

Lecture Notes in Electrical Engineering 988

Muhammad Amirul Abdullah ·

Ismail Mohd. Khairuddin · Ahmad Fakhri Ab. Nasir ·

Wan Hasbullah Mohd. Isa ·

Mohd. Azraai Mohd. Razman ·

Mohd. Azri Hizami Rasid ·

Sheikh Muhammad Hafiz Fahami Zainal ·

Barry Bentley · Pengcheng Liu *Editors*

Advances in Intelligent Manufacturing and Mechatronics

Selected Articles from the Innovative
Manufacturing, Mechatronics &
Materials Forum (iM3F 2022), Pahang,
Malaysia

Lecture Notes in Electrical Engineering

Volume 988

Series Editors

Leopoldo Angrisani, Department of Electrical and Information Technologies Engineering, University of Napoli Federico II, Naples, Italy

Marco Arteaga, Departament de Control y Robótica, Universidad Nacional Autónoma de México, Coyoacán, Mexico

Bijaya Ketan Panigrahi, Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, Delhi, India

Samarjit Chakraborty, Fakultät für Elektrotechnik und Informationstechnik, TU München, Munich, Germany

Jiming Chen, Zhejiang University, Hangzhou, Zhejiang, China

Shanben Chen, 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, Humanoids and Intelligent Systems Laboratory, Karlsruhe Institute for Technology, Karlsruhe, Germany

Haibin Duan, Beijing University of Aeronautics and Astronautics, Beijing, China

Gianluigi Ferrari, Università di Parma, Parma, Italy

Manuel Ferre, Centre for Automation and Robotics CAR (UPM-CSIC), Universidad Politécnica de Madrid, Madrid, Spain

Sandra Hirche, Department of Electrical Engineering and Information Science, Technische Universität München, Munich, Germany

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, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Alaa Khamis, German University in Egypt El Tagamoa El Khames, New Cairo City, Egypt

Torsten Kroeger, Stanford University, Stanford, CA, USA

Yong Li, 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

Sebastian Möller, Quality and Usability Laboratory, TU Berlin, Berlin, Germany

Subhas Mukhopadhyay, School of Engineering and Advanced Technology, Massey University,

Palmerston North, Manawatu-Wanganui, New Zealand

Cun-Zheng Ning, Electrical Engineering, Arizona State University, Tempe, AZ, USA

Toyoaki Nishida, Graduate School of Informatics, Kyoto University, Kyoto, Japan

Luca Oneto, Department of Informatics, BioEngineering, Robotics and Systems Engineering, University of Genova, Genova, Genova, Italy

Federica Pascucci, Dipartimento di Ingegneria, Università degli Studi "Roma Tre", Rome, Italy

Yong Qin, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China

Gan Woon Seng, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore

Joachim Speidel, Institute of Telecommunications, Universität Stuttgart, Stuttgart, Germany

Germano Veiga, Campus da FEUP, INESC Porto, Porto, Portugal

Haitao Wu, Academy of Opto-electronics, Chinese Academy of Sciences, Beijing, China

Walter Zamboni, DIEM—Università degli studi di Salerno, Fisciano, Salerno, Italy

Junjie James Zhang, Charlotte, NC, USA

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. ****

Muhammad Amirul Abdullah ·
Ismail Mohd. Khairuddin ·
Ahmad Fakhri Ab. Nasir ·
Wan Hasbullah Mohd. Isa ·
Mohd. Azraai Mohd. Razman ·
Mohd. Azri Hizami Rasid ·
Sheikh Muhammad Hafiz Fahami Zainal ·
Barry Bentley · Pengcheng Liu
Editors

Advances in Intelligent Manufacturing and Mechatronics

Selected Articles from the Innovative
Manufacturing, Mechatronics & Materials
Forum (iM3F 2022), Pahang, Malaysia

 Springer

Editors

Muhammad Amirul Abdullah
Faculty of Manufacturing and Mechatronic
Engineering Technology
Universiti Malaysia Pahang
Pekan, Malaysia

Ismail Mohd. Khairuddin
Faculty of Manufacturing and Mechatronic
Engineering Technology
Universiti Malaysia Pahang
Pekan, Malaysia

Ahmad Fakhri Ab. Nasir
Faculty of Computing
Universiti Malaysia Pahang
Pekan, Malaysia

Wan Hasbullah Mohd. Isa
Faculty of Manufacturing and Mechatronic
Engineering Technology
Universiti Malaysia Pahang
Pekan, Malaysia

Mohd. Azraai Mohd. Razman
Faculty of Manufacturing and Mechatronic
Engineering Technology
Universiti Malaysia Pahang
Pekan, Malaysia

Mohd. Azri Hizami Rasid
Faculty of Manufacturing and Mechatronic
Engineering Technology
Universiti Malaysia Pahang
Pekan, Malaysia

Sheikh Muhammad Hafiz Fahami Zainal
Faculty of Manufacturing and Mechatronic
Engineering Technology
Universiti Malaysia Pahang
Pekan, Malaysia

Barry Bentley
Department of Computer Science
Cardiff Metropolitan University
Cardiff, UK

Pengcheng Liu
Department of Computer Science
University of York
York, UK

ISSN 1876-1100

ISSN 1876-1119 (electronic)

Lecture Notes in Electrical Engineering

ISBN 978-981-19-8702-1

ISBN 978-981-19-8703-8 (eBook)

<https://doi.org/10.1007/978-981-19-8703-8>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

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

Preface

The third edition forum of The Innovative Manufacturing, Mechatronics & Materials Forum 2022 (iM3F 2022) organized by Universiti Malaysia Pahang through its Faculty of Manufacturing and Mechatronic Engineering Technology was held on 20 July 2022. The main field focuses on manufacturing, mechatronics as well as materials.

More than 141 submissions were received during iM3F 2022 and were reviewed in a single-blind manner, and 30 papers were advocated by the reviewers to be published in this Lecture Notes in Electrical Engineering. The editors would like to express their gratitude to all the authors who submitted their papers. The paper published in this proceeding has been thoroughly reviewed by the appointed technical review committee consists of various experts in the field of mechatronics engineering.

The conference had brought a new outlook on cutting-edge issues shared through keynote speeches by Assoc. Prof. Ir. Dr. Faiz Mohd Turan, Prof. Dr. Hasbullah Idris and Dr. Barry Bentley.

Finally, the editors hope that readers find this volume informative as we thank LNEE for undertaking this volume publication. We also would like to thank the conference organization staff and the International Program Committees' members for their hard work.

Pekan, Pahang, Malaysia
November 2022

Muhammad Amirul Abdullah
Ismail Mohd. Khairuddin
Ahmad Fakhri Ab. Nasir
Wan Hasbullah Mohd. Isa
Mohd. Azraai Mohd. Razman
Mohd. Azri Hizami Rasid
Sheikh Muhammad Hafiz Fahami
Zainal
Barry Bentley
Pengcheng Liu

Contents

A Computational Time Analysis of Discrete Simulated Kalman Filter Optimizer	1
Suhazri Amrin Rahmad, Zuwairie Ibrahim, and Zulkifli Md Yusof	
A Real-Time Social Distancing and Face Mask Detection System Using Deep Learning	13
Suet Nam Wai, Sew Sun Tiang, Wei Hong Lim, and Koon Meng Ang	
A Systematic Review for Robotic for Cognitive Speech Therapy for Rehabilitation Patient	23
Junbo Qi, Esyin Chew, and Jiaji Yang	
An Estimation Steering Feedback Torque in Vehicle Steer by Wire System	39
S. M. H. Fahami, Faiz Mohd Turan, and M. A. Zakaria	
An Implementation of Sliding Mode Voltage Control Controlled Buck-Boost Converter for Solar Application	53
Nursabrina Athirah Mohd Mustakin, Mohd. Shafie Bakar, and Mazyah Mat Noh	
An Optimized Deep Learning Model for Automatic Diagnosis of COVID-19 Using Chest X-Ray Images	61
Suhaim Parvez Wadekar, Koon Meng Ang, Nor Ashidi Mat Isa, Sew Sun Tiang, Li Sze Chow, Chin Hong Wong, Meng Choung Chiong, and Wei Hong Lim	
Automatic Vehicle Location (AVL): Evaluation on the Punctuality Index of City Public Bus Service	75
Haziman Zakaria, Diyana Kamarudin, Faiz Azizul, Mohammad Fitri Idrus, Nor Rokiah Hanum Md Haron, and Norhana Mohd Aripin	

Bearing Fault Diagnosis Using Extreme Learning Machine Based on Artificial Gorilla Troops Optimizer	87
M. Firdaus Isham, M. S. R. Saufi, M. D. A. Hasan, W. A. A. Saad, M. Salman Leong, M. H. Lim, and Z. A. B. Ahmad	
Classifying Ethnicity of the Pedestrian Using Skin Colour Palette	105
Syahmi Syahiran Ahmad Ridzuan, Zaid Omar, and Usman Ullah Sheikh	
Cluster Analysis Based on Image Feature Extraction for Automated OMA	117
Muhammad Danial Bin Abu Hasan, Syahril Ramadhan Saufi, M. Firdaus Isham, Shaharil Mad Saad, W. Aliff A. Saad, Zair Asrar Bin Ahmad, Mohd Salman Leong, Lim Meng Hee, and M. Haffizzi Md. Idris	
Detection of Lead with IoT Water Monitoring System Using Microstrip Antenna-Based Sensor	127
Abelle Chin Tze Hui, Sew Sun Tiang, Kah Hou Teng, Wei Hong Lim, and Mastaneh Mokayef	
Emotion Recognition Using Ultra-Short-Term ECG Signals with a Hybrid Convolutional Neural Network and Long Short-Term Memory Network	139
Vui Chee Chang, Jee-Hou Ho, Bee Ting Chan, and Ai Bao Chai	
Enhancement of Morlet Mother Wavelet in Time-Frequency Domain in Electroencephalogram (EEG) Signals for Driver Fatigue Classification	151
Rafiuddin Abdubrani, Mahfuzah Mustafa, and Zarith Liyana Zahari	
Fabrication of Aneurysm Biomodel Using 3D Printing Technology	163
Jamil Ahmad Hisam, Muhamad Yusof Salehudin, Muhammad Ismail Mat Lizah, Muhammad Izzat Ahmad Suhaimi, Muhammad Haqim Muhammad Hisham, Ismayuzri Ishak, and Mohd Jamil Mohamed Mokhtarudin	
Feature Selection of Medical Dataset Using African Vultures Optimization Algorithm	175
Wy-Liang Cheng, Koon Meng Ang, Sew Sun Tiang, Kah Yung Yap, Li Pan, Chin Hong Wong, Mahmud Iwan Solihin, and Wei Hong Lim	
Flow Direction Algorithm for Feature Selection	187
Wy-Liang Cheng, Koon Meng Ang, Wei Hong Lim, Sew Sun Tiang, Meng Choung Chiong, Chun Kit Ang, Li Pan, and Chin Hong Wong	

Fuzzy Logic Controller by Particle Swarm Optimization Discoverer for Semi-Active Suspension System 199
 Mat Hussin Ab Talib, Nur Hafiezul Mohd. Rosli, Intan Zaurah Mat Darus, Hanim Mohd. Yatim, Muhamad Sukri Hadi, Mohd. Ibthisham Ardani, Mohd. Syahril Ramadhan Mohd. Saufi, and Ahmad Hafizal Mohd. Yamin

Optimized Machine Learning Model with Modified Particle Swarm Optimization for Data Classification 211
 Kah Sheng Lim, Koon Meng Ang, Nor Ashidi Mat Isa, Sew Sun Tiang, Hameedur Rahman, Balaji Chandrasekar, Eryana Eiyada Hussin, and Wei Hong Lim

Performance Comparison of Kalman Filter and Extended Kalman Filter for Human Tracking and Prediction with Particle Swarm Optimisation 225
 Abiodun Afis Ajasa and Nawawi Sophan Wahyudi

Stability and Bifurcation Analysis of Rössler System in Fractional Order 239
 Ibrahim Mohammed Sulaiman, Abiodun Ezekiel Owoyemi, Mohamad Arif Awang Nawi, Sadiya Salisu Muhammad, U. R. Muhammad, Ali Fareed Jameel, and Mohd Kamal Mohd Nawawi

SUAS-Based NDVI and RGB Image for Remote Landscape and Environmental Monitoring on University Campus 251
 Ahmad Anas Yusof, Mohd Faid Yahya, Mohd Khairi Mohamed Nor, and Muhammad Fahmi Miskon

A Mathematical Model of PD Controller-Based DC Motor System Using System Identification Approach 263
 Nur Naajihah Ab Rahman and Nafrizuan Mat Yahya

The Classification of Wafer Defects: An Evaluation of Different Feature-Based ResNet Transfer Learning Models with Support Vector Machine 277
 Lim Shi Xuen, Ismail Mohd Khairuddin, Mohd Azraai Mohd Razman, Jessnor Arif Mat Jizat, Edmund Yuen, Eng Hwa Yap, Andrew Huey Ping Tan, and Anwar P. P. Abdul Majeed

The Correlation Between Peltier Module, Solution Volume and Temperature in IoT-Controlled Hydroponic Nutrient Solution Management 285
 Hamdan Sulaiman, Ahmad Anas Yusof, and Mohd Khairi Mohamed Nor

The Statistical Impact of Artificial Intelligence Towards the Price Change of Financial Instrument 293
 Lim Guo Huang, Choong Kah Wei, Nor Aziyatul Izni, Loh Yue Fang, Tan Sher Lyn, and Sarah Atifah Saruchi

Total Harmonic Distortion Study for Improvement of AC-AC Converter Under Buck-Type 305
Mohd. Shafie Bakar, Nurul Amira Ibrahim, and Abu Zaharin Ahmad

Training Feedforward Neural Networks Using Arithmetic Optimization Algorithm for Medical Classification 313
Koon Meng Ang, Wei Hong Lim, Sew Sun Tiang, Hameedur Rahman, Chun Kit Ang, Elango Natarajan, Mohamed Khan Afthab Ahamed Khan, and Li Pan

Various Type of Crops and Trees Detection Using Clustering Technique Through Image Processing 325
Mohd Izzat Mohd Rahman, Mohd Azraai Mohd Razman, Ismail Mohd Khairuddin, Anwar P. P. Abdul Majeed, Muhammad Amirul Abdullah, and Wan Hasbullah Mohd Isa

Transient Pressure Analysis in Water Hydraulics Machine Using Induced Pressure Effect from the Compression of Different Materials 333
Ahmad Anas Yusof, Suhaimi Misha, Faizil Wasbari, Mohamed Hafiz Bin Md Isa, Mohd Qadafie Ibrahim, and Mohd Shahir Kasim

X-Ray Baggage Object Detection Using Neural Networks Approach for Safety Purpose 341
Samuel Ato Gyasi Otabir, Sew Sun Tiang, Wei Hong Lim, Hung Yang Leong, and Bo Sun

A Computational Time Analysis of Discrete Simulated Kalman Filter Optimizer



Suhazri Amrin Rahmad , Zuwairie Ibrahim , and Zulkifli Md Yusof 

Abstract Simulated Kalman filter (SKF) is a population-based optimization algorithm based on the Kalman filter framework. To find the global optimum, the SKF applies a Kalman filter process that involves prediction, measurement, and estimation. However, the SKF can only operate in numerical search space. In literature, many techniques and modifications have been made to the SKF algorithms to function in a discrete search space. An example of the modified SKF is the discrete simulated Kalman filter optimizer (DSKFO). However, little research has been conducted on the DSKFO. This paper studies the computational time complexity of the DSKFO to acquire a better understanding of the algorithm's complexity. The analysis is done by comparing the computational time of the DSKFO against four combinatorial SKFs. The findings show that the DSKFO is the fastest algorithm for solving all TSP instances. The DSKFO requires just 13 s to solve the smaller TSP instance *eil51*, whereas SEDESKF, BSKF, DESKF, and AMSKF need 14, 34, 36, and 42 s, respectively. DSKFO solves the larger TSP instance *dsj1000* in 79 s, whereas SEDESKF, BSKF, DESKF, and AMSKF need 182, 1104, 1125, and 1167 s, respectively.

Keywords Combinatorial · Simulated Kalman filter · Computational analysis

S. A. Rahmad (✉) · Z. Ibrahim
College of Engineering, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang,
Kuantan, Pahang, Malaysia
e-mail: mek18007@stdmail.ump.edu.my

Z. M. Yusof
Faculty of Manufacturing and Mechatronic Engineering Technology, Universiti Malaysia Pahang,
26600 Pekan, Pahang, Malaysia

1 Introduction

Combinatorial optimization is a branch of optimization problems that has applications in a variety of fields. It is frequently utilized in a wide range of areas, including applied mathematics, artificial intelligence, computer science, and electronic engineering. An example of a combinatorial optimization problem is a travelling salesman problem (TSP).

The travelling salesman problem (TSP) is a well-known combinatorial optimization routing problem. It has piqued academics interest because it is both simple to comprehend and difficult to solve. The TSP can be stated as follows: A salesman begins his or her journey in one city before moving on to the next set of cities. The objective of TSP is to determine the shortest and most cost-effective path travelled by the travelling salesman.

Many metaheuristic algorithms, such as genetic algorithm (GA) [1], ant colony optimization (ACO) [2], and simulated annealing (SA) [3], have been proposed to tackle combinatorial problems throughout the last decades. A numerical optimization algorithm is one that operates in the numerical search space. The algorithm must be modified or additional computations must be performed in order to operate in discrete search space.

A numerical search space is a set of all feasible solutions in which the variables are all real numbers, whereas a discrete search space is a set of all feasible solutions in which the variables are all integers. For example, an objective function, $f(x_1, x_2)$ contains two variables, x_1 and x_2 , with an interval of $[0, 4]$. Thus, a feasible solution of the objective function in numerical search space is illustrated in Fig. 1a, where the variables can be any set of real numbers. On the other hand, a feasible solution in discrete search space is shown in Fig. 1b, where the variables consist of integers only.

The simulated Kalman filter (SKF) [4, 5] is an optimization algorithm originally introduced for numerical optimization problems. The SKF has been improved by the addition of a computation for the purpose of solving combinatorial problems. Few existing combinatorial algorithm that are developed based on the SKF are binary SKF (BSKF) [6], distance evaluated SKF (DESKF) [7], angle modulated SKF (AMSKF) [8], and state encoded distance evaluated SKF (SEDESKF) [9].

Recently, a new discrete variant of the SKF called the discrete simulated Kalman filter optimizer (DSKFO) [10] has been introduced. The algorithm provides impressive experimental results in the literature. However, it is currently uncertain how effectively the computational time of the algorithm can scale as the problem size increases. In this paper, the computational analysis of the DSKFO algorithm is conducted by comparing the runtime of the DSKFO against the runtime of four existing combinatorial SKFs: the BSKF, DESKF, AMSKF, and SEDESKF.

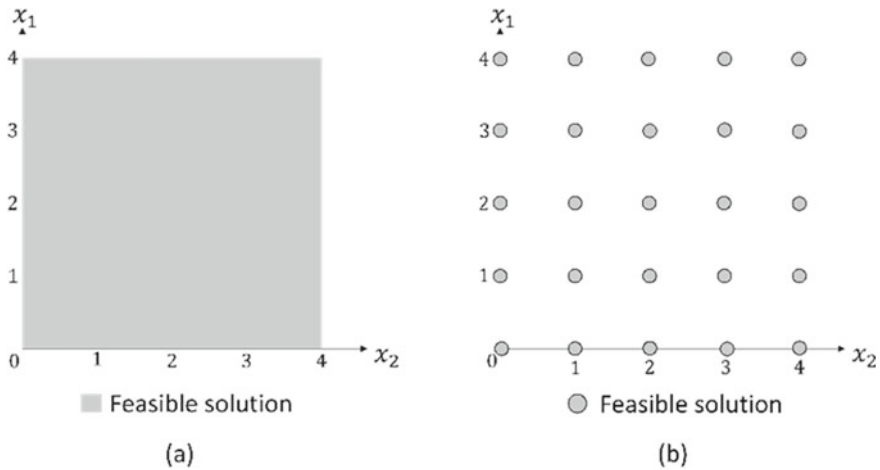


Fig. 1 Example of **a** numerical search space and **b** discrete search space

2 Methodology

2.1 Discrete Simulated Kalman Filter Optimizer (DSKFO)

The discrete simulated Kalman filter optimizer (DSKFO) algorithm is illustrated in Fig. 2. The algorithm begins by assigning a random sequence to N agents. The number of iterations is represented by t and the stopping condition for the algorithm is set at the maximum number of iterations, t_{\max} . The initial value of the error covariance estimate, $P(0)$, the process noise, Q , and the measurement noise, R , all of which required for Kalman gain calculation are also set, where $\{P, Q, R\} \in \mathbb{R}$. Each solution is comprised of D -dimensional vector. The first dimension as well as all other dimensions in a solution is in the form of state, in which each state is subset of all states. Each state may consist of integer number ranging from 1 until D . In other word, a state vector, X , can be expressed as $X = \{x_1, x_2, \dots, x_D\}$, $\{x, D\} \in \mathbb{Z}$, $x \in [1, D]$. The state vector, X , of the i th agent at time t in the DSKFO is shown as (1).

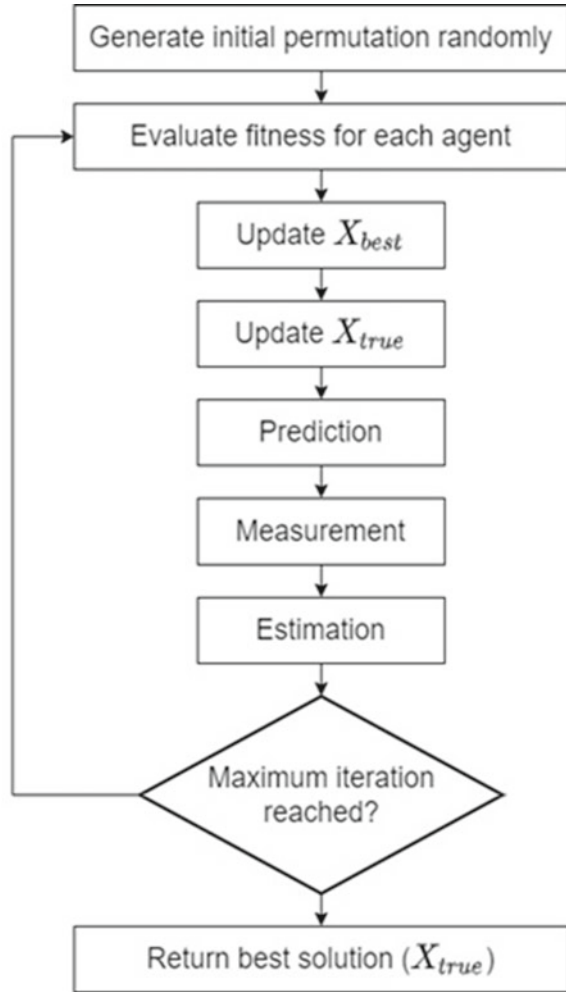
$$X_i(t) = \{x_i^1(t), x_i^2(t), \dots, x_i^d(t), \dots, x_i^D(t)\} \quad (1)$$

After that, each agent is put through an evaluation to determine their fitness value. The fitness values are compared, and the agent with the best fitness value at each iteration is set as $X_{\text{best}}(t)$. For minimization problem,

$$X_{\text{best}}(t) = \min_{i \in \{1, \dots, n\}} \text{fit}_i(X(t)) \quad (2)$$

and for maximization problem,

Fig. 2 Discrete simulated Kalman filter optimizer (DSKFO) algorithm



$$X_{best}(t) = \max_{i \in \{1, \dots, n\}} \text{fit}_i(X(t)) \quad (3)$$

The best solution obtained so far is called X_{true} . If a better fitness value is discovered, it will be taken as the value of X_{true} . The X_{true} is updated if the $X_{best}(t)$ is better than the X_{true} depending on the type of problem being evaluated ($X_{best}(t) < X_{true}$ for minimization problem, or $X_{best}(t) > X_{true}$ for maximization problem).

In the prediction stage, the time-update equations are computed as follows:

$$X_i(t|t+1) = X_i(t) \quad (4)$$

$$P(t|t+1) = P(t) + Q \quad (5)$$

where $X_i(t|t+1)$ and $X_i(t)$ represent the predicted state and the current state, respectively, and $P(t|t+1)$ and $P(t)$ are the predicted error covariant estimate and the current error covariant estimate, respectively.

The next step is measurement. In DSKFO, a substitution mutation utilized by [11] is used in the measurement step. The procedure for the measurement step is shown in Pseudocode 1.

Pseudocode 1. Procedure for Mutation in the Measurement Step

```

01: for each agent,  $i$ 
02:   for each dimension,  $d$ 
03:     generate  $rand$  with range of  $[0,1]$ 
04:     if  $rand > 0.5$ 
05:        $Z_i^d(t) = X_i^d(t|t+1)$ 
06:     else
07:        $Z_i^d(t) = X_{true}^d$ 
08:     end
09:   end
10: end

```

The final step is estimation. A substitution mutation mechanism from [11] is also used in this step. The Kalman gain, $K(t)$, is computed as follows:

$$K(t) = \frac{P(t|t+1)}{P(t|t+1) + R} \quad (6)$$

The Kalman gain is a weight assigned to the measurements and the current state estimation. High gain places more weight towards the measurement and lower gain follows more closely to the prediction. The measurement residual is another element that influences the mutation during the estimate step. The measurement residual is the difference between the measurement, $Z_i(t)$, and the predicted state, $X_i(t|t+1)$. In DSKFO, the Hamming distance is used to compute the difference between these two variables. The Hamming distance is then converted into a range of $[0, 1]$ using following equation.

$$y_i(t) = \frac{H(Z_i(t), X_i(t|t+1))}{D} \quad (7)$$

where $y_i(t)$ is the measurement residual as $y_i(t) \in [0, 1]$, and the Hamming distance between the measurement and predicted state is represented as $H(Z_i(t), X_i(t|t+1))$. Next, the measurement residual, $y_i(t)$, is multiplied by the Kalman gain, $K(t)$, to produce a correction, $\hat{K}(t)$, for the predicted state. The $\hat{K}(t)$ value will determine the probability of mutation in each dimension. High $\hat{K}(t)$ value leads more dimensions to take a value from measurement, $Z_i(t)$, whereas low $\hat{K}(t)$

value allows more dimension takes a value from predicted state, $X_i(t|t+1)$. Mutations are then occurred based on correction, $\widehat{K}(t)$ as shown in Pseudocode 2 to produce the estimated states for following iteration, $X_i(t+1)$.

Pseudocode 2. Procedure for Mutation in the Estimation Step

```

01: for each agent,  $i$ 
02:   for each dimension,  $d$ 
03:     generate  $rand$  with range of  $[0,1]$ 
04:     if  $rand > \widehat{K}(t)$ 
05:        $X_i^d(t+1) = X_i^d(t|t+1)$ 
06:     else
07:        $X_i^d(t+1) = Z_i^d(t)$ 
08:     end
09:   end
10: end

```

In early iteration, the algorithm promotes exploration process as the mutation in estimation step occurs in many dimensions. As the iteration progresses, the Hamming distance between the measurement value and the predicted state decreases, lowering the chance of mutation in each dimension in the estimation step. The reduction of chosen dimensions for mutation in estimation step causes the algorithm to proceed to the exploitation process. After that, the estimated error covariance for the following iteration, $P(t+1)$ is computed as follows:

$$P(t+1) = (1 - K(t))P(t|t+1) \quad (8)$$

Finally, the procedures are performed for the following iteration until the maximum number of iterations is achieved.

3 Experiment, Result, and Discussion

A TSP benchmark set consists of 47 TSP instances that are used to evaluate the computational time of the algorithms. The computational time of the DSKFO is compared against four existing combinatorial simulated Kalman filter (SKF) algorithms, which are binary SKF (BSKF), distance evaluated SKF (DESKF), angle modulated SKF (AMSKF), and DESKF with state encoded (SEDESKF).

Table 1 shows the experimental parameter settings of the algorithms for solving the TSP. All algorithms use the same value for every parameter to provide fair comparison for the experiment. The computational analysis of the algorithms is then assessed based on the runtime values of 1 trial in minimizing the total distance of the TSP.

Table 2 shows the fitness value obtained by the algorithms for solving the TSP. Note that the algorithm with best fitness value for each instance is bolded. Based on

Table 1 Experimental parameter settings

Parameter	DSKFO	BSKF	AMSKF	DESKF	SEDESKF
Iteration	1000	1000	1000	1000	1000
Number of agents	30	30	30	30	30
Number of trials	1	1	1	1	1
Initial error covariance estimate	1000	1000	1000	1000	1000
Process noise	0.5	0.5	0.5	0.5	0.5
Measurement noise	0.5	0.5	0.5	0.5	0.5

the table, the DSKFO outperformed other combinatorial SKF algorithms for solving every TSP instances.

The comparison between the runtime of DSKFO against the BSKF, DESKF, AMSKF, and SEDESKF is shown in Table 3. Based on the findings, the DSKFO performs the fastest despite having more steps according to Pseudocode 6. A clear justification for the outperformance of the DSKFO and SEDESKF compared to the BSKF, DESKF, and AMSKF is because of the input types. The BSKF, DESKF, and AMSKF represent its solution in binary, whereas the DSKFO and SEDESKF represent its solution as a state. A binary input causes the algorithm to generate more dimensions, resulting in a greater number of operational steps.

4 Conclusion

This paper investigates the computational time required by the DSKFO algorithm for solving the travelling salesman problem (TSP). The algorithm's computational complexity is determined by comparing its execution time to that of four combinatorial SKFs. According to the findings, the DSKFO performs the quickest compared to the BSKF, AMSKF, DESKF, and SEDESKF. The SEDESKF ranks second, followed by the BSKF, the DESKF, and the AMSKF. Further research of the DSKFO can be considered for future studies.

Table 2 Performance of the algorithms for solving TSP

Instance	DESKF	AMSKF	BSKF	DSKFO	SEDESKF
berlin52	22,932.2	22,874.86	22,847.64	19,033.89	22,406.98
bier127	544,106.7	544,059.5	542,440	489,010.7	536,858.9
ch130	39,254.37	39,357.7	39,267	35,321.95	39,426.15
ch150	46,270.79	46,168.05	46,174.03	41,839.67	46,136.63
d198	1,645,013	1,646,428	1,648,227	1,597,967	1,640,368
d493	157,618.5	158,018.6	158,476.9	136,081.7	143,879.8
d657	411,998.9	411,931.2	411,621	380,277.4	405,901.5
d1291	796,175.3	796,174.9	796,929.4	757,082	794,117
dsj1000	5.24E + 08	5.23E + 08	5.24E + 08	5.02E + 08	5.20E + 08
eil51	2845.66	2856.43	2853.754	2555.159	2840.658
eil76	1268.42	1266.809	2127.613	1102.025	1267.804
eil101	2052.86	2039.967	23,782.28	1822.935	2043.157
gil262	23,846.46	23,851.59	23,853.9	21,832.33	23,718.7
kroA100	137,043	136,954.9	137,188.7	120,715.6	135,675.3
kroA150	216,442.1	215,813.7	215,796.9	194,157	214,278.5
kroA200	291,940.4	291,098.8	291,063.8	272,885.1	289,059.9
kroB100	134,923.4	134,818.2	134,786.5	128,634	133,147.7
kroB200	285,802.7	285,558.9	286,095.5	248,305.7	283,920
kroC100	135,469.5	135,858.8	135,539.3	119,472	133,605.2
kroD100	131,622.3	131,561.2	131,396.8	118,427	130,181.9
kroE100	138,503.9	137,716.4	138,610.7	120,180.4	136,381.9
lin105	99,036.19	98,766.64	99,045.13	85,650.8	98,295.8
lin318	527,049.5	528,817.1	529,112.7	497,228.2	527,431.1
p654	1,845,492	1,848,103	1,849,637	1,722,673	1,835,950
pcb442	1,333,055	1,335,124	1,335,923	1,309,015	1,332,669
pcb1173	708,486.4	707,728.3	708,016.9	671,691.3	706,661.8
pr76	6,085,013	6,078,577	6,079,543	5,854,900	6,070,399
pr107	446,386.8	446,571.5	449,263.3	399,598.7	438,474.9
pr124	580,257.8	573,148.5	579,691.2	493,549.5	572,756.4
pr136	690,108.3	689,959.7	689,880.4	623,854.7	689,707.8
pr144	682,605.3	686,191.1	682,410.8	591,661.7	679,453.9
pr152	886,217.2	886,369	886,457.3	777,614.8	880,010
pr226	1,479,082	1,477,167	1,482,490	1,319,486	1,472,488
pr264	954,199	954,069.8	958,776.8	881,561.7	945,954.9
pr299	664,537	667,263.2	666,494.6	616,103.6	663,592.6
pr439	1,737,005	1,732,577	1,731,523	1,623,659	1,714,496

(continued)

Table 2 (continued)

Instance	DESKF	AMSKF	BSKF	DSKFO	SEDESKF
pr1002	461,023.3	461,176	461,949.7	406,375.4	459,240.7
rat99	19,422.96	19,441.65	19,461.39	17,846.2	19,329.74
rat195	103,909.3	104,311.9	104,248	99,569.35	103,555.9
rat575	166,512.9	167,018.8	166,983	158,673.6	165,840
rat783	6696.176	6718.733	6732.771	5939.013	6632.39
rd100	46,096.3	45,664.31	45,944.33	40,210.75	45,651.72
rl1304	8,917,743	8,908,134	8,916,299	8,625,869	8,880,536
rl1323	9,303,794	9,303,447	9,302,486	9,050,985	9,275,698
rl1889	14,171,974	14,159,546	14,157,634	13,702,898	14,114,830
st70	2882.996	2902.092	2890.875	2401.752	2887.399
ts225	1,411,955	1,410,333	1,409,169	1,285,080	1,409,269

Table 3 Runtime of the algorithms

Instance	DESKF (s)	AMSKF (s)	BSKF (s)	DSKFO (s)	SEDESKF (s)
berlin52	43	50	41	15	16
bier127	109	130	108	21	28
ch130	115	134	113	24	29
ch150	121	154	126	24	31
d198	168	193	164	27	39
d493	457	505	449	46	79
d657	683	729	668	55	104
d1291	1569	1584	1565	102	216
dsj1000	1125	1167	1104	79	182
eil51	36	42	34	13	14
eil76	66	76	63	17	20
eil101	83	105	77	18	22
gil262	239	275	235	29	47
kroA100	81	95	79	20	24
kroA150	129	150	126	23	32
kroA200	165	191	161	26	38
kroB100	82	95	78	20	22
kroB200	164	190	160	26	36
kroC100	81	94	78	19	23
kroD100	82	95	80	20	24

(continued)

Table 3 (continued)

Instance	DESKF (s)	AMSKF (s)	BSKF (s)	DSKFO (s)	SEDESKF (s)
kroE100	80	93	77	19	23
lin105	86	100	83	21	23
lin318	288	322	284	32	53
p654	666	720	659	54	102
pcb442	405	453	399	40	73
pcb1173	1393	1395	1365	91	189
pr76	66	76	64	18	20
pr107	87	104	86	21	25
pr124	96	113	94	22	27
pr136	117	136	116	24	30
pr144	122	141	118	22	27
pr152	130	150	126	23	28
pr226	185	216	183	27	43
pr264	238	275	235	30	47
pr299	269	306	265	32	52
pr439	398	446	396	39	72
pr1002	1066	1118	1051	86	155
rat99	81	95	78	20	23
rat195	162	188	160	25	37
rat575	581	633	577	50	91
rat783	810	861	804	63	122
rd100	82	95	79	20	22
rl1304	1565	1570	1556	98	213
rl1323	1587	1590	1574	101	208
rl1889	2536	2537	2418	139	344
st70	61	71	59	18	18
ts225	186	218	185	27	42

Acknowledgements The authors would like to thank the Ministry of Higher Education for providing financial support under Fundamental Research Grant Scheme (FRGS) No. FRGS/1/2018/TK04/UMP/02/9 (University reference RDU190176).

References

1. Holland JH (1992) Genetic algorithms. *Sci Am* 66–72
2. Colomi A, Dorigo M, Maniezzo V (1992) An investigation of some properties of an ‘Ant algorithm.’ *Ppsn* 92, 509–520

3. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220(4598):671–680
4. Ibrahim Z, Abdul Aziz NH, Nor NA, Razali S, Mohamad MS (2016) Simulated Kalman filter: a novel estimation-based metaheuristic optimization algorithm. *Adv Sci Lett* 22(10):2941–2946. <https://doi.org/10.1166/asl.2016.7083>
5. Ibrahim Z et al (2015) A Kalman filter approach for solving unimodal optimization problems. *ICIC Express Lett* 9(12):3415–3422
6. Yusof ZM, Ibrahim I, Satiman SN, Ibrahim Z, Aziz NHA, Aziz NAA (2015) BSKF: simulated Kalman filter. In: Proceedings of the—AIMS 2015, 3rd international conference on artificial intelligence, modelling and simulation, pp 77–81. <https://doi.org/10.1109/AIMS.2015.23>
7. Yusof ZM et al (2016) Distance evaluated simulated Kalman filter for combinatorial optimization problems. *ARNP J Eng Appl Sci* 11(7):4911–4916
8. Yusof ZM et al (2016) Angle modulated simulated Kalman filter algorithm for combinatorial optimization problems. *ARNP J Eng Appl Sci* 11(7):4854–4859
9. Yusof ZM et al (2018) Distance evaluated simulated kalman filter with state encoding for combinatorial optimization problems. *Int J Eng Technol* 7(4):22–29. <https://doi.org/10.14419/ijet.v7i4.27.22431>
10. Rahmad SA, Ibrahim Z, Md Yusof Z (2022) Simulated Kalman filter with modified measurement, substitution mutation and hamming distance calculation for solving traveling salesman problem. In: Enabling industry 4.0 through advances mechatronics. Lecture notes in electrical engineering, vol 900, pp 309–320
11. Ab Rahman T, Ibrahim Z, Ab Aziz NA, Zhao S, Abdul Aziz NH (2018) Single-agent finite impulse response optimizer for numerical optimization problems. *IEEE Access* 6(c):9358–9374. <https://doi.org/10.1109/ACCESS.2017.2777894>

A Real-Time Social Distancing and Face Mask Detection System Using Deep Learning



Suet Nam Wai, Sew Sun Tiang, Wei Hong Lim, and Koon Meng Ang

Abstract It has been more than two years since the transmission of COVID-19 virus has affected the public health globally. Due to its natural characteristic, the virus is very likely to undergo mutation over time and consistently changes to a new variant with higher severity and transmission rate. The pandemic is expected to prolong with the increment in number of daily cases which leads to why preventive measures like practising distance apart rule and wearing facemask are still mandatory in the long run. This paper is prepared to develop a social distancing model using deep learning for COVID-19 pandemic. The tracking accuracy of the proposed model is discussed in the paper and compared with other deep learning methods as well. The efficiency of the detection model is observed and evaluated by performing quantitative metrics. The monitoring model is trained by implementing YOLOv4 algorithm and has achieved an accuracy of 93.79% with F1-score of 0.87 in detecting person and facemask. The model is applicable for real-time and video detection to monitor social distance violation as an effort to flatten the curve and slow down the transmission rate in the community.

Keywords Deep learning · Social distancing · Mask detection · YOLOv4

1 Introduction

Even though vaccines for COVID-19 are now available worldwide to fight against the pandemic, the fundamentals of preventive measures are still highly anticipated. Vaccination is just an additional step in reducing the severity effect of the disease and death. The extend of how much it can protect a person from the infection and transmitting the virus to others is still unknown [1]. The term, social distancing, can be described as a public health practice that limits any in-person contact with anyone by staying at home and away from public spaces to reduce the airborne transmission

S. N. Wai · S. S. Tiang (✉) · W. H. Lim · K. M. Ang
Faculty of Engineering, Technology and Built Environment, UCSI University, 56000 Kuala Lumpur, Malaysia
e-mail: tiangss@ucsiuniversity.edu.my

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
M. A. Abdullah et al. (eds.), *Advances in Intelligent Manufacturing and Mechatronics*,
Lecture Notes in Electrical Engineering 988,
https://doi.org/10.1007/978-981-19-8703-8_2

[2, 3]. Ainslie et al. revealed that the number of new cases dropped significantly during the imposition of strict social distancing and movement restrictions towards mainland China and Hong Kong SAR from late January to early February 2020 [4]. Prem et al. had studied the effectiveness of physical distancing in Wuhan whereby it decreased the median number of infections by more than 92% in middle of 2020 and 24% at the end of 2020 [5]. Fong et al. and Kahalé had also verified that social distancing is an effective preventive measure in combatting the pandemic [6, 7]. On the other side, deep learning is a universal learning approach that can perform in almost all application domains in cases where humans do not have to be present in the scene to conduct the specific task. It can be defined as a subset of machine learning that uses neural networks with many layers and is introduced to mimic the function of the human brain in data processing [8], object detection [9, 10], and fault detection [11, 12]. It has been evolving for the past decades with improvised algorithms to produce higher accuracy percentage and generate data concurrent with the present situations. Developing a social distancing monitoring model using deep learning can contribute to slowing down the virus transmission rate that is affecting the public health by identifying social distance violation through person detection.

2 Related Work

A summary of other similar works in using deep learning for object detection to monitor the practice of preventive measures is shown in Table 1.

Uddin et al. [13] used ResNet50 as the CNN architecture to develop an intelligent model that categorized people based on body temperature which resulted in person tracking accuracy at 84%. Saponara et al. [14] applied YOLOv2 to monitor social distance and body temperature through thermal camera using two different datasets and achieved accuracy detection of 95.6% and 94.5%, respectively. Pun et al. [15] utilized YOLOv3 framework with the addition of Deepsort approach that can track the identified people by assigning them with unique IDs. The proposed model had

Table 1 Comparison of quantitative analysis data based on different social distancing models

Paper	Methods	Dataset size	Accuracy (%)	Precision	Recall	F1-Score
[13]	CNN (ResNet50)	11 880	84	0.84	0.82	0.82
[14]	YOLOv2	775	95.6	0.95	0.96	0.95
		800	94.5	0.94	0.95	0.94
[15]	YOLOv3 and deepsort	800	84.6	–	–	–
[16]	YOLOv3 with transfer learning	–	95	0.86	0.83	0.84
[17]	YOLOv4	7 363	97.84	0.85	0.97	0.91
[18]	YOLOv4	3.76M	99.8	0.998	0.976	0.99

84.6% accuracy. Ahmed et al. [16] proposed his model to detect human from overhead perspective by implementing YOLOv3 adopted with transfer learning which in return achieving 95% accuracy. Rahim et al. [17] developed a social distancing monitoring model specifically for low-light environment targeting night-time using YOLOv4 algorithm. Despite the limitation of having the proposed model to focus in the environment temporarily before monitoring, the accuracy result was 97.84%. Razaeei and Azarmi [18] aimed to have a viewpoint-independent human classification algorithm to monitor social distancing that can overcome limitation of light condition and challenging environment without needing to consider the angle and position of the camera. Their proposed model was built on YOLOv4 algorithm and obtained an accuracy of 99.8%.

3 Methodology

3.1 Dataset Preparation

A total of 530 images are collected randomly from various online sources shown in Google Images as well as selectively from raw images published by X. zhangyang's GitHub [19] and Prajnash's GitHub [20]. These images are taken with people from all ages and gender in different situations like walking, standing, sitting, and other possible body positions to maximize the stimulated conditions for detecting person with and without facemask. The dataset consists of both closed-up and distant images with 200 images focusing on single person with mask only, 160 images focussing on single person without mask only, and 170 images mix with a group of people with and without mask. They are pre-processed by resizing and orienting to establish a base size and orientation to be fed into the framework. This helps in improving the quality and consistency of the data for feature extraction as shown in Fig. 1.

3.2 Model Training

The model training process is conducted via Google Colab to utilize their GPU acceleration for extra computational power. The reason behind choosing YOLOv4 algorithm rather than other deep learning methods is that it is the only framework that can run in a conventional GPU that is easily accessible with minimal cost from home. Besides, its performance in speed and accuracy has been proven with astonishing outcomes, and it suits the real-time application for the proposed model [21]. The network size for model training is 416×416 . The hyperparameters, which cannot be inferred by the model, are set such that momentum is configured to 0.949, weight decay is configured to 0.0005, and the learning rate is at 0.001. The classification model is trained to predict three classes namely person, with mask, and without mask.

3.3 Performance Evaluation

Quantitative metrics are the measurements of how robust the model is and act as a form of feedback to determine which aspects of the model can be improved. Since the proposed model focuses on classification performance, the metrics used for performance evaluation are precision, recall, F1-score, mean average precision (mAP), and intersection over union (IoU).

Precision is used to measure the ratio of true positives (TP) to the total positives predicted as expressed in Eq. (1). It is more on how many predictions did the model capture correctly.

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

Recall, also known as sensitivity, is used to measure the ratio of TP to the actual number of positives as expressed in Eq. (2). It is more on how many predictions did the model miss.

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

Both precision and recall can be represented in a single score called F1-score. It takes the harmonic mean of those two metrics as expressed in Eq. (3).

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Meanwhile, average precision (AP) is the result of the area under the precision–recall curve and can be calculated using Eq. (4). This is where mean average precision (mAP) comes into the picture to calculate the average of AP for all the classes as shown in Eq. (5).

$$\text{AP} = \frac{1}{11} \sum_{\text{Recall}_i} \text{Precision}(\text{Recall}_i) \quad (4)$$

$$\text{mAP} = \frac{1}{N} \times \sum_{i=1}^N \text{AP}_i \quad (5)$$

3.4 Deployment of Classifier Model

The model deployment is conducted in PyCharm Community Edition 2021.2.1. The overview workflow of the classifier model can be seen in Fig. 2. The model is

begun by reading the input video and converting it into frames. The ability of object detector is then applied to classify three classes based on the confidence value. If the predicted object is a person, the model would proceed with evaluating the inter-distance measurement. If the predicted object is with mask, purple bounding box is generated with “mask” text labelled on top of it. If the predicted object is without mask, red bounding box is generated with “no mask” text labelled on top of it. A mini dashboard is updated at the top left corner of the output video to show the monitoring status according to the number of bounding boxes generated per frame.

The inter-distance calculation is performed by measuring the distance between the centre point of every bounding box of predicted person. The centroid coordinate of the bounding box can be obtained by adding the lowest and highest value of the same axis and divide them by two as expressed in Eq. (6). C_i , which is also equivalent to (X_i, Y_i) , represents the centroid coordinate. X_{\min} and X_{\max} are the lowest and highest x-coordination of the bounding box, respectively. Likewise, Y_{\min} and Y_{\max} are the lowest and highest y-coordination of the bounding box, respectively.

$$C_i = (X_i, Y_i) = \left(\frac{X_{\min} + X_{\max}}{2}, \frac{Y_{\min} + Y_{\max}}{2} \right) \quad (6)$$

After that, Euclidean distance criterion is applied here to translate the distance between the pixels in the input frame to metric distance format. The equation of Euclidean formula is shown in Eq. (7). The distance between two centroid points of the bounding boxes is represented as $D(C_1, C_2)$. X_{\max} and Y_{\max} represent the coordinates from either one of the centroid points that has the largest value. X_{\min} and Y_{\min} represent the coordinates from the other the centroid point that has the smallest value.

$$D(C_1, C_2) = \sqrt{(X_{\max} - X_{\min})^2 + (Y_{\max} - Y_{\min})^2} \quad (7)$$



Fig. 1 Samples of images for dataset preparation

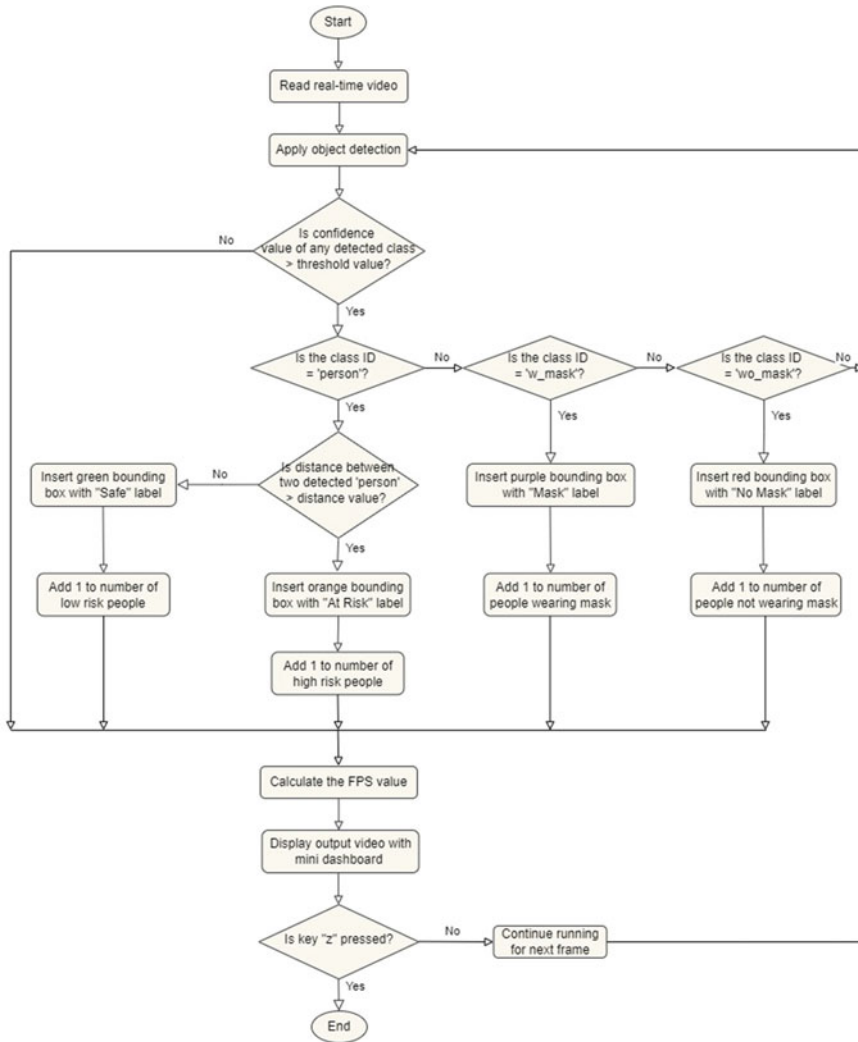


Fig. 2 Flowchart of social distancing monitoring model

Initially, the bounding boxes will not be drawn first when they are detected. Once the inter-distance calculation is computed, the model will decide whether the bounding boxes will be in green or red. The violation distance is denoted as the violation threshold value in this case. If $D(C_1, C_2)$ is more than or equal to the violation threshold value, then the bounding boxes will be drawn in green with the text “safe” as the label on top of them. If $D(C_1, C_2)$ is smaller than the violation threshold value, the bounding boxes be drawn in green with the text “at risk” as the

Table 2 Comparison analysis between three different training models

	YOLOv4	YOLOV3	YOLOv2
mAP@0.50 (%)	93.79	93.07	92.29
Precision	0.82	0.85	0.79
Recall	0.94	0.91	0.93
F1-score	0.87	0.88	0.85

label on top of them. This process is repeated in loop for every frame in real-time video.

4 Results and Discussion

4.1 Quantitative Analysis of Deep Learning Methods

In this work, three different deep learning models are pre-trained with the same dataset and hyperparameters for comparisons. The labelled images are split into 80% of training set and 20% of testing set to measure the robustness of the models.

Based on the quantitative metrics tabulated in Table 2, it is analysed that YOLOv2 model has the lowest performance out of the three training models. On the contrary, the overall robustness of both YOLOv4 and YOLOv3 models are quite similar as their precedence is the other's flaw and vice versa. YOLOv4 model has the upper hand in terms of accuracy and recall whereas YOLOv3 model has the upper hand in terms of precision and F1-score. After some considerations, YOLOv4 model is selected to be deployed as the classifier in the proposed social distancing monitoring model due to having the highest accuracy detection of 93.79% when compared to the other two models. It also has the best sensitivity in not missing out any true positives with recall value at 0.94 and a fair F1-score at 0.87.

4.2 Performance of Social Distancing Monitoring Model

The experiment is done at public areas that have potential widespread of COVID-19 transmission. Hence, the videos are captured from three different cases. Figure 3a, b shows the results at one of the rest stops beside Lebuhraya Utara-Selatan in Perak as an example of open space area. The second case is aimed at enclosed space, for example, like Mid Valley Megamall, and the results are shown in Fig. 3c, d. The third case is aimed at public semi-enclosed place like KL Sentral Transit Hub as shown in Fig. 3e, f.

By referring to the output frames in Fig. 3, it is observed that the overall result of object classification and localization is executed well towards detecting objects that

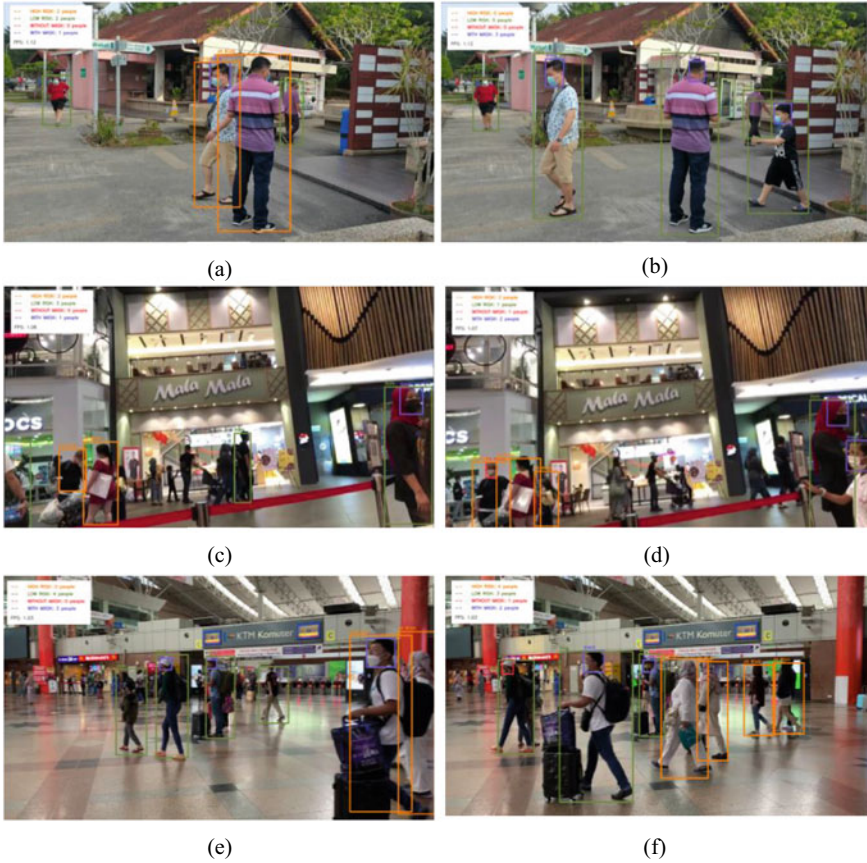


Fig. 3 Visualization of classification and localization as well as monitoring social distancing

are close to the camera. Besides, the monitoring of social distance violation is well performed as expected and the mini dashboard is updated correctly for every single frame according to the number of bounding boxes generated. It can be interpreted that camera position should be taken into considerations as the monitoring performance is able to execute better when the camera is positioned at eye level rather than at lower angle assuming at sitting position level. The performance of social-distance monitoring model, in terms of number of high risks, number of low risks, number of individuals without mask, and number of individuals with mask, is summarized in Table 3.

Table 3 Performance of social-distance monitoring model at Perak Rest Stop, Mid Valley Megamall, and KL Sentral Transit Hub

Venues	Perak Rest Stop		Mid Valley Megamall		KL Sentral Transit Hub	
	Figure 3a	Figure 3b	Figure 3c	Figure 3d	Figure 3e	Figure 3f
# High risks	2	0	2	3	2	4
# Low risks	2	5	3	1	4	3
# Without mask	0	0	0	1	0	1
# With mask	1	3	1	2	3	2

5 Conclusion

The development of social distancing monitoring model using deep learning and the analysis of the model performance are covered in this paper. The effectiveness of social distancing is studied before building the model to understand better in relation to the objective of the project. The process of model training using YOLOv4 method is discussed so that the proposed model can work with real-time and video detection. As a result, the model has achieved accuracy detection of 93.79% and F1-score of 0.87. In terms of deployment performance, it is shown that the object classification and localization as well as the evaluation of social distance violation are executed well towards predicted objects that are close to the camera at eye level position. The outcome of the social distancing monitoring model can be implemented in situations where public health is emphasized corresponds to the practice of preventive measures during COVID-19 pandemic. An additional feature of facemask detection is included too in an effort to mitigate the transmission rate of airborne virus in public places. Nevertheless, improvements can be made in future work to detect a wider range of the crowds since the proposed model only works with objects that are close to the camera.

Acknowledgements This work was supported by the Ministry of Higher Education Malaysia under the Fundamental Research Schemes with project codes of Proj-FRGS/1/2019/TK04/UCSI/02/1 and the UCSI University Research Excellence & Innovation Grant (REIG) with project code of REIG-FETBE-2022/038.

References

1. COVID-19 Vaccines Advice. <http://www.who.int/emergencies/diseases/novel-coronavirus-2019/covid-19-vaccines/advice>. Accessed 22 July 2021
2. What is social distancing and how can it slow the spread of COVID-19? | Hub, Mar. <http://hub.jhu.edu/2020/03/13/what-is-social-distancing/>. Accessed 27 July 2021
3. Coronavirus, Social and Physical Distancing and Self-Quarantine | Johns Hopkins Medicine. <http://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/coronavirus-social-distancing-and-self-quarantine>. Accessed 27 July 2021

4. Ainslie KEC et al (2020) Evidence of initial success for China exiting COVID-19 social distancing policy after achieving containment. *Wellcome Open Res* 2020 5(5):81
5. Prem K, Liu Y, Russell TW, Kucharski AJ, Eggo RM, Davies N (2020) The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *Lancet Public Health* 5(5):e261–e270
6. Fong MW, Gao H, Wong JY, Xiao J, Shiu EYC, Ryu S, Cowling BJ (2020) Nonpharmaceutical measures for pandemic influenza in nonhealthcare settings—social distancing measures. *Emerg Infect Dis* 26(5):976
7. Kahl   N (2020) On the economic impact of social distancing measures. SSRN Electron J
8. Jdid B, Lim WH, Dayoub I, Hassan Kais, Rizon M (2021) Robust automatic modulation recognition through joint contribution of Hand-crafted and conceptual features. *IEEE Access* (9):104530–104546
9. Voon YN, Ang KM, Chong YH, Lim WH, Tiang SS (2022) Computer-vision-based integrated circuit recognition using deep learning. In: Zain MZ et al (eds) *Proceedings of the 6th international conference on electrical, control and computer engineering, LNEE*, vol 842. Springer, Singapore, pp 913–925
10. Low JW, Tiang SS, Lim WH, Chong YH, Voon YN (2022) Tomato leaf health monitoring system with SSD and MobileNet. In: Zain MZ et al (eds) *Proceedings of the 6th international conference on electrical, control and computer engineering, LNEE*, vol 842. Springer, Singapore, pp 795–804
11. Alrifay M, Lim WH, Ang CK, Natarajan E, Solihin MI, Rizon M, Tiang SS (2022) Hybrid deep learning model for fault detection and classification of grid-connected photovoltaic system. *IEEE Access* 10:13852–13869
12. Alrifay M, Lim WH, Ang CK (2021) A novel deep learning framework based RNN-SAE for fault detection of electrical gas generator. *IEEE Access* 9:21433–21442
13. Uddin MI, Shah SAA, Al-Khasawneh MA (2020) A novel deep convolutional neural network model to monitor people following guidelines to avoid COVID-19. *J Sens*
14. Saponara S, Elhanashi A, Gagliardi A (2021) Implementing a real-time, AI-based, people detection and social distancing measuring system for Covid-19. *J Real-Time Image Process* 1–11
15. COVID-19 social distancing with person detection and tracking via fine-tuned YOLO v3 and Deepsort techniques, <https://arxiv.org/abs/2005.01385v4>. Accessed 27 July 2021
16. Ahmed I, Ahmad M, Rodrigues JJPC, Jeon G, Din S (2021) A deep learning-based social distance monitoring framework for COVID-19. *Sustain Cities Soc* 65:102571
17. Rahim A, Maqbool A, Rana T (2021) Monitoring social distancing under various low light conditions with deep learning and a single motionless time of flight camera. *PLoS ONE* 16(2):e0247440
18. Rezaei M, Azarmi M (2020) DeepSOCIAL: social distancing monitoring and infection risk assessment in COVID-19 pandemic. *Appl Sci* 10(21):7514
19. GitHub—X-zhangyang/Real-World-Masked-Face-Dataset: real-World Masked Face Dataset, 口罩人脸数据集, <http://github.com/X-zhangyang/Real-World-Masked-Face-Dataset>. Accessed 24 Feb 2022
20. GitHub—Prajnasb/observations. <http://github.com/prajnasb/observations>. Accessed 24 Feb 2022
21. YOLOv4: Optimal Speed and Accuracy of Object Detection. <http://arxiv.org/abs/2004.10934v1>. Accessed 24 Feb 2022

A Systematic Review for Robotic for Cognitive Speech Therapy for Rehabilitation Patient



Junbo Qi , Esyin Chew , and Jiaji Yang 

Abstract Stroke is a disease with a very high disability rate in the world. In recent years, more and more young people have suffered from stroke. Aphasia is one of the more common complications. When patients suffer from aphasia, the brain's speech, cognitive, literacy, and comprehension skills decline. As the number of patients increases, the number of nurses is not enough, and patients have fewer and fewer opportunities to receive treatment. In addition, the cost of treatment prevents many patients from receiving timely treatment. Therefore, new and cheaper treatments are needed to improve this situation, which can cover more patients. Since surgical robots have been introduced into the surgical field to participate in the surgical process, there is no corresponding robot for rehabilitation training in aphasia. Through literature review, this project hopes to provide patients with speech training, cognitive training and understanding training through Sanbot ELF, and gradually restore the patient's language ability.

Keywords The robot of artificial intelligence · Stroke and aphasia · Aphasia rehabilitation · The robot of artificial intelligence in medical field

1 Introduction

1.1 Background

Stroke is considered as an emergency disease with a short onset time and requires immediate treatment. Stroke is a disease that can lead to death, and it has a high mortality rate all over the world. There are three main types of stroke. The first is hemorrhagic stroke. About 85% of stroke patients have this type of stroke. It can also be subdivided into two subgroups. The first is caused by a blood clot in an artery in the neck or brain. The second is caused by blood clots from the heart that reach the

J. Qi · E. Chew · J. Yang (✉)
Cardiff Metropolitan University, Western Avenue, Llandaff, Cardiff, UK
e-mail: JYang@cardiffmet.ac.uk

brain and block the flow of blood. The second type is ischemic stroke. Fifteen percent of stroke patients have it. It was caused by a brain hemorrhage. The third is a transient ischemic attack (TIA) [1]. The condition is mild, usually lasting only a few minutes and up to 24 h. There are about three main symptoms of a stroke. First, the stroke patient cannot control the muscles of the half face, causing the half face to have no expression. Second, a stroke patient has an arm that cannot be lifted and held. Third, stroke patients are unable to express themselves correctly, their speech is slurred, and they are unable to understand what others are saying. If someone is found to exhibit the above symptoms, then you need to contact the doctor immediately [2].

Complications of stroke include dyskinesia, cognitive impairment, and communication problems. This also led to the main research content of this project, using robots to help stroke patients to resume communication skills. Because communication barriers can be restored through rehabilitation training. Aphasia patients often have trouble understanding and expressing words, such as showing partial or complete loss of ability to speak, understand spoken language or gestures, read, calculate or write [3]. Aphasia in stroke patients can be recovered, and the golden period of treatment is between 3 and 6 months, so early treatment is more necessary. The training of aphasia is better at home. Because there is no interference from other factors, it can also be combined with daily life. And because of the different social and cultural backgrounds of stroke elderly people, one-on-one language rehabilitation training will be better. But there are not many families who have the conditions to care at home in real life. Although it sounds ruthless, it is impossible to let the family do nothing, just to accompany the patient at home. Because the family also has a job, it is impossible to stay with the patient, they can only do as much time as possible to accompany the patient. So, the rehabilitation center usually takes care of the training. But as more people have strokes, so does the workload of nurses. Therefore, if the robot is used to assist the nurse to provide the patient with some simple rehabilitation training, the burden on the nurse and family members can be greatly reduced.

Sanbot ELF was used in this project. Three robot is an intelligent service robot (humanoid robot) developed by qihan technology, which is showed in Fig. 1. The Sanbot ELF platform will unleash the power of cloud computing robots and artificial intelligence for hotels, retail, security, education, health care, and many other customer-oriented industries. Therefore, Sanbot ELF can provide more intelligent and personalized services. It has powerful functions such as motion interaction, voice interaction, and induction interaction. Sanbot ELF has rich semantic understanding and execution ability, autonomous walking and obstacle avoidance, automatic charging, face detection and recognition, autonomous sound source positioning, and other activities. Connected to the Internet, its powerful computing power and self-learning ability are more capable of super-complex tasks. In order to meet the needs of the medical field, the design team of Sanbot ELF input more than 80,000 pieces of health knowledge into the robot system. In addition, Sanbot ELF can also connect with hospitals, mobile phones of patients' families, and other platforms to provide more comprehensive nursing services [4].



Fig. 1 Photo of Sanbot ELF

1.2 Problem Definitions and Research Motivation

- (1) Will people accept and trust nursing robots to provide treatment for patients?
- (2) What kind of training should robots provide?
- (3) How to make training easy?

Stroke used to happen in older people, but data shows that many young people are suffering from stroke now. Four hundred children a year suffer strokes in the UK. So, people should pay more attention to stroke. Usually, if a family has a stroke patient, the family usually takes the patient to a hospital for rehabilitation, and there are some nurses in the hospital to take care of them. However, as the number of patients increases, the working pressure of nurses in hospitals also increases. Sometimes, nurses are too busy to take care of all patients at the same time. Therefore, if the robot is used to assist the nurse to provide the patient with some simple rehabilitation training, the burden on the nurse can be greatly reduced. So, the motivation for this program is to help nurses provide simple rehabilitation training, reduce their work stress, and help more stroke patients.

1.3 Aim and Objectives

Aims: The aim of the program is to help nurses at the rehabilitation center train stroke patients to recover from speech problems and to be able to communicate with people. And there are three objectives in this project.

- (1) To research, design, develop, and pilot training stroke patients with cognitive robotics to resume communication skills.

- (2) To propose a Rehab robotics model to address existing problem in cognitive and speech rehabilitation.

Objectives: In order to achieve these aims, four objectives were designed.

- (1) Through case study and literature analysis, find out the deficiencies and problems of the current robot application in stroke aphasia rehabilitation training.
- (2) Through case analysis, summarize people's attitudes toward the application of robots in the medical field and which factors are related, such as gender, age, interaction mode.

2 Literature Review

2.1 Critical Literature Review Search Method

The search for this project started in February 2019. Find valuable papers and materials at Cardiff Metropolitan University library and Google Scholar. Sources of papers include Scopus, IEEE, ACM, Taylor and Francis Journals, and more. Literature review has no restrictions on the time of publication of the paper. Therefore, it is possible to collect a wider range of treatments related to it.

Search result: When searching for robot of artificial intelligence in Scopus, IEEE, ACM, etc., the results are as follows. The IEEE has 16,551 results, as shown in Fig. 2. ACM has 557,190 results, as shown in Fig. 3. Scopus has 15,665 results, as shown in Fig. 4. Among them, in the Scopus results, the number of articles with this topic has increased dramatically since 2014, as shown in Fig. 5. And in these articles, the number of articles published in the USA is far ahead, the second is China, and the third is Japan, as shown in Fig. 6.

Among these results, there are re-selected articles related to stroke and aphasia rehabilitation, such as 1,650 articles in ACM. The article on stroke aphasia was published in the IEEE with 27 articles. There are 1310 articles in ACM. Scopus has 6,404 articles.

According to JBI Critical Appraisal Checklist for Systematic Reviews and Research Syntheses and my own Research direction [5], the following five criteria are summarized to help select articles. Finally, a total of 11 articles were selected in combination with all the keywords, and the selection process is shown in Fig. 7.

- (1) Whether the literature research is related to my research.
- (2) Have the research questions been solved?
- (3) Whether the paper was published too long ago and whether it played a guiding role in the current research.
- (4) If the study includes experiments, whether the experimental process is rigorous and whether the results have bias.
- (5) Whether there is evidence to support the conclusion that robots have positive or negative effects on improving aphasia.

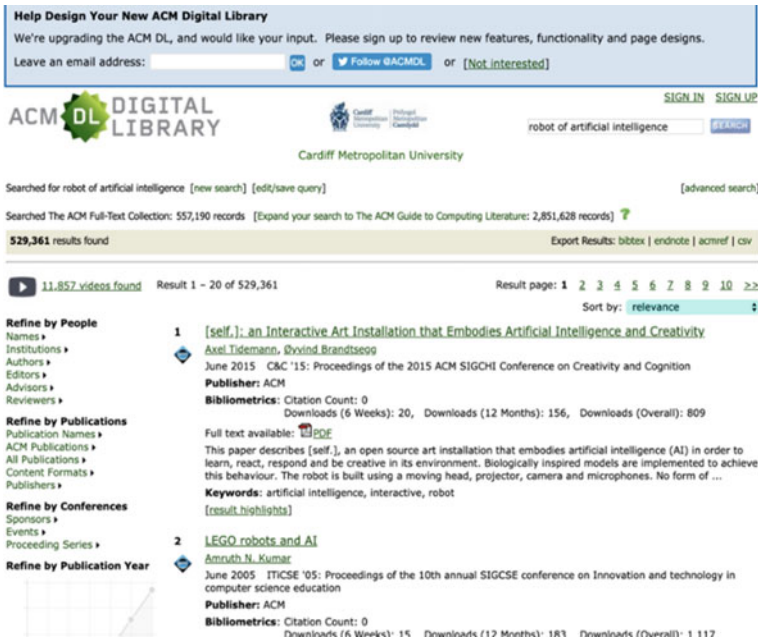


Fig. 2 Number of articles about robot of artificial intelligence in ACM

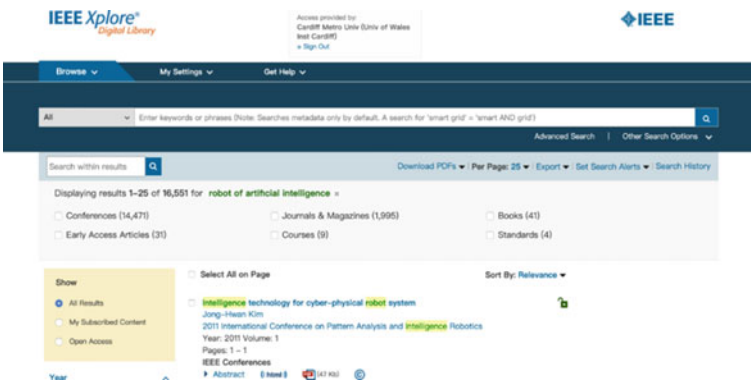


Fig. 3 Number of articles about robot of artificial intelligence in IEEE

2.2 Stroke and Aphasia

With the progress of science and technology and economy, people’s life quality has been greatly improved, and the number of the elderly is increasing. According to the data of Australian Bureau of Statistics, the population of Australia is growing rapidly, and the proportion of the elderly in the total population is increasing [6]. This

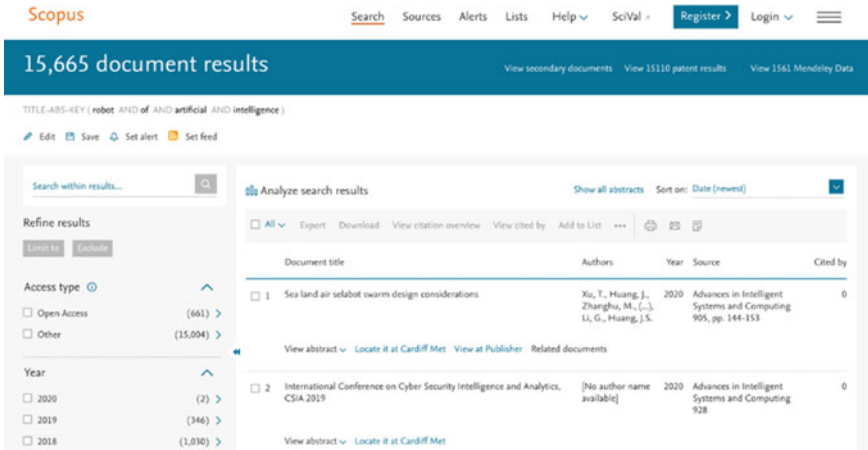


Fig. 4 Number of articles about robot of artificial intelligence in Scopus

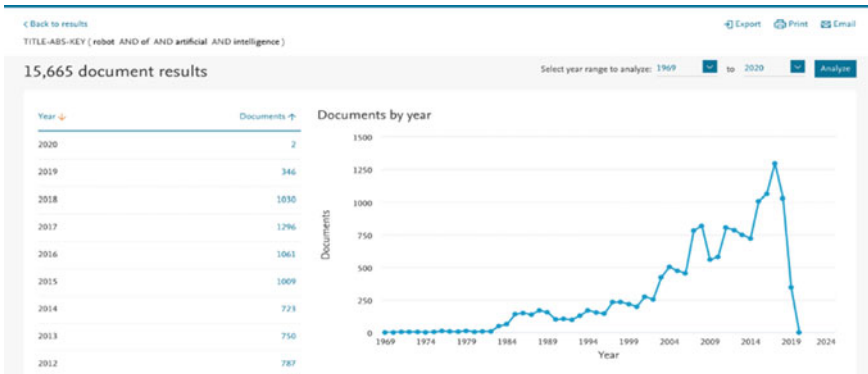


Fig. 5 Number of articles about robot of artificial intelligence in IEEE

is followed by increased rates of stroke and aphasia. Because stroke mostly occurs in the elderly, 59% of stroke patients in the UK occur in the elderly [7]. According to a survey by relevant institutions in the UK. Although the data are from Australia and the UK, this is not an isolated case, as the same happens in many countries around the world.

Aphasia is a disease that affects communication skills. It can influence both verbal and written communication skills as well as the ability to understand written content and what is heard. There are many causes of aphasia, such as brain injury, brain trauma, stroke, and brain tumor suppression. Stroke is the leading cause of aphasia, with approximately 25–40% of stroke patients suffering from aphasia [8].

Stroke is one of the leading causes of disability in adults. The British Stroke Association found that stroke was the biggest cause of disability in adults, affecting some

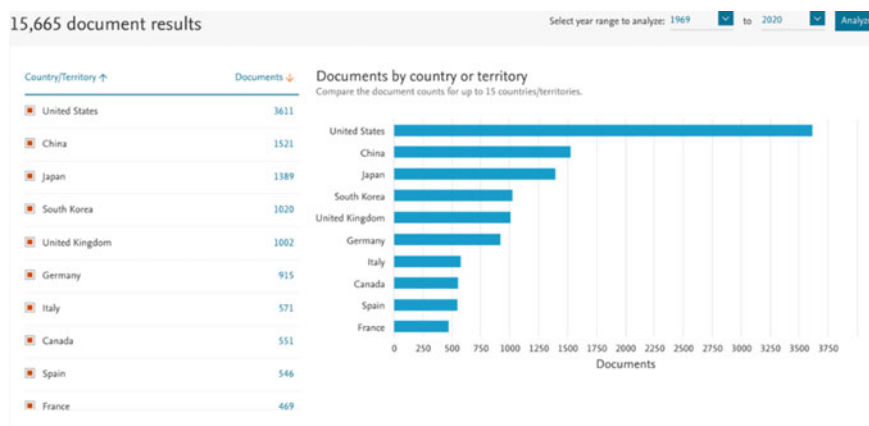


Fig. 6 Number of papers published each year in Scopus's results, by country

350,000 people. The American Stroke Association says stroke is the leading cause of long-term disability in the USA. At the same time, the Australian National Stroke Foundation also claims that stroke is one of the biggest causes of adult disability in Australia. And according to the Swedish aphasia association, 35% of the country's 12, 000 new aphasia patients are of working age, meaning they will not be able to work or communicate properly until they are back to normal [9].

In addition, although aphasia is a common disease in life, many people do not understand its harm. One of the biggest dangers of aphasia is that it can have a huge impact on a patient's daily life. Lam and Wodchis looked at the impact of more than 60 diseases on patients' quality of life and health in 15 ways. The results showed that aphasia had a significant impact on patients' lives, which was ranked highly [10].

2.3 Method of Treating Aphasia That Already Exists or is Being Studied

By reading other people's papers, find out about several existing or ongoing treatments for aphasia. This chapter selects the following six methods. The first is a treatment provided by a nurse. The second is to ask the patient to describe a story in the form of a prop given to the patient to help him practice his speech skills. The third is the use of computer programs to help patients recover. The fourth is using a tablet to help patients recover. The fifth is the use of smartphone apps to provide treatment for aphasia. The sixth is to help patients recover from aphasia by making them play games.

Nurse: Everyone's first thought after they get sick is to go to the hospital, so the most common way to treat aphasia is to send them to a rehabilitation center where they are cared for by professional nurses. Usually in hospitals or rehabilitation centers,

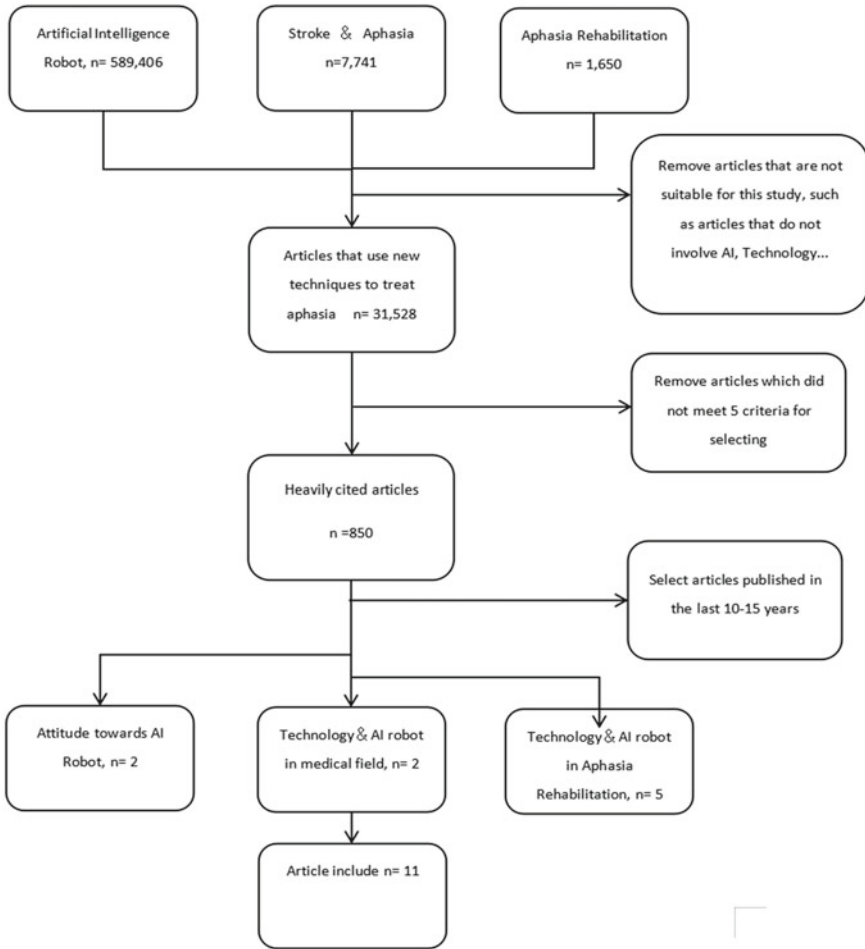


Fig. 7 Flow diagram of the research methodology

nurses will train patients in this way. The nurse first presents a picture and asks the patient to say what is in the picture. If the patient cannot answer correctly, the nurse will provide some suggestive sentences to help the patient answer, and the process will continue until the patient can answer correctly. Of course, although the patient may not be able to answer the correct answer after the nurse's guidance, in this case the nurse will directly tell him the correct answer and ask the patient to repeat the answer to deepen the impression [8].

Storytelling: Christopher Stapleto and his colleagues treat aphasia by asking patients to draw a story from a photo or drawing on a table. According to their findings, there are three ways to treat aphasia. The first is to immerse the patient in training, the second is to involve the patient in tasks similar to real world tasks (such

as making breakfast), and the third is to encourage the patient to communicate with others through storytelling to achieve the training purpose. What all three approaches have in common is that the patient is trained in context, rather than in a simple, boring, repetitive way. So, storytelling is an effective way to help patients recover from aphasia. Because when the patient is telling the story, they focus on the relevant props and the relevant situations. The patient pays less attention to the words and more to the story. This reduces anxiety when patients speak and increases their confidence and interest in treatment. Patients can stimulate their creativity by telling stories, which helps to improve the language barrier [11].

Computer: A number of researchers are currently working on using computers to provide rehabilitation training for patients with aphasia. Because the use of computer therapy can increase the intensity of treatment, it can also reduce the burden of nursing. There are two advantages to using computer therapy. The first is that patients are free to choose the time of treatment, and the second is that individuals can do it independently, which can increase patients' autonomy [12].

At present, using computer to treat aphasia has made some good achievements. According to Archibald et al., studies using computers to provide rehabilitation training for a specific language skill, or a set of language skills have achieved good results in reading, spelling, expression, and understanding semantics [13–15].

In an experiment conducted by Katz and Wertz, they compared patients who participated in computer training with those who did not receive rehabilitation training. The results showed that patients with computer aphasia performed better in providing reading, writing, expression, and comprehension training than those who did not receive any training [16]. Also, in Cherney's experiment, aphasia patients were divided into two groups, one receiving computer therapy and the other receiving professional nurses. According to Cherney's experiments, the number of words per minute that the computer-treated patients spoke increased, indicating the performance of the computer-treated patients. However, the Western Aphasia Battery (WAB) values of the experimental group and the control group were negative, indicating that the effect of using the computer in this test was not as good as that of receiving the nurse's treatment [17].

Western Aphasia Battery is a test that classifies aphasia and assesses its severity. The purpose of this test is to reflect the patient's language skills and non-verbal skills. The language ability includes four. The first one is to test whether the patient can express his or her own thoughts in a fluent manner. The second is to test whether the patient can correctly understand what is heard. The third is to test the patient's ability to repeat the naming. The fourth test patient is the ability to read and write. Non-linguistic abilities include three points, the first point of drawing ability, the second is computing power, and the third is block design and apraxia. Through this series of tests, not only can the severity of the disease be known but also the ability of the patient to perform more prominently, which can guide the development of appropriate treatment plans in the future [18].

Therefore, the use of computer to provide treatment for aphasia patients is also a better solution. Its advantage lies in the free choice of treatment time and the individual's independent completion, which increases the patient's autonomy and

reduces the workload of nurses. It has two obvious disadvantages. The first one is that it requires some computer skills, which may cause trouble to some people. Second, poor portability makes it impossible for patients and their families to travel with computers all the time.

Tablet: With the development of technology, the tablet's working ability is no less than a computer. So, some researchers have designed a treatment plan on a tablet. This solves the problem that the computer is not convenient to carry. Compared to computers, tablets are small, lightweight, and portable, allowing patients to be treated anywhere, anytime. The tablet is easy to use and control, and it also provides high quality video and audio for enhanced training. In the experiment conducted by Sonia Routhier et al., they administered self-administered semantic and phonological therapy to two patients with aphasia using a tablet. Results one of the participants had an obvious therapeutic effect, while the other participant had a better therapeutic effect though not as good as the first one. This demonstrates that the use of the tablet for the treatment of aphasia patients is effective [19]. Therefore, tablet has become a new way to motivate patients to exercise.

Smart Phone App: Now, almost everyone has a smart phone, the function of the phone is becoming more and more powerful, and people not only use the phone to make calls. More applications cover all aspects of people's life, bringing a lot of convenience to people. So, developing apps for treating aphasia on mobile phones could lead to more patients being treated. It is cheaper and more convenient for patients to use mobile apps for treatment. So, Cheng-Lin Shih and his colleagues developed a mobile app. They selected 60 words from the "Word discover" card. Then divide the 60 words into two groups. The first group is that the treatment group contains 30 words, and the second group contains the other 30 words, called the generalized group. They asked two patients, 55 and 62, to participate in the study. The results showed that the correct rate of the therapeutic words increased from 57 to 63% but decreased from 63 to 47% in the generalized words. In addition, in the treatment terms, the patient's response time decreased from 268 to 213 s, while in the generalized words. The response time decreased from 251 to 248 s. In this experiment, the correct answer rate of both patients was increased, and the response time was reduced [8].

Games: Another effective way to treat aphasia is to have patients play games. Cristina Romani et al. believed that using games as a method of rehabilitation training could enhance the enthusiasm of participants and play a positive role in the recovery of aphasia [20]. Stahl et al. designed two sets of rehabilitation training content. The first is ILAT, which lets participants play a game called "Go Fish." This is a card game that discards the cards when the patient gets a pair of identical cards and wins the game when the patient does not have a card. The other is traditional naming therapy, which asks the patient to say what is on the card. Stahl et al. invited 18 participants to participate in the experiment. The results show that the ILAT program has many improvements in the subscale of the Aachen aphasia test [8].

2.4 AI Robot in Health Care

Robots are taking more and more responsibility in the medical field. The robots used in the medical industry are roughly divided into two categories; one is a surgical robot that can perform surgery for patients, and the other is an auxiliary robot that can accompany patients, provide rehabilitation training, etc. [21]. There have been many studies that have proven that rehabilitation robots have achieved very good results in the medical industry [22], for example, Zora robots for elderly care services [23], and SLT language robots for improving cerebral palsy and communication disorders [24] and MAKRO robots for the treatment of cerebral palsy and similar movement disorders [25]. These robots have proven that robots play a positive role in the medical industry and believe that more types of robots will emerge in the future.

Nursing robots are not all perfect, and they have both advantages and disadvantages. There are two obviously advantages. The first one is let patients receive rehabilitation training at home, which will save time for patients to adapt to the environment, because patients are better trained in a more familiar environment. The second one is using robots can reduce the workload of nurses and improve the quality of care. The disadvantage of using robot to care elderly is using a robot to receive treatment at home may reduce the patient's exposure to the outside world, and the patient may be overly dependent on the robot [26].

2.5 People's Attitude Toward AI Robots

Some studies have found that gender affects people's attitude toward robots. The next two experiments verify this problem. Maartje and Somaya first put the Nao in a room and then let the participants answer some basic questions, including age, gender, and name, and so on. Next, participants were invited to enter the room to interact with the Nao. Next, touch the robot for 30 s according to Nao's requirements. Finally, fill out the questionnaire according to the interaction process. They invited 60 students from a faculty of behavioral sciences in the Netherlands. They were between 18 and 28 years old with an average age of 20.6 years. Among the invited testers were 28 males and 32 females. In addition, 30 participants were Dutch, and 30 participants were German. And almost all participants have no previous experience of interacting with robots. According to their experimental results, the negative attitude toward interaction with robots indicates that women's interaction with robots is more negative than that of men. The robotic interactions indicate that participants feel more anxious after interaction, with women being more anxious than men [27].

The experimental content of Tatsuya Nomura is basically the same as that of Maartje. The difference is that the robot used by Tatsuya is Robovie. A total of 400 participants were invited to the experiment, including 197 male participants, 199 female participants, and 4 unknown participants. The average age of participants was 21.4 years. According to the results of the Tatsuya Nomura experiment, they found

that men who hold high negative attitudes and anxiety with robot interactions try to avoid talking and talking with robots. On the other hand, the results of the experiment also reflect that female participants who have a high negative attitude toward robot interaction while being anxious will stay away from the robot's location. In addition, women who have a high negative attitude toward emotional interaction with robots are reluctant to communicate with robots [28].

In addition, studies have shown that different roles in healthcare centers have different attitudes toward robots. Different attitudes will directly affect the efficiency of the robot. A study by Broadbent et al. found that nurses and healthcare center staff responded more positively to the use of medical robots than older people, meaning that older people still prefer human care rather than robots [29]. In addition, some studies have shown that older people are more likely to accept home care robots when they live alone. People who have had contact with robots at an early age are more able to accept robots to participate in life [30]. Many scientists believe that the use of artificial intelligence robots in medical care will become more common in the future. Therefore, in order to better cooperate with the robot, the medical staff and patients need to adjust their mentality. Therefore, for medical personnel, artificial intelligence robots appear as helpers rather than human competitors. For patients, accepting a robot means that the treatment will be better.

2.6 Results of Secondary Data Collections

Stroke is the most common disease among the elderly according to data from the UK and Australia [6], and this disease has a younger trend [9]. Strokes usually have a variety of complications. Aphasia, as one of the complications that can have a great impact on the life of patients [10], has a 25–40% incidence rate [8]. The optimal treatment period for aphasics is within 3–6 months of onset, and good training during the valuable treatment period is the key. The common point among the three more effective approaches to aphasia is that patients are trained in a specific environment, rather than in a simple, boring, and repetitive way [11].

Throughout the literature review, the treatment of aphasia can be divided into two categories, shown in Fig. 8. One is to achieve therapeutic goals by changing the equipment that provides training, including using a computer, tablet, and smart phone. The second category is by changing the training content, including storytelling, games. Each of these methods has its pros and cons, but according to the results of their own experiments, these methods have played a positive role in the treatment of aphasia.

Receiving nurse care is the basic treatment of aphasia, and the use of other methods to improve the efficiency of treatment while the patient is being treated by the nurse. Nurses train patients by guiding patients to answer questions [8]. This method is the most direct and effective way to strengthen the patient's ability to pronounce through repeated exercises and finally achieve the purpose of free speech. However,

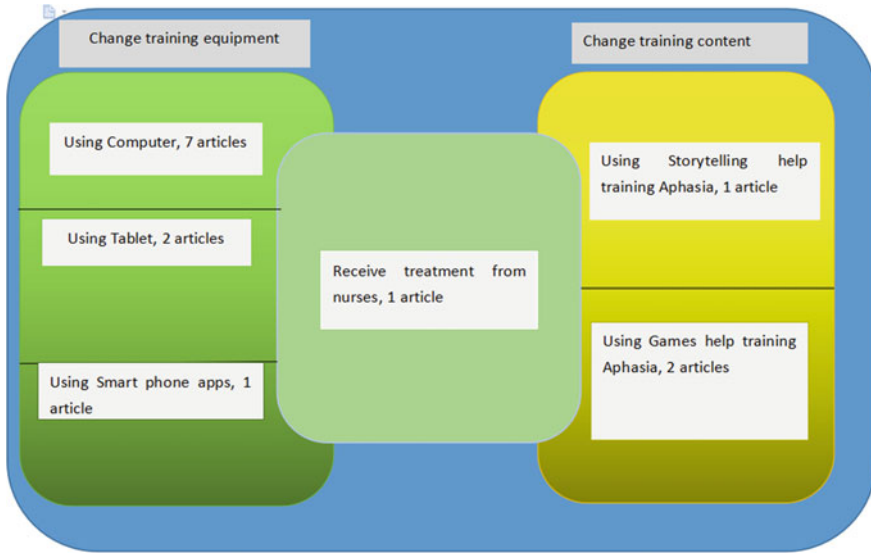


Fig. 8 Aphasia therapy classification

this training method also has some drawbacks. For example, long-term work may cause patients and nurses to get bored and affect the training effect.

As the number of patients increases, it is not enough to rely on nurses to treat aphasia. Therefore, the intervention of science and technology is an inevitable result. The use of computers, mobile phones, tablets, and robots is to improve treatment efficiency and cover more patients. By installing different types of programs, not only can aphasia be treated, but patients' information can be managed more comprehensively. Compared with computers, tablets and smart phones are more convenient carriers for patients to receive treatment anytime and anywhere. This can dramatically increase the time patients spend on treatment. But these methods are limited by machine performance. However, intelligent robots can combine the advantages of the above devices with more convenience, such as larger battery capacity can increase training time, more powerful CPU can enhance data processing ability, movable limbs can conduct demonstration training for patients and so on.

Storytelling is a way of telling a story through the association of multiple objects in real life. In this way, patients can shift their attention from objectives to stories so as to reduce the anxiety when they cannot speak a certain word, while stimulating the patient's interest, improving the patient's participation in training, and ultimately achieve better therapeutic effects [11]. But the treatment works well for the recovering patient, who can tell a coherent story. Severely ill patients may not be able to speak a full sentence, and this is obviously not for them.

Compared with storytelling, games are a more widely applicable training method. It can stimulate the enthusiasm of patients to receive treatment through simple games and improve the therapeutic effect. Stahl et al. designed two rehabilitation training

games [8]. The combination of these two games can improve patients' cognitive and speech disorders at the same time, and Stahl et al. found that games play a positive role in the recovery of aphasia.

3 Conclusion

This paper briefly summarizes some work on the use of robots in the direction of rehabilitation training for aphasia caused by stroke by reviewing the literature. Through the retrieval and induction of these literatures and studies, the conclusions of this study are as follows:

- (1) Some studies have shown that it is effective to use some electronic devices such as computers, mobile phones, and tablets for the rehabilitation of aphasia caused by stroke. This proves that the use of robots in this field is feasible. Some research results also show that electronic equipment is insufficient in rehabilitation training, such as inability to carry or move easily, and prolonged use of equipment can make patients feel bored. The emergence of robots can improve these deficiencies.
- (2) Some studies have shown that women and the elderly are more likely to have negative emotions about robots, which may be due to the greater need for emotional attention and support for women and the elderly. The monotonous and boring mode of interaction between robots and humans makes it easier for them to fail to feel the emotional attention and support of robots for them. Therefore, the use of robots for rehabilitation training should consider providing feedback to the patient's emotions as much as possible in the interactive mode, such as using facial expression recognition to capture the patient's facial expressions and give different feedback during the training process.
- (3) Some games can be designed on robots to help with rehabilitation training. In addition, some studies have demonstrated the importance of nurses' care in the whole rehabilitation training, so the authors believe that robots should be used as an auxiliary treatment tool to help nurses and doctors to complete rehabilitation training rather than replace them.

References

1. American Stroke Association (2019) Stroke symptoms <https://www.stroke.org/en/about-stroke/stroke-symptoms>. Accessed 28 Dec 2019
2. American Stroke Association (2019) Types of stroke. <https://www.stroke.org/en/about-stroke/types-of-stroke>. Accessed 28 Dec 2019
3. Mayo Clinic (2019) Aphasia-symptoms and cause. <https://www.mayoclinic.org/diseases-conditions/aphasia/symptoms-causes/syc-20369518>. Accessed 30 Dec 2019
4. Sanbot (2019). <http://www.sanbot.com/sanbots1/Healthcare>. Accessed 10 Aug 2019

5. Joanna Briggs Institute (2017) The Joanna Briggs Institute Critical Appraisal tools for use in JBI systematic reviews-checklist for systematic reviews and research syntheses. <https://joannabriggs.org/research/critical-appraisal-tools.html>. Accessed 10 Sep 2019
6. Australian Bureau of Statistics (2018) Population Projections, Australia, 2017 (base)—2066. <http://www.abs.gov.au/Ausstats/abs@.nsf/mf/3222.0>. Accessed 4 May 2019
7. Steve Brine (2018) New Figs show larger proportion of strokes in the middle aged. <https://www.gov.uk/government/news/new-Fig.s-show-larger-proportion-of-strokes-in-the-middle-aged>. Accessed 4 May 2019
8. Shih C-L, Cheng K-S, Wang J-L, Jhang B-S, Yang C-H (2013) Smart phone based assistive speech therapeutic system for Aphasia. In: Proceedings of 2013 IEEE third international conference on consumer electronics
9. Adamson J, Beswick A, Ebrahim S (2004) Is stroke the most common cause of disability? Paper Presented at J Stroke Cerebrovasc Dis: Official J Natl Stroke Assoc 13(4):171–177
10. Jonathan MCL, Wodchis WP (2010) The relationship of 60 disease diagnoses and 15 conditions to preference-based health-related quality of life in Ontario hospital based long-term care residents. *Med Care* 48(4):80–387
11. Stapleton C, Whiteside Phd J, Davies Phd J, Mott D, Vick J (2014) Transforming lives through story immersion: innovation of aphasia rehabilitation therapy through storytelling learning landscapes, pp 29–34
12. Wade J, Mortley J, Enderby P (2003) Talk about IT: views of people with aphasia and their partners on receiving remotely monitored computer-based word finding therapy. *Aphasiology* 17:1031–1056
13. Archibald L, Orange J, Jamieson D (2009) Implementation of computer-based language therapy in aphasia. *Ther Adv Neurol Disord* 2:99–311
14. Wallesch C, Johannsen-Horbach H (2004) Computers in aphasia therapy: effects and side effects. *Aphasiology* 18:223–228
15. Wertz R, Katz RC (2004) Outcomes of computer provided treatment for aphasia. *Aphasiology* 18:229–244
16. Katz R, Wertz R (1992) Computerized hierarchical reading treatment in aphasia. *Aphasiology* 6:165–177
17. Cherney LR (2010) Oral reading for language in aphasia (ORLA): evaluating the efficacy of computer-delivered therapy in chronic non-fluent aphasia. Paper Presented at Top Stroke Rehabil 17:423–431
18. Sublett K (2013) Copy of Western aphasia battery-revised. <https://prezi.com/b9q2fcepzdw/copy-of-western-aphasia-battery-revised/> Accessed 4 May 2019
19. Routhier S, Bier N, Macoir J (2014) Smart tablet for smart self-administered treatment of verb anomia: two single-case studies in aphasia, 269–289
20. Romani C, Thomas L, Olson A, Lander L (2018) Playing a team game improves word production in poststroke aphasia, 1–36
21. Kim J, Gu GM, Heo P (2016) Robotics for Healthcare. In: Jo H et al (eds) *Biomedical engineering: frontier research and converging technologies*. Springer International Publishing, Cham, pp 489–509
24. Chew and Turner (2019) Can a robot bring your life back? A systematic review for robotics in rehabilitation. Springer Nature Book Chapter: *Robotics in Healthcare: Field Examples and Challenges*
23. Tuisku O et al (2018) Robots do not replace a nurse with a beating heart. *Inf Technol People* 32(1):47–67
24. Ochoa-Guaraca M, Pulla-Sánchez D, Robles-Bykbaev V, López-Nores M, Carpio-Moreta M, García-Duque J (2017) A hybrid system based on robotic assistants and mobile applications to support in speech therapy for children with disabilities and communication disorders. *Campus Virtuales* 6(1):77–87
25. Gnjatovic M et al (2018) Pilot corpus of child-robot interaction in therapeutic settings. In: Proceedings of 8th IEEE international conference on cognitive Infocommunications, CogInfoCom 2017—Proceedings (CogInfoCom), pp 253–258

26. Marti P, Bacigalupo M, Giusti L, Mennecozzi C, Shibata T (2006) Socially assistive robotics in the treatment of behavioural and psychological symptoms of dementia. In: Proceedings of the First IEEE/RAS-EMBS international conference on biomedical robotics and biomechatronics, BioRob, Pisa, Italy
27. Maartje MA, Allouch SB (2013) The relation between people's attitude and anxiety towards robots in human-robot interaction. In: Proceedings of 2013 IEEE ROMAN
28. Nomura T, Kanda T, Suzuki T, Kato K (2008) Prediction of human behavior in human—Robot interaction using psychological scales for anxiety and negative attitudes toward robots. *IEEE Trans Robot* 24(2):442–451
29. Broadbent E, Stafford R, MacDonald B (2009) Acceptance of healthcare robots for the older population: review and future directions. *Int J Soc Robot* 1:319–330
30. Crumble R, Sedghdelli S, Khosla R, Chu M-T (2014) Socially assistive robots in elderly care: a mixed-method systematic literature review. *Int J Hum-Comput Interact* 30(5):369–393

An Estimation Steering Feedback Torque in Vehicle Steer by Wire System



S. M. H. Fahami, Faiz Mohd Turan, and M. A. Zakaria

Abstract A steering feedback torque in conventional steering system is generated between a tire and ground contact. This steering torque is for a driver steering feel in parallel confident level during maneuver. In steering by wire (SBW) system, the absence of the column shaft requires the system to generate the torque and should equal to conventional steering system for a realistic driver steering feel. This paper propose an algorithm to create an estimation steering feedback torque control for vehicle SBW system. The estimation torque consists of a steering and front axle model. While by adding a phase compensation torque, driver will have realistic steering feel. Moreover, to control the torque, the LQR + GS control method is used. To investigate the effectiveness of proposed control algorithm, the MATLAB tools software is used to analyze the response. Based on the finding results, the proposed algorithm is able to create estimation feedback torque, and the phase compensation torque provides a realistic driver steering feel.

Keywords SBW · Feedback · Torque · LQR

1 Introduction

The next generation of steering system is steering by wire (SBW) system, whereby the need of column shaft is absent and it was replaced with a sensor, actuators, and controller unit as shown in Fig. 1. The SBW system offers an advantage such as enhance response of the vehicle maneuver and stability [1–3]. Moreover, it gives larger space interior cabin and ergonomic due to the absence of column shaft [4, 5].

The steering feedback torque for driver steering feel [1, 14] is a challenging issue in SBW system as main characteristics [7]. The main function of the torque is for steering feel and returnability of steering wheel [8]. A study on how to create a torque has been conducted. Oh et al. [9] create a torque map element of elements steering

S. M. H. Fahami (✉) · F. M. Turan · M. A. Zakaria
Faculty of Manufacturing and Mechatronic Engineering Technology, University Malaysia Pahang,
Pekan, Pahang 26600, Malaysia
e-mail: hafizfahami@ump.edu.my

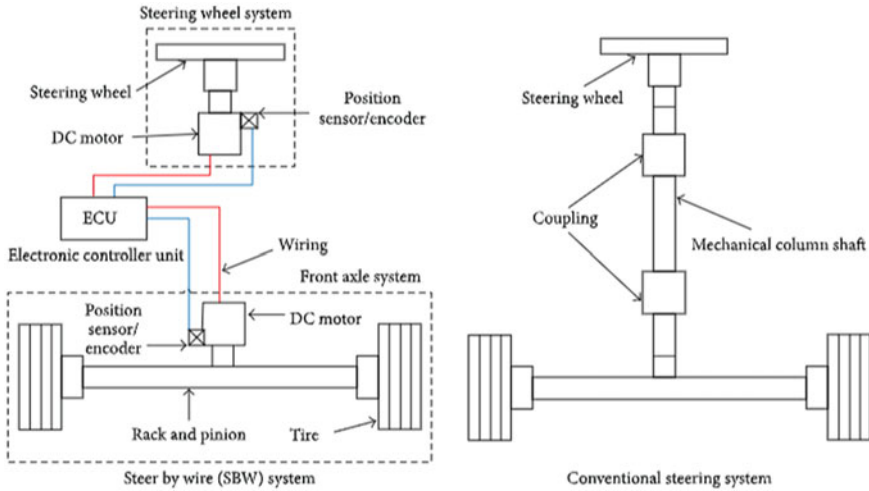


Fig. 1 Conventional steering and SBW system

angle and speed of vehicle. At a high speed, the torque is increased and vice versa at lower speed. The controls gain of the torque map is to vary the torque. On the others, author, Kim et al. [10] improve the torque map by adding the damping torque for a steering feel. Authors, Amberkare et al. [6], used parameter of steering angle with an appropriate system model to generate the torque that change the desired steering feel behavior, while a disturbance observer method with current of front model system by Asai et al. [11] is to create the torque. A model reference with factor of inertia and damping element based on steering system is introduced by Park et al. [12], while a model matching approach is applied by Odenhtal et al. [13] which combination of electrical and mechanical parts create the torque.

This paper is to propose and create an estimation steering feedback torque with the control algorithm in for SBW system. The subsystem component composed of steering wheel, front axle system, and linear vehicle model with a control method is applied. To analyze the effectiveness of the proposed control algorithms, the electric power steering (EPS) system [15] is used as reference of the system.

2 The Modeling of Steer by Wire (SBW) System

The models of SBW system consist of subsystem of steering wheel system, front axle system, and linear vehicle model as a main component system.

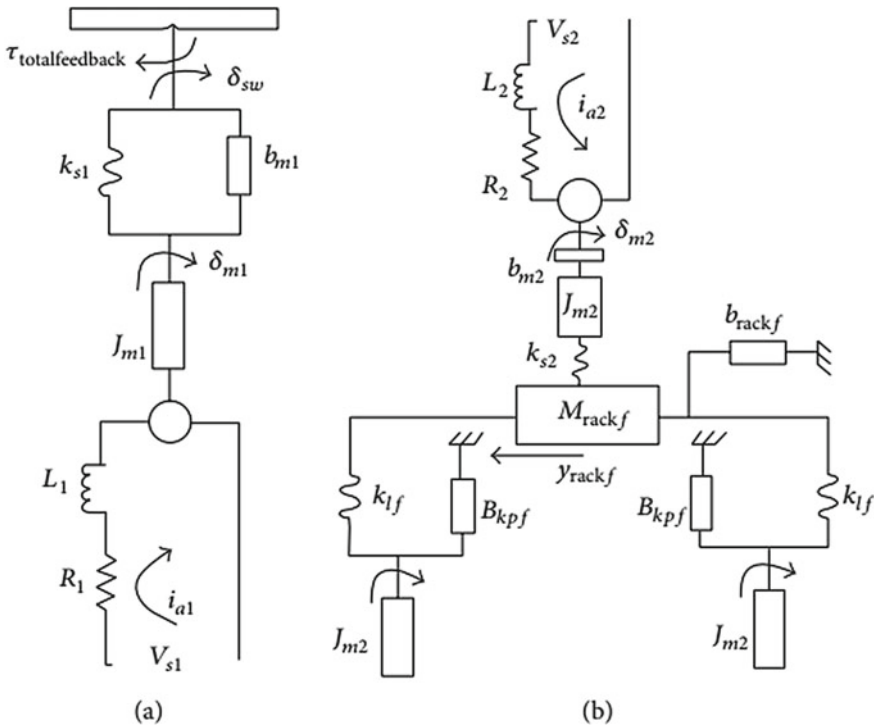


Fig. 2 System diagram of **a** steering wheel and **b** front axle system

2.1 The Steering Wheel System Model

Figure 2A shows a system diagram of steering wheel system.

As shown in Fig. 2a, purpose of steering wheel is to create a steering torque for steering feel and returnability of steering. The estimation total feedback torque ($\tau_{totalfeedback}$) is the input to the system. The rate change of steering wheel motor angular displacement ($\dot{\delta}_{m1}$), the motor angular displacement (δ_{m1}), and the current of steering wheel motor (i_{a1}) are outputs of the system. Table 1 shows parameters of steering wheel system.

2.2 The Front Axle System Model

The model front axle system is shown in Fig. 3. The front axle component consists of front axle motor, system of rack-pinion, and dynamic of wheel model. The steering angle (δ_{sw}) is input to the system, and front tire angle (δ_f) is output of the system. The parameters of front axle wheel are illustrated in Table 2.

Table 1 Parameter of steering wheel system

Parameters	Descriptions	Values	Units
R_1	Motor resistance	5.54	Ω
L_1	Motor inductance	0.016	H
K_{sm}	Steering motor constant	0.025	Nm
J_{m1}	Steering motor inertia	0.0035	Kgm^2
b_{m1}	Steering motor damping	0.0071	$\text{Nm}/(\text{rad}/\text{s})$
k_{s1}	Torque stiffness	0.024	Nm
V_{s1}	Steering motor voltage	–	V

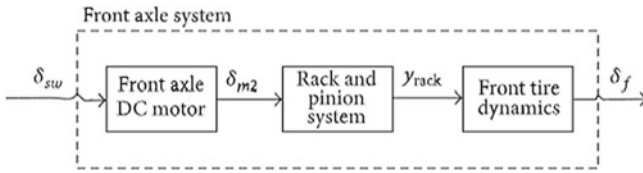


Fig. 3 Front axle wheel system

Table 2 Parameter of front wheel system

Parameters	Descriptions	Values	Units
R_2	Motor resistance	5.54	Ω
L_2	Motor inductance	0016	H
K_{fm}	Front motor constant	0.025	Nm
J_{m2}	Front motor inertia	0.0035	Kgm^2
b_{m2}	Front motor damping	0.0071	$\text{Nm}/(\text{rad}/\text{s})$
k_{s2}	Front torque stiffness	0.024	Nm
B_{rack}	Rack damping coefficient	0.016	
M_{rack}	Rack lumped coefficient	0.031	Nm
k_{lf}	Rack linkage stiffness	0.00063	Kgm^2
r_L	Offset of king pin axis	0.00035	m
r_p	Pinion gear radius	0.026	m
B_{kp}	King pin damping coefficient	0.00061	Kgm^2
I_f	Lumped front wheel inertia	0.00035	Kgm^2

2.3 The Linear Vehicle Model

This paper used a linear vehicle model shown below in Fig. 4, [16] to track the dynamics response such as yaw rate and body slip angle as input to generate a self-aligning torque. The input to this model is front tire angle and vehicle speed. The parameters of the model are illustrated in Table 3.

The assumptions of the model considered for normal driving maneuverer are written as follows:

- Negligible—friction force direction,
- Vehicle at constant speed.
- Maximum vehicle speed applied at 120 km/h.
- Steering ration 15:1.

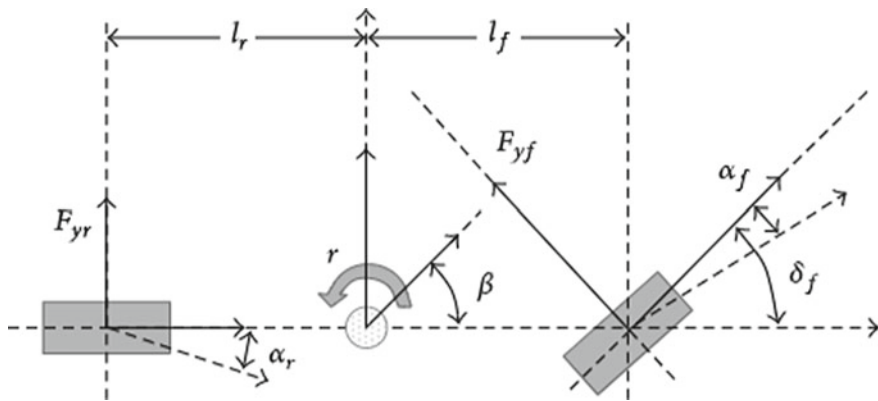


Fig. 4 Linear vehicle model [16]

Table 3 Parameter of single-track linear vehicle model

Parameters	Descriptions	Values	Units
C_f	Front cornering stiffness	35,000	N/rad
C_r	Rear cornering stiffness	25,000	N/rad
l_f	Length front center wheel	1.5	m
l_r	Length rear center wheel	1.1	m
m	Vehicle mass	1600	kg
I_s	Vehicle inertia	2100	$\text{Kg}\cdot\text{m}^2$
V	Vehicle speed	-	Km/h

3 Estimation Feedback Torque and the Control Method

It is known that the steering feedback torque in SBW system has to be created for driver steering feel and steering returnability. Therefore, the steering wheel in SBW system is composed of DC motor and sensor used to create and control the feedback torque ($\tau_{\text{feedbacktotal}}$). A block diagram as shown in Fig. 5 shows the proposed estimation feedback torque control.

Based on Fig. 5 proposed control, the estimation feedback torque consists of average torque of the front axle motor (τ_{fm}), steering motor (τ_{sm}), and estimation self-aligning torque (τ_a). Furthermore, add the phase compensation torque elements which is inertia (τ_{inertia}) and damping torque (τ_{damp}) for a realistic steering torque, while a steering feel gain (k_{feel}) is to vary the torque depending on steering wheel angle and vehicle speed response and integration between gain scheduling and LQR controller is used to control the torque. To show an influence of the tire ground contact is represented by self-aligning torque (τ_a) and acts as disturbance input as written in Eq. 1.

$$\tau_a = -C_f(t_p + t_m) \left[\beta + \frac{l_f r}{v} - \delta_f \right] \mu \quad (1)$$

whereby coefficient of a dry road condition—(μ), pneumatic trail—(t_p), mechanical trail—(t_m) body slip angle—(β), and yaw rate—(r) of vehicle. Therefore, the estimation feedback torque for the steering system is written in Eq. 2.

$$\tau_{\text{feedback}} = (\tau_a + \tau_{\text{sm}} + \tau_{\text{fm}}) K_{\text{feel}} \quad (2)$$

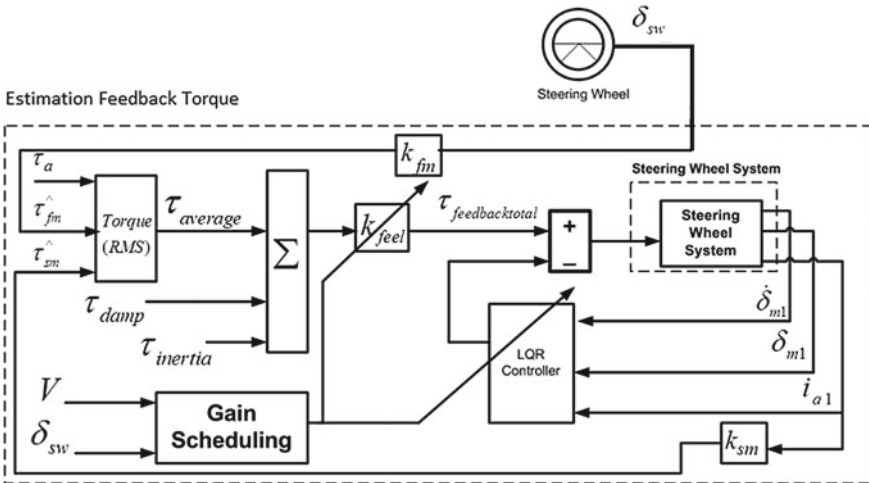


Fig. 5 Estimation feedback torque control algorithm of SBW system

The relationship of torque between current and motor constant can be defined [17]. Thus, the torque of front axle motor (τ_{fm}) is the product of current of front axle motor ($i_{\alpha 2}$) and front axle motor constant (k_{fm}) written below in Eq. 3.

$$\tau_{fm} = i_{\alpha 2} k_{fm} \quad (3)$$

While the steering motor torque (τ_{sm}) is product of steering motor current ($i_{\alpha 1}$) and steering motor constant (k_{sm}), it is written in Eq. 4.

$$\tau_{sm} = i_{\alpha 1} k_{sm} \quad (4)$$

Figure 6 shows a torque response between the steering wheel and front axle motor, when input to the system is a profile of lane change maneuver. The results shows, a front axle motor gives high in torque and this is come from the element of self-aligning torque effected from tire ground contact compare to the torque of steering angle.

The phase compensation torque has the potential to vary the steering feel. The driver could sense tight at steering wheel at high speed and vice versa at lower speed. Moreover, it is able to reduce a vibration and stabilize the system [5] by controlling the gains. For this reason, the element is taken in this propose control to provide better response of feedback torque and in parallel for a driver feel during manoeuver. This phase compensation torque elements consist of damping and inertia torque.

The damping torque (τ_{damp}) is written in Eq. 5, whereby a (δ_{m1}) is a rate of change steering wheel motor angle and (k_{damp}) is a damping gain.

$$\tau_{damp} = \delta_{m1} k_{damp} \quad (5)$$

Fig. 6 Steering wheel and front axle motor torque

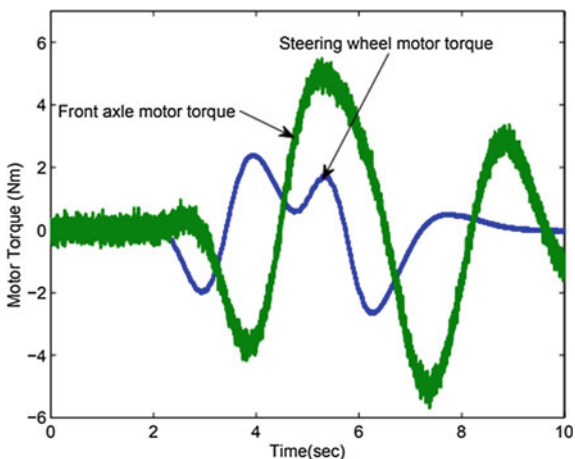
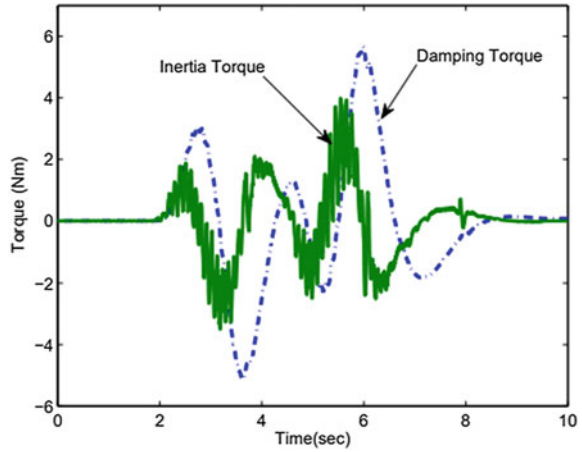


Fig. 7 Phase compensation torque



For the inertia torque ($\tau_{inertia}$) is a product of acceleration of the steering motor angle ($\ddot{\delta}_{m1}$) and the inertia gain ($k_{inertia}$), written in Eq. 6.

$$\tau_{inertia} = \ddot{\delta}_{m1} k_{inertia} \quad (6)$$

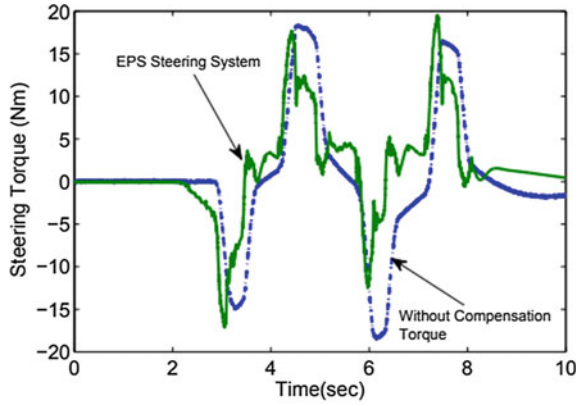
The responses of this phase compensation torque are shown in Fig. 7. The damping torque provides a smooth response whereby it is able to reduce a vibration.

While the inertia elements provide a high magnitude from the inertia feel itself, the combination between damping and inertia factor could provide an average of a realistic driver steering feel. This is shown in Fig. 8, whereby the compensation torque gives an effort interest to follow the steering torque of EPS system. However, wide adjustment gain of damping and inertia can improve the torque response.

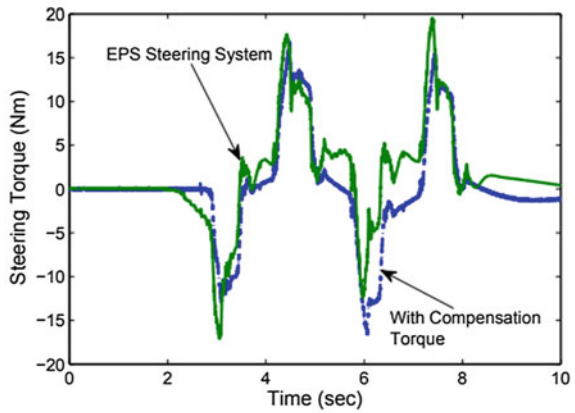
4 The LQR Controller and LQR + Gain Scheduling (GS)—Control Feedback

There are two controllers proposed to control the feedback torque which are LQR controller and LQR + GS. Both controllers are then comparing to analyze the response performance that is adequate with a change of road condition. The justification has been made that both controllers should have acceptable RMS value 10% from the reference value.

Fig. 8 Steering torque with ad w/o phase compensation torque



(a) Steering Torque without compensation torque



(b) Steering Torque with compensation torque

4.1 The LQR Controller

For the LQR controller, the Bryson’s rule method is used in order to define a gain (K_x) of the controller. This gain is illustrated in Table 4.

Table 4 Gain parameter for LQR controller

LQR gain (K_x)	K_1	K_2	K_3
	0.45	0.03	0.001

Table 5 Gain parameter for LQR controller + GS

V (km/h)	δ_{sw}	k_1	k_2	k_3	k_{feel}
0–30	$(\pm 180^\circ < \delta_{sw} < \pm 405^\circ)$	0.02	0.55	0.001	0.13
31–100	$(\pm 46^\circ < \delta_{sw} < \pm 179^\circ)$	0.02	0.74	0.001	0.55
101–120	$(+45^\circ < \delta_{sw} < -45^\circ)$	0.04	0.85	0.001	0.90

4.2 The LQR Controller with Gain—Scheduling

The process behavior changes depending on operating condition [6] in many situations. Therefore then, the controller gain is possible to change by monitoring the condition of the process. The gain scheduling method in these studies is used to change the gain of LQR controller and adjustable feel gain accordance to the input of vehicle speed and driver steering angle. By doing this, it could provide a better response in torque control to achieve a realistic driver steering feel. The speed of vehicle are categorized as between (0–30) km/h, lower speed, (31–100) km/h, medium speed, and (101–120) km/h, high speed. While for a steering angle, it is between $(\pm 180^\circ < \delta_{sw} < \pm 405^\circ)$, low speed, $(\pm 46^\circ < \delta_{sw} < \pm 179^\circ)$, medium speed, and $(+45^\circ < \delta_{sw} < -45^\circ)$, high speed. The controller gains of (k_1, k_2, k_3) and k_{feel} are illustrated in Table 5.

The steering torque comparison at different control approach is shown below in Fig. 9. The system input is the response of lane change maneuver at vehicle speed of 80 km/h. Table 6 illustrated a steering torque comparison.

Fig. 9 Comparison between controller LQR and LQR + GS

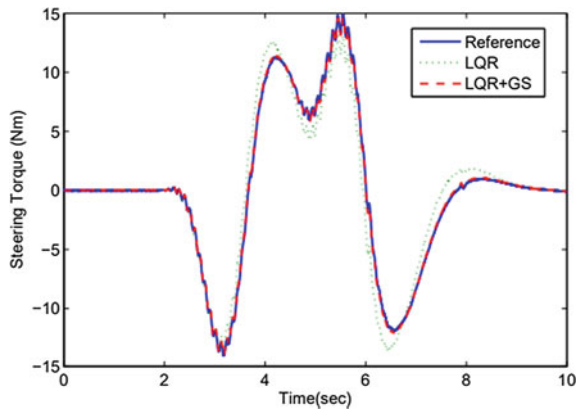
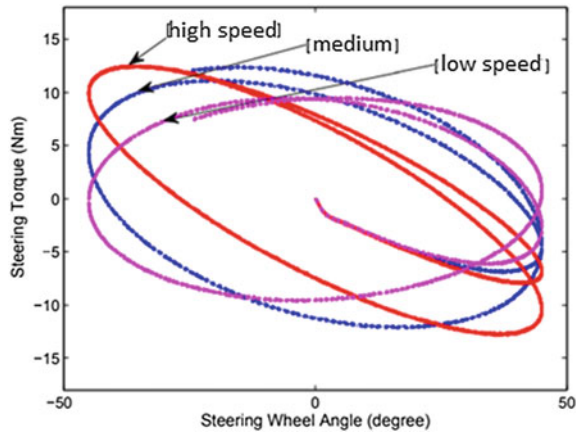


Table 6 RMS steering torque at different control

	References	LQR	LQR + GS
RMS (Nm)	5.313	4.685	5.236
Similarity (%)	–	88.2%	98.55%

Fig. 10 Steering torque and steering angle at different speed



Based on the results, the RMS values steering torque of reference torque, LQR, and LQR + GS torque are 5.313 Nm, 4.685 and 5.236 Nm. The method of LQR + GS offers 98.33% improvement due to change of gain of vehicle speed and steering angle compared using method of LQR controller, while the responses between a steering torque and steering angle at different speed are shown in Fig. 10.

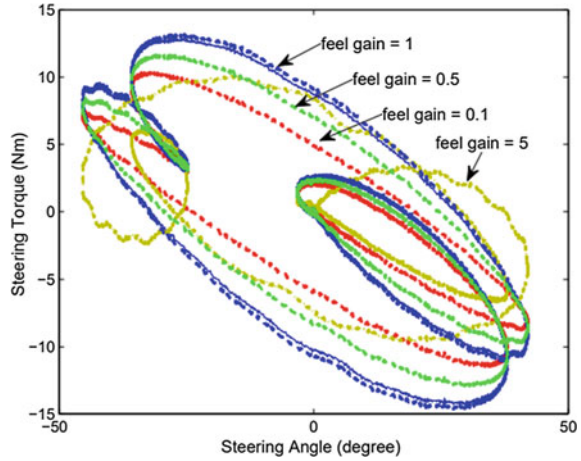
The driver input steering angle is sine wave response at different speed. Based on the obtained results, it is clearly described, even at equal input steering angle, that the drivers will feel a oppose driver steering torque by varying gains of the controller.

By increasing the steering torque, it will cause to sense a stiff on steering wheel and vice versa. Furthermore, Fig. 11 shows other results at different steering feel gain at equal high speed. The range of steering feel gain is between 0.1 and 1 defined through optimization. When the gain value is 1, the steering torque increases at almost 14 Nm. Thus, driver senses hard to turn the steering wheel. By decreasing the gain value to 0.1, the steering torque is reduced to 10 Nm. On the other hand, when the feel gain is more than their maximum limit value which is 1, the response of steering torque is become unstable. And this in parallel driver could sense different feel at same speed.

5 Conclusion

Based on the proposed architecture algorithm, the estimation feedback torque is based on element of the torque of steering and front axle motor. While in order to improve steering feel during maneuver, the phase compensation torque is added and the steering feel gain is used to vary the range sense of realistic of steering feel. Furthermore, the LQR + GS control method is used to control the feedback torque. Based on the result findings, the LQR + GS approach gives a better improvement to control the steering torque response compared to LQR controller whereby a

Fig. 11 Different steering feel gain at high speed



percentage performance is 98% similar to reference differ to LQR controller whereby is 88%. On the others, a wide range adjustment gains of the feel and phase compensation, its able contributes to improves the steering torque to during maneuver for driver steering feel.

References

1. Yao Y (2006) Vehicle steer-by-wire system control. SAE Technical Paper 2006-01-1175
2. Kumar EA, Dinesh D, Kamble N (2012) An overview of active front steering system. *Int J Sci Eng Res* 3
3. Fahami SMH, Zamzuri H, Mazlan SA, Zulkarnain NB (2013) The design of vehicle active front steering based on steer by wire system. *Adv Sci Lett* 19(1):61–65
4. Fahami SMH, Zamzuri H, Mazlan SA, Saruchi SA (2014) The variable steering ratio for vehicle steer by wire system using hyperbolic tangent method. *Appl Mech Mater* 575:781–784
5. Kaufmann T, Millsap S, Murray B, Petrowski J (2001) Development experience with steer by wire. In: *Proceedings of the SAE international congress and exhibition*, SAE 2001-01-2479
6. Amberkare S, Bolourchi F, Demerly J, Millsap S (2004) A control system methodology for steer by wire systems. SAE Technical Paper 2004-01-1106
7. Baviskar A, Wagner JR, Dawson DM, Braganza D, Setlur P (2009) An adjustable steer-by-wire haptic interface tracking controller for ground vehicles. *IEEE Trans Veh Technol* 58(2):546–554
8. Segawa M, Kimura S, Kada T, Nakona S (2002) A study of reactive torque control for steer by wire system. In: *Proceedings of the international symposium on advanced vehicle control (AVEC'02)*, Hiroshima, Japan, pp 653–658
9. Oh SW, Chae HC, Yun SC, Han CS (2004) The design of a controller for the steer-by-wire system. *JSME Int J, Ser C* 47(3):896–907
10. Kim CJ, Jang JH, Oh SK, Lee JY, Hedrick JK (2008) Development of a control algorithm for a rack-actuating steer-by-wire system using road information feedback. *Proc Inst Mech Eng Part D: J Automob Eng* 222(9):1559–1571
11. Asai S, Kuroyanagi H, Takeuchi (2004) Development of a steer-by-wire system with force feedback using a disturbance observer. In: *Proceedings of SAE world congress & exhibition on steering & suspension technology*, Detroit, Mich, USA

12. Park TJ, Han CS, Lee SH (2005) Development of the electronic control unit for the rack-actuating steer-by-wire using the hardware-in-the-loop simulation system. *Mechatronics* 15(8):899–918
13. Odenhtal D, Bunte T, Heitzer HD, Eicker C (2000) How to make steer-by-wire feel like power steering. In: *Proceedings of the 15th IFAC World Congress, Barcelona, Spain*
14. Na H, Zong C, Hu D (2009) Investigations on cornering control algorithm design and road feeling optimization for a steer-by-wire vehicle. In: *Proceedings of the IEEE international conference on mechatronics and automation (ICMA'09), Changchun, China*, pp 3246–3251
15. Ancha S, Baviskar A, Wagner JR, Dawson DM (2007) Ground vehicle steering systems: modelling, control, and analysis of hydraulic, electric and steer-by-wire configurations. *Int J Veh Des* 44(1–2):188–208
16. Yih P, Gerdes JC (2005) Modification of vehicle handling characteristics via steer-by-wire. *IEEE Trans Control Syst Technol* 13(6):965–976
17. Pastorino R, Naya MA, Pérez JA, Cuadrado J (2011) Geared PM coreless motor modelling for driver's force feedback in steer-by-wire system. *Mechatronics* 21(6):1043–1054

An Implementation of Sliding Mode Voltage Control Controlled Buck-Boost Converter for Solar Application



Nursabrina Athirah Mohd Mustakin, Mohd. Shafie Bakar,
and Mazyah Mat Noh

Abstract The control of direct current to direct current (DC–DC) converter is a key step to guarantee a fixed output voltage despite of load and input voltage variations. DC–DC converters are an important element in solar application to attain desired level of voltage and to shape it according to the demand. This paper presents fully on the performance and comparison of the sliding mode control (SMC) and SMI-PID methodologies for buck-boost DC–DC converter for solar applications for three types of modules; user-defined, Solartech energy, and Suntech energy module. In this context, the SMC presents an adequate choice due to its robustness and efficiency. The comprehensive operating principle under continuous conduction mode (CCM) was obtained. With duty cycle varied at 0.7 (boost operation) and at 0.4 (buck operation), the system gained at output voltage -52.4 V and -15.11 V, respectively, under steady-state condition. Also, the study obtained that the higher switching frequency (10 kHz) all three types of solar modules carry out the output voltage ripple at below 5%.

Keywords DC–DC converter · Sliding mode control · Robustness

1 Introduction

Direct current to direct current (DC–DC) converters are high-frequency power conversion circuits that smooth out switching noise into regulated DC voltage using inductors, transformers, and capacitors. Basically, the photovoltaic (PV) system is generally intermittent in nature which makes it is critical to stabilize the output voltage. Hence, buck-boost converter can be used as the regulator. Model-based controller which is sliding mode voltage control (SMVC) or as known as sliding mode control (SMC) able to cope with their wide input voltage and load variations as it is well known for their robustness, stability, and easy implementation [1]. To analyze the performance of buck-boost converter with nonlinear load for

N. A. M. Mustakin · Mohd. S. Bakar (✉) · M. M. Noh
FTKKE, Universiti Malaysia Pahang, Pekan, Pahang, Malaysia
e-mail: shafie@ump.edu.my

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
M. A. Abdullah et al. (eds.), *Advances in Intelligent Manufacturing and Mechatronics*,
Lecture Notes in Electrical Engineering 988,
https://doi.org/10.1007/978-981-19-8703-8_5

solar application is the objective that will be achieved throughout this research. The output voltage is being examined when PV array being placed at the input circuit of buck-boost converter. The response of converter agrees with the range of duty ratio (5–90%) and 1 kHz as minimum frequency. Being buck-boost converter, this converter is capable of both stepping-up and stepping-down the voltage with the voltage variation problems can be reduced. It is transforming a positive DC voltage at the input to a negative DC voltage at the output. For a pulse width modulation (PWM) duty cycle, $D \rightarrow 0$, the output voltage equals to zero, and for $D \rightarrow 1$, the output voltage grows toward negative infinity. It is operated in continuous conduction mode (CCM) which given by

$$V_{\text{out}} = -\frac{D}{1-D} \cdot V_{\text{in}} \quad (1)$$

Using the control approach, it could be able to provide a method to design a controller for a system to be insensitive to parameter variations and external load disturbances. One of the main features of this method is that one only needs to drive the error to a switching surface, after which the system is in sliding mode and robust against modeling uncertainties and disturbances. The output of a solar PV system is maximized either by mechanically tracking the sun and orienting the panel in such a way that it receives the maximum solar irradiance under changing conditions of temperature, or by mechanically tracking the sun and orienting the panel in such a way that it receives the maximum solar irradiance [2]. This paper is organized in the following steps. The system modeling of conventional SMVC and SMVC with PID controller is depicted separately in Sect. 2. In Sect. 3, the output voltage simulation results and analysis of output voltage ripple are reported under three types of solar modules specifications such as user-defined, Solartech energy, and Suntech energy. In Sect. 4, the simulation presents the conclusion and findings.

2 Methodology

2.1 System Modeling of Conventional SMVC Controller

For designing such a controller, the moving average of the output voltage which significantly simplifies its design is used. Due to limitations of the system parameters such as duty cycle, it is not possible to increase the convergence factor beyond a certain value [3]. The SM controller has a switching function:

$$u = \begin{cases} 1 & \text{when } S > 0 \\ 0 & \text{when } S < 0 \end{cases} \quad (2)$$

where S is the instantaneous state variable's trajectory.

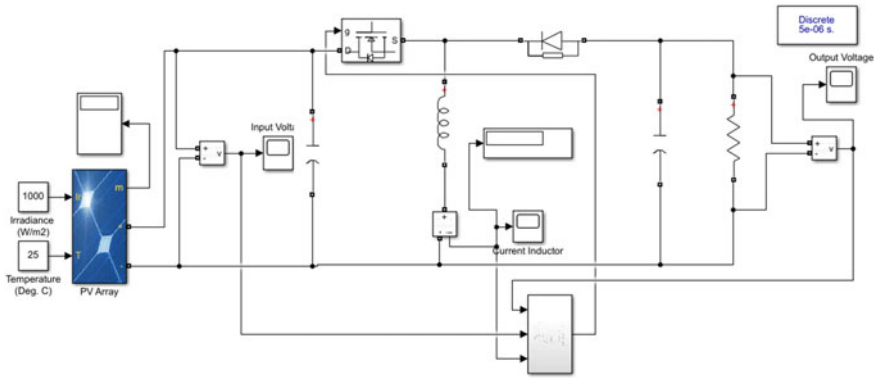


Fig. 1 Design circuit of SMVC-based buck-boost converter with PV array in CCM

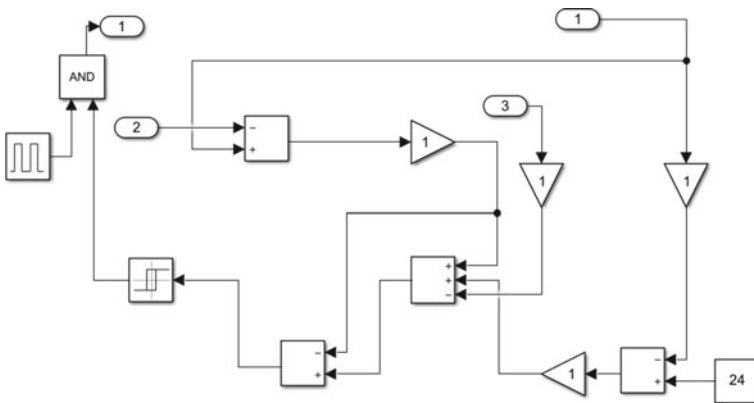


Fig. 2 SMVC design circuit in subsystem

Figure 1 shows the MATLAB/Simulink program for the modeling of DC–DC buck-boost converter based on SMVC. The closed loop control using SMC in the subsystem is represented on Fig. 2.

2.2 System Modeling-Based PID SMVC Controller

The first step to design a SMVC is to develop the converter model in terms of the desired control variables. The controller under study is a second-order proportional integral derivative (PID) SM voltage controller [4]. A second-order PID type of SM voltage controller is adopted [5].

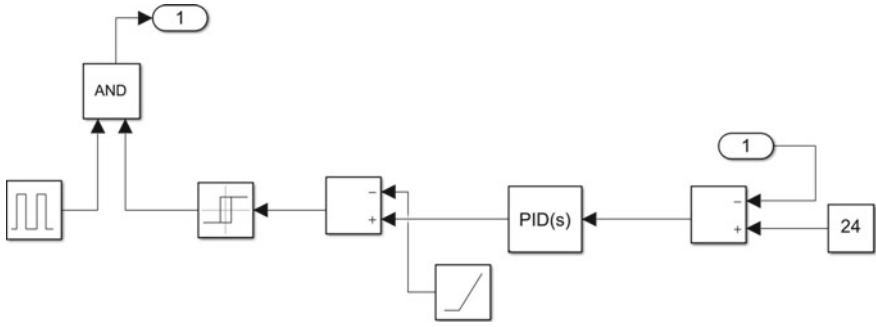


Fig. 3 SMVC design circuit in subsystem-based PID SMVC converter

Table 1 Specifications of buck-boost converter

Parameter	Description	Nominal value
V_{in}	Input voltage	24 V
f	Switching frequency	1–10 kHz
L	Inductance	25 μ H
C	Capacitance	1000 μ F
R_L	Load resistance	12 Ω
K_p	Proportional component	0.8758
K_I	Integral component	84.013
K_D	Derivative component	7e-4
D	Duty cycle	0.4 and 0.7

The PID controller uses the formula:

$$G_c(s) = \frac{U(s)}{E(s)} = (K_P + \frac{K_I}{s} + K_D s) \tag{3}$$

where K_p is proportional gain, K_I is integral gain, and K_D is derivative gain.

Figure 3 illustrated the closed loop control of buck-boost converter using SMVC consists of PID controller. The default values of the converter parameters are shown in Table 1 that adopted from [3].

3 Results and Discussion

3.1 Simulation Results

The performance of the PWM-based SMVC and PID controller buck-boost converter is as shown in Figs. 4 and 5, respectively. During duty cycle at 0.7 as shown in Fig. 4,

there has been a steady-state condition in step-up operation with output voltage – 52.4 V. Besides, Fig. 5 reveals that there has been also a steady pattern but in step-down operation (duty cycle 0f 0.4) with output voltage – 15.11 V. The steady pattern which is the proof of the system is operating in CCM mode which being measured during the system located at inductor current. It is complying the theoretical buck-boost operation; where for duty cycle above 0.5, the operation is under boost mode, and for duty cycle below 0.5, the operation is under buck mode.

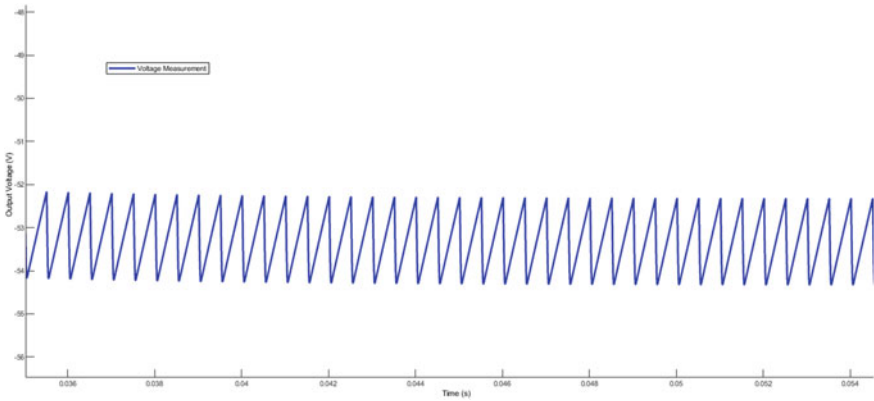


Fig. 4 Simulation result for buck-boost converter using SMVC in MATLAB/simulink

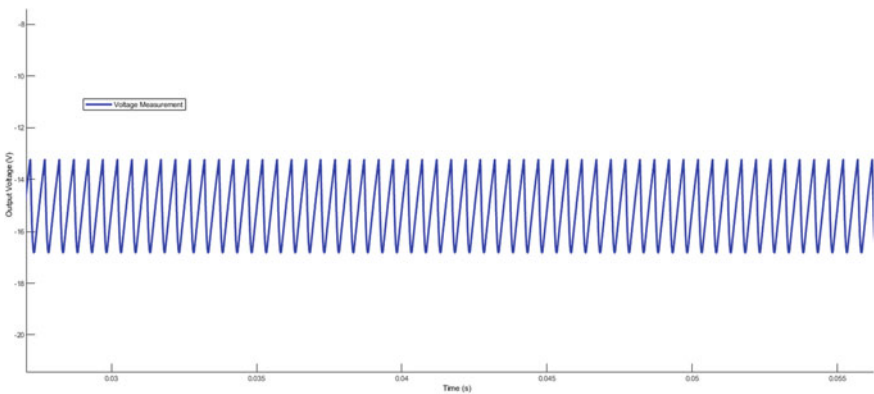


Fig. 5 Simulation result for buck-boost converter using SMVC with PID controller in MATLAB/simulink

3.2 Analysis with Three Different Solar Modules Using SMVC and PID Controller

In Figs. 6 and 7, the general pattern of output voltage will higher than input voltage (boost mode) when $D > 0.5$. While during $D < 0.5$, the output voltage will less than input voltage (buck mode). All the three solar modules shown that the higher the duty cycle, the higher output voltage of the converter either with PID controller or not.

Figures 8 and 9 illustrate the output voltage ripple of the system by using three different solar modules with varied switching frequency from 1 to 10 kHz under duty cycle 0.4 and 0.7, respectively. The output voltage ripple of the system is expected to decline when the switching frequency is higher. The notation $D, R, C,$ and f are the duty cycle, resistance, capacitance, and frequency of the converter, respectively. The formula been used for calculating the output voltage ripple, r , is given by

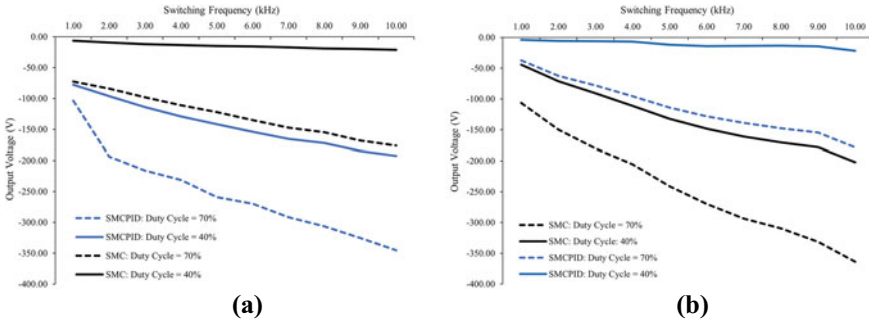
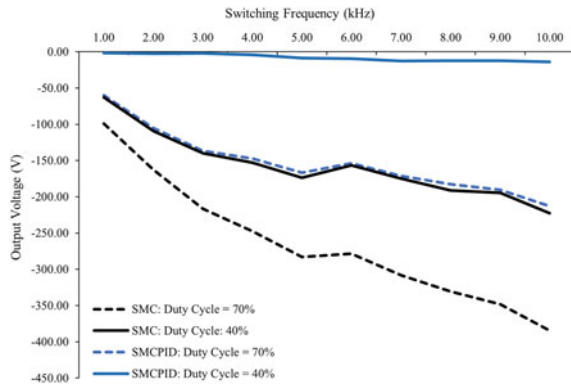


Fig. 6 Analysis of output voltage data result of **a** user-defined solar module and **b** Solartech energy solar module for buck-boost converter with SMVC and PID controller by varied the switching frequency

Fig. 7 Suntech energy module. Analysis of output voltage data result for buck-boost converter with SMVC and PID controller by varied the switching frequency



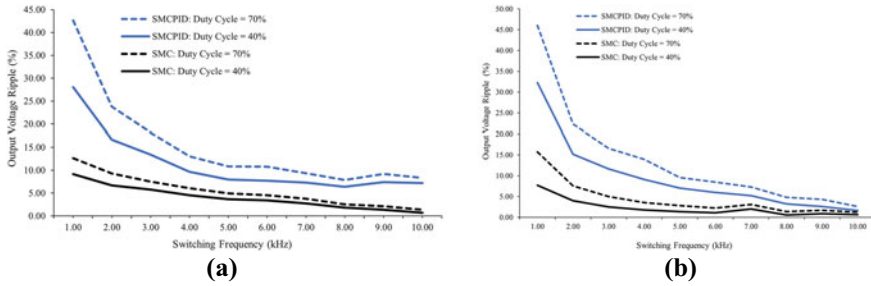
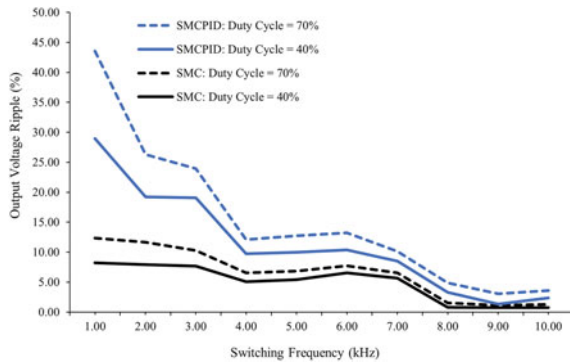


Fig. 8 Analysis of output voltage ripple data result: **a** user-defined solar module and **b** Solartech energy solar module for buck-boost converter with SMVC and PID controller by varied the switching frequency

Fig. 9 Analysis of output voltage ripple data result for buck-boost converter with SMVC and PID controller by varied the switching frequency under Suntech energy solar module



$$r = \frac{D}{RCf} = \frac{\Delta V_O}{V_O} \tag{4}$$

From the overall analysis for those solar modules, the switching frequency above 5 kHz shows the low output voltage ripple with PID controller which has overshoot output ripple voltage than sliding mode. It shows that the suitable type of DC–DC converter and solar system in any system is important to produce a stable output voltage. Taken together, the output voltage results can be concluded that the higher the duty cycle, the higher output voltage of converter either with PID controller or not. These results indicate that it is a robust and stable system which could be proven by the static performances. The output voltage ripple shows the overshoot voltage compared to conventional SMVC, but when it reaches to 2 kHz, the result shows the static and stable data.

4 Conclusion

Issue in DC–DC converter for fixed output voltage with and input voltage variations was studied and analyzed. A simple approach to design of fixed duty cycle (0.4 and 0.7) and varied switching frequency (1–10 kHz)-based SM voltage controller with and without PID controller for buck-boost converter type operating in CCM is presented. It can be concluded from the results that the output voltage of PWM-based SMVC is feasible for common conversion purposes such as solar application (i.e., user-defined, Solartech energy, and Suntech energy modules). The higher switching frequency (at 10 kHz), the output voltage ripple achieved below 5%.

References

1. Al-Qaisi MAF, Shehab MA, Al-Gizi A, Al-Saadi M (2019) High performance DC/DC buck converter using sliding mode controller. *Int J Power Electron Drive Syst* 10(4):1806
2. Arjyadhara P, Chitralkha J (2013) Analysis of solar PV cell performance with changing irradiance and temperature. *Int J Eng Comput Sci* 2(1):214–220
3. Tan S-C, Lai YM, Tse CK (2005) Design of PWM based sliding mode voltage controller for DC-DC converters operating in continuous conduction mode. In: 2005 European conference on power electronics and applications, vol 2005, pp 10
4. Aseem K, Selva KS (2020) Closed loop control of DC-DC converters using PID and FOPID controllers *11(3):1323–1332*
5. *Engineering I* (2015) Design of PWM-based sliding-mode control of boost converter with improved, pp 817–824

An Optimized Deep Learning Model for Automatic Diagnosis of COVID-19 Using Chest X-Ray Images



Suhaim Parvez Wadekar, Koon Meng Ang, Nor Ashidi Mat Isa, Sew Sun Tiang, Li Sze Chow, Chin Hong Wong, Meng Choung Chiong, and Wei Hong Lim

Abstract COVID-19 has caused havoc throughout the world in the last two years by infecting over 455 million people. Development of automatic diagnosis software tools for rapid screening of COVID-19 via clinical imaging such as X-ray is vital to combat this pandemic. An optimized deep learning model is designed in this paper to perform automatic diagnosis on the chest X-ray (CXR) images of patients and classify them into normal, pneumonia and COVID-19 cases. A convolutional neural network (CNN) is employed in optimized deep learning model given its excellent performances in feature extraction and classification. A particle swarm optimization with multiple chaotic initialization scheme (PSOMCIS) is also designed to fine tune the hyperparameters of CNN, ensuring the proper training of network. The proposed deep learning model, namely PSOMCIS-CNN, is evaluated using a public database consists of the CXR images with normal, pneumonia and COVID-19 cases. The proposed PSOMCIS-CNN is revealed to have promising performances for automatic diagnosis of COVID-19 cases by producing the accuracy, sensitivity, specificity, precision and F1 score values of 97.78%, 97.77%, 98.8%, 97.77% and 97.77%, respectively.

Keywords Automatic diagnosis · COVID-19 · Convolutional neural network · Chest X-ray · Hyperparameter learning · Particle swarm optimization

S. P. Wadekar · K. M. Ang · S. S. Tiang · L. S. Chow · M. C. Chiong · W. H. Lim (✉)
Faculty of Engineering, Technology and Built Environment, UCSI University, 56000 Kuala Lumpur, Malaysia
e-mail: limwh@ucsiuniversity.edu.my

N. A. M. Isa
School of Electrical and Electronics Engineering, Engineering Campus, Universiti Sains Malaysia, 14300 Nibong Tebal, Pulau Pinang, Malaysia

C. H. Wong
Maynooth International Engineering College, Fuzhou University, Fuzhou 350108, China

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
M. A. Abdullah et al. (eds.), *Advances in Intelligent Manufacturing and Mechatronics*,
Lecture Notes in Electrical Engineering 988,
https://doi.org/10.1007/978-981-19-8703-8_6

1 Introduction

COVID-19 is a highly infectious and contagious viral disease caused by a novel virus of SARS-CoV-2. The first COVID-19 case emerged from Wuhan, China, and was reported on December 2019. Over the past two years, this disease has rampaged throughout the world by causing substantial damages on economy as well as human life [1]. Referring to the updates provided by World Health Organization (WHO) as of March 2022, a total of 455,969,731 infection cases and 6,059,991 death cases are reported worldwide [2]. Some typical symptoms of COVID-19 include high fever, loss of taste and smell, low oxygen level, cough, dyspnea. Although reverse transcription polymerase chain reaction (RT-PCR) is considered as a golden test used for diagnosing COVID-19 accurately, it is a time-consuming process that can take up to 48 h to confirm the infected cases. The feasibilities of various clinical imaging techniques such as computed tomography (CT), magnetic resonance imaging (MRI) and chest X-ray (CRX) in diagnosing COVID-19 were subsequently explored to address the drawback of PCR test. While useful for the diagnose of COVID-19 cases, both of CT and MRI are more expensive procedures and not recommended to use for routine scanning due to the safety hazard of high radiation exposure [3]. Meanwhile, CXR is considered as a more affordable clinical imaging technique for most hospitals and radiology laboratories to achieve the good accuracy of COVID-19 diagnosis within shorter timeframe.

A computer-aided diagnosis (CAD) system incorporated with optimized deep learning model is designed in this paper to perform rapid diagnosis of COVID-19 cases by referring to the CXR images of patients. Deep learning is a subfield of artificial intelligence that enables the computers to extract meaningful information from various input sources (e.g., images, videos, speech, etc.) and perform the designated tasks such as classification [4, 5], fault detection [6, 7], etc. Convolutional neural network (CNN) is a popular deep learning model, and it is adopted in current study to diagnose the CXR images of patients and classify them into normal, pneumonia and COVID-19 cases. A crucial factor that governs the classification performance of CNN is the hyperparameter settings used during the network training process. For instance, CNNs may overlook some important features from inputs and produce inaccurate models if excessively low learning rates are set, whereas CNNs trained with excessively high learning rates might suffer with premature convergence issue. There are needs of developing systematic approaches to optimize the hyperparameter settings of CNN during training process for achieving good classification performance.

Metaheuristic search algorithms (MSAs) inspired by different natural phenomena are envisioned as promising approaches to tackle various real-world optimization problems [8–16] due to their desirable characteristics such as excellent global search ability and independent of gradient information. Particle swarm optimization (PSO) [17] is a popular MSA due to its simple implementation and fast convergence rate. Despite of its popularity, original PSO still suffers with premature convergence issue when solving complex optimization problems such as hyperparameter tuning of CNN. The random initialization scheme adopted by original PSO tends to produce

initial population with poor quality without considering surrounding information of search environment. Numerous enhanced PSO variants [18–23] were developed in past decades to achieve better balancing of exploration and exploitation searches.

The contributions of this paper are explained herein. A PSO with multiple chaotic initialization scheme (PSOMCIS) is first proposed to optimize the hyperparameters of CNN, where both ideas of multiple chaotic maps (MCM) and oppositional-based learning (OBL) are integrated to produce an initial population with better quality. An optimized deep learning model known as PSOMCIS-CNN is then developed to perform rapid diagnosis on the CXR images of patients and classify them into normal, pneumonia and COVID-19 cases with better accuracy. Finally, the classification performances of PSOMCIS-CNN in diagnosing COVID-19 cases are analyzed and proven more competitive than those of deep learning models optimized by other PSO variants.

2 Related Works

2.1 Existing Techniques for COVID-19 Diagnosis

Some existing works of applying AI techniques for COVID-19 diagnosis are presented in this subsection. In [24], five pretrained CNN models (i.e., DarkNet-19, ResNet-101, SqueezeNet, VGG-16 and VGG-19) were trained using transfer learning and reported to classify COVID-19 and normal cases with accuracies above 90%. An ensemble deep learning method was designed in [25] to diagnose COVID-19 by considering six pretrained networks (i.e., Xception, DenseNet201, ResNet152V2, InceptionResNetV2, NASNetLarge and VGG-16) via majority voting. A novel CNN known as STM-RENet was designed in [26] to identify the homogeneity and inhomogeneity regions through systematic use of convolutional process as well as region and edge implementations. A hybrid deep learning model based on CNN and gated recurrent unit (GRU) was proposed in [27], and it achieved the precision, recall and F1 score values of 0.96, 0.96 and 0.95, respectively. An optimized CNN known as OptCoNet was proposed in [3] and it was reported to perform COVID-19 diagnosis with accuracy, specificity, precision and F1 score values of 97.78%, 97.75%, 96.25%, 92.88% and 95.25%, respectively. In [28], a pretrained network of Xception trained with COVID-19 and pneumonia datasets via transfer learning can deliver the accuracy, precision and recall rate of 89.6%, 93% and 98.2%, respectively.

2.2 Original PSO

PSO is a popular MSA, and its search mechanisms are inspired by collective behavior of bird flocking to locate food source [17]. Suppose that I and

D refers to population size of PSO and dimensional size of problem, respectively. Each i th particle with velocity of $V_i = [V_{i,1}, \dots, V_{i,d}, \dots, V_{i,D}]$ and position of $X_i = [X_{i,1}, \dots, X_{i,d}, \dots, X_{i,D}]$ is considered as the candidate solution of given problem, where $i = 1, \dots, I$ and $d = 1, \dots, D$. Meanwhile, the personal best position of each i th particle is represented as $P_{\text{best},i} = [P_{\text{best},i,1}, \dots, P_{\text{best},i,d}, \dots, P_{\text{best},i,D}]$, and the global best particle in population is denoted as $G_{\text{best}} = [G_{\text{best},1}, \dots, G_{\text{best},d}, \dots, G_{\text{best},D}]$. At any $(t + 1)$ th iteration, the velocity and position of each i th particle in every d th dimension can be updated as follows:

$$V_{i,d}^{t+1} = \omega V_{i,d}^t + c_1 r_1 (P_{\text{best},i,d}^t - X_{i,d}^t) + c_2 r_2 (G_{\text{best},d}^t - X_{i,d}^t) \quad (1)$$

$$X_{i,d}^{t+1} = X_{i,d}^t + V_{i,d}^{t+1} \quad (2)$$

where ω is an inertia weight; c_1 and c_2 are acceleration coefficients; $r_1, r_2 \in [0, 1]$ are two random numbers generated with uniform distribution. For every i th particle, the fitness value of its updated position is evaluated as $f(X_i^{t+1})$ and compared with those of $P_{\text{best},i}^t$ and G_{best}^t that are evaluated as $f(P_{\text{best},i}^t)$ and $f(G_{\text{best}}^t)$, respectively. Both of $P_{\text{best},i}^t$ and G_{best}^t are replaced by X_i^{t+1} if the latter solution is more superior. The search process of PSO using Eqs. (1) and (2) is repeated iteratively until predefined termination conditions are met. Finally, the global best position G_{best} is returned, and its decision variables are decoded to solve the given optimization problem.

3 Proposed PSOMCIS-CNN for COVID-19 Diagnosis

3.1 Datasets

The image datasets used for training the proposed PSOMCIS-CNN are obtained from public database mentioned in [3], and they consist of 2,700 CXR images that are equally divided into 900 images for each category of normal, pneumonia and COVID-19. These image datasets are randomly partitioned with the ratio of 70% to 30% for training and testing the proposed PSOMCIS-CNN, respectively. All CXR images are resized to the resolutions of $224 \times 224 \times 3$ using data augmentation and converted into color images before they are used for the training or testing of model.

3.2 Formulation of Hyperparameter Tuning as Optimization Problem

A CNN architecture shown in Table 1 is incorporated into proposed PSOMCIS-CNN model to perform automatic diagnosis of COVID-19. Unlike the typical machine learning methods such as artificial neural network (ANN), CNN has deeper architecture that enables it to simultaneously perform feature extraction and classification. The feature extractor component of CNN contains the convolutional (Conv) layers, batch normalization (BN) layers, activation functions known as nonlinear rectified linear unit (ReLU) and maximum pooling (Max_Pool) layers. For each Conv layer, multiple numbers of filters are used to perform convolutional operation on input images in order to extract the desired features and map them into feature space via ReLU. The BN layers inserted between Conv layer and ReLU prevent overfitting issue by normalizing the gradients and activations, whereas Max_Pool layers are used for downsampling the feature maps obtained from Conv layers without compromising the crucial information of input images. Meanwhile, the classification component of CNN contains the fully connected (FC) layer and Softmax layer used to classify the input features into particular number of classes, i.e., three classes (normal, pneumonia and COVID-19) for current study.

An optimizer such as stochastic gradient descent (SGD) is typically used to train the PSOMCIS-CNN model based on given datasets by minimizing the predefined cost function via the adjustment of learnable parameters (i.e., weights and biases)

Table 1 CNN architecture incorporated into proposed PSOMCIS-CNN model

Layer type	Filter size	Numbers of filter	Stride size
Input	$224 \times 224 \times 3$	–	–
Conv + BN + ReLu	3×3	64	1×1
Max_Pool	2×2	–	2×2
Conv + BN + ReLu	3×3	64	1×1
Max_Pool	2×2	–	2×2
Conv + BN + ReLu	3×3	32	1×1
Max_Pool	2×2	–	2×2
Conv + BN + ReLu	3×3	16	1×1
Max_Pool	2×2	–	2×2
Conv + BN + ReLu	3×3	8	1×1
FC	3 classes	–	–
Softmax	–	–	–

contained in Conv and FC layers in iterative manners. Before training the PSOMCIS-CNN model using SGD, a number of hyperparameters need to be set appropriately to ensure proper training of PSOMCIS-CNN that can lead to robust classification performances. In this paper, the hyperparameters to be optimized when training PSOMCIS-CNN model are: (a) momentum $M \in [0.5, 0.9]$, (b) initial learning rate $L^{\text{ini}} \in [0.01, 0.1]$, (c) maximum epoch number $EP^{\text{Max}} \in [5, 20]$ and (d) L2 regularization $R^{\text{L2}} \in [1 \times 10e^{-4}, 5 \times 10e^{-4}]$. These hyperparameters to be optimized are encoded in the vector form of $X = [M, L^{\text{ini}}, EP^{\text{Max}}, R^{\text{L2}}]$. It is nontrivial to search for the best combination of these four hyperparameters that can provide proper training of PSOMCIS-CNN model to achieve robust classification performance due to its computational expensive nature. Therefore, a new PSO variant known as PSOMCIS is designed to optimize these hyperparameters in order to achieve the best network performance. A fitness function known as error rate $f^{\text{error}}(X)$ is defined to evaluate the quality of each PSOMCIS particle as follow:

$$f^{\text{error}}(X) = \frac{FP + FN}{TP + TN + FP + FN} \quad (3)$$

where TP , FP , TN and FN are the true positive, false positive, true negative and false negative values produced by the PSOMCIS-CNN model trained based on the hyperparameter settings as encoded in particle with position vector of X .

3.3 Hyperparameter Tuning Using PSOMCIS

Random initialization scheme is adopted by original PSO population without considering information of surrounding environment. It is possible to generate initial solutions at local optima that can lead to premature convergence or initial solutions far away from global optimum that slow down the convergence speed of algorithm [29]. To address this issue, both concepts of multi-chaotic maps (MCM) and oppositional-based learning (OBL) are incorporated into the initialization scheme of PSOMCIS, aiming to produce initial population with better fitness and diversity levels.

The ergodicity and non-repetitive characteristics of MCM are first adopted to generate a chaotic population $\mathbf{P}^{\text{CS}} = [X_1^{\text{CS}}, \dots, X_i^{\text{CS}}, \dots, X_I^{\text{CS}}]$ of PSOMCIS. Let $\gamma_0 \in [0, 1]$ be a randomly generated initial chaotic sequence and γ_z is chaotic sequence produced by chosen chaotic map at the z th sequence, where $z = 1, \dots, Z$ and Z is the final sequence number. Depending on the value of γ_0 , one of the following four chaotic maps is chosen to calculate the next sequence when initializing \mathbf{P}^{CS} , i.e., (a) circle map is selected if $\gamma_0 \leq 0.25$, (b) Gauss map is selected if $0.25 \leq \gamma_0 < 0.5$, (c) Singer maps is selected if $0.5 \leq \gamma_0 < 0.75$ and (d) sinusoidal maps is selected if $0.75 \leq \gamma_0 < 1$, where,

$$\mathbf{Circle} : \gamma_{z+1} = \left| \gamma_z + b - \left(\frac{a}{2\pi} \right) \sin(2\pi \gamma_z), 1 \right| \quad (4)$$

$$\mathbf{Gauss} : \gamma_{z+1} = \begin{cases} 1, \gamma_z = 0 \\ \frac{1}{|\gamma_z|}, \text{ otherwise} \end{cases} \quad (5)$$

$$\mathbf{Singer} : \gamma_{z+1} = \mu(7086\gamma_z - 23.32\gamma_z^2 + 28.75\gamma_z^3 - 13.302875\gamma_z^4), \mu = 1.07 \quad (6)$$

$$\mathbf{Sinusoidal} : \gamma_{z+1} = a\gamma_z \sin(\pi\gamma_z), a = 2.3 \quad (7)$$

Suppose that X_d^{\min} and X_d^{\max} are the lower and upper limits of decision variables in d th dimension. Define $X_{i,d}^{\text{CS}}$ is the d th dimension of every i th chaotic particle in \mathbf{P}^{CS} and γ_Z is the output of chosen chaotic map in final sequence, then,

$$X_{i,d}^{\text{CS}} = \gamma_z(X_d^{\max} - X_d^{\min}) + X_d^{\min} \quad (8)$$

An opposition population of $\mathbf{P}^{\text{OBL}} = [X_1^{\text{OBL}}, \dots, X_i^{\text{OBL}}, \dots, X_I^{\text{OBL}}]$ is subsequently generated using OBL. Given $X_{i,d}^{\text{CS}}$, the opposite solution $X_{i,d}^{\text{OBL}}$ is obtained as:

$$X_{i,d}^{\text{OBL}} = X_{i,d}^{\min} + X_{i,d}^{\max} - X_{i,d}^{\text{CS}} \quad (9)$$

After all solution members of \mathbf{P}^{CS} and \mathbf{P}^{OBL} are generated, a combined population $\mathbf{P}^{\text{Merge}}$ with the population size of $2I$ is generated by merging \mathbf{P}^{CS} and \mathbf{P}^{OBL} , i.e., $\mathbf{P}^{\text{Merge}} = \mathbf{P}^{\text{CS}} \cup \mathbf{P}^{\text{OBL}}$. All particles of $\mathbf{P}^{\text{Merge}}$ are sorted from the best to worst based on fitness values calculated using Eq. (3). The best I solution members of $\mathbf{P}^{\text{Merge}}$ are finally selected as initial population of PSOMCIS, i.e., $\mathbf{P}^{\text{Initial}} = [X_1, \dots, X_i, \dots, X_I]$.

The procedures used for optimizing the hyperparameters of PSOMCIS-CNN are presented in Fig. 1. Given $\mathbf{P}^{\text{Initial}}$, the personal best positions of all PSOMCIS particles and global best position of population are initialized. During the searching process, the velocities and positions of all particles are updated iteratively using Eqs. (1) and (2), respectively. The fitness value of new position generated by each PSOMCIS particle is calculated using Eq. (3) and compared with those of personal and global best positions. The latter two positions and their fitness values are replaced by those of new solution if it is more superior. The optimization process is terminated when the stopping criterion $\zeta > \zeta^{\max}$ is met, where ζ is the fitness evaluation counter and ζ^{\max} is the predefined maximum fitness evaluation number. Finally, the optimal hyperparameter settings of PSOMCIS-CNN can be obtained by decoding the global best position.

Algorithm 1: Hyperparameter Optimization Using PSOMCIS	
Inputs: $l, D, X_i^{max}, X_i^{min}, Z, \zeta^{max}$	
01:	Initialize $P^{CS} = \emptyset$ and $P^{OBL} = \emptyset$;
02:	for each i -th particle do
03:	for each d -th dimension do
04:	Randomly generate $\gamma_0 \in [0, 1]$ and set $z = 1$;
05:	if $\gamma_0 \leq 0.25$ then choose the Circle map as Eq. (5);
06:	else if $0.25 < \gamma_0 \leq 0.5$ then choose Gauss map as Eq. (6);
07:	else if $0.5 < \gamma_0 \leq 0.75$ then choose Singer map as Eq. (7);
08:	else if $0.75 < \gamma_0 \leq 1$ then choose Sinusoidal map as Eq. (8);
09:	end if
10:	while $z \leq Z$ do
11:	Update the chaotic variable, γ_z with the chaotic map chosen;
12:	Update chaotic sequence with $z \leftarrow z + 1$;
13:	end while
14:	Calculate $X_{i,d}^{CS}$ and $X_{i,d}^{OBL}$ using the Eqs. (9) and (10), respectively;
15:	end for
16:	Update $P^{CS} \leftarrow P^{CS} \cup X_i^{CS}$ and $P^{OBL} \leftarrow P^{OBL} \cup X_i^{OBL}$;
17:	end for
18:	Merge the P^{CS} and P^{OBL} to form P^{Merge} ;
19:	Perform evaluation for all the members in P^{Merge} and sort from best to worst based on their fitness values calculated using Eq. (4);
20:	Extract the first best l members from sorted P^{Merge} to form $P^{Initial}$;
21:	for each i -th particle do
22:	Randomly initialize the velocity V_i in all dimensions;
23:	Update $P_{best,i}, G_{best}, f^{error}(P_{best,i})$ and $f^{error}(G_{best})$;
24:	end for
25:	while $\zeta \leq \zeta^{max}$ do
26:	for each i -th particle do
27:	Update V_i and X_i using Eqs. (1) and (2), respectively;
28:	Fitness evaluation of new X_i using Eq. (4) to obtain $f^{error}(X_i)$;
29:	Update $P_{best,i}, G_{best}, f^{error}(P_{best,i})$ and $f^{error}(G_{best})$;
30:	$\zeta \leftarrow \zeta + 1$;
31:	end for
21:	end while
Output: $G_{best}, f^{error}(G_{best})$	

Fig. 1 Pseudocode of hyperparameter optimization using PSOMCIS

4 Performance Analysis of PSOMCIS-CNN

The performance of proposed PSOMCIS-CNN to classify the normal, pneumonia and COVID-19 cases based on given CXR images is evaluated and compared with four other CNN models optimized by original PSO [17], PSO without velocity

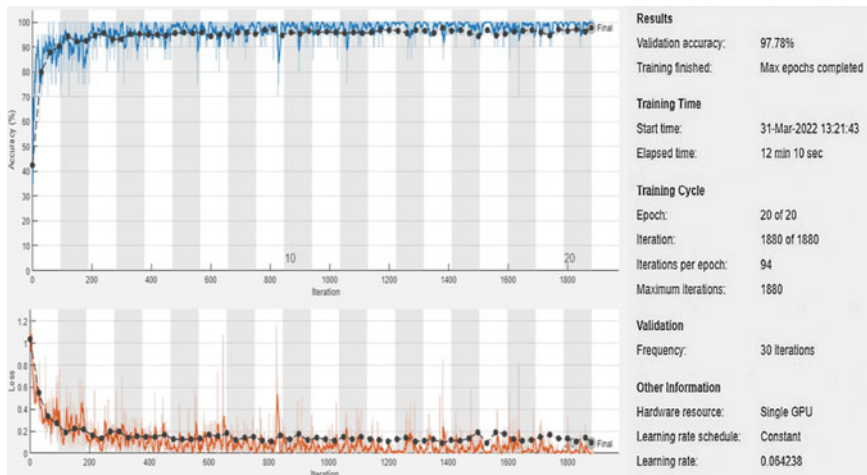


Fig. 2 Training progress of PSOMCIS-CNN

(PSOWV) [18], accelerated PSO (APSO) [19] and PSO with modified initialization scheme (PSOMIS) [20]. The CNN models optimized by different PSO variants are trained using 70% of datasets and tested using 30% of datasets. During the training process, the learnable parameters of CNNs (i.e., biases and weights) are adjusted using SGD, whereas the hyperparameters are optimized by the respective PSO variants. Figure 2 shows the training progress of proposed PSOMCIS-CNN and variation of its loss function value throughout the training process. Accordingly, PSOMCIS-CNN achieves a promising validation accuracy level of 97.79% at the end of training process.

Confusion matrices produced by different optimized CNN models when diagnosing COVID-19 cases with CXR images are shown in Fig. 3 to compare their classification performances. It is reported that the proposed PSOMCIS-CNN can completely dominate PSO-CNN and PSOWV-CNN by making more correct classification on all three cases (i.e., normal, pneumonia and COVID-19). Although PSOMIS-CNN can make slightly better classification for COVID-19 case than that of PSOMCIS-CNN, our proposed optimized deep learning model is much better when classifying the remaining two cases, i.e., normal and pneumonia. Finally, it is also observed that APSO-CNN has much worse performance than PSOMCIS-CNN when classifying normal case.

Five performance metrics known as accuracy, sensitivity, specificity, precision and F1 score are further used to compare the performances of CNN models optimized with PSOMCIS and other PSO variants as shown in Table 2, where the best result of each metric is highlighted as bold font. Evidently, the proposed PSOMCIS-CNN can produce best results in terms of accuracy, sensitivity, specificity, precision and F1 score in diagnosing COVID-19 cases from CXR images, suggesting the feasibility

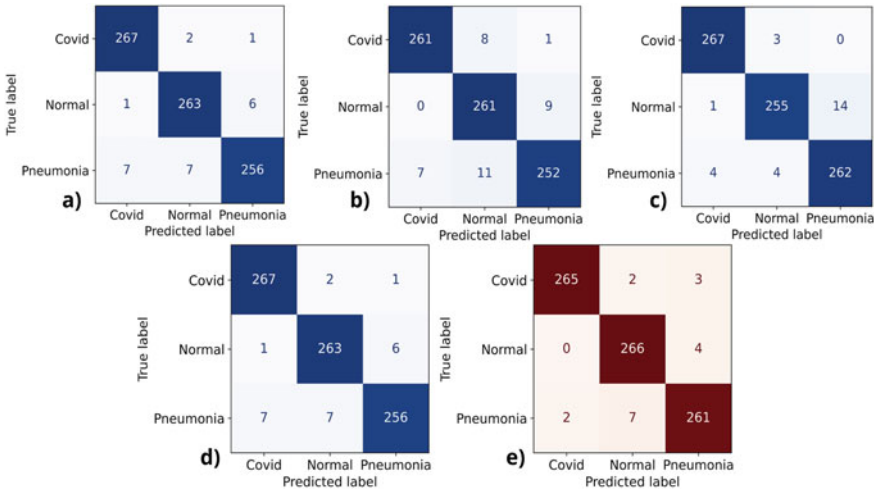


Fig. 3 Confusion matrices of **a** PSO-CNN, **b** PSOVV-CNN, **c** APSO-CNN, **d** PSOMIS-CNN and **e** proposed PSOMCIS-CNN

Table 2 Performances of different optimized deep learning models in COVID-19 diagnosis

Method	Accuracy	Sensitivity	Specificity	Precision	F1 Score
PSO-CNN	96.91	96.91	98.45	96.91	96.91
PSOVV-CNN	95.56	96.42	97.77	95.56	95.55
APSO-CNN	96.79	96.79	98.39	96.79	96.79
PSOMIS-CNN	97.03	97.03	98.51	97.03	97.03
PSOMCIS-CNN	97.78	97.77	98.88	97.77	97.77

of proposed PSOMCIS-CNN to assist radiologist and reduce burden of healthcare system.

Table 3 further compared the performances of PSOMCIS-CNN and other state-of-the-art methods in diagnosing COVID-19. The datasets, network architectures and results of these compared methods are extracted from their respective literatures, where “-” means the results were not reported. From Table 3, PSOMCIS-CNN can outperform all state-of-the-art methods in terms of accuracy, sensitivity, specificity, precision and F1 score, implying the robustness and effectiveness of proposed optimized deep learning model in COVID-19 diagnosis.

Table 3 Comparison of PSOMCIS-CNN with state-of-the-art methods in COVID-19 diagnosis

Method	Accuracy	Sensitivity	Specificity	Precision	F1 score
DarkNet-53 [24]	90.38	89.51	91.22	90.78	90.14
ResNet-101 [24]	91.75	91.49	92.00	91.49	91.49
SqueezeNet [24]	90.72	88.51	93.01	92.91	90.66
VGG-16 [24]	94.16	94.93	93.46	92.91	93.91
VGG-19 [24]	90.72	95.24	87.27	85.11	89.89
CB-STM-ReNet [26]	96.53	97.00	96.00	93.00	95.00
CNN-GRU [27]	96.00	–	–	96.00	96.00
Xception [28]	89.50	–	–	98.20	–
Proposed	97.78	97.77	98.88	97.77	97.77

5 Conclusions

An optimized deep learning model known as PSOMCIS-CNN is proposed to perform automatic diagnosis of COVID-19. A PSOMCIS is designed by leveraging the concepts of MCM and OBL to produce initial population with better quality for effective searching of optimal CNN hyperparameters. Simulation studies show that PSOMCIS-CNN can outperform other compared methods by producing the accuracy, sensitivity, specificity, precision and F1 score of 97.78%, 97.77%, 98.88%, 97.77% and 97.77%, respectively, when classifying normal, pneumonia and COVID-19 cases.

Acknowledgements This work was supported by the Ministry of Higher Education Malaysia under the Fundamental Research Schemes with project codes of FRGS/1/2019/TK04/UCSI/02/1 and FRGS/1/2020/TK0/UCSI/02/4. This work is also supported by the UCSI University Research Excellence & Innovation Grant (REIG) with project code of REIG-FETBE-2022/038.

References

- Ozili PK, Arun T (2020) Spillover of COVID-19: impact on the global economy. SSRN J
- COVID Live - Coronavirus Statistics - Worldometer. https://www.worldometers.info/coronavirus/?utm_campaign=Advegas1?. Accessed 12 Mar 2022
- Goel T, Murugan R, Mirjalili S, Chakrabarty DK (2021) OptCoNet: an optimized convolutional neural network for an automatic diagnosis of COVID-19. *Appl Intell* 51(3):1351–1366
- Voon YN, Ang KM, Chong YH, Lim WH, Tiang SS (2022) Computer-vision-based integrated circuit recognition using deep learning. In: Zain ZMd et al (ed) *Proceedings of the 6th international conference on electrical, control and computer engineering*, vol 842. Springer Singapore, Singapore, pp 913–925
- Jdid B, Lim WH, Dayoub I, Hassan K, Mohamed Juhari MRB (2021) Robust automatic modulation recognition through joint contribution of hand-crafted and contextual features. *IEEE Access* 9:104530–104546

6. Alrifayy M, Lim WH, Ang CK (2021) A novel deep learning framework based RNN-SAE for fault detection of electrical gas generator. *IEEE Access* 9:21433–21442
7. Alrifayy M et al (2022) Hybrid deep learning model for fault detection and classification of grid-connected photovoltaic system. *IEEE Access* 10:13852–13869
8. Yao L, Lai C-C, Lim WH (2015) Home energy management system based on photovoltaic system. In: 2015 IEEE international conference on data science and data intensive systems, Sydney, Australia, Dec 2015, pp 644–650
9. Yao L, Lim WH (2018) Optimal purchase strategy for demand bidding. *IEEE Trans Power Syst* 33(3):2754–2762
10. Yao L, Lim W, Tiang S, Tan T, Wong C, Pang J (2018) Demand bidding optimization for an aggregator with a genetic algorithm. *Energies* 11(10):2498
11. Yao L, Chen Y-Q, Lim WH (2015) Internet of Things for electric vehicle: an improved decentralized charging scheme. In: 2015 IEEE international conference on data science and data intensive systems, Sydney, Australia, Dec 2015, pp 651–658
12. Ang KM, Lim WH, Tiang SS, Ang CK, Natarajan E, Ahamed Khan MKA (2022) Optimal training of feedforward neural networks using teaching-learning-based optimization with modified learning phases. In: Isa K et al (ed) *Proceedings of the 12th national technical seminar on unmanned system technology 2020*, vol 770. Springer Singapore, Singapore, pp 867–887
13. Solihin MI, Lim WH, Tiang SS, Ang CK (2021) Modified particle swarm optimization for robust anti-swing gantry crane controller tuning. In: Zain ZMd et al (ed) *Proceedings of the 11th national technical seminar on unmanned system technology 2019*, vol 666. Springer Singapore, Singapore, pp 1173–1192
14. Hassan CS et al (2018) Crash performance of oil palm empty fruit bunch (OPEFB) fibre reinforced epoxy composite bumper beam using finite element analysis. *Int J Automot Mech Eng* 15(4):5826–5836
15. Hassan CS, Durai V, Sapuan SM, Abdul Aziz N, Mohamed Yusoff MZ (2018) Mechanical and crash performance of unidirectional oil palm empty fruit bunch fibre-reinforced polypropylene composite. *BioResources* 13(4):8310–8328
16. Hassan CS et al (2018) Effect of chemical treatment on the tensile properties of single oil palm empty fruit bunch (OPEFB) fibre. *Trends Text Eng Fash Technol* 3(2):1–7
17. Kennedy J, Eberhart R (1995) Particle swarm optimization. In: *Proceedings of ICNN'95—international conference on neural networks*, Perth, WA, Australia, vol 4, pp 1942–1948
18. El-Sherbiny MM (2011) Particle swarm inspired optimization algorithm without velocity equation. *Egypt Inform J* 12(1):1–8
19. Zhang H, Yang Z (2018) Accelerated particle swarm optimization to solve large-scale network plan optimization of resource-leveling with a fixed duration. *Math Probl Eng* 2018:1–11
20. Cheng W-L et al (2022) Particle swarm optimization with modified initialization scheme for numerical optimization. In: Zain ZMd et al (ed) *Proceedings of the 6th international conference on electrical, control and computer engineering*, vol 842. Springer Singapore, Singapore, pp 497–509
21. Karim AA, Mat Isa NA, Lim WH (2020) Modified particle swarm optimization with effective guides. *IEEE Access* 8:188699–188725
22. Lim WH et al (2018) A self-adaptive topologically connected-based particle swarm optimization. *IEEE Access* 6:65347–65366
23. Ang KM et al (2022) Modified particle swarm optimization with unique self-cognitive learning for global optimization problems. In: Nasir AFAb et al (ed) *Recent trends in mechatronics towards industry 4.0*, vol 730. Springer Singapore, Singapore, pp 263–274
24. Tang GS, Chow LS, Solihin MI, Ramli N, Gowdh NF, Rahmat K (2021) Detection of COVID-19 using deep convolutional neural network on chest X-ray (CXR) images. In: 2021 IEEE Canadian conference on electrical and computer engineering (CCECE), ON, Canada, Sept 2021, pp 1–6
25. Mun NW, Solihin MI, Chow LS, Machmudah A (2022) Pneumonia identification from chest X-rays (CXR) using ensemble deep learning approach. In: Zain ZMd et al (ed) *Proceedings of the 6th international conference on electrical, control and computer engineering*, vol 842. Springer Singapore, Singapore, pp 1139–1151

26. Khan SH, Sohail A, Khan A, Lee Y-S (2022) COVID-19 detection in chest X-ray images using a new channel boosted CNN. *Diagnostics* 12(2):267
27. Shah PM et al (2021) Deep GRU-CNN model for COVID-19 detection from chest X-rays data. *IEEE Access* 1
28. Khan AI, Shah JL, Bhat M (2020) CoroNet: a deep neural network for detection and diagnosis of COVID-19 from chest x-ray images. *Comput Methods Programs Biomed* 196:105581
29. Ahmad MF et al (2022) Differential evolution with modified initialization scheme using chaotic oppositional based learning strategy. *Alex Eng J* 61(12):11835–11858

Automatic Vehicle Location (AVL): Evaluation on the Punctuality Index of City Public Bus Service



Haziman Zakaria, Diyana Kamarudin, Faiz Azizul, Mohammad Fitri Idrus,
Nor Rokiah Hanum Md Haron, and Norhana Mohd Aripin

Abstract This study evaluates on-time bus service in ‘City A’, Malaysia using the punctuality index. Besides, the bus service reliability performance will also identify bus service punctuality characteristics for various operation conditions and routes in ‘City A’. City ‘A’ has been selected as this city lacks commuter and rail services, unlike Kuala Lumpur and the Klang Valley; thus, City A’s public bus service is in high demand. This article collected bus data using automatic vehicle location (AVL). Buses use SIM cards and mobile data networks to transmit location and time. GPS and fleet tracking system measure bus time and speed (FTS). Sampled bus system data was analysed to calculate punctuality indices for all routes. Different operating conditions affected bus punctuality. The punctuality index measures a stage bus’s service quality. Finally, the result of the study can be used to evaluate and improve public bus service.

Keywords Automatic vehicle location · Fleet tracking system · Punctuality index · Transit capacity and quality of service manual (TCQSM)

1 Introduction

Public transportation reduces traffic congestion, saves money, and saves time. Most neighbourhood residents take public transportation to work, school, or shopping. Buses are popular due to their low cost and large service area [1]. ‘City A’ is one of

H. Zakaria (✉) · D. Kamarudin · F. Azizul · M. F. Idrus · N. R. H. M. Haron · N. M. Aripin
Faculty of Industrial Management, Universiti Malaysia Pahang, 26300 Gambang, Pahang,
Malaysia
e-mail: haziman@zoho.com

D. Kamarudin
Faculty of Education and Liberal Studies, City University, 46100 Petaling Jaya, Selangor,
Malaysia

CoE for Artificial Intelligence and Data Science, Universiti Malaysia Pahang, 26600 Pekan,
Pahang, Malaysia

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
M. A. Abdullah et al. (eds.), *Advances in Intelligent Manufacturing and Mechatronics*,
Lecture Notes in Electrical Engineering 988,
https://doi.org/10.1007/978-981-19-8703-8_7

Peninsular Malaysia's fastest-growing commercial centres. The 2005 National Physical Plan identified 'City A' as Peninsular Malaysia's future growth centre [2]. Most bus companies in 'City A' are closing due to profit concerns and declining ridership. 'City bus services' were introduced in 'City A' in response to public transportation demand. There was no superior intercity service in 'City A' before 'city bus services'. Private bus companies around 'City A' ran at random times were late and provided ineffective services. Few people use this bus service and its accompanying services as a mode of transportation.

Even so, public transportation in most Malaysian cities, including 'City A', has not been able to compete with private cars. Most people would rather drive their car than take public transportation. Few of them say they are unhappy with how bad the public transportation services are. People did not use bus services because they did not have good amenities (like clean seats and air conditioning) and because being on time was important [3].

So, even though the government built Rapid KL, Rapid Penang, and Rapid Kuantan, people did not use public transportation. People who take public transportation are still few. If public transportation is not used much by many people, the bus company has to reduce service to save money. There is also what this means for people who usually take public transportation.

Sticking to its route is crucial for making sure customers are happy with its service. Even though buses often ran in 'City A', people had trouble getting where they needed to go because the buses were late. The fact that people on public buses are often uncomfortable makes them less reliable, making them less attractive than private cars. So, it is essential to find out if the public transportation system is working well enough. The output can be seen as a way for society to get the improvements it wants to make services better, meet needs, and make everyone happy.

Automatic vehicle location (AVL) systems allow authorities to track and observe moving targets surreptitiously. Using GPS satellite data, AVL systems can provide investigators with information on a vehicle's location, speed, and stopping point. One of Malaysia's integrated AVL systems is used in the 'City A' buses fleet tracking system. A complete AVL system uses GPS and other technologies (like CCTV video recording/online streaming and a passenger information system) to find buses and the mobile data network (MDN) to move data. The fleet tracking system keeps an eye on 'City A' buses. Every 15 s, it sends the location of each bus to the AVL centre. This information is then sent to the central server, which is used to improve service management and reliability, and the real-time passenger information system tells passengers.

Each cabin bus's on-board unit (OBU) is needed to find out where the bus is. It is a computer setup to get information about bus locations and events from different sources and store it. A GPS receiver on the roof of the bus sends the location of the vehicle every second. The system uses this information and algorithms for optimization and route/map matching to figure out exactly where the buses are.

The OBU keeps track of where a bus is along its route and what it is doing. It keeps track of things like how long the trip is and how many people are on it. Using a mobile data network, the OBU sends information about where the bus is to the master data

server (MDN). The location and activity data are then sent to the reporting central server through local area communications networks for post-processing, historical storage, and management reports.

One thing that affects the quality of bus service is how reliable it is. Reliability is a complicated idea that many things can define. The Malaysian Ministry of Transportation defines the reliability of public transportation as the number of morning rush-hour trips that are finished in less than an hour [4]. There is much evidence that shows how critical dependability is to the quality of transportation services [5]. Reliability can be described and judged from both the bus company's and the riders' points of view [6].

Several studies have also shown that bus service reliability is still an essential part of figuring out how good it is. Allen et al. [7] say that the reliability of the bus is vital for waiting time. They stressed how important it is for transportation companies to be on time and leave enough space between buses.

Accessibility is a measure of how easy it is to use transit to get around. It can be measured by the distance between transit stops or how long it takes to get from one stop to another. To find the right metric for public transportation, we need to find the ones that focus on how easy it is for people to get to transit where they live, where they work, and along routes that connect the two. As shown in the Table 1, two types of measures are made just for public transportation.

Using the transit capacity and quality of service manual (TCQSM), on-time performance and headway adherence can be used to measure dependability. Headway adherence means, amongst other things, how regular or even the time between bus trips is, how often trips are missed, and how many trips are passed up. People think that each bus has its schedule for when it will arrive at each bus stop. The quality of service framework for bus operations is shown in the Table 2. It shows the many ways that transit TCQSM can judge the quality of bus service.

Punctuality is measured by how much time has passed between the actual and expected arrival times. Timeliness is related to being on time. The headway evenness or adherence of interval service reliability criterion measures reliability in the same way a customer would [6].

Table 3 [8] shows how to make an index for how on-time bus operations are. Several types of punctuality indices were found. Each one was based on the number of bus stops and routes and included on-time performance and headway adherence.

In conclusion, the punctuality index (PI), which is also called 'on-time performance', is a metric that can be used in this study to measure reliability. The area chosen for this study is 'City A', a proliferating popular place for investment and tourism. Also, since bus companies have a monopoly, service reliability and integration are more accessible than in cities with more than one bus company. The specific research question that this study is trying to answer is:

RQ1: How well does the 'City A' bus service show up on time?

RQ2: How on time are the bus routes that 'City A' runs?

RQ3: Do you think there is a link between being on time and raining?

RQ4: How did other things affect how on-time the 'City A' bus service was?

2 Methodology

Automatic vehicle location (AVL) datasets were used to evaluate punctuality of bus route which collected through global positioning system (GPS) receiver inside on-board unit, which has been installed in 51 units buses. AVL dataset contains data for all trips that had scheduled departure time from the terminal between 5.20 a.m. and 11:00 p.m. The on-board unit is located in the driver's cabin and will transmit the data to bus control centre via mobile data network (MDN). The data will then be stored in fleet tracking system reporting database. In addition, the driver provided information on (1) the number of passengers alighting and boarding, (2) actual departure times and the scheduled and the actual, and (3) scheduled arrival times at each stop until the final destination. The passenger load, punctuality index, and travel time will be calculated using these data. Data collected is from 51 bus that covering 16 route and 580 trip daily for a period 7 month. The dataset was also divided into weekday and weekend to determine whether traffic condition has relationship with punctuality.

This study used CRISP-DM to analyse and process data of fleet tracking system (FTS) which contains punctuality, location, and the performance of the journey. CRISP-DM has 6 main phases (i.e. business understanding, data understanding, data preparation, modelling, evaluation, and deployment) which helps to narrow down the result and focus to the business user. The collected data was analysed to determine the transit capacity and quality of service manual standard (TCQSM), characteristics of bus service, and the punctuality index of each route.

To analyse the data, Microsoft Excel Analysis Tool was used to find the relationship between dependent variable (punctuality) and factors that affected the bus punctuality. The data is analysed by plotting the graph using Microsoft Excel Data.

3 Result

Punctuality refers to 5 min early or late in actual compare to scheduled time, and the punctuality percentage for the day has been calculated for each route. The quality of service was determined by comparing the punctuality value with TCQSM. The average punctuality performance percentage for 29 lines routes ranges from 37.1 to 91.8% with 76.4% on average. The monthly average punctuality route was recorded at 75.1% (June), 76.0% (July), 74.8% (August), 75.6% (September), 76.8% (October), 76.7% (November), and 78.4% (December). For level of service (LOS), only month of August was rated as F, whilst the rest of the month was rated as E. It is revealed that the average of punctuality index from June to December of 'City A' route is 76.4%, which is LOS E, means that every day there is one late transit service vehicle.

Using the percentage of punctuality service for each month, we can determine the level of service (LOS) rate for each route. According to the results, the lowest level of service (LOS) for most routes is C. Every week, more than one vehicle will be late for the average passenger at LOS 'C'. Line route which is always in Top 5 punctual

route from month June to December is line route 4002, followed by 3032, 2002, 4001, and 5001. Compressively, the punctuality indexes are not the same because traffic, driver, and passenger factors change randomly during the week. With a 52.4% punctuality rate, Route 200 is the most reliable route for being on time. For Route 401, the punctuality index is going down, whilst it changes throughout the month for other routes.

Bus route 400 is always on time. However, it varies greatly, which shows that the service is not always on time. Route 602 has the worst on-time performance because of traffic jams and many people taking it. These things make it take longer to get somewhere, so the bus might be late or miss the next trip. This situation proves that TCQSM's claim about the effect of traffic characteristics is valid [9]. Several drivers got to the station ahead of schedule and then left early or late. This behaviour goes against being on time and hurts your credibility. How on-time a public bus depends on the road conditions, the length of the route, the number of stops, the operations control strategies, the availability of the vehicle and crew, and the driving skills of the operators [10].

Hypothesis 1 measures the relationship between punctuality index and rainfall, and Fig. 1 shows the punctuality index versus the rainfall. Based on regression analysis from Fig. 1, bus operation punctuality has positive correlation with rainfall. When more rainfall, the punctuality also increases. This is could be when there is heavy rain, not many people are planning to going out using bus and also no traffic congestion. R2 is utilised as an indicator of fit quality. It indicates the number of points on the regression line. The result indicates that R2 is 0.83, indicating an excellent fit. In other words, the independent variables explain 83% of the dependent variables (y-values) over the independent variable (x-values). R2 is utilised as an indicator of fit quality. It indicates the number of points on the regression line. The result indicates that R2 is 0.83, indicating an excellent fit. In other words, the independent variables explain 83% of the dependent variables (y-values) over the independent variable (x-values). In addition, based on ANOVA regression analysis output, the significance F value is 0.004 which is less than 0.05 (5%). Figure 1 shows the result is reliable and statistically significant.

Second hypothesis measures the relationship between punctuality index and ridership. The one week punctuality data is use on this test to analyse the effect of traffic condition to the punctuality. P-value is the probability value in hypothesis testing to accept or reject null hypothesis. The alpha values used at 0.05 or 5% significant level. Based on t-test, one-tailed p-value = 0.004 and two-tailed p-value = 0.008. If p-value is less than 0.05, it suggests a significant differences between punctuality on weekend and weekday. Based on the result, Fig. 2 shows that the punctuality index on weekend is statistically higher than weekday.

Hypothesis 3 measures the relationship between punctuality index and ridership. Based on regression analysis from Fig. 3, the correlation coefficient result is 0.19 which shows a very weak linear relationship. So, the number of riders seems to have a negligible effect on the punctuality index. In addition, significance F value is 0.479 which is higher than 0.05 (5%) and resulted not significant between ridership and punctuality.

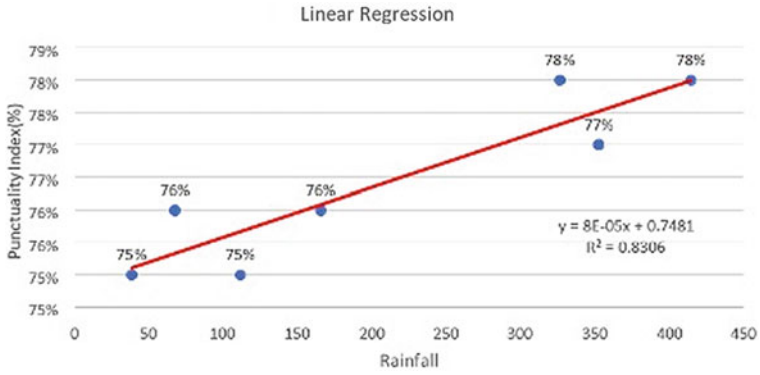


Fig. 1 Punctuality index (%) versus rainfall

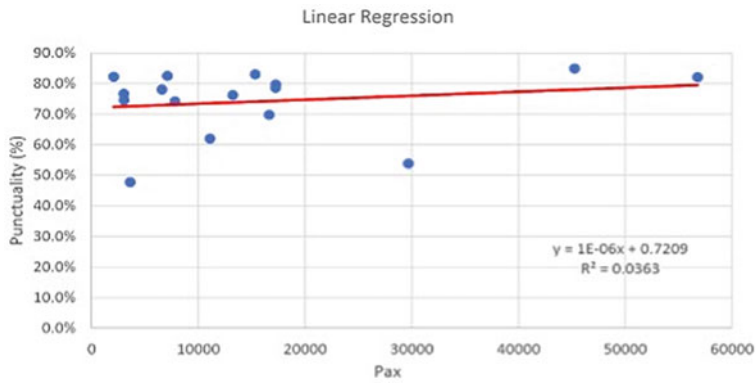


Fig. 2 Punctuality index (%) versus ridership

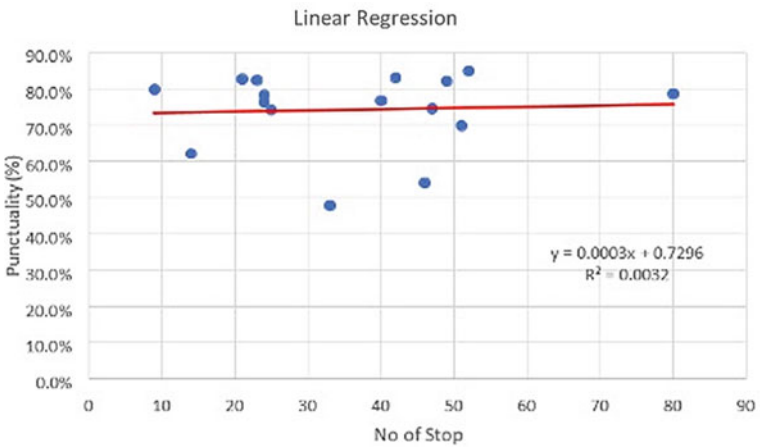


Fig. 3 Punctuality index (%) versus number of stops

The fourth hypothesis measures the relationship between punctuality index and number of stops. Based on regression analysis, the result of correlation coefficient is 0.05 showing no linear relationship since it is close to 0. As a result, the punctuality index has no significant different with the number of stops. Additionally, the significance F value is 0.83 which is higher than 0.05 (5%) and significantly no correlation.

4 Discussion

From a passenger's point of view, reliability is the ability of a system to stick to a schedule or keep regular headways and a predictable trip time. It is also known as 'arrival as scheduled' [11]. Unexpected or unscheduled vehicle arrivals make it hard to keep a steady pace or let vehicles pass through stations regularly [12]. There are some stops or stations where buses may be late. Consequently, they often leave later than planned. Because of these things, the bus service is not very reliable. Two important part that will help passenger when dealing with bus delay is customer-focused attitude from bus drivers and fast information.

Firstly, people in 'City A' complained that buses were always late and long wait times on some bus routes. They said that the company was running out of buses because so many of them were broken. Since there are fewer buses on some routes, the difference in headways is more considerable. Customers started to use taxis and e-hailing to get where they needed to go. Over the past year, the number of people taking the bus in 'City A' has steadily decreased. However, people on several routes were happy with how on-time and consistent the service was.

Secondly, customers know that the buses leave on time from what they have seen on the schedule. Because there are not enough buses, there is still a long wait between buses. Because there is much demand, some buses may leave before their scheduled time. A bus might be late if many kids wait at a station or frequently stop. Most people who take the bus are students, so the schedules are based on when they go to school.

Thirdly is driver behaviour. When the bus was full, some of the drivers left before they were supposed to. Because of this, people had to wait longer for the next trip. This problem will result in more people on board for the next trip. On the second trip, the speed was slowed because the bus would frequently stop to load and unload. For example, a bus trip that was supposed to leave at 12:30 p.m. left 30 min early because so many people were on it. The following bus will not come until 2:00 p.m. for those who arrive in the afternoon. These people will have to wait longer and will be added to the people on the 2 p.m. flight. People who cannot afford to wait for so long might choose a different way to get around, making public transportation less appealing in the long run.

Lastly is bus condition. The excellent condition bus must be in good shape and has 30 seats that recline and 34 places to stand air-conditioned. There are two automatic doors in the front and middle of the bus during rush hours that help people get on

and off. All buses use electronic ticketing machines (ETMs) to sell tickets. These machines print out a small piece of paper, and you can pay with cash or a credit card.

Most bus riders know that a bus's ability to arrive on time depends on traffic, road conditions, and things that cannot be helped, like accidents. On the other hand, passengers think that bus companies should focus on things they can control to cut down on delays. Bus delays can cause people to be late to work, pick up their kids late, or miss medical or other appointments. When delays and other problems happen, bus customers will benefit from two changes: getting necessary information and having bus drivers who care more about their customers. When the bus is late, people feel helpless. Many people on the plane think they cannot find out what is going on. 'Sometimes, it is hard to decide whether to stay or leave at bus stops', said one passenger. There is a big difference between what passengers have experienced and what they want from bus drivers. During delays and other problems, passengers want bus drivers to act as the company's customer service representative by giving information, showing empathy, and apologising. To close this gap, we need to teach drivers how to provide good customer service and look for ways to reduce the stress on drivers by getting them to talk to passengers directly.

To a lesser extent, planned roadwork and typical peak traffic congestion are seen by passengers as delays or disruptions, as are severe weather, accidents, emergency roadwork, and technical problems with the bus. Passengers also complain about delays caused by buses leaving early, driving by stops without picking up passengers, taking too long to buy tickets, and what they see as poorly designed schedules that cause services to 'cluster'. Some examples of disruptions the industry could expect were delays and changes to services. Some roads were closed or detoured because of roadwork or important local events. Traffic jams (especially during rush hours) and planned road work were the most common reasons for delays or other problems.

Passengers also mentioned other things they think cause problems and delays with bus service. These had to do with the bus's business, the driver, or other people on the bus most of the time.

(1) Buses often stop for short periods along the way because the driver needs a break or because they need to switch drivers, though this is not always the case. Passengers do not always know what is going on or why. Furthermore, the explanation is not always thought to make sense even when they do the problem. (2) Buses depart early from the schedule. Participants said that the driver might not have seen all of the people waiting for their specific vehicle and that they had seen buses drive by bus stops without stopping whilst people were waiting. (3) People who take too long to get on or pay are doing so because many people are getting on simultaneously, and they need to get payment change. (4) A human error happens, for example, is a driver who takes the wrong turn. Some of the people who took part in the study thought that timetables are not always set up well, which causes services to be grouped together and then spread out. This problem is especially true when more than one company serves the same area.

The drivers who took part in this study said that traffic jams, road closures, bad weather, accidents, mechanical problems, and passengers who were hard to deal with were all significant sources of disruption. When customers' trips are delayed or

have other problems, drivers often feel sorry for them, even if the customers do not know it. Drivers do not want to give accurate information if the situation is likely to change, so they do not. This situation means that the information is not correct, which could hurt them in the future. This problem is a real worry, but the people who took part in this study said they were aware of how road travel affects the ability of bus companies and drivers to give accurate information. Passengers thought that having some information was better than having none, as long as it was based on the best information that was available.

Both passengers and drivers suggested other ways to deal with delays and other problems and keep passengers informed. (1) Roadwork or events in the area that close roads or cause traffic jams. Local governments and bus companies need to talk to each other more about scheduling maintenance that causes minor inconveniences. Drivers and passengers, for example, would prefer that work on a stretch of road not be done all at once but instead at different times. Also, they would rather take a different route around road construction than drive through it. (2) Traffic jams and temporary traffic lights are a problem. Some passengers and drivers think that the lights could be timed better to give cars enough time to pass before letting traffic flow in the other direction. Several passengers and drivers asked that timetables include a 'rush-hour contingency' to make peak-hour schedules more realistic. Bus priority lanes were supported by drivers everywhere because they saw how well they worked where they were already in place. (3) Bus has mechanical issues. Some passengers think that technical problems are unacceptable because cars should be well-kept (whilst it might be naive to think nothing will ever break down, this is a genuine perception that bus companies need to be aware of). (4) Boarding time too long for one customer. Most passengers like new ways to buy tickets, like smartcards, which make the process faster. Many customers think that there should be more buses at certain times to reduce the number of people waiting to get on, which causes delays. In particular, these passengers think that more specialised school buses are needed so that other passengers do not have to wait when school is in session.

5 Conclusion

Punctuality is one of the most important ways to measure how well a bus service works. 76.4% of buses on the level of service E route arrive on time from the analysis. In December, Route 400 has a 91.8% (LOS B) punctuality index, whilst in June, it has a 37.1% (LOS F) index. These results show that the 'City A' bus service needs to get better at being on time.

The punctuality index ranges from 37.1 to 91.8%, with 76.4.8%. Statistical T-test result shows that punctuality on weekends is statistically higher than on weekdays. P-value with one tail is 0.004164, and P-value with two tails is 0.008328. In both cases, P-value is less than the alpha value, i.e. 0.05, thus can reject the null and assume a significant difference between punctuality on weekends and weekdays.

Thus from the result, it can be conclude that a common reason for unreliable service is that too many people are on the route. It makes travel times longer, makes bus rides unpredictable, raises costs, makes people less likely to trust buses, and reinforces negative ideas about them.

As an effect of rainfall on punctuality, based on regression analysis output, punctuality of bus operation has a positive correlation with the rainfall. This positive correlation could be when there is heavy rain, not many people plan to go out using the bus. When it is raining, many people will choose to travel by car because it is more comfortable.

Based on regression analysis output of effect ridership towards punctuality, the result correlation coefficient is 0.19 means a fragile linear relationship. The punctuality index seems to increase with the number of ridership slightly. Several things effect this result, such as traffic, weather, dwell time, and the number of people getting on or off the bus, can affect how well a transit bus system works along its route. Dwell time is the amount of time a vehicle stops for passenger service. It includes the time between when the doors opened and when they closed. The method of payment can affect how long the bus stays in one place [13, 14]. Paying fares affect how long a bus stays in one place. City A bus will take both cash and credit cards. Since the bus only has one driver and no helper, they only can provide limited ways to pay. The driver must be the one to make money and give out tickets. Customers have to wait in line at the door, making boarding take longer.

As an effect of the number of stops towards punctuality, regression analysis output indicates that the correlation coefficient is 0.05 means no linear relationship since it is very close to 0. A study suggested that the platform crowding pattern significantly affects dwell time [15].

It makes it hard for people to move around and hard to see approaching buses. Also, when more people are waiting on the platform, there is a higher chance that more people will get on the same bus. This condition causes bus stops to get crowded, and people have to wait longer for the bus. Also, if the station is full of people, the bus may not be able to see passenger waiting. This problem could make people slower to react when the regular bus comes, which would make them stay longer.

The study says that the punctuality index can measure how reliable mixed-traffic fixed-route bus services are [16]. Punctuality index studies are often used to measure how well bus routes and bus companies serve their customers. If bus companies could lose government subsidies based on how on time they were, they would try to improve how on time they were [17]. When figuring out how reliable public bus service is, the punctuality index is just one of many things to think about.

More factors should be looked at to expand the scope of bus reliability research, such as how often or how often buses run. In addition, improving punctuality is helpful to passengers to reduce the waiting time at bus stops or make reasonable travel arrangements before making a trip. However, for this to be effective, the information provided to passengers should be reliable and accurate.

Acknowledgements This research is funded by UMP PGRS 2003162: Determinant of Malaysia SME Innovation Action Based on The Triple Helix Model.

Appendix

Table 1 Type of metrics from guide to sustainable transportation performance measures, EPA

Type of metrics	Description	Metrics identified
Distance to transit stops	These metrics capture the main office commuters based on locations, population, trip origins, or trip destinations within a certain radius of a transit stop	(1) Per cent of daily/peak period trips (origins and destinations) starting or ending within 100 m of a transit stop. (2) Per cent of population and employment within 100 m of transit
Destinations accessible by transit	These metrics capture not just the accessibility of transit stops, but the connection that transit provides to various destinations	(1) Number of households within a 30-min transit ride of major employment centres/city centre. (2) Percentage of work and education trips accessible in less than 30 min transit travel time. (3) Percentage of workforce that can reach their workplace by transit within one hour with no more than one transfer

Table 2 Quality of service framework

	Service measure		
	Transit stops	Route segments	System
Availability	Frequency	Hours of service	Service coverage
Comfort and convenience	Passenger load	Reliability; (1) on-time performance and (2) headway adherence	Transit-auto travel time

Table 3 Punctuality indexes

Punctuality index	Description
P1	Shows how much time has passed between the actual arrival time and the time that was planned (adherence)
P2	Shows how much time has passed between the actual headway and the planned one (regularity) destinations
P3	An index that shows how long it takes between the average headway of a day and the average headway of the next bus (evenness)

References

1. Hakimi Ibrahim AN, Borhan MN, Mat Yazid MR, Rahmat RA, Yukawa S (2021) Factors influencing passengers' satisfaction with the light rail transit service in alpha cities: evidence from Kuala Lumpur, Malaysia using structural equation modelling. *Mathematics* 9(16). <https://doi.org/10.3390/math9161954>
2. Abu Bakar MF, Norhisham S, Ming Fai C, Baharin NL (2021) Evaluating the quality of service for bus performance in Kuantan. *Int J Acad Res Bus Soc Sci* 11(2):1342–1351. <https://doi.org/10.6007/IJARBS/v11-i2/9209>
3. Ponrahono Z, Bachok S, Osman MM, Ibrahim M (2016) Sustaining existing and prospective passengers on urban public buses: the case study of rapid Kuantan, Pahang, Malaysia. *Int J Sustain Futur Hum Secur* 4(2):22–29. <https://doi.org/10.24910/JSUSTAIN/4.2/2229>
4. Malaysian Government (2012) National land public transport masterplan final draft. Putrajaya. <https://govdocs.sinarproject.org/documents/prime-ministers-department/land-public-transport-commission/national-land-public-transport-master-plan-final-draft.pdf/view>. Accessed 30 May 2022
5. Batley R et al (2019) New appraisal values of travel time saving and reliability in Great Britain. *Transportation (Amst)* 46(3):583–621. <https://doi.org/10.1007/S11116-017-9798-7>
6. Napiah M, Kamaruddin I, Suwardo (2011) Punctuality index and expected average waiting time of stage buses in mixed traffic. *WIT Trans Built Environ* 116:215–226. <https://doi.org/10.2495/UT110191>
7. Allen J, Muñoz JC, de Dios Ortúzar J (2019) Understanding public transport satisfaction: using Maslow's hierarchy of (transit) needs. *Transp Policy* 81:75–94. <https://doi.org/10.1016/J.TRA.NPOL.2019.06.005>
8. Young Kho S, Sik Park J, Ho Kim Y, Ho Kim E (2005) A development of punctuality index for bus operation. *J East Asia Soc Transp Stud* 6:492–504. https://www.researchgate.net/publication/237260026_A_development_of_punctuality_index_for_bus_operation. Accessed 30 May 2022
9. Tang T, Fonzone A, Liu R, Choudhury C (2021) Multi-stage deep learning approaches to predict boarding behaviour of bus passengers. *Sustain Cities Soc* 73. <https://doi.org/10.1016/J.SCS.2021.103111>
10. Pitka P, Simeunović M, Tanackov I, Savković T (2017) Deterministic model of headway disturbance propagation along an urban public transport line. *Teh Vjesn* 24(4):1147–1154. <https://doi.org/10.17559/TV-20151126111613>
11. Sukor NSA, Airak S, Hassan SA (2021) 'More than a free bus ride'—exploring young adults' perceptions of free bus services using a qualitative approach: a case study of Penang, Malaysia. *Sustainability* 13(6). <https://doi.org/10.3390/SU13063294>
12. Morri N, Hadouaj S, Ben Said L (2020) Intelligent regulation system to optimize the service performance of the public transport. In: *ICEIS 2020—Proceedings of 22nd international conference on enterprise information systems*, vol 1, pp 416–427. <https://doi.org/10.5220/0009416104160427>
13. Fletcher G, El-Geneidy A (2013) Effects of fare payment types and crowding on dwell time. *Transp Res Rec* 2351:124–132. <https://doi.org/10.3141/2351-14>
14. Chen G, Chen W, Zhang S, Zhang D, Liu H (2020) Influence of mobile payment on bus boarding service time. *J Adv Transp* 2020. <https://doi.org/10.1155/2020/9635853>
15. Moosavi SMH, Yuen CW, Yap SP, Onn CC (2020) Simulation-based sensitivity analysis for evaluating factors affecting bus service reliability: a big and smart data implementation. *IEEE Access* 8:201937–201955. <https://doi.org/10.1109/ACCESS.2020.3036285>
16. Jilu Joseph V (2015) Punctuality index for the city bus service. *Int J Eng Res* 4(4):206–208
17. Norhisham S et al (2019) Service frequency and service hours evaluation for bus service in West Klang Valley. *IOP Conf Ser Mater Sci Eng* 636(1). <https://doi.org/10.1088/1757-899X/636/1/012008>

Bearing Fault Diagnosis Using Extreme Learning Machine Based on Artificial Gorilla Troops Optimizer



M. Firdaus Isham, M. S. R. Saufi, M. D. A. Hasan, W. A. A. Saad, M. Salman Leong, M. H. Lim, and Z. A. B. Ahmad

Abstract Bearing diagnosis is important to ensure smooth machinery operation and safety. Machine learning methods have been used widely in bearing diagnosis study, and one of the recent methods used is an extreme learning machine (ELM). The ELM method offers ease of implementation, rapid learning rate, and better generalization performance. However, the ELM method may lead to inaccurate diagnosis due to inappropriate value selection for neuron number, input weight, and hidden layer bias. Therefore, this paper proposed a new bearing diagnosis using ELM-based gorilla troops optimizer (GTO) method, named as GTO-ELM. The GTO method was used to select an optimized parameter for the ELM method. Two sets of bearing datasets from experimental work and online database were used in this study for evaluation on the proposed method. Both datasets have four different type of operation condition which are normal (baseline), inner race fault, outer race fault, and ball fault. Based on diagnosis result, the proposed GTO-ELM was able to surpass whale optimization algorithm (WOA)–ELM in term of convergence speed and conventional ELM methods in term of diagnosis performance with almost 10–12% better performance.

Keywords Fault diagnosis · Extreme learning machine · Optimization

1 Introduction

Bearing is one of the most important components in rotating machinery equipment in many industries such as aviation, energy, transportation, construction, and oil and gas. In order to optimize the production rate, bearing component may subject

M. F. Isham (✉) · M. S. R. Saufi · M. D. A. Hasan · W. A. A. Saad · Z. A. B. Ahmad
School of Mechanical Engineering, University Teknologi Malaysia, Sultan Ibrahim Chancellery Building, Jalan Iman, 81310 Skudai, Johor, Malaysia
e-mail: mfirdausisham@gmail.com

M. S. Leong · M. H. Lim
Institute of Noise and Vibration, Universiti Teknologi Malaysia Kuala Lumpur, Kampung Datuk Keramat, 54000 Kuala Lumpur, W. P. Kuala Lumpur, Malaysia

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
M. A. Abdullah et al. (eds.), *Advances in Intelligent Manufacturing and Mechatronics*,
Lecture Notes in Electrical Engineering 988,
https://doi.org/10.1007/978-981-19-8703-8_8

to long operation under harsh environment and sometimes exceeding its designed lifecycle [1]. This will lead to bearing failure that can cause serious harm to people, organization, and economy. Generally, bearing may fail due to inner race fault, outer race fault, ball fault, etc. [2, 3]. Vibration signal is usually used in order to detect bearing failure due to its ability to distinguish normal and abnormal conditions by referring to the amplitude and vibration pattern [4]. In this study, vibration signals were used in the proposed bearing diagnosis method known as GTO-ELM for bearing fault detection and to improve safety, reliability, and durability of rotating machinery equipment.

Traditional bearing diagnosis was mainly depending on human interpretation on frequency spectrum [1]. This approach mainly leads to inaccurate interpretation due to human error. Recently, automated bearing diagnosis has gained attention due to its ability to provide a good diagnosis result without human intervention. Hence, a lot of bearing diagnosis approaches have been proposed such as WOA-ELM and VMDEA [1, 5, 6]. However, the demands to produce more advanced and more accurate diagnosis models had been gained attention within researchers with a lot of new algorithms had been proposed to be tested in different studies [5, 7–11].

In 2006, a novel single-hidden layer feedforward neural network known as ELM method had been proposed by Huang et al. [12]. The ELM method provides better generalization performance and faster learning rate as compared with support vector machine (SVM) and artificial neural network (ANN) methods [13]. The ELM method has been applied in many areas of study which are machinery diagnosis [14–18], air quality forecasting [19], electric load forecasting [20], and machining [21]. This ELM method also had been used for prognostic study with its ability to solve regression problems [21]. However, the performance of the ELM method is depending on its three main parameters which are number of hidden neurons, input weight, and hidden layer bias. To date, there are numerous studies that have been done to solve this problem by combining the ELM method with meta-heuristic or optimizer methods. Gray wolf optimizer (GWO) is one of the meta-heuristic algorithms that have been used to solve ELM problems. For example, Thammasakorn et al. have used GWO-ELM for imbalance data classification [22], Zhou et al. have proposed GWO-ELM for carbon price forecasting [23], Xiao et al. have proposed GWO-ELM for shaft orbit identification [24], Sales et al. have proposed GWO-ELM for water lake depth modeling [25], Shariati et al. have proposed GWO-ELM for compressive strength of concrete strength [26], Li et al. have proposed GWO-ELM for multi-domain gearbox fault diagnosis [27], and Yao et al. have proposed GWO-ELM for hybrid gearbox fault diagnosis [28]. Whale optimization algorithm (WOA) is the latest optimizer which was used in order to solve the ELM problem. For instance, Nayak et al. have proposed WOA-ELM method for pathological brain detection [29], Sun and Zhang have proposed WOA-ELM method for carbon price forecasting [30], Isham et al. have proposed WOA-ELM for gearbox fault diagnosis [6], and Sun and Wang have proposed WOA-ELM for CO₂ prediction and analysis [9].

Therefore, this paper aims to propose the combination GTO and ELM where the GTO method is mainly used to correctly select the three parameter values of ELM method. The GTO method was recently proposed by Abdollahzadeh et al. [31]. The

GTO method was able to provide better convergence and provide better or competitive performance as compared with other optimizer [31]. To the best of our knowledge, the GTO method only had been applied in electrical system optimization [32]. The GTO method was inspired by the behaviors of gorillas group and described in two main strategies which are exploration phase and exploitation phase. Three strategies are used in exploration phase which are migration to undiscovered area, moving to a different gorilla group and migration to an identified location [31]. Two strategies are used in exploitation phase which are follow the silverback and competition for adult females [31]. To the best of our knowledge, this is the first paper applying the GTO method to improve the ELM method for bearing diagnosis application.

This paper is described from the following aspect. The theoretical background of the methods used was briefly explained in Sect. 2. Experimental procedure and bearing datasets were described in Sect. 3. Parameter optimization and bearing diagnosis study presented in Sect. 4. In Sect. 5, the conclusion and future direction are finally given.

2 Methods

2.1 The Proposed Bearing Diagnosis Method

The proposed diagnosis method is basically a combination of GTO and ELM method which have been fully described in Fig. 1. A modification on fitness function for GTO method is required in order to hybrid this method with ELM method. The minimization fitness function for GTO method is basically based on the training and testing accuracy as described in Eq. 1, where $Tr_{accuracy}$ is training accuracy, and $Te_{accuracy}$ is testing accuracy.

$$fitness_{minimization} = 1 - ((Tr_{accuracy} + Te_{accuracy})/2) \quad (1)$$

2.2 Extreme Learning Machine

ELM is a simple and efficient method to train single-hidden layer feedforward neural networks (SLFNs), as proposed by Huang et al. [12]. The ELM method helps to resolve the problem of conventional gradient-based algorithms for SLFNs [12]. The ELM method mainly consists of input layer, hidden layer, and output layer as shown in Fig. 2, where x_i is a feature, w_i is input weight, b is bias. Then, H matrix was produced as described in Eqs. 2 and 3.

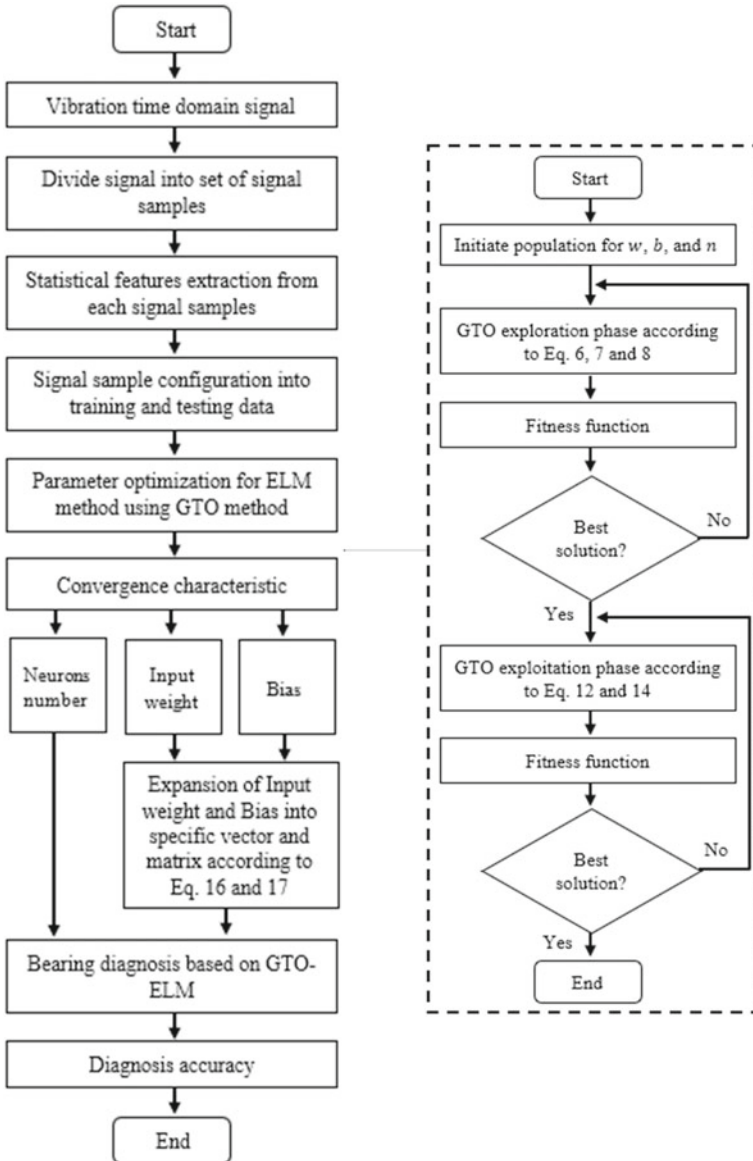
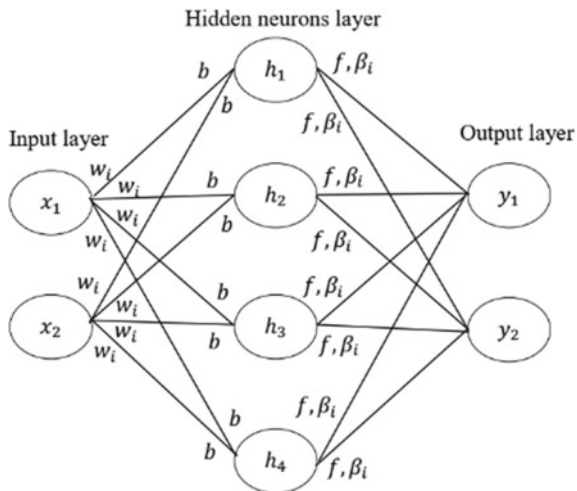


Fig. 1 Proposed GTO-ELM method

Fig. 2 ELM topology



$$h_1 = (w_1x_1 + b) \times f \tag{2}$$

$$H = (w_{m,n}x_{m,n} + b) \times f \tag{3}$$

$$y = H\beta \tag{4}$$

$$y_i = \sum \beta_i f[w_i x_i + b] \tag{5}$$

Then, the output matrix was formed as described in Eqs. 4 and 5 where y_i is an output layer, and β_i is an output weight.

2.3 Artificial Gorilla Troops Optimizer

The GTO method was inspired by the behavior of gorillas group social intelligence in nature. In this GTO method, all gorillas considered as a possible solution in GTO. The best possible solution at each optimization stage is considered as a silverback gorilla. Basically, GTO consists of two phase which are exploration phase and exploitation phase. For exploration, there are three strategies used. First, the migration of gorillas to undiscovered area which mainly to increase the exploration of GTO as described in Eq. 6.

$$GX(t + 1) = (UL - LL) \times r_1 + LL, \text{Rand} < p \tag{6}$$

Second, moving to a different group of gorillas which mainly to balance between exploration and exploitation as described in Eq. 7.

$$GX(t + 1) = (r_2 - C) \times X_r(t) + L \times H, \text{Rand} \geq 0.5 \quad (7)$$

Third, migration to the identified location which mainly to increase GTO capability to search for diverse optimization spaces as described in Eq. 8.

$$GX(t + 1) = X(i) - L \times (L \times (X(t) - GX_r(t)) + r_3 \times (X(t) - GX_r(t))), \text{Rand} < 0.5 \quad (8)$$

where $GX(t + 1)$ is the possible solution position vector of gorilla in the following t iteration, $X(t)$ is the current vector position of gorilla, while Rand , r_1 , r_2 , and r_3 are the random values between 0 and 1. The parameter, p , is the probability of choosing the migration strategy to an unidentified position and must be specified between 0 and 1 before optimization process start. X_r is a random selected member of gorilla, and GX_r is a random selected vector of gorilla possible solution position. LL and UL present lower limit and upper limits of the variables accordingly. For C , L and H can be mathematically represented according to Eqs. 9–11.

$$C = (\cos(2 \times r_4) + 1) \times (1 - t/\max(t)) \quad (9)$$

$$L = C \times l \quad (10)$$

$$H = Z \times X(t) \quad (11)$$

where r_4 is a random value between 0 and 1, l is a random value between -1 and 1, and Z is a random value between $-C$ and C . If $\text{Rand} < p$, the first strategy is chosen, if $\text{Rand} \geq 0.5$, the second strategy is chosen, and if $\text{Rand} < 0.5$, the third strategy is choosing accordingly. The best solution from exploration phase will become the silverback.

For exploitation, there are two strategies which were used. First strategy is following the silverback (chosen gorilla). This strategy will be selected if $C > W$, where W is the random parameter set. The chosen gorilla (silverback) is a leader that makes decision and guide other to find a food. This behavior can be described in Eqs. 12 and 13, where $g = 2^L$.

$$GX(t + 1) = L \times M \times (X(t) - X_{\text{silverback}}) + X(t) \quad (12)$$

$$M = \left(\left| (1/N) \sum_{i=1}^N GX_i(t) \right|^g \right)^{1/g} \quad (13)$$

Second strategy is the competition for adult female gorilla. This strategy will be selected if $C < W$. In nature, young male gorilla will compete violently with other males in selecting female gorilla. It can be described in Eq. 14.

$$GX(t + 1) = X_{\text{silverback}} - (X_{\text{silverback}} \times Q - X(t) \times Q) \times A \quad (14)$$

$$Q = 2 \times r_5 - 1 \quad (15)$$

$$A = \beta \times E \quad (16)$$

where Q is the impact force, r_5 is a random value between 0 and 1, A is a vector which indicates the degree of violence in case of conflict, β is specified set value before optimization maneuver, and E is the violence impact on the solution dimensions. The best solution from exploitation will become the new silverback (chosen gorilla either from exploration phase or new gorilla selected). The full flow of GTO method can be accessed in Fig. 1.

3 Procedure

Two sets of bearing datasets were used in order to validate the proposed GTO-ELM method. One set from experimental data and another one is from online bearing data from Case Western Reserve University (CWRU). Both datasets consist of four type of bearing condition which are healthy (baseline), inner race fault, outer race fault, and ball fault.

3.1 Experimental Bearing Data

The experimental data were extracted from Machinery Fault and Rotor Dynamics Simulator (MFS-RDS) test rig as shown in Fig. 3.

The test rig was operated at 30 Hz with three accelerometer sensors located on bearing housing. The sampling frequency was set at 24.6 kHz. The faults were artificially induced with a defect size of 1.5 mm on inner race, outer race, and ball. Figure 4a–c presents the components with the induced fault accordingly.

Vibration signals were then extracted for healthy (baseline), inner race fault, outer race fault, and ball fault condition. Then, the signals were divided into 150 signal samples for each bearing condition with 1000 data point for each sample. Total of 600 signal samples were then divided into training and testing data for diagnosis purpose. The data configuration for training and testing was shown in Table 1 accordingly. Each bearing condition was also labeled with a specific class accordingly. A ratio

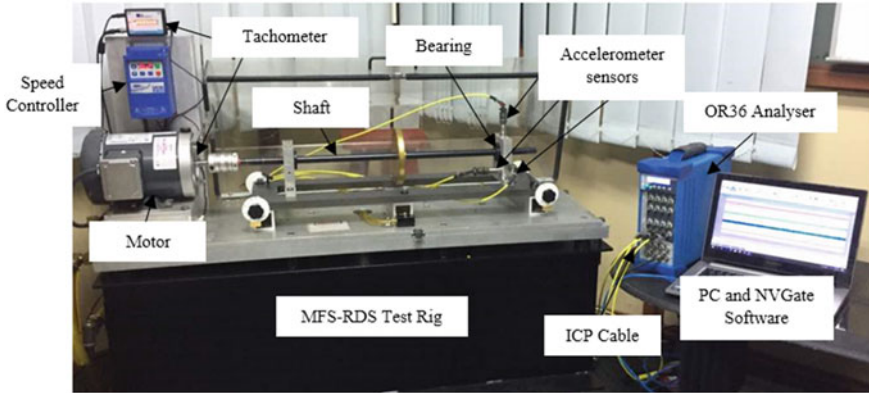
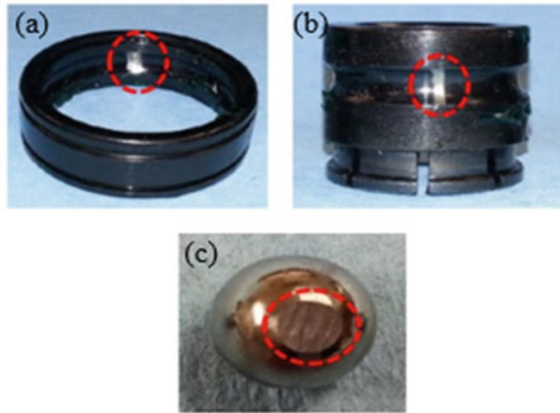


Fig. 3 Machinery Fault and Rotor Dynamic Simulator (MFS-RDS)

Fig. 4 Bearing components with induced fault, **a** inner race fault, **b** outer race fault, and **c** ball fault



of 70% training and 30% testing data distribution were used to distribute the data accordingly.

Figure 5 shows the vibration signal generated from MFS-RDS test rig for each bearing condition.

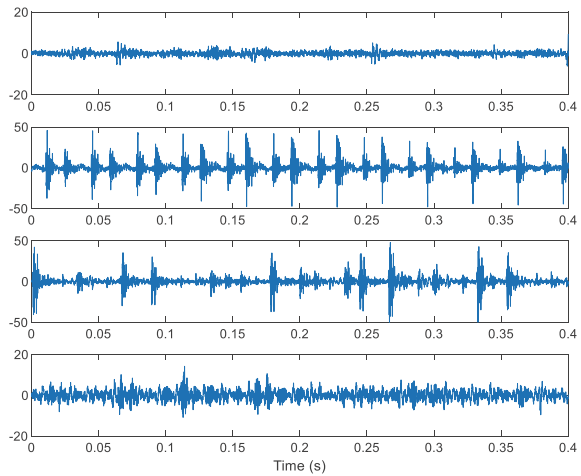
3.2 CWRU Bearing Data

The datasets were downloaded from CWRU data center Website. The datasets have same condition as experiment data where it consist of four types of bearing condition which are healthy (baseline), inner race fault, outer race fault, and ball fault. The data were extracted from test rig shown in Fig. 6. All the fault was artificially induced

Table 1 Signal sample configuration

Data	Samples	Condition	Label
Training	105	Healthy	0.2
	105	Inner race fault	0.4
	105	Outer race fault	0.6
	105	Ball fault	0.8
Testing	45	Healthy	0.2
	45	Inner race fault	0.4
	45	Outer race fault	0.6
	45	Ball fault	0.8

Fig. 5 Bearing vibration signal generated from MFS-RDS, **a** healthy, **b** inner race fault, **c** outer race fault, and **d** ball fault



with 0.007 inches in diameter on inner raceway, outer raceway, and rolling element. The details of the experiment can be referred to CWRU data center Websites.

The same procedure as experiment bearing data was used for this CWRU bearing data where the vibration signal from each bearing condition was then divided into 150 signal samples with 1000 data point for each. Total of 600 signal samples were then

Fig. 6 CWRU experimental test rig

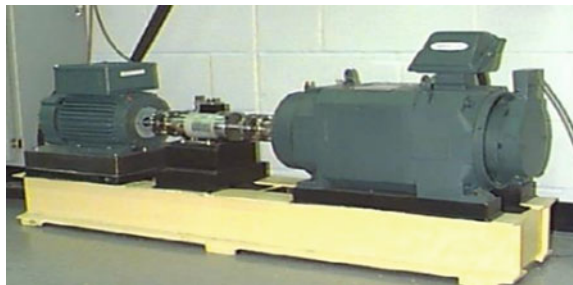
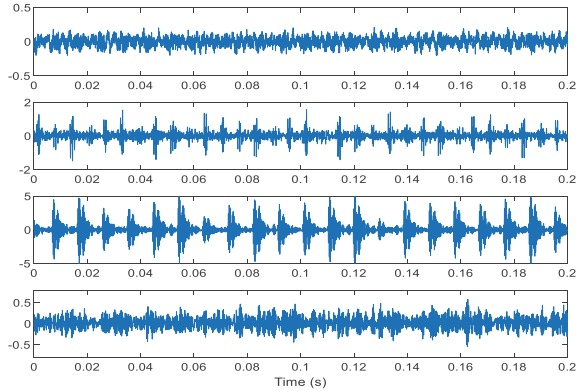


Fig. 7 Bearing vibration signal from CWRU data center, **a** healthy, **b** inner race fault, **c** outer race fault, and **d** ball fault



divided into training and testing data according to 70% training data and 30% testing data. The same signal sample configuration can be referred to Table 1 accordingly. Figure 7 shows the vibration signals for each bearing condition from CWRU data center.

4 Results and Discussion

In this section, there are two stages of result which will be presented. First result presented is about parameter optimization for ELM method based on GTO method. Second result presented the diagnosis performance of the proposed method as compared with conventional and WOA-ELM method. In this study, eight statistical features listed in Table 2 were used. The features were extracted from each vibration signal samples from both datasets. The features were then feed into the proposed method for parameter optimization and diagnosis.

4.1 ELM Parameter Optimization

GTO algorithm will be used to select the optimized parameters value for the ELM method. There are some parameters need to be defined in order to run the GTO algorithm which are number of maximum iterations, population size, number of variables, lower bound, and upper bound as stated in Table 3.

All the features from both datasets were then analyzed by GTO-ELM and also analyzed by WOA-ELM for comparison study. Figure 8a, b shows the convergence curve of GTO and WOA using the same features data from the experimental datasets. Figure 9a, b shows the convergence curve of GTO and WOA using the same features data from the CWRU datasets. From the result, it shows that the GTO method was

Table 2 Statistical features

Parameter	Equation
RMS	$\sqrt{\sum_{n=1}^N x(n)^2 / N}$
Range	$\max(x) - \min(x)$
Skewness	$(\sum_{n=1}^N (x(n) - \bar{x})^3 / N) / (\sqrt{\sum_{n=1}^N (x(n) - \bar{x})^2 / N})^3$
Kurtosis	$(\sum_{n=1}^N (x(n) - \bar{x})^4 / N) / (\sqrt{\sum_{n=1}^N (x(n) - \bar{x})^2 / N})^4$
Crest factor	$\max x / \sqrt{\sum_{n=1}^N x(n)^2 / N}$
Shape factor	$(\sqrt{\sum_{n=1}^N x(n)^2 / N}) / (\sum_{n=1}^N x(n) / N)$
Impulse factor	$\max x / (\sum_{n=1}^N x(n) / N)$
Margin factor	$\max x / ((\sum_{n=1}^N \sqrt{ x(n) }) / N)^2$

Table 3 GTO initial parameter

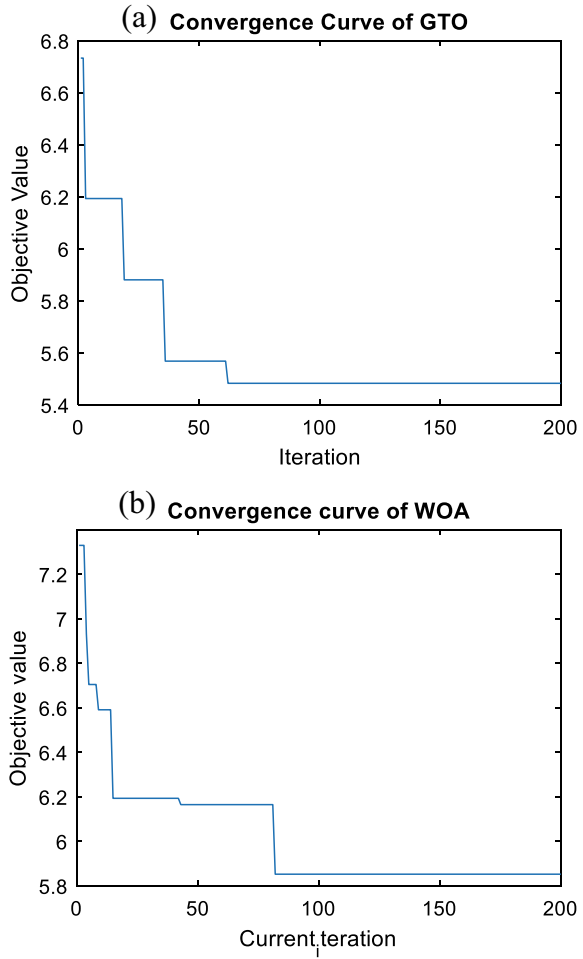
GTO parameter	Value
Population size	150
Iteration	100
Variable	3
Lower bound	[0, 0, 1]
Upper bound	[1, 1, 200]

managed to reduce the convergence speed with better objective function minimization value as compared with WOA method. For experimental datasets, WOA recorded 85 iteration to converge whereas GTO recorded only 60 iteration to converge. For CWRU datasets, WOA recorded 100 iteration to converge, whereas GTO recorded only 55 iteration to converge. Less iteration to converge means the optimization method was able to find the minimum solution faster. This result can be reflected back to the statement made by Benyamin, where the GTO method was able to improve convergence characteristic as compared with another optimizer.

4.2 Bearing Fault Diagnosis Based on GTO-ELM

By having the optimized parameters value for number of hidden neurons, input weight, and bias, the diagnosis process can be proceeded accordingly. Before testing the proposed method, a quick study on diagnosis performance when number of neurons was used wrongly in ELM algorithm. Figure 10a, b shows the training and testing performance of conventional ELM without fixing the number of neurons. As

Fig. 8 Convergence curve for ELM parameter optimization for experimental datasets, **a** GTO and **b** WOA

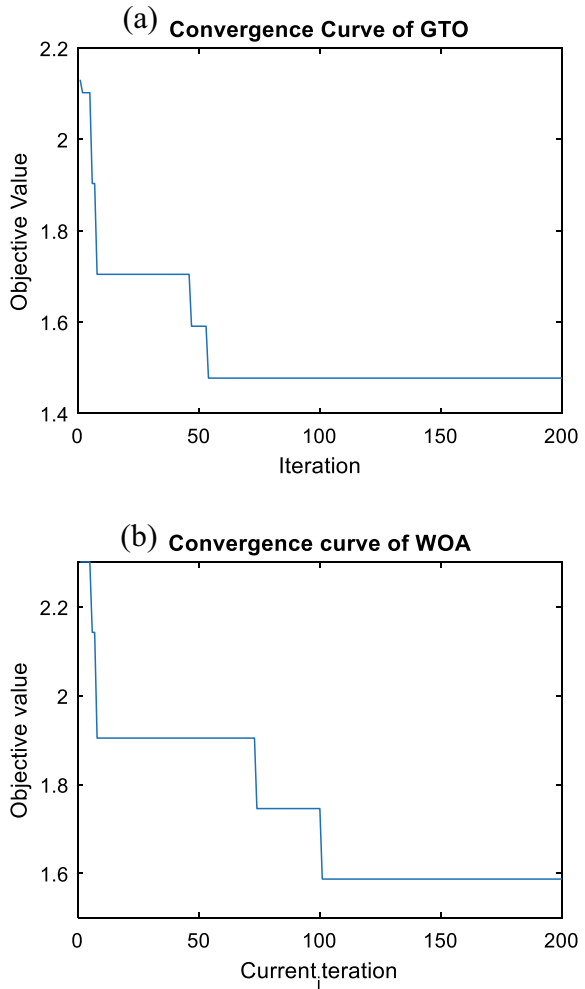


observe this plot, it clearly sees that the testing performance was degraded when the number of neurons set was higher. Shown in Fig. 8b, the diagnosis accuracy start to degrade when the number of neurons reach almost 200. This is basically the main reason for the upper limit of the GTO parameter for variables number three (number of neurons) was set to 200.

The comparison study also has been done by comparing the proposed method with conventional ELM and WOA-ELM method. Figure 11a–c presents the diagnosis performance result. For conventional ELM, the number of neurons were set to range of 1–200 due to no specific of fix value of neuron number can be used. The weight and bias are randomly distributed for conventional ELM in range value of 0–1.

From the result shown in Fig. 9a–c, WOA-ELM and GTO-ELM were able to provide significance increase of diagnosis performance as compared with conventional ELM about 10–12% better diagnosis accuracy for both datasets. The proposed

Fig. 9 Convergence curve for ELM parameter optimization for CWRU datasets, **a** GTO and **b** WOA



method provides a competitive diagnosis performance as compared with the WOA-ELM. This is expected result as both methods were able to find good and optimal solution for the ELM parameters. The result also has been summarized in Table 4.

5 Conclusions

This study has proposed bearing diagnosis method using the combination of GTO and ELM method, so called GTO-ELM method. The proposed method has been tested with experimental bearing datasets and CWRU bearing datasets where each

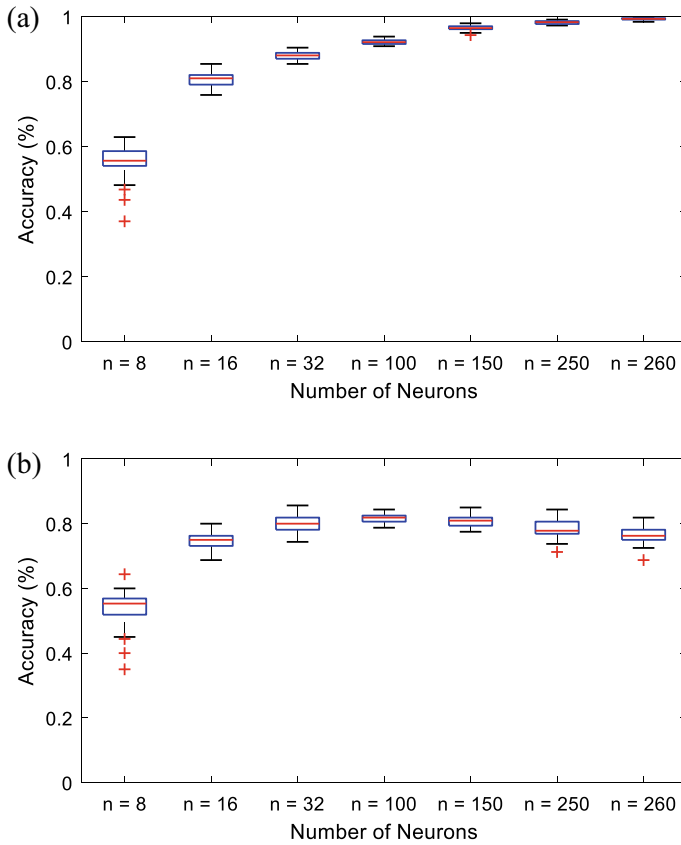


Fig. 10 Training and testing data with different value of neurons number for experimental datasets, **a** training and **b** testing

dataset consists of four different fault condition. The study can be summarized as follows:

1. The GTO method has proved its capability to select an accurate number of neurons, input weight, and bias for the ELM method.
2. The GTO method has also proved its ability to provide better and fast selection method as compared with WOA method where the GTO method surpasses WOA in term of convergence and solving minimization problem.
3. The proposed GTO-ELM provides a competitive diagnosis approach in diagnosing bearing application. This has been proved based on the result for both datasets where GTO-ELM surpasses conventional ELM with almost 10–12% better accuracy and provides competitive almost 1–2% better diagnosis accuracy as compared with WOA-ELM.

Fig. 11 Comparison diagnosis performance, **a** conventional ELM, **b** GTO-ELM, and **c** WOA-ELM

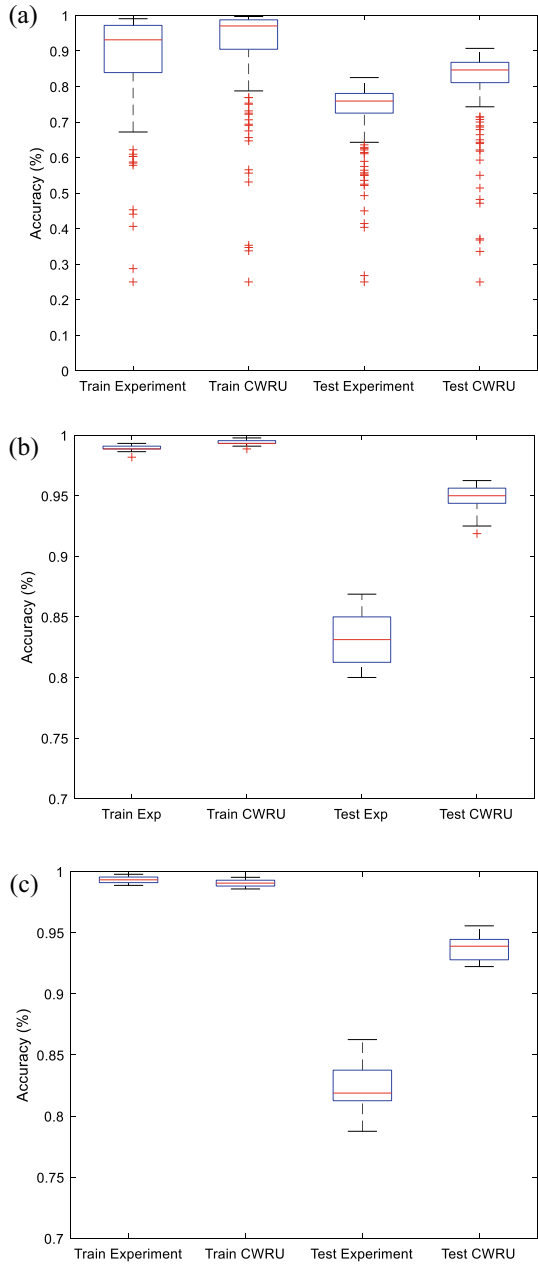


Table 4 Bearing diagnosis performance

Method	Datasets	Training (%)	Testing (%)
GTO-ELM	Experiment	98.9	83.2
	CWRU	99.4	94.8
WOA-ELM	Experiment	98.9	82.3
	CWRU	99.3	94.1
Conventional ELM	Experiment	88.3	73.1
	CWRU	91.9	81.6

For future work, we will test our proposed GTO-ELM approach on other rotating machinery application such as gear, shaft, and blades. We will also be focusing to further improve the GTO algorithm as there are a lot of random parameters were sets in this algorithm which not so very practical in some cases and also may affect the effectiveness of this method.

Acknowledgements The author would like to extend their greatest gratitude to the Institute of Noise and Vibration UTM for funding the study under the Higher Institution Center of Excellence (Hi-CoE) Grant Scheme (R.K130000.7843.4J227 and R.J130000.7824.4J234) and UTM Fundamental Research Grant Scheme (UTM-FR), Q.J130000.3851.22H06.

References

1. Saufi SR et al (2019) Challenges and opportunities of deep learning models for machinery fault detection and diagnosis: a review. *IEEE Access* 7:122644–122662
2. Li Z et al (2018) Multi-dimensional variational mode decomposition for bearing-crack detection in wind turbines with large driving-speed variations. *Renew Energy* 116:55–73
3. Zhang X et al (2018) Bearing fault diagnosis using a whale optimization algorithm-optimized orthogonal matching pursuit with a combined time–frequency atom dictionary. *Mech Syst Signal Process* 107:29–42
4. Li J et al (2017) Rolling bearing fault diagnosis based on time-delayed feedback monostable stochastic resonance and adaptive minimum entropy deconvolution. *J Sound Vib* 401:139–151
5. Dong W et al (2021) Intelligent fault diagnosis of rolling bearings based on refined composite multi-scale dispersion q-complexity and adaptive whale algorithm-extreme learning machine. *Measurement* 176:108977
6. Isham MF et al Optimized ELM based on whale optimization algorithm for gearbox diagnosis. *MATEC Web Conf* 255 (2019)
7. Bai R et al (2021) Rolling bearing fault diagnosis based on multi-channel convolution neural network and multi-scale clipping fusion data augmentation. *Measurement* 184:109885
8. Li C et al (2021) Meta-learning for few-shot bearing fault diagnosis under complex working conditions. *Neurocomputing* 439:197–211
9. Sun W, Wang Y (2021) Prediction and analysis of CO₂ emissions based on regularized extreme learning machine optimized by adaptive whale optimization algorithm. *Pol J Environ Stud* 30(3):2755–2767
10. Wang J et al (2021) Accuracy-improved bearing fault diagnosis method based on AVMD theory and AWPSO-ELM model. *Measurement* 181:109666

11. Zhang T et al (2021) A novel feature adaptive extraction method based on deep learning for bearing fault diagnosis. *Measurement* 185:110030
12. Huang GB et al (2006) Extreme learning machine: theory and applications. *Neurocomputing* 70(1–3):489–501
13. Huang G-B et al (2012) Extreme learning machine for regression and multiclass classification. *IEEE Trans Syst, Man, Cybern Part B, Cybern* 42(2):513–529
14. Chen Z et al (2019) Mechanical fault diagnosis using convolutional neural networks and extreme learning machine. *Mech Syst Signal Process* 133:106272
15. Isham MF et al (2019) Iterative variational mode decomposition and extreme learning machine for gearbox diagnosis based on vibration signals. *J Mech Eng Sci* 13(1):4477–4492
16. Li Y et al (2016) Fault diagnosis of rolling bearing based on permutation entropy and extreme learning machine. In: *Proceedings of 28th Chinese control and decision conference (CCDC 2016)*, pp 2966–2971
17. Liang M et al (2018) A novel faults diagnosis method for rolling element bearings based on ELCD and extreme learning machine. *Shock Vib* 2018
18. Mao W et al (2017) Online sequential prediction of bearings imbalanced fault diagnosis by extreme learning machine. *Mech Syst Signal Process* 83(suppl C):450–473
19. Wang D et al (2017) A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine. *Sci Total Environ* 580:719–733
20. Chen Y et al (2018) Mixed kernel based extreme learning machine for electric load forecasting. *Neurocomputing* 312:90–106
21. Benkedjough T, Rechak S (2016) Intelligent prognostics based on empirical mode decomposition and extreme learning machine. In: *2016 8th international conference on modelling, identification and control (ICMIC)*, pp 943–947
22. Thammasakorn C et al (2018) Optimizing weighted ELM based on gray wolf optimizer for imbalanced data classification. In: *2018 10th international conference on information technology and electrical engineering (ICITEE)*, pp 512–517
23. Zhou J et al (2019) Forecasting the carbon price using extreme-point symmetric mode decomposition and extreme learning machine optimized by the grey wolf optimizer algorithm
24. Xiao J et al (2018) Identification of shaft orbit based on the grey wolf optimizer and extreme learning machine. In: *2018 2nd IEEE advanced information management, communicates, electronic and automation control conference (IMCEC)*, pp 1147–1150
25. Sales AK et al (2021) Urmia lake water depth modeling using extreme learning machine-improved grey wolf optimizer hybrid algorithm. *Theor Appl Climatol* 146(1):833–849
26. Shariati M et al (2020) A novel hybrid extreme learning machine–grey wolf optimizer (ELM-GWO) model to predict compressive strength of concrete with partial replacements for cement. *Eng Comput*
27. Li H et al (2020) Research on multi-domain fault diagnosis of gearbox of wind turbine based on adaptive variational mode decomposition and extreme learning machine algorithms
28. Yao G et al (2021) A hybrid gearbox fault diagnosis method based on GWO-VMD and DE-KELM
29. Nayak DR et al (2017) Pathological brain detection using extreme learning machine trained with improved whale optimization algorithm. In: *2017 ninth international conference on advances in pattern recognition (ICAPR)*, pp 1–6
30. Sun W, Zhang C (2018) Analysis and forecasting of the carbon price using multi-resolution singular value decomposition and extreme learning machine optimized by adaptive whale optimization algorithm. *Appl Energy* 231:1354–1371
31. Abdollahzadeh B et al (2021) Artificial gorilla troops optimizer: a new nature-inspired metaheuristic algorithm for global optimization problems. *Int J Intell Syst* 36(10):5887–5958
32. Ginidi A et al (2021) Gorilla troops optimizer for electrically based single and double-diode models of solar photovoltaic systems

Classifying Ethnicity of the Pedestrian Using Skin Colour Palette



Syahmi Syahiran Ahmad Ridzuan, Zaid Omar, and Usman Ullah Sheikh

Abstract Nowadays, with the emergence of big data, there is an increasing desire to analyze and understand them. In this area, the focus is on the challenge of identifying pedestrian attributes from CCTV images, especially in terms of ethnicity. Due to a lack of necessary characteristics, ethnicity classification is nigh impossible in this instance. Facial landmarks are a requirement for the existing approaches. Therefore, it is suggested to use the individual's skin tones as features instead. Segmenting the skin area of each unique face adds multiple dominant colours to the colour palette, which are later employed as characteristics during classification. The P-DESTRE dataset, which provides pedestrian dataset and their properties, including their ethnicities, is used to demonstrate the viability of the suggested method. The accuracy percentage for distinguishing between Caucasian and Indian pedestrians using the P-DESTRE dataset and skin colour palette is 98%. The outcome demonstrates that ethnicity classification is possible when utilizing a colour palette as a feature. On this premise, it is still possible to identify a pedestrian's ethnicity from CCTV footage even without the use of face landmarks.

Keywords Pedestrian attribute · Content-based video retrieval · Ethnicity classification

S. S. A. Ridzuan (✉) · Z. Omar · U. U. Sheikh
School of Electrical Engineering, Universiti Teknologi Malaysia, Johor Bahru, Malaysia
e-mail: syahmisyahiran.ahmadridzuan@gmail.com

Z. Omar
e-mail: zaid@fke.utm.my

U. U. Sheikh
e-mail: usman@fke.utm.my

1 Introduction

Ethnicity is what makes the people around the world have different traits but also allows them to form their own distinct groups [1]. Being of the same ethnicity also entails sharing many comparable physical characteristics, such as skin tone, blood type, build, head shape, and hair texture. However, it is difficult to accurately identify someone's ethnicity due to mixed marriages, and the children tend to recognize their ethnicity by their parents' blood or heritage, rather than their appearance [2]. Since police identify someone based on the person's appearance [3], it is normal to stick with the stereotypes of a race. For example, a black person can be easily grouped as African and a white person as Caucasian.

As the big data analysis becoming the norm nowadays, especially with massive images and videos available in public, it is essential to have a content-based image or video retrieval or pedestrian attribute recognition when one wants to query an object-of-interest (OOI) or person-of-interest (POI) from the images or videos. One of the examples of information retrieval that uses computer vision techniques to address issues with searching and managing massive picture or video collections is called content-based image retrieval (CBIR), also known as content-based video retrieval (CBVR) for the video counterpart [4]. Meanwhile, pedestrian attributes recognition (PAR) is a process of extracting the visible features of the pedestrian from CCTV footage which is useful in visual surveillance, with applications in re-identifying the individual, recognizing the face, and identifying human [5]. As CBIR or CBVR creates a seamless and simpler query system to retrieve OOI or POI, PAR prepares the attributes needed for the query. The ethnicity classification enables to determine the pedestrian ethnicity which is one of the important traits during person identification. The availability of ethnicity and other attributes will make the person's query much more accurate and relieves the burden of the law enforcer of scouring and analyzing the large database manually.

Normally, the skin colour first comes into mind when distinguishing a person's ethnicity. Afterwards, it is the anthropological aspects or the physical traits of the person [6]. Thus, some research works suggested using facial landmarks instead of colour [7]. It allows the police officer to look beyond the skin colour and possibly avoid misidentification like during the chasing of the infamous serial killer, Zodiac Killer. The two patrol officers near the crime scene mistakenly let off the real serial killer due to the mix-up by the police which led them to look for African descent suspect instead [8]. They drove past the real perpetrator who actually fits the physical appearance and managed to successfully escape from the scene.

However, in the context of closed circuit television (CCTV) footage, there are several challenges that make it harder to use the current techniques to classify ethnicity. The first hurdle is the low-resolution video which limits the details of facial features [9]. The second is the person is moving which requires tracking to ensure that the moment the person's face is facing the front is captured [10]. Now counting the other difficulties due to low illumination, occlusion, and the lack of the required ground truth.

2 Literature Review

The methods that are utilizing the facial landmarks need to have a clear sight of the face. Guo and Mu [11] proposed classifying ethnicity using biologically inspired features (BIF). The Gabor filter is used to generate layers of simple (S) and complex (C) cell units for BIF. The extracted BIF is then combined with various learning techniques to reduce dimensionality before being classified using SVM. Gutta et al. [12] introduced hybrid classifiers made up of RBF networks and inductive DT. This method provides the user with an average accuracy rate of 94% for ethnic classification. Lin et al. [13] extracted facial features using Gabor filter banks and AdaBoost learning and then classified those features using an SVM classifier. The authors divided ethnic recognition into 3 categories: yellow (Mongoloid), black (African), and white (Caucasian). The authors also included pre-processing to help improve accuracy. He achieved an accuracy rate of 94.58% when comparing yellow to white, 95.59% when comparing yellow to black, and 96.21% when comparing white to black. Mohammad and Al-Ani [14] started by combining features extracted using the HOG and LBP methods. Then, four classifiers are compared before deciding which classifiers are best in classification: SVM, multi-layer perceptron (MLP), linear discriminant analysis (LDA), and quadratic discriminant analysis (QDA). The authors demonstrated that SVM with the polynomial kernel outperformed other classifiers, achieving an overall accuracy rate of 98.5%. These techniques, according to [7], require a dataset with visible frontal faces, such as mugshots, such as the FERET and MORPH II datasets. The lack or absence of facial landmarks in CCTV footage is due to low-resolution images, as well as other variations such as illumination, obstruction, and camera angles.

The other solution focussed on gait recognition. Gait recognition, according to Zhang et al. [15], has been widely used for person and gender identifications, but it is also capable of ethnicity identification. As a result, they attempt to determine ethnicity based on the fusion of multi-view gait. The highest classification rate obtained from their fusion schemes is 84.4% when using multilinear principal component analysis (MPCA). Samangoeei and Nixon [16] also used gait biometrics to create human content-based retrieval by using the dataset of 2000 videos of people walking in laboratory conditions. Amongst the semantic traits discernible by humans at a distance in their experiment is ethnicity where they managed to obtain $mAP = 9\%$ above random. This solution, however, needs a long recording of the individual walking from multiple view angles. It might not work well for a single camera positioned at the same angle and in a crowded environment where the other individuals might obstruct the individual recording.

There is also another workaround by classifying using clothing. Rajput and Aneja [17] focus on Indian Ethnic Clothing by defining 15 classes of attire and reached 88.43% classification accuracy by using Resnet-50. Although the authors are classifying within Indian style clothing, it can be used to differentiate Indians wearing ethnic attire apart from the other ethnicities. Washburn [18] studies the design categories on Bakuba Raffa Cloth where different ethnicities in Bakuba has different

style and type of clothing representing the ethnic. Ginige and Yasas Mahima [19] implement classification models that can predict the cloth type and the colour of the clothes based on consumers' ethnicities in Sri Lanka, namely the Sinhalese and the Tamils. It helps the online sellers to get a clear understanding of the buying behaviours and the expectations of the consumers based on their ethnicities. These solutions are working in a specific condition where there are distinctive attires worn by a specific ethnicity. However, most people in the community nowadays are wearing western clothing and the same dress code which makes them indistinguishable from others.

As a solution, it is imperative to look back at the skin colour as an option due to the limitations of the current technique. The use of colour to classify ethnicity is not a novelty, but there will be a need for adaptation and improvement of the previous works.

Some previous works utilize pixel intensity and colour histogram [20–23] to differentiate ethnicity. The reason why many avoid using the colour itself to classify ethnicity is because of several factors such as illumination variation. Therefore, instead of using a single colour for ethnicity classification using skin colour, it is much more accurate to create a colour palette consisting of multiple prominent colours representing that individual. Using a single colour as a parameter is not enough, therefore, several colours can be set as several parameters to define the ethnicity, taking into account the changes caused by different illumination for non-controlled environments.

For reference, there are already some established classification systems for skin types that can be used to differentiate between different ethnicities such as Fitzpatrick's skin type [24], Von Luschan's chromatic scale [25, 26], Lancer Ethnicity Scale [27], and Goldman World Classification of Skin Types [28]. Fitzpatrick skin type is initially developed to estimate the response of different skin types towards UV light which is to describe sun tanning behaviour. Afterwards, a 6-point subjective classification system is developed to assess the propensity of the skin to burn during phototherapy sessions. It is found that Type I which is the palest skin colour type always burns, whilst Type VI which is the darkest skin colour type never burns. In contrast, Von Luschan's chromatic scale is established to establish racial classification using skin colour. Instead of a 6-point skin type, this system uses 36 opaque glass tiles which are compared to a person's skin colour. Lancer Ethnicity Scale accounts for 5 different skin types based on geography and heredity. As for Goldman's World Classification of Skin Types, which is based on Fitzpatrick skin type, suggested adding other stimuli of melanocytic function. Therefore, a skin colour classification system is based on response to skin tanning, and post-inflammatory hyperpigmentation (PIH) is based on ethnicity.

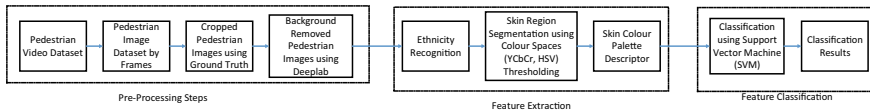


Fig. 1 Block diagram contains the overall process from the acquired pedestrian video dataset to the classification result

3 Methodology

3.1 Dataset

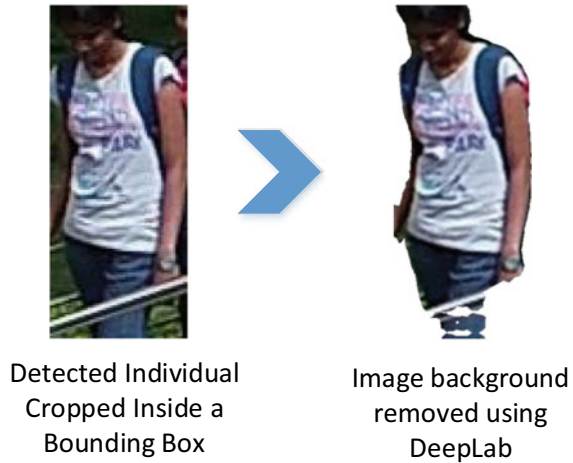
P-DESTRE [29] is the dataset used in this paper, and it is the result of a collaborative effort between researchers at the University of Beira Interior (Portugal) and the JSS Science and Technology University (India). To enable the research on pedestrian identification from aerial data, a group of “DJI Phantom 4” drones piloted by humans flew over various scenes on both universities’ campuses, collecting data that simulated everyday conditions in outdoor urban crowded environments. All of the subjects in the dataset were offered explicitly as volunteers, and they were asked to simply ignore the UAVs (Fig. 1), which were flying at heights ranging from 5.5 to 6.7 m, with camera pitch angles ranging from 45 to 90°. 269 volunteers were mostly between the ages of 18 and 24 (>90%), roughly divided into two halves for gender (175 males (65%) and 94 females (35%), and ethnicity (“White” and “Indian”). Around 28% of the volunteers were wearing glasses, and 10% were wearing sunglasses. Data were captured at 30 frames per second with a spatial resolution of 4K (3,840 2,160) and saved in “mp4” format with H.264 compression.

3.2 Pre-processing

Before getting into the feature extraction step, the dataset needs to undergo some pre-processing steps. First, the video is split into multiple frames. The provided ground truth already contained the coordinates of the bounding box containing the person-of-interest. It contains x (top left column of the box), y (top left row of the box), h (height), and w (width). By using the ground truth of the dataset, each detected individual in the frame is cropped. To separate the person from the surrounding, a background subtraction technique is needed, which in this case, DeepLab is used.

DeepLab [30], an effective background removal technique, is used to separate the person from the background to refine the feature extraction later. DeepLab is a state-of-the-art deep learning model for semantic image segmentation, where the goal is to assign semantic labels such as a person, a dog, or a cat to every pixel in the input image. Atrous convolution is used to explicitly control the resolution at which feature responses are computed within deep convolutional neural networks.

Fig. 2 The image of a pedestrian in the bounding box (left) undergoes DeepLab background removal technique which created the pedestrian without background image (right) as the output



Atrous convolution allows enlarging the field-of-view of filters to incorporate a larger context. It thus offers an efficient mechanism to control the field-of-view and finds the best trade-off between accurate localization (small field-of-view) and context assimilation (large field-of-view) (Fig. 2).

3.3 Feature Extraction

After removing the background of each individual, it is now possible to create the skin colour palette. To achieve this, two methods are required, namely multiple thresholding and Annesley's colour detector. Multiple thresholding is used to segment only the skin region to avoid the clothing colour to be included in the colour palette. Annesley's colour detector is then implemented to create the skin colour palette by determining the most prominent colours.

HSV and YCbCr Multiple Thresholding. In order to segment the skin region, multiple threshold [31], and zero-sum game theory model [32], the colour space ranges are modified, or different colour spaces are combined to complement the type of dataset used. Two colour spaces are used: HSV and YCbCr. The threshold for each colour space which was based on [31] is defined in Eq. 1.

$$HSV : V \geq 40; 0.2 < S < 0.6; 0^\circ < H < 25^\circ \text{ or } 335^\circ < H < 360^\circ \quad (1)$$

$$YCbCr : 0 \leq Y \leq 255; 77 < C_B < 127; 133 < C_R < 173$$

The masks obtained using the two colour spaces are combined using AND logical operation into a single mask to crop the skin region as shown in Fig. 3.

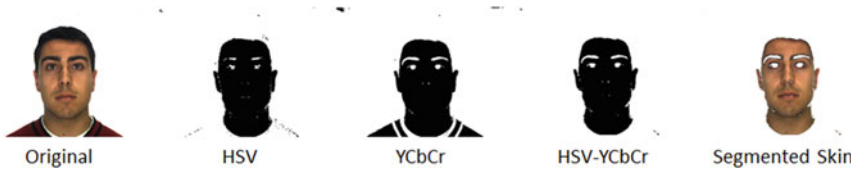


Fig. 3 The HSV and YCbCr Multiple Thresholding steps where the HSV and YCbCr masks created using threshold values are combined using AND operation which is then used to determine the skin region

This process is repeated for each individual, creating a skin region only collection. The segmented skin is then available to be applied with Annesley's colour detector to create the colour palette.

Annesley's Colour Detector. The idea of using skin tone as an ethnicity classifier is from Fitzpatrick [24] and von Luschan [25, 26] skin colour chart that was initially used to determine skin burnt. Therefore, Paul Annesley colour detector method [33] is the proposed solution to create skin colour palette. Originally, the method can create a colour palette of any RGB image, but it is then fine-tuned to produce a skin colour palette which is suitable for ethnicity classification. Using the skin pixel extracted using HSV and YCbCr Multiple Thresholding up to five major colours excluding the background are determined which later can be used as features during the classification.

The algorithm is as follows:

1. Load the image with removed background and segmented skin region
2. Collect all the skin tones available and assign their prominence values
3. Determine the background colour
4. Create a skin colour palette containing up to 5 most dominant colours in hexadecimal (Fig. 4)

Before proceeding into the classification step, the hexadecimal value of the colours is converted into RGB values as the hexadecimal value does not correspond well to colour similarity.

3.4 Feature Classification-Support Vector Machine

Support vector machine (SVM) is supervised learning for classification and regression analysis which are developed by Cortes and Vapnik [34]. It is selected due to its reliability as it is used prevalently in image classification [35] and is able to produce a high classification rate. Guo and Mu [11], Lin et al. [13], and Mohammad and Al-Ani [14] use SVM to classify ethnicity and managed to obtain a higher than 90% accuracy rate. It also does not require a massive sample size like deep learning methods, and the classification steps are much simpler; thus, it needs lesser processing power.

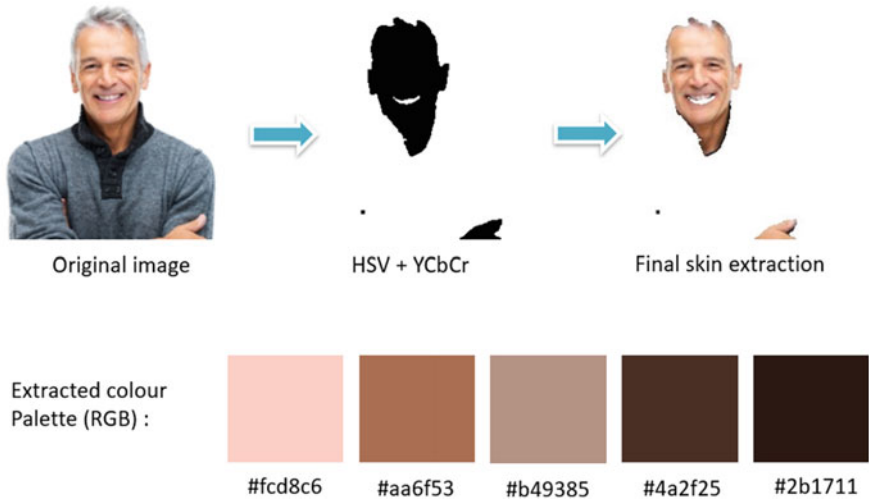


Fig. 4 The original image containing a person has its skin region extracted using the HSV and YCbCr multiple thresholding technique and then, Annesley’s colour detector is applied to create the skin colour palette

In order to classify the ethnicity of the pedestrian, the skin colour palette of the individuals is used as a feature. By default, the RGB values of each prominent colour are stored and converted into features with the possibility of changing them into other colour spaces. The samples are divided into 80% for the training set and 20% for the testing set. The ethnicity consists of two classes: Caucasian (White) from the University of Beira Interior (Portugal) and Indian (Brown) from the JSS Science and Technology University (India). It is a linear SVM that uses regularization parameter = 1 and linear kernel. The other parameters are kept at default values.

4 Result

Using the P-Destre dataset, 2000 sample images (1000 images each for Caucasian and Indian classes) are used for classification. Although the technique allows up to 5 colours as features, most of the Caucasian images managed to get up to 4 colours in RGB. Each colours are represented in single-channel (R1,G1,B1,R2,G2,B2,..., R5,G5,B5) as features for the classification step via SVM. The result of the classification is as follows:

From Table 1, it is evident that as more colours are used as features, the fewer available sample images for classification purposes. The classification rate also increases as more colours are added as features. This result shows that there is an interclass separation between the classes. Next is to determine how long it takes to finish the classification by using the same number of images.

Table 1 Performance of each classification model with different number of colours as features in terms of classification rate

No. of colours	No. of images	Classification rate (%)
1 colour	2000	92
2 colours	1960	97
3 colours	1632	98
4 colours	296	100

Table 2 Performance of each classification model with different number of colours as features in terms of processing time

No. of colours	Processing time (seconds)
1 colour	0.10662961006164551
2 colours	0.08076214790344238
3 colours	0.03198099136352539
4 colours	0.028976917266845703

From Table 2, it seems that an increase in colour used helps a bit in reducing the processing time even though more features needed to be processed. The time taken to finish the processing is too small to be a deciding factor in choosing the number of colours.

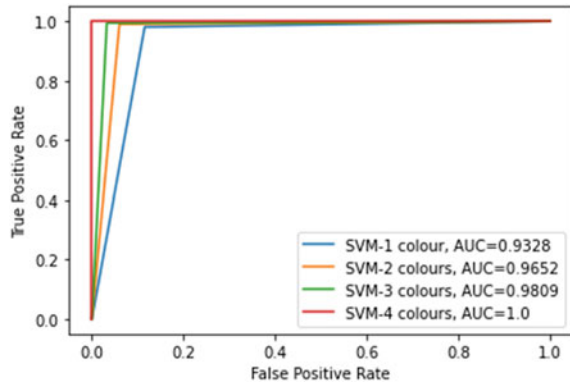
Due to the limitations of the dataset, it is very difficult to implement the classification using 5 colours for comparison. The video footage of the students from JSS Science and Technology University (India) which forms Indian classes has better illumination, allowing more skin colour extraction as shown in Fig. 5. Meanwhile, the video footage of the University of Beira Interior (Portugal) which forms Caucasian class has lower brightness, and the shadow affects the available skin colour for extraction.

Based on Fig. 5, it is shown that each class has their own set of skin colour tones which allows for creating an interclass separation, thus, a higher classification rate. According to Fitzpatrick’s skin pigmentation scale, the skin colour pertaining to the respective group allows the classification of the ethnicity. Utilizing this idea of where



Fig. 5 The skin colour tones of Indian pedestrian (left) and Caucasian pedestrian (right) from the dataset

Fig. 6 ROC-AUC curve of all classification models



a group of people of the same ethnicity share very close sets of colour tones, it allows the classification based on ethnicity. The same reason why the other authors who use gait or facial landmarks to classify ethnicity as the chosen features are able to define their unique traits.

ROC-AOC Curve. To compare the performance of the different number of colour palette classifications, the area under the curve (AUC) and receiver operating characteristics (ROC) curve are used. ROC is a probability curve, and AUC represents the degree or measure of separability.

From Fig. 6, it is clear that the red curve which represents the SVM classification with 4 colours hugs the top left corner of the plot the most. It means that the model is better at classifying the data into categories. Then, the red curve was followed by the green curve (SVM with 3 colours), the orange curve (SVM with 2 colours) and the blue curve (SVM with 1 colour) according to the classification rates. As fewer colours are used during the classification, the capability of classifying the data starts to diminish. The AUC values of each classification also support this argument. The red curve has the highest value which means it has the best capability of separating between classes, creating the best interclass separation.

Even though, the classification model using SVM with 4 colours tops in the performance, the limitation of the dataset, particularly the Caucasian, causes a fewer number of sample images available for classification. Therefore, the SVM model with 3 colours would be a better choice as it has a bigger sample size and ensure the classification result is more convincing.

5 Conclusion

The objective of this paper is to be able to accurately classify the ethnicity of the pedestrian from CCTV footage, and the implementation of the proposed approach shows that it is doable even without using facial features. By using 3 RGB colours as

features, the accuracy rate of classifying Caucasian and Indian pedestrians reaches 98% and can be increased if more prominent colours are used.

In the future works, it is possible to explore other colour spaces as features and solve the illumination problem in certain classes which will allow more sample images for the classification process. Currently, there are not many pedestrian datasets with ethnicity in the ground truth like P-DESTRE, and it is limited to two ethnicities. It is encouraged to implement this approach on other similar pedestrian datasets with more ethnicities and also on the mugshot dataset to see its limitation and to evaluate if it can replicate the same performance in other types of datasets.

Acknowledgements The authors wish to thank the Ministry of Higher Education (MOHE) Malaysia and Universiti Teknologi Malaysia for supporting this work through the Fundamental Research Grant Scheme (FRGS), vote number 5F437.

References

1. Williams RM (2001) Ethnic conflicts. In: International encyclopedia of the social & behavioral sciences. Pergamon, pp 4806–4810
2. Lichter DT, Qian Z (2018) Boundary blurring? Racial identification among the children of interracial couples. *Ann Am Acad Polit Soc Sci* 677(1):81–94
3. Chicago Police Department (2013) How to describe a suspect. <https://portal.chicagopolice.org/portal/page/portal/ClearPath/Get%20Involved/Hotlines%20and%20CPD%20Contacts/How%20to%20Describe%20a%20Suspect>
4. Mai NTL, Ahmad SSB, Omar ZB (2018) Content-based image retrieval system for an image gallery search application. *Int J Electr Comput Eng* 8(3):1903
5. Wang X, Zheng S, Yang R, Zheng A, Chen Z, Tang J, Luo B (2022) Pedestrian attribute recognition: a survey. *Pattern Recogn* 121:108220
6. Jenkins R (2001) Ethnicity: anthropological aspects. In: International encyclopedia of the social & behavioral sciences, pp 4824–4828
7. Ridzuan SSA, Omar Z, Sheikh UU (2019) A review of content-based video retrieval techniques for person identification. *ELEKTRIKA-J Electr Eng* 18(3):49–56
8. Rodelli M (2005) 4th interview with Don Fouke—thoughts on the Zodiac Killer. [https://web.archive.org/web/20060501193603, http://www.mikerodelli.com/4interview.html](https://web.archive.org/web/20060501193603/http://www.mikerodelli.com/4interview.html)
9. Semertzidis T, Axenopoulos A, Karadimos P, Daras P (2016) Soft biometrics in low resolution and low quality CCTV videos. 24–25
10. Bouma H, Borsboom S, den Hollander RJM, Landsmeer SH, Worring M (2012) Re-identification of persons in multi-camera surveillance under varying viewpoints and illumination. In: Sensors, and command, control, communications, and intelligence (C3I) technologies for homeland security and homeland defense XI, vol 8359. International Society for Optics and Photonics, p 83590Q
11. Guo G, Mu G (2010) A study of large-scale ethnicity estimation with gender and age variations. In: 2010 IEEE computer society conference on computer vision and pattern recognition-workshops. IEEE, pp 79–86
12. Gutta S, Huang JRJ, Jonathon P, Wechsler H (2000) Mixture of experts for classification of gender, ethnic origin, and pose of human faces. *IEEE Trans Neural Netw* 11(4):948–960
13. Lin H, Lu H, Zhang L (2006) A new automatic recognition system of gender, age and ethnicity. In: 2006 6th world congress on intelligent control and automation, vol 2. IEEE, pp 9988–9991
14. Mohammad AS, Al-Ani JA (2017) Towards ethnicity detection using learning based classifiers. In: 2017 9th computer science and electronic engineering (CEECE). IEEE, pp 219–224

15. Zhang D, Wang Y, Bhanu B (2010) Ethnicity classification based on gait using multi-view fusion. In: 2010 IEEE computer society conference on computer vision and pattern recognition-workshops. IEEE, pp 108–115
16. Samangoeei S, Nixon MS (2010) Performing content-based retrieval of humans using gait biometrics. *Multimed Tools Appl* 49(1):195–212
17. Rajput PS, Aneja S (2021) IndoFashion: apparel classification for Indian ethnic clothes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3935–3939
18. Washburn DK (1990) Style, classification and ethnicity: design categories on Bakuba raffia cloth. *Trans Am Philos Soc* 80(3)
19. Ginige TNDS, Yasas Mahima KT (2021) Ethnicity based consumer buying behavior analysis and prediction on online clothing platforms in Sri Lanka. In: 2021 the 5th international conference on information system and data mining, pp 121–127
20. Demirkus M, Garg K, Guler S (2010) Automated person categorization for video surveillance using soft biometrics. In: Biometric technology for human identification VII, vol 7667. International Society for Optics and Photonics, p 76670P
21. Farinella G, Dugelay J-L (2012) Demographic classification: do gender and ethnicity affect each other?. In: 2012 international conference on informatics, electronics & vision (ICIEV). IEEE, pp 383–390
22. Abirami B, Subashini TS (2020) Automatic race estimation from facial images using shape and color features. In: Data engineering and communication technology. Springer, Singapore, pp 173–181
23. Putriany DM, Rachmawati E, Sthevanie F (2021) Indonesian ethnicity recognition based on face image using gray level co-occurrence matrix and color histogram. In: IOP conference series: materials science and engineering, vol 1077, no 1. IOP Publishing, p 012040
24. Fitzpatrick TB (1975) Soleil et peau [Sun and skin]. *J Méd Esthét (in French)* 2:33–34
25. Thomas N (1905) Hautfarbentafel by von Luschan. *Man* 5:160
26. von Luschan E, von Luschan F (1914) Anthropologische Messungen an 95 Engländern (SS “Durham Castle”. Brit. Association 1905). *Z Ethnol* 46(H. 1):58–80
27. Lancer HA (1998) Lancer ethnicity scale (LES). *Lasers Surg Med: Off J Am Soc Laser Med Surg* 22(1):9–9
28. Goldman MP (2008) Universal classification of skin type. In: Simplified facial rejuvenation. Springer, Berlin, Heidelberg, pp 47–50
29. Kumar SVA, Yaghoubi E, Das A, Harish BS, Proença H (2020) The P-DESTRE: a fully annotated dataset for pedestrian detection, tracking, and short/long-term re-identification from aerial devices. *IEEE Trans Inf Forensics Secur* 16:1696–1708
30. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans Pattern Anal Mach Intell* 40(4):834–848
31. Gasparini F, Schettini R (2006) Skin segmentation using multiple thresholding. In: Internet imaging VII, vol 6061. International Society for Optics and Photonics, p 60610F
32. Dahmani D, Cheref M, Larabi S (2020) Zero-sum game theory model for segmenting skin regions. *Image Vis Comput* 99:103925
33. Hotson D, Lars Y (2019) Colorific (v 0.3.0) [source code]. <https://github.com/99designs/colorific>
34. Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20(3):273–297
35. Chandra MA, Bedi SS (2021) Survey on SVM and their application in image classification. *Int J Inf Technol* 13(5):1–11

Cluster Analysis Based on Image Feature Extraction for Automated OMA



Muhammad Danial Bin Abu Hasan , Syahril Ramadhan Saufi, M. Firdaus Isham, Shaharil Mad Saad, W. Aliff A. Saad, Zair Asrar Bin Ahmad, Mohd Salman Leong, Lim Meng Hee, and M. Haffizzi Md. Idris

Abstract This study introduces a new method for automated operational modal analysis (OMA) that uses image feature extraction on stabilisation diagrams to cluster data in parametric models. The implementation of automated OMA, a modal analysis that does not require as much human engagement as traditional methods, is a difficult challenge. Without requiring user input, the stabilisation diagram and clustering tools separate real poles from spurious (noise) poles. However, the maximum within-cluster distance between representations of the same physical mode from different system orders is required by existing clustering algorithms, and additional adaptive approaches must be used to optimise the selection of cluster validation criteria, as a consequence of a significant computational work. The proposed image clustering procedure is based on an input stabilisation diagram image that was constructed and displayed independently at a pre-defined interval frequency, and standardised image features in MATLAB were utilised to extract image features from each generated stabilisation diagram image. The image feature extraction was then used to create an image clustering diagram with a predetermined fixed threshold for classifying physical modes. Even for closely spaced modes, image clustering has been shown to give reliable output results that can recognise actual modes in stabilisation diagrams using image feature extraction, without the need for any calibration, user-defined parameter at start-up, or additional adaptive approach for cluster validation criteria.

Keywords Automated OMA · Clustering · Operational modal analysis · Stabilisation diagram

M. D. B. Abu Hasan (✉) · S. R. Saufi · M. F. Isham · S. Mad Saad · W. A. A. Saad · Z. A. B. Ahmad

Faculty of Mechanical Engineering, Universiti Teknologi Malaysia (UTM), 81310 Skudai, Johor Bahru, Malaysia
e-mail: muhammaddanial.ah@utm.my

M. D. B. Abu Hasan · M. S. Leong · L. M. Hee · M. H. Md. Idris
Institute of Noise and Vibration, Universiti Teknologi Malaysia (UTM), 54100 Kuala Lumpur, Malaysia

1 Introduction

Automated OMA approaches are becoming more and more prevalent as a result of recent advancements in modal damage detection and vibration-based monitoring. This is a significant step towards completely removing all physical interaction, since typical OMA calls for a high level of expert user involvement. It is generally used for repeating tests or using many data sources for the same OMA test. Because the input data must be automatically analysed, thus modal parameter identification changes may be quickly recognised [1], this automated approach is essential for SHM applications. Successful OMA automation depends on a broad range of modal signals to provide a more precise indication of physical modes.

The estimation of modal parameters using a nonparametric model required direct estimate from frequency response or PSD peaks chosen from the complex mode indicator function (CMIF) [2] or the averaged normalised power spectral density [3]. Recently, a method for automated peak selection was developed that mainly dependent on MAC index selection and peak picking method [4–8]. Due to the strength of the signal “leaking” out to surrounding frequencies, the estimation of modal parameters in the frequency domain is typically exaggerated, particularly modal damping ratio. The modal peaks of the spectral density functions are widened by the spectral leakage phenomenon [9]. The majority of research efforts are currently concentrated on automated parametric methods since simulation analysis has shown that the identifying dynamic parameters obtained from the parametric method’s state space model are significantly more accurate than nonparametric estimates [10, 11].

The model order is usually oversized in traditional parametric modal estimation to estimate all actual modes of interest in the frequency band. Model oversizing is necessary since models typically have biases and do not include noise. To differentiate between actual and spurious modes, a qualified analyst must put up significant effort. To discern between real and spurious modes, a strong tool is required, such as a stabilisation diagram. Stabilisation diagrams must be constructed for each modal analysis to clarify and determine whether a mode is physical or not within a predetermined range of model order [1, 3].

Additionally, using stabilisation diagram tools is not appropriate for real-time applications since it requires a lot of human interaction, is costly, takes a long time, and requires extensive human engagement. However, in recent years, computer methods for analysing stabilisation diagrams have been developed in a manner resembling human decision-making. Since the user must set consistency thresholds for each modal parameter, choosing a physical mode may be challenging [12]. Early attempts to automate modal estimation relied on selection criteria and clustering algorithms to distinguish between actual poles and others because so much of the human participation is for monitoring.

To be clear, in order to accelerate the process, stabilisation diagram analysis is commonly divided into two phases: (a) eliminating noise modes and (b) clustering physical modes to provide precise parameter estimates for each mode [13]. Many distinct clustering techniques have been introduced in recent years.

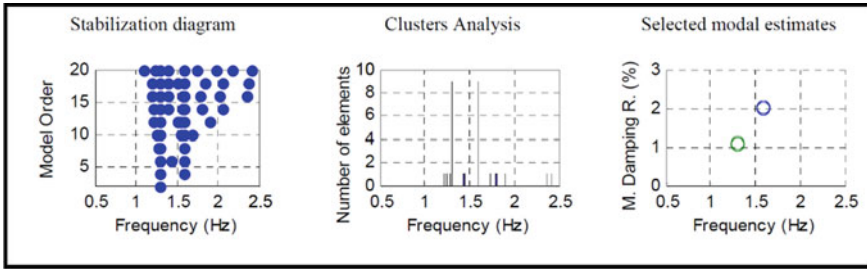


Fig. 1 Steps of proposed automated OMA procedure

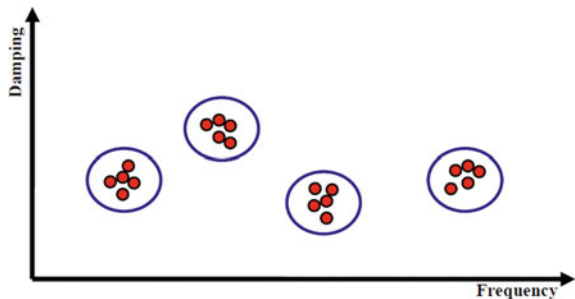
Therefore, the automated OMA entails the following procedures, which are succinctly shown in Fig. 1:

- Take measurements of the structural responses and extract the modal parameters using an oversizing model order of n modes.
- Determine the poles of increasing model order and then construct a stability diagram indicating whether the mode is real or spurious.
- Differentiating between actual and spurious modes amongst n modes using clustering method.

A method for categorising or grouping objects according to their qualities is cluster analysis. The items should thus have a high degree of internal (inside the cluster) homogeneity and a high degree of external (between the cluster) heterogeneity [14]. Algorithms for numerical-order parametric recognition attempt to cluster estimated modes with similar physical modes. To perform cluster analysis, the simplest method is to reduce the complexity of the figure by presenting only the physical modes. In reality, more, more or less randomly scattered spots would be seen.

Cluster analysis is typically employed to associate objects that are similar to one another (shown by circles in Fig. 2). A scenario with just two variables was visually depicted by the theories. Examples of variables or modal quantities include modal parameters [1].

Fig. 2 Framework for explaining the use of clustering algorithms



There are three categories of common cluster tools: histogram analysis, partitioning techniques, and hierarchical. The following flaws have an impact on the current cluster tools for automated OMA techniques [15]:

- There are a number of predetermined set parameters that must be used in order to estimate real structural modes.
- Each analysis of the data set requires a time-consuming calibration procedure at start-up.
- The values for thresholds and parameters vary naturally because of wear and tear or other environmental influences on buildings' modal characteristics.
- Cluster validation requirements for existing clustering algorithms require an additional adaptive approach.

In order to avoid tuning analytic parameters at launch, an alternative approach is required. Other researchers then realised the importance of discarding pre-defined parameters [16]. The goal of this study is to establish a novel cluster analysis technique for automated OMA that is based on image feature extraction. The key concepts of these methods are discussed, along with the advantages and disadvantages of employing them.

2 Cluster Analysis Based on Image Feature Extraction

2.1 Stabilisation Diagram

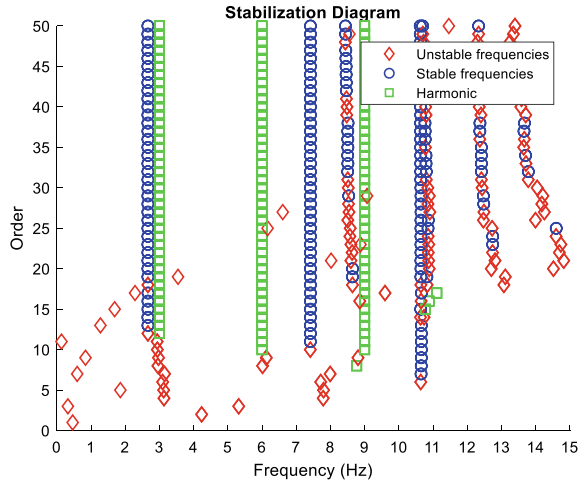
The stabilisation diagram can be built once the modal parameters have been computed parametrically to determine the finest state space dimension. By predicting poles with higher model orders, this approach is frequently used to differentiate between stable, unstable, and noisy modes. The modal model sometimes has 5–10 times as many modes as the experimental data. Model oversizing is necessary since models are typically biased and do not have noise modelling. Though noise modes are distributed, stable physical modes may be identified by their vertical alignment. This is as a result of the poles' similarity to the lower-order one-order model [17].

Equations (1) and (2) [18] are used to compare the natural frequencies and damping ratio of poles from two orders:

$$\frac{|f(n-1) - f(n)|}{f(n-1)} < x \quad (1)$$

$$\frac{|\zeta(n-1) - \zeta(n)|}{\zeta(n-1)} < y \quad (2)$$

Fig. 3 Example of stabilisation diagram plot with shape recognition



where x and y are the modal damping ratio and natural frequency thresholds for models of consecutive orders, respectively. For variance between models of consecutive orders, modes must meet the following (x and y) thresholds in order to be considered stable: The modal damping ratio is less than 5%, and natural frequency fluctuation is less than 1% [19]. A stabilisation threshold for harmonic components is defined as variation lower than 0.1% across models of consecutive orders, as evaluated by damping ratios, and is used as a prerequisite criteria to distinguish between harmonics and structural poles. Whilst harmonic components have exceptionally low damping ratios because of their appearance as sharp peaks, actual poles often have damping ratios that range between 0.1 and 2% [20].

This information can be used to eliminate negative modes. These thresholds enable a significant separation between harmonic components and stable modes (which represent vibration modes) (which indicate harmonic components). Examples of stabilisation plots are shown in Fig. 3.

2.2 Image Feature Extraction for Clustering

The following subsections discuss the procedure for using image clustering in stabilisation diagrams with equivalent physical poles.

Input image

The input image of the stabilisation diagram that has been reduced to a preset interval frequency is necessary for the image clustering process. The stabilisation diagram in this work is created and presented separately for each 0.01 Hz frequency interval using Eqs. (3) and (4):

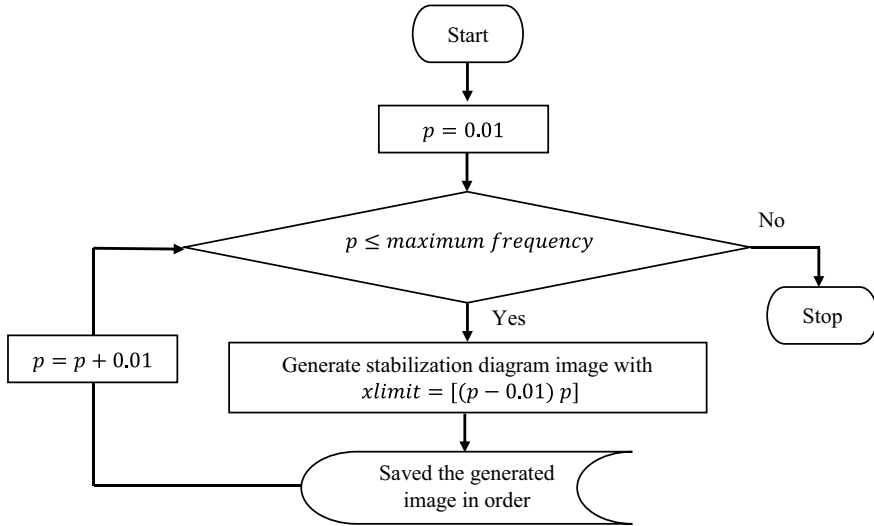


Fig. 4 Process flowchart for the input images produced by a stabilisation diagram with a 0.01 Hz interval frequency

$$(\text{maximum frequency})/0.01 = \text{total number of images} \quad (3)$$

$$xlim = [(r - 0.01)r] \quad (4)$$

where r is the natural frequency. Therefore, each image represents a frequency range of 0.01 Hz. Even for closely spaced modes, this selected frequency interval value provides sensitivity and accuracy. According to the literature [1], one decimal place of frequency is insufficient to discriminate between two closely spaced modes. The steps are depicted in Figs. 4 and 5. The steps are represented in Figs. 4 and 5.

Extraction of image features

The image features in each image of the resulting stabilisation diagrams are then extracted using MATLAB's standard image processing package. These characteristics represent the parameters of modal parameters (natural frequencies, damping ratios) under various stable and unstable conditions. This study employed the maximally stable external regions technique, which employs regions as the characteristic value to capture all modes of interest. The image feature extraction procedure is depicted in Fig. 6.

This method was configured to cluster each of the stabilisation diagram's physical poles before applying image feature extraction to build an image clustering plot. The used of the MATLAB function -find and a threshold is for separating the weak modes and leaving just the dominant modes. The threshold in the image clustering plot is set at 20 features; thus, 10 poles or more are considered dominant modes. An illustration

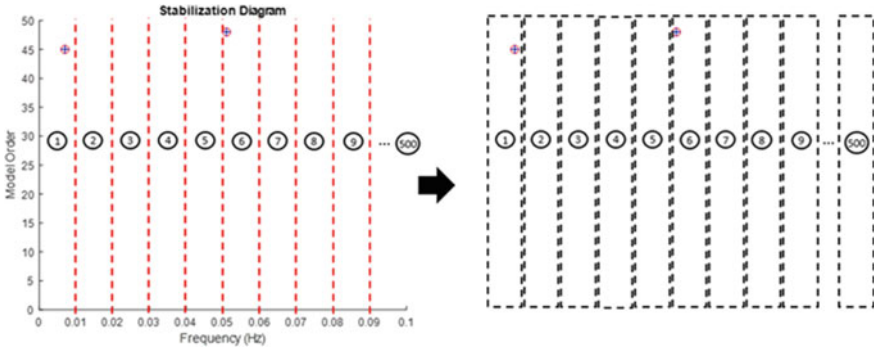


Fig. 5 Example generated input images for feature extraction from a stabilisation diagram with a 0.01 Hz interval frequency. The maximum frequency is 5 Hz

of a generated image that was identified as a physical mode by the clustering plot is shown in Fig. 7.

3 Conclusions

Instead of employing a typical clustering technique, any calibration or user-defined parameter at launch may be avoided by using image-based feature extraction for clustering, resulting in effective identification of structural modes from a stabilisation diagram. Using image feature extraction and image clustering, it has been demonstrated that it is possible to accurately identify physical modes in stabilisation diagrams, even for closely spaced modes, without the need for calibration, user-defined start-up parameters, or additional adaptive cluster validation criteria. This study will serve as the foundation for future research aimed at improving the automation of the OMA approach in structural health monitoring (SHM) systems.

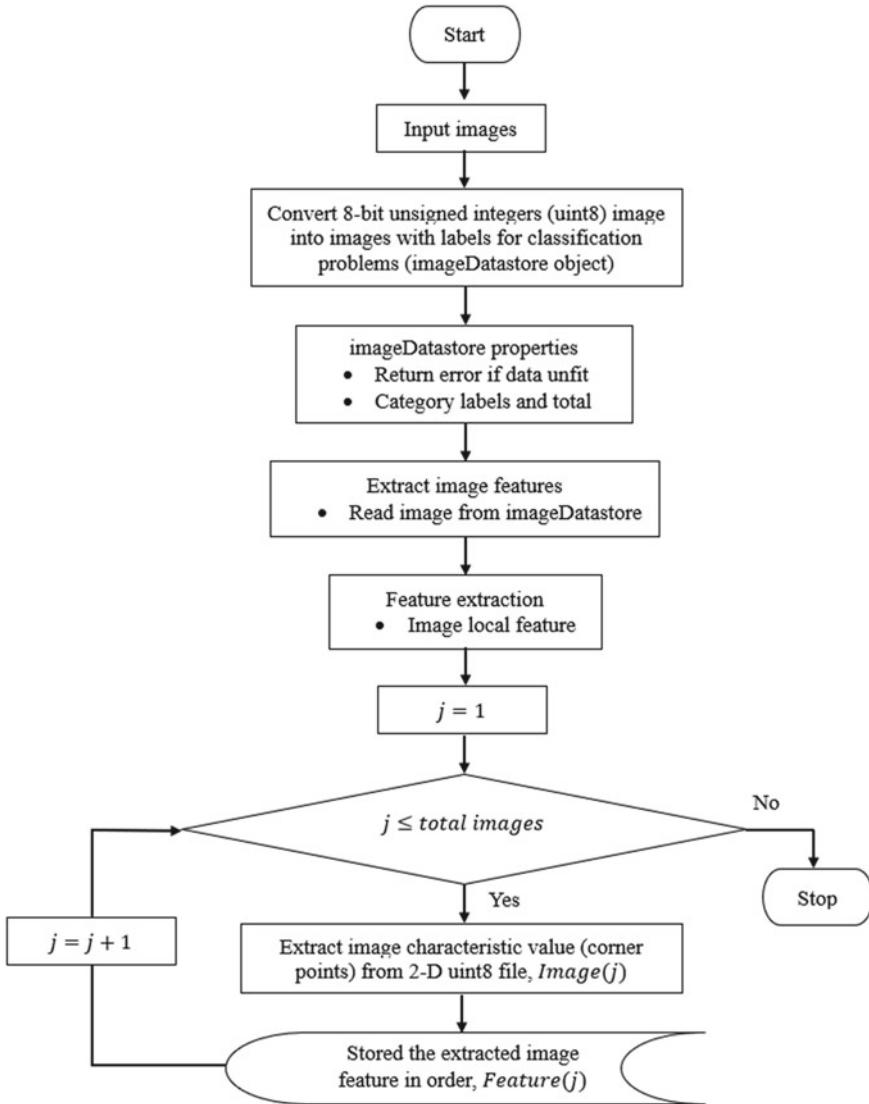
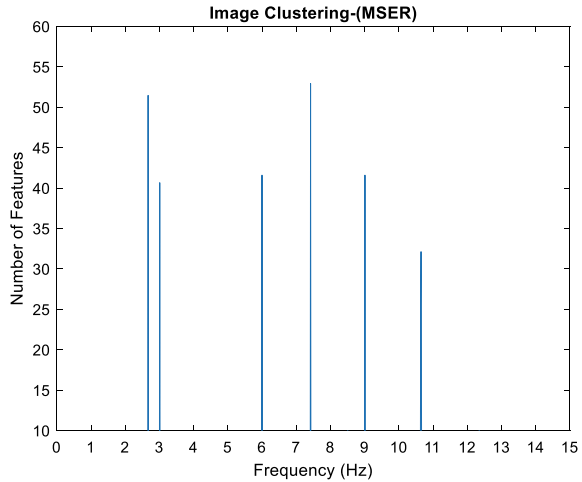


Fig. 6 Process flowchart for extracting image features

Fig. 7 Extraction of image features from maximally stable external regions is used to build an image clustering plot



Acknowledgements The authors would like to extend their greatest gratitude to the Institute of Noise and Vibration UTM for funding the current study under the Higher Institution Centre of Excellence (HICoE) Grant Scheme (R.K130000.7843.4J227). Additional funding for this research came from the UTM Research University Grant (Q.J130000.3824.31J47) and the Fundamental Research Grant Scheme (R.K130000.7840.4F653) from The Ministry of Higher Education, Malaysia.

References

1. Bricker R, Venture C (2015) Introduction to operational modal analysis. Wiley, Chichester
2. Shih CY, Tsuei YG, Allemang RJ, Brown DL (1988) Complex mode indication function and its applications to spatial domain parameter estimation. *Mech Syst Signal Process* 2(4):367–377. [https://doi.org/10.1016/0888-3270\(88\)90060-X](https://doi.org/10.1016/0888-3270(88)90060-X)
3. Peeters B (2000) System identification and damage detection in civil engineering. Katholieke Universiteit Leuven, Belgium
4. Brincker R, Andersen P, Jacobsen NJ (2007) Automated frequency domain decomposition for operational modal analysis. In: Conference proceedings of the society for experimental mechanics series, pp 1–7
5. Rainieri C, Fabbrocino G (2010) Automated output-only dynamic identification of civil engineering structures. *Mech Syst Signal Process* 24(3):678–695
6. Rainieri C, Fabbrocino G, Cosenza E (2007) Automated operational modal analysis as structural health monitoring tool: theoretical and applicative aspects. 347:479–484
7. Piodi F, Rizzi E (2017) A refined frequency domain decomposition tool for structural modal monitoring in earthquake engineering. *Earthq Eng Eng Vib* 16(3):627–648. <https://doi.org/10.1007/s11803-017-0394-9>
8. Piodi F, Ferrari R, Rizzi E (2017) Earthquake structural modal estimates of multi-storey frames by a refined frequency domain decomposition algorithm. *JVC/J Vib Control* 23(13):2037–2063. <https://doi.org/10.1177/1077546315608557>
9. Brandt A (2011) Noise and vibration analysis: signal analysis and experimental procedures. Wiley, Chichester

10. Peeters B, De Roeck G (2001) Stochastic system identification for operational modal analysis: a review. *J Dyn Syst, Meas, Control* 123(4):659. <https://doi.org/10.1115/1.1410370>
11. Reynders E (2009) System identification and modal analysis in structural mechanics. PhD thesis, K.U. Leuven
12. Piersol AG, Paez TL (2010) Harris' shock and vibration handbook, 6th edn. McGraw-Hill, New York, USA
13. Cabboi A (2012) Automatic operational modal analysis: challenges and applications to historic structure and infrastructures. PhD thesis, Università degli Studi di Cagliari
14. Hair JF, Anderson RE, Tatham RL, Black WC (1998) Multivariate data analysis. Upper Saddle River
15. Hasan MDA, Ahmad ZAB, Leong MS, Hee LM, Haffizzi Md. Idris M (2019) Cluster analysis for automated operational modal analysis: a review. *MATEC Web Conf* 255:02012. <https://doi.org/10.1051/mateconf/201925502012>
16. Reynders E, Houbrechts J, Roeck GD (2012) Fully automated (operational) modal analysis. *Mech Syst Signal Process* 29:228–250. <https://doi.org/10.1016/j.ymsp.2012.01.007>
17. Rainieri C, Fabbrocino G (2014) Operational modal analysis of civil engineering structures. Springer, New York. <https://doi.org/10.1007/978-1-4939-0767-0>
18. Schanke SA (2015) Operational modal analysis of large bridges. Master thesis, Norwegian University of Science and Technology (NTNU)
19. Magalhães F (2010) Operational modal analysis for testing and monitoring of bridges and special structures. PhD thesis, University of Porto
20. Gagnol V, Le TP, Ray P (2011) Modal identification of spindle-tool unit in high-speed machining. *Mech Syst Signal Process* 25(7):2388–2398. <https://doi.org/10.1016/j.ymsp.2011.02.019>

Detection of Lead with IoT Water Monitoring System Using Microstrip Antenna-Based Sensor



Abelle Chin Tze Hui, Sew Sun Tiang, Kah Hou Teng, Wei Hong Lim, and Mastaneh Mokayef

Abstract This paper presents a microstrip antenna-based sensor for detecting lead in water. The proposed antenna consists of a simple rectangular patch with inset feeding with an overall size of 50 mm × 45 mm. CST Studio Suite and COMSOL Multiphysics are used to simulate the characteristics of the antenna and analyze the reflection coefficients in different environment conditions. It can achieve a 10 dB reflection coefficient with VSWR <2 at 2.4 GHz. Results show that the reflection coefficient decreases with the increment of lead contents in the solution from 0.1 mg/L to 0.5 mg/L. With the salient antenna performances, the sensor demonstrates good sensitivity to detect lead content in water and upload the data using IoT analytic platform which is called ThingSpeak for data monitoring purpose. The proposed design is suitable to be used in IoT smart water quality monitoring system for lead detection as well as other heavy metal detection in water.

Keywords Lead detection · Smart water quality · Microstrip Patch Antenna · ThingSpeak

1 Introduction

In recent years, the global population has been widely increased, and the healthcare monitoring systems attracted considerable attention [1, 2]. There is a great demand for the development of cost-effective remote health monitoring that could be easy to use for daily life, especially for water quality. A smart water quality monitoring system is a system which enable user to monitor the water quality with aid of Internet of Things (IoT). It is a type of network which able to connect the devices with presence of Internet to gather data obtained from sensors embedded in devices or machines. Water quality monitoring system can be designed with different sensors and microcontrollers to detect desired parameters in water samples, such as pH values, turbidity, conductivity, and others [3, 4]. Most past research has focused on

A. C. T. Hui · S. S. Tiang (✉) · K. H. Teng · W. H. Lim · M. Mokayef
Faculty of Engineering, Technology and Built Environment, UCSI University, Lumpur, Malaysia
e-mail: tiangss@ucsiuniversity.edu.my

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
M. A. Abdullah et al. (eds.), *Advances in Intelligent Manufacturing and Mechatronics*,
Lecture Notes in Electrical Engineering 988,
https://doi.org/10.1007/978-981-19-8703-8_11

127

the smart water monitoring system by using Arduino or Raspberry Pi, pH sensor, flow sensor, turbidity sensor and ultrasonic sensor [4]. For more detailed water monitoring, some systems are specifically designed to detect specific content or heavy metals in water [5]. With the aid of IoT, the user able to detect excessive harmful content in water. Lead is also known as one of the heavy metals inside the water which will cause lead poisoning although it might take years for the effect to be shown. Different degree of lead exposure will affect series function in human body such as fertility disorders, high blood pressure, mood disorder, fatigue, or even death according to different level of lead exposure. As per World Health Organization (WHO) recommendation in 1993, a guideline value of maximum 0.5 mg/L for lead in drinking water is given. Lead is more common for water quality detection as the source of lead in water comes from lead pipe, faucets, or plumbing fixture. The increase of lead in drinking water will cause infection or diseases.

2 Literature Review

Microstrip patch antenna is a sensor with a thin metallic patch fabricated on dielectric substrate with a ground plane. Currently, microstrip technology is extensively studied in various fields such as communication, medicine, agriculture, and sensor detection. In wireless communication, antennas are typically used to transmit and receive electromagnetic waves; however, it can also be applied as sensors to detect the changes in the dielectric properties of the material under test in the sensing area. Several microstrip antenna-based sensor has been widely used to detect salt and sugar content in water [6–9].

Various microstrip patch antenna designs have been proposed for salt and sugar detection in water content. A rectangular-shaped sensor design with 51.3 mm × 51.3 mm, which is operated at 2.4 GHz, has been investigated for detection of 20%, 50%, and 80% of salt and sugar concentration [6]. In [7], a crescent-shaped microstrip sensor has been proposed with a FR4 substrate to detect different concentrations of salt and sugar solutions at different frequencies. A S-shaped resonator-based ring with Rogers RT/duroid 5880 by using inset feeding method was designed and able to detect different concentration of salt and sugar solution at resonant frequency of 6.20 GHz [8]. A tuning fork-shaped antenna with Rogers RT/duroid 5880 has been proposed as a sensing device resonating at 9.70 GHz to detect different concentrations of salt and sugar in water [9]. Table 1 shows the comparison of the various microstrip-based antenna sensors for salt and sugar detection in water. In this paper, a novel microstrip antenna with dimension of 50 mm × 45 mm is presented. The proposed antenna covers the frequency band from 1.0 GHz to 4.0 GHz, thereby maintaining a voltage standing wave ratio <2. The proposed antenna with FR4 material is used to track different concentrations of lead in water using microwave signals. The performance of the antenna sensor is optimized by performing parameter studies using CST Studio Suite and COMSOL Multiphysics software.

Table 1 Comparison of proposed antenna for salt and sugar detection in water

Authors, Year	Patch design	Dimension (mm)	Resonant frequency (GHz)	Limitation
[6]	Rectangular shaped	51.3 × 51.3	2.40	Poor reflection coefficient (<15 dB); Large microstrip antenna size
[7]	Crescent shaped	32.0 × 22.0	3.20, 8.70, 11.60, 14.50	Poor reflection coefficient, <21 dB
[8]	S-shaped resonator-based ring	24.0 × 18.0	6.20	Complex design and fabrication process
[9]	Tuning fork shaped	24.0 × 18.0	9.70	Complex design and fabrication process

3 Methodology

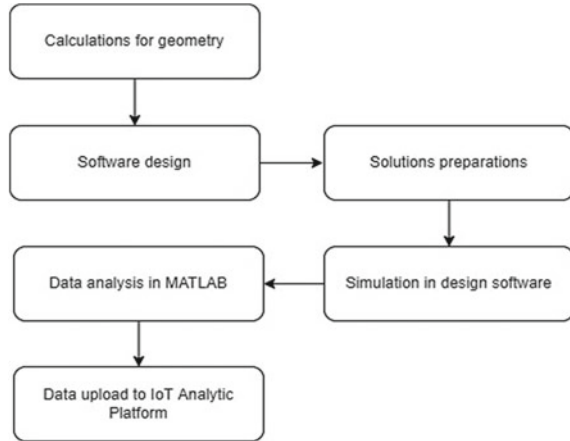
3.1 Overview

The purpose of this research work is to design a novel microstrip sensor for lead detection in water at 2.4 GHz and upload the data to IoT analytic platform. Figure 1 shows overview of the proposed smart water quality monitoring system. First, the initial geometrical parameters are estimated by using an empirical formula based on a resonant frequency. The resonant frequency is set at 2.4 GHz as smaller signal bandwidth resulted in small signal interference [10]. The input impedance of the sensor is 50Ω to maximize the power and voltage [11]. The parameter values are altered iteratively until satisfactory results are obtained over the entire operating frequency band. The sensor is designed in two different software to ensure the sensor is workable and simulate with different environment conditions when sensor was placed in. To obtain the properties of the solution required for testing in COMSOL Multiphysics, lab preparation and testing is performed after the antenna optimization process. The collected data are exported into MATLAB software and uploaded to IoT analytic platform.

3.2 Sensor Design

First, the proposed antenna is designed and simulated with CST Studio Suite at 2.4 GHz operating frequency. The geometry of the proposed antenna-based sensor is presented in Fig. 2, which is composed of a rectangular patch antenna with simple inset feeding. The optimized geometrical parameters of the proposed antenna sensor are shown in Table 2. The antenna performance depends on several factors such as

Fig. 1 Flowchart of the overview of the proposed smart water quality monitoring system



shape and patch size. The patch size and ground plane directly affect the performance of the patch and change the resonance frequency of the microstrip antenna. Also, the inset feeding directly attached to the patch can reduce the fraudulent radiation. The optimized antenna in CST Studio Suite is inserted into COMSOL Multiphysics for further simulation. COMSOL Multiphysics is a software which able to simulate the sensor under different environment conditions so that the user able to view the changes under different conditions. The validation process was done in COMSOL Multiphysics by changing the volume of the tested solution. Figure 2a and b show the sensor design in CST Studio Suite and COMSOL Multiphysics.

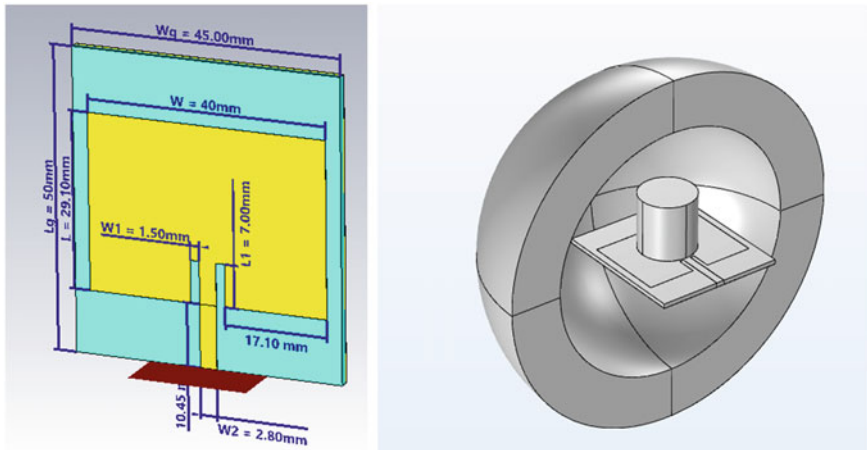


Fig. 2 Isometric view of proposed antenna-based sensor in **a** CST Studio Suite and **b** COMSOL Multiphysics

Table 2 Optimized parameters of proposed antenna sensor

Parameters	Values (mm)
Patch width, W	40.00
Patch length, L	29.10
Inset length, L_1	7.00
Inset width, W_1	1.50
Transmission line width, W_2	2.80
Ground width, W_g	45.00
Ground length, L_g	50.00

Fig. 3 Solution preparation and testing using AB200 Benchtop pH/conductivity meters



3.3 Solution Preparation

The electrical conductivity of the solution is obtained from the solution preparation process for further simulation in COMSOL Multiphysics. Lead (II) nitrate solution is prepared, and the concentration of the solution requirements adhere the Drinking Water Quality Standard of Group III with the lead concentration must be less than 0.5 mg/L. Hence, all the concentration of lead (II) nitrate solutions are prepared from 0.1 mg/L to 0.5 mg/L. AB200 Benchtop pH/conductivity meter is used to obtain the electrical conductivity of each concentration of solution. The data obtained and collected from AB200 Benchtop pH/conductivity meter are then inserted into COMSOL Multiphysics to simulate the sensor under different environment conditions. Figure 3 shows the equipment used to detect electrical conductivity of the solution.

3.4 Data Analyzing and Uploading

MATLAB Software was selected for analyzing and uploading data to IoT analytic platform. The finalized data from COMSOL Multiphysics are exported into

Table 3 Reflection coefficient of different ground plane dimension at 2.4 GHz

Dimensions (mm)	Reflection Coefficient (dB)
70 × 70	-10.16
60 × 60	-13.57
50 × 50	-24.76
50 × 45	-31.79
45 × 45	-29.81

MATLAB Software. The analyzed data are uploaded to IoT analytic platform, which is known as ThingSpeak. ThingSpeak is preferred for data monitoring in this project as it has connection with MATLAB Software, which will ease the task of uploading data.

4 Results and Discussion

4.1 Parameter Study of Microstrip Antenna-Based Sensor in CST Studio Suite

The dimension of the sensor is optimized in CST Studio Suite before inserting into COMSOL Multiphysics. The ground plane size plays a vital role in increasing impedance bandwidth and reflection coefficient because of the electromagnetic coupling between the radiating patch and the ground plane [12]. Table 3 shows the reflection coefficients with different sizes of ground plane.

The optimized dimension of ground plane was selected to be 50 mm × 45 mm. Thus, due to the further size reduction, the reflection coefficient decreases.

4.2 Validation results

Table 4 shows the comparison reflection coefficient at two resonant frequencies, 2.4 and 3.5 GHz when the sensor is designed with two different software, CST Studio Suite, and COMSOL Multiphysics 5.6. The environment condition in COMSOL Multiphysics is set to air, and the ground plane size is 50 mm × 45 mm. The peak resonant frequency 2.4 GHz is chosen as the design frequency [10]. According to the simulation result, the proposed antenna achieved impedance bandwidth from 1 to 4 GHz to maintain a VSWR < 2, with resonant frequency 2.4 GHz with a reflection coefficient of -31.79 dB for CST and -15.84 dB for COMSOL, respectively. Table 5 shows the simulation of different ground plane dimensions,

Table 4 Comparison reflection coefficient for CST Studio and COMSOL Multiphysics

Software	Frequencies	
	2.4 GHz (dB)	3.5 GHz (dB)
CST studio suite	-31.78	-5.85
COMSOL multiphysics 5.6	-15.84	-7.10

Table 5 Comparison of different ground plane dimension from COMSOL Multiphysics

Ground Plane Dimensions	Reflection coefficient (dB)
50 mm × 45 mm	-15.84
70 mm × 70 mm	-7.10

which is 70 mm × 70 mm and 50 mm × 45 mm. The reflection coefficient decreases with the increase in the size of ground plane, from -15.84 dB to -7.10 dB.

To ensure consistency of results, the proposed antenna is tested with two solutions with different volumes. The geometry of the cylinder is designed using COMSOL Multiphysics. Table 6 shows the comparison of simulated reflection coefficient of the proposed antenna sensor with different size of volume and solutions. The main resonance frequency is 2.4 GHz. For water solution, the reflection coefficient is -13.89 dB for 10 mm × 5 mm while -14.27 dB for 10 mm × 20 mm. From the results, the reflection coefficient decreases with the increase in the size of the volume of the water solution. The proposed antenna has been investigated using concentrated solutions of lead (II) nitrate solution of 5S/m electrical conductivity. The reflection coefficient variation at the peak resonance frequency at 2.4 GHz is -13.68 dB. This process is to prove that the ability of sensor to detect different kinds of solution. Figure 4 shows the reflection coefficient variation with the different size of the lead or water. For water solution, it is clearly observed that the resonant frequency shifts downward as the size increases. Hence, the change in resonant shift is due to the variation of the size and type of the solutions.

Table 6 Comparison of simulated reflection coefficient of the proposed antenna sensor with different type and size of solutions

Solutions	Cylinder geometry (Radius × height)	Frequencies	
		2.4 GHz (dB)	3.5 GHz (dB)
Water	10 mm × 5 mm	-13.89	-3.96
	10 mm × 20 mm	-14.27	-4.08
Lead (5S/m)	10 mm × 20 mm	-13.68	-3.93

Fig. 4 Reflection coefficient of different size and type of solutions

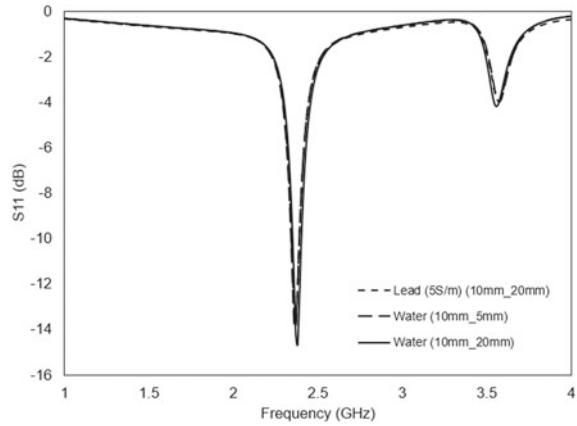


Table 7 Electrical conductivity and normalized data

Concentrations (mg/L)	EC ($\mu\text{S/cm}$)	Normalized EC ($\mu\text{S/cm}$)
0.1	3.26	4.9754
0.2	5.95	5.0191
0.3	5.08	5.0672
0.4	4.97	5.1064
0.5	5.97	5.1500

4.3 Solution Preparation

The lead (II) nitrate solution is prepared from concentration of 0.1 mg/L to 0.5 mg/L, and the electrical conductivity obtained is recorded using AB200 Benchtop pH/conductivity meters. The normalization of the results is performed as the electrical conductivity obtained from lab is inaccurate due to human error during preparation of solution. Table 7 shows the electrical conductivity and normalized electrical conductivity for different concentration of lead.

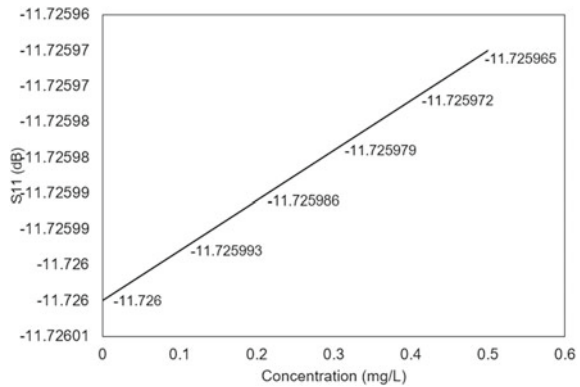
4.4 COMSOL Simulation Results

The normalized electrical conductivity was incorporated into COMSOL Multiphysics to obtain the reflection coefficient of each concentration. Table 8 shows the reflection coefficients of each concentration and normalized reflection coefficient for better analysis in MATLAB Software. The table also shows the differences of reflection coefficient comparing each concentration with distilled water (0mg/L). The mean of the differences of normalized reflection coefficient is $m = 21\mu\text{ dB}$. The

Table 8 Reflection coefficient and normalized reflection coefficient obtained from COMSOL Multiphysics

Concentration (mg/L)	Reflection coefficient (dB)	Normalized reflection coefficient (dB)	Differences compared to 0 mg/L (dB)
0.1	-11.72582601	-11.725993	7μ
0.2	-11.72581897	-11.725986	14μ
0.3	-11.72581123	-11.725979	21μ
0.4	-11.72580492	-11.725972	28μ
0.5	-11.72579790	-11.725965	35μ
			Mean, m = 21μ

Fig. 5 Small differences on reflection coefficient of each concentration



proposed antenna presented a good impedance matching with reflection coefficient below -10 dB. To ensure consistency of results, the proposed antenna is tested with different concentrations from 0.1 mg/L to 0.5 mg/L. From the results, the reflection coefficient decreases when the concentration of the solution increases, as the polarization of water decreases [7]. Although the difference of reflection coefficient between each concentration is small, but the decrement still able to detect as shown in Fig. 5.

4.5 MATLAB and IoT Analytic Platform Results

The normalized reflection coefficient was exported into MATLAB software for analysis. The exported data from MATLAB were first converted into array; then, Eq. (1) was used for the system to analyze the data. Due to the normalized reflection coefficient with 0.000001 differences, the decimal point in MATLAB was changed to 'long' to display and calculate 15 digits after the decimal points. The system is

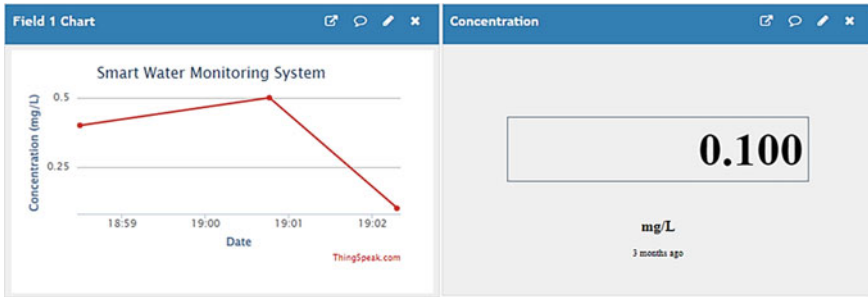


Fig. 6 Analyzed results displayed in ThingSpeak when user input reflection coefficient

designed for user input reflection coefficient and upload the data to IoT analytic platform, which is ThingSpeak as shown in Fig. 6.

$$y = mx + c \quad (1)$$

5 Conclusion and Recommendations

This paper presents a proof of concept for a smart water quality monitoring system using a microstrip patch antenna-based sensor to detect different concentrations of lead using microwave signals. The proposed antenna has a compact size of 50 mm × 45 mm on a low-cost FR4 substrate. The optimized parameters are simulated based on the changes of dimension of ground plane and volume of solution. The simulated of reflection coefficient demonstrates that the reflection coefficient decrease with small changes with the increment in percentages of concentrations of lead in water. Further improvements are required for sensor design to detect significant changes with different solutions and concentrations. The low-cost antenna-based sensor can be used in the smart water quality monitoring system with real-time monitoring in ThingSpeak platform.

Acknowledgements This work was supported by the Ministry of Higher Education Malaysia under the Fundamental Research Schemes with project codes of Proj-FRGS/1/2019/TK04/UCSI/02/1 and the UCSI University Research Excellence & Innovation Grant (REIG) with project code of REIG-FETBE-2022/038.

References

1. Yang Y, Wang H, Jiang R, Guo X, Cheng J, Chen Y (2022) A review of IoT-enabled mobile healthcare: technologies, challenges, and future trends. *IEEE Access*. Early Access
2. Han WT, Tiang SS, Lim WH, Mastaneh M, Ang KM, Liu Yanan (2021) Wearable sensors based remote patient monitoring using IoT data analytics. In: Chew E et al (eds) *Proceedings of the 8th international conference on robot intelligent technology and applications, LNME*. Springer, Singapore, pp 340–350
3. Lezzar F, Benmerzoug D, Kitouni I (2020) IoT for monitoring and control of water quality parameters. *Int J Interact Mob Technol* 14(16):4–19
4. Doni A, Murthy C, Kurian MZ (2018) Survey on multi sensor based air and water quality monitoring using IoT. *Indian J Sec Res* 17(2):147–153
5. Menon G, Ramesh M, Divya P (2017) A low cost wireless sensor network for water quality monitoring in natural water bodies. In: *2017 IEEE global humanitarian technology conference*. IEEE, pp 1–8
6. Njokweni S, Kumar P (2020) Salt and sugar detection system using a compact microstrip patch antenna. *Int J Smart Sens Intell Syst* 13(1):1–9
7. Islam MT, Rahman M, Singh M, Samsuzzaman M (2018) Detection of salt and sugar contents in water on the basis of dielectric properties using microstrip antenna-based sensor. *IEEE Access* 6:4118–4126
8. Rahman MN, Islam MT, Sobuz MS (2018) Microwave measurement system to detect salt and sugar concentration. *Microw Opt Technol Lett* 1772–1774
9. Rahman MN, Islam MT, Samsuzzaman M (2018) Detection of different concentrated salt and sugar solution based on dielectric properties using microstrip technology. *Microw Opt Technol Lett* 60(6):1573–1577
10. Marini S, Asrika S, Hasad A, Bivani M, Nisfani M (2021) Microstrip antenna 2.4 Ghz U-Slot patch dual slit vertical with ground square design for Zigbee technology. In: *Proceedings of the 2nd Borobudur international symposium on science and technology (BIS-STE 2020)*. Atlantis Press, pp 455–460
11. Chakravarthy S, Sarveshwaran N, Sriharini S, Shanmugapriya M (2016) Comparative study on different feeding techniques of rectangular patch antenna. In: *2016 thirteenth international conference on wireless and optical communications networks (WOCN)*. IEEE, pp 1–6
12. Khan MU, Sharawi MS, Mitra R (2015) Microstrip patch antenna miniaturisation techniques: a review. *IET Microwaves Antennas Propag* 9(9):913–922

Emotion Recognition Using Ultra-Short-Term ECG Signals with a Hybrid Convolutional Neural Network and Long Short-Term Memory Network



Vui Chee Chang, Jee-Hou Ho , Bee Ting Chan , and Ai Bao Chai 

Abstract This research aims to investigate emotion recognition using ultra-short-term electrocardiogram (ECG) signals with a hybrid convolutional neural network and long short-term memory (CNN-LSTM) network. DREAMER dataset consists of 23 subjects was used in this study. Raw data were recorded in the form of audio-visual stimuli during affect elicitation. ECG signals acquired from this dataset were pre-processed to filter noises. Single ECG cycle with one R-R peak interval was extracted using Pan–Tompkins algorithm and fed to the hybrid CNN-LSTM network. Another type of input in the form of decomposed ECG signals via empirical mode decomposition (EMD) was also investigated. The network was trained to classify high/low valence, arousal and dominance, respectively. Results show that hybrid CNN-LSTM network outperformed the basic CNN and LSTM network with classification accuracy ranges from 60 to 88% compared to 40.6% to 86.8% using basic configuration. Meanwhile, using EMD as the input achieved better recognition rates.

Keywords Emotion recognition · Electrocardiogram · Neural network

1 Introduction

Heart rate variability (HRV) can be defined as the fluctuations in the time interval between normal consecutive heartbeats. These beat-to-beat alternations can be quantified by measuring the distance between two consecutive R-peaks (interval between two R-peaks) in an electrocardiogram (ECG) signal. These intervals could reveal the variation of heart rhythm and the balance of autonomic nervous system (sympathetic and parasympathetic) activities [1, 2]. Traditionally, ECG was used to diagnose cardiac diseases such as epilepsy and arrhythmia. In recent years, ECGs are one of the commonly used physiological signals due to the information contained on the heart

V. C. Chang · J.-H. Ho (✉) · B. T. Chan · A. B. Chai
Department of Mechanical, Materials and Manufacturing Engineering, Faculty of Science and Engineering, University of Nottingham Malaysia, Jalan Broga, 43500 Semenyih, Selangor, Malaysia
e-mail: JeeHou.Ho@nottingham.edu.my

activity, where the heart rhythm changes could be correlated to the emotion. Non-invasive and versatility are amongst the good characteristics of using ECG signals in emotion recognition.

Automated affective recognition is an active research field, particularly within the healthcare industry. It relies on robust algorithms to automatically recognize human emotions by studying their involuntary physiological responses such as ECG. The recognition systems often incorporate artificial intelligence (AI) technology that focuses on detecting and responding to the affective state of a human, which is commonly known as computational empathy. The entire AI systems are primarily built upon the analysis of a physiological response to a stimulus, triggered by the SNS. During this response, there will be numerous physiological changes in the body that causes the human being to show a voluntary or involuntary physical reaction. For instance, when peoples are under overwhelming stresses in a short period of time, their bodies might release a mixture of hormones such as cortisol or adrenaline to prepare them for a physical reaction ('fight-or-flight'). However, outward expression is less reliable compared to the involuntary responses such as electroencephalogram (EEG) and ECG. Many external factors such as lighting, temperature, humidity, odour quality, surrounding noises, accessories wore on the subjects must be considered while extracting external voluntary modalities of expression [3]. These factors severely disturb and greatly prolong the duration of the experiment. Moreover, facial expressions, gestures and tone of voice can be manipulated easily (such as peoples hiding their shameful expression intentionally); these are the challenges for vision-based emotion recognition systems [4, 5]. As a non-invasive approach, using HRV or ECG signals acquired from biosensors could be a viable alternative in developing an automated affective recognition system.

In some biomedical applications, using HRV or gold standard of ECGs (5 min) may fail to work for intended applications in a real-time constrained environment. The AI-based applications require huge data processing, using long duration signals may consume more computational resources. This leads to the motivation of using ultra-short-term HRV or ECG signals in such biomedical applications. Liang et al. [6] validated the reliability of using ultra-short-term HRV (less than 5 min) analysis under rest and three post-exercise conditions. In emotion and stress recognition applications, Hwang et al. [7] proposed a deep learning framework Deep ECGNet using recurrent and convolutional neural network architecture for monitoring stressful states with ultra-short-term ECG signals. Improvement of recognition accuracy was observed comparing to conventional HRV method. Nita et al. [8] proposed a new data augmentation convolutional neural network (CNN) for human emotion recognition. In this study, ECG samples from DREAMER database [9] were augmented through randomization, concatenation and resampling and then fed to 7-layer CNN classifier. Classification accuracy of 77.54% to 95.16% was achieved. Nevertheless, this method requires considerable amount of data to be processed and augmented. Meanwhile, Dar et al. [10] proposed a CNN and long short-term memory-based (LSTM) emotion classifiers with physiological signal inputs such as EEG, ECG and galvanic skin response (GSR). The classifiers were tested on two public datasets of DREAMER and AMIGOS [11] and achieved average accuracy of 98.73% using

short-term ECG signals, 76.65% using EEG signals and 63.67% using GSR signals in emotion elicitation recognition. There are other published studies on emotion recognition using ECG signals [12–15]. Majority of the studies extracted combination of various features from long ECG signals to train with different machine learning schemes.

In this study, an emotion recognition scheme is developed using a hybrid CNN-LSTM neural network architecture. The network is fed with a single-cycle ECG signal segmented and interpolated into one R-R peak interval. DREAMER dataset is used to test the proposed network scheme classifying emotion into high and low states of valence, arousal and dominance.

2 Methodology

2.1 Dataset

DREAMER [9] is a publicly available dataset that consists of 23 healthy volunteers aged between 22 and 33 years old. Eighteen film clips were presented to the individuals in the form of audio and visual stimuli for affect elicitation. The length of the film clips ranged between 65 and 393 s. The subjects were presented with a neutral film clip at the beginning and between the 18 film clips to establish baseline signals. After viewing the film clip, each subject was asked to evaluate their emotion with a scale of 1–5 for three categories of emotions, low–high valence (unpleasant/stressed vs joy/relax), low–high arousal (bored/clueless vs excited/alert), low–high dominance (helpless vs empowered). This method is also known as Self-Assessment Manikins (SAM) [16].

All data were collected by SHIMMER™ ECG sensor [17] with a sampling frequency of 256 Hz. This dataset contained EEG and ECG signals. There are two channels of ECG recorded by the sensor: Lead-I (RA to LL) and Lead-II (LA to LL). Only, the lead-I ECG was used in this study. According to the establisher, there are multiple emotions recorded in each stimulus database. Hence, only, the last 60 s signals recording from each film clip was used to avoid unnecessary noises in the analysis.

2.2 ECG Signal Processing

Figure 1 shows the block diagram of the proposed methodology and processing sequence applied in this study. The detailed description of each processing step is further explained in the following subsections.

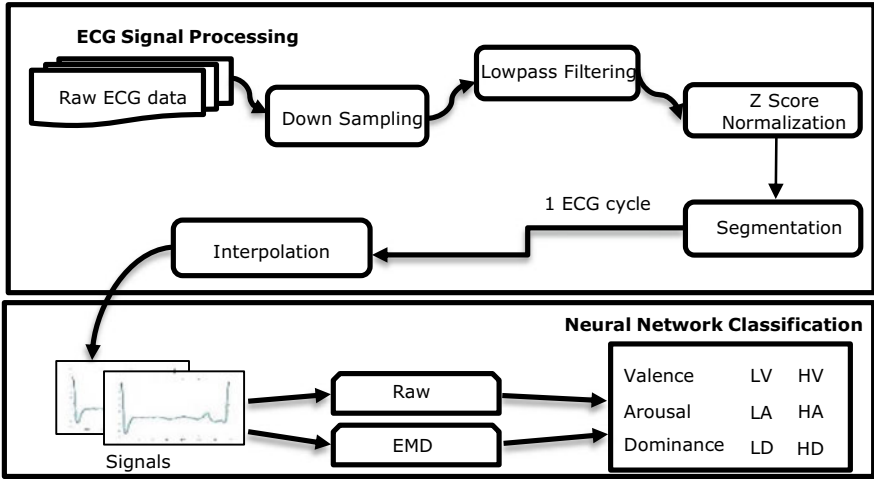


Fig. 1 Block diagram of the ECG signal processing and neural network classification

ECG Signal Downsampling and Filtering

Raw ECG data were extracted from the DREAMER database, and the sampling frequency was downsampled from 256 to 128 Hz. Next, a low-pass filter with a passband frequency of 60 Hz was applied to remove high-frequency noise artefacts from the ECG signals. After the filtering process, the last 64 s of the recording was extracted to avoid unnecessary emotion components during the classification. Due to the effect of filtering, there were spikes observed in the first and last 2 s of the extracted signal, and these spikes were then removed. The final length of the ECG signal was 60 s for each stimulus.

Z-Score Normalization

Normalization is commonly employed in physiological signals pre-processing as a general approach to standardize the magnitude and range of the signals. As such, Z-score normalization was applied to each of the 60 s filtered ECG signal. The fluctuations effects caused by the voltage amplitude may decrease the recognition rates to a large extent [18]. The common scale of standard deviation and zero mean allows the features extracted from the ECG signals to be comparable.

Segmentation and Interpolation

For the normalized ECG signals, the Pan–Tompkins algorithm was used to detect the locations of R-peaks and extract the R-R intervals. The R-R peaks interval covers the QRS complexes of a normal heartbeat. After normalizing the signal to a range between -3 to 6, the R-peak was assumed to be at least three sampling points above zero (half of the highest) and 65 sampling points for the R-to-R peak interval. This setup is the default setting for R-peak detection. After the segmentation, some anomaly signals were detected. There were missing or low R-peaks which caused

the algorithm failed to detect a complete R-R interval, and hence, the signal was stretched to the next detectable R-peak. These abnormal segmentations were manually removed. The remaining correctly segmented R-R intervals was used to represent as one ECG cycle; the data points vary in a range from 80 to 110 as the duration of each ECG cycle varies. Each ECG cycle was then interpolated to 128 data points such that consistent data points could be fed into the neural network classifier. This process is demonstrated in Fig. 2.

Empirical Mode Decomposition (EMD)

The empirical mode decomposition is a form of Hilbert–Huang transform, which assumes a signal could be represented by a series of intrinsic mode functions (IMF). EMD is popular in analysing non-linear and non-stationary time series [19] and eliminate unnecessary noise by decomposing them into n number of IMFs and one residual. In this study, ECG signals were decomposed into 5 IMFs and one residual. Figure 3 shows the original ECG signals acquired from a low valence (LV) case and the corresponding 4 IMFs after EMD decomposition.

The frequency spectrum obtained from the fast Fourier transform is used to identify the attributes in each IMF. IMF 1 is used in this study as the input to the neural

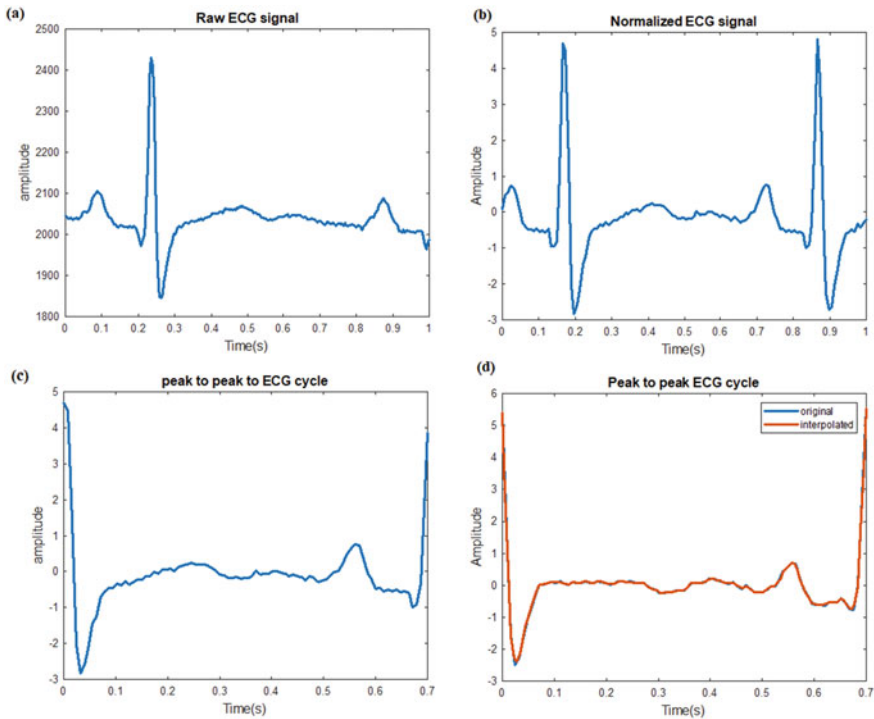


Fig. 2 a Raw ECG signal; b Normalized ECG signal; c 1 R-R peak interval ECG signal segment; d Original (blue) and interpolated (red) ECG signal

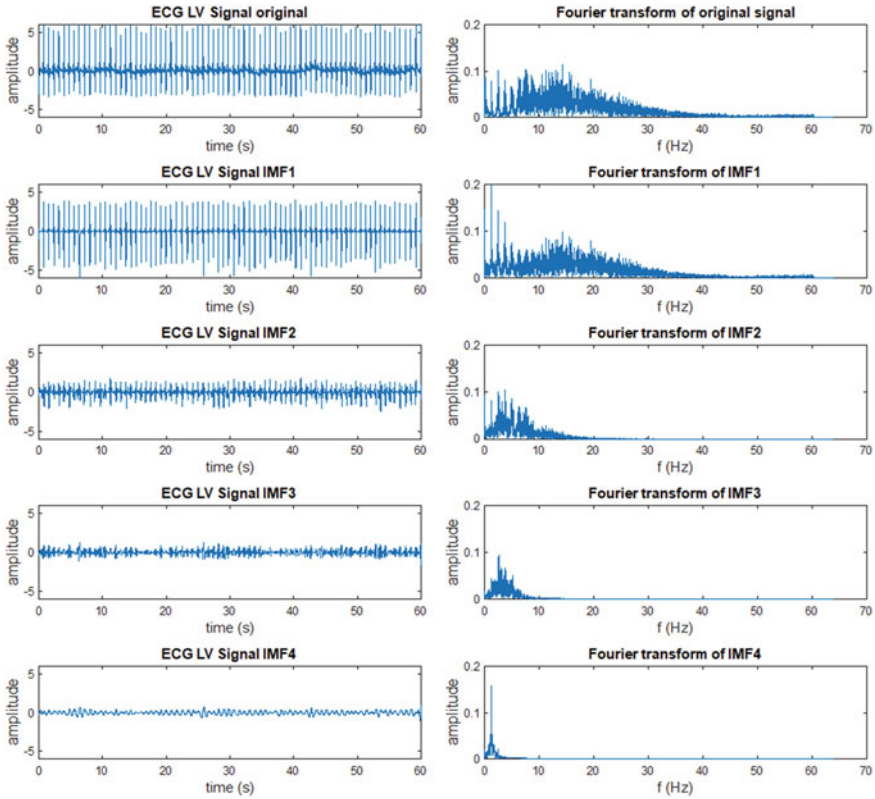


Fig. 3 Time series and frequency spectrum of the original ECG LV signal and its EMD decomposition

network classifiers as it exhibits similar characteristics of the ECG signal whereas the lower frequency noises are contained in IMF 2 to IMF 4.

2.3 Hybrid CNN-LSTM Network Architecture

In this study, a hybrid CNN-LSTM network is proposed as the emotion classifier. Two types of input were investigated, (i) the original 1-cycle ECG signal (Fig. 2d) and (ii) the decomposed ECG signal via EMD. Besides that, basic LSTM and 1D-CNN network were also tested in order to compare with the performance of the hybrid network.

LSTM network is a form of a recurrent neural network, which could learn the long-term dependencies between time steps of sequence data. This characteristic is suitable for the training of time series ECG signals. In particular, bi-directional

(BiLSTM) network was used to get a better recognition rate because it preserves the information from both past and future (captured the sequence in two directions) instead of one.

Two layers of convolution were used to extract the spatial features of the signals and condense it by using dot products multiplication. The weights and bias were used as the filters to remove unnecessary features and condense the meaningful instead. The gradient generated by the loss function in the end of each of the iteration will update the weights and bias via backpropagation. This allows the neural network to properly identify the meaningful features from the noise.

One ECG cycle data were captured by the sequence input layer as a sequence. The sequence then passed through the BiLSTM layer with 128 hidden layers. The state activation function and gate activation function are *tanh* and *sigmoid*, respectively. The output has linked to a fully connected layer of size 2. The output from the fully connected layer was connected to the SoftMax layer which lead to the classification output layer for the classification of low and high valence (LV/HV), low and high arousal (LA/HA), low and high dominance (LD/HD). These 3 emotions (valence, arousal, dominance) were tested separately and classified into 2 classes (low and high states).

The MiniBatchSize is 256 with an initial learning rate of 0.01 with Adam (derived from adaptive moment estimation) optimizer; the maximum Epoch is 60 and the gradient squared decay factor of 0.99. The training and testing of the neural network ran on a Core-i7 machine.

As shown in Fig. 4, the hybrid 1D-CNN and LSTM in sequential order is presented in the flowchart by following the branch of the blue arrows (series setup), while the parallel setup follows the branch of the orange arrows. The concatenation layer is used to concatenate both learnable outputs from previous layers and input the data into the fully connected layer.

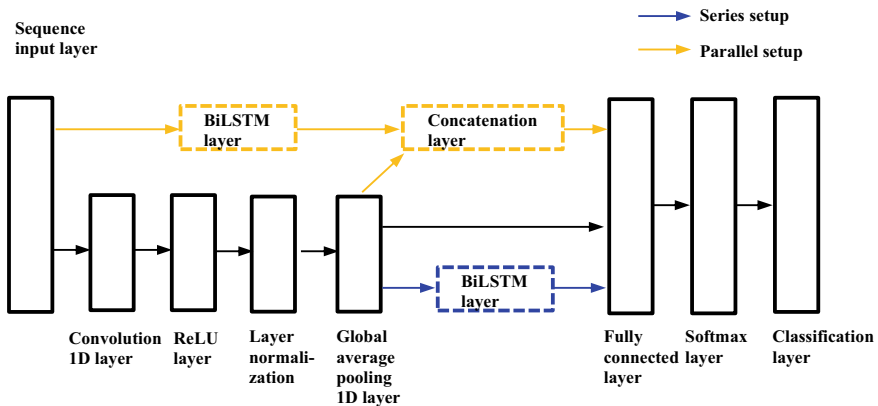


Fig. 4 Hybrid CNN-LSTM network architecture

The hybrid model has the advantage of extracting time-in-variant features of the input due to the convolution and still able to accurately capture the long-term dependency of the input. It is believed that the parallel setup is better than the series due to the input passes through both architectures equally before concluding which feature is more crucial for the classification. During the backpropagation, the weights and biases are updated to in both architectures at the same time.

3 Results and Discussions

Out of the total number of datasets, 70% was randomly split into training data; 15% were randomly split into validation, and 15% was randomly split into testing. With the correct setup, the training loss is expected to gradually decrease along with the training progress as the network accuracy is improving without overfitting the training data. Table 1 shows the classification accuracies of original ECG and EMD-decomposed ECG signals (using IMF 1).

As shown in Table 1, EMD-ECG decomposition generally achieved better classification accuracies in all neural network configurations. The improvement for basic LSTM network ranges from 4.6% to 15.7%. For the basic CNN network, improvement of 0.4% to 18.6% was achieved. For the hybrid CNN-LSTM networks, the improvements were mostly in a single-digit percentage range (slight decreases were observed in some cases). Notably, the hybrid network achieved good classification accuracies when the original ECG signals are used as the input, and hence, the improvements of using EMD-decomposed signals are relatively less. Nevertheless, the results showed that IMF 1 from the decomposed ECG signal was able to retain the important characteristics of ECG signal in response to the emotional stimuli. Meanwhile, since lower frequency noises were contained in other IMFs, IMF 1 may reveal better feature for emotion classification compared to the original ECG signal. Besides, the underwhelming performance of using the original ECG signals as the input (e.g. 40.6% accuracy in LV case with basic LSTM network) could be due to the acquired signals being highly interfered with the instrumental noises and emotional artefacts, as reported in other studies that used DREAMER database [10, 20]. A combination of signal pre-processing and the ability of the neural network to extract deep learning features could be vital to address these issues. On another note, the imbalanced distribution of emotion class instances amongst the subject could also result in the low recognition rates. Some subjects might not exhibit strong emotion states, or they might feel neutral or might not focus on the video contents during some parts of the experiment.

For the comparison of different neural network architectures, the hybrid CNN-LSTM network clearly outperformed both basic LSTM and CNN networks. In the hybrid network, the parallel setup performed slightly better in HV, HA and HD classes; meanwhile, the performance was comparable with the series setup in LV, LA and LD classes. The results prove that the hybrid network architecture is more robust in extracting the time-invariant features with its convolution layer while its

Table 1 Emotion classification accuracy

Classes	Basic LSTM		Basic CNN		Hybrid CNN-LSTM (Series/Parallel)			
	ECG (%)	EMD-ECG (%)	ECG (%)	EMD-ECG (%)	ECG (%)	EMD-ECG (%)	ECG (%)	EMD-ECG (%)
Low valence	40.6	56.3	45.6	64.2	63.9	63.6	63.6	65.7
High valence	60.8	65.7	70.6	83.1	71.2	82.5	85.2	88
Low arousal	50.4	64.2	58	65	62	64	63	70.5
High arousal	66.4	71	76.8	77.2	78.8	79.6	87.9	86.4
Low dominance	45.2	60.4	51.2	58	61	60	60	68.5
High dominance	62.2	75	72.1	86.8	74.8	76.6	83.9	81.4

LSTM layer recognizes the sequence input characteristics of the ECG signals. As for the computational resources, the time consumed in the basic 1D-CNN network was the least (average 197 s) as the network used condensed features for the classification. The computational time taken for the hybrid network in parallel setup was the longest (average 1240 s) as it feeds input into both 1D-CNN and LSTM for learning before concatenating for classification.

4 Conclusions

In this study, a hybrid CNN-LSTM network has been implemented for emotion classification by using ultra-short-term ECG signals. It has the potential to be a good surrogate of the gold standard 5 min ECG signals to assess emotion states. Besides that, while deep learning CNN network is famous for automatically extracting useful features for classification, the ECG signals could still be further processed with EMD decomposition before feeding into the deep learning network to further enhance the classification performance. Future studies should focus on testing the hybrid network with dataset of more balanced distribution of class instances with larger population and less artefact noises. Other physiological signals such as EEG and GSR could be investigated as well for emotion recognition.

References

1. Stein PK, Domitrovich PP, Hui N, Rautaharju P, Gottdiener J (2005) Sometimes higher heart rate variability is not better heart rate variability: results of graphical and nonlinear analyses. *J Cardiovasc Electrophysiol* 16(2005):954–959
2. Stein PK, Kleiger RE (1999) Insights from the study of heart rate variability. *Annu Rev Med* 50(1):249–261
3. Calvo RA, D’Mello S (2010) Affect detection: an interdisciplinary review of models, methods, and their applications. *IEEE Trans Affect Comput* 1(1):18–37
4. Ghali ALI, Kurdy MB (2018) Emotion recognition using facial expression analysis. *J Theor Appl Inf Technol* 96(18):6117–6129
5. Busso C, Deng Z, Yildirim S, Bulut M, Lee CM (2012) Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the 6th international conference on multimodal interfaces, Sorrento, Italy, pp 205–211
6. Wu L, Shi P, Yu H, Liu Y (2020) An optimization study of the ultra-short period for HRV analysis at rest and post-exercise. *J Electrocardiol* 63(2020):57–63
7. Hwang B, You J, Vaessen T, Myin-Germeys I, Park C, Zhang BT (2018) Deep ECGNet: an optimal deep learning framework for monitoring mental stress using ultra short-term ECG signals. *Telemedicine and e-Health* 24(10):753–772
8. Nita S, Bitam S, Heidet M, Mellouk A (2022) A new data augmentation convolutional neural network for human emotion recognition based on ECG signals. *Biomed Signal Process Control* 75:103580
9. Katsigiannis S, Ramzan N (2017) DREAMER: a database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE J Biomed Health Inform* 22(1):98–107

10. Dar MN, Akram MU, Khawaja SG, Pujari AN (2020) CNN and LSTM-based emotion charting using physiological signals. *Sensors* 20(16):4551
11. Miranda-Correa JA, Abadi MK, Sebe N, Patras I (2018) Amigos: a dataset for affect, personality and mood research on individuals and groups. *IEEE Trans Affect Comput* 12(2):479–493
12. Dissanayake T, Rajapaksha Y, Ragel R, Nawinne I (2019) An ensemble learning approach for electrocardiogram sensor based human emotion recognition. *Sensors* 19(20):4495
13. Sepúlveda A, Castillo F, Palma C, Rodriguez-Fernandez M (2021) Emotion recognition from ECG signals using wavelet scattering and machine learning. *Appl Sci* 11(11):4945
14. Panahi F, Rashidi S, Sheikhani A (2021) Application of fractional Fourier transform in feature extraction from Electrocardiogram and Galvanic Skin Response for emotion recognition. *Biomed Signal Process Control* 69:102863
15. Chen T, Yin H, Yuan X, Gu Y, Ren F, Sun X (2021) Emotion recognition based on fusion of long short-term memory networks and SVMs. *Digit Signal Process* 117:103153
16. Morris JD (1995) Observations: SAM: the self-assessment Manikin; an efficient cross-cultural measurement of emotional response. *J Advert Res* 35(6):63–68
17. Burns A, Greene BR, McGrath MJ, O’Shea TJ, Kuris B, Ayer SM, Stroiescu F, Cionca V (2010) SHIMMER™—a wireless sensor platform for noninvasive biomedical research. *IEEE Sens J* 10(9):1527–1534
18. Cimtay Y, Ekmekcioglu E (2020) Investigating the use of pretrained convolutional neural network on cross-subject and cross-dataset EEG emotion recognition. *Sensors* 20(7):2034
19. Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Yen NC, Tung CC, Liu HH (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc R Soc Lond Ser A: Math Phys Eng Sci* 454(1971):903–995
20. Li W, Zhang Z, Hou B, Song A (2021) Collaborative-set measurement for ECG-based human identification. *IEEE Trans Instrum Meas* 70:1–8

Enhancement of Morlet Mother Wavelet in Time–Frequency Domain in Electroencephalogram (EEG) Signals for Driver Fatigue Classification



Rafiuddin Abdubrani, Mahfuzah Mustafa, and Zarith Liyana Zahari

Abstract Driving is hazardous due to various factors, including driving attitudes, road type, and driving perceptual environment. These influences factors may cause a fatigue condition. Moreover, less driving experience and lack of alertness can also be contributed to dangerous accidents. Fatigued driving is a key factor in car accidents worldwide because of sleep disorders and driving durations. An EEG signal is used to determine changes in brain activity for diagnosing driver fatigue states. Artifacts were removed using independent component analysis (ICA) in the preprocessing stage. Then, features are extracted from the temporal region of the brain using eight channels (Fp1, Fp2, O1, O2, F4, F3, P4, and P3). The frequency bands used are alpha, delta, and theta. In continuous wavelet transform analysis, the Morlet wavelet is a fast wavelet transform in time–frequency analysis. Still, it has shift sensitivity and lacks phase information, affecting the frequency resolution analysis. This study proposes the enhancement of the Morlet mother wavelet for frequency resolution in the time–frequency domain using independent component analysis to overcome the drawbacks of the Morlet wavelet. The proposed technique can increase the percentage of driver fatigue classification accuracy of EEG signals. Then, the artificial neural network (ANN) classifier with Levenberg–Marquardt (LM) training algorithm gives the highest accuracy of the classification results with 97.40%, followed by the k-nearest neighbor (KNN) with 95.83% and the support vector machine (SVM) with 83%.

Keywords Driver Fatigue · Electroencephalogram (EEG) · K-Nearest Neighbor (k-NN) · Support Vector Machine (SVM) · Artificial Neural Network (ANN) · Morlet mother wavelet

R. Abdubrani (✉) · M. Mustafa · Z. L. Zahari

Faculty of Electrical and Electronics Engineering Technology, Universiti Malaysia Pahang, 26600 Pekan, Pahang, Malaysia
e-mail: rafi128@yahoo.com

Z. L. Zahari

Electronic Section, Universiti Kuala Lumpur British Malaysian Institute, 53100 Gombak, Selangor, Malaysia

1 Introduction

Fatigued driving is a key factor in car accidents worldwide. It can cause mental and physical disorders that lead to observable driving performance changes and increase the risk of road accidents. An electroencephalogram (EEG) can be used to observe the brain activities of a driver to determine the normal and fatigued state. Raw EEG signals contain artifacts, such as muscle activities, eye blinking, and unwanted noises. This paper presents a method for removing the noise that may distort the signal to some degree.

In this research, the independent component analysis (ICA) technique will draw the eye-blinking artifact without disturbing the EEG signal. Wavelet convolution is more computationally efficient than other methods and requires less code because it involves the fewest computations, most of which are performed using the fast Fourier transform [1]. A further benefit of the Morlet mother wavelet is that it offers good time resolution at high frequencies as well as good frequency resolution at low frequencies [2]. However, a phase shift happens in different EEG frequency bands and leads to bad frequency resolution in the scalogram region. The Morlet formula's time should be centered on the EEG frequency bands to avoid introducing the phase shift.

This work proposes an enhancement of the Morlet mother wavelet for frequency resolution in the time–frequency domain, independent component analysis-continuous wavelet transform (ICA-CWT) in preprocessing of EEG signals. The ICA-CWT removed unnecessary artifacts and produces clean EEG signals. Then, enhancement of the Morlet mother wavelet centers the frequency so that all the energy in the scalogram was calculated, and the results are more accurate during the classification process. Machine learning classifiers can be used to identify different stages of driver drowsiness. Classifiers such as artificial neural network (ANN), k-nearest neighbors (KNN), and support vector machine (SVM) are extensively employed in a variety of pattern recognition, text classification, ranking models, and object identification and event detection applications.

2 Related Works

EEG is an electrical brain impulse captured in the scalp by a multi-channel device for acquiring and storing information. Amplitude varies from $2\ \mu\text{V}$ to $100\ \mu\text{V}$ and frequency from 0.5 Hz to 100 Hz. Phases are associated with similar frequency ranges. Instead, electrical signals such as brain reactions to external stimuli such as visual or audio stimuli are evoked potential/event-related potential (EP/ERP). The frequency range is between 1 Hz and 3 kHz. Signals amplitudes range from $0.1\ \mu\text{V}$ to $20\ \mu\text{V}$ [3].

Raw EEG signals are contaminated with ocular, cardiac, muscle, and extrinsic artifacts. The artifacts are undesired electrical potentials originating from sources

other than the brain, such as the electrocardiogram (ECG), electromyogram (EMG), electrocardiogram (EOG), power line, and amplifier noise, poor electrode contact with the scalp, and current drift [4]. These artifacts need to be removed in preprocessing to get a clean EEG signal. Individual components of signals can be separated from a signal using ICA [5] and one of the removals of the effective artifact to remove eye blink [6] and muscle artifacts [7] from EEG signals.

Previously reported studies investigated the impact of continuous wavelet transform (CWT) to identify features from signals and denoise them [8]. The signal correlating function for the time scale is measured using the wavelet coefficients method to assess the energy features [9].

The machine learning algorithms are used for classification, such as k-nearest neighbor (KNN), artificial neural network (ANN), and support vector machine (SVM), to classify the subjects normal and fatigue conditions. It is also important to identify suitable machine learning to classify the fatigue states accurately to separate the data into categories. SVM is resistant to overfitting and delivers excellent generalization performance when applied to various time series forecasting problems encountered in the field of time series forecasting [10]. A previous study found that the SVM classifier-based sleep apnea classifier was more accurate than other classifiers [11]. Furthermore, the previous researcher used the three-type classifier for the emotion recognition and the accuracy rates of KNN with 71.7%, SVM with 78.75%, and ANN with 82.03%, respectively [12]. The ANN classifier shows the best machine learning algorithm for their research.

3 Methodology

The research design is divided into these major parts, data acquisition, preprocessing, feature extraction, and classification.

As shown in Fig. 1, achieving the proposed work enhances frequency resolution to better energy conservation results. The ICA technique was implemented by preprocessing EEG signals to eliminate unwanted noises and artifacts such as muscle activities and eye blinking. Other purposes of using ICA are to gain an EEG amplitude range between $-100\mu\text{V}$ to $100\mu\text{V}$ and locate the signal in the proper coordinate. Therefore, the EEG amplitude range from $-100\mu\text{V}$ to $100\mu\text{V}$ is considered the clean EEG signal. Clean EEG signals are divided into three sub-bands, alpha, delta, and theta. Then, in the feature extraction section, CWT is the most widespread technique for time–frequency domain analysis to extract energy conservation from EEG signals [13]. The mother wavelet is the Morlet wavelet, which comprises a complex sinusoid within a Gaussian envelope. The Morlet mother wavelet has a time scale of t , and the connection between scale and frequency is inverse, meaning that frequency increases as scale decreases [14]. Finally, the KNN algorithm in the tenfold cross-validation strategy is used in the classification process to classify between fatigue and normal states.



Fig. 1 Design of the research

3.1 Data Acquisition

The dataset is prospectively available online from an Internet database by a previous researcher [15]. The experiment uses a Neuroscan device with 30 electrodes channel, including two reference channels as shown in Fig. 2, digitized at 1000 Hz. The red boxes are the channel selected for further analysis. The dataset was divided into normal and fatigue states with five minutes duration. The subjects are twelve young, healthy men aged 19–24 years who participated in a simulator for driving. The ZY-31D vehicle driving simulator was developed by Pe-king ZhongYu CO., LTD, based in Beijing's Daxing district, and features a wide-screen made of three 24-inch displays. A low traffic density scenario was created using a Peking ZIGUANGJIYE program ZG-601 as the driving environment. The driving task started at 0900 h, and subjects continuously drove for up to 2 h. There have two phases, normal state and fatigue state. The last 5 min after subjects drive for 20 min will categorize as a normal state and after 40–100 min after subjects, self-reported fatigue questionnaire results indicated a fatigued state.

Fig. 2 International 10–20 system for electrode placement [16]



3.2 *Electrode Selection*

One of the channel selections for this paper is channel O1, which improves driver fatigue detection system performance because it is the highest correlation compared to other channels. EEG electrodes, O1 and O2, are chosen subjectively based on topographic maps of average between centrality under two mental states, mostly found in the occipital brain [17]. The other active electrode is Fp1 which positively impacts the classification performance of driver fatigue. Using a combined Fp1 and Fp2 electrode with 85% accuracy was more accurate than using Fp1 alone or Fp2 alone to classify fatigue driving. This was compared to 79% for Fp1 and 68% for Fp2, respectively [18]. In the brain's frontal area, F4 was selected for further analysis because it can provide the highest performance in the classification process and electrode Fp1. Another electrode which is P4 was found to be more beneficial in drowsiness and poor driving performance study and will be used in this analysis. In addition, electrodes O2, Fp2, F3, and P3 are used in further research. This is how the previous researcher got 62.3% of the time. Ateke Goshvarpour and Atefeh Goshvarpour used ten EEG channels and some channels, like Fp1, Fp2, F3, and F4, to get the best results [19]. The most active EEG electrodes and optimum EEG channels are Fp1 and P3. These channels were previously successfully used in EEG emotion recognition [20].

3.3 *Preprocessing with ICA*

The programs are written entirely using MATLAB software. There are always noises and other interferences in the raw EEG signals that need to be removed before processing further. In preprocessing, the ICA technique was applied to remove unwanted noises and artifacts, isolating source signals (IC) from observed signals without knowing the mixture beforehand. The ICA mixing model can be stated in vector–matrix notation as follows:

$$x = As \tag{1}$$

The unknown mixing matrix A is represented by each row of the matrix x, while each row of the matrix s represents the observed mixed signal.

3.4 *Feature Extraction*

One of the feature extractions to reduce the amount of redundant data from the dataset is continuous wavelet transform (CWT). It is a fundamental transformation that forms a suitable distribution of time frequencies. It is particularly well suited

for the intended tasks due to its high computing performance; provided sufficient wavelets of analytics are used. The CWT is defined by;

$$CWT\{h(x)\} = S_{CWT}(x, \omega) = \int_{-\infty}^{\infty} h(\varepsilon)\psi * ((\varepsilon - x)\omega)d\varepsilon \quad (2)$$

This produces a location-dependent spectral decomposition, such as the spectral response for both spatial frequency ω and spatial location x . ψ is referred to as the mother wavelet.

According to the Morlet wavelet, the impulse voltage signal can be used as an excitation to extend the frequency range and enhance sensitivity performance. The mother wavelet must show a non-zero value or finite energy to meet a suitability requirement. Equations (3)–(5) are spelled out these criteria. The description of the Morlet wavelet is given in Eq. (7), and it is important to align frequency in the center and required bandwidth in signal processing. The researcher improved the mother wavelet parameter from $\Psi(\omega)$ to $\psi(t)$ in the Fourier transform method [21].

$$\int_{-\infty}^{\infty} \psi(t)dt = 0 \quad (3)$$

$$\int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty \quad (4)$$

$$C_{\psi} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{|\omega|} d\omega < \infty \quad (5)$$

$$\psi(t) = \exp\left(-\frac{t^2}{2}\right)\cos(5t) \quad (6)$$

$$\psi(t) = \exp\left(-\frac{t^2}{25}\right)\cos(2\pi t) \quad (7)$$

3.5 Classification

The k-nearest neighbors (KNNs) algorithm is a data classification and regression technique that is often used to find patterns and consistency in data. KNN is a method for supervised learning that is part of a family of algorithms. KNN is a non-parametric technique used to evaluate statistical data and data analysis [22]. The various distance metrics accessible include Euclidean, city block, cosine, correlation, and hamming.

Specifically, the distance metric employed in this work is the Euclidean distance. The ideal value of k is obtained by conducting a series of tests with the variable from one to ten of the k value number. In this paper, the experiment evaluated the number of nearest neighbors, k from one to ten, to find that the highest classification accuracy, and number of cross-validation is ten. According to the previous researcher, the best classifier was discovered to be a fine KNN classifier with a classification accuracy of 100.0%. Furthermore, the best category is KNN because all classifiers in this category attained a classification accuracy of more than 90.0% [23].

Then, the other classifier used for classification is artificial neural network (ANN) models that perform biological neural network organization and functions. The dataset is divided into 70% of the training dataset, 15% of the testing dataset, and the other 15% to the validation dataset. Then, ten hidden layers and Levenberg–Marquardt (LM) algorithm are used in this paper. The mathematical expression for ANN computation is as follows:

$$y_j = \sum_{i=1}^n w_{ij} x_i + \theta_j \tag{8}$$

y_j denotes parameters that have been shifted to the following layer, and j represents a node. In this case, n is the number of moving edges to node j ; x_i stand for those things that are entered through a unit to the node of node, j , θ_j equals bias node, j .

A support vector machine (SVM) is used for the separation of data into two or more categories. Then, a tenfold cross-validation is used to estimate the SVM’s performance. The dataset is divided into 90% of the training dataset and 10% of the testing dataset. The separation can distinguish invisible data with a sufficient capacity for generalization during the study to produce a linear with a high margin for the classes. The SVM classifier can be defined as;

$$g(x) = w \cdot \Phi(x) + b \tag{9}$$

$$w = \sum_{i=1}^n a_i y_i \Phi(x_{pi}) \tag{10}$$

$$b = \sum_{i=1}^n a_i y_i \Phi(x_{pj}) + y_i \tag{11}$$

where w is normal to separating hyperplane defined by $\Phi(x)$.

4 Results and Discussions

In preprocessing part, channels of Neuroscan, Fp1, Fp2, O1, O2, F4, F3, P4, and P3, were used for driver fatigue classification. The original EEG signal from recording systems was contaminated with the artifacts, as shown in Fig. 3. ICA technique was applied to remove the artifacts, and the noise was eliminated, as shown in Fig. 4. EEG data acquisition was divided into three sub-bands, which are alpha (8–13 Hz), delta (0.5–4 Hz), and theta (4–8 Hz). Alpha brainwaves are the most immediately observable and were the first to be found. They can be discerned when the eyes are closed and the mind is at ease. Theta brainwaves are most visible when sleeping or near the stage of sleep. Delta brainwaves are the slowest of all brainwaves and most active while sleeping deeply and dreamlessly. Healing and rejuvenation are also stimulated in this condition, which is why getting enough sleep every night is so important.

As part of the feature extraction process, the continuous wavelet transforms using the Morlet as the mother wavelet was employed to extract the features. The upper and lower bound value for effective support in the Morlet wavelet needs to identify for better energy conservation results. As shown in Fig. 5, the result of energy conservation is good with the default parameter of the Morlet mother wavelet. However, improving frequency resolution is important for delta and theta bands to get better energy conservation and avoid inaccurate classification results. As shown in Figs. 6 and 7, energy conservation for both delta and theta bands is not within the scalogram region. The phase shift of the Morlet mother wavelet was applied to improve the frequency resolution of the delta and theta band, as shown in Figs. 8 and 9.

The KNN, SVM, and ANN are utilized as classifiers to classify the normal and fatigue state of the drivers. The class is divided into normal (1) and fatigue (2). The k value for the KNN classifier used in this research is up to 10 to find the most suitable

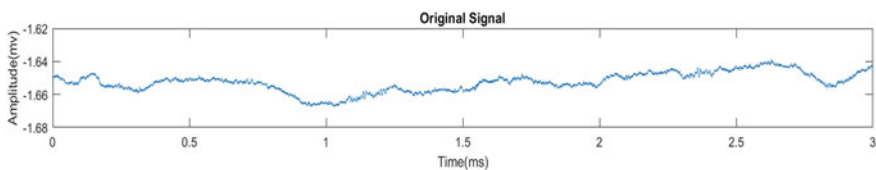


Fig. 3 Original EEG signal

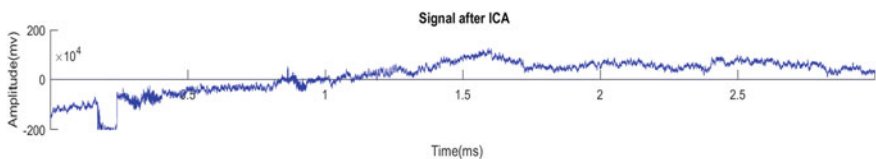


Fig. 4 Clean EEG signal after ICA applied

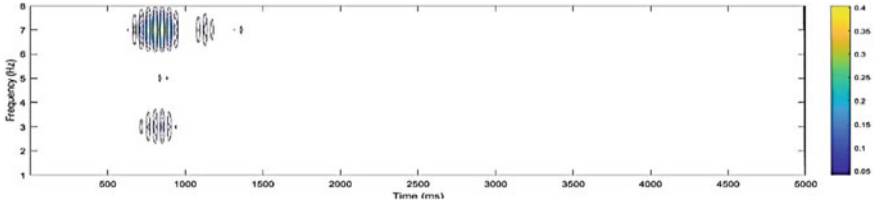


Fig. 5 Energy distribution of alpha band in normal state

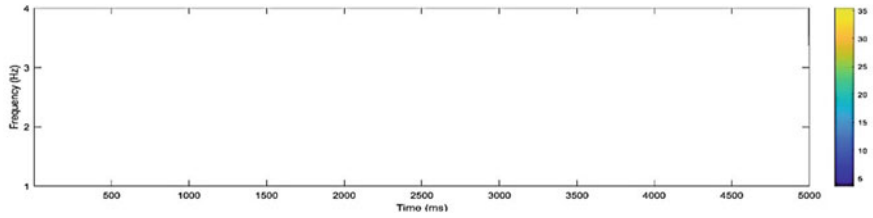


Fig. 6 Energy distribution of delta band before the enhancement of Morlet wavelet

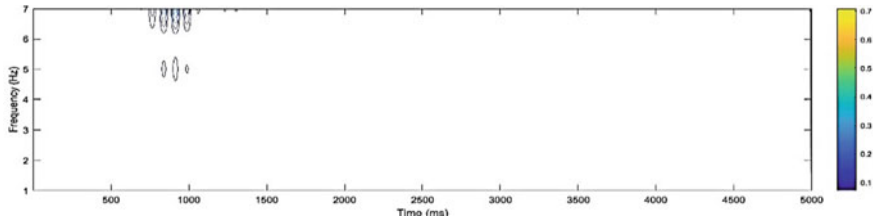


Fig. 7 Energy distribution of theta band before the enhancement of Morlet wavelet

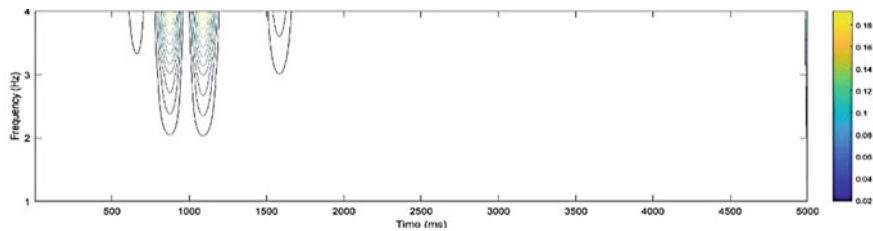


Fig. 8 Energy distribution of delta band after the enhancement of Morlet wavelet

value with high accuracy results. The cross-validation of KNN is tenfold. The highest accuracy of the KNN classifier at k value is 1 with 95.83%.

Then, the highest accuracy of the SVM classifier is 83% with tenfold cross-validation, and ANN is the highest result for the classification with 97.40% accuracy. The analysis found that ANN gives the most accurate classification result for driver

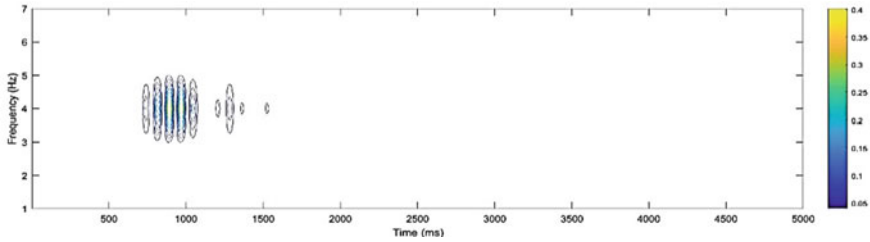


Fig. 9 Energy distribution of theta band after enhancement of Morlet wavelet

Table 1 Classifier results

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)
KNN	95.83	93.10	98.28
SVM	83.00	83.00	83.00
ANN	97.40	96.61	98.25

fatigue states detection compared to KNN and SVM. Levenberg–Marquardt’s (LM) algorithm was used for the ANN modeling with ten hidden layers. The sensitivity and specificity of ANN are 96.61% and 98.25%. Table 1 shows the results for KNN, SVM, and ANN accordingly.

5 Conclusion

This paper proposes enhancing the Morlet mother wavelet in CWT and the ICA technique to increase the percentage of driver fatigue classification accuracy. The enhancement helps classify fatigue and normal state precisely for driver fatigue classification. Then, the ANN classifier is the classifier that gives accurate classification results with 97.40%. The evaluation results show that the enhancement method can improve and provide better results in terms of accuracy. For future work, implementing various feature extraction algorithms can determine the most suitable features for driver fatigue classification.


References

1. Cohen MX (2019) A better way to define and describe Morlet wavelets for time-frequency analysis. *Neuroimage* 199(April):81–86
2. Darnila E, Ula M, Tarigan K, Limbong T, Sinambela M (2018) Waveform analysis of broadband seismic station using machine learning Python based on Morlet wavelet. *IOP Conf Ser Mater Sci Eng* 420:012048
3. Chakraborty S, Aich S, Il Joo M, Sain M, Kim HC (2019) A multichannel convolutional neural network architecture for the detection of the state of mind using physiological signals from

- wearable devices. *J Healthc. Eng*
4. Monteiro TG, Skourup C, Zhang H (2019) Using EEG for mental fatigue assessment: a comprehensive look into the current state of the art. *IEEE Trans Human-Mach Syst* 49(6):599–610
 5. Lai CQ, Ibrahim H, Abdullah MZ, Abdullah JM, Suandi SA, Azman A (2018) Artifacts and noise removal for electroencephalogram (EEG): a literature review. *IEEE Symp Comput Appl Ind Electron (ISCAIE)* 2018:326–332
 6. Tharwat A (2018) Independent component analysis: an introduction. *Appl Comput Inform* 17(2):222–249
 7. Jiang X, Bin Bian G, Tian Z (2019) “Removal of artifacts from EEG signals: a review. *Sensors (Switzerland)* 19(5):1–18
 8. Lestari FPA, Pane ES, Suprpto YK, Purnomo MH (2018) Wavelet based-analysis of alpha rhythm on EEG signal. In: 2018 International conference on information and communication technology ICOIACT 2018, vol 2018-Janua, pp 719–723
 9. Karuppusamy NS, Kang BY (2020) Multimodal system to detect driver fatigue using EEG, gyroscope, and image processing. *IEEE Access* 8:129645–129667
 10. Zahari ZL, Mustafa M, Zain ZM, Abdubrani R, Naim F (2021) The enhancement on stress levels based on physiological signal and self-stress assessment. *Trait Signal* 38(5):1439–1447
 11. Vimala V, Ramar K, Ettappan M (2019) An intelligent sleep apnea classification system based on EEG signals. *J Med Syst* 43(2)
 12. Mert A, Akan A (Oct.2018) Emotion recognition based on time–frequency distribution of EEG signals using multivariate synchrosqueezing transform. *Digit Signal Process* 81:106–115
 13. Signal MEEG, Matthew U (2021) *Jurnal Teknologi music-based EEG signal using Matthew correlation coefficient*, vol 6, pp 53–61
 14. Mattar EA, Al-Junaid HJ, Al-Mutib KN (2019) Electroencephalography features extraction and deep patterns analysis for robotics learning and control through brain-computer interface. In: 2019 International conference on innovation and intelligence for informatics, computing, and technologies 3ICT 2019, pp 1–6
 15. Min J, Wang P, Hu J (2017) Driver fatigue detection through multiple entropy fusion analysis in an EEG-based system. *PLoS ONE* 12(12):1–19
 16. Huang Q, Zhang Z, Yu T, He S, Li Y (2019) An EEG-/EOG-based hybrid brain-computer interface: application on controlling an integrated wheelchair robotic arm system. *Front Neurosci* 13(November):1–9
 17. Wang F, Wu S, Ping J, Xu Z, Chu H (2021) EEG driving fatigue detection with PDC-based brain functional network. *IEEE Sens J* 21(9):10811–10823
 18. Luo H, Qiu T, Liu C, Huang P (2019) Research on fatigue driving detection using forehead EEG based on adaptive multi-scale entropy. *Biomed Signal Process Control* 51:50–58
 19. Islam MR, Ahmad M (2019) Wavelet analysis based classification of emotion from EEG signal. In: 2nd international conference on electrical, computer and communication engineering ECCE 2019, pp 1–6
 20. Goshvarpour A, Goshvarpour A (2020) A novel approach for EEG electrode selection in automated emotion recognition based on lagged poincare’s indices and sLORETA. *Cognit Comput* 12(3):602–618
 21. Wu JJ, Huang JJ, Qian T, Tang WH (2019) Study on nanosecond impulse frequency response for detecting transformer winding deformation based on Morlet wavelet transform. In: 2018 international conference on power system technology POWERCON 2018 - no. 201804270000428, pp 3479–3484
 22. Devika R, Avilala SV, Subramaniaswamy V (2019) Comparative study of classifier for chronic kidney disease prediction using naive bayes, KNN and random forest. In: Proceedings of 3rd international conference on computing methodologies and communication, ICCMC 2019, no. ICCMC, pp 679–684
 23. Tuncer T, Dogan S, Subasi A (2021) EEG-based driving fatigue detection using multi-level feature extraction and iterative hybrid feature selection. *Biomed Signal Process Control* 68(2020):102591

Fabrication of Aneurysm Biomodel Using 3D Printing Technology



Jamil Ahmad Hisam, Muhamad Yusof Salehudin,
Muhammad Ismail Mat Lizah, Muhammad Izzat Ahmad Suhaimi,
Muhammad Haqim Muhammad Hisham, Ismayuzri Ishak ,
and Mohd Jamil Mohamed Mokhtarudin 

Abstract Performing endovascular treatment requires highly skilled surgeons to avoid surgical errors. The development of an in vitro training tool for endovascular treatment is essential and requires the development of an artificial blood vessel or a biomodel. In this project, an aneurysm biomodel is fabricated using 3D printing technology. Firstly, an idealized saccular-type aneurysm geometry is developed. Then, a mould is fabricated using 3D printing following the geometry. The biomodel must be transparent and hollow to ease the visualization while performing fluid flow experiment. In order to fabricate this, the lost core method is used. The mould core is fabricated using poly-vinyl alcohol (PVA), which can easily be dissolved when soaked in water. Meanwhile, other parts of the mould are fabricated using poly-lactic acid (PLA). Then, an agar–water mixture is used to make the biomodel by pouring into the mould and then froze at 0 °C for 30 min. The biomodel produced has about 5% shrinkage from the original geometry. In addition, the biomodel fabricated is flexible but is easily teared depending on the agar–water ratio used, which prevents it from being used for the in vitro experiment. Improvement of the biomodel materials could overcome the limitations from the current technique.

Keywords Aneurysm · Biomodel · 3D printing

1 Introduction

Endovascular treatment such as stenting is used to treat blood vessel diseases such as aneurysm. Performing the treatment procedure is quite challenging for medical

J. A. Hisam · M. Y. Salehudin · M. I. M. Lizah · M. I. A. Suhaimi · M. H. M. Hisham · I. Ishak · M. J. M. Mokhtarudin (✉)
Faculty of Manufacturing and Mechatronics Engineering Technology, Universiti Malaysia Pahang,
26600 Pekan, Pahang, Malaysia
e-mail: mohdjamil@ump.edu.my

M. J. M. Mokhtarudin
Centre for Research in Advanced Fluid and Processes (Fluid Centre), Universiti Malaysia Pahang,
Lebuhraya Tun Razak, 26300 Kuantan, Pahang, Malaysia

surgeons, and this requires the appropriate training tools in order to improve their technical skills [1]. In vitro training tools use a geometrically accurate blood vessel model to train surgeons performing the endovascular treatment operation [1]. This artificial blood vessel or a biomodel must be able to resemble the mechanical properties of the actual blood vessel and also be transparent to allow for the visualization of the fluid within the biomodel [1, 2].

Materials such as poly- (vinyl alcohol) hydrogel (PVA-H) [3], silicon [4] and polydimethylsiloxane (PDMS) [5, 6] have been used extensively in producing the blood vessel biomodels. However, to fabricate the biomodels require three major steps. Firstly, the geometry of the blood vessel must be constructed using computer-aided drawing (CAD). The geometry can either be in idealized form [7, 8] or in patient-specific form [5, 6]; the latter is harder to fabricate due to its irregular structure [9]. Secondly, a mould is created before the material can be poured into it and follow the shape of the biomodel according to the CAD geometry. Lastly, the biomodel is produced through a series of chemical processes.

The production of a mould is important to allow for a sustainable production of the biomodel. Rapid prototyping or additive manufacturing is a technology that enables the production of a complex structure by converting information from CAD 3D model into a physical object in a layer-by-layer manner [10]. There are numerous rapid prototyping techniques, for examples the stereolithography (SLA), selective laser sintering (SLS), fused deposition modelling (FDM), direct metal laser sintering (DMLS) and inkjet 3D printing. All of these methods have their advantages and disadvantages in terms of operating cost, ease of use, printing resolution and printing speed [11]. Although the production of mould is one of the important steps in fabricating the biomodel, the biomodel produced must be hollow to allow for fluid to flow within. Therefore, the mould must be designed specifically in order to produce this hollow structure.

In this article, an aneurysm biomodel will be fabricated using 3D printing technology. Aneurysm is chosen because it possesses a complex structure that requires a specialized mould in order to produce it. The main problem to be addressed in this article is to design a suitable mould for the fabrication of the aneurysm biomodel. In addition, the final mould is expected to have reusable parts. After the mould is produced, similar fabrication strategy can then be applied to produce other types of blood vessel biomodels.

2 Methodology

This section will explain the processes involve in fabricating the mould for aneurysm biomodel. These include designing the geometry of the aneurysm blood vessel, designing suitable mould geometry, selecting suitable materials for the mould and lastly, fabricating the biomodel using available materials.

Table 1 Desired dimension of the aneurysm biomodel

Property	Desired dimension
Thickness	5 mm
Blood vessel diameter	33 mm
Aneurysm diameter	35 mm
Aspect ratio	1.51
Dome-to-neck ratio	1.28

2.1 Criteria of Aneurysm Biomodel Geometry

The aneurysm biomodel to be fabricated here is based on an idealized geometry. This is because it is easier to produce compared to fabricating a patient-specific geometry. However, fabricating a patient-specific blood vessel geometry will be subjected as a future work.

The biomodel geometry to be fabricated is based on a saccular-type aneurysm, which has dimensions (see Table 1) according to the criteria proposed by [12, 13]. Generally, the blood vessel diameter chosen is bigger than 30 mm based on an abdominal aorta, and the thickness is 5 mm. The aspect ratio (i.e. the aneurysm height over then neck ratio) should be lower than 1.6. Lastly, the dome-to-neck ratio should be less than 2.

Figure 1 shows the aneurysm diameter of the biomodel, while Fig. 2 shows the completed aneurysm biomodel CAD geometry drawn using SolidWorks. This geometry will be used to design the shape of the mould in the following section.

2.2 Mould Design

The mould must be able to produce a hollow biomodel that will allow fluid to flow within the biomodel once it is fabricated. The lost core moulding process is implemented to make a cavity inside the biomodel. Therefore, there are 3 parts of the mould, namely: (1) the core, (2) the cope and (3) the drag. Figure 3 shows the initial design of the mould. Here, the transfer tube is a space to allow the biomodel material to be injected into the mould. Meanwhile, the vent is used to let any air bubble formed during the material injection out from the mould.

However, the initial design has several limitations. Firstly, the addition of hinges creates some gaps between the drag and cope that make them movable and make the removal of biomodel from the mould difficult. Further, the addition of bolt and nut on the mould will require additional tools to tighten the drag and cope, which can be time-consuming.

In order to overcome the limitations of the initial design, new improved design has been proposed. Figure 4 shows the new mould design. In this design, a notch is added at the sides of the cope that ensures the drag to not move and reduces their

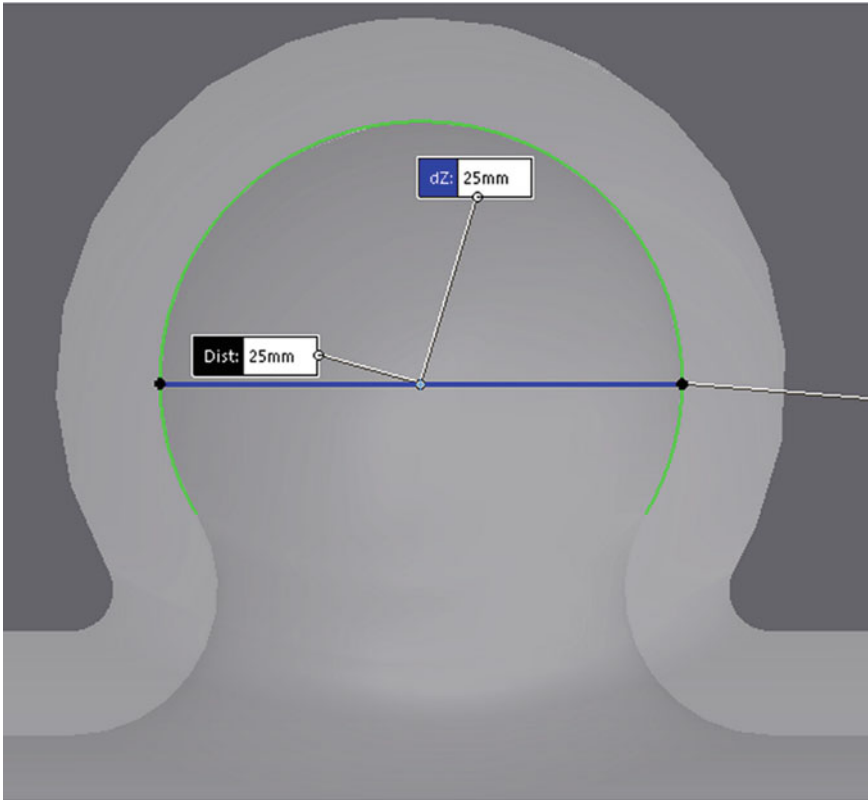


Fig. 1 Diameter of aneurysm biomodel

gaps. The assembled mould can be clamped by G-clamp and removes the bolt and nut from the design. Lastly, the core is redesigned such that a flat part is added at both ends to ensure the core fits correctly within the mould and makes it unmovable.

2.3 Mould Material and Fabrication

The cope and drag are fabricated using poly-lactic acid (PLA) 3D printing filament because of it is high strength and produces sturdy object. Meanwhile, the core uses poly-vinyl alcohol (PVA) 3D printing filament. This is because the core must be removed later. PVA can dissolve in water at room temperature, thus enable it to be removed later [13].

The mould will be fabricated using Ultimaker 2+ . The printer is set to print the core and the mould according to the material properties of the filaments. In order

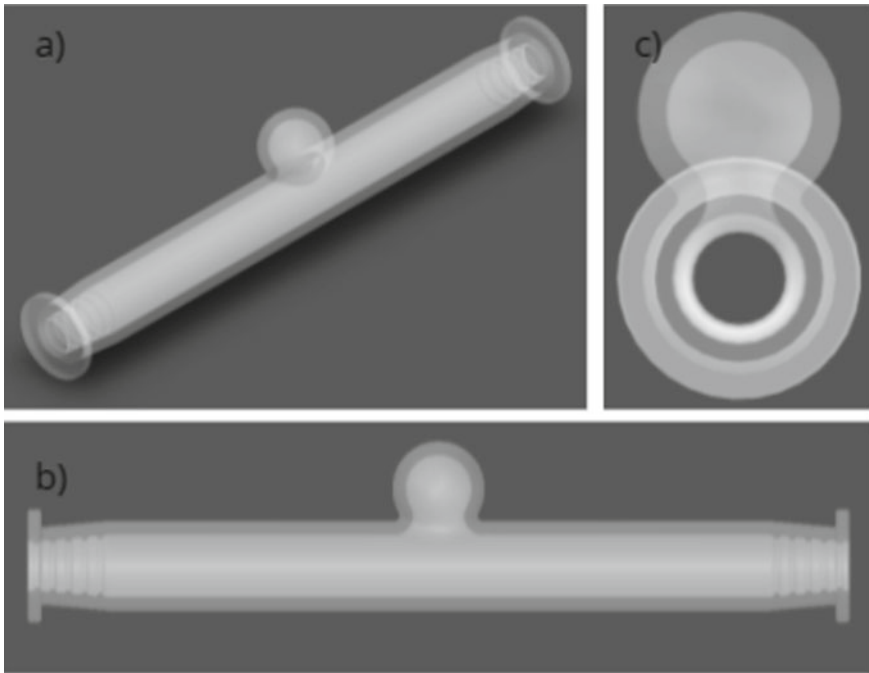


Fig. 2 Completed aneurysm biomodel CAD drawing in **a** isometric view, **b** left view and **c** front view

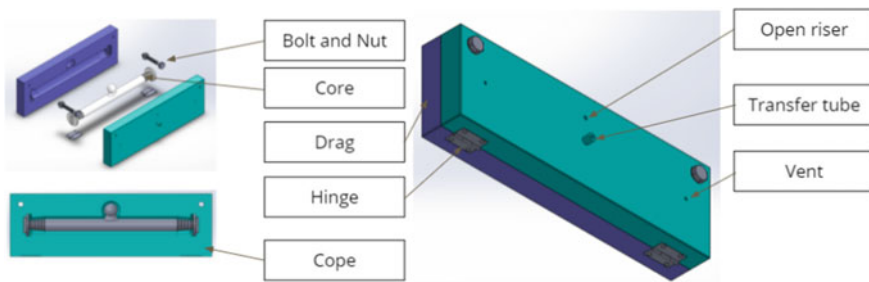


Fig. 3 Initial mould design in exploded and assembled forms

to make the printing process quicker, the core is printed in 2 parts and then stuck together using a glue. Figures 5 and 6 show the final fabricated core, cope and drag.

Then, after the parts are fabricated, post-processing steps are performed to produce a better surface finishing. Sand paper is used to remove burrs on the cope and drag. This is also to ensure the inner parts of the cope and drag are smooth so that the biomodel created later is also having a smooth outer surface.

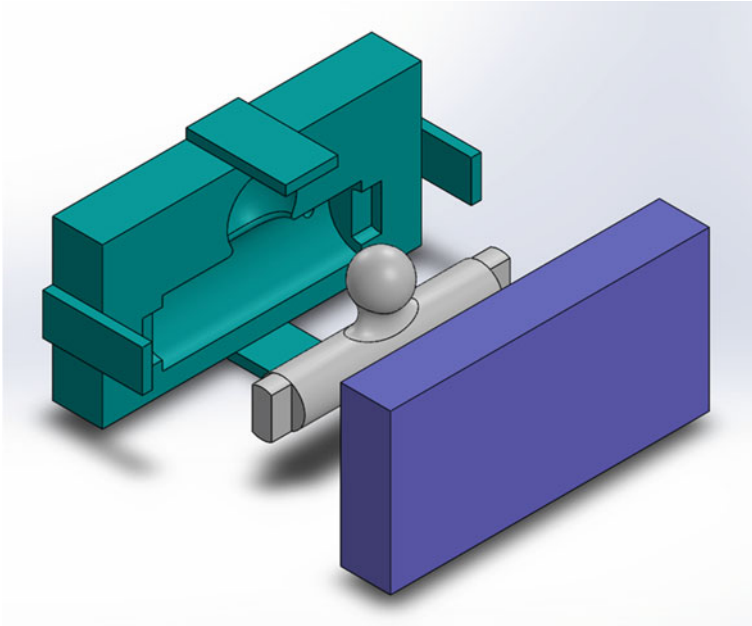


Fig. 4 New mould design considering the limitations on the initial design

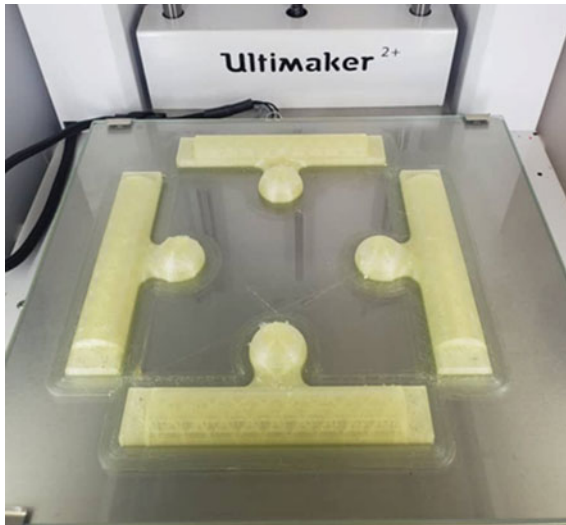


Fig. 5 Two fabricated cores in half and later are glued together

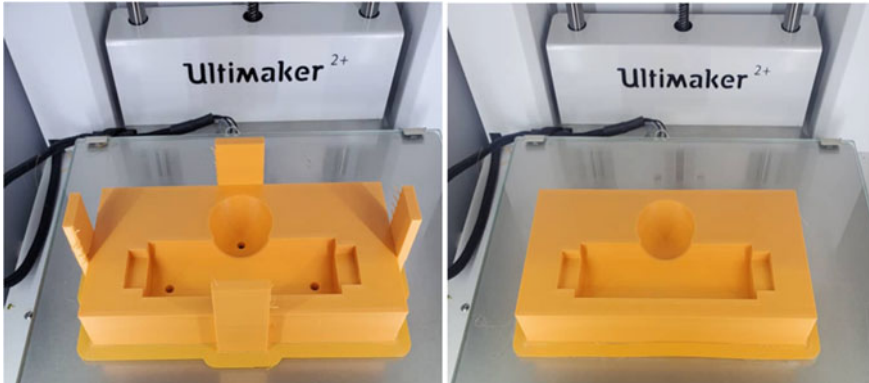


Fig. 6 Fabricated cope (left) and drag (right)

2.4 Biomodel Material and Fabrication

The biomodel must be transparent to allow for fluid flow visualization in a mock-circulation experiment and flexible so that it mimics the real blood vessel compliance. The best material for the biomodel is PVA-H [3] and silicon [4]. However, due to some limitations involving proper equipment, agar mixture is used as an alternative to evaluate the effectiveness of the mould created. Agar has been used in biomedical engineering applications such as in producing a nanocomposite based for cell culture [14].

60 g of agar powder is mixed with 150 ml of distilled water and heated until all agar is diluted. Then, before the mixture can be poured into the mould, the cope and drag with the core in between are clamped tightly. The mixture is then poured into the mould using a standard syringe. The mould is placed in a freezer at 0°C for 30 min. Different agar and water combinations produce biomodel with different mechanical properties. In this article, different water volume is used to ensure different agar-water ratio, and the mechanical properties are observed.

After the freezing process, the agar is removed from the mould and soaked into distilled water in order to remove the core. This process takes a few hours until the core can be completely removed. Figure 7 shows the final agar with the core inside. Here, the core is not removed in order to maintain the structure of the biomodel as it is easily collapsed.

3 Results

This section discusses the final mould and the aneurysm biomodel. The two designs are evaluated in terms of the dimension accuracy and product flaws.

Fig. 7 Fabricated biomodel with PVA core



3.1 Mould and Biomodel Shrinkage

The dimension of the aneurysm biomodel is measured and compared with the desired. Table 2 shows the dimension percentage difference between the actual dimension and the desired dimension. Here, the percentage differences are negative indicating shrinkage in the mould and the biomodel. The shrinkage in the biomodel is less than 5%.

The PLA used for creating the mould has a certain degree of shrinkage. The shrinkage occurs when the temperature changes rapidly during the printing and

Table 2 Comparison between desired and actual dimensions of the aneurysm biomodel

Property	Actual dimension	Percentage difference (%)
Thickness	4.8 mm	-4.00
Blood vessel diameter	31.6 mm	-4.24
Aneurysm diameter	33.5 mm	-4.29
Aspect ratio	1.48	-1.99
Dome-to-neck ratio	1.23	-3.91

Table 3 Agar–water ratio and resulting properties

Agar–water ratio	Mechanical properties
60 g agar powder; 150 ml water	Gelatine state achieved, but the biomodel teared easily
60 g agar powder, 200 ml water	Gelatine state achieved, but the biomodel is less flexible
60 g agar powder; 250 ml water	Gelatine state achieved, and the biomodel is durable with slight flexibility

cooling processes. In addition, it occurs as a result of uneven cooling post-printing. The PLA has a shrinkage rate in between 2 to 2.5%.

3.2 Biomodel Mechanical Properties

Different agar–water mixture results in different mechanical properties of the aneurysm biomodel. Table 3 shows the agar–water ratio and the mechanical properties observed. The water volume is varied from 150 to 250 ml while the agar is maintained at 60 g. All agar–water mixture achieved gelatine state but has different flexibility. The desired biomodel must be flexible and durable so that it can withstand the fluid pressure from the heart-mimicking pump. In addition, the agar biomodel fabricated is easily collapsed in atmospheric pressure, which shows that the agar biomodel is very flexible.

We also tested to produce the biomodel using PVA-H. However, the resulting biomodel is unstable and easily melted upon putting in a room temperature. Based on the previous work using PVA-H [3], successful production of biomodel using PVA-H requires efficient drying and reformation processes. In addition, the temperature used during the reformation process must be within the range found within the blood vessel in order to produce a suitable biomodel mimicking the mechanical properties of the actual blood vessel [3].

4 Discussion

Fabricating an aneurysm biomodel requires the development of a suitable mould. In this study, a lost core mould was chosen to fabricate a saccular-type aneurysm biomodel. The lost core mould is used to fabricate a hollow biomodel so that fluid flow within the biomodel can be visualized. Once the biomodel mixture has been removed from the mould, it is soaked in water to dissolve the internal core. However, the soaking process takes long time for the core to completely dissolve. Other solvent such as acetone can be used to dissolve the core [4].

The mould created using PLA and 3D printer has staggered internal surface, which requires post-processing to smoothen it. This can be done using sand paper. However,

the smoothing process might change the final dimension of the mould, subsequently affecting the final biomodel dimension. In addition, the smoothness of the mould internal surface may affect the dry-freezing process of creating the biomodel [15]. Other materials such as metals can be replaced PLA for making the mould, but the fabrication process will require different technique such as SLA and SLS. The drag and cope parts are reusable, meanwhile the core must be refabricated in order to produce additional biomodel. However, the mould design is only specific for the aneurysm biomodel used in this study. Improving the reusability of the mould can ensure the sustainability of the materials.

The biomodel must be transparent and flexible. In this study, the biomodel is created by using agar-water mixture. The final biomodel is transparent, but is less durable and easily teared. Previous works use PVA-H [3] and silicon [4] as the material for fabricating the biomodel. In order to perform fluid flow experiment in this biomodel, it must be fixed on a test rig using a specialized clamp [15]. If the biomodel is not durable enough, it can easily tear during the experiment. Adjustment on the biomodel geometry, especially at both ends is needed to ensure the biomodel, can be fixed and not teared during the experiment.

5 Conclusion

The aneurysm biomodel can be created using the lost core method. The core of the mould must be dissolved in order to obtain a hollow biomodel. The mould is reusable with the exception that the core must be fabricated every time the biomodel needs to be fabricated. The biomodel created is not suitable for the fluid flow experiment because it can easily tear. Improvement of the biomodel materials is needed to ensure the biomodel is durable, flexible, as well as transparent to allow for the visualization during the fluid flow experiment.

Acknowledgements This research is funded by the Universiti Malaysia Pahang Prototype Development Grant (PDU213212).

References

1. Kono K, Shintani A, Okada H, Terada T (2013) Preoperative simulations of endovascular treatment for a cerebral aneurysm using a patient-specific vascular silicone model—technical note. *Neurol Med Chir* 53(5):347–351
2. Paramasivam S, Baltasvias G, Psatha E, Matis G, Valavanis A (2014) Silicone models as basic training and research aid in endovascular neurointervention—a single-center experience and review of the literature. *Neurosurg Rev* 37(2):331–337
3. Shimizu Y, Putra NK, Ohta M (2018) Reproduction method for dried biomodels composed of poly (vinyl alcohol) hydrogels. *Sci Rep* 8(1):1–9
4. Kaneko N, Mashiko T, Ohnishi T, Ohta M, Namba K, Watanabe E, Kawai K (2016) Manufacture of patient-specific vascular replicas for endovascular simulation using fast, low-cost method. *Sci Rep* 6(1):1–7
5. Jewkes R, Burton HE, Espino DM (2018) Towards additive manufacture of functional, spline-based morphometric models of healthy and diseased coronary arteries: in vitro proof-of-concept using a porcine template. *J Funct Biomater* 9(1):15
6. Costa PF, Albers HJ, Linssen JE, Middelkamp HH, Van Der Hout L, Passier R, Van Der Meer AD (2017) Mimicking arterial thrombosis in a 3D-printed microfluidic in vitro vascular model based on computed tomography angiography data. *Lab Chip* 17(16):2785–2792
7. Doutel E, Carneiro J, Oliveira MSN, Campos JBLM, Miranda JM (2015) Fabrication of 3d milli-scale channels for hemodynamic studies. *J Mech Med Biol* 15(01):1550004
8. Brunette J, Mongrain R, Tardif JC (2004) A realistic coronary artery phantom for particle image velocimetry: featuring injection-molded inclusions and multiple layers. *J Vis* 7:241–248
9. Bonfanti M, Franzetti G, Homer-Vanniasinkam S, Díaz-Zuccarini V, Balabani S (2020) A combined in vivo, in vitro, in silico approach for patient-specific haemodynamic studies of aortic dissection. *Ann Biomed Eng* 48(12):2950–2964
10. Kalaskar DM (2017) 3D Printing in medicine. Elsevier, First
11. Carvalho V, Gonçalves I, Lage T, Rodrigues RO, Minas G, Teixeira SF, Lima RA (2021) 3D printing techniques and their applications to organ-on-a-chip platforms: a systematic review. *Sensors* 21(9):3304
12. Haccin-Bey L, Origitano TC, Biller J (2009) Subarachnoid hemorrhage in young adults. In: Philadelphia PA (ed) *Stroke in children and young adults*, 2nd edn. Butterworth-Heinemann, pp 289–314
13. Mix DS, Yang L, Johnson CC, Couper N, Zarras B, Arabadjis I, Richards MS (2017) Detecting regional stiffness changes in aortic aneurysmal geometries using pressure-normalized strain. *Ultrasound Med Biol* 43(10):2372–2394
14. Souza RM, Santos TQ, Oliveira DP, Alvarenga AV, Costa-Felix RPB (2016) Standard operating procedure to prepare agar phantoms. *J Phys: Conf Ser* 733(1):012044. IOP Publishing
15. Hubakri MF, Zamri MAS, Hamzah MNA, Roslan RA, Wan Ab Naim WN, Mohamed Mokhtarudin MJ (2022) Design and development of a flexible test rig for biomedical engineering PIV experiment. In: *Enabling industry 4.0 through advances in manufacturing and materials*. Lect Notes Mech Eng. Springer, Singapore. https://doi.org/10.1007/978-981-19-2890-1_10

Feature Selection of Medical Dataset Using African Vultures Optimization Algorithm



Wy-Liang Cheng, Koon Meng Ang, Sew Sun Tiang, Kah Yung Yap, Li Pan, Chin Hong Wong, Mahmud Iwan Solihin, and Wei Hong Lim

Abstract Feature selection is one of the popular techniques used to reduce the number of features by eliminating noisy, unreliable, and unnecessary data without affecting the classification accuracy. Metaheuristic algorithms were widely incorporated by researchers to search for the best possible features in simplifying and enhancing dataset feature. This is because the traditional optimization techniques have drawback of suffering from entrapment into local optima when handling a dataset with large number of features. In this study, the capability of African vultures optimization algorithm (AVOA) in conducting feature selection on medical datasets while preserving classification accuracy is investigated. Eight medical datasets retrieved from UCI machine learning repository are used to evaluation performance of AVOA in feature selection and compare with other algorithms known as opposition-based differential evolution algorithms (CO-DE), particle swarm optimization (PSO), hybrid canonical differential evolutionary particle swarm optimization (hC-DEEPSO), and multi-verse optimizer (MVO). Comparative study reports that AVOA produces the best mean accuracy (i.e., 82.9%) in six out of eight medical datasets and lowest number of features (i.e., around 24 features) in four out of eight medical datasets. AVOA can outperform other competitive algorithm in the selected medical dataset as it has most of the mean accuracy and lowest number of features.

Keywords African Vultures Optimization Algorithm (AVOA) · Feature selection · Medical classification · Optimization

W.-L. Cheng · K. M. Ang · S. S. Tiang · L. Pan · M. I. Solihin · W. H. Lim (✉)
Faculty of Engineering, Technology and Built Environment, UCSI University, 56000 Kuala Lumpur, Malaysia
e-mail: limwh@ucsiuniversity.edu.my

K. Y. Yap
School of Energy and Chemical Engineering, Xiamen University Malaysia, 43900 Sepang, Malaysia

C. H. Wong
Maynooth International Engineering College, Fuzhou University, Fuzhou 350108, China

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
M. A. Abdullah et al. (eds.), *Advances in Intelligent Manufacturing and Mechatronics*,
Lecture Notes in Electrical Engineering 988,
https://doi.org/10.1007/978-981-19-8703-8_15

1 Introduction

Feature selection is a crucial technique used to retrieve important information from raw data of a complex problem. The irrelevant and unwanted features in a dataset can cause negative impacts on the efficiency and computation time of a system. For instance, real-world medical dataset with large number of irrelevant information that increase the dimensionality of the problems, leading to the loss of cost, speed and accuracy of the learning method [1]. The diagnosis of disease requires high sensitivity and accuracy to avoid bad consequences caused by misdiagnosis [2]. Feature selection technique is a potential solution to be implemented into medical diagnosis devices to examine the symptoms and characteristics of infections with higher accuracy and efficiency [3]. However, the optimal selection of feature/attribute subset from large data possess difficulties that require a better solution for feature selection to interpret the raw data into a better construed data [4].

Feature selection is a technique implemented to decrease the size of input data when training a predictive model [5]. This technique has significant influence in performing machine learning tasks such as object detection [6, 7] and fault detection [8, 9]. Existing feature selection approaches can be classified as filter, embedded, and wrapper approaches [10]. For filter approach [11], feature subsets are assessed based on predefined metrics or given information content. For embedded approach [11], information regarding the specific structure of the classification algorithm used by a certain learning algorithm is integrated. For wrapper approach [10], the predictive accuracy is considered as the most valuable and decisive factor. The characteristic of wrapper approach being dependence on certain classifier causes the loss of generality and bias. Although the embedded approach requires lesser computational power than wrapper approach, the selected feature subset is greatly affected by the learning algorithms. Additionally, method with optimal classification performance and dimensional space is worth to be studied [12]. Wrapper methods are reported to have better classification accuracy in contrast to filter methods [13].

Metaheuristic search algorithm (MSA) emerges as a promising method to perform feature selection. MSA can be categorized based on the inspiration of search behaviors [14], known as (i) swarm intelligence (SI) algorithms that mimic group behaviors of animals, (ii) evolutionary algorithms inspired by Darwin's theory of evolution, (iii) human-characteristic algorithm that mimics the characteristic of human in learning, observing, and socializing, (iv) physic-based algorithms motivated by scientific principle. Due to its promising global search capability, different MSAs were proposed to solve various types of optimization problems as those reported in [15–22]. However, majority of MSAs is affected by No-Free-Lunch (NFL) theory which describes that no optimization technique can solve all optimization problem [23]. African vultures optimization algorithm (AVOA) [24] is one of the swarm intelligence algorithms that mimics African vultures scavenging and navigation behavior. The original literature of AVOA [24] reported that it has good results when solving engineering design problems and single-objective optimization problems, but there is no result regarding

feature selection application. In this paper, AVOA is incorporated as a feature selection algorithm to remove irrelevant features from original datasets to enhance the classification accuracy. The performance of AVOA is investigated, and comparison is conducted with four different MSAs in solving eight datasets retrieved from UCI machine learning repository [25].

The following sections of this paper are organized as follows. The inspiration of AVOA and feature selection problems is described in Sect. 2. The overall framework of AVOA in feature selection is described in Sect. 3. The simulation settings and comparative studies are reported in Sect. 4. The conclusion and future recommendations are summarized in Sect. 5.

2 Background

2.1 Inspiration of AVOA

Vultures are a type of hunting bird that can be found in Africa, America and Europe. The characteristic of vultures is bald without regular feathers which is to avoid contamination when consuming carcasses. However, recent research reported that bare skin also helps to regulate body heat. This is the reason for the vultures to dip their head into their body during the cold season. These birds do not create nests unlike normal birds and rarely attack animals except when their prey is injured or has health issues. They are also beneficial to avoid stinging and infecting carcasses, especially in the tropics which makes them to be an important factor in the ecosystem.

The population of African vulture species has been reduced over the years, especially in African countries as they are essential to handle animal carcasses in the wild. There are different species of vultures in Africa, but their searching food pattern and lifestyle are similar [26]. The type of vulture can be separate into three classes [27]. The first class is vulture that is physically stronger than regular vultures such as lappet-faced vulture. In this class, these vultures have a higher chance of acquiring food sources when compare with other classes. The second classes represent vultures that are physically weaker than the first classes such as white-backed vulture gyps Africans [27]. For the third class, the vulture is weaker when compare with first and second class such as hooded vulture *Necrosyrtes monachus*. The similarity of these classes can be observed when traveling a long distance for food sources [28]. During their travel, vultures tend to have a rotational flight during their flight form. While looking for food, vultures travel to find other species of vultures that have obtain their food sources and might cause other species of vulture to that direction which result the vulture to have dispute with other vulture to obtain the food source [29]. Weaker vultures will surround healthy vultures and obtain food by tiring the stronger vultures, and some starving vulture will be more aggressive [30]. The characteristic of vulture when searching for food sources is the inspiration for the algorithm.

2.2 Feature Selection Problem

Feature selection is designed as a bi-objective optimization problem, where the classification accuracy and number of selected features to be maximized or minimized, respectively. The objective of feature selection is achieved by formulating the fitness function as follow:

$$f(X) = \gamma\rho + \tau \frac{|F_{sub}|}{|F_{ori}|} \quad (1)$$

where $\gamma \in [0, 1]$ and $\rho \in [1 - \gamma]$ define the parameters of the weightage of classification quality and subset length, respectively, the variable τ represents the classification error rate, $|F_{sub}|$ and $|F_{ori}|$ describe the selected feature subset and overall features in original datasets, respectively.

3 Feature Selection Using African Vultures Optimization Algorithm (AVOA)

Based on the biological life of vultures, the algorithm is designed with 4 different stages. Specifically, the best vulture of the swarm is identified in stage 1; the satiation rate of vultures is calculated in stage 2; explorative and exploitative search strategies are promoted in stages 3 and 4, respectively.

During the stage 1 of the algorithm, the initial population are stochastically generated with total N vultures and evaluated with objective function. The best solution in the population is identified as the best vulture of the first swarm, and the second-best solution is identified as the second-best vulture in the second swarm. Based on the biological life of vultures, vultures tend to search for food sources and go further if they have higher stamina. However, the vulture could be aggressive in searching for food with stronger vultures when hungry and lack stamina to search further. Stage 2 of the algorithm emulates this behavior of vultures by formulating the satiation rate; a mathematical equation is formed to change from the exploration stage to the exploitation stage that mimics the rate at which the vultures are hungry or satisfied. The satiation rate tends to be decreasing trend as formulated as follow:

$$h = (2 \times r_1 + 1) \times \sigma \times \left(1 - \frac{\mu}{\mu_{max}}\right) + g \quad (2)$$

where r_1 indicates the random number between 0 and 1, σ refers to a random number between -1 and 1, variable μ and μ_{max} define as fitness evaluation number and maximum fitness evaluation number, respectively, parameter g that used to enhance the performance in dealing with complex optimization problems by increasing the probability of escaping from local optima is formulated as follows:

$$g = r_2 \times \left(\sin^p \left(\frac{\pi}{2} \times \frac{\mu}{\mu_{max}} \right) + \cos \left(\frac{\pi}{2} \times \frac{\mu}{\mu_{max}} \right) - 1 \right) \quad (3)$$

where r_2 defines random number between -2 and 2 , variable p is a parameter with a predefined number representing the optimization operation disrupts the exploration and operation states, μ and μ_{max} indicate the fitness evaluation number and maximum fitness evaluation number, respectively. At the end of stage 2, roulette wheel selection is performed to randomly select a random vulture X_r from either best vulture or second-best vulture.

If $|h|$ calculated in stage 2 is larger than or equal to 1, the searching process proceeds to stage 3 of the algorithm, where the explorative search is promoted. In stage 3, vultures analyze different random regions with two different strategies based on a probability P_1 with a predefined value between 0 and 1. The search strategy practiced by each n th vulture in updating its position X_n is randomly selection based on probability P_1 as follows:

$$X_n^{new} = \begin{cases} X_r - |(2 \times r_3) \times X_r - X_n| \times h, & \text{if } P_1 \geq r \text{ and} \\ X_r - h + r_3 \times ((X^{UB} - X^{LB}) \times r_4 + X^{LB}), & \text{otherwise} \end{cases} \quad (4)$$

where h is the satiation rate of vulture calculated using Eq. (2), X^{LB} and X^{UB} represent the lower and upper boundary, X_r is the randomly selected vulture using roulette wheel, variables r_3 , r_4 , and $rand$ represent random numbers between 0 and 1.

If $|h|$ calculated in stage 2 is smaller than 1, the searching process proceeds to the stage 4 of the algorithm, where exploitative search strategies are adopted. Stage 4 consists of two subphases with 2 different search strategies in each subphase. If $|h|$ is larger than or equal to 0.5, the searching process proceeds to the first subphase. In the first subphase, the search strategy adopted by each n th vulture is randomly selected based on a probability P_2 with a predefined value between 0 and 1. Given the value of P_2 , the position X_n of n th vulture is calculated as follows:

$$X_n^{new} = \begin{cases} (A + B)/2, & P_2 > r \text{ and} \\ X_r - |X_r - X_n| \times h \times L^f, & \text{otherwise} \end{cases} \quad (5)$$

where A and B emulate the movements of the best vulture X_1^{best} and second-best vulture X_2^{best} in competing for one food source that can be formulated as follows, respectively:

$$A = X_1^{best} - \frac{X_1^{best} \times X_n}{X_1^{best} \times (X_n)^2} \times h \quad (6)$$

$$B = X_2^{best} - \frac{X_2^{best} \times X_n}{X_2^{best} \times (X_n)^2} \times h \quad (7)$$

Define L^f as the Levy flight [31] patterns used to enhance the search effectiveness of algorithm, and it can be calculated as follow:

$$L^f = 0.01 \times \left(u \times \left(\frac{\Gamma(1 + \alpha) \times \sin(\frac{\pi\alpha}{2})}{\Gamma(1 + 2 \times \alpha) \times \alpha \times 2 \times (\frac{\alpha-1}{2})} \right)^{\frac{1}{\alpha}} \right) / |v|^{\frac{1}{\alpha}} \quad (8)$$

where u and v are random numbers between 0 and 1, α is a predefined and fixed number equals to 1.5, $\Gamma(\cdot)$ refers to the gamma function.

If $|h|$ is lesser than 0.5, the searching process proceeds to the second subphase. In the second subphase, the search strategy adopted by each n th vulture is randomly selected based on a probability P_3 with a predefined value between 0 and 1. Given the value of P_3 , the position X_n of n th vulture is calculated as follows:

$$X_n^{new} = \begin{cases} |(2 \times r_5) \times X_r - X_n| \times (h + r_6) - X_r - X_n, & P_3 > r \text{ and} \\ X_n - (S_1 - S_2), & \text{otherwise} \end{cases} \quad (9)$$

where r_5 and r_6 represent random numbers between 0 and 1, S_1 and S_2 are calculated as follows, respectively, with r_7 and r_8 represent random numbers between 0 and 1:

$$S_1 = X_n \times \left(\frac{r_7 \times X_n}{2\pi} \right) \times \cos(X_n) \quad (10)$$

$$S_2 = X_n \times \left(\frac{r_8 \times X_n}{2\pi} \right) \times \sin(X_n) \quad (11)$$

The searching process of the AVOV for identifying the best combinations of features is repeated for μ_{max} fitness evaluations and terminated when $\mu > \mu_{max}$, where μ is the fitness evaluation counter and μ_{max} is the maximum fitness evaluation counter. The best vulture X_1^{best} obtained at the end of the searching process is returned, and it is decoded to obtain the optimal feature subsets. The overall framework of using AVOA for feature selection is illustrated in Fig. 1.

4 Result Analysis

4.1 Simulation Settings

The capability of AVOA in solving feature selection problems is investigated using eight medical datasets obtained from UCI machine learning repository [25] as presented in Table 1. Comparative study is conducted between AVOA and other four MSAs, known as opposition-based differential evolution algorithms (CO-DE)

Algorithm: AVOA-based Feature Selection Algorithm	
Inputs: $N, D, \mu, \mu_{max}, X^{UB}, X^{LB}, P_1, P_2, P_3$	
01:	Parameter initialization and load the selected medical dataset;
02:	Initialize the AVOA population, i.e., position of vultures;
03:	Evaluate fitness of each vultures for feature selection task using Eq. (1)
04:	Select the first X_1^{best} and second-best vultures X_2^{best} ;
05:	while $\mu \leq \mu_{max}$ do
06:	for each n -th do
07:	Calculate h using Eq. (2);
08:	Perform roulette wheel on X_1^{best} and X_2^{best} to select vultures X_r ;
09:	if $ h \geq 1$ then
10:	Calculate X_n^{new} using Eq. (4);
11:	else if $ h < 1$ then
12:	if $ h \geq 0.5$ then
13:	Calculate X_n^{new} using Eq. (5);
14:	else if $ F < 0.5$ then
15:	Calculate X_n^{new} using Eq. (9);
16:	end if
17:	end if
18:	Boundary check of X_n^{new} based on X^{UB} and X^{LB} ;
19:	Evaluate fitness of X_n^{new} for feature selection task using Eq. (1);
20:	Update the X_1^{best} and X_2^{best} ;
21:	$\mu \leftarrow \mu + 1$;
22:	end for
23:	end while
Output: X_1^{best}	

Fig. 1 Overall framework of AVOA for feature selection problem

[32], particle swarm optimization (PSO) [33], hybrid canonical differential evolutionary particle swarm optimization (hC-DEEPSO) [34], and multi-verse optimizer (MVO) [35], in terms of mean accuracy and the number of feature selected. To ensure fair comparisons, the optimal parameters of these compared algorithms are set based on the recommendations of their original literatures. During the simulation process, the dimension size D is set to be equal as the number of features of each dataset and population number N is set as 20. The simulation of each dataset is conducted for 30 independent runs with maximum fitness evaluation number $\mu_{max} = 2000$ for each run.

Table 1 List of datasets

No	Dataset	Instances	Number of Features	Classes
1	Lymphography	148	18	3
2	Statlog (Heart)	270	13	2
3	Diabetes	768	8	2
4	Ovarian	216	4000	2
5	Echocardiogram	61	8	2
6	Liver Disorders	345	6	2
7	Parkinsons	195	22	2
8	Arrhythmia	452	279	13

4.2 Comparisons Between Selected Algorithms

The performance of each compared algorithm, in terms of mean accuracy Acc_{mean} and the average number of selected features $N_{feature}$, in solving each dataset is reported in Tables 2 and 3, respectively. The mean accuracy is based on Eq. (12).

Table 2 Mean accuracy Acc_{mean} produced by each compared algorithm

Dataset	AVOA	CO-DE	PSO	hC-DEEPSO	MVO
1	5.897E-01	<u>5.862E-01</u>	5.115E-01	5.854E-01	5.356E-01
2	9.037E-01	8.673E-01	8.222E-01	8.652E-01	<u>8.938E-01</u>
3	7.375E-01	7.880E-01	<u>7.660E-01</u>	6.959E-01	7.540E-01
4	1.000E + 00	9.946E-01	9.899E-01	<u>9.990E-01</u>	9.783E-01
5	9.577E-01	8.077E-01	<u>9.167E-01</u>	8.750E-01	8.551E-01
6	8.362E-01	6.812E-01	<u>7.744E-01</u>	7.408E-01	7.459E-01
7	8.991E-01	<u>8.974E-01</u>	8.957E-01	8.737E-01	8.692E-01
8	<u>7.104E-01</u>	6.570E-01	7.633E-01	6.998E-01	6.648E-01

Table 3 Average number of selected features $N_{feature}$ produced by each compared algorithm

Dataset	AVOA	CO-DE	PSO	hC-DEEPSO	MVO
1	3.3667	6.3333	7.3667	<u>4.6667</u>	6.7667
2	4.9000	3.3333	<u>4.8333</u>	5.2000	5.0000
3	2.8333	<u>2.9333</u>	4.1000	3.0000	4.0667
4	<u>139.7667</u>	1672.8667	1735.8667	103.7667	1789.9667
5	3.8667	3.0667	2.9000	2.2333	<u>2.5667</u>
6	3.0667	2.1000	3.2000	3.0000	<u>2.1333</u>
7	1.7667	3.1667	4.1000	<u>2.0667</u>	4.5667
8	38.1333	112.3667	102.3667	<u>39.3333</u>	117.4667

$$\text{Mean accuracy} = \text{mean}(1 - \text{Error}_{\text{mean}}) \quad (12)$$

The variable $\text{Error}_{\text{mean}}$ is the final value of the classification error for every run which can be obtain when performing evaluation. The average number of selected features represents the chosen of the best possible features, and the lower number represents a good result as it requires less feature for future processing. The value with boldface indicates the best result among the compared algorithms while the italic and underlined value represents the second-best result. In Table 2, the higher value of Acc_{mean} implies that the better performance in classifying the given dataset with obtained feature subset. For Table 3, the best result is considered when the lowest number of selected features is obtained to solving the given dataset. Hence, Table 2 reports that AVOA has the best classification accuracy by producing six best Acc_{mean} values in solving datasets 1, 2, 4, 5, 6, and 7. It is followed by PSO, CO-DE, hC-DEEPSO, and MVO. Moreover, Table 3 reports that AVOA has the best performance in minimizing the number of features by producing four best N_{feature} values and one second-best N_{feature} value. It is followed by hC-DEEPSO, CO-DE, MVO, and PSO. Tables 2 and 3 report that AVOA can solve feature selection problems with better classification accuracy and smaller feature subsets.

5 Conclusion

In this paper, a new metaheuristic algorithm known as AVOA is used to perform feature selection to discard unwanted features from medical datasets without compromising the classification accuracy. Extensive simulation study has proven that AVOA can produce the best mean accuracy with a minimum number of features when dealing with majority of selected medical datasets. Particularly, AVOA produces the highest average accuracy (i.e., 82.9%) in solving all eight selected datasets and the lowest average number of features (i.e., 24) in solving all eight selected datasets. This approach improves the accuracy and efficiency of diseases diagnosis and medical treatments. As future improvement, AVOA can be incorporated with adaptive searching operators to enhance the performance in dealing with feature selection problems with higher complexity level.

Acknowledgements This work was supported by the Ministry of Higher Education Malaysia under the Fundamental Research Schemes with project codes of FRGS/1/2019/TK04/UCSI/02/1 and FRGS/1/2020/TK0/UCSI/02/4. This work is also supported by the UCSI University Research Excellence & Innovation Grant (REIG) with project code of REIG-FETBE-2022/038.

References

1. Zamani H, Nadimi-Shahraki MH (2016) Feature Selection Based on Whale Optimization Algorithm for Diseases Diagnosis, pp 1243–1247
2. Szumski NR, Cheng EM (2009) Optimizing algorithms to identify Parkinson's disease cases within an administrative database. *Movement Disorders* 24(1):51–56. Accessed 15 Jan 2009
3. Dash M, Liu H (1997) Feature selection for classification. *Intelligent Data Analysis* 1(1):131–156. <https://www.sciencedirect.com/science/article/pii/S1088467X97000085>. Accessed 1 Jan 1997
4. Neggaz N, Houssein EH, Hussain K (2020) An efficient henry gas solubility optimization for feature selection. *Expert Syst Appl* 152:113364. Accessed 18 Aug 2020
5. Ang KM et al (2022) New Hybridization algorithm of differential evolution and particle swarm optimization for efficient feature selection 27:5. <https://doi.org/10.5954/ICAROB.2022.OS22-1>
6. Low JW, Tiang SS, Lim WH, Chong YH, and Voon YN (2022) Tomato leaf health monitoring system with SSD and MobileNet. In: Zain ZMd, Sulaiman MH, Mohamed AI, Bakar MS, Ramli MS, (eds) *Proceedings of the 6th international conference on electrical, control and computer engineering*, Singapore. Springer Singapore, pp 795–804
7. Voon YN, Ang KM, Chong YH, Lim WH, Tiang SS (2022) Computer-vision-based integrated circuit recognition using deep learning. In: Zain ZMd, Sulaiman MH, Mohamed AI, Bakar MS, Ramli MS (eds) *Proceedings of the 6th international conference on electrical, control and computer engineering*, Singapore. Springer Singapore, pp 913–925
8. Alrifayy M et al (2022) Hybrid deep learning model for fault detection and classification of grid-connected photovoltaic system. *IEEE Access* 10:13852–13869
9. Alrifayy M, Lim WH, Ang CK (2021) A novel deep learning framework based RNN-SAE for fault detection of electrical gas generator. *IEEE Access* 9:21433–21442
10. Stańczyk U (2015) Feature evaluation by filter, wrapper, and embedded approaches. In: Stańczyk U, Jain LC (eds) *Feature selection for data and pattern recognition*. Springer, Berlin, Heidelberg, pp 29–44
11. Hancer E, Xue B, Zhang M (2018) Differential evolution for filter feature selection based on information theory and feature ranking. *Knowl-Based Syst* 140:103–119. <https://www.sciencedirect.com/science/article/pii/S0950705117304987>
12. Huan L, Lei Y (2005) Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans Knowl Data Eng* 17(4):491–502
13. Maldonado S, Weber R (2009) A wrapper method for feature selection using support vector machines. *Inf Sci* 179(13):2208–2217. <https://www.sciencedirect.com/science/article/pii/S00205509000917>
14. Mohamed AW, Hadi AA, Mohamed AK (2020) Gaining-sharing knowledge based algorithm for solving optimization problems: a novel nature-inspired algorithm. *Int J Mach Learn Cybern* 11(7):1501–1529. <https://doi.org/10.1007/s13042-019-01053-x>
15. Jamaludin FA, Ab-Kadir MZA, Izadi M, Azis N, Jasni J, Rahman MSA (2016) Considering the effects of a RTV coating to improve electrical insulation against lightning. In: 2016 33rd international conference on lightning protection (ICLP), 25–30 Sept 2016, pp 1–5. <https://doi.org/10.1109/ICLP.2016.7791414>
16. Jamaludin FA et al (2018) Effect of RTV coating material on electric field distribution and voltage profiles on polymer insulator under lightning impulse. In: 2018 34th international conference on lightning protection (ICLP), 2–7 Sept. 2018, pp 1–6. <https://doi.org/10.1109/ICLP.2018.8503296>
17. Shaari M et al (2020) Supervised evolutionary programming based technique for multi-DG installation in distribution system. *IAES Int J Artif Intell (IJ-AI)* 9:11. <https://doi.org/10.11591/ijai.v9.i1.pp11-17>
18. Solihin MI, Lim WH, Tiang SS, Ang CK Modified Particle Swarm Optimization for Robust Anti-swing Gantry Crane Controller Tuning,”. In: Zain ZMd et al (eds) *Proceedings of the*

- 11th national technical seminar on unmanned system technology 2019, Singapore. Springer Singapore, pp 1173–1192
19. Zè H, Ang CK, Lim WH, Yu LJ, Solihin MI (2020) Development of an artificial intelligent approach in adapting the characteristic of polynomial trajectory planning for robot manipulator, vol 9, pp 408–414. <https://doi.org/10.18178/ijmerr.9.3.408-414>
 20. Sharma A et al (2021) Opposition-based tunicate swarm algorithm for parameter optimization of solar cells. *IEEE Access* 9:125590–125602. <https://doi.org/10.1109/ACCESS.2021.3110849>
 21. Sharma A, Mathur S (2018) Comparative analysis of ML-PSO DOA estimation with conventional techniques in varied multipath channel environment. *Wirel Personal Commun* 100(3):803–817. <https://doi.org/10.1007/s11277-018-5350-0>
 22. Sharma A, Dasgotra A, Tiwari SK, Sharma A, Jatly V, Azzopardi B (2021) Parameter extraction of photovoltaic module using tunicate swarm algorithm. *Electronics* 10(8). <https://doi.org/10.3390/electronics10080878>
 23. Yu-Chi H, Pepyne DL (2001) Simple explanation of the no free lunch theorem of optimization. In: *Proceedings of the 40th IEEE conference on decision and control* (Cat. No.01CH37228), 4–7 Dec 2001, vol 5, pp 4409–4414
 24. Abdollahzadeh B, Gharehchopogh FS, Mirjalili S (2021) African vultures optimization algorithm: a new nature-inspired metaheuristic algorithm for global optimization problems. *Comput Ind Eng* 158:107408. <https://www.sciencedirect.com/science/article/pii/S0360835221003120>
 25. Dua D, Graff C (2017) UCI machine learning repository
 26. Meteyer CU, Rideout BA, Gilbert M, Shivaprasad HL, Oaks JL (2005) Pathology and proposed pathophysiology of diclofenac poisoning in free-living and experimentally exposed oriental white-backed vultures (*Gyps bengalensis*). *J Wildl Dis* 41(4):707–716
 27. Houston DC (1974) The role of griffon vultures *Gyps africanus* and *Gyps ruppellii* as scavengers. *J Zool* 172(1):35–46
 28. Attwell RIG (1963) Some observations on feeding habits, behaviour and inter-relationships of Northern Rhodesian vultures. *Ostrich* 34(4):235–247
 29. BosÈ M, Sarrazin F (2007) Competitive behaviour and feeding rate in a reintroduced population of Griffon Vultures *Gyps fulvus*. *Ibis* 149(3):490–501
 30. Anderson DJ, Horwitz RJ (1979) Competitive interactions among vultures and their avian competitors. *Ibis* 121(4):505–509
 31. Yang X-S (2010) Firefly algorithm, Levy flights and global optimization. In: *Research and development in intelligent systems XXVI*. Springer, pp 209–218
 32. Rahnamayan S, Tizhoosh HR, Salama MMA (2006) Opposition-based differential evolution algorithms. In: *2006 IEEE international conference on evolutionary computation*, 16–21 July 2006, pp 2010–2017
 33. Kennedy J, Eberhart R (1995) Particle swarm optimization. In: *Proceedings of ICNN'95 - international conference on neural networks*, 27 Nov-1 Dec 1995, vol 4, pp 1942–1948
 34. Marcelino CG et al (2018) Solving security constrained optimal power flow problems: a hybrid evolutionary approach. *Appl Intell* 48(10):3672–3690
 35. Mirjalili S, Mirjalili SM, Hatamlou A (2016) Multi-verse optimizer: a nature-inspired algorithm for global optimization. *Neural Comput Appl* 27(2):495–513

Flow Direction Algorithm for Feature Selection



Wy-Liang Cheng, Koon Meng Ang, Wei Hong Lim, Sew Sun Tiang,
Meng Choung Chiong, Chun Kit Ang, Li Pan, and Chin Hong Wong

Abstract Feature selection is a method used to decrease the number of features by removing unwanted, noisy and inconsistent data while maintaining classification accuracy. Most researchers have focused on using metaheuristic algorithms to select the best possible features to improve and simplify the dataset quality. However, the traditional optimization method tends to suffer from local optimality problems as the increasing of features in datasets. In this paper, an investigation is conducted to assess the performance of flow direction algorithm (FDA) in enhancing the classification accuracy by performing feature selection. Eight datasets obtained from UCI machine learning repository are used to perform comparative studies with existing algorithms known as differential evolution (DE), biogeography-based learning particle swarm optimization (BLPSO), henry gas solubility optimization (HGSO) and African vulture optimization algorithm (AVOA). The results reported that FDA obtains best mean accuracy in the comparative studies with the selected algorithms and the number of features selected.

Keywords Flow direction algorithm (FDA) · Feature selection · Metaheuristic search algorithms · Optimization

1 Introduction

In recent years, data-driven approaches emerged as promising methods to perform machine learning tasks. Datasets of different domains were created along with the advancement of technology to train machine learning models. However, the datasets with higher complexity level cause the traditional machine learning method to be obsolete due to dimensionality dilemma [1]. These datasets tend to have unrelated,

W.-L. Cheng · K. M. Ang · W. H. Lim (✉) · S. S. Tiang · M. C. Chiong · C. K. Ang · L. Pan
Technology and Built Environment, UCSI University, 56000 Kuala Lumpur, Malaysia
e-mail: limwh@ucsiuniversity.edu.my

C. H. Wong
Maynooth International Engineering College, Fuzhou University, Fuzhou 350108, China

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
M. A. Abdullah et al. (eds.), *Advances in Intelligent Manufacturing and Mechatronics*,
Lecture Notes in Electrical Engineering 988,
https://doi.org/10.1007/978-981-19-8703-8_16

redundant, noisy and irrelevant information that can badly impact on the approximation accuracy of the system [2]. The efficiency of an algorithm, in terms of computational cost, generalization capability and accuracy, can be improved by removing these redundant features in dataset. Nevertheless, the task of removing unnecessary features from a dataset to improve the efficiency of a machine learning method remains arduous. Moreover, the selection of features data subset from large information poses several challenges that require a better method for feature selection to greatly decrease the initial data into an improve interpreted data [3].

Feature selection is a method applied to reduce the number of input data when training a predictive model [4]. This method has been a great contribution toward different fields to solve classification problems such as fault detection [5, 6], automatic modulation recognition [7], object detection [8, 9], etc. Classification approaches can be classified as filter, embedded and wrapper approaches, based on the evaluation criteria [10]. In filter approach [11], feature subsets are evaluated based on predefined metrics or information content. In embedded approach, knowledge about the specific structure of the classification algorithm used by a certain learning algorithm is incorporated. Although embedded approach requires lower computational effort than wrapper approach, the selected feature subset is highly dependent on the learning algorithm. Moreover, a better trade-off between the lower dimensional space and higher classification performance is worth to be investigated [12]. Wrapper methods are observed to have better classification accuracy than filter methods [13].

Metaheuristic search algorithm (MSA) is one of the options to handle feature section problem. MSA can be classified based on their searching method and inspiration sources [14], such as (i) evolutionary algorithms inspired by Darwin's theory of evolution, (ii) swarm intelligence (SI) algorithms inspired by the behavior of animals in a group, (iii) physics-based algorithms that mimic scientific theory, (iv) human-characteristic algorithm inspired by human behaviors in socializing, learning and observing. Various approach of MSAs were introduced in recent years to solve different real-world applications such as smart grid energy management [15–18], neuroevolution [19], composite materials machining [20–22] and other engineering applications [23–30]. Nevertheless, the effectiveness of these algorithms can be restricted by no-free-lunch (NFL) theory [31] that states there is no optimization method will be able to solve all optimization problems. Flow direction algorithm (FDA) [32] is one of the physics-based algorithms that imitates the flow direction from the highest position to the lowest position in a drainage system. FDA was initially proposed and reported to have significant performance in solving single-objective optimization problems and engineering design problems. However, the capability of FDA in tackling features selection problems remains questionable. In this study, FDA is designed as a feature selection algorithm to remove redundant features from original datasets, to improve the classification accuracy. The performance of FDA, in terms of classification accuracy and number of selected features, is evaluated and compared with four existing MSAs in solving eight datasets retrieved from UCI machine learning repository.

The remaining sections of this paper are organized as follows. The search mechanisms of FDA and feature selection problems are described in Sect. 2. Overall FDA frameworks in feature selection are described in Sect. 3. The simulation parameters setting and comparative studies are reported in Sect. 4. The conclusion and future works are summarized in Sect. 5.

2 Background

2.1 Flow Direction Algorithm Structure

In drainage basin, the excessive rainfall is defined as rainfall flowed over the surface of ground and does not penetrate the soil. The water left over on the ground surface after precipitation and losses are known as direct runoff. Technique has been introduced to determine direct runoff, known as ϕ —index method [33].

The index ϕ is defined as the average number of water loss during rainfall with a unit of centimeter per hour, such that the excessive water will turn into runoff. The direct runoff value can be produced by the difference between the index ϕ at each time interval and the rainfall. The amount of direct runoff r_d can be calculated as follow:

$$r_d = \sum_{m=1}^M (R_m - \phi \Delta t) \quad (1)$$

where the variable Δt and M represent the time interval and the number of time steps, respectively.

The number of rainfalls is transformed into direct runoff after deducting the rainfall loss factors. The movement of the runoff toward the outlet of basin is affected by the aspect of slope and can be imitated by separating the drainage basin into different set of cells. The amount of runoff in each cell is shifted to other cells based on the slope and height of nearby cell. D8 method [34] was introduced to determine the variable of runoff direction. It stated that the flow moves to one of the eight surrounding cells [35]. This technique supposes that every cell has eight neighbors, each cell has distance and height to the neighboring cells. The difference between the height and distance of each cell with adjacent cells is determined to identify the flow direction. Then, each cell slope is computed and the flow in cell flows to the cell with highest slope. The flow direction is identified by initializing D8 algorithm for the basin. After the flow direction is specified, a variable equal to the number of cells that flow into that cell is considered. Hence, the greatest number is appointed to the outlet point of basin.

2.2 Feature Selection Problem

Feature selection is formulated as a bi-objective optimization problem, where the number of selected features and classification accuracy to be minimized and maximized, respectively. To achieve the objective, fitness function is formulated to define the quality of each candidate solution as follows:

$$f(g) = \omega\varphi + \gamma \frac{|F_s|}{|F_t|} \quad (2)$$

where $\omega \in [0, 1]$ and $\gamma \in [1 - \omega]$ represent the parameters indicating the weightage of classification quality and subset length, respectively; the variable φ is defined as classification error rate; $|F_s|$ and $|F_t|$ refer to the chosen feature subset and overall features in original datasets, respectively.

3 Flow Direction Algorithm (FDA) for Feature Selection

The search mechanism of FDA simulates the flow direction in a drainage basin identified by D8 method, after converting rainfall to runoff [32]. At the early stage of FDA algorithm, the initial population is randomly generated in the drainage basin (i.e., search space). Then, the flows (i.e., candidate solutions) flow to locations with lower positions in the drainage basin. Hence, the lowest position of location is achieved by the flows as the global optimum after several iterations.

At the beginning of the algorithm, the predefined input parameters, known as population number, number of neighbor and neighborhood radius, are set. Then, the initial position of flows (i.e., candidate solutions) is generated based on the following equation:

$$x_f(i) = b^{low} + r_1 * (b^{up} - b^{low}) \quad (3)$$

where $X_f(i)$ defines the position of the i -th flow; b^{low} and b^{up} refer to the lower and upper boundary of dimensional components, respectively; r indicates a number generated by uniform distribution between 0 and 1.

During the optimization process, it is assumed that there are β neighborhoods around each flow, where the position is defined as follows:

$$X_n(j) = X_f(i) + r_2 * \Delta \quad (4)$$

where $X_n(j)$ defines the position of each j -th neighbor; r_2 indicates a random number generated by normal distribution, a mean of zero and A standard deviation of 1; the symbol Δ represents a parameter controlling the search behavior of the algorithms. Specifically, a smaller value of Δ tends to search in a smaller range of search space,

and a greater value of Δ tends to search in a greater range. Greater value of Δ results in higher probability to achieve global optimal solution (i.e., exploration), and lower value of Δ leading to the higher precision in locating the global optimal solution (i.e., exploitation). The parameter Δ plays a crucial role in affecting the searching performance of FDA algorithm, by achieving better balancing of exploration and exploitation behaviors. In this algorithm, the value of parameter Δ is linearly decreased with the increasing of fitness evaluation number, leading to the flows flowing toward random positions for better diversity. The parameter Δ is calculated as follows:

$$\Delta = (r_3 * X_r - r_4 * X_f(i)) * ||X_{best} - X_f(i)|| * W \quad (5)$$

where r_3 and r_4 are random numbers generated by uniform distribution in range of 0 to 1; X_r refers to a random position that calculated based on Eq. (1); W indicates a nonlinear weight with random number between 0 and infinite. The first component of Eq. (5) presents that the i -th flow moves toward a random position. Meanwhile, the second component describes that the i -th flow moves toward the best solution found, implying that the Euclidean distance between the best solution and i -th solution is decreased along with the increment of iteration. The third component of Eq. (5), known as parameter W , is formulated as

$$W = \left(\left(1 - \frac{\mu}{\mu_{max}} \right)^{(2*r_5)} \right) * \left(\overline{r_6} * \frac{\mu}{\mu_{max}} \right) * \overline{r_7} \quad (6)$$

where r_6 and r_7 refer to random numbers generated by uniform distribution between zero to one; r_5 defines a random number generated by normal distribution; the variable μ and μ_{max} define the fitness evaluation counter and maximum fitness evaluation number, respectively.

Furthermore, the flow moves at a velocity of V toward the neighbor with the lowest fitness value. The velocity of the flow is related to its slope as formulated as

$$V_f = r_8 * S_0 \quad (7)$$

where S_0 defines the slope vector between the neighbor and the current position of the flow, r_8 represents a random number generated from normal distribution. The velocity of each flow is limited with lower and upper boundaries of $v^{low} = -0.1 * (b^{up} - b^{low})$ and $v^{up} = -v^{low}$, respectively. The slope vector of i -th flow is calculated as

$$S_0(i, j, d) = \frac{f(X_f(i)) - f(X_n(j))}{||X_f(i, d) - X_n(j, d)||} \quad (8)$$

where $f(X_f(i))$ and $f(X_n(j))$ define the fitness value of i -th flow and j -th neighbor; d represents the index of dimensional component. Given the velocity vector, the new position of i -th flow is calculated as

$$X_f^{new}(i) = X_f(i) + V_f * \frac{X_f(i) - X_n(j)}{||X_f(i) - X_n(j)||} \quad (9)$$

Furthermore, it is worth to simulate a condition, where the fitness value of any neighbor is higher than current flow, by randomly selecting another flow. If the fitness value of the random flow is lower than the current flow, the random flow will move toward the same direction with current flow, otherwise, the flow moves in the dominant slope direction. The flow direction is simulated as

$$X_f^{new}(i) = \begin{cases} X_f(i) + \bar{r}_9 * (X_f(k) - X_f(i)), & \text{if } f(X_f(k)) < f(X_f(i)) \\ X_f(i) + r_{10} * (X_{best} - X_f(i)), & \text{otherwise} \end{cases} \quad (10)$$

where k refers to a random integer.

The overall mechanism of FDA is described in Fig. 1. The algorithm is iterated until the termination conditions $\mu > \mu_{max}$ is achieved, where μ represents current fitness evaluation counter and μ_{max} maximum fitness evaluation number.

4 Result

4.1 Simulation Settings

The performance of FDA in solving feature selection problems is investigated using eight datasets obtained from UCI machine learning repository [36] as summarized in Table 1. FDA is compared with other MSAs known as differential evolution (DE) [37], biogeography-based learning particle swarm optimization (BLPSO) [38], henry gas solubility optimization (HGSO) [39] and African vultures optimization algorithm (AVOA) [40]. The performance comparison among the algorithms considers the mean accuracy and the number of selected features. The input parameters are predefined as, population number $N = 20$, dimension D equals to the number of features of each dataset, maximum fitness evaluations number $\mu_{max} = 2000$. The simulation of each dataset is conducted for 30 independent runs.

4.2 Performance Metrics

Hold-out method is used to distribute the datasets for simulation studies. The original datasets are separated into 80% of training datasets and 20% of the testing datasets. Mean accuracy is applied as one of the performance metrics in measuring the actual and predicted values produced by the proposed method. The mean accuracy produced by each method in solving a given dataset can be obtained as follow:

<p>Algorithm: Flow Direction Algorithm (FDA)</p> <p>Inputs: $N, D, \mu, \mu_{max}, b^{low}, b^{up}, v^{low}, v^{up}, \beta$</p> <p>01: Define fitness function and input variable;</p> <p>02: Initialize the position of flows using Eq. (3);</p> <p>03: Evaluate fitness of flows and select the best flow with lowest fitness value;</p> <p>04: while $\mu < \mu_{max}$ do</p> <p>05: for each i-th flow do</p> <p>06: Update W based on Eq (6);</p> <p>07: for each j-th neighbor do</p> <p>08: Calculate Δ using Eq. (5);</p> <p>09: Calculate $X_n(j)$ using Eq. (4);</p> <p>10: Limit position with lower boundary (b^{low}) and upper boundary (b^{up});</p> <p>11: Evaluate $f(X_n(j))$ using Eq. (2);</p> <p>12: $\mu \leftarrow \mu + 1$;</p> <p>13: end</p> <p>14: Sort X_n from best to worst based on fitness values;</p> <p>15: if $f(X_n(1)) < f(X_f(i))$ then</p> <p>16: Calculate $S_0(i, j, d)$ using Eq. (8);</p> <p>17: Update velocity of each flow V_f with Eq. (7);</p> <p>18: Limit velocity for V_f based on v^{low} and v^{up};</p> <p>19: Calculate $X_f^{new}(i)$ using Eq. (9);</p> <p>20: else</p> <p>21: Generate random integer number of r;</p> <p>22: Calculate $X_f^{new}(i)$ using Eq. (10);</p> <p>23: end if</p> <p>24: Limit position with lower boundary (b^{low}) and upper boundary (b^{up});</p> <p>25: Evaluate $f(X_f^{new}(i))$ using Eq. (2);</p> <p>26: $\mu \leftarrow \mu + 1$;</p> <p>27: Update current flow and best flow.</p> <p>28: end for</p> <p>29: end while</p> <p>Output: $X_{best}, f(X_{best})$</p>

Fig. 1 Pseudocode for FDA

$$Mean\ accuracy = mean(1 - Error_{mean}) \quad (11)$$

where $Error_{mean}$ represents the final value of the classification error produced by each run. Besides that, average number of selected features is formulated as another performance metric to measure the performance of the method in minimizing the number of features of a given dataset. A method tends to have better performance when the number of selected features is lower. The average number of selected features produced by each method in solving a given dataset can be calculated as follow:

Table 1 List of datasets

No.	Dataset	Instances	Number of features	Classes	Area
1	Dermatology	358	34	6	Life
2	Glass identification	214	9	8	Physical
3	Lymphography	148	18	3	Life
4	Statlog (heart)	270	13	2	Life
5	Ionosphere	351	34	2	Physical
6	connectionist bench (Sonar, mines vs. rocks)	208	60	2	Physical
7	Parkinson	195	22	2	Life
8	Waveform database generator (version 1)	5000	21	3	Physical

$$\text{Average number of selected features} = \text{find}(X_{best} > \tau) \quad (12)$$

where X_{best} refers to the best solution in the search space and τ represents the threshold values of the removal of redundant features from the original datasets.

4.3 Comparisons Between Selected Algorithms

The performance result, in terms of mean accuracy Acc_{mean} and average number of selected features $N_{feature}$, in solving eight datasets are reported in Tables 2 and 3, respectively. The value with boldface represents the best result while underlined, and italic values indicate the second-best result. In Table 2, the higher Acc_{mean} value implies that the algorithm has better performance in classifying the given dataset. Meanwhile in Table 3, the lower $N_{feature}$ value indicates that the algorithm has better performance in selecting minimum number of features from the given dataset. Based on Table 2, it shows that FDA has the best classification performance as it able to produce best Acc_{mean} in solving datasets 2, 3, 4, 6 and 7. It is followed by DE, HGSO, AVOA and BLPSO. By referring to Table 3, AVOA the best performance in selecting the least number of features followed by HGSO, BLPSO, DE and FDA. It shows that FDA is not effective in selecting minimum number of features. However, FDA has the best accuracy when solving datasets with relatively lesser instances, known as Datasets 2, 3, 4, 6 and 7. This behavior might be caused by the inherent drawbacks of FDA in dealing with large number of features. Hence, as potential future study, additional mechanisms are required to be implemented into FDA to enhance its performance in reducing number of selected features. However, the performance comparisons show that FDA can produce higher accuracy when solving the selected datasets.

Table 2 Mean accuracy Acc_{mean}

Dataset	FDA	DE	BLPSO	HGSO	AVOA
Dataset 1	<u>9.964E-01</u>	1.000E+00	9.601E-01	9.732E-01	9.901E-01
Dataset 2	8.299E-01	7.857E-01	6.563E-01	<u>7.897E-01</u>	7.571E-01
Dataset 3	6.885E-01	4.828E-01	4.632E-01	<u>6.115E-01</u>	5.897E-01
Dataset 4	9.095E-01	8.907E-01	8.586E-01	8.309E-01	<u>9.037E-01</u>
Dataset 5	<u>9.437E-01</u>	9.719E-01	9.314E-01	8.995E-01	9.371E-01
Dataset 6	9.680E-01	9.187E-01	8.902E-01	8.829E-01	<u>9.211E-01</u>
Dataset 7	9.780E-01	8.949E-01	<u>9.453E-01</u>	9.436E-01	8.991E-01
Dataset 8	<u>8.420E-01</u>	8.456E-01	8.234E-01	8.245E-01	8.313E-01

Table 3 Average number of selected features $N_{feature}$

Dataset	FDA	DE	BLPSO	HGSO	AVOA
Dataset 1	12.2000	10.8333	16.9667	13.6000	<u>11.4333</u>
Dataset 2	<u>4.0000</u>	5.0000	4.5667	5.1333	3.9000
Dataset 3	5.6000	4.5667	8.1667	<u>3.6667</u>	3.3667
Dataset 4	5.6333	5.7667	5.5000	3.6000	<u>4.9000</u>
Dataset 5	10.2667	7.1667	16.8000	<u>2.3333</u>	1.5000
Dataset 6	23.0667	18.3333	30.2667	9.9333	<u>12.7000</u>
Dataset 7	3.7333	2.3667	10.1667	<u>2.2667</u>	1.7667
Dataset 8	15.0333	15.5667	12.0000	<u>14.2667</u>	16.9667

5 Conclusion

FDA is designed as a feature selection algorithm to remove redundant features from the datasets. The comparative study reports that FDA is able to obtain the best mean accuracy against other MSAs. However, the effectiveness of FDA in selecting minimum number of features has room of improvement. As future studies, additional searching mechanism can be implemented to improve the performance in selecting lesser features without affecting the accuracy.

Acknowledgements This work was supported by the Ministry of Higher Education Malaysia under the Fundamental Research Schemes with Project codes of FRGS/1/2019/TK04/UCSI/02/1 and FRGS/1/2020/TK0/UCSI/02/4. This work is also supported by the UCSI University Research Excellence & Innovation Grant (REIG) with project code of REIG-FETBE-2022/038.

References

1. Hira ZM, Gillies DF (2015) A review of feature selection and feature extraction methods applied on microarray data. *Adv Bioinformat* 2015
2. Emary E, Zawbaa HM (2019) Feature selection via Lévy Antlion optimization. *Pattern Anal Appl* 22(3):857–876
3. Neggaz N, Houssein EH, Hussain K (2020) An efficient henry gas solubility optimization for feature selection. *Expert Syst Appl* 152:113364. <https://www.sciencedirect.com/science/article/pii/S0957417420301895>
4. Koon Meng Ang MRBMJ, Lim WH, Tiang SS, Ang CK, Hussin EE, Pan L, Chong PH (2022) New hybridization algorithm of differential evolution and particle swarm optimization for efficient feature selection. In: *Proceedings of international conference on artificial life & robotics (ICAROB2022)*, vol 27, pp 148–152, 2022/01/22 January 20
5. Alrifay M, Lim WH, Ang CK (2021) A novel deep learning framework based RNN-SAE for fault detection of electrical gas generator. *IEEE Access* 9:21433–21442
6. Alrifay M et al (2022) Hybrid deep learning model for fault detection and classification of grid-connected photovoltaic system. *IEEE Access* 10:13852–13869
7. Jdid B, Lim WH, Dayoub I, Hassan K, Juhari MRBM (2021) Robust automatic modulation recognition through joint contribution of hand-crafted and contextual features. *IEEE Access* 9:104530–104546
8. Low JW, Tiang SS, Lim WH, Chong YH, Voon YN (2022) Tomato leaf health monitoring system with SSD and MobileNet. In: Zain ZMd, Sulaiman MH, Mohamed AI, Bakar MS, Ramli MS (eds) *Proceedings of the 6th international conference on electrical, control and computer engineering*, Singapore. Springer, Singapore, pp 795–804
9. Voon YN, Ang KM, Chong YH, Lim WH, Tiang SS (2022) Computer-vision-based integrated circuit recognition using deep learning. In: Zain ZMd, Sulaiman MH, Mohamed AI, Bakar MS, Ramli MS (eds) *Proceedings of the 6th international conference on electrical, control and computer engineering*, Singapore. Springer, Singapore, pp 913–925
10. Stańczyk U (2015) Feature evaluation by filter, wrapper, and embedded approaches. In: Stańczyk U, Jain LC (eds) *Feature selection for data and pattern recognition*. Springer, Berlin, pp 29–44
11. Hancer E, Xue B, Zhang M (2018) Differential evolution for filter feature selection based on information theory and feature ranking. *Knowl Based Syst* 140:103–119. <https://www.sciencedirect.com/science/article/pii/S0950705117304987>
12. Huan L, Lei Y (2005) Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans Knowl Data Eng* 17(4):491–502
13. Maldonado S, Weber R (2009) A wrapper method for feature selection using Support Vector Machines. *Inf Sci* 179(13):2208–2217
14. Cheng W-L et al. (2022) Particle swarm optimization with modified initialization scheme for numerical optimization. In: Zain ZMd, Sulaiman MH, Mohamed AI, Bakar MS, Ramli MS (eds) *Proceedings of the 6th international conference on electrical, control and computer engineering*, Singapore. Springer, Singapore, pp 497–509
15. Yao L, Chen YQ, Lim WH (2015) Internet of things for electric vehicle: an improved decentralized charging scheme. In: *2015 IEEE International conference on data science and data intensive systems*, 11–13 December 2015, pp 651–658
16. Yao L, Lim WH, Tiang SS, Tan TH, Wong CH, Pang JY (2018) Demand bidding optimization for an aggregator with a genetic algorithm. *Energies* 11(10)
17. Yao L, Lai CC, Lim WH (2015) Home energy management system based on photovoltaic system. In: *2015 IEEE International conference on data science and data intensive systems*, 11–13 December 2015, pp 644–650
18. Yao L, Lim WH (2018) Optimal purchase strategy for demand bidding. *IEEE Trans Power Syst* 33(3):2754–2762

19. Ang KM, Lim WH, Tiang SS, Ang CK, Natarajan E, Ahamed Khan MKA (2022) Optimal training of feedforward neural networks using teaching-learning-based optimization with modified learning phases. In: Isa K et al. (ed) Proceedings of the 12th national technical seminar on unmanned system technology 2020, Singapore. Springer, Singapore, pp 867–887
20. Suresh S, Elango N, Venkatesan K, Lim WH, Palanikumar K, Rajesh S (2020) Sustainable friction stir spot welding of 6061-T6 aluminium alloy using improved non-dominated sorting teaching learning algorithm. *J Mater Res Technol* 9(5):11650–11674. <https://www.sciencedirect.com/science/article/pii/S2238785420316501>
21. Natarajan E, Kaviarasan V, Lim WH, Tiang SS, Tan TH (2018) Enhanced multi-objective teaching-learning-based optimization for machining of Delrin. *IEEE Access* 6:51528–51546
22. Natarajan E, Kaviarasan V, Lim WH, Tiang SS, Parasuraman S, Elango S (2020) Non-dominated sorting modified teaching-learning-based optimization for multi-objective machining of polytetrafluoroethylene (PTFE). *J Intell Manuf* 31(4):911–935
23. Hassan C, Durai V, Sapuan S, Nuraini AA, Mohamed Yusoff MZ (2018) Mechanical and crash performance of unidirectional oil palm empty fruit bunch fibre-reinforced polypropylene composite. *Bioresources* 13:8310–8328. <https://doi.org/10.15376/biores.13.4.8310-8328>
24. Hassan C, Sapuan S, Nuraini AA, Mohamed Yusoff MZ (2018) Effect of chemical treatment on the tensile properties of single oil palm empty fruit bunch (OPEFB) fibre. 3. <https://doi.org/10.31031/TTEFT.2018.03.000560>
25. Hassan C, Pei Q, Sapuan S, Nuraini AA, Mohamed Yusoff MZ (2018) Crash performance of oil palm empty fruit bunch (OPEFB) fibre reinforced epoxy composite bumper beam using finite element analysis. *Int J Autom Mech Eng* 15:5826–5836. <https://doi.org/10.15282/ijame.15.4.2018.9.0446>
26. Solihin MI, Lim WH, Tiang SS, Ang CK (2019) Modified particle swarm optimization for robust anti-swing gantry crane controller tuning. In: Zain ZMd et al (eds) Proceedings of the 11th National technical seminar on unmanned system technology 2019, Singapore. Springer, Singapore, pp 1173–1192
27. Zè H, Ang CK, Lim WH, Yu LJ, Solihin MI (2020) Development of an artificial intelligent approach in adapting the characteristic of polynomial trajectory planning for robot manipulator. 9:408–414. <https://doi.org/10.18178/ijmerr.9.3.408-414>
28. Sharma A, Dasgotra A, Tiwari SK, Sharma A, Jatly V, Azzopardi B (2021) Parameter extraction of photovoltaic module using tunicate swarm algorithm. *Electronics* 10(8). <https://doi.org/10.3390/electronics10080878>
29. Sharma A et al (2021) Opposition-based tunicate swarm algorithm for parameter optimization of solar cells. *IEEE Access* 9:125590–125602. <https://doi.org/10.1109/ACCESS.2021.3110849>
30. Sharma A, Mathur S (2018) Comparative analysis of ML-PSO DOA estimation with conventional techniques in varied multipath channel environment. *Wirel Personal Commun* 100(3):803–817. <https://doi.org/10.1007/s11277-018-5350-0>
31. Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. *IEEE Trans Evol Comput* 1(1):67–82
32. Karami H, Anaraki MV, Farzin S, Mirjalili S (2021) Flow direction algorithm (FDA): a novel optimization approach for solving optimization problems. *Comput Ind Eng* 156:107224
33. Chow VT (1964) Handbook of applied hydrology: a compendium of water-resources technology
34. O’Callaghan JF, Mark DM (1984) The extraction of drainage networks from digital elevation data. *Comput Vision Graph Image Process* 28(3):323–344
35. Jenson SK, Domingue JO (1988) Extracting topographic structure from digital elevation data for geographic information system analysis. *Photogramm Eng Remote Sens* 54(11):1593–1600
36. Dua D, Graff C (2017) UCI machine learning repository
37. Price K, Storn RM, Lampinen JA (2006) Differential evolution: a practical approach to global optimization. Springer Science & Business Media
38. Chen X, Tianfield H, Mei C, Du W, Liu G (2017) Biogeography-based learning particle swarm optimization. *Soft Comput* 21(24):7519–7541

39. Hashim FA, Houssein EH, Mabrouk MS, Al-Atabany W, Mirjalili S (2019) Henry gas solubility optimization: a novel physics-based algorithm. *Fut Gener Comput Syst* 101:646–667. <https://www.sciencedirect.com/science/article/pii/S0167739X19306557>
40. Abdollahzadeh B, Gharehchopogh FS, Mirjalili S (2021) African vultures optimization algorithm: a new nature-inspired metaheuristic algorithm for global optimization problems. *Comput Ind Eng* 158:107408

Fuzzy Logic Controller by Particle Swarm Optimization Discoverer for Semi-Active Suspension System



Mat Hussin Ab Talib, Nur Hafiezul Mohd. Rosli, Intan Zaurah Mat Darus, Hanim Mohd. Yatim, Muhamad Sukri Hadi, Mohd. Ibthisham Ardani, Mohd. Syahril Ramadhan Mohd. Saufi, and Ahmad Hafizal Mohd. Yamin

Abstract Semi-active suspension systems utilizing magneto-rheological damper have been used especially in the vehicle due to their simple design and control with the effective outcome. Nevertheless, the FL controller design without considering the intelligent algorithm utilizing the FL gain scaling leads to the undesirable condition of the vehicle body. Thus, this study is conducted to develop and evaluate the performance of the particle swarm optimization discoverer (PSOD) in tuning the fuzzy logic (FL) controller in a semi-active suspension system while being compared to the original particle swarm optimization (PSO) and passive system. Taking an acceleration of the suspension system response as an objective function, the PSOD strategy is an attempt to find and search for an optimum value of the gains that able to be a sort of contact information for improving the targeted value obtained from the FL controller. The application of this system is simulated in MATLAB Simulink. The effectiveness of the PSOD was shown by the simulation result with as high as 63.79% and 59.82% of improvement in terms of sprung displacement and sprung acceleration, respectively. This result indicates that the PSOD could provide improvement for vehicle ride comfort and effective improvement solution over the PSO.

Keywords Particle swarm optimization discoverer · Semi-active suspension system · Magneto-rheological damper · Fuzzy logic controller

1 Introduction

Since the early development of car, suspension system is one of the most important part required to sustain the increasing speed and weight. Using direct connection between the car body and tire will put very high stress toward the tire and thus

M. H. Ab Talib (✉) · N. H. Mohd. Rosli · I. Z. M. Darus · H. Mohd. Yatim · Mohd. I. Ardani · Mohd. S. R. Mohd. Saufi · A. H. Mohd. Yamin
Faculty of Mechanical Engineering, Universiti Teknologi Malaysia, 81310 Bahru, Johor, Malaysia
e-mail: mathussin@utm.my

M. S. Hadi
Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
M. A. Abdullah et al. (eds.), *Advances in Intelligent Manufacturing and Mechatronics*,
Lecture Notes in Electrical Engineering 988,
https://doi.org/10.1007/978-981-19-8703-8_17

compromising the structural integrity of the vehicle. This will provide an uncomfortable driving experience and poor handling performance. To overcome these issues, suspension system has evolved rapidly from leaf springs to shock absorber and coil springs. Passive suspension system is basically a combination of these two components. While this setup provides a low-cost, efficient and simple solution [1], however, this system cannot fulfill both requirement of comfort and handling at the same time. This system comes in two variations, soft and hard setup. Soft setup provides comfort, while the hard setup provides good handling. Thus, users are fixed with one setup that focused on only one benefit while compromising the other.

Further, development of the suspension system has come out with a system that is able to fit between the passive and active suspension system, known as semi-active suspension system. This system performs almost the same as the active system as both systems use electronic control to work. However, unlike active system, semi-active system is much less complicated in term of design as it does not require the integration of hydraulic system, but instead it uses a variable damper. This has resulted in cheaper cost and easier implementation as replacing the hydraulic system has removed the need for huge power demand. Furthermore, less mechanical parts will result in less overall weight, easier maintenance and more durable system, as can be seen from the passive system. In short, semi-active suspension system can offer the durability of passive system combined with the performance of the active system altogether in the most cost-effective solution.

This research is a simulation for semi-active suspension system in a vehicle. Magneto-rheological (MR) damper that has been equipped in most semi-active suspension system that can be found today is used to conduct the simulation. MR damper uses MR fluid, which is a non-Newtonian fluid that is reacting to magnetic field. The fluid acts as a fluid in normal condition and changes to semisolid state when there is a presence of magnetic field provided by electrical signal. This will affect the suspension stiffness to change from soft to hard in order to fulfill both comfort and handling requirement in a matter of milliseconds. Other liquid such as electro-rheological liquid also has been used in the semi-active damper [2]. However, due to its characteristics that require high electric field strength but could only give slight changes to the viscosity and also easily affected by temperature, MR fluid is preferred to be used in the damper [3]. Various controllers have been developed by multiple researchers for this purpose, and fuzzy logic (FL) controller is one of the controllers that can be found. However, the controller needs input such as the damper gain to change the suspension stiffness which is always changing from time to time. Thus, optimization algorithm such as particle swarm optimization (PSO) is used to calculate the gain values in real time. Recent research that has been done has developed a new optimized PSO namely particle swarm optimization discoverer (PSOD). The FL controller optimized by PSOD will be the main focus for this research.

There are wide researches that have been done on the optimization algorithm for the controller implemented in the semi-active suspension system. However, research that focuses on the optimized PSO such as the PSOD in FL controller still have not been done. The potential of PSOD over the PSO in performance could not be proved especially in the suspension application. Although there are some algorithms that

offer better performance than PSO such as firefly algorithm (FA) and advanced firefly algorithm (AFA) [4, 5], PSOD might be able to improve the PSO-based algorithm and overcome the problems faced by the original PSO such as easily moved to local optima [6] and less accuracy of control on its speed and direction due to the partial optimism [7]. These disadvantages affect the performance of the suspension system to provide the best solution for both comfort and handling requirements. However, the simplicity of this algorithm together with less parameters requirement compared to other algorithms [8] have become important factors for this algorithm to remain as a consideration to be continuously utilized. Thus, this study will analyze the performance of the PSOD in term of improvement over the original PSO according to the application in vehicle semi-active suspension system in order to provide faster and accurate results that can accommodate both comfort and handling performance.

The paper is organized as follows. Section 2 describes the modeling of suspension system and MR damper system. In Sect. 3, the development and integration process for FL controller and PSOD algorithm technique is elaborated. Section 4 shows the analysis and discussion, and in Sect. 5, the conclusion is described.

2 Semi-Active Suspension Modeling

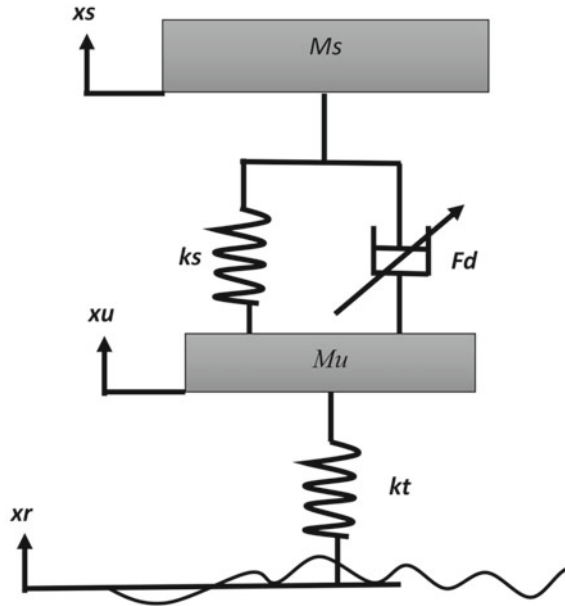
The suspension system is first developed using a passive suspension system since semi-active suspension system has the same configuration except for variable damper. The mathematical model for the suspension system is derived from two degrees of freedom suspension in quarter car model such as in Fig. 1 that shows an example of passive suspension system configuration. It should be noted that the damping of unsprung mass is assumed to be negligible. The mathematical equation for semi-active suspension system is as follows:

$$M_s \ddot{x}_s + F_d - k_s(x_u - x_s) = 0 \quad (1)$$

$$M_u \ddot{x}_u + F_d - k_s(x_u - x_s) - k_t(x_r - x_u) = 0 \quad (2)$$

where the sprung and unsprung mass is symbolized as M_s and M_u , respectively. The damper is represented as F_d , k_s is a spring stiffness and k_t is a tire stiffness. Other parameters like body acceleration, \ddot{x}_s , tire acceleration, \ddot{x}_u , body displacement, x_s , tire displacement, x_u , and road profile displacement, x_r are the important elements that can be defined as the main parameters of interests that were investigated in this study. The parameter value for sprung mass and unsprung mass are set as 80.5 kg and 18.5 kg, respectively, whereas for spring stiffness and tire stiffness, the values are defined as 45,409 N/m and 274,680 N/m, respectively. These scale-down values are taken experimentally regarding the real elements in the suspension structure.

Fig. 1 Semi-active suspension model



The damper system used in this study is based on the magneto-rheological (MR) system and is modeled using parametric approach called the spencer model. The said model is also developed using MATLAB simulation block diagram at its parameters are defined from the previous study by other researchers [9].

3 FL Controller Design and Optimization

Fuzzy logic controller consists of four basic concepts. The first one is fuzzification, next is fuzzy inference engine followed by fuzzy rule base and finally defuzzification. Fuzzification is the process where linguistic variables are converted from raw values. Then, these values will be taken by the fuzzy inference engine for phase of decision-making while being processed through the fuzzy rule base. The finalized linguistic variables are converted back to raw values by the defuzzification before it is fed into the system. In this case, the system is referring to the suspension system.

To integrate the FL controller into the semi-active suspension system, this controller requires two inputs from the system which are sprung velocity and relative velocity from the suspension output. These inputs will be processed by the controller by applying its concepts to produce an output in term of damper constant that will be translated into voltage by force tracking control. Two membership functions are used for each input into the controller as shown in Figs. 2 and 3 that consist of Gaussian shape in negative (N) and positive (P) regions while a bell shape in zero (Z) region. The output of the controller uses six sets of damper constants that has been listed

in Table 1 as its membership function since it uses Sugeno fuzzy inference system. The Sugeno inference system types are used in this study is due to the output of the damper coefficient values are need to be continuous in order to posse more flexibility of the values. Based on these membership functions for both input and output of the controller, the fuzzy rules to evaluate the optimal output are defined as shown in Table 2. This configuration works well as has been proposed by Ab Talib et al. [5] in his study.

Fig. 2 Membership function for sprung velocity

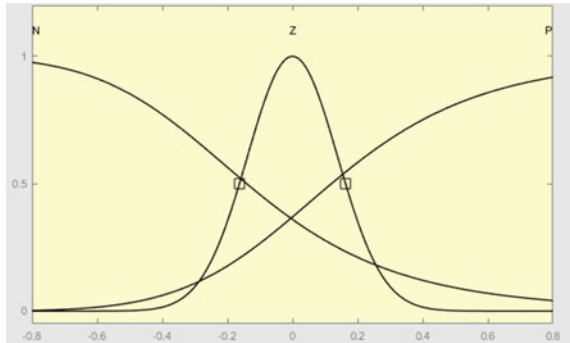


Fig. 3 Membership function for relative velocity

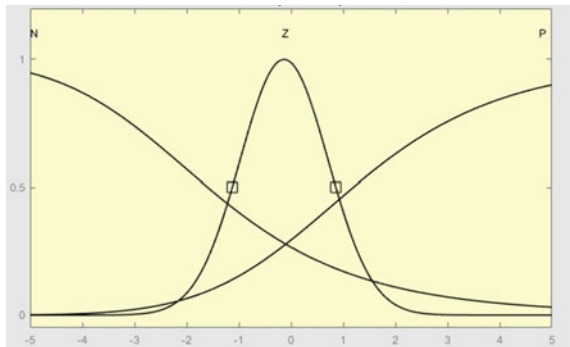


Table 1 Coefficient of fuzzy logic output

Coefficient	Value
C_{min}	700
C_{d1}	3 000
C_{d2}	6 000
C_{d3}	9 000
C_{d4}	12 000
C_{max}	15 000

Table 2 Output evaluation based on fuzzy rules

		Relