computer vision, WACV 2021, pp 1547–1557. <https://doi.org/10.1109/WACV48630.2021.00159>
4. Karkkainen K, Joo J (2019) FairFace: face attribute dataset for balanced race, gender, and age
5. Azzopardi G, Foggia P, Greco A, Saggese A, Vento M (2018) Gender recognition from face images using trainable shape and color features, pp 1983–1988
6. Molina D, Causa L, Tapia J (2020) Toward to reduction of bias for gender and ethnicity from face images using automated skin tone classification. In: Lecture notes informatics (LNI), proceedings of the series gesellschaft fur information P-306, pp 281–289
7. Demirezen H, Erdem CE (2018) Remote photoplethysmography using nonlinear mode decomposition. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings. IEEE, pp 1060–1064
8. Muthukumar V, Pedapati T, Ratha N, Sattigeri P, Wu C, Kingsbury B, Kumar A, Thomas S, Mojsilovi A, Varshney KR (2019) Color-theoretic experiments to understand unequal gender classification accuracy from face images, pp 2286–2295. <https://doi.org/10.1109/CVPRW.2019.00282>
9. Benkaddour MK (2021) CNN based features extraction for age estimation and gender classification proposed method convolutional neural networks. *Informatica* 45:697–703
10. Zhou Y, Ni H, Recognition AF (2019) Face and gender recognition system based on convolutional neural networks, pp 1091–1095
11. Nagpal C, Dubey SR (2019) A performance evaluation of convolutional neural networks for face anti spoofing. In: Proceedings of the international joint conference on neural networks. IEEE, pp 1–8
12. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018) MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE computational society conference on computer visual pattern recognition, pp 4510–4520. <https://doi.org/10.1109/CVPR.2018.00474>
13. Darabant AS, Borza D, Danescu R (2021) Recognizing human races through machine learning—a multi-network, multi-features study. *Mathematics* 9:1–19. <https://doi.org/10.3390/math9020195>
14. Kaur G, Sinha R, Tiwari PK, Yadav SK, Pandey P, Raj R, Vashisth A, Rakhra M (2021) Face mask recognition system using CNN model. *Neurosci Inform* 53:100035. <https://doi.org/10.1016/j.neuri.2021.100035>
15. Krishnapriya K, King MC, Bowyer KW (2021) Analysis of manual and automated skin tone assignments for face recognition applications
16. Patel R, Chaware A (2020) Transfer learning with fine-tuned MobileNetV2 for diabetic retinopathy, pp 2020–2023
17. Asif S, Wenhui Y, Tao Y, Jinhai S, Amjad K (2021) Real time face mask detection system using transfer learning with machine learning method in the era of Covid-19 pandemic. In: Proceedings of the 2021 4th international conference on artificial intelligence on big data, ICAIBD, pp 70–75. <https://doi.org/10.1109/ICAIBD51990.2021.9459008>

Literature Survey on Edge Detection-Based Methods for Blood Vessel Segmentation from Retinal Fundus Images



Nazish Tariq, Shadi Mahmoodi Khaniabadi, Soo Siang Teoh, Shir Li Wang, Theam Foo Ng, Rostam Affendi Hamzah, Zunaina Embong, and Haidi Ibrahim

Abstract Retinal vessel segmentation is an essential step in the diagnosis of various retinal diseases. Edge detection-based methods have shown promising results for retinal vessel segmentation due to their ability to identify the boundaries of the vessels. In this paper, we surveyed several edge detection-based methods for retinal vessel segmentation from three main databases: PubMed, IEEEExplore, and Google Scholar. The outcomes from the literature search were filtered based on inclusion and exclusion criteria. From the selected literature, information about the edge detection techniques, the image datasets used, and the evaluation measures, are extracted. From this literature survey, we can see that there are many approaches that have been proposed by researchers to segment the blood vessel edges from the retinal fundus images. Most of them are using the traditional approaches, such as Sobel operators, and Canny edge detector. Recently, deep learning-based approaches have been proposed for this purpose. Some of the commonly used databases for retinal

N. Tariq · S. M. Khaniabadi · S. S. Teoh · H. Ibrahim (✉)

School of Electrical and Electronic Engineering, Engineering Campus, Universiti Sains Malaysia, 14300 Nibong Tebal, Penang, Malaysia
e-mail: haidi_ibrahim@ieee.org

S. L. Wang

Faculty of Computing and Meta-Technology, Universiti Pendidikan Sultan Idris (UPSI), Tanjung Malim, Perak, Malaysia

T. F. Ng

Centre of Global Sustainability Studies (CGSS), Level 5, Hamzah Sendut Library, Universiti Sains Malaysia, 11800 Minden, Penang, Malaysia

R. A. Hamzah

Fakulti Teknologi Kejuruteraan Elektrik Dan Elektronik, Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia

Z. Embong

Department of Ophthalmology, School of Medical Sciences, Health Campus, Universiti Sains Malaysia, 16150 Kubang Kerian, Kelantan, Malaysia

fundus images have also been reported in this review. Several evaluation measures that have been utilized by researchers have also been identified.

Keywords Edge detector · Retinal vessel segmentation · Fundus images

1 Introduction

Retinal vessels are one of the most important structures in the human eye as they play a crucial role in the diagnosis and treatment of various eye diseases such as diabetic retinopathy, glaucoma, and hypertension [1]. To detect these abnormalities, it is necessary to segment the retinal vessels from fundus images. The segmentation of retinal vessels is a challenging task due to the complex structure of the retinal vasculature and the presence of noise and artifacts in the images.

Several methods have been proposed for retinal vessel segmentation. Edge detection-based methods have shown promising results due to their ability to capture vessel edges accurately [2]. The edges represent the boundaries between different regions in the image, such as the boundaries between the retinal vessels and the background.

In this literature survey, we aim to investigate the performance of edge detection-based methods for retinal vessel segmentation. This survey will contribute to the development of more accurate and efficient tools for retinal vessel segmentation.

2 Methodology

Literature search was conducted in three well-known databases, which are PubMed, IEEEXplore, and Google Scholar during 2nd April 2023 to 27th April 2023. The keywords used are “retinal vessels segmentation”, “edge detection”, and “medical image analysis”. The inclusion criteria were studies that focused on edge detection-based methods for retinal blood vessel segmentation, published in English language. Studies that do not report detailed description on the method, or do not use retina fundus images were excluded. The data extraction process involved recording information such as the name of the author, publication year, journal or conference proceeding, edge detection method, performance measure, and databases used.

3 Results and Discussions

3.1 Related Works on Retinal Vessel Segmentations

Table 1 presents a few works on retinal vessel segmentations. This table shows that there are different approaches to segment the retinal blood vessels. These methods vary from traditional to deep-learning approaches.

Table 1 Approaches for retinal blood vessels segmentation

Authors	Summary	Approach
Hoover et al. [8]	Locate retinal blood vessels using a matched filter response	Traditional
Michal and Stewart [9]	Extract vessel centerlines from retinal images by using multiscale matched filters	Traditional
Quinn and Krishnan [10]	Segment retinal blood vessel using curvelet transform and morphological reconstruction	Traditional
Yin et al. [11]	Automatically segment and measure of blood vessels in fundus images using a probabilistic formulation	Traditional
Nguyen et al. [12]	Segment retinal blood vessel using multi-scale line detector	Traditional
Melinscak et al. [13]	Uses a deep neural network (DNN) approach for retinal vessel segmentation. The method consists of two stages: vessel classification and vessel delineation	Deep learning
Fu et al. [14]	Segment vessel using deep learning networks and fully connected conditional random fields (FC-CRF)	Deep learning
Chakraborty et al. [15]	Involves a pre-processing using a median filter to remove noise, gradient approximation using the Sobel operators, and post-processing to remove false edges and filling the gaps	Traditional
Hu et al. [16]	Segment retinal vessel using a multi-scale convolutional neural network (MCNN) with an improved cross-entropy loss function	Deep learning
Jiang et al. [17]	Segment retinal vessel by using a dilated multi-scale convolutional neural network (DCNN)	Deep learning
Orujov et al. [18]	Uses a new fuzzy logic-based edge detection algorithm for blood vessel detection in retinal images	Traditional
Ooi et al. [19]	Proposed a semi-automatic segmentation approach, which is based on Canny edge detector	Traditional
Chatterjee et al. [20]	Edge is detected using Sobel edge detector. A post-processing is applied to refine the edges	Traditional
Zhang et al. [21]	Uses Edge-Aware U-Net model for the retinal vessel segmentation	Deep learning

3.2 Commonly Available Retinal Fundus Datasets

Researchers have used different datasets for color retinal fundus images (Table 2). The images are acquired from different patients with different types of pathologies. The datasets are publicly available and have been used by many researchers for retinal vessel segmentation. The followings are the five commonly used datasets:

- Digital Retinal Images for Vessel Extraction (DRIVE) dataset [3]: This dataset contains 40 color fundus images for blood vessel segmentation.
- STructured Analysis of the REtina (STARE) dataset [4]: This database provides 400 raw fundus images. For blood vessel segmentation, this databasae provides 40 hand labeled images as the ground truth.
- High-Resolution Fundus (HRF) dataset [5]: It contains 45 color fundus images, of which 15 are from healthy persons, 15 images from glaucomatous patients, and the remaining 15 images from diabetic retinopathy patients.
- Child Heart and Health Study in England (CHASE) BD1 dataset [6]: This dataset is acquired from 14 children. Fundus images are captured from both eyes, producing a total of 28 retinal fundus images.
- Standard Diabetic Retinopathy Database (DIARETDB1) dataset [7]: This dataset contains 89 color fundus images.

These are just a few examples of the datasets used by the researchers for retinal vessel segmentation. There are many other datasets available as well.

3.3 Performance Evaluation for Retinal Vessel Segmentation

Table 2 shows the performance matrices used by researchers to evaluate the performance of the edge detection methods. As shown by Table 2, there is no standard evaluation. This finding agrees with our previous review [22].

4 Conclusion

In conclusion, edge detection techniques can be effective for retinal vessel segmentation. However, the choice of technique should depend on the specific application and the characteristics of the image data. The traditional methods such as Sobel operator and Canny edge detector are popular as they can give good results and are relatively simple to implement. On the other hand, as deep learning techniques continue to advance, they may offer even more accurate and efficient methods for retinal vessel segmentation. However, the development of deep learning methods will require access to large amounts of annotated data and sufficient computational resources. Overall, further research is needed to fully explore the potential of edge

Table 2 Datasets and performance evaluation measures used by researchers

Authors	Dataset	Performance Measures
Hoover et al. [8]	20 retinal fundus slides captured by a TopCon TRV-50 fundus camera	True positive rate (TPR) and false positive rate (FPR)
Michal and Stewart [9]	Uses STARE and DRIVE datasets	ROC curves, (1-Precision)-Recall plots, likelihood ratio, and Chernoff bound
Quinn and Krishnan [10]	Evaluated using real images from hospital	Peak signal-to-noise ratio (PSNR), mean squared error (MSE), and time elapsed
Yin et al. [11]	Uses STARE, DRIVE, REVIEW, and KPIS datasets	Vessel's width
Nguyen et al. [12]	Uses STARE and DRIVE datasets	Accuracy (ACC), mean absolute error (MAE) for vessel width measurement, Wilcoxon sign rank test
Melinscak et al. [13]	Uses DRIVE dataset	ACC, TPR, FPR, area under the receiver operating characteristic (ROC) curve
Fu et al. [14]	Uses STARE and DRIVE datasets	ACC, sensitivity (Sen)
Chakraborty et al. [15]	(No dataset used)	Pratt score, F-measure, false alarm count, ACC, sen, precision, true negative rate, TPR, recall, G-mean, specificity (Sp)
Hu et al. [16]	Uses STARE and DRIVE datasets	Sen, Sp, ACC, area under the ROC curve (AUC)
Jiang et al. [17]	Uses STARE, DRIVE and CHASE datasets	ROC curve, PR curves, ACC, Sen, Sp, F1 score
Orujov et al. [18]	Uses STARE, DRIVE and CHASE datasets	Sen, Sp, ACC, Dice coefficient, Jaccard similarity
Ooi et al. [19]	Uses STARE, DRIVE and CHASE datasets	Pratt's figure of merit (PFOM)
Chatterjee et al. [20]	Uses DRIVE dataset	ACC, Sp, Sen
Zhang et al. [21]	Uses STARE, DRIVE and CHASE datasets	Sen, Sp, ACC, F1 score, AUC, Matthew's correlation coefficient, intersection over union, processing time

detection and deep learning techniques for retinal fundus images, and to develop methods that can be applied effectively in clinical settings.

Acknowledgements This work is supported by the Ministry of Higher Education (MoHE), Malaysia, under the Fundamental Research Grant Scheme (FRGS), with grant number FRGS/1/2019/TK04/USM/02/1.

References

1. Aswini S, Suresh A, Priya S, Santhosh Krishna BV (2018) Retinal vessel segmentation using morphological top hat approach on diabetic retinopathy images. In: Proceedings of the 2018 fourth international conference on advances in electrical, electronics, information, communication and bio-informatics (AEEICB). IEEE, Chennai, pp 1–5
2. Li Q, Feng B, Xie L, Liang P, Zhang H, Wang T (2016) A cross-modality learning approach for vessel segmentation in retinal images. *IEEE Trans Med Imag* 35(1):109–118
3. DRIVE Homepage. <https://drive.grand-challenge.org>. Accessed 20 April 2023
4. Structured Analysis of the Retina. <https://cecas.clemson.edu/~ahoover/stare/>. Accessed 21 April 2023
5. High-Resolution Fundus (HRF) Image Database. <https://www5.cs.fau.de/research/data/fundus-images>. Accessed 20 April 2023
6. CHASE_DB1 retinal vessel reference dataset. <https://researchdata.kingston.ac.uk/96/>. Accessed 20 April 2023
7. DIARETDB1. <http://www2.it.lut.fi/project/imageret/diaretdb1/>. Accessed 10 April 2023
8. Hoover AD, Kouznetsova V, Goldbaum M (2000) Locating blood vessels in retinal images by piecewise threshold probing a matched filter response. *IEEE Trans Med Imag* 19(3):203–210
9. Michal S, Stewart CV (2006) Retinal vessel centerline extraction using multiscale matched filters, confidence and edge measures. *IEEE Trans Med Imag* 25(12):1531–1546
10. Quinn EAE, Krishnan KG (2013) Retinal blood vessel segmentation using curvelet transform and morphological reconstruction. In: Proceedings of the 2013 IEEE international conference on emerging trends in computing, communication and nanotechnology (ICECCN). IEEE, Tirunelveli, pp 570–575
11. Yin Y, Adel M, Bourennane S (2013) Automatic segmentation and measurement of vasculature in retinal fundus images using probabilistic formulation. *Comput Math Methods Med* 13:260410
12. Nguyen UTV, Bhuiyan A, Park LAF, Ramamohanarao K (2013) An effective retinal blood vessel segmentation method using multi-scale line detection. *Pattern Recogn* 46(3):703–715
13. Melinscak M, Prentasac P, Loncaric S (2015) Retinal vessel segmentation using deep neural networks. In: Proceedings of the 10th international conference on computer vision theory and applications (VISAPP 2015). SCITEPRESS. Berlin, pp 11–14
14. Fu H, Xu Y, Wong DWKW, Liu J (2016) Retinal vessel segmentation via deep learning network and fully-connected conditional random fields. In: Proceedings of the 2016 IEEE 13th international symposium on biomedical imaging (ISBI). IEEE, Prague, pp 698–701
15. Chakraborty S, Chatterjee S, Dey N, Ashour AS, Shi F (2017) Gradient approximation in retinal blood vessel segmentation. In: Proceedings of the 2017 4th IEEE Uttar Pradesh section international conference on electrical, computer and electronics (UPCON). IEEE, Mathura, pp 618–623
16. Hu K, Zhang Z, Niu X, Zhang Y, Cao C, Xiao F, Gao X (2018) Retinal vessel segmentation of color fundus images using multiscale convolutional neural network with an improved cross-entropy loss function. *Neurocomputing* 309:179–191
17. Jiang Y, Tan N, Peng T, Zhang H (2019) Retinal vessels segmentation based on dilated multi-scale convolutional neural network. *IEEE Access* 7:76342–76352
18. Orujov F, Maskeliunas R, Damasevicius R, Wei W (2020) Fuzzy based image edge detection algorithm for blood vessel detection in retinal images. *Appl Soft Comput* 94:106452
19. Ooi AZH, Embong Z, Hamid AIA, Zainon R, Wang SL, Ng TF, Hamzah RA, Teoh SS, Ibrahim H (2021) Interactive blood vessel segmentation from retinal fundus image based on Canny edge detector. *Sensors* 21(19):6380
20. Chatterjee S, Suman A, Gaurav R, Banerjee S, Singh AK, Ghosh BK, Mandal RK, Biswas M, Maji D (2021) Retinal blood vessel segmentation using edge detection method. *J Phys Confer Ser* 1717:012008
21. Zhang Y, Fang J, Chen Y, Jia L (2022) Edge-aware U-net with gated convolution for retinal vessel segmentation. *Biomed Sig Process Control* 73:103472

22. Tariq N, Hamzah RA, Ng TF, Wang SL, Ibrahim H (2021) Quality assessment methods to evaluate the performance of edge detection algorithms for digital image: a systematic literature review. *IEEE Access* 9:87763–87776

Near Infrared Remote Sensing of Vegetation Encroachment at Power Transmission Right-of-Way



Pei Yu Lim, David Bong, Kung Chuang Ting, Florence Francis Lothai, Annie Joseph, and Tengku Mohd Afendi Zulcaffle

Abstract The event of electricity outage could cause huge financial losses for the industry and inconvenience to the consumers. Vegetation encroachment at the power transmission right-of-way is one of the main causes. Transmission line fault could occur when a tree falls into the vicinity of power transmission line. Conventional inspection method such as ground inspection is the simplest approach to counter vegetation encroachment. However, technical personnel is required to travel on site to perform the inspection manually. This process is often time consuming and prone to human error. Airborne Light Detection and Ranging (LiDAR) system and satellite imagery are remote sensing approaches to inspect the power transmission right-of-way. These approaches could reduce reliance on physical site inspection and remove human error. However, large dataset needs to be processed and specialist equipment is needed for this method which also increases the overall cost. In this research, a simple yet cost effective method is used to detect vegetation encroachment by using near aerial infrared (NIR) image processing approach. The process is divided into two parts. First, detect the inconspicuous power transmission line by utilizing Radon Transform (RT) in vertical derivative image and detect the peaks of the Radon Transform. Next, detect the vegetation encroachment in the clearance zone by using green normalized difference vegetation index (GNDVI) algorithm to differentiate between trees and glassy plains. Preliminary experiment results show a satisfactory performance in detecting vegetation encroachment at the power transmission right-of-way.

Keywords Vegetation encroachment · Power transmission right-of-way · Near infrared (NIR)

P. Y. Lim · D. Bong (✉) · A. Joseph · T. M. A. Zulcaffle
Faculty of Engineering, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Malaysia
e-mail: bldavid@unimas.my

K. C. Ting · F. F. Lothai
Faculty of Computing and Software Engineering, i-CATS University College, 93350 Kuching, Malaysia

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
N. S. Ahmad et al. (eds.), *Proceedings of the 12th International Conference on Robotics, Vision, Signal Processing and Power Applications*, Lecture Notes in Electrical Engineering 1123, https://doi.org/10.1007/978-981-99-9005-4_64

507

1 Introduction

Highly elevated structures are used in the transmission of high voltage electrical energy generated from the generation plant to the consumer area. The clearance for overhead line with aluminium or copper conductors must be maintained as it is predisposed to flashover. This causes unnecessary power interruptions for the consumer and revenue losses to the utility companies. The consequence of a blackout can be catastrophic when it affects a large businesses area that relies on electricity for its operations [1].

Transmission network should be monitored to guard against any failure of continuous electricity supply. The preservation of electrical distribution networks should be examined regularly as the highest failure is due to tree encroachment [2, 3]. Better surveillance information can be provided by using online or remote monitoring as any fault could be located precisely and the process of power supply recovery could be facilitated. This can help to decrease the interruption time. Trees could be trimmed, and specified clearance of grasslands and bushes at the power transmission right-of-way should be determined in order to keep the vegetation from affecting the conductors.

Since the high voltage transmission lines are cascade connected, vegetation encroachment at any part of the right-of-way will potentially affect the entire cascading network [4]. Therefore, for commercial and legal reasons, electric utility companies have to regularly inspect their power transmission lines. One of the conventional methods involves site surveys or ground inspection, where a team is responsible for visual inspection of power transmission right-of-way by foot patrols or using vehicles [5]. Any vegetation that encroaches into the high voltage right-of-way are identified and trimmed after the visual inspection. Unfortunately, this approach is costly because a high number of technical personnel are needed. It is also time consuming because it requires long hours to travel to the actual site. Moreover, it is prone to judgmental errors during the visual inspection.

Another method is by using airborne Light Detection And Ranging (LiDAR) scanning system [1]. LiDAR is an active remote sensing technique. This technique can find the distant target range or further information by measuring the scattered light property. Vegetation encroachment of power lines could be detected by satellite imagery too [6, 7]. However, costly specialist equipment is required for this kind of remote sensing.

In this research, a low-cost solution is designed to monitor the vegetation encroachment to the power transmission right-of-way by using near infrared (NIR) imaging technique. A DJI Phantom 3 with Blue-Green-NIR advanced camera drone is used to capture image combination of Blue, Green and NIR channels. Image processing technique is then used to identify vegetation encroachment at the power transmission right-of-way. NIR wavelength between 680 and 800 nm is used in this technique.

2 Methodology

2.1 Region of Interest

The red channel of the advanced NIR camera could record NIR image/video. The camera has the feature to differentiate vegetation from non-vegetation remotely by capturing colour NIR-Green-Blue (NIR-G-B) image of wavelength between 680 and 800 nm. It is built with 20 mm focal length lens. Figure 1 shows the process of extracting region-of-interest (ROI) from an NIR-G-B image.

NIR-G-B image is different from conventional RGB image where the red channel is replaced by NIR channel. Figure 2a shows a sample NIR-G-B image where the brownish areas indicate the presence of vegetation such as trees or grass plains.

A vertical mask $w(i, j)$ [8] is used for linear feature extraction of the image. The image itself has to be aligned to achieve 90° vertical orientation of the transmission line. In drone mission planning, the direction of flight is planned along the power lines, which could result in the alignment of the power transmission line being imaged as intended. Through convolution process, the output grayscale pixel $G(x, y)$ can be obtained as follows:

$$G(x, y) = \sum_{i=0}^n \sum_{j=0}^n f(x + i, y + j) * w(i, j) \quad (1)$$

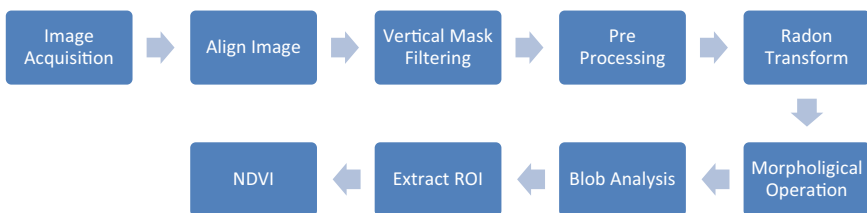


Fig. 1 Extraction of ROI from the NIR image

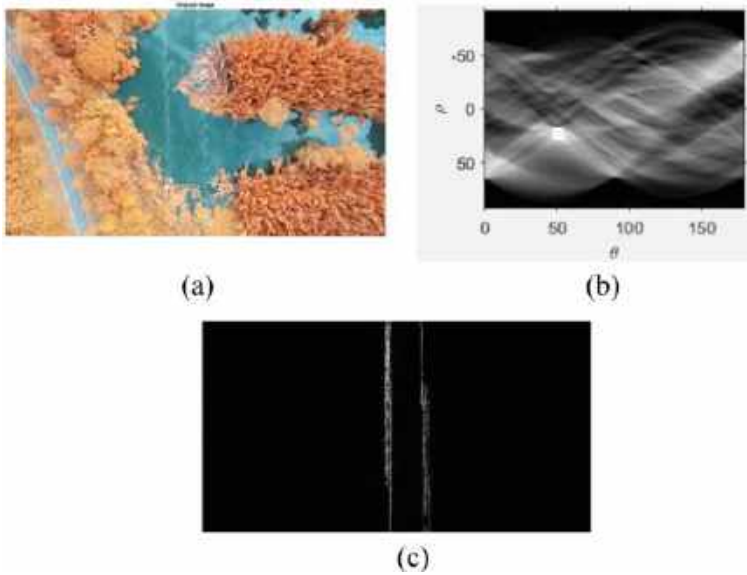


Fig. 2 a A sample of NIR-G-B image, b the result of Radon Transform, c output image after performing blob analysis

where $f(x, y)$ is the grayscale value of input image and n is the total number of pixels. Radon Transform is then applied to the output image. From the Radon Transform, peak values could be obtained which indicate the positions of the lines from the image. Figure 2b shows the result of Radon Transform.

Thresholding is then applied to eliminate image noise. The value is set to zero if it is less than the threshold, and set to one if more than the threshold. Inverse Radon Transform is applied to obtain the line image. Morphological operation and blob analysis are used to trace the boundary of the power transmission line.

A region of interest (ROI) is identified after the blob analysis to select the boundaries of the clearance zone where the power transmission line is detected.

2.2 Green Normalized Difference Vegetation Index (GNDVI)

Figure 3 shows the Green Normalized Difference Vegetation Index (GNDVI) algorithm to extract and process the NIR and green channels of an image for the purpose of quantifying vegetation. Figure 4 shows the NIR and green channels of a sample image.

GNDVI quantifies vegetation by using

$$GNDVI = \frac{NIR - G}{NIR + G} \quad (2)$$

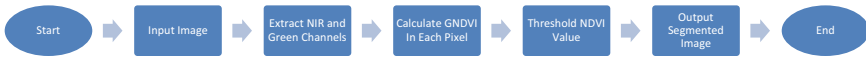


Fig. 3 The GNDVI algorithm

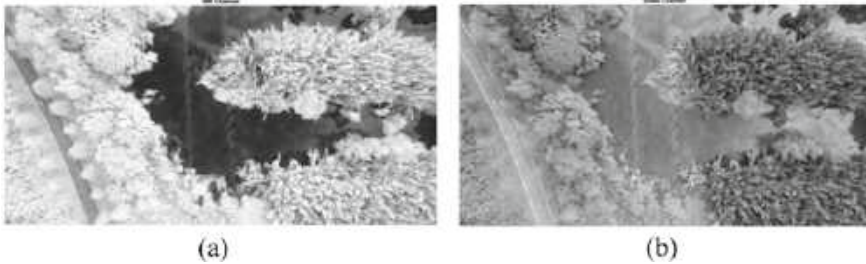


Fig. 4 **a** NIR channel, **b** green channel

Values that range from 0 to 0.1 resemble sterile areas of rock, sand, or snow. Values in between 0.2 and 0.4 indicates a low positive value that signify shrub and grassland, while high values that are in between 0.5 and 0.9 or close to 1 signify temperate and tropical rainforests [9]. Therefore, the value of GNDVI can be used to classify trees and grass plains. This is important to avoid false positive detection of grass plains as potential threat.

3 Results and Discussion

Figure 5a shows the aligned image of power transmission lines by using the algorithm from Fig. 1. Image ROI is then determined (red box in Fig. 5c). Clearance of the power transmission right-of-way is determined from the image ROI. Figure 5d shows the final result with the blue lines indicate the detected power transmission line and the red segmented areas represent the vegetation encroachment which located inside the clearance zone.

The overlapped image of ground truth with GNDVI output image is as shown in Fig. 5e. The green segmented areas in the image represent vegetation encroachment which is not detected (False Negative). The purple segmented areas indicate detected vegetation encroachment which is not present in the ground truth image (False Positive). The white segmented areas represent True Positive and black segmented areas represent True Negative. From the preliminary experiment, the accuracy of the proposed algorithm for a single location is about 91%.

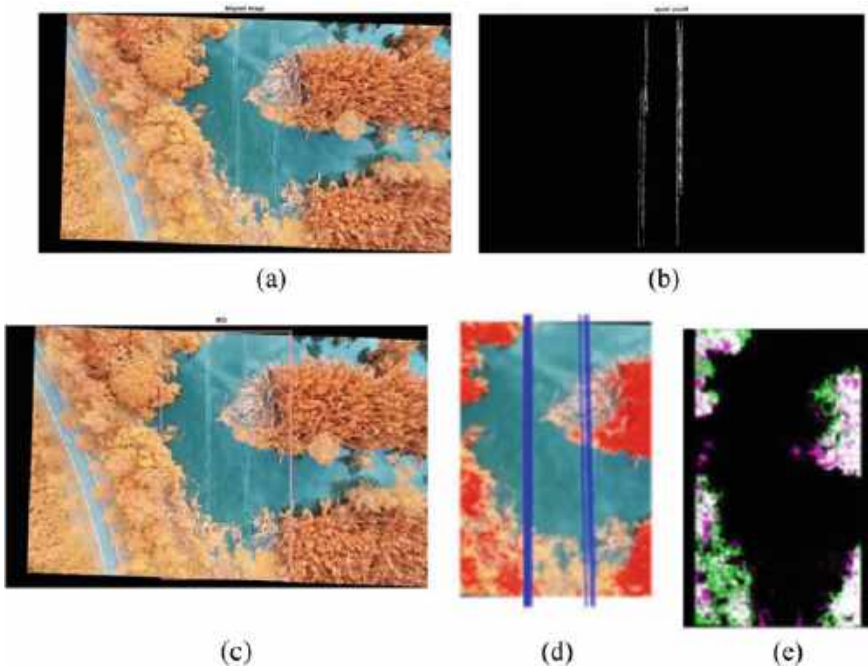


Fig. 5 a Aligned image, b boundaries of the power transmission line, c image ROI, d final image, e overlapped image

4 Conclusion

This research proposed detection of vegetation encroachment at power transmission right-of-way by using low cost NIR image. Inconspicuous power transmission line is detected by using Radon Transform (RT) in vertical derivative image and the vegetation encroachment is detected by using GNDVI. Preliminary experiment showed satisfactory result in detecting vegetation encroachment and differentiating types of vegetation. For future work, multisensorial UAV systems could be used, which combine NIR-G-B camera and spectrometry with laser scanning.

Acknowledgements Kementerian Pengajian Tinggi Malaysia, Fundamental Research Grant Scheme (FRGS), FRGS/1/2020/TK0/UNIMAS/02/14.

References

1. Ahmad J, Malik AS, Xia L, Ashikin N (2013) Vegetation encroachment monitoring for transmission lines right-of-ways: a survey. *Electr Power Syst Res* 95:339–352
2. Louit D, Pascual R, Banjevic D (2009) Electrical power and energy systems optimal interval for major maintenance actions in electricity distribution networks. *Int J Electr Power Energy Syst* 31(7–8):396–401
3. Sittithumwat A, Soudi F, Tomsovic K (2004) Optimal allocation of distribution maintenance resources with limited information. *Electr Power Syst Res* 68(3):208–220
4. Zheng C, Sun D (2006) Recent applications of image texture for evaluation of food qualities—a review. *Trends Food Sci Technol* 17(3):113–128
5. Luque-Vega LF, Castillo-Toledo B, Loukianov A, Gonzalez-Jimenez LE (2014) Power line inspection via an unmanned aerial system based on the quadrotor helicopter. In: *MELECON 2014—2014 17th IEEE Mediterranean electrotechnical conference, Beirut, Lebanon*, pp 393–397
6. Gazzea M, Pacevicius M, Dammann DO, Sapronova A, Lunde TM, Arghandeh R (2022) Automated power lines vegetation monitoring using high-resolution satellite imagery. *IEEE Trans Power Delivery* 37(1):308–316
7. Haroun FME, Deros SNM, Din NM (2021) Detection and monitoring of power line corridor from satellite imagery using RetinaNet and K-Mean clustering. *IEEE Access* 9:116720–116730
8. Chan TS, Yip R (1996) Line detection algorithm. In: *Proceedings of international conference on pattern recognition, Vienna, Austria*, pp 126–130
9. Baltsavias EP (1999) A comparison between photogrammetry and laser scanning. *ISPRS J Photogramm Remote Sens* 54(2–3):83–94

Evaluation of Three Variants of LBP for Finger Creases Classification



Nur Azma Afiqah Salihin, Imran Riaz , and Ahmad Nazri Ali 

Abstract Biometric technology improves security and authentication, especially in sensitive systems like attendance systems. The common traits for biometrics are typically from the iris, fingerprints, face, etc. Another trait that possibly comes from the finger creases and the research work evaluating the finger crease's capability for biometric classification is proposed in this paper. Various local binary patterns (LBP) are employed to extract the features, and the classification performance is evaluated using Support Vector Machines (SVM) on two different kernels. From the evaluation, an accuracy of up to 94% with a percentage of FAR and FRR less than 3 is observed for all the proposed LBP methods.

Keywords Finger based biometric · Local binary pattern · Support Vector Machines

1 Introduction

Biometric identity systems provide better anti-spoofing capabilities than older verification methods like ID cards. An example of biometrics is using a person's unique physical features and behavioral patterns for identification [1]. Due to the fact that the conventional identification method has flaws that third parties may exploit, biometrics identification has brought unquestionably extensive use in user-verification frameworks with an increased emphasis on data security. Many different biometric features

N. A. A. Salihin · I. Riaz · A. N. Ali (✉)

School of Electrical and Electronic Engineering, University Science Malaysia, Engineering Campus, 14300 Nibong Tebal, Penang, Malaysia
e-mail: nazriali@usm.my

I. Riaz

e-mail: imran.ee@must.edu.pk

I. Riaz

Mirpur Institute of Technology, Mirpur University of Science and Technology, 10250 Mirpur, Azad Kashmir, Pakistan

Fig. 1 An example of a finger image and the region of interest



have been widely used as a recognition system due to their distinctiveness and precision. Iris, voice, fingerprint, hand geometry, and facial features are among examples [2]. Hand-based biometric identification is a rapidly expanding study topic due to its cheap cost in gathering data, its reliability in identifying persons, and its degree of acceptability by users compared to the other biometric identification systems discussed [3].

One trait that can be used for hand-based biometrics is the creases of the fingers, where the middle phalanx is more distinguishable and traceable compared to the distal and proximal phalanx. Fine wrinkles and textural variation are particularly abundant in the middle phalanx of the fingers [4]. It is essential to understand the fundamental anatomy of the middle phalanx of the finger and the fingerprint to identify them correctly. Phalanges are singular “phalanx.” The middle phalanges (hand) are a set of bones in the fingers. Each intermediate phalanx connects to the same side’s proximal and distal phalanges. Each proximal phalanx is linked to the metacarpal bone in the palm, while each distal phalanx is the fingertip and fingernail home. Except for the thumb, which only contains proximal and distal phalanges, the middle phalanx has two joints and allows the finger to flex in two directions. The finger middle phalanx is flexible and contains multiple wrinkles or creases. This region has several apparent wrinkles, also known as creases, as depicted in Fig. 1 in the red rectangle box. This project will be divided into three major stages: image preprocessing, feature extraction, and classification.

The remainder of this paper is subdivided into the following sections: Sect. 2 summarizes the related hand-based biometrics classification. Section 3 describes the methodology of the proposed work. Section 4 discusses the experimental approach and results, and finally, Sect. 5 is for the conclusion.

2 Related Works

Fingerprint, hand geometry, and finger vein patterns are all examples of hand-based biometric traits used in the working system. Among all hand-based biometric qualities, the fingerprint is the most ancient and widely utilized due to its distinct properties and pattern. The uniqueness of the fingerprint pattern, known as friction ridges used

to verify and authorize users. However, since this fingerprinting technique is associated with a criminal identification, it may introduce inaccuracies due to aged, dry, and filthy fingers [5]. Additionally, fingerprint systems use a significant amount of computing resources, which is one of their drawbacks when used for identification [6].

Besides that, hand geometry is a straightforward approach among the hand-based biometric features. This low-cost technology offers benefits independent of environmental conditions such as dry weather or individual abnormalities such as dry skin. It might provide high precision while having no adverse influence on that. However, this hand geometry technique may result in injuries or scratches to the hands and the wearing of jewelry such as rings and bracelets, causing high FAR and FRR [7]. The immense physical size of the hand geometry-based device may hinder its usability in specific applications too.

Furthermore, the finger vein is a more recent biometric technology that records the unique vein patterns within the finger by shining infrared light near it and recording the effects using a sensor such as a charge-coupled device (CCD). On the other hand, the image acquisition step is the most difficult in uncontrolled situations because of the heat-emitting surfaces. Finger vein biometric systems, on the other hand, have advantages over different types of biometric identifiers in terms of accuracy, contactless ness, and hygienic operation [7].

3 Methodology

This research aims to determine the feasibility of using the middle phalanx finger creases via various local binary patterns. This project is centered on software using MATLAB R2021b for image processing, feature extraction, and classification. Three Local Binary Pattern (LBP) variants named Conventional LBP, Local Directional Pattern (LDP), and Local Optimal Oriented Pattern (LOOP) are utilized to extract the features [8]. At the same time, the SVM classifier is used to categorize the combinations. The performance of finger creases is compared using two different SVM kernels.

Before furthering the extraction process, several main pre-processing stages are performed: ROI cropping and grayscale conversion. All images are manually cropped to obtain the region of interest due to the wide finger size range. The cropped images are then set into the fixed dimension of 300×300 pixels to cover most of the abundant information. The original image is in RGB format, then a stage for converting the image into grayscale is also performed.

Each of the pre-processed images will go through Conventional LBP, LDP, and LOOP LBP algorithms to extract the features. Conventional LBP is defined in 3×3 rectangular neighborhoods using Eq. (1).

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p \quad (1)$$

where here (x_c, y_c) , R , g_c and g_p denote the coordinates, radius, gray value of the central pixel, and gray pixel value of its neighbors, respectively. For LDP, the algorithm is defined using Eqs. (2) and (3). An LDP operator calculates edge response values at each pixel location in eight directions and generates a binary code based on relative strength magnitude.

$$LDP_k(x_c, y_c) = \sum_{n=0}^7 s(m_n - m_k) \cdot 2^n \quad (2)$$

where

$$s(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

The primary formulation of LOOP LBP is shown in Eqs. (4) and (5), in which the algorithm will perform rotation invariance into the formulations of conventional LBP and LDP.

$$LOOP(x_c, y_c) = \sum_{n=0}^7 s(i_n - i_c) \cdot 2^{w_n} \quad (4)$$

where

$$s(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

After the images are extracted, the generated features vector is processed for classification. This paper used the Support Vector Machine (SVM) method with two primary kernels: Linear and Polynomial. For an initial evaluation, we used 400 samples from 20 subjects, where each of the subjects provided 20 images which are 10 images for training and the remaining 10 samples for testing.

SVM classification is chosen in this research due to its high texture analysis and recognition systems performance. The remarkable accuracy of the SVM classifier enables it to make accurate decisions about whether the incoming data matches any of the database's datasets. High accuracy is predicted as an output from a model that utilized both train and test data as input. The data is compared subject by subject and then repeated from 1 to 20 for each subject.

4 Experimental Results and Discussion

To justify the effectiveness of the proposed methods, pre-processing of the images and verification to determine the system’s performance and accuracy system are performed. The images are resized into 300×300 pixels at the pre-processing stage, where an example of the result is shown in Fig. 2a. After that, the image is converted to grayscale, as shown in Fig. 2b.

Figure 3a–d are an example of the extracted images of finger creases on the middle phalanx from the original gray image and three LBP versions for conventional LBP, LDP, and LOOP LBP, respectively. The histogram of each LBP variant image is displayed on the right, illustrating the output structure of the image over a wide range from 0 to 255. From the figures, each uniform pattern is assigned to a unique bin, while non-uniform patterns are assigned to a single bin.

We used the generated histogram bin as the features vector for the training dan testing dataset. The training data is trained into the model first, followed by the testing data using the model’s prediction. False Acceptance Rate (FAR) and False Rejection Rate (FRR) are used to evaluate the performance of the algorithms. Table 1 shows



Fig. 2 An example of an image, a original and resized image, b grayscale conversion

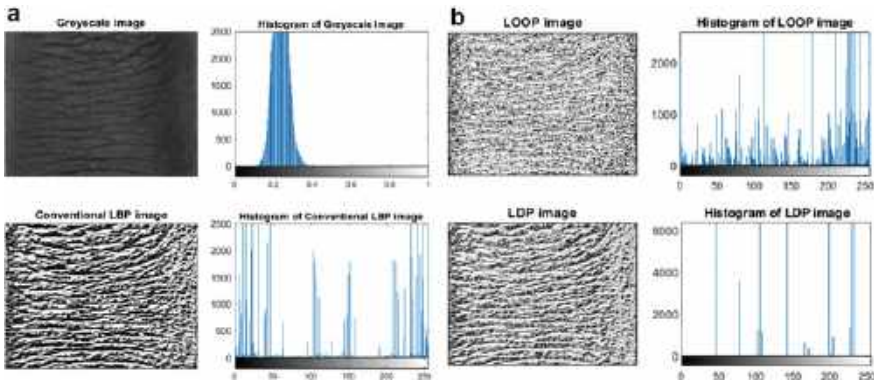


Fig. 3 LBP image with histogram, a original image, b conventional LBP, c LDP LBP, and d LOOP LBP

Table 1 Overall FAR and FRR for linear kernel and polynomial kernel SVM for 3 LBP variants

Testing sample		FAR (%)	FRR (%)
Linear kernel	LBP	2.0	2.0
	LDP	1.0	1.0
	LOOP	0.5	0.5
Polynomial kernel	LBP	2.375	2.375
	LDP	1.375	1.375
	LOOP	0.875	0.875

Table 2 Overall comparison for three LBP variants

Evaluation metrics	Linear kernel			Polynomial kernel		
	LBP	LDP	LOOP	LBP	LDP	LOOP
Accuracy (%)	95.56	97.78	98.89	94.44	96.94	98.06
FAR (%)	2.0	1.0	0.5	2.5	1.375	0.875
FRR (%)	2.0	1.0	0.5	2.5	1.375	0.875

linear and polynomial kernels' overall FAR and FRR performance for three different LBP. Whereas, Table 2 shows the performance in terms of accuracy for both kernels and three LBP variants. From both tables, it is found that the LOOP LBP is able to achieve higher performance, followed by LDP and LBP.

5 Conclusion

In conclusion, a study about the performance of three different variants of LBP for finger creases classification is reported in this paper. With the Support Vector Machines classifier, the experimental results have shown a significant accuracy score with low FAR and FRR. The LOOP variant has achieved higher accuracy than the other two variants. It believes that using particular LBP algorithms can distinguish between individuals, and the performance can be further improved if more samples are available for each subject and the number of subjects.

Acknowledgements Acknowledgment to Ministry of Higher Education Malaysia for the Fundamental Research Grant Scheme with Project Code: FRGS/1/2021/ICT02/USM/02/1 for the financial support of this research. The images used in this study are acquired through ethical approval protocol with the study protocol code USM/JEPeM/21100657.

References

1. Jain AK, Feng J, Nandakumar K (2010) Fingerprint matching. *Computer* 43(2):36–44
2. Okereafor K, Ekong I, Markson IO, Enwere K (2020) Fingerprint biometric system hygiene and the risk of COVID-19 transmission. *JMIR Biomed Eng* 5(1):e19623
3. Barra S, De Marsico M, Nappi M, Narducci F, Riccio D (2019) A hand-based biometric system in visible light for mobile environments. *Inf Sci* 479:472–485
4. Roopa JJ, Veluchamy S (2018) Person validation system using hand based biometric modalities. In: 2018 IEEE international conference on communication, computing and internet of things (IC3IoT), pp 42–47
5. Chetana K, Aditya A (2021) Iris-fingerprint multimodal biometric system based on optimal feature level fusion model. *AIMS Electron Electr Eng* 5(4):229–250. <https://doi.org/10.3934/electreng.2021013>
6. Mahmud CAM, Masudul HI (2022) Contactless fingerprint recognition using deep learning—a systematic review. *J Cybersecur Priv* 2:714–730. <https://doi.org/10.3390/jcp2030036>
7. Sabhanayagam T, Prasanna Venkatesan V, Senthamaraikannan K (2018) A comprehensive survey on various biometric systems. *Int J Appl Eng* 13(5):2276–2297
8. Chakraborti T, McCane B, Mills S, Pal U (2018) LOOP descriptor: local optimal-oriented pattern. *IEEE Signal Process Lett* 25(5):635–639. <https://doi.org/10.1109/lsp.2018.2817176>

R-Peaks and Wavelet-Based Feature Extraction on K-Nearest Neighbor for ECG Arrhythmia Classification



A. M. Khairuddin , K. N. F. Ku Azir , and C. B. M. Rashidi 

Abstract The aim of this research is to classify 17 types of arrhythmias by applying the algorithm developed from combining the morphological and the wavelet-based statistical features. The proposed arrhythmia classification algorithm consists of four stages: pre-processing, detection of R-peaks, feature extraction, and classification. Seven morphological features (MF) that were retrieved from the R-peak locations. Following this, another nine wavelet-based statistical features (SF) were gathered by decomposing wavelets in level 4 from the Daubechies 1 wavelet (Db1). These 16 features are then applied to the k-nearest neighbor (k-NN) algorithm. The accuracy (ACC) of the suggested classification algorithm was assessed by using the MIT-BIH arrhythmia benchmark database (MIT-BIHADB). The experimental results of this work attained an average accuracy (ACC) of 99.00%.

Keywords Arrhythmia · Wavelet · Electrocardiography · Classification · Algorithm

1 Introduction

The healthy heart rate for adults normally ranges from 60 to 100 beats per minute. Nonetheless, when the heart beats too rapidly, slowly, or inconsistently, these abnormal heartbeats can lead to an interference to the electrical signals. This caused

A. M. Khairuddin · K. N. F. Ku Azir (✉) · C. B. M. Rashidi
Faculty of Electronic Engineering and Technology (FKTEN), Universiti Malaysia Perlis (UniMAP), 02600 Arau, Perlis, Malaysia
e-mail: fazira@unimap.edu.my

A. M. Khairuddin · K. N. F. Ku Azir
Centre of Excellence for Advanced Computing (AdvComp), Universiti Malaysia Perlis (UniMAP), 02600 Arau, Perlis, Malaysia

C. B. M. Rashidi
Centre of Excellence for Advanced Communication Engineering (ACE), Universiti Malaysia Perlis (UniMAP), 02600 Arau, Perlis, Malaysia

the heart to undergo systole state which resulted in arrhythmia [1]. While all arrhythmias suggest some forms of disorder of the heartbeats, some irregular heartbeats are lethal and dangerous that they can reduce cardiac output as well as may cause sudden death. Fortunately, arrhythmias can be detected through monitoring the symptoms that they produced. Monitoring of the arrhythmia has often been done by utilizing the electrocardiogram (ECG) [2]. The typical ECG test records the electrically generated signals of the heart. The test records of the electrical activity of the heart are used to identify potentially fatal arrhythmias.

The review of the literature reveals that there are many different types of arrhythmias. The types of arrhythmias are classified according to where they originated in the heart (atria or ventricles) as well as by how they influence the speed of the heart rate. As an example, the earlier study by [3] identified five general types of arrhythmia. However, the other study by [4] uncovered 15 general types of arrhythmias. The 15 arrhythmias comprised: (1) ventricular tachycardia (VTC); (2) left bundle branch block (LBBB); (3) atrial fibrillation (AFIB); (4) atrial premature beat (APB); (5) idioventricular rhythm (IR); (6) premature ventricular contraction (PVC); (7) pre-excitation (PE); (8) AV block: 2nd degree (AVB2); (9) right bundle branch block (RBBB); (10) atrial flutter (AFL); (11) supraventricular tachyarrhythmia (SVT); (12) ventricular bigeminy (VB); (13) ventricular flutter (VFL); (14) ventricular fusion beat (VFB); and (15) ventricular trigeminy (VT).

It is essential to identify arrhythmia early and classify them precisely to effectively avert arrhythmias. Prior studies have proposed arrhythmia classification algorithms to accurately categorize arrhythmia [4–6]. However, the evaluation of past studies seems to suggest limitation in terms of the classification of numerous types of arrhythmias. For instance, although there are at least 17 different types of arrhythmias, past studies have mainly focused on classifying only a few selected classes of arrhythmias. In view of this, there is a need to develop an algorithm which can classify the 17 types of arrhythmias. In an attempt to develop the algorithm, this study used the k-nearest neighbor (k-NN) algorithm with the combination of the morphological features of the R-peaks and the wavelet-based statistical features.

2 Methodology

In this work, the classification algorithm was developed in four stages: (1) signal denoising; (2) detection of R-peaks; (3) feature extraction; and (4) arrhythmia classification. The four stages of the algorithm created in this work are shown in Fig. 1.

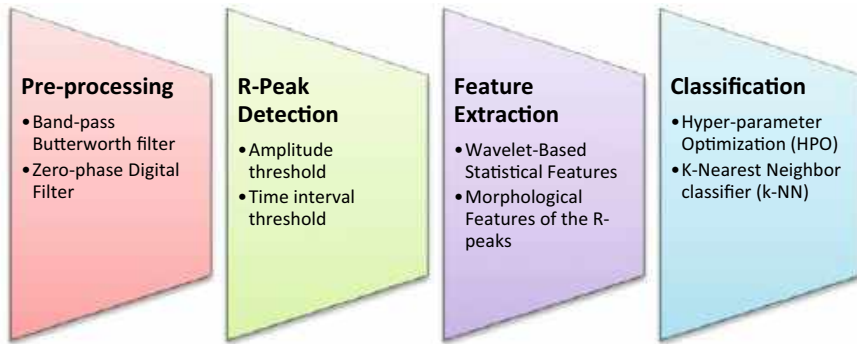


Fig. 1 Four stages of the proposed algorithm

3 Experimental Setup

The development and testing of the algorithm were carried out on a desktop with the specifications of AMD Ryzen 5 5600X CPU, 16 GB of RAM, 64-bit Windows 10 Home Basic. This desktop was installed with MATLAB software (version 2018a).

The original ECG signals from the MIT-BIHADB were gathered in order to implement and evaluate the algorithm's performance. The database contained 48 ECG records (30 min) that were sampled at 360 Hz. Both limb lead V1 (MLV1) and lead II (MLLII) and were both present and modified in each record. In this case, the MLLII was strictly utilized for algorithm development and testing. The ECG signals were chosen at random from the 10 s ECG segments.

The 17 ECG heartbeats classes in the MIT-BIH arrhythmia database (MIT-BIHADB) included the following: (1) fusion of ventricular and normal beat (FVN); (2) pacemaker rhythm (PR); (3) ventricular bigeminy (VB); (4) ventricular tachycardia (VTC); (5) right bundle branch block (RBBB); (6) AV block: 2nd degree (AVB2); (7) atrial premature beat (APB); (8) pre-excitation (PE); (9) left bundle branch block beat (LBBBB); (10) idioventricular rhythm (IR); (11) atrial flutter (AFL); (12) ventricular trigeminy (VT); (13) premature ventricular contraction (PVC); (14) normal sinus rhythm (NSR); (15) ventricular flutter (VFL); (16) atrial fibrillation (AFIB); and (17) supraventricular tachyarrhythmia (SVT). The algorithm suggested in this paper was created to classify these 17 different types of ECG heartbeats.

This study's findings were assessed by using four assessment metrics. The four assessment metrics comprised of sensitivity (SEN), positive predictivity (+P), accuracy (ACC), and mean squared error (MSE). The mean squared error (MSE) was used to evaluate the performance of the filter in denoising the ECG signals. The SEN, +P, and ACC were used to evaluate the performance of the algorithm in identifying the R-peaks. The ACC metric was used to evaluate the performance of the algorithm in classifying the 17 arrhythmia classes.

4 Results and Discussions

The R-peaks of these 17 ECG heartbeats were detected by adjusting the thresholding parameters for the identification of the R-peaks that involved the *amp_t* and the *time_t*. The default values for these thresholding parameters were set to 0.20 and 30 respectively. These parameters were adjusted with different values. Table 1 indicates record 102 and its thresholding parameters. The performance of R-peak detection for record 102 was also evaluated in terms of SEN, +P, and ACC.

The results of the accuracy of the algorithm by utilizing the k-nearest neighbor (k-NN) classifier are evaluated with three different types of features and summarized in Table 2. The findings as shown in Table 2 denote that the combination of morphological and statistical features that have 1 number of neighbors, distance metric of Euclidean, and distance weight of equal able to obtain the highest accuracy (ACC) of 99.00%. However, the morphological feature which also has 1 number of neighbors (NON), distance metric (DM) of Euclidean (E), and distance weight (DW) of equal (Eq), obtained the lowest accuracy (ACC) of 80.10%.

Table 1 The performance of R-peak detection (record 102)

<i>T_{amp}</i>		0.10	0.15	0.20	0.25	0.30
SEN	20	100.00	100.00	97.12	85.19	73.62
+P		95.80	93.86	93.04	93.81	96.29
ACC		95.80	93.86	90.42	80.65	71.59
SEN	25	100.00	100.00	97.12	85.19	73.62
+P		97.68	97.24	95.03	93.48	96.29
ACC		97.68	97.24	92.43	80.41	71.59
SEN	30	100.00	100.00	100.00	85.19	73.62
+P		97.55	97.50	100.00	96.83	97.40
ACC		97.55	97.50	100.00	82.87	72.20

The bolded values represents the best statistically significant results

Table 2 Classification accuracy obtained by different type of feature

Type of feature	NON	DM	DW	Accuracy (%)
Morphological	1	E	Eq	80.10
	10	E	Eq	71.70
Statistical	1	E	Eq	96.90
	10	E	Eq	88.80
Morphological + statistical	1	E	Eq	99.00
	10	E	Eq	93.70

The bolded values represents the best statistically significant results

The statistical feature (SF) that possesses number of neighbors (NON) of 1, distance metric (DM) of Euclidean, and distance weight (DW) of equal, obtained higher accuracy (ACC) of 96.90% as compared to the morphological feature (MF). These results appeared to indicate that the combination of features can increase the precision of the classification. Furthermore, it seemed that the number of neighbors of 1 provided the best classification accuracy in comparison to the number of neighbors of 10 and 100. The results of Table 2 indicate that the accuracy (ACC) of the algorithm can be influenced by the type of feature, number of neighbors, distance metric and distance weight.

The classification performance of the algorithm was determined by evaluating the accuracy of each of the 17 arrhythmia classes. There are 14 specific categories of arrhythmias obtained ACC of 100.00%. The 14 categories included NSR, AFL, AFIB, SVT, PE, PVC, VB, VT, VTC, IR, VFB, RBBB, AVB2 and PR. However, three categories of arrhythmias obtained accuracy (ACC) of less than 100.00%. These three classes of arrhythmias included APB, VFL, and LBBBB. The lowest accuracy (ACC) was attained by the VFL class, which was 85.70%. Overall, the algorithm used in this study achieved an average ACC of 99.00%.

This study's algorithm was also assessed against prior published ECG arrhythmia classification algorithms with their respective number of features, features extracted, and classifiers used. Table 3 provides the comparison of the performance of the algorithms.

Table 3 Comparison of the performance of the algorithm developed in this work and the prior documented algorithms

Literature	No. of classes	Features extracted	Classifier	ACC (%)
[7]	5	ICA, PCA, linear features and non-linear features	ANN	98.91
[8]	4	DWT, temporal features and MF	SVM	98.39
[9]	5	EMD, temporal features, MF and SF	RBF-NN	99.89
[10]	16	Permutation entropy, energy, RR-interval, STD and KR	SVM	99.21
[11]	5	PCA	SVM	97.77
[12]	5	CNN and LSTM	DL	98.10
[13]	2	Temporal features	DL	99.68
[14]	5	Coupled-convolution layer structure	CNN	99.43
[15]	15	Parametric and VP	k-NN	97.70
This study	17	MF and SF	k-NN	99.00

ICA independent component analysis; *PCA* principal component analysis; *DWT* discrete wavelet transform; *MF* morphological features; *SF* statistical features; *VFF* VF-filter leakage measure; *STD* standard deviation; *KR* kurtosis; *EMD* empirical mode decomposition; *RBF-NN* radial basis function neural network; *LSTM* long short-term memory; *WF* wave features; *GF* Gabor features; *VP* visual pattern; *CNN* convolutional neural network; *DBN* deep belief network

The results of the comparison presented in Table 3 appear to indicate that the algorithm developed in this study achieved higher accuracy (ACC) than those in the past studies. The results in Table 3 show that the algorithm developed in this study attained higher classification accuracy (99.00%) as compared to the other 5 prior studies. Nonetheless, the minor difference in the classification performance of the algorithms may be due to the following factors: (1) different R-peak detection methods; (2) types and number of features extracted; (3) distinct feature extraction and feature selection techniques; (4) discrete choice of ECG signals from database; (5) uncommon classes of ECG heartbeats; and (6) different selection of machine learning classifier.

5 Conclusion

This study was focused on classifying 17 different types of arrhythmias by using an algorithm that included the morphological features of the R-peaks, wavelet-based statistical features, as well as the k-nearest neighbor (k-NN) algorithm. The algorithm's performance was tested by using the MIT-BIH arrhythmia benchmark database (MIT-BIHADB). The study's findings indicate that the proposed algorithm achieved an average accuracy (ACC) of 99.00%. The comparison of the findings of this work with earlier studies shows that the achievement of the algorithm produced in this work is in accord with the algorithms described in previous research. For future research, it is suggested that different filter design specifications be utilized and tested on the raw ECG signals to denoise the noises and artifacts more efficiently. It is also recommended that different combinations of morphological, statistical, and other essential features be used to improve the model to classify arrhythmias.

References

1. WebMD (2020) What is bigeminy? WebMD. <https://www.webmd.com/heart-disease/atrial-fibrillation/bigeminy-arrhythmia#>. Accessed 30 July 2020
2. Stanford Medicine (2020) What is an electrocardiogram? Stanford Medicine. <https://stanfordhealthcare.org/medical-tests/e/ekg.html>. Accessed 12 Mar 2020
3. Stanford Medicine (2020) Types of arrhythmia. Stanford Medicine. <https://stanfordhealthcare.org/medical-conditions/blood-heart-circulation/arrhythmia/types.html>. Accessed 12 Mar 2020
4. Martis RJ, Acharya UR, Lim CM, Mandana KM, Ray AK, Chakraborty C (2013) Application of higher order cumulant features for cardiac health diagnosis using ECG signals. *Int J Neural Syst* 23(4):1–19. <https://doi.org/10.1142/S0129065713500147>
5. Faziludeen S, Sabiq PV (2013) ECG beat classification using wavelets and SVM. In: 2013 IEEE conference on information and communication technologies. ICT 2013, pp 815–818. <https://doi.org/10.1109/CICT.2013.6558206>
6. Das MK, Ari S (2014) ECG beats classification using mixture of features. *Int Sch Res Not* 2014:1–12. <https://doi.org/10.1155/2014/178436>

7. Elhaj FA, Salim N, Harris AR, Swee TT, Ahmed T (2016) Arrhythmia recognition and classification using combined linear and nonlinear features of ECG signals. *Comput Methods Programs Biomed* 127:52–63. <https://doi.org/10.1016/j.cmpb.2015.12.024>
8. Sahoo S, Kanungo B, Behera S, Sabut S (2017) Multiresolution wavelet transform based feature extraction and ECG classification to detect cardiac abnormalities. *Meas J Int Meas Confed* 108:55–66. <https://doi.org/10.1016/j.measurement.2017.05.022>
9. Sahoo S, Mohanty M, Behera S, Sabut SK (2017) ECG beat classification using empirical mode decomposition and mixture of features. *J Med Eng Technol* 41(8):652–661. <https://doi.org/10.1080/03091902.2017.1394386>
10. Raj S, Ray KC (2018) Automated recognition of cardiac arrhythmias using sparse decomposition over composite dictionary. *Comput Methods Programs Biomed* 165:175–186. <https://doi.org/10.1016/j.cmpb.2018.08.008>
11. Yang W, Si Y, Wang D, Guo B (2018) Automatic recognition of arrhythmia based on principal component analysis network and linear support vector machine. *Comput Biol Med* 101:22–32. <https://doi.org/10.1016/j.compbiomed.2018.08.003>
12. Oh SL, Ng EYK, Tan RS, Acharya UR (2018) Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. *Comput Biol Med* 102:278–287. <https://doi.org/10.1016/j.compbiomed.2018.06.002>
13. Sannino G, De Pietro G (2018) A deep learning approach for ECG-based heartbeat classification for arrhythmia detection. *Future Gener Comput Syst* 86:446–455. <https://doi.org/10.1016/j.future.2018.03.057>
14. Xu X, Liu H (2020) ECG heartbeat classification using convolutional neural networks. *IEEE Access* 8:8614–8619. <https://doi.org/10.1109/ACCESS.2020.2964749>
15. Yang H, Wei Z (2020) Arrhythmia recognition and classification using combined parametric and visual pattern features of ECG morphology. *IEEE Access* 8:47103–47117. <https://doi.org/10.1109/ACCESS.2020.2979256>

Ground Truth from Multiple Manually Marked Images to Evaluate Blood Vessel Segmentation



Nazish Tariq, Michael Chi Seng Tang, Haidi Ibrahim, Teoh Soo Siang, Zunaina Embong, Aini Ismafairus Abd Hamid, and Rafidah Zainon

Abstract Blood vessel segmentation from digital images is one of the valuable processes for medical diagnosis. Many researchers have proposed blood vessel segmentation algorithms, which can segment the blood vessels automatically or with minimum human interventions. One of the popular blood vessel segmentation branches is edge-based segmentation. In this approach, only the edges are detected by the algorithm. While developing edge segmentation algorithms, researchers must evaluate their proposed methods' performance. If full-reference-based quality measures are utilized, the ground truth, which shows the targetted segmentation output, is needed. This ground truth is commonly generated manually, where human experts identify and draw the edges. However, the manually segmented edges may differ depending on the experts due to several factors, including individual preference. The work in this paper aims to give some insight into how to combine these images. This paper suggests that the edges be classified as useful edges, weak edges, and unintentional edges.

Keywords Blood vessels · Edge segmentation · Ground truth

N. Tariq · M. C. S. Tang · H. Ibrahim (✉) · T. S. Siang
School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Engineering Campus,
14300 Nibong Teba, Penang, Malaysia
e-mail: haidi@usm.my

Z. Embong
Department of Ophthalmology, School of Medical Sciences, Universiti Sains Malaysia, Health
Campus, 16150 Kota Bharu, Malaysia

A. I. A. Hamid
Department of Neurosciences, School of Medical Sciences, Universiti Sains Malaysia, Health
Campus, 16150 Kota Bharu, Malaysia

Brain and Behaviour Cluster, School of Medical Sciences, Universiti Sains Malaysia, Health
Campus, 16150 Kota Bharu, Malaysia

R. Zainon
Department of Biomedical Imaging, Advanced Medical and Dental Institute, Universiti Sains
Malaysia, SAINS@BERTAM, 13200 Kepala Batas, Penang, Malaysia

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
N. S. Ahmad et al. (eds.), *Proceedings of the 12th International Conference on
Robotics, Vision, Signal Processing and Power Applications*, Lecture Notes in Electrical
Engineering 1123, https://doi.org/10.1007/978-981-99-9005-4_67

531

1 Introduction

Image segmentation is commonly used as a part of diagnosis or prognosis procedures [1, 2]. This procedure will divide the image into several parts, where each part (or segment) has some properties in common. For an automatic diagnostic system, image segmentation prepares the input image by identifying the critical region or structures [3]. On the other hand, for the observations done by humans, image segmentation helps highlight the important areas or abnormalities, such as tumours, in the medical image [4].

The edges of important structures or organs are usually drawn by human experts. However, this process is tedious, usually requires a long time to be completed, and the segmented results may vary, depending on the human who traces the edges [5]. Thus, researchers have suggested many approaches that can do the segmentation automatically [6–9], or semi-automatic with minimum human intervention [10–12]. One of the popular approaches to segmenting the blood vessels from medical images is using edge detectors, where the segmentation process marks only the edges of the objects or organs on the image. This approach usually works by considering the locations where there are abrupt changes in intensity values.

While developing edge detection algorithms, researchers need to evaluate the performance of their proposed method. Many quality assessment methods are available for this purpose [13]. For a full-reference image quality assessment method, ground truth is required. The ground truth shows how the segmented output should look like. A common way to generate the ground truth is by asking the expert to manually marks the edges.

However, as mentioned before, manual segmentation output may not be consistent and has intra-human variations. For example, the ground truth produced by one person may not be exactly the same as the one produced by another person. Figure 1 shows an example of the reason why there are some variations on the marked edges. Some researchers mark the edges of the bright object, as shown in Fig. 1b. Some researchers mark the edges on the darker side of the object, as shown in Fig. 1c. On the other hand, some other researchers may mark freely anywhere near the edges, as shown in Fig. 1d.

The paper aims to propose a method to combine several manually marked images, to generate one ground truth. This paper is organized as follows. Section 2 describes our methodology. Section 3 presents our results and discussions. Section 4 concludes our work.

2 Methodology

By considering two input images, one will be taken as the base of the ground truth (G_A), whereas another image will be used as a reference (G_B). This work suggests of using the image with more edges as G_A . This can be done by counting the number

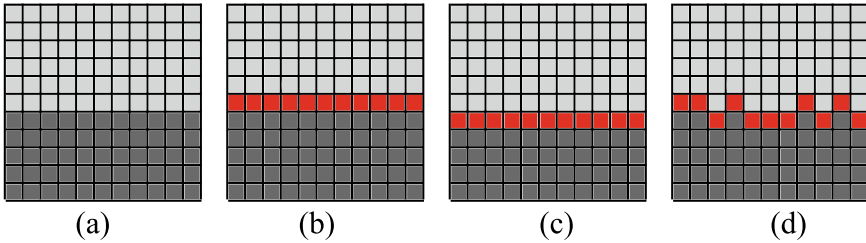


Fig. 1 An example to show the variations on the marking of the edges (indicated by the red-coloured cells). Each square presents a pixel. **a** A part of the object’s border. **b** The edge is marked on the bright pixels. **c** The edge is marked on the dark pixels. **d** The edge is marked freely around the edges

of pixels that are marked as edges in both images and identifying which image gives us the highest count. The reason is that if observer *A* detect more edges than observer *B*, it is considered that more information about the edges can be obtained from G_A . However, this work will classify these edges into three categories; useful edges, weak edges, and unintentional edges. All pixels in the final ground truth G_F are initially defined as non-edges.

The following classifications are done to each edge candidate in G_A . If an edgel in G_A can be found in G_B , this edgel is considered as useful edgel. Else, the following procedure is taken:

- Useful edge: Dilate the edge candidate pixel in G_A by an $n \times n$ square structuring element. Take the result of this dilation process as a mask (i.e., a mask with size of $(2n - 1) \times (2n - 1)$ pixels), and count the number of the edge elements (i.e., edgels) in G_B that are covered by this mask. If the number of edgels is equal to or more than $2n - 1$ pixels, this candidate will be considered as a useful edge. The reason why the structuring element of $n \times n$ is being chosen is that this work wants to confirm the similarity of the edges in G_A and G_B . The same edges that are identified should not be too far located from each other. Then, the reason why this work chooses a threshold of $2n - 1$ edgels is because, if we consider a straight horizontal or vertical edge within this window, the minimum number of the edgels to define this edge is $2n - 1$ pixels.
- Weak edge: If the number of edgels defined by the windows is greater than one but less than $2n - 1$, it is considered as a weak edge.
- Unintentional edges: If there is no edgel found within the window, this edgel is considered an unintentional edge.

3 Results and Discussions

This paper is using the open dataset known as STructured Analysis of the REtina (STARE) [14, 15]. This dataset contains retinal fundus images. This work used dataset in the category of blood vessel segmentation. The input images are colour images, with size 700×605 pixels, and stored in PPM file format. In this dataset, each fundus image has two ground truth, produced by two human experts. The ground truths are grayscale images, stored in PGM file format. In this dataset, the blood vessels areas are marked as white pixels, with intensity value of 255. To extract the edges from this image, this work generates two images. The first one is the dilated version of the image I_d , using a structuring element 2×2 pixels. Another image is the eroded version I_e , also by using a structuring element 2×2 pixels. The edges I_E is obtained by subtracting I_e from I_d .

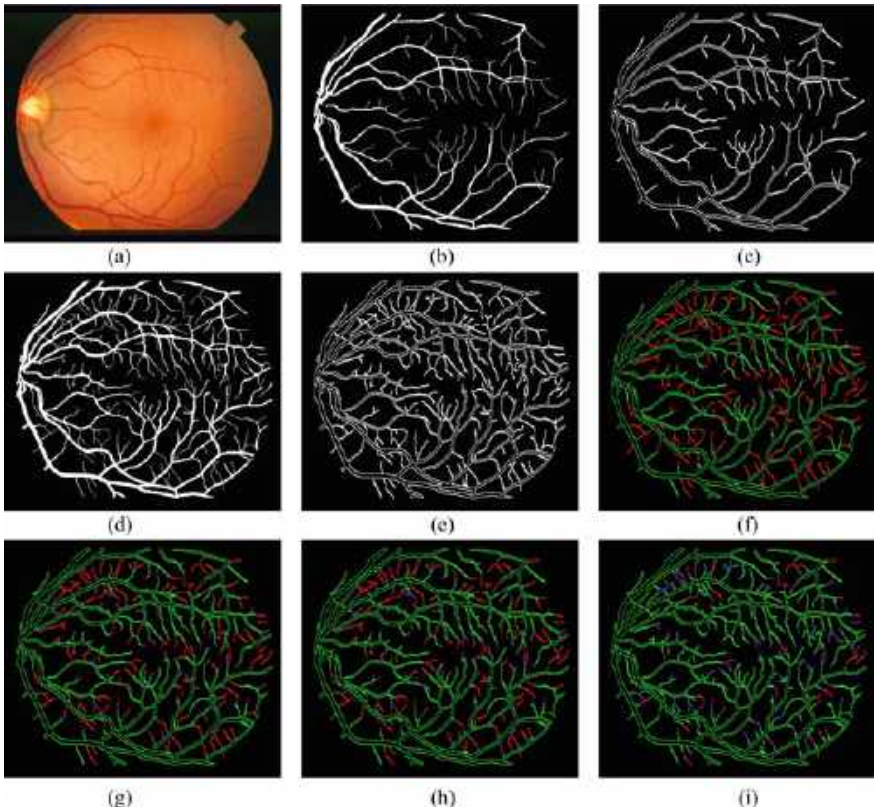


Fig. 2 **a** Input image 0077. **b** Blood vessels marked by the human observer 1. **c** I_E obtained from (b). **d** Blood vessels marked by the human observer 2. **e** I_E obtained from (d). **f** G_F obtained with $n = 2$. **g** G_F obtained with $n = 5$. **h** G_F obtained with $n = 7$. **i** G_F obtained with $n = 15$

Figure 2a shows an example of the input image (i.e., image 0077). The ground truth provided by the first observer is shown in Fig. 2b, whereas the ground truth provided by the second observer is given in Fig. 2d. The corresponding edges are shown in Fig. 2c, e, respectively. In this example, Fig. 2e will be taken as G_A , and Fig. 2c will be taken as G_B . This is because more edges in Fig. 2e. For G_F , this paper marks the useful edge with green colour, weak edge with blue colour, and unintentional edges with red colour. Results from several sizes of the structuring element (i.e., $n \times n$) are shown in Fig. 2f–i. As shown by this figure, more edgels from unintentional edges will be changed to useful edge or weak edge, by increasing the value of n .

4 Conclusion

This work gives some insight into combining two manually marked images to become one ground truth. The results shows that the edges can be classified into three categories, which are useful edges, weak edges and unintentional edges. Having only one ground truth will help us to simplify our evaluation process. For future development, investigations on how to set the proper value of n can be further investigated. Besides, the ground truth now has three classes of edges, which also requires modification on the assessment formulas. We would like to suggest introducing the weights for each class into the assessment formulas.

Acknowledgements This work is supported by the Ministry of Higher Education (MoHE), Malaysia, under the Fundamental Research Grant Scheme (FRGS), with grant number FRGS/1/2019/TK04/USM/02/1.

References

1. Salahuddin Z, Chen Y, Zhong X, Woodruff HC, Rad NM, Mali SA, Lambin P (2023) From head and neck tumour and lymph node segmentation to survival prediction on PET/CT: an end-to-end framework featuring uncertainty, fairness, and multi-region multi-modal radiomics. *Cancers* 15(7). ID:1932. <https://doi.org/10.3390/cancers15071932>
2. Tang Z, Duan J, Sun Y, Zeng Y, Zhang Y, Yao X (2023) A combined deformable model and medical transformer algorithm for medical image segmentation. *Med Biol Eng Comput* 61:129–137. <https://doi.org/10.1007/s11517-022-02702-0>
3. Singh L, Janghel RR, Sahu SP (2023) An empirical review on evaluating the impact of image segmentation on the classification performance for skin lesion detection. *IETE Tech Rev* 40(2):190–201. <https://doi.org/10.1080/02564602.2022.2068681>
4. Poonkodi S, Kanchana M (2023) MSCAUNet-3D: Multiscale Spatial Channel Attention 3D-UNet for lung carcinoma segmentation on CT image. In: 2023 international conference on advances in intelligent computing and applications (AICAPS). IEEE, Kochi, India, pp 1–5. <https://doi.org/10.1109/AICAPS57044.2023.10074322>

5. Yepes-Calderon F, McComb JG (2023) Eliminating the need for manual segmentation to determine size and volume from MRI. A proof of concept on segmenting the lateral ventricles. *PLoS ONE* 18(5):e0285414. <https://doi.org/10.1371/journal.pone.0285414>
6. Salahuddin Z, Chen Y, Zhong X, Rad NM, Woodruff HC, Lambin P (2023) HNT-AI: an automatic segmentation framework for head and neck primary tumors and lymph nodes in FDG-PET/CT images. In: Andrearczyk V, Oreiller V, Hatt M, Depeursinge A (eds) *Head and neck tumor segmentation and outcome prediction. HECKTOR 2022. Lecture notes in computer science*, vol 13626. Springer, Cham. https://doi.org/10.1007/978-3-031-27420-6_21
7. Kulseng CPS, Nainamalai V, Grovik E, Geitung JT, Aroen A, Gjesdal KI (2023) Automatic segmentation of human knee anatomy by a convolutional neural network applying a 3D MRI protocol. *BMC Musculoskelet Disord* 24. ID:41. <https://doi.org/10.1186/s12891-023-06153-y>
8. Tang MCS, Teoh SS, Ibrahim H (2022) Retinal vessel segmentation from fundus images using DeepLabv3+. In: *2022 IEEE 18th international colloquium on signal processing & applications (CSPA)*. IEEE, Selangor, Malaysia, pp 377–381. <https://doi.org/10.1109/CSPA55076.2022.9781891>
9. Onder M, Evli C, Turk E, Kazan O, Bayrakdar IS, Celik O, Costa ALF, Gomes JPP, Ogawa CM, Jagtap R, Orhan K (2023) Deep-learning-based automatic segmentation of parotid gland on computed tomography images. *Diagnostics* 13(4). ID:581. <https://doi.org/10.3390/diagnostics13040581>
10. Bruzadin A, Boaventura M, Colnago M, Negri RG, Casaca W (2023) Learning label diffusion maps for semi-automatic segmentation of lung CT images with COVID-19. *Neurocomputing* 522:24–38. <https://doi.org/10.1016/j.neucom.2022.12.003>
11. Vidal L, Biscaccianti V, Fragnaud H, Hascoet JY, Crenn V (2023) Semi-automatic segmentation of pelvic bone tumors: usability testing. *Ann 3D Print Med* 9. ID:100098. <https://doi.org/10.1016/j.stlm.2022.100098>
12. Ooi AZH, Embong Z, Abd Hamid AI, Zainon R, Wang SL, Ng TF, Hamzah RA, Teoh SS, Ibrahim H (2021) Interactive blood vessel segmentation from retinal fundus image based on Canny edge detector. *Sensors* 21(19). ID:6380. <https://doi.org/10.3390/s21196380>
13. Tariq N, Hamzah RA, Ng TF, Wang SL, Ibrahim H (2021) Quality assessment methods to evaluate the performance of edge detection algorithms for digital image: a systematic literature review. *IEEE Access* 9:87763–87776. <https://doi.org/10.1109/ACCESS.2021.3089210>
14. *STRUCTURED Analysis of the RETina*. <https://cecas.clemson.edu/~ahoover/stare/>. Accessed 21 Apr 2023
15. Hoover A, Kouznetsova V, Goldbaum M (2000) Locating blood vessels in retinal images by piece-wise threshold probing of a matched filter response. *IEEE Trans Med Imaging* 19(3):203–210. <https://doi.org/10.1109/42.845178>

A Comparative Study of Noise Reduction Techniques for Blood Vessels Image



Shadi Mahmoodi Khaniabadi, Haidi Ibrahim, Ilyas Ahmad Huqqani, Harsa Amylia Mat Sakim, and Soo Siang Teoh

Abstract The accurate analysis and interpretation of blood vessel images are essential for diagnosing and monitoring various medical conditions. However, these images often suffer from the presence of noise, which can hinder proper visualization and lead to erroneous interpretations. In this paper, we present a comprehensive comparative study of noise reduction techniques for blood vessel images, by a literature survey. The study encompasses both traditional and new methods, evaluating their performance, benefits, and challenges. Traditional methods, such as Anisotropic Diffusion Filtering and Wavelet Transform, have proven effective in preserving blood vessel structures and retaining fine details. However, they require careful parameter selection and may be computationally intensive. On the other hand, new techniques, including Contrast Limited Adaptive Histogram Equalization (CLAHE), Non-Local Mean Filter (NLM), and deep learning-based approaches, offer promising advancements in noise reduction capabilities with reduced computational complexity. The choice between traditional and new methods depends on specific application requirements, noise characteristics, and available computational resources. Our findings highlight the need for further research in parameter tuning, computational efficiency optimization, and hybrid approaches to enhance the noise reduction process in blood vessel images. This study contributes to the advancement of medical imaging by providing valuable insights for researchers and practitioners, enabling improved diagnostic accuracy and patient care.

Keywords Blood vessels · Noise reduction · Medical image

S. M. Khaniabadi · H. Ibrahim (✉) · I. A. Huqqani · H. A. Mat Sakim · S. S. Teoh
School of Electrical and Electronic Engineering, Universiti Sains Malaysia, 14300 Nibong Tebal,
Penang, Malaysia
e-mail: haidi@usm.my

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
N. S. Ahmad et al. (eds.), *Proceedings of the 12th International Conference on Robotics, Vision, Signal Processing and Power Applications*, Lecture Notes in Electrical Engineering 1123, https://doi.org/10.1007/978-981-99-9005-4_68

537

1 Introduction

Blood vessel imaging plays a crucial role in various medical fields, including cardiology, neurology, and ophthalmology, providing essential information for diagnosing and monitoring diseases [1]. However, the accurate analysis and interpretation of blood vessel images can be challenging due to the presence of noise, which can obscure important details and impact diagnostic accuracy [2, 3]. Therefore, the development of effective noise reduction techniques is vital to improve the quality and reliability of blood vessel images [1].

This paper presents a literature survey aimed at evaluating and analyzing several noise reduction techniques used for blood vessel images. The study encompasses both conventional and novel methods, with a focus on their respective advantages and limitations. By conducting a thorough investigation and comparison of these techniques, valuable insights can be gained to guide future research and facilitate the development of effective noise reduction strategies for blood vessel imaging.

2 Noise Reduction Techniques for Blood Vessels Images

Traditional methods of noise reduction techniques in blood vessel images encompass a range of approaches [4]. These methods aim to suppress noise while preserving the structural integrity and fine details of blood vessels, providing clear and accurate images for analysis [4, 5]. Techniques such as anisotropic diffusion filtering, wavelet transform, modified Gaussian filter, non-local means filtering, block-weighted total variation, Wiener filtering, Gaussian filtering, and total variation denoising offer different strategies for mitigating noise in blood vessel images [6–10], which is shown in Table 1. Recently, several new methods have been developed to address the challenge of noise reduction in blood vessel images [11, 12]. The new noise reduction methods in blood vessel images are presented in Table 2.

A comparative analysis was conducted between traditional and new methods. Both approaches share a common goal of reducing noise while preserving vital image features [13, 17, 21]. Traditional techniques have been extensively studied and widely implemented [6, 7, 9]. However, challenges arise in the form of parameter selection for optimal noise reduction, potential blurring or artifacts, and computational intensity for larger images [7]. On the other hand, new methods introduce innovative approaches to noise reduction [15, 20, 22]. Yet, they also pose challenges such as the need for accurate parameter tuning and ensuring the preservation of essential image features [22]. Choosing between traditional and new methods depends on the specific requirements of the application, the nature of the noise present in blood vessel images, and the available computational resources [1]. Researchers and practitioners should carefully consider the trade-offs between effectiveness, computational complexity, and the preservation of crucial image details when selecting the most suitable noise reduction technique [1, 4].

Table 1 Traditional methods of noise reduction techniques in blood vessels image

Ref.	Method	Description	Benefits	Difficulties
2019 [6]	Anisotropic diffusion filtering	Suppresses noise and preserves edges. Applies diffusion based on local image gradients	<ul style="list-style-type: none"> ✓ Preserves blood vessel structures and edges ✓ Robust against noise with varying spatial characteristics and retains fine details 	<ul style="list-style-type: none"> × Challenging to set appropriate parameters × Sensitive to noise near edges × Time-consuming for large images
2019 [7]	Wavelet transform	Reduces noise by decomposing the image into frequency bands. Reconstruct the image using selected bands	<ul style="list-style-type: none"> ✓ Reduces noise effectively while maintaining image quality ✓ Provides a balance between noise reduction and detail preservation 	<ul style="list-style-type: none"> × Requires careful selection of parameters × Computationally expensive × Not effective images with extreme noise or structures
2019 [8]	Modified Gaussian filter	Variants of Gaussian are applied to reduce noise	<ul style="list-style-type: none"> ✓ Smooths noise and preserves vessel structures 	<ul style="list-style-type: none"> × Proper selection of parameters for optimal noise reduction
2018 [5]	Non-local means filtering	Considers the similarity between image patches and estimates denoised pixel values based on this information	<ul style="list-style-type: none"> ✓ Effectively reduces noises and reserves vessel structures ✓ Adaptively suppresses noise at different scales and utilizes redundancy in natural images 	<ul style="list-style-type: none"> × Requires careful tuning of parameters for optimal results × Sensitive to variations in image content and noise characteristics and requires more computational time
2017 [13]	Block-weighted total variation	Reduces noise while preserving image quality for 3D coronary vessel wall MR imaging	<ul style="list-style-type: none"> ✓ Improved image quality ✓ Better visualization of small blood vessels ✓ Reduced artifacts 	<ul style="list-style-type: none"> × Computationally intensive × Requires parameter tuning × It may not be effective for all types of noise
2016 [14]	Wiener filtering	Estimate an optimal filter, minimizing the mean square error between the filtered and input image	<ul style="list-style-type: none"> ✓ Effective when noise characteristics are known or estimated ✓ Preserves blood vessel details and adapts to different noise levels 	<ul style="list-style-type: none"> × Requires accurate estimation of noise power spectrum × Performance is dependent on the noise model's accuracy

(continued)

Table 1 (continued)

Ref.	Method	Description	Benefits	Difficulties
2015 [9]	Gaussian filtering	Applies a Gaussian kernel attenuates noise while preserving blood vessel structures	<ul style="list-style-type: none"> ✓ Simple and computational efficient ✓ Preserves blood vessels and effective in reducing random noise 	<ul style="list-style-type: none"> × Challenging to select an appropriate kernel × Limited effectiveness against non-Gaussian noise
2014 [10]	Total variation denoising	Minimizes the total variation of the image while preserving its edges	<ul style="list-style-type: none"> ✓ Preserves blood vessels and fine details ✓ Robust against noise 	<ul style="list-style-type: none"> × Requires careful selection of parameters × Can lead to staircase artifacts × Computationally intensive for large images

3 Conclusion

In this research, a comprehensive analysis was conducted to compare traditional and new noise reduction techniques for blood vessel images. The strengths and limitations of each approach were identified, providing valuable insights for researchers and practitioners. Traditional methods were found to effectively preserve blood vessel structures and retain fine details but faced challenges in parameter selection and computational complexity. New methods showed promising advancements in noise reduction capabilities and adaptability to different noise characteristics. The choice between traditional and new methods should be based on specific application requirements, noise nature, and available computational resources. Future research should address parameter tuning challenges, optimize computational efficiency, and explore hybrid approaches. Overall, this study contributes to the understanding of noise reduction techniques for blood vessel images, aiding in improved diagnostic accuracy and patient care in medical imaging.

Table 2 Recent methods of noise reduction techniques in blood vessels image

Ref.	Method	Description	Benefits	Difficulties
2023 [15]	Non-local mean filter (NLM)	Reduces noise by taking weighted average of nearby pixel intensities	<ul style="list-style-type: none"> ✓ Robust towards variations in image characteristics ✓ Preserves important structures of the blood vessels 	<ul style="list-style-type: none"> × Difficult to find the right balance between noise reduction and details preservation × High computational complexity
2022 [11]	Denosing filter	It removes noise while preserves image features	<ul style="list-style-type: none"> ✓ The method is effective in handling different types and levels of noise 	<ul style="list-style-type: none"> × Requires careful tuning of parameters × High computational complexity
2022 [16]	Dynamic preprocessing and mathematical morphology	Reduce noise by applying a series of filters and operations (thresholding, erosion, dilation) to enhance the image contrast and remove unwanted elements	<ul style="list-style-type: none"> ✓ Improved image quality ✓ Reduced computational complexity 	<ul style="list-style-type: none"> × Include potential loss of important details × Sensitivity to parameter selection × Need to careful tuning to achieve optimal results
2021 [17]	Relative total variation-based	It reduces noise in the blood vessels by applying a regularization method the promote smoothness in the image while preserving important vessel structures	<ul style="list-style-type: none"> ✓ The technique effectively reduces noise in blood vessels and enhance the image quality ✓ The method selectively smooths the image while preserving important vessel details 	<ul style="list-style-type: none"> × High computational costs × Trade-off between noise reduction and preserving vessel structures

(continued)

Table 2 (continued)

Ref.	Method	Description	Benefits	Difficulties
2021 [18]	Median filtering	Replace each pixel value with the median value of its local neighborhood, reducing salt-and-pepper noise	<ul style="list-style-type: none"> ✓ Effectively removes salt-and-pepper noise ✓ Preserves blood vessel structures ✓ Robust against outliers 	<ul style="list-style-type: none"> × Can blur fine details and blood vessel boundaries × Limited effectiveness against Gaussian noise × Challenging to select a suitable window size
2020 [19]	Dual-tree complex wavelet transform (DTCWT)	Decompose the image into complex wavelet coefficients, threshold them to remove noise, and reconstruct the image	<ul style="list-style-type: none"> ✓ Effectively preserves edges and structures ✓ Robust against various types of noise, making it suitable for different imaging conditions 	<ul style="list-style-type: none"> × Computational expensive
2020 [20]	Deep learning-based technique	A deep learning algorithm iteratively reconstruct the image, gradually reduce the noise in the process	<ul style="list-style-type: none"> ✓ Improved image quality ✓ Better visualization of small blood vessels ✓ Reduced artifacts 	<ul style="list-style-type: none"> × Computational expensive × Requires parameter tuning × May not effective for all types of noise

Acknowledgements This work is supported by the Ministry of Higher Education (MoHE), Malaysia, under the Fundamental Research Grant Scheme (FRGS), with grant number FRGS/1/2019/TK04/USM/02/1.

References

1. Huang Q, Tian H, Jia L, Li Z, Zhou Z (2023) A review of deep learning segmentation methods for carotid artery ultrasound images. *Neurocomputing* 545:126298
2. Dash S, Verma S, Kavita, Bevinakoppa S, Wozniak M, Shafi J, Ijaz MF (2022) Guidance image-based enhanced matched filter with modified thresholding for blood vessel extraction. *Symmetry* 14(2)
3. Zheng C, Zhou H, Zhao K, Kong D, Ji T (2023) Internal carotid artery dissection with different interventions and outcomes: two case reports. *J Int Med Res* 51(2)
4. Guney G, Uluc N, Demirkiran A, Aytac-Kiperçil E, Unlu MB, Birgul O (2019) Comparison of noise reduction methods in photoacoustic microscopy. *Comput Biol Med* 109:333–341

5. Siregar S, Nagaoka R, Haq IU, Saijo Y (2018) Non local means denoising in photoacoustic imaging. *Jpn J Appl Phys* 57(7S1):07LB06
6. Wang Y, Wang Y (2019) Anisotropic diffusion filtering method with weighted directional structure tensor. *Biomed Signal Process Control* 53:101590
7. Lopez-Tiro F, Peregrina-Barreto H, Rangel-Magdaleno J, Ramirez-San-Juan JC (2019) Visualization of in-vitro blood vessels in contrast images based on discrete wavelet transform decomposition. In: *IEEE international instrumentation and measurement technology conference (I2MTC)*, pp 1–6
8. Halder A, Ghose S (2019) Blood vessel extraction from retinal images using modified Gaussian filter and bottom-hat transformation. In: *Computational intelligence in pattern recognition*
9. Wang S, Yin Y, Cao G, Wei B, Zheng Y, Yang G (2015) Hierarchical retinal blood vessel segmentation based on feature and ensemble learning. *Neurocomputing* 149:708–717
10. Selesnick IW, Graber HL, Pfeil DS, Barbour RL (2014) Simultaneous low-pass filtering and total variation denoising. *IEEE Trans Signal Process* 62(5):1109–1124
11. Shodiq MN, Yuniarno EM, Nugroho J, Purnama IKE (2022) Ultrasound image segmentation for deep vein thrombosis using UNet-CNN based on denoising filter. In: *2022 IEEE international conference on imaging systems and techniques (IST)*, pp 1–6
12. Bhutto JA, Tian L, Du Q, Sun Z, Yu L, Tahir MF (2022) CT and MRI medical image fusion using noise-removal and contrast enhancement scheme with convolutional neural network. *Entropy* 24(3)
13. Chen Z, Zhang X, Shi C, Su S, Fan Z, Ji JX, Xie G, Liu X (2017) Accelerated 3D coronary vessel wall MR imaging based on compressed sensing with a block-weighted total variation regularization. *Appl Magn Reson* 48(4):361–378
14. Subbaraya CK, D'sa AG, Manohar TV, Nanjesh BR (2016) Novel approach for extraction of blood vessels using morphology and filtering techniques. In: *2016 international conference on electrical, electronics, and optimization techniques (ICEEOT)*, pp 4315–4319
15. Iqbal S, Naveed K, Naqvi SS, Naveed A, Khan TM (2023) Robust retinal blood vessel segmentation using a patch-based statistical adaptive multiscale line detector. *Digit Signal Process* 139:104075
16. Chakour E, Mrad Y, Mansouri A, Elloumi Y, Bedoui MH, Andaloussi IB, Ahaitouf A (2022) Blood vessel segmentation of retinal fundus images using dynamic preprocessing and mathematical morphology. In: *8th international conference on control, decision and information technologies (CoDIT)*, vol 1, pp 1473–1478
17. Kim K, Jeong HW, Lee Y (2021) Performance evaluation of dorsal vein network of hand imaging using relative total variation-based regularization for smoothing technique in a miniaturized vein imaging system: a pilot study. *Int J Environ Res Public Health* 18(4)
18. Dikkala U, Joseph MK, Alagirisamy M (2021) A comprehensive analysis of morphological process dependent retinal blood vessel segmentation. In: *2021 international conference on computing, communication, and intelligent systems (ICCCIS)*, pp 510–516
19. Liu H, Lin S, Ye C, Yu D, Qin J, An L (2020) Using a dual-tree complex wavelet transform for denoising an optical coherence tomography angiography blood vessel image. *OSA Contin* 3(9):2630
20. Hong JH, Park EA, Lee W, Ahn C, Kim JH (2020) Incremental image noise reduction in coronary CT angiography using a deep learning-based technique with iterative reconstruction. *Korean J Radiol* 21(10):1165–1177
21. Kaur P, Singh RK (2023) A review on optimization techniques for medical image analysis. *Concurr Comput Pract Exp* 35(1)
22. Kumar KS, Singh NP (2023) Retinal disease prediction through blood vessel segmentation and classification using ensemble-based deep learning approaches. *Neural Comput Appl* 35(17):12495–12511

Survey on Blood Vessels Contrast Enhancement Algorithms for Digital Image



Shadi Mahmoodi Khaniabadi, Harsa Amylia Mat Sakim, Haidi Ibrahim, Ilyas Ahmad Huqqani, Farzad Mahmoodi Khaniabadi, and Soo Siang Teoh

Abstract This paper surveys blood vessel contrast enhancement algorithms in digital images, aiming to optimize imaging techniques for accurate analysis and interpretation of vascular structures. Various contrast enhancement techniques, including global and local approaches, are employed to improve the visibility and differentiation of blood vessels from the surrounding background. The investigation reveals that both global and local enhancement techniques play vital roles in enhancing blood vessel contrast. Global enhancement methods, such as spatial and frequency domain approaches, focus on enhancing overall contrast and visibility throughout the entire image. Yet, local enhancement techniques selectively enhance contrast and visibility in specific regions of interest, while preserving overall image quality. By combining global and local enhancement approaches, researchers can achieve comprehensive and targeted enhancement of blood vessel visibility and analysis. The findings emphasize the significance of utilizing suitable enhancement techniques to optimize blood vessel contrast in digital images and advance the field of medical imaging. This research contributes valuable insights for the development of optimized imaging techniques and algorithms for accurate blood vessel analysis and diagnosis.

Keywords Contrast enhancement · Digital images · Blood vessels

S. M. Khaniabadi · H. A. Mat Sakim · H. Ibrahim (✉) · I. A. Huqqani · S. S. Teoh
School of Electrical and Electronic Engineering, Universiti Sains Malaysia, 14300 Nibong Tebal,
Penang, Malaysia
e-mail: haidi@usm.my

F. M. Khaniabadi
School of Electrical and Electronic Engineering, Safahan University, Janbazan, 43431-81747
Isfahan, Iran

1 Introduction

This paper surveys various techniques specifically designed for blood vessel enhancement in digital images. Spatial Domain techniques operate directly on the pixel values of the image, utilizing filters and transforming to enhance blood vessel visibility [1, 2]. Frequency Domain techniques, on the other hand, employ frequency analysis to enhance blood vessel contrast by manipulating the frequency components of the image [2]. Understanding the characteristics and capabilities of these techniques is essential for optimizing the contrast enhancement process based on the specific requirements of blood vessel analysis [3, 4].

2 Contrast Enhancement Techniques

The techniques are divided into two main categories: global enhancement and local enhancement. The global enhancement category further branches into two subcategories: spatial domain and frequency domain techniques. The local category also splits into spatial domain and frequency domain methods.

2.1 Global Spatial Domain Techniques

These methods contribute significantly to the enhancement of blood vessel contrast, facilitating improved interpretation and examination of vascular structures in digital images [5, 6]. Erwin [7] employed contrast stretching as a global spatial domain technique to enhance blood vessel contrast in digital images. This method expands the range of pixel intensity values, resulting in improved visibility by increasing the contrast between dark and light regions. The benefit lies in enhancing the overall visibility of blood vessels. However, there is a potential risk of over-enhancement and loss of detail in certain areas, necessitating careful consideration for optimal results. Similarly, Suma and Saravana Kumar [6] focused on Histogram Equalization as another global spatial domain technique. By redistributing pixel intensity values across the entire range, this method uniformly enhances the visibility of blood vessels. It improves the differentiation between blood vessels and surrounding tissues, facilitating image interpretation and analysis. Yet, achieving a balance between contrast enhancement and preserving overall image quality poses a challenge. There is a possibility of losing fine details and over-amplification in specific regions. These global spatial domain techniques serve as valuable tools for the analysis and diagnosis of vascular structures, contributing to the development of optimized imaging techniques for accurate blood vessel analysis [8], which is shown in Table 1.

Table 1 Summary of recent studies in blood vessels contrast enhancement

Ref.	Method	Description	Advantages	Drawbacks
2023 [4]	Wavelet transform	Utilizes a set of Gabor filters to analyze image features at different orientations and frequencies	<ul style="list-style-type: none"> ✓ Multi-resolution analysis ✓ Orientation selectivity ✓ Frequency selectivity 	<ul style="list-style-type: none"> × Need careful parameter tuning × High computational complexity × Sensitive to noise
2022 [9]	Retinex-based techniques	Effectively enhance the contrast of blood vessels in digital images, irrespective of image resolution, by compensating for illumination variations and improving visibility	<ul style="list-style-type: none"> ✓ Enhances visibility and contrast of the blood vessels 	<ul style="list-style-type: none"> × Proper adjustment of parameters and fine-tuning is required
2021 [10]	Top-hat transform	Enhances the visibility of blood vessels by subtracting the morphological opening of the image from the original image	<ul style="list-style-type: none"> ✓ Enhances the visibility of fine structures ✓ Differentiate blood vessels from the surrounding tissues 	<ul style="list-style-type: none"> × Sensitive to noise × Requires careful selection of structuring elements and parameters
2021 [11]	Homomorphic filtering	Separating illumination and reflectance components. Use frequency domain processing	<ul style="list-style-type: none"> ✓ Simultaneously improves contrast and brightness ✓ Effective in varying lighting conditions 	<ul style="list-style-type: none"> × Exhibit low contrast compared to the surrounding background
2020 [7]	Contrast stretching	Expanding the range of pixel intensity values	<ul style="list-style-type: none"> ✓ Visibility of blood vessels has improved 	<ul style="list-style-type: none"> × May lead to over-enhancement and loss of details in some areas of the image
2020 [12]	2D Gabor wavelet	Utilizes a set of Gabor filters to analyze image features at different orientations and frequencies, effectively enhancing blood vessels at various scales	<ul style="list-style-type: none"> ✓ Multi-resolution analysis ✓ Orientation selectivity ✓ Frequency selectivity 	<ul style="list-style-type: none"> × Requires careful parameter tuning × High computational complexity × Sensitive to noise

(continued)

Table 1 (continued)

Ref.	Method	Description	Advantages	Drawbacks
2019 [3]	Histogram equalization	Redistribute pixel intensity values across the entire range, enhancing their visibility in digital images	✓ Enhances the visibility of blood vessels uniformly throughout the image	× Potential loss of fine details and over amplification in certain regions
2019 [8]	Local histogram equalization	Redistributing pixel intensity values across the entire range, resulting in improved visibility and differentiation between blood vessels and surrounding tissues	✓ Uniform enhancement of blood vessels visibility throughout the image ✓ Facilitation of the interpretation and analysis of vascular structures	× Retinal images used for detecting diabetic retinopathy often have low radiance levels

2.2 Local Spatial Domain Techniques

These techniques offer significant benefits, including the improvement of blood vessel visibility and contrast, thereby enabling precise analysis and diagnosis of vascular structures. The enhancement of blood vessel contrast in digital images, regardless of resolution, was discussed by Almalki et al. [9]. The utilization of Retinex-based Techniques, a form of Local Spatial Domain Techniques, effectively compensates for illumination variations and improves the visibility of blood vessels. Similarly, Ramos-Soto et al. [10] explored the use of the Top-Hat Transform as another local spatial domain technique, which aids in highlighting finer structures and differentiating blood vessels from the background and surrounding tissues. Additionally, Singh et al. [8] focused on Local Histogram Equalization, which uniformly enhances blood vessel visibility and improves differentiation from surrounding tissues. These techniques offer valuable tools for the analysis and interpretation of vascular structures, contributing to the development of optimized imaging techniques for accurate blood vessel analysis. However, careful parameter adjustment, consideration of noise sensitivity, and balancing contrast enhancement with image quality are important factors to address in order to achieve optimal results [8].

2.3 Global Frequency Domain Techniques

The impact of image resolution on blood vessel contrast enhancement in digital images has been extensively explored through the utilization of global frequency

domain techniques in the field of image processing [2]. These techniques are widely recognized and utilized due to their ability to operate in the frequency domain, facilitating thorough analysis and manipulation of the frequency components present in an image [4, 12, 13]. Notable methods in this domain include Fourier Transform (FT), Discrete Fourier Transform (DFT), Fast Fourier Transform (FFT), Wavelet Transform (WT), Gabor Filters, 2D Gabor Wavelet, Laplacian of Gaussian (LoG) Filter, and Bandpass Filtering [4, 12, 14–16]. As shown in Table 1, da Rocha et al. [12] using the 2D Gabor Wavelet technique, a Global Frequency Domain Technique. This method utilizes a set of Gabor filters to analyze image features at different orientations and frequencies, resulting in improved blood vessel visibility at various scales. The technique offers benefits such as multi-resolution analysis, orientation selectivity, and frequency selectivity. However, achieving optimal contrast enhancement requires careful parameter tuning, particularly for the Gabor filter parameters. In addition, Paul and Shan in [4] focused on the Wavelet Transform as a global frequency domain technique for enhancing blood vessel contrast. This method utilized Gabor filters to analyze image features at different orientations and frequencies, resulting in improved visibility at various scales. Yet, achieving optimal contrast enhancement required precise parameter tuning, particularly for the Gabor filter parameters.

2.4 Local Frequency Domain Techniques

Local frequency domain techniques are valuable tools in image processing for enhancing contrast and visibility in specific regions of an image by manipulating their frequency components [2, 17]. They offer the advantage of selectively enhancing desired features, such as blood vessels, while maintaining the integrity of surrounding areas. However, challenges arise in determining optimal region sizes and shapes, as well as fine-tuning parameters for optimal results [2]. Nonetheless, local frequency domain techniques contribute to the exploration of image resolution effects on blood vessel visibility and analysis, supporting the development of optimized imaging techniques in this field [2, 11]. As depicted in Table 1, the study by Tao [11] employed the Homomorphic Filtering method, categorized under local frequency domain techniques, to enhance the contrast of blood vessels in digital images. Through the separation of the illumination and reflectance components, this technique improved image resolution and enhanced the visibility of blood vessels. Notably, Homomorphic Filtering demonstrated the ability to simultaneously enhance both contrast and brightness, rendering it suitable for images captured under varying lighting conditions. The study utilized this method to investigate the impact of image resolution on blood vessel analysis and enhance the interpretation of vascular structures in digital images.

3 Conclusion

The study findings demonstrated the crucial roles of both global and local enhancement techniques in improving blood vessel contrast. Global enhancement methods focused on enhancing overall contrast and visibility across the entire image. Through comprehensive manipulation of pixel values or frequency components, these techniques effectively differentiated blood vessels from the background. Furthermore, the effectiveness of local enhancement techniques in targeting specific regions of interest within the image was evident. These techniques selectively enhanced contrast and visibility in localized areas while preserving overall image quality. These results underscore the importance of integrating global and local enhancement approaches to optimize blood vessel contrast enhancement. The combination of these techniques provides a comprehensive framework for enhancing blood vessel visibility and facilitating analysis. This integrated approach contributes to the development of optimized imaging techniques that enhance the accuracy and reliability of blood vessel analysis and diagnosis in various applications. By considering the advantages and characteristics of both global and local enhancement methods, researchers can advance the field of medical imaging and diagnosis through optimized imaging techniques for accurate blood vessel analysis. The findings of this study offer valuable insights for future research and development in the field of blood vessel contrast enhancement.

Acknowledgements This work is supported by the Ministry of Higher Education (MoHE), Malaysia, under the Fundamental Research Grant Scheme (FRGS), with grant number FRGS/1/2019/TK04/USM/02/1.


References

1. Luisi JD, Lin JL, Ameredes BT, Motamedi M (2022) Spatial-temporal speckle variance in the en-face view as a contrast for optical coherence tomography angiography (OCTA). *Sensors* 22(7)
2. Mustafa WA, Yazid H, Jaafar M (2016) A systematic review: contrast enhancement based on spatial and frequency domain. *J Adv Res Appl Mech* 28:1–8
3. Sukanya A, Rajeswari R (2019) Enhancement of coronary blood vessels based on Frangi's vesselness filter and morphological operations. *Int J Innov Technol Explor Eng* 8(10):274–281
4. Paul P, Shan BP (2023) Preprocessing techniques with medical ultrasound common carotid artery images. *Soft Comput*
5. Vijayalakshmi D, Nath MK, Acharya OP (2020) A comprehensive survey on image contrast enhancement techniques in spatial domain. *Sens Imaging* 21(1):40
6. Suma KG, Saravana Kumar V (2019) A quantitative analysis of histogram equalization-based methods on fundus images for diabetic retinopathy detection. Springer Singapore, Singapore, pp 55–63
7. Erwin (2020) Improving retinal image quality using the contrast stretching, histogram equalization, and CLAHE methods with median filters. *Int J Image Graph Signal Process* 12(2):30–41

8. Singh N, Kaur L, Singh K (2019) Histogram equalization techniques for enhancement of low radiance retinal images for early detection of diabetic retinopathy. *Eng Sci Technol Int J* 22(3):736–745
9. Almalki YE, Jandan NA, Soomro TA, Ali A, Kumar P, Irfan M, Keerio MU, Rahman S, Alqahtani A, Alqhtani SM, Hakami MAM, Alqahtani Saeed S, Aldhabaan WA, Khairallah AS (2022) Enhancement of medical images through an iterative McCann Retinex algorithm: a case of detecting brain tumor and retinal vessel segmentation. *Appl Sci* 12(16)
10. Ramos-Soto O, Rodríguez-Esparza E, Balderas-Mata SE, Oliva D, Hassanien AE, Meleppat RK, Zawadzki RJ (2021) An efficient retinal blood vessel segmentation in eye fundus images by using optimized top-hat and homomorphic filtering. *Comput Methods Programs Biomed* 201:105949
11. Tao J (2021) A method of blood vessel segmentation in fundus images based on image enhancement. *J Phys Conf Ser* 1955(1):12043
12. da Rocha DA, Barbosa ABL, Guimarães DS, Gregório LM, Gomes LHN, da Silva Amorim L, Peixoto ZMA (2020) An unsupervised approach to improve contrast and segmentation of blood vessels in retinal images using CLAHE, 2D Gabor wavelet, and morphological operations. *Res Biomed Eng* 36(1):67–75
13. Shahnawaz M, Choudhry A, Wadhvani R (2016) Analysis of digital image filters in frequency domain. *Int J Comput Appl* 140(6):12–19
14. Iwanowski M (2020) Image contrast enhancement based on Laplacian-of-Gaussian filter combined with morphological reconstruction. In: Burduk R, Kurzynski M, Wozniak M (eds) *Progress in computer recognition systems*. Springer International Publishing, Cham, pp 305–315
15. Tseng CC, Lee SL (2021) Frequency selective filtering of graph signal in directed graph Fourier transform domain. In: 2021 IEEE international conference on consumer electronics-Taiwan (ICCE-TW), pp 1–2
16. Manglik T, Axel L, Pai W, Kim D (2004) Use of bandpass Gabor filters for enhancing blood-myocardium contrast and filling-in tags in tagged MR images. *Proc Int Soc Magn Reson Med ISMRM* 3(c):1793
17. Usman B, Ayuba S (2015) Practical digital image enhancements using spatial and frequency domains techniques. *Int Res J Comput Sci (IRJCS)* 5

Face Image Authentication Scheme Based on Cohen–Daubechies–Feauveau Wavelets



Muntadher H. Al-Hadaad, Rasha Thabit , Khamis A. Zidan , and Bee Ee Khoo 

Abstract The recent interest of face image manipulation detection has been directed towards providing the ability of detecting various types of manipulations. In the best scenario, the available methods can detect the manipulations and localize the manipulated face region. The ability of recovering the face region after manipulation localization will be very useful in practical applications, however, this has not been highlighted in the previous researches. In this paper, a new face image authentication (FIA) scheme is presented based on image watermarking and Cohen–Daubechies–Feauveau (CDF) wavelets. In the proposed scheme, the CDF is used to generate the recovery bits from the face region in order to be used for recovering the face region when manipulations exist. Several experiments have been conducted to evaluate the performance of the proposed scheme which proved its efficiency in generating high quality watermarked images, detecting various types of manipulations, localizing the manipulated blocks in the face region, and recovering the face region with good visual quality. The comparison with the state-of-the-art detection schemes proved the superiority of the proposed scheme.

Keywords DeepFakes reveal · Face image manipulation detection · Face image authentication · Multimedia forensics

M. H. Al-Hadaad · R. Thabit

Computer Engineering Department, College of Engineering, Al-Iraqia University, Baghdad, Iraq
e-mail: muntazir3haydar@gmail.com

R. Thabit

e-mail: rasha.thabit@aliraqia.edu.iq

R. Thabit

Computer Techniques Engineering, Dijlah University College, Baghdad, Iraq

K. A. Zidan

Vice Rector of Al-Iraqia University for Scientific Affairs, Al-Iraqia University, Baghdad, Iraq
e-mail: khamis_zidan@aliraqia.edu.iq

B. E. Khoo (✉)

School of Electrical and Electronic Engineering, Engineering Campus, Universiti Sains Malaysia, 14300 Nibong Tebal, Penang, Malaysia
e-mail: beekhoo@usm.my

1 Introduction

The rapid development of the easy-to-use image processing applications facilitates the process of capturing and modifying the personal digital images. Nowadays, an enormous number of personal images are exchanged through open access channels such as internet [1]. The ability of exchanging personal images through internet has many benefits, however, the security and integrity of the shared images are the main concerns of the images' owners and their recipients [2].

The continuous improvements in the artificial intelligence field and the deep-learning based algorithms provide a fertile ground to present various face image manipulation (FIM) applications [3–5]. The face images can be modified for detrimental or benign intentions [5]. The detrimental FIM methods have been considered as a threat and accordingly different FIM detection (FIMD) techniques have been introduced to disclose the manipulated face images [6–8]. The deep-learning based FIMD techniques obtained good results in specific situations but they may face several limitations in others [8–13]. Recently, some review papers highlighted the restrictions and challenges that can be encountered by deep-learning based FIMD techniques such as the false detection cases, the long time required for training and processing the input datasets, the need for high quality images datasets, the need for a prior knowledge of the applied FIM method to choose the suitable detection technique, and many others [14–16].

In [17], a new direction in FIMD field has been suggested which is based on image watermarking technology. The detection scheme starts by detecting the face region followed by dividing the image into non-overlapping blocks and classifying the blocks into two groups that are either belong to the face region or belong to the non-face region. The tamper detection data have been generated from the face region's blocks and embedded in the non-face region's blocks using Slantlet transform (SLT)-based watermarking algorithm [18–20]. The scheme in [17] obtained promising results compared to previous FIMD techniques especially in terms of accuracy, time, and ability to detect different types of manipulations, however, it cannot recover the original face region when manipulations are detected.

To current date, the available FIMD techniques have focused on revealing manipulations in the face images while the ability to recover the original face region after revealing manipulations has not been highlighted. In this paper, a new face image authentication (FIA) scheme is presented which can reveal manipulations and recover the original face region. Since the watermarking-based scheme obtained promising results, the proposed scheme in this paper depends on the use of image watermarking technology. To generate the recovery bits from the face region, the Cohen–Daubechies–Feauveau (CDF) wavelets have been applied and the best family has been chosen to be adopted in the implementation of the proposed FIA scheme.

The rest of the paper contains: Sect. 2 which presents the proposed embedding and extraction algorithms; Sect. 3 illustrates the experimental results and their discussion; and Sect. 4 presents the conclusions of this work.

2 Proposed FIA Scheme

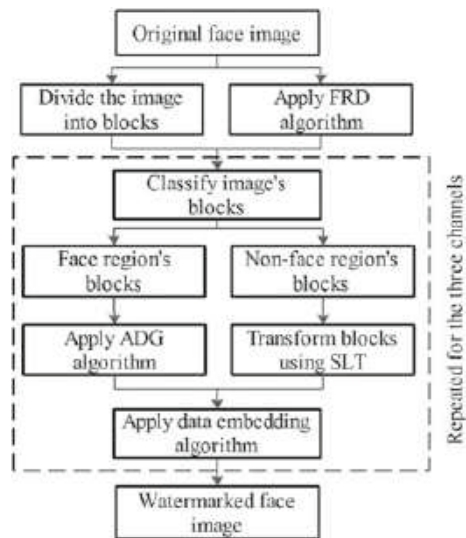
The proposed FIA scheme consists of two main algorithms that are applied at the embedding side and the extraction side. Each main algorithm consists of sub-algorithms which are required to obtain the targeted outputs. The following subsections presents the details of the proposed algorithms.

2.1 The Proposed Algorithm at the Embedding Side

The block diagram of this algorithm is shown in Fig. 1 which starts by reading the original face image and ends by the output watermarked face image. The procedure of this algorithm is explained in the following subsections.

- (1) *First stage:* The first stage of the embedding algorithm starts by reading the original face image (I_{OR}) of size $(M \times N \times 3)$. Then, the face region detection (FRD) algorithm is applied to detect and localize the face window. The FRD starts by applying the Multi-task Cascaded Neural Network (MTCNN) algorithm [17, 21] which gives the information about the face window including the top left corner of the face region, in addition to the height and width of the face window. The output of MTCNN is modified to select the pixels of the face region. After specifying the pixels' positions of the face region, a binary image (I_{Bin}) of size $(M \times N)$ is generated. The I_{Bin} pixels at the face window's positions are set to ones while the other pixels are set to zeros.

Fig. 1 Block diagram of the proposed embedding procedure



The I_{OR} has three channels called RGB channels (i.e., Red, Green, and Blue). Each channel image is divided into non-overlapping blocks of size $(B_s \times B_s)$ where B_s refers to the side length of the block. The I_{Bin} is also divided into non-overlapping blocks of size $(B_s \times B_s)$.

- (2) *Second satge*: The second stage starts by reading the blocks of one channel from I_{OR} and their corresponding blocks in I_{Bin} . The average of each I_{Bin} block is calculated to classify the block in the channel image ($I_{channel}$) at the same positions into two classes. If the average of I_{Bin} block equals to zero, the $I_{channel}$ block belongs to non-face region's blocks. If the average of I_{Bin} block is more than zero, the $I_{channel}$ block belongs to the face region's blocks.

The blocks belong to the face region are used as the input to the authentication data generation (ADG) algorithm. In the proposed ADG algorithm, the average of each block is calculated to be used for manipulation detection and localization. The face recovery information is generated from the approximation coefficients after transforming the blocks using Cohen–Daubechies–Feauveau (CDF) wavelets. The choice of this wavelet type depended on several experiments that have been conducted to find the best wavelet type that can be used to recover the face region with the best visual quality. The authentication information is converted to binary and concatenated to form a binary sequence that is ready to be embedded into the non-face blocks. The SLT-based watermarking algorithm [18] is applied to embed the authentication binary sequence in the non-face blocks. The output of embedding process is the watermarked non-face blocks.

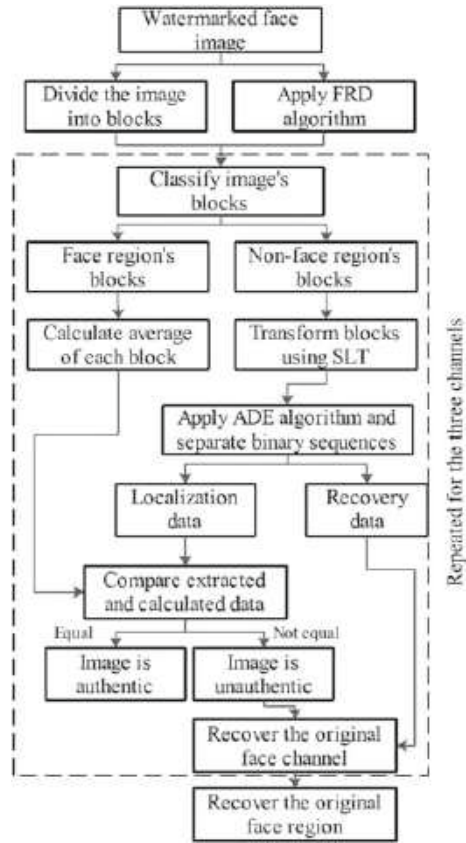
- (3) *Third satge*: In the third stage, the watermarked channel image is generated by rearranging the face region's blocks and the watermarked non-face region's blocks. The final watermarked face image (I_{WM}) is generated from the three watermarked channel images.

2.2 The Proposed Algorithm at the Extraction Side

The block diagram of this algorithm is shown in Fig. 2 which starts by reading the watermarked face image and ends by the authentication result and the recovered face region when manipulations exist. The procedure of this algorithm is explained in the following subsections.

- (1) *First stage*: The first stage of the extraction algorithm starts by reading the watermarked face image (I_{WM}) of size $(M \times N \times 3)$. Then, the face region detection (FRD) algorithm is applied to detect and localize the face window as explained in the embedding procedure. The FRD output is the binary image (I_{Bin}) of size $(M \times N)$ which is used to classify the blocks of the RGB channels of I_{WM} . The channel images and I_{Bin} are divided into non-overlapping blocks of size $(B_s \times B_s)$ where B_s refers to the side length of the block.

Fig. 2 Block diagram of the proposed extraction procedure



- (2) *Second stage:* The second stage starts by reading the blocks of one channel from I_{OR} and their corresponding blocks in I_{Bin} . The average of each I_{Bin} block is calculated to classify the block in the channel image $I_{channel}$ as explained in the embedding procedure. The average of each block belongs to the face region is calculated and saved to be used in the third stage of this algorithm. The blocks belong to the non-face's region are used as the input to the authentication data extraction (ADE) algorithm. In the proposed ADE algorithm, the embedded binary sequence which consists of the manipulation localization and recovery information is extracted from the non-faces' blocks using SLT-based watermark extraction algorithm [18]. The extracted binary sequence is separated into two parts that are localization bits and recovery bits. The localization bits are converted to decimal values to recover the original average values of the blocks. The recovery bits are used to recover the original approximation coefficients of the CDF wavelets.
- (3) *Third stage:* In the third stage, the calculated average values from the face region's blocks and the extracted average values from the non-face region's

blocks are compared to detect and localize manipulations. If the compared values are identical, the block is considered authentic, else the block is considered unauthentic and the recovery procedure starts. The extracted approximation coefficients from the previous stage are used to recover the face region. The horizontal, vertical, and details coefficients are considered zeros and the inverse CDF transform is applied to generate the recovered face region.

3 Experimental Results

The experimental work has been conducted first to evaluate the generated recovery data from the proposed ADG algorithm. The second part of the experimental work has been conducted to evaluate the proposed FIA scheme in terms of visual quality, payload, capacity, time complexity, and the authentication capability.

Extensive experiments were done to choose the best B_s and the results proved that the $B_s = 16$ obtained a good compromise between visual quality and robustness of the embedded data. The final subsection, presents the comparison with the state-of-the-art FIMD schemes.

3.1 Test of Recovery Data

To choose the best wavelet family, the visual quality of the recovered face region using the proposed recovery algorithm has been tested by calculating the Peak Signal-to-Noise Ratio (*PSNR*) between the original face region and the recovered face region. Thirty types of wavelet families have been tested and the CDF3.5 obtained the highest *PSNR* values, therefore, it has been adopted in the implementation of the proposed FIA algorithms. Samples of the obtained results from this test are presented in Table 1.

3.2 Test of FIA Performance

In this test, different face images with various sizes have been used to evaluate the performance of the proposed FIA scheme, samples of the test images are shown in Fig. 3. The *PSNR* between I_{OR} and I_{WM} has been calculated to test the visual quality of the watermarked images. Table 2 presents samples of the visual quality test results.

The time complexity has been calculated for the embedding and extraction algorithms using tic toc command in MATLAB software. The computer used in this experiment has the following properties: 2.60 GHz Intel® Core TM i7 CPU and 8 GB memory. The generated payload and the total embedding capacity have been also devalued. Table 3 presents samples of the results obtained from the payload, embedding capacity, and time complexity tests.

Table 1 Visual quality test for recovered face region

Wavelet type	PSNR (dB) for sample test images				
	Image 1	Image 2	Image 3	Image 4	Image 5
BIOR5.5	24.3623	18.0065	18.8657	17.415	28.1301
CDF1.5	31.8989	37.0062	33.3642	34.8596	38.3699
CDF2.2	32.0336	37.3676	33.2941	33.263	38.6792
CDF3.3	32.9486	38.3417	34.7419	35.5681	39.4507
CDF3.5	32.9982	38.394	34.788	35.6123	39.5158
CDF4.2	31.9031	37.0988	33.3159	33.397	38.181
CDF5.3	30.5142	35.307	32.6954	33.0712	35.5718
CDF6.2	28.9913	33.5113	30.3685	30.808	33.8572
CDF6.4	30.1049	34.5366	31.3998	31.7553	34.7971
DB3	32.5144	36.8807	34.222	33.9149	37.2864
DB6	31.274	34.1777	32.7085	30.8518	34.096
DB8	30.1001	32.0581	30.9866	29.0037	31.9616
Haar	32.0368	37.178	33.5153	34.9857	38.555
Sym3	14.6255	10.4511	10.2051	9.8947	18.1427
Sym6	24.8445	26.9725	24.1389	20.751	26.666
Sym7	25.0551	27.7253	24.8635	21.3963	27.48
Max. PSNR	32.9982	38.394	34.788	35.6123	39.5158

The bold is just to highlight that the highest PSNR value has been obtained using CDF 3.5.

To test the ability of the proposed scheme in revealing manipulations in the face region and recovering the original face region when manipulations exist, the watermarked face images have been manipulated using different attacks. Samples of the results are shown in Figs. 4, 5, 6 and 7 which proved that the proposed scheme can successfully localize the manipulated region and recover the original face region regardless the size of the manipulated area.

3.3 Comparison with the State-of-the-Art

As mentioned in the introduction section, the deep-learning based FIMD schemes [8–13] faced some limitations and most of the available schemes can detect only one type of manipulations. On the other hand, the watermarking-based technique in [17] can detect different types of manipulations but it cannot recover the original face region after manipulations revealing process. The proposed scheme outperforms various state-of-the-art schemes because it can reveal different types of manipulations, localize the manipulated blocks in the face region, and recover the face region if manipulations exist. Table 4 presents a general comparison of the proposed scheme



Fig. 3 Sample test images

Table 2 Visual quality test for watermarked images

Image name	Image size	Size of face region	PSNR (dB)
Img_1	$570 \times 1014 \times 3$	$160 \times 128 \times 3$	45.0877
Img_2	$600 \times 1200 \times 3$	$112 \times 80 \times 3$	54.0449
Img_3	$1152 \times 2048 \times 3$	$432 \times 384 \times 3$	44.3761
Img_4	$1152 \times 2048 \times 3$	$288 \times 224 \times 3$	48.2253
Img_5	$536 \times 1024 \times 3$	$144 \times 144 \times 3$	42.1705
Img_6	$600 \times 1200 \times 3$	$224 \times 192 \times 3$	46.1548
Img_7	$640 \times 800 \times 3$	$80 \times 64 \times 3$	53.0178
Img_8	$1669 \times 2500 \times 3$	$256 \times 224 \times 3$	56.5061
Img_9	$455 \times 728 \times 3$	$128 \times 128 \times 3$	38.336
Img_10	$608 \times 1080 \times 3$	$224 \times 176 \times 3$	46.1699

with various face image manipulation detection schemes. Table 5 presents a comparison between the proposed scheme and FIMD scheme in [17] in terms of payload and time complexity. The same test images shown in Fig. 3 have been used in this experiment. The payload generated in [17] is lower than the payload generated in the proposed scheme because in [17] only localization data are used while in the

Table 3 Payload, embedding capacity, and time complexity test results

Image name	Payload (bits)	Capacity (bits)	Embedding time (s)	Extraction time (s)
Img_1	170,880	404,352	0.872964	11.360993
Img_2	74,880	523,584	0.960417	7.56287
Img_3	1,379,520	1,639,872	3.070259	23.72201
Img_4	537,216	1,714,752	2.479163	37.787807
Img_5	172,992	386,304	0.881936	17.123695
Img_6	358,464	495,360	1.094932	24.019591
Img_7	43,008	378,240	0.76407	10.367643
Img_8	477,120	3,069,312	3.642619	35.756902
Img_9	136,896	226,368	0.867687	13.792615
Img_10	328,512	454,272	1.021984	24.837479



Fig. 4 Manipulation localization and recovery for test image ‘Img_1’



Fig. 5 Manipulation localization and recovery for test image ‘Img_2’



Fig. 6 Manipulation localization and recovery for test image ‘Img_3’



Fig. 7 Manipulation localization and recovery for test image ‘Img_7’

proposed scheme the data is generated from the localization and recovery data. Since the payload is higher in the proposed scheme, it is logical to require more time from embedding and extracting data.

Table 4 General comparison with the state-of-the-art schemes

Scheme	Method	Various manipulations	Manipulation detection ability	Manipulation localization ability	Face region recovery
[8–13]	Deep-learning	×	✓	×	×
[17]	Watermarking	✓	✓	✓	×
Proposed	Watermarking + CDF	✓	✓	✓	✓

Table 5 Comparison between the proposed scheme and the scheme in [17]

Scheme	Scheme in [17]			Proposed scheme		
	Payload (bits) localization data	Embedding time (s)	Extraction time (s)	Payload (bits) localization and recovery data	Embedding time (s)	Extraction time (s)
Img_1	3264	0.550777	0.206595	170,880	0.872964	11.360993
Img_2	1920	0.671195	0.177495	74,880	0.960417	7.56287
Img_3	23,040	1.458834	0.919955	1,379,520	3.070259	23.72201
Img_4	10,560	1.455500	0.559119	537,216	2.479163	37.787807
Img_5	3456	0.524380	0.175104	172,992	0.881936	17.123695
Img_6	6912	0.649806	0.291740	358,464	1.094932	24.019591
Img_7	1536	0.505593	0.119490	43,008	0.76407	10.367643
Img_8	8448	2.059830	0.640550	477,120	3.642619	35.756902
Img_9	2688	0.459536	0.138916	136,896	0.867687	13.792615
Img_10	6528	0.622006	0.275500	328,512	1.021984	24.837479

4 Conclusions

This paper presents a new face image authentication (FIA) scheme which can localize manipulations and recover the original face region. The proposed recovery algorithm based on the use of CDF wavelet which has been included in the FIA steps to recover the face region. The preliminary tests have been conducted to choose the best wavelet type and the results proved that the CDF3.5 obtained the best results. Extensive experiments have been conducted to evaluate the proposed FIA scheme and the results proved that it can detect different face image manipulations and recover the face region after manipulations localization. The general comparison with the state-of-the-art schemes proved the superiority of the proposed scheme and it can be applied to ensure the intactness and safety of the digital face images in different practical applications especially in the image forensics and data security fields.

Acknowledgements The authors would like to thank their institutions for encouraging and supporting their researches.

References

1. Hodeish ME, Bukauskas L, Humbe VT (2022) A new efficient TKHC-based image sharing scheme over unsecured channel. *J King Saud Univ Comput Inf Sci* 34(4):1246–1262. <https://doi.org/10.1016/j.jksuci.2019.08.004>
2. Devipriya M, Brindha M (2022) Secure image cloud storage using homomorphic password authentication with ECC based cryptosystem. *Adv Syst Sci Appl* 22(1):92–116

3. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J (2020) Deepfakes and beyond: a survey of face manipulation and fake detection. *Inf Fusion* 64:131–148. <https://doi.org/10.1016/j.inffus.2020.06.014>
4. Verdoliva L (2020) Media forensics and deepfakes: an overview. *IEEE J Sel Top Signal Process* 14:910–932
5. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J (2022) An introduction to digital face manipulation. In: Rathgeb C, Tolosana R, Vera-Rodriguez R, Busch C (eds) *Handbook of digital face manipulation and detection: from deepfakes to morphing attacks*. Springer International Publishing, Cham, pp 3–26. https://doi.org/10.1007/978-3-030-87664-7_1
6. Vamsi VVNS et al (2022) Deepfake detection in digital media forensics. *Glob Transit Proc*. <https://doi.org/10.1016/j.gltip.2022.04.017>
7. Korshunov P, Marcel S (2018) Deepfakes: a new threat to face recognition? Assessment and detection. arXiv:1812.08685
8. Matern F, Riess C, Stamminger M (2019) Exploiting visual artifacts to expose deepfakes and face manipulations. In: 2019 IEEE winter applications of computer vision workshops (WACVW), pp 83–92. <https://doi.org/10.1109/WACVW.2019.00020>
9. Hu S, Li Y, Lyu S (2021) Exposing GAN-generated faces using inconsistent corneal specular highlights. In: ICASSP 2021—2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 2500–2504. <https://doi.org/10.1109/ICASSP39728.2021.9414582>
10. Han X, Ji Z, Wang W (2020) Low resolution facial manipulation detection. In: 2020 IEEE international conference on visual communications and image processing (VCIP), pp 431–434. <https://doi.org/10.1109/VCIP49819.2020.9301796>
11. Yang X, Li Y, Qi H, Lyu S (2019) Exposing GAN-synthesized faces using landmark locations. In: *Proceedings of the ACM workshop on information hiding and multimedia security*
12. McCloskey S, Albright M (2019) Detecting GAN-generated imagery using saturation cues. In: 2019 IEEE international conference on image processing (ICIP), pp 4584–4588. <https://doi.org/10.1109/ICIP.2019.8803661>
13. Li H, Li B, Tan S, Huang J (2018) Detection of deep network generated images using disparities in color components. arXiv:1808.07276
14. Vaccari C, Chadwick A (2020) Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Soc Media Soc* 6(1):2056305120903408. <https://doi.org/10.1177/2056305120903408>
15. Salih ZA, Thabit R, Zidan KA, Khoo BE (2022) Challenges of face image authentication and suggested solutions. In: 2022 international conference on information technology systems and innovation (ICITSI), pp 189–193. <https://doi.org/10.1109/ICITSI56531.2022.9970797>
16. Pashine S, Mandiya S, Gupta P, Sheikh R (2021) Deep fake detection: survey of facial manipulation detection solutions. *Int Res J Eng Technol* 8(8):4441–4449. [Online]. Available: <http://arxiv.org/abs/2106.12605>
17. Salih ZA, Thabit R, Zidan KA, Khoo BE (2022) A new face image manipulation reveal scheme based on face detection and image watermarking. In: 2022 IEEE international conference on artificial intelligence in engineering and technology (IICAJET), no 1001, pp 1–6. <https://doi.org/10.1109/iicaiet55139.2022.9936838>
18. Thabit R, Khoo BE (2017) Medical image authentication using SLT and IWT schemes. *Multimed Tools Appl* 76(1):309–332. <https://doi.org/10.1007/s11042-015-3055-x>
19. Thabit R, Khoo BE (2014) A new robust reversible watermarking method in the transform domain. In: *Lecture notes in electrical engineering LNEE*, vol 291. https://doi.org/10.1007/978-981-4585-42-2_19
20. Thabit R, Khoo BE (2015) A new robust lossless data hiding scheme and its application to color medical images. *Digit Signal Process* 38. <https://doi.org/10.1016/j.dsp.2014.12.005>
21. Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process Lett* 23(10):1499–1503. <https://doi.org/10.1109/LSP.2016.2603342>

A Finger Knuckle Print Classification System Using SVM for Different LBP Variants



Imran Riaz , Ahmad Nazri Ali , and Ilyas Ahmad Huqqani 

Abstract Finger knuckle print is one of the most important biometric traits and plays a vital role in a secure identification system. In this paper, performance evaluation of local binary pattern (LBP) and its variants center symmetric local binary pattern (CS-LBP) and median local binary pattern (MLBP) are investigated. After feature extraction, a support vector machine (SVM) with the linear kernel is used for the performance evaluation of two different datasets named the Poly-U FKP dataset and the USM-FKP dataset. The experimental results show that CS-LBP performs better for the USM-FKP dataset with an accuracy of 86.2% which demonstrates the potential of the FKP classification system.

Keywords Finger knuckle print · Local binary pattern · Support vector machine

1 Introduction

Hand based biometric identification has drawn an increasing attention due to its low cost of data acquisition, reliability, acceptability and user friendly [1]. Among different hand based biometrics, fingerprint is the first modality used for recognizing a person. Finger dorsal surface called finger knuckle print has emerged as an alternative biometric with numerous advantages; uniqueness for each individual, invariant and

I. Riaz · A. N. Ali (✉) · I. A. Huqqani
School of Electrical and Electronic Engineering, University Science Malaysia, Engineering Campus, 14300 Nibong Tebal, Penang, Malaysia
e-mail: nazriali@usm.my

I. Riaz
e-mail: imran.ee@must.edu.pk

I. A. Huqqani
e-mail: ilyashuqqani@gmail.com

I. Riaz
Mirpur Institute of Technology, Mirpur University of Science and Technology, MUST, Mirpur Azad Kashmir, Pakistan

discriminant features, convenient image acquisition and reliable for person recognition. Therefore, finger knuckle based recognition system has emerged as an active research frontier. The skin pattern of FKP is highly rich in texture due to skin folds and is unique to all people. Feature extraction plays vital role in biometric recognition system. Since, FKP images contain high texture information, therefore, a high-quality feature extraction technique is the core requirement for better recognition accuracy. The success of LBP and its variants in different applications motivated us to evaluate it in the FKP classification system. Although, researchers considered LBP in FKP classification, but majority of the research for FKP using LBP is conducted either on multi-algorithm system or combined with some other biometric trait named as multimodal system. Based on these findings, we have focused our research towards FKP classification system with a comparative analysis of different LBP variants.

In this paper, four different LBP texture descriptors named uniform LBP, average LBP, center symmetric LBP and compound LBP have been used for feature extraction process. For classification, we have employed a support vector machine using linear kernel. A self-collected database of USM-FKP images is also developed. The structure of the paper is as follows as: Related work is described in Sect. 2. Image preprocessing and feature extraction approaches are described in Sect. 3. Experimental results conducted are reported in Sect. 4. Finally, concluding remarks are presented in Sect. 5.

2 Related Work

Compared to other biometrics, several techniques for feature extraction of finger knuckle prints have been proposed in the literature. In [2], authors proposed a multi algorithm system for FKP recognition, where texture patterns were extracted using Gabor with Exception-Maximization algorithm and Scale Invariant Feature Transform (SIFT) algorithm was employed to extract the feature vectors from these obtained texture patterns. In a recent research study [3], authors proposed a feature optimization approach for identifying the best feature vectors of finger knuckle print.

In literature, some efforts for FKP recognition system based on local binary pattern and its variants have employed. In [4], authors presented LBP based FKP recognition system in multi resolution domain and experimented on PolyU FKP database by using nearest neighbor and extreme learning machine classifiers. Raid et al. [5] proposed the horizontal and vertical lines-LBP for feature extraction and experiments conducted on IITD-FKP dataset with 85.97% identification accuracy using city block distance metric. Zahra et al. [6] presented a technique based on Gabor filter and LBP for extracting features and finally Bio-Hashing approach is applied on the extracted feature vectors. In another research study [7], basic LBP with the addition of dividing the FKP image into blocks is used for personal identification and reported recognition rate of 87.63%. El-Tarhouni et al. [8], extended completed LBP (CLBP) descriptor which employs only magnitude and sign components and named as dynamic threshold CLBP (dTCLBP) for FKP recognition. Recently, Heidari et al.

[9] proposed a technique by combining the LBP entropy-based histogram of texture patterns at different levels with a set of statistical texture features. The genetic algorithm (GA) was applied to find the superior features and support vector machine (SVM)-based feedback approach was used for classification.

3 Proposed Method

In this section, the methodology of the proposed system including data acquisition, image pre-processing, feature extraction and classification is discussed. Figure 1a shows the block diagram of the proposed FKP classification system. A special image acquisition device, Logitech Pro web-cam C920 has been constructed to capture the finger knuckle images. Each subject is requested to place the right finger on the support in such a way that back side is facing the sensor. Place the right finger on the support in such a way that back side is facing the sensor. The acquisition of a sample finger knuckle image is shown in Fig. 1b, c. Pre-processing of the acquired FKP images is an important stage for recognition system. There are three major steps in the pre-processing stage, color segmentation, background removal and finally extracting the region of interest area.

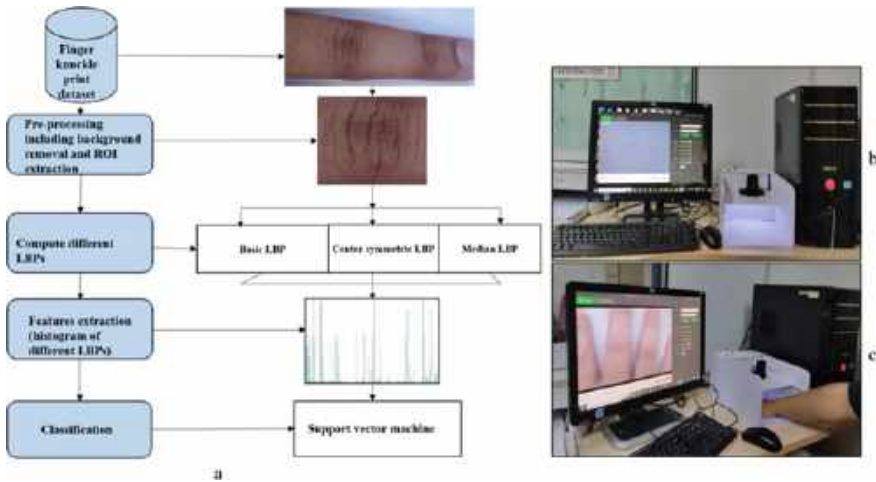


Fig. 1 a Proposed FKP classification system. b FKP image acquisition setup. c Acquisition of a sample FKP image

3.1 Feature Extraction

Feature extraction plays a vital role in the performance of the FKP recognition system. To extract the global or local information of the image, is the main goal of the feature extraction stage. Three different variants of the local binary pattern are used in the process of feature extraction, namely center symmetric LBP, average LBP and median LBP.

The LBP operator calculation for a given window having a number of pixels' P and radius R, is given in Eqs. 1 and 2.

$$LBP_P^R(x_c - y_c) = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \tag{1}$$

$$s(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases} \tag{2}$$

3.1.1 Center Symmetric Local Binary Pattern (CSLBP)

The basic idea of center symmetric LBP (CSLBP) [10] is to compare only pairs of horizontal, vertical and diagonal pixels depicted in Fig. 2a, instead of thresholding each pixel against the central pixel in basic LBP. The CSLBP descriptor is computed as follows in Eq. 3.

$$CSLBP_{P,R,T} = \sum_{i=0}^{\frac{P}{2}-1} s(g_i - g_{i+\frac{P}{2}})2^i, \text{ and } s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

The dimensionality of the extracted feature histogram is reduced from 28 = 256 to 24 = 16 dimensions, which also reduces the storage space.

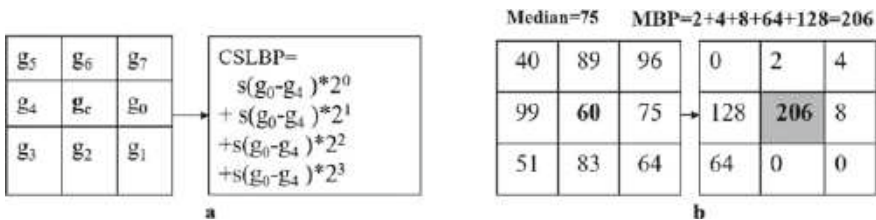


Fig. 2 An illustration for computing. a Center symmetric LBP. b Median binary pattern (MBP)

3.1.2 Median Binary Pattern (MBP)

Hafiane et al. [11], introduced median binary pattern (MBP), another extension of the basic LBP operator. The MBP is computed using Eq. 4.

$$MBP = \sum_{i=0}^P s(g_i) \times 2^i, \text{ and } s(g_i) = \begin{cases} 1, & \text{if } g_i \geq Med \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

In Fig. 2b, computation of MBP is illustrated where central pixel is also included and median is 120.

4 Experimental Results and Analysis

In this section, results on the self-collected USM-FKP dataset and the PolyU FKP database are discussed for basic LBP, CSLBP, ALBP and MBP. The subsequent subsections provide a brief description of: (1) the FKP datasets; (2) performance metrics and the obtained results.

4.1 Finger Knuckle Print Datasets

4.1.1 Self-collected USM-FKP Dataset

In this research study, we collected FKP dataset images using Logitech Pro webcam C920, which are taken from 80 people including male and female students of University Science Malaysia. These samples are compiled in one session only, containing 10 images for right index finger only and the dataset contains 800 images.

4.1.2 PolyU-FKP Dataset

The PolyU-FKP dataset is considered as a benchmark dataset for FKP recognition system, it contains 2515 images of middle finger. In this research, images of 100 subjects have been utilized for comparing with the self-collected USM-FKP dataset.

4.2 Performance Metrics and Obtained Results

To analyze the different LBP descriptors, experiments are conducted on two different datasets which are self-collected FKP dataset and PolyU-FKP dataset. The feature

vectors obtained from the LBP descriptors are randomly divided with a ratio of 70:30 into training and testing features respectively. Support vector machine is applied for classification where precision, recall, F1-score and accuracy are used for performance evaluation.

Experimental results for both USM and PolyU FKP databases are shown in Fig. 5. The comparison results for precision, recall, F1-score and accuracy showed that different LBP descriptor have higher accuracy on the FKP dataset than PolyU. Also, this higher performance reveals that the implementation of the proposed finger knuckle acquisition system demonstrates the advantages of this system. It is also evident from the Fig. 5 that the best recorded accuracy for FKP dataset with center symmetric LBP is 86.25% compared with other LBP descriptors.

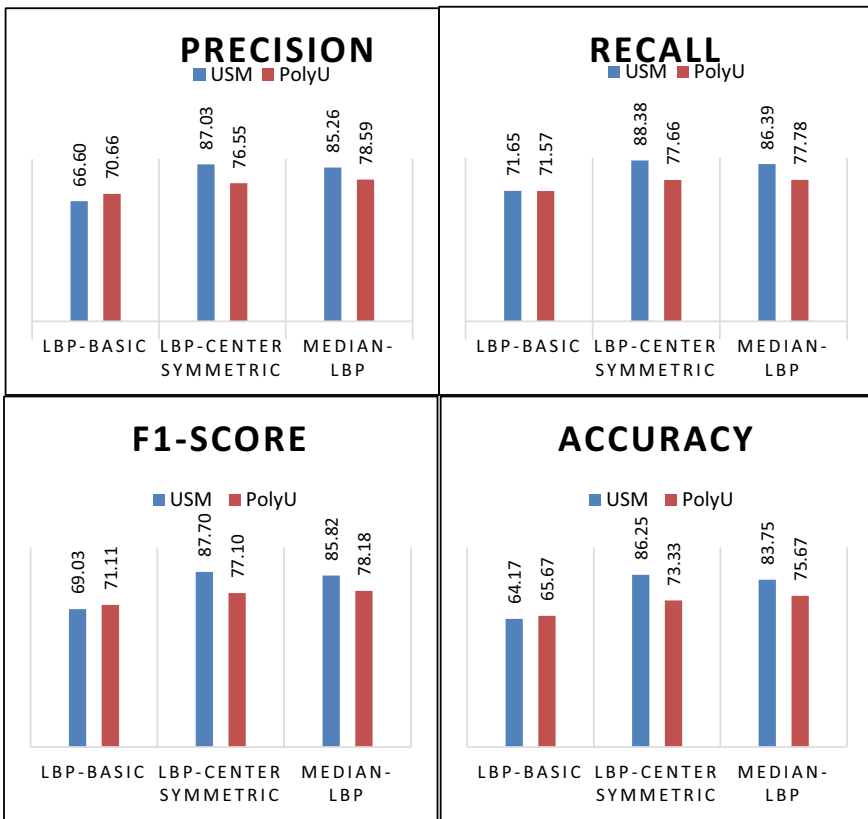


Fig. 3 Comparison of precision, recall, F1-score and accuracy

5 Conclusion

In this paper, LBP and its two variants CS-LBP and MLBP are analyzed for FKP recognition system. The proposed research study also involves developing a dataset of FKP images using Logitech Pro web-cam C920. Feature extraction using basic LBP, CS-LBP and MLBP was performed on two different datasets named Poly-U FKP and USM-FKP. Experimental results have revealed that CS-LBP outperforms and gives an accuracy of 86.25% for USM-FKP dataset with SVM classifier. In future, FKP will be combined with other biometrics to make a multi modal biometric system as well LBP feature extraction can also be combined with other feature extraction technique for multi-algorithm biometric system.

Acknowledgements Acknowledgment to Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project Code: FRGS/1/2021/ICT02/USM/02/1 for the financial support of this research. The images used in this study are acquired through approval ethical protocol with the study protocol code USM/JEPeM/21100657.

References

1. Papers DE (2004) Biometric-based technologies, no 101
2. Vidhyapriya R, Lovelyn Rose S (2019) Personal authentication mechanism based on finger knuckle print. *J Med Syst* 43(8). <https://doi.org/10.1007/s10916-019-1332-3>
3. Jayapriya P, Umamaheswari K (2022) Finger knuckle biometric feature selection based on the FIS_DE optimization algorithm. *Neural Comput Appl* 34(7):5535–5547. <https://doi.org/10.1007/s00521-021-06705-0>
4. Arun DR, Columbus CC, Meena K (2016) Local binary patterns and its variants for finger knuckle print recognition in multi-resolution domain. *Circuits Syst* 07(10):3142–3149. <https://doi.org/10.4236/cs.2016.710267>
5. Al-Nima RRO, Jarjes MK, Kasim AW, Sheet SSM (2020) Human identification using local binary patterns for finger outer knuckle. In: *Proceeding—2020 IEEE 8th conference on systems, process and control. ICSPC 2020, Dec 2020*, pp 7–12. <https://doi.org/10.1109/ICSPC50992.2020.9305779>
6. Shariatmadar ZS, Faez K (2013) Finger-knuckle-print recognition via encoding local-binary-pattern. *J Circuits Syst Comput* 22(6):1–16. <https://doi.org/10.1142/S0218126613500503>
7. Yu PF, Zhou H, Li HY (2014) Personal identification using finger-knuckle-print based on local binary pattern. *Appl Mech Mater* 441:703–706. <https://doi.org/10.4028/www.scientific.net/AMM.441.703>
8. El-Tarhouni W, Boubchir L, Bouridane A (2016) Finger-knuckle-print recognition using dynamic thresholds completed local binary pattern descriptor. In: *2016 39th international conference on telecommunications and signal process. TSP 2016*, pp 669–672. <https://doi.org/10.1109/TSP.2016.7760967>
9. Heidari H, Chalechale A (2020) A new biometric identity recognition system based on a combination of superior features in finger knuckle print images. *Turk J Electr Eng Comput Sci* 28(1):238–252. <https://doi.org/10.3906/elk-1906-12>
10. Heikkilä M, Pietikäinen M, Schmid C (2009) Description of interest regions with local binary patterns. *Pattern Recognit* 42(3):425–436. <https://doi.org/10.1016/j.patcog.2008.08.014>
11. Hafiane A, Seetharaman G, Zavidovique B (2007) Median binary pattern for textures classification. In: *Lecture notes in computer science (including subseries lecture notes in artificial*

intelligence and lecture notes in bioinformatics). LNCS, vol 4633, pp 387–398. https://doi.org/10.1007/978-3-540-74260-9_35

Assessment of Real-World Fall Detection Solution Developed on Accurate Simulated-Falls



Abdullah Talha Sözer, Tarik Adnan Almohamad, and Zaini Abdul Halim

Abstract One of the urgent and popular research areas is wearable devices-based fall detection (FD). Over the past 20 years, researchers have conducted many experiments in which falls and activities of daily living were simulated. Researchers inferred that real-world fall data is in demand rather than simulated fall data, but this inference still lacks comparisons. In this study, an assessment of a simulated fall dataset and a real-world fall dataset is proposed. The assessment investigates the efficacy of simulated data for developing an FD solution. Comparisons were conducted between FD methods developed on simulated and real-world data to observe the effectiveness of simulated falls. The experiments showed that the method with real-world data offered similar performances to the method with simulated data. In contrast to existing solutions, the provided comparison revealed that accurate simulated data are beneficial for developing a real-world FD solution.

Keywords Fall detection · Real-world fall · SVM · Simulated fall · Accelerometer · Machine learning

A. T. Sözer · T. A. Almohamad
Electrical-Electronics Engineering Department, Faculty of Engineering, Karabuk University,
Karabuk 78050, Türkiye
e-mail: talhasozer@karabuk.edu.tr

T. A. Almohamad
e-mail: tarikalmohamad@karabuk.edu.tr

Z. A. Halim (✉)
School of Electrical and Electronic Engineering, Universiti Sains Malaysia, 14300 Nibong Tebal,
Penang, Malaysia
e-mail: zaini@usm.my

1 Introduction

A fall event describes an uncontrollable movement of a person's body position. Falls may cause physiological health problems such as bruises, head trauma, hip fractures, disability, death; and psychological health problems like social isolation and negative economic outcomes [1]. Because of these, detection methods of fall events have attracted remarkable attention from researchers in the field.

Fall detection (FD) research comprises the distinction of daily activities from falling. The FD is achieved by collecting and processing data coming from various sources with the incorporation of thresholding, traditional machine learning, or deep learning tools. In the literature, two types of detection approaches are used, i.e., wearable-based and environment-based methods. The former approach relies on utilising sensors for acceleration, pressure, direction, magnetic field, and heart rate. It provides several advantages of cost, usability, and better privacy that can make FD systems broadly preferable. In the latter approach, devices such as standard cameras, infrared cameras, Kinect, motion, radar, and vibration sensors are used [1–6].

In wearable-based approaches, most studies use data that are collected from simulated actions to develop and validate FD methods. To generate simulated data, a group of subjects performs predetermined activities of daily living (ADL) and fall actions. *SisFall* [7], *ASLH* [8], *UMAFall* [9] are examples of publicly available fall datasets. However, some existing studies have shown that the performance of FD methods degrades when tested on real-world falls [10–14]. These studies claimed that simulated falls do not entirely represent real-world falls. The reason for this may be that it is not easy to simulate an involuntary and sudden action such as a fall. In addition, several predetermined fall types are performed to build up simulated fall data. However, real-world falls can be in many different forms.

Nevertheless, since obtaining real-world fall data is a difficult process, the usefulness of an *accurate* and comprehensive simulated fall dataset to develop a real-life FD solution should be further investigated. The term *accurate* refers to a fall that accurately mimics a real-world fall.

In this study, a comprehensive simulated dataset called *FallAllD* [15] and a real-world fall dataset [13] have been utilised to develop and test FD solutions. FD methods were developed and tested on both datasets and comparisons between the real-world data and the simulated data were conducted. Results show that the accurate simulated data are useful for developing a real-world FD solution. This reduces the need to generate real-world fall data.

The main contributions of the work can be summarised as follows:

1. Providing a comparison between the real-world fall dataset and the accurate simulated fall dataset.
2. Measuring the effectiveness of the simulated dataset to develop a real-world FD solution.

The rest of this paper is organised as follows: Sect. 2 explains the datasets, pre-processing of data, and machine learning method. Section 3 presents experimental

procedures and results. Section 4 discusses the results. Section 5 summarises the work and proposes some future works.

2 Material and Method

The efficacy of simulated data to obtain a real-world FD solution is studied in this section. The pattern recognition steps and experiments carried out are described.

2.1 Datasets

Two different datasets were used for the experiment. The first one is *FallAIIID*, which includes simulated fall and ADL data. The dataset contains 20 s signals obtained by the waist, neck, and wrist-worn accelerometer, gyroscope, barometer, and magnetometer over 15 individuals. It includes 35 types of falls and 44 types of ADL. *FallAIIID* has covered all possible types of falls, such as lateral fall and rotation while falling. The ADL data includes periodic usual movements such as walking and transient actions such as jumping, squatting, and lying on a bed. This makes *FallAIIID* a generic dataset that covers almost all daily actions. Thus, it is an excellent candidate to check if simulated falls can be used to develop a real-world FD solution.

The second dataset represents real-world fall and ADL data obtained from people with MS. The dataset contains fall and ADL signals from 25 individuals. The signals were obtained from a waist-worn accelerometer device during daily life activities. The individuals were tracked for eight weeks, and 54 real-world falls have been recognised from only 12 individuals out of the 25 ones [13]. However, there is no given information about types of falls and ADL in the abovementioned dataset.

2.2 Data Pre-processing and Future Extraction

3D-Accelerometer signals from waist-worn devices, which are frequently preferred for FD detection and included in both datasets, were used in the experiments. 6 s signal windows where falls occurred in the middle of them were obtained for both datasets. Figure 1 shows simulated and real-world fall signals.

Gravitational acceleration signal components were filtered by a 0.3 Hz high-pass filter. The frequently preferred features of FD were calculated for each axis [16]. The feature vector contained mean, median, range, standard deviation, median absolute deviation, skewness, kurtosis, and spectral power in 0–5, 5–10, 10–15, 15–20, 20–25 Hz bands.

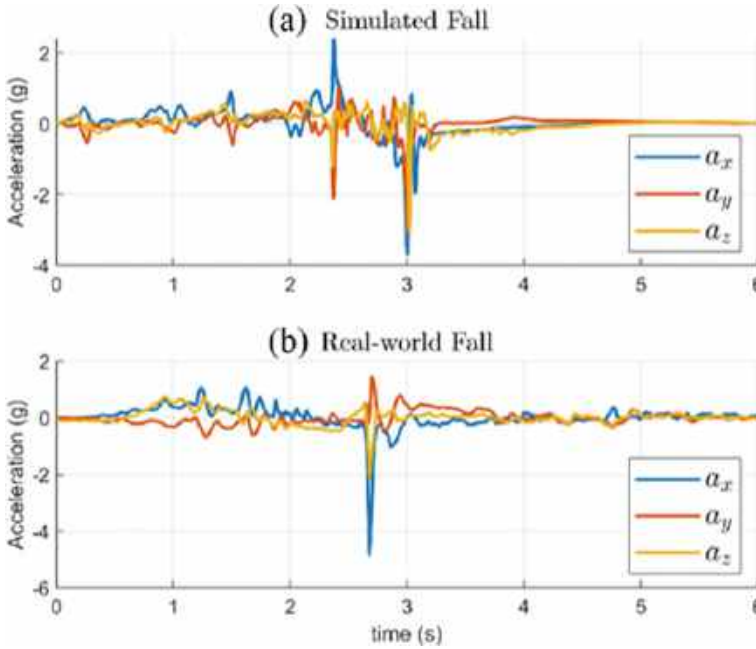


Fig. 1 Acceleration signals of **a** simulated and **b** real-world falls

2.3 Machine Learning Model

Artificial intelligence tools, such as deep neural networks (DNNs) and support vector machines (SVMs), have been widely utilised to perform FD solutions. Due to the significant detection capabilities of SVMs for small sizes of datasets [17, 18], SVM has been considered in this work. In the MATLAB software, the `fitsvm` function was utilised to implement the SVM tool and the Gaussian kernel function was used. The kernel scale was determined empirically to give the optimal FD performance. To evaluate the model performance, sensitivity, specificity, and accuracy measurements were used.

$$Sensitivity = \frac{True\ positive}{True\ positive + False\ negative} \quad (1)$$

Sensitivity refers to the FD capacity of the method.

$$Specificity = \frac{True\ negative}{True\ negative + False\ positive} \quad (2)$$

Specificity measurement refers to minimising false fall alarms in the performance of the FD solution.

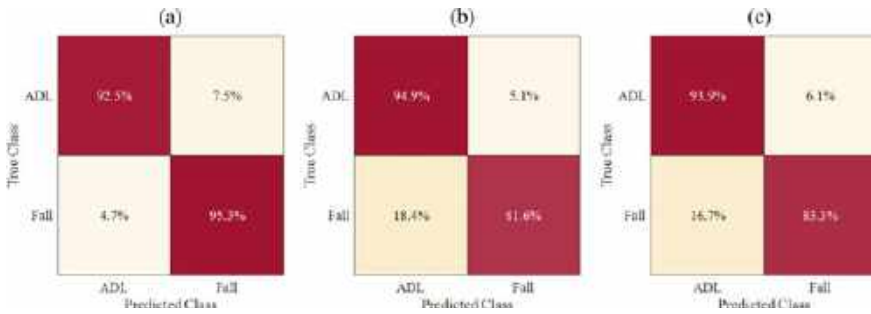


Fig. 2 a Results of the first experiment: sensitivity 95.3%, specificity 92.5%, accuracy: 93.9%.
 b Results of the second experiment: sensitivity: 81.6%, specificity: 94.9%, accuracy: 88.3%.
 c Results of the third experiment: sensitivity: 83.3%, specificity: 93.9%, accuracy: 88.6%

$$Balanced\ Accuracy = \frac{Specificity + Sensitivity}{2} \tag{3}$$

3 Experimental Procedures and Results

Three experiments were conducted to examine the possibility of developing a real-world fall detection solution using a comprehensive simulated fall dataset. In the first experiment, this dataset was divided into training and testing sections, which were then input into SVM for model evaluation. In the second experiment, SVM was trained using the simulated fall dataset and then tested on a real-world fall dataset. In the final experiment, the real-world fall dataset itself was split into training and testing parts, which were used to train and test an SVM model. We have considered the second experiment (training using simulated, testing using real-world datasets) to measure the effectiveness of the simulated dataset for developing a real-world FD solution.

The test results in Fig. 2 have shown that the models trained by the simulated dataset yielded similar FD performance on the real-world fall dataset.

4 Discussion

The experiments conducted with simulated and real-world fall datasets demonstrate remarkable results. Table 1 summarises the results of the experiments.

Existing methods in the literature showed that the FD performance of a model developed on simulated data is degraded when evaluated on real-world falls. When experiments 1 and 2 are compared, similar results are encountered. Contrarily, we assert that these results are insufficient to determine that simulated data is inadequate

Table 1 Results of all experiments

	Sensitivity (%)	Specificity (%)	Accuracy (%)
Experiment 1	95.3	92.5	93.9
Experiment 2	81.6	94.9	88.3
Experiment 3	83.3	93.9	88.6

for developing real-world FD solutions. This is because the outcomes of experiments 2 and 3 demonstrate that the model trained with the simulated fall dataset and the model trained with the real-world fall dataset exhibit the same performance when tested with real-world falls. In addition, when the results were examined, the same level of specificity was obtained in all models. However, sensitivity decreased in real-world fall tests. This shows that simulated falls can be more easily distinguished from real-world falls.

Based on the experiments, simulated datasets could be helpful in the development of real-world FD solutions. Also, we recommend that data acquisition from an accurate simulation of falls is inevitable. Fall simulation environments should be as identical as possible to real-life conditions. Conducted simulations should consider pre-fall ADL activities and locations where real-world falls often occur.

5 Conclusion and Future Work

This study explored the effectiveness of simulated data for generating a real-world FD solution. Our experiments were conducted under comprehensive, accurate simulated, and real-world datasets. The experiments demonstrated that accurate simulated data are beneficial for developing a real-world FD solution.

In future work, we envisioned that solid features for better representing the difference between ADL and fall are demanded. Moreover, new classes, such as near-fall, can be developed to evaluate current methods' performance better.

References

1. Rajagopalan R, Litvan I, Jung TP (2017) Fall prediction and prevention systems: recent trends, challenges, and future research directions. *Sensors (Switzerland)* 17(11):1–17
2. Kerdjidj O, Ramzan N, Ghanem K, Amira A, Chouireb F (2020) Fall detection and human activity classification using wearable sensors and compressed sensing. *J Ambient Intell Humaniz Comput* 11(1):349–361
3. Saleh M, Jeannes RLB (2019) Elderly fall detection using wearable sensors: a low cost highly accurate algorithm. *IEEE Sens J* 19(8):3156–3164
4. Wang C et al (2016) Low-power fall detector using triaxial accelerometry and barometric pressure sensing. *IEEE Trans Ind Inform* 12(6):2302–2311
5. Wang X, Ellul J, Azzopardi G (2020) Elderly fall detection systems: a literature survey. *Front Robot AI* 7

6. Wang Y, Wu K, Ni LM (2017) WiFall: device-free fall detection by wireless networks. *IEEE Trans Mob Comput* 16(2):581–594
7. Sucerquia A, López JD, Vargas-Bonilla JF (2017) SisFall: a fall and movement dataset. *Sensors (Switzerland)* 17(1)
8. Özdemir AT (2016) An analysis on sensor locations of the human body for wearable fall detection devices: principles and practice. *Sensors (Switzerland)* 16(8)
9. Casilari E, Santoyo-Ramón JA, Cano-García JM (2017) UMAFall: a multisensor dataset for the research on automatic fall detection. *Procedia Comput Sci* 110:32–39
10. Lipsitz LA et al (2016) Evaluation of an automated falls detection device in nursing home residents. *J Am Geriatr Soc* 64(2):365–368
11. Aziz O et al (2017) Validation of accuracy of SVM-based fall detection system using real-world fall and non-fall datasets. *PLoS ONE* 12(7):1–11
12. Sucerquia A, López JD, Vargas-Bonilla JF (2018) Real-life/real-time elderly fall detection with a triaxial accelerometer. *Sensors (Switzerland)* 18(4):1–18
13. Mosquera-Lopez C et al (2021) Automated detection of real-world falls: modeled from people with multiple sclerosis. *IEEE J Biomed Health Inform* 25(6):1975–1984
14. Palmerini L, Klenk J, Becker C, Chiari L (2020) Accelerometer-based fall detection using machine learning: training and testing on real-world falls. *Sensors (Switzerland)* 20(22):1–15
15. Saleh M, Abbas M, Le Jeannes RB (2021) FallAllID: an open dataset of human falls and activities of daily living for classical and deep learning applications. *IEEE Sens J* 21(2):1849–1858
16. Liu KC, Hsieh CY, Huang HY, Hsu SJP, Chan CT (2020) An analysis of segmentation approaches and window sizes in wearable-based critical fall detection systems with machine learning models. *IEEE Sens J* 20(6):3303–3313
17. Almohamad TA, Salleh MFM, Mahmud MN, Karas IR, Shah NSM, Al-Gailani SA (2021) Dual-determination of modulation types and signal-to-noise ratios using 2D-ASIQH features for next generation of wireless communication systems. *IEEE Access* 9:25843–25857
18. Almohamad TA, Mohd Salleh MF, Mahmud MN, Sa’D AHY (2018) Simultaneous determination of modulation types and signal-to-noise ratios using feature-based approach. *IEEE Access* 6:9262–9271

Deep Learning Based Distance Estimation Method Using SSD and Deep ANN for Autonomous Braking/Steering



Siti Nur Atiqah Halimi, Mohd Azizi Abdul Rahman,
Mohd Hatta Mohammed Ariff, Yap Hong Yeu, Nor Aziyatul Izni,
Mohd Azman Abas, and Syed Zaini Putra Syed Yusoff

Abstract The Automatic Emergency Braking (AEB) system is a mechanism that enables drivers to leverage the capabilities of their vehicles by warning them of potential collisions and assisting them in averting them. Autonomous Emergency Steering (AES) is one of the active safety systems that can assist with evasive steering. It will make it simpler for the driver to avoid an accident that could have been prevented. Concerns include the distance necessary to prevent a collision when turning or reversing and the required space when braking and turning. Given such inquiries, developing a system to estimate the distance between the vehicles is necessary. Consequently, this study suggested utilizing deep learning for AEB and AES to estimate the distance between vehicles using a monocular vision sensor. In addition, the object distance estimation method is employed as a distance estimation method. Experiments are conducted to determine the precision of the proposed method for estimating the distance between the target vehicle and the camera using LiDAR distances. The result indicates that the proposed method for estimating distance has an accuracy of 92% compared to LiDAR distance. As a result, the findings of this research have the potential to contribute to the methodological foundation for further understanding drivers' behavior, with the ultimate objective of lowering the number of accidents involving rear-end crashes.

S. N. A. Halimi · M. A. A. Rahman (✉) · M. H. M. Ariff
Malaysia Japan International Institute of Technology, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia
e-mail: azizi.kl@utm.my

Y. H. Yeu
Emoovit Technology Sdn Bhd, Persiaran APEC, Cyberjaya, Selangor, Malaysia

N. A. Izni
Centre of Foundation Studies, Universiti Teknologi MARA, Cawangan Selangor Kampus, Dengkil, Selangor, Malaysia

M. A. Abas
Faculty of Mechanical Engineering, Universiti Teknologi Malaysia, Johor Bahru, Johor, Malaysia

S. Z. P. S. Yusoff
Techcapital Resources Sdn Bhd, UPM-MTDC, Serdang, Selangor, Malaysia

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024 581
N. S. Ahmad et al. (eds.), *Proceedings of the 12th International Conference on Robotics, Vision, Signal Processing and Power Applications*, Lecture Notes in Electrical Engineering 1123, https://doi.org/10.1007/978-981-99-9005-4_73

Keywords Distance estimation · Deep learning · Artificial intelligence · Autonomous braking · Autonomous steering

1 Introduction

When an obstruction appears out of nowhere in front of the car, the driver is forced to make many split-second decisions and selections. However, a driver may be unable to react quickly enough to avoid the risk when an accident is about to occur. In this scenario, the Autonomous Emergency Braking (AEB) system can apply the emergency brakes automatically if they are activated. Moreover, the Autonomous Emergency Steering (AES) system will activate its emergency steering if it determines that a collision cannot be avoided based on the time-to-collision measurement. It will cause the vehicle to swerve to the left or right to avoid the collision.

The Automatic Emergency Braking (AEB) [1] system is a mechanism that enables drivers to maximize their vehicles' capabilities by alerting them to potential collisions and assisting them in avoiding them. Autonomous Emergency Steering (AES) [2] is one of the active safety systems that can aid in evasive steering. In contrast to AEB, if a potential collision is detected, the AES system automatically adjusts the steering to prevent a crash. With this aid, the driver will have an easier time avoiding an accident that could have been avoided. Concerns include the distance necessary to prevent a collision when turning or reversing and the required space when braking and turning. Given such inquiries, developing a system to estimate the distance between the vehicles is necessary.

Two methods that are currently available use deep learning for determining distance. The first approach is object distance estimation which calculates the distance to an object based on the object's size. Meanwhile, the term "depth estimation" refers to the second technique, which relies on a depth map as its source of ground truth. Alternately, deep learning-based depth estimation methods analyze the depth of the whole image but need to estimate the distance between cars precisely. DepthNet [3], DenseDepth [4], and MonoDepth [4, 5] are examples of depth estimation methods. These approaches have a risk of generalization, which become problematic if the driving situations observed on video during training and testing are different. In addition, using depth estimation methods in combination with object detection, especially with deep neural networks (DNNs) in real-world applications like autonomous vehicles, would not be possible. The approaches for estimating depth estimation based on deep learning need a lot of CPU power and processing.

Object distance estimation is another vision-based approach for measuring distance. This approach extracts the input characteristics from the bounding box of the identified object retrieved from the camera [6]. For instance, DistNet [7] was a recent attempt to construct a network for distance estimation. The authors used a CNN-based model (YOLO) for bounding box prediction rather than learning image features for distance estimation. The ANN is the most complicated and has the highest variance since it generates a nonlinear mapping between each detection parameter

and predicted output. Besides that, finding the correct configuration and architecture for ANN is the most challenging aspect of ANN. However, since deep learning-based depth prediction methods are computationally expensive, and using them with object detection in real-time applications may be challenging, this research proposed using Deep ANN as a distance estimation algorithm. Meanwhile, SSD with MobileNet architecture is used for object detection in the proposed Distance Estimation method.

The following outlines this paper's format: Sect. 2 explains the experimental conditions for data collecting and provides information on the proposed method. The outcomes of the experiments were analyzed more thoroughly in Sect. 3. Finally, Sect. 4 summarizes this study's findings and suggests further research.

2 Data Collection

This research presents a technique for measuring the distance between vehicles predicted on monocular vision and uses an approach based on deep learning. The vision sensor was chosen to use in this study because it is capable of high-level processing when paired with software. In addition, compared to stereo vision, monocular vision utilized for estimating distances provides advantages such as lower cost, greater ease of use, and faster processing speed. Using a vision sensor allows for object segmentation and the detecting of desired objects. Since methods for depth prediction based on deep learning require a significant amount of computing, and it can become difficult to apply these methods together with object detection in real-time applications, this study proposed using the object distance estimation method as a distance estimation method. Figure 1 shows the framework that consists of Deep ANN as a distance estimation algorithm. Meanwhile, SSD with MobileNet architecture is used for object detection in the Monocular Vision Distance Estimation method.

The KITTI dataset is by the most popular choice in mobile vehicles and autonomous transportation usage. The KITTI dataset's raw data for distance estimation includes the test images, train images and train annotations. The train annotations files include roughly 51,000 pieces of data. However, the KITTI dataset has just about 7000 pictures. The KITTI dataset has nine different Classes. The dataset is split into train and test using just the class 'Car' so that the focus may remain on the rear-end accident that occurred between two vehicles. More specifically, the dataset contains 70% of the train and 30% of the test data.

The modified KITTI dataset is trained and evaluated using this study's proposed Deep ANN distance estimation algorithms. The proposed method's input variables are Xmin, Ymin, Xmax, and Ymax, corresponding to the minimum and maximum object coordinate values from the bounding box data. This study's output variable generated by the Deep ANN architecture is the distance from the camera (Zloc). The Hidden Layers are the neuronal layers positioned between the input and output layers. Hidden Layers permit a neural network's function to be subdivided into distinct data processing. To prevent undertraining, each Hidden Layer is composed of eight neurons.

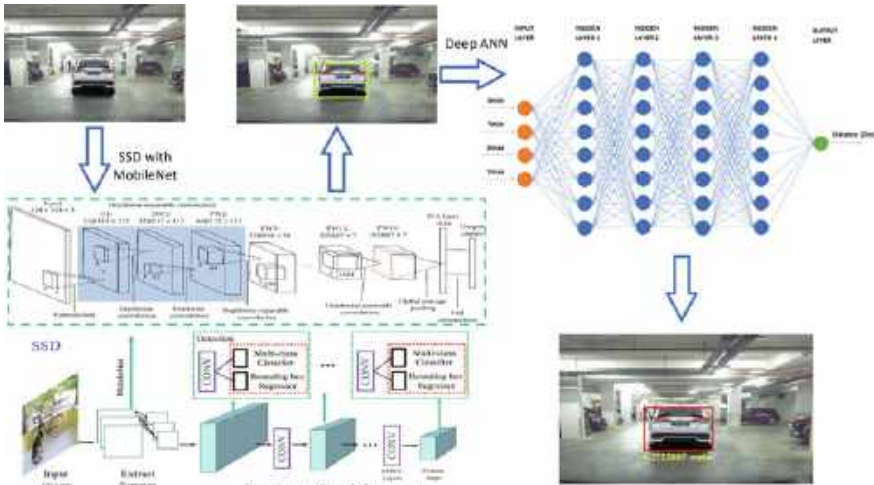


Fig. 1 A framework of monocular vision distance estimation method

This study used the LiDAR distances as the gold standard for determining whether or not the suggested approach was accurate. Cyberjaya, Malaysia, is the experiment’s location to collect the LiDAR distance data. Before gathering data, the camera and LiDAR must first be set up at the host vehicle, as illustrated in Fig. 2. Then, the host vehicle was driven around Cyberjaya to collect data. The data obtained includes an image of the target vehicle and the LiDAR distance. The picture of the target vehicle was obtained through the accessibility of a single camera positioned on the host vehicle. Following that, the photos were preprocessed using the proposed Distance Estimation method. Using SSD in conjunction with MobileNet architecture, the bounding box of the target vehicle was determined. After that, Deep ANN was used to estimate the distance between the camera and the target vehicle by utilizing the bounding box data obtained from the object detection. The outcome of the proposed method was compared with the LiDAR distance.

3 Results and Discussion

An image of the target vehicle is taken using a camera installed on the host vehicle, and the proposed method for estimating distance is applied to the photos taken. Figure 3 illustrates an example of the outcome of the proposed Distance Estimation method. The left image corresponds to a Car Type A, whereas the right image corresponds to a Car Type B at a LiDAR distance of 5 m. The yellow text in the figure indicates the distance estimation value the proposed method provides, while the red box contains information about detecting vehicles. Due to the importance of time in real-time applications, the proposed method only detects one car at a time.



Fig. 2 The position of the camera and LiDAR at the host vehicle



Fig. 3 The example of the collected data

In Fig. 4, the graph shows the relationship between the proposed Distance Estimation and the LiDAR distance. The distance from the proposed method is similar to the LiDAR distance in the 1–10 m range. On the other hand, the proposed method distance has an absolute error distance of ± 7 m compared with LiDAR distance between the range of 11–20 m. The image’s illumination exposure could be one of the factors contributing to the difference error obtained in the final analysis.

The outcome demonstrates that the performance of the proposed method can eventually replace LiDAR for determining the distance between vehicles. The proposed method yields excellent results and can accurately estimate distance. If the proposed method inaccurately estimates the distance, it will negatively impact the effectiveness of the AEB and AES in the vehicle’s system. Consequently, vehicle collisions are possible. The result also demonstrates that the proposed method for distance estimation has an accuracy of 92% compared with LiDAR distance. Nevertheless, the distance error for all types of cars is acceptable. An acceptable distance error of a few meters or less is usual for self-driving vehicles traveling at highway speeds. It allows the vehicle to avoid accidents with other vehicles or obstacles safely.

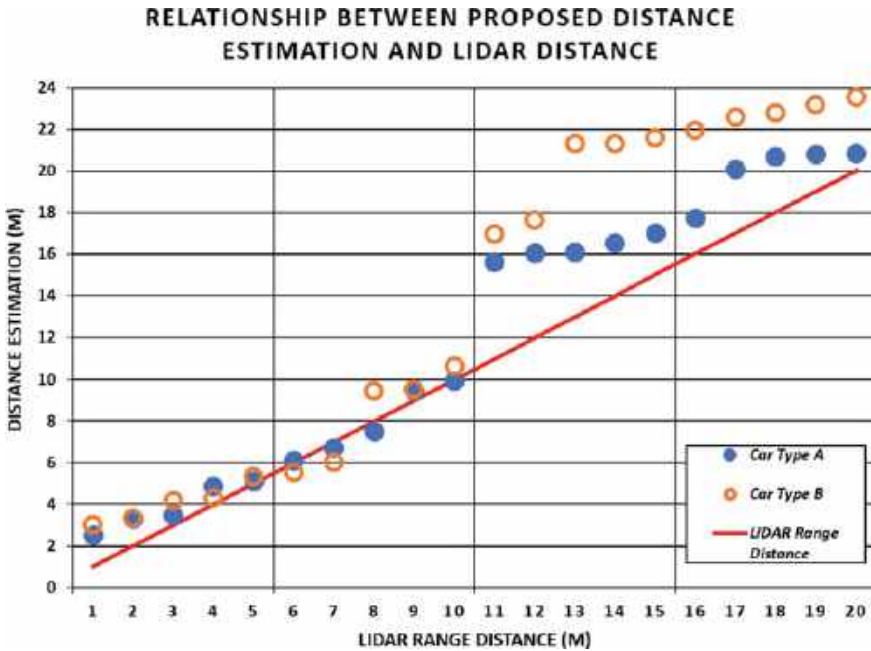


Fig. 4 The relationship between proposed distance estimation and LiDAR distance

4 Conclusion

This study presented a Monocular Vision Distance Estimation method employing SSD with MobileNet architecture for object detection and Deep ANN for distance estimation. This study utilizes a vision sensor instead of a radar system because it can have high-level processing when coupled with software. Using a vision sensor enables the segmentation of objects and the detection of desired objects. The experiment is then conducted to evaluate the accuracy of the proposed method for estimating the distance between the target vehicle and the camera with LiDAR. An image of the target vehicle is taken using a camera installed on the host vehicle, and the proposed method for estimating distance is applied to the photos taken. The result indicates that the proposed method for estimating distance has an accuracy of 92% compared to LiDAR distance. As a result, the findings of this research have the potential to contribute to the methodological foundation for further understanding drivers' behavior, with the ultimate objective of lowering the number of accidents involving rear-end crashes.

Acknowledgements This work was supported by the Ministry of Education Malaysia under Fundamental Research Grant Scheme (FRGS/1/2021/TK0/UTM/02/42).

References

1. Yang L, Yang Y, Wu G, Zhao X, Fang S, Liao X, Wang R, Zhang M (2022) A systematic review of autonomous emergency braking system: impact factor, technology, and performance evaluation. *J Adv Transp* 2022
2. Lowe E, Guvenç L (2021) A review of autonomous road vehicle integrated approaches to an emergency obstacle avoidance maneuver. arXiv preprint [arXiv:2105.09446](https://arxiv.org/abs/2105.09446)
3. Masoumian A, Marei DG, Abdulwahab S, Cristiano J, Puig D, Rashwan HA (2021) Absolute distance prediction based on deep learning object detection and monocular depth estimation models. Paper presented at the CCIA
4. Adz-Dzikri AA, Virgono A, Dirgantara FM (2021) Advance driving assistance systems: object detection and distance estimation using deep learning. Paper presented at the 2021 8th international conference on electrical engineering, computer science and informatics (EECSI)
5. Back M, Lee J, Bae K, Hwang SS, Chun IY (2021) Improved and efficient inter-vehicle distance estimation using road gradients of both ego and target vehicles. Paper presented at the 2021 IEEE international conference on autonomous systems (ICAS)
6. Karthika K, Adarsh S, Ramachandran K (2020) Distance estimation of preceding vehicle based on mono vision camera and artificial neural networks. Paper presented at the 2020 11th international conference on computing, communication and networking technologies (ICCCNT)
7. Vajgl M, Hurtik P, Nejezchleba T (2022) Dist-YOLO: fast object detection with distance estimation. *Appl Sci* 12(3):1354

Wood Defect Inspection on Dead Knots and Pinholes Using YOLOv5x Algorithm



Liew Pei Yi, Muhammad Firdaus Akbar, Bakhtiar Affendi Rosdi, Muhamad Faris Che Aminudin, and Mohd 'Akashah Fauthan

Abstract Accurate detection of wood defects is crucial for ensuring the quality and reliability of wood pieces in various industries such as construction and furniture production. Some challenging defects such as dead knots and pinholes vary in size and shape, complex textures, and the presence of wood grains makes the inspection process more complex. Thus, this paper evaluates the performance of the YOLOv5x algorithm in detecting and localizing wood defects, especially dead knots, and pinholes. Using an augmented custom dataset and trained with transfer learning from a pre-trained model with COCO dataset, the algorithm achieves outstanding results. With a precision score of 96.5%, a recall score of 91.8%, and a mAP@0.5 score of 95.5%, it indicates highly accurate defect detection and localization. However, the model's performance slightly dropped when considering mAP@0.5–0.95 which scored 66.1%, indicating challenges in detecting defects with higher IoU thresholds. Visual examples demonstrated the algorithm's capabilities as well as instances of incorrect detections and failed detections. The findings of this study can contribute to the field of wood inspection systems and highlight areas for further improvement in defect detection algorithms.

Keywords Dead knots · Pinholes · Transfer learning · Wood defect inspection · YOLOv5x algorithm

L. P. Yi · M. F. Akbar (✉) · B. A. Rosdi · M. F. C. Aminudin
School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Engineering Campus,
14300 Nibong Tebal, Penang, Malaysia
e-mail: firdaus.akbar@usm.my

M. 'A. Fauthan
Machinery Technology Center, SIRIM Bhd, 44200 Hulu Selangor, Malaysia

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
N. S. Ahmad et al. (eds.), *Proceedings of the 12th International Conference on Robotics, Vision, Signal Processing and Power Applications*, Lecture Notes in Electrical Engineering 1123, https://doi.org/10.1007/978-981-99-9005-4_74

589

1 Introduction

Wood may contain defects such as cracks, knots, wood grading, holes, resin pockets, and more [1]. These defects can result from the wood's natural growth, environmental factors, or the industries' pre-processing methods. They can reduce the strength and durability of wood, leading to potential safety issues in various industries such as the collapse of buildings or injuries caused by broken wood furniture. Thus, wood quality control is critical to ensure that wood products meet the required standards and specifications. Early detection of wood defects has become an important area of research today due to the significant impact of defects on wood quality.

Wood defect inspection has been an active area of research, employing various techniques including image processing methods, and machine learning techniques such as Decision Trees, Random Forests, Support Vector Machines (SVM) [2], Artificial Neural Networks (ANN) [3], K-Means Clustering [4], and more. However, these methods have limitations, including poor accuracy, human intervention, and sensitivity to lighting and environmental conditions.

Recent advancements in deep learning have shown great promise in terms of accuracy and performance on defect detection. In [5], a Deep Convolutional Neural Network (DCNN) with 16 trainable layers is used to train on the augmented wood images consisting of knots, cracks, and mildew stains. Dropout and regularization methods are applied during training, the Xavier initialization is used to initialize the model's weight. In [6], Mask R-CNN is used with a glance network to detect and classify the defects of wood veneer. In the combined network, Mask R-CNN provides the rectangular region proposals, deriving more complex feature maps on the features extracted from the glance network. In [7], a pre-trained ResNet model with 18 layers is used in wood knot detections by applying the transfer learning method. Although these algorithms offer high accuracy and performance in defect detection, the complexity of their architecture makes training and inference slower, requiring more computational resources.

To achieve a good balance between accuracy and speed, making a good wood defect inspection system with limited computational resources, a wood defect inspection system is proposed by utilizing the YOLOv5x algorithm. YOLO (You Only Look Once) is a state-of-the-art object detection algorithm with high accuracy, lightweight, real-time performance, and the ability to handle complex object detection tasks. By using its strength, this paper extends the existing research on wood defect detection by focusing specifically on detecting dead knots and pinholes. Inspecting both defects can be challenging due to the great variation in size and shape, from small pin-sized holes to large irregular knots. The rough and uneven wood surface texture, including the complexity of the wood grain, further complicates the inspection.

Therefore, a novel approach to wood defect inspection on dead knots and pinholes by utilizing YOLOv5x algorithms is proposed. Besides, this paper contributes a custom dataset that consists of various dead knots and pinholes images that capture the complexity and variability, augmented by using various augmentation methods.

Additionally, transfer learning is employed during model training to allow faster convergence and improve generalization performance. Furthermore, an evaluation of the proposed system is provided by the metrics such as precision, recall, and mean Average Precision (mAP). The inference of the model to new unseen data will be analyzed. The content of this paper is as follows: Sect. 2 explains the fundamental principles of the YOLOv5x algorithm; Sect. 3 illustrates the method used in the model training and validation; Sect. 4 provides the analysis of the results of the trained model and its performance. Finally, Sect. 5 gives the conclusion of the paper and future improvements.

2 Fundamental Principles of the YOLOv5x Algorithm

To address the challenge of detecting defects with various shapes and sizes, YOLOv5 uses anchor boxes, also known as prior boxes. Anchor boxes are the predefined bounding box templates of different scales and aspect ratios. They serve as reference templates for detecting objects of various sizes. These anchor boxes are present in different detection layers at varying scales. The layers closer to the input layer tend to capture the spatial information of larger objects, while the deeper layers have smaller receptive fields, enabling them to capture the information of smaller objects.

Figure 1 depicts the working principle of the YOLOv5x algorithm which involves dividing an image into grid cells, where each grid cell is responsible for detecting objects within it. During the training process, the algorithm generates predicted bounding boxes based on the anchor box templates and further refines these predictions.

In the figure, an object is presented in grid 1, bounded by a predicted bounding box generated by YOLOv5x as (b_x, b_y, b_w, b_h) . To evaluate the alignment between

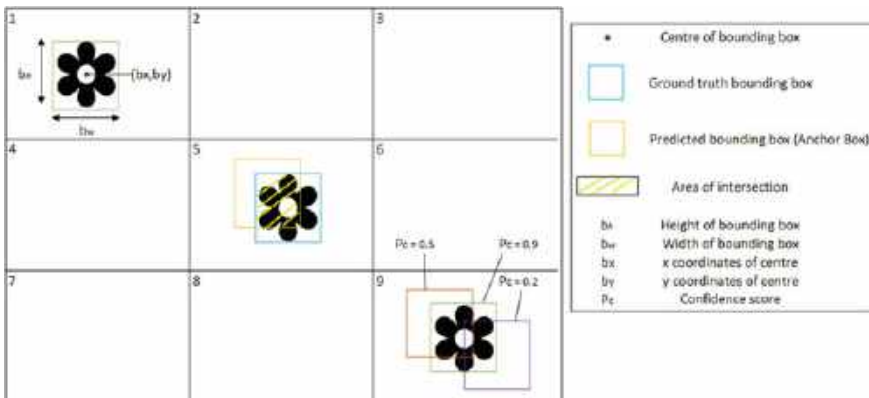


Fig. 1 An image with 3×3 grids

the predicted bounding box and the object, the Intersection over Union (IoU) is calculated using the following equation:

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Sum of Area of Predicted Bounding Box \& Ground Truth Bounding Box}} \quad (1)$$

Based on the IoU calculation, YOLOv5 selects the predicted bounding box with the highest IoU and assigned it as a positive detection. Moreover, YOLOv5 predicts the objectness score and class probabilities of predefined object classes that contributed to the confidence score of the detected object, denoted as P_c , which represents the overall confidence in the presence of an object and its prediction accuracy.

During the training process, multiple bounding boxes may be generated for an object, as shown in Grid 9. The detection layer applies the non-maximum suppression (NMS) technique to select the most confident and non-overlapping bounding box at the final object detection. The YOLOv5x algorithm's unique architecture which utilizes anchor boxes and features fusion efficiently, makes it well-suited for defect detection tasks.

3 Proposed Methodology

A methodology is proposed to enable faster training and better performance in the detection of dead knots and pinholes, which incorporates transfer learning. Firstly, a pre-trained model with COCO dataset, the YOLOv5x model, has been chosen. Using the initial weights of the pre-trained model as starting point, the model is further modified by replacing the classifier to match the number of classes ($nc = 2$). A custom wood dataset that consists of 891 dead knots and 1081 pinholes is used. The dataset is applied with data augmentation techniques to enhance data quality and variability, including image resizing, flipping, and rotation with clockwise and anticlockwise directions. Figure 2 shows the annotated dead knots and pinholes in the dataset.

To generalizes better and robust to new unseen data, augmentation techniques used during training are applied to the samples on the fly during the training process, including Hue, Saturation, Value (HSV) color space, rotation, translation, scaling,

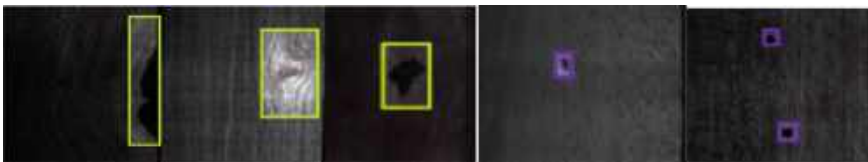


Fig. 2 Samples of the annotated dataset with dead knots and pinholes

flipping and mosaic augmentation. The augmented dataset is split into training, validation, and testing sets with a ratio of 6:3:1. The modified model is trained using the augmented data and optimized hyperparameters with input size 640, 300 epochs, batch size with 2, Stochastic Gradient Descent (SGD) as the optimizer and IoU threshold with 0.20. During training, the model's performance on the validation sets is monitored and adjustments are made to prevent overfitting. Once trained, the model's performance is evaluated on the validation dataset using metrics such as precision, recall, and mAP to assess its effectiveness in detecting defects. Finally, the trained model is applied to unseen images for inference, providing defect classification and localization for quality control and defect analysis in wood inspection systems.

4 Result and Discussion

To measure the performance of the trained YOLOv5x algorithm, the evaluation was conducted on the validation dataset. Table 1 presents the evaluation results. Firstly, considering all classes, the model achieved a precision of 0.965, indicating that 96.5% of the predicted defects were correct. The recall score of 0.918 suggests that the model identified 91.8% of the actual defects presented in the dataset. Additionally, mAP@0.5 is 0.955, indicating high accuracy in detecting and localizing defects. However, the mAP@0.5–0.95 is 0.661, suggesting that the model may have difficulty in accurately detecting defects with higher IoU thresholds, which could lead to some false negatives.

Figure 3 dictates the ability of the trained YOLOv5x model in detecting dead knots and pinholes in the new unseen wood images. Figure 3a, b shows that the trained model successfully detects the dead knots and pinholes in the wood samples. The bounding boxes tightly bound the identified defects and were able to distinguish the small defects from the surrounding wood grain. However, the model still occasionally makes incorrect detections as shown in Fig. 3c, where the model identifies a non-defective area as the dead knots. These false positives can occur due to lighting problems and complex backgrounds. Furthermore, Fig. 3d depicts the samples with failed detections, in which the model failed to detect a dead knot presented in the wood samples. This false negative indicates that the model may be challenging in detecting some defects due to their variation in sizes and orientation.

Table 1 Evaluation metrics of trained YOLOv5x

Class	Precision	Recall	mAP@0.5	mAP@0.5–0.95
All	0.965	0.918	0.955	0.661
Dead knots	0.959	0.984	0.938	0.683
Pinholes	0.971	0.932	0.972	0.639

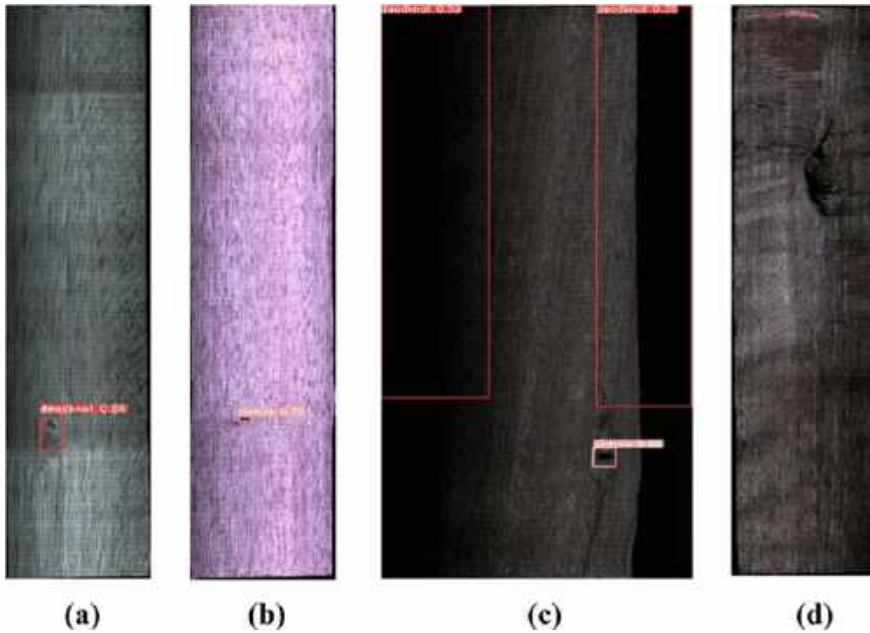


Fig. 3 Result of detection with **a** dead knot, **b** pinhole, **c** wrong detection, and **d** fail detection

Therefore, it is important to reduce the false negatives and false positives caused by the model. False negatives can result in compromised quality control in wood inspection systems, leading to financial losses or safety concerns. Additionally, false positives can result in unnecessary rejection and lead to a waste of resources and increase the time for manual inspection. However, the overall precision of the model is high (96.5%), indicating a low false positives rate. Therefore, further analysis and fine-tuning of the model are needed, including increasing the size of the dataset with more diverse samples, employing advanced data augmentation techniques, adjusting hyperparameters, and further training by using the trained model.

5 Conclusion

In conclusion, the trained YOLOv5x algorithm shows strong performance in detecting and localizing dead knots and pinholes with high scores of precision, recall, and mAP@0.5. However, it still faces the challenge of detecting the defects with higher IoU thresholds, leading to a lower score of mAP@0.5–0.95. The visual examples presented also highlight the capabilities and limitations of the trained model. To improve the model's performance, it is important to address the false negatives

and false positives by considering the dataset diversity, advanced data augmentation techniques, adjustments of hyperparameters, and further training with transfer learning to ensure higher quality control in the related fields.

Acknowledgements This work was supported by Universiti Sains Malaysia (USM), Bridging GRA Grant with Project No: 304/PELECT/6316607.

References

1. Kryl M, Danys L, Jaros R, Martinek R, Kodytek P, Bilik P (2020) Wood recognition and quality imaging inspection systems. *J Sens* 2020. <https://doi.org/10.1155/2020/3217126>
2. Deng ZY, Wang YZ, Zhang HR (2020) Detection method of wood skin defects based on bag-of-words model. In: ACM international conference proceeding series. Association for computing machinery, Oct 2020, pp 125–130. <https://doi.org/10.1145/3438872.3439068>
3. Chun TH et al (2021) Identification of wood defect using pattern recognition technique. *Int J Adv Intell Inform* 7(2):163–176. <https://doi.org/10.26555/ijain.v7i2.588>
4. Zhang Y, Jiang D, Zhang Z, Chen J (2022) Three-dimensional inversion of knot defects recognition in timber cutting. *J For Res*. <https://doi.org/10.1007/s11676-022-01532-y>
5. He T, Liu Y, Yu Y, Zhao Q, Hu Z (2020) Application of deep convolutional neural network on feature extraction and detection of wood defects. *Measurement* 152. <https://doi.org/10.1016/j.measurement.2019.107357>
6. Hu K, Wang B, Shen Y, Guan J, Cai Y (2020) Defect identification method for poplar veneer based on progressive growing generated adversarial network and MASK R-CNN model
7. Gao M, Chen J, Mu H, Qi D (2021) A transfer residual neural network based on ResNet-34 for detection of wood knot defects. *Forests* 12(2):1–16. <https://doi.org/10.3390/f12020212>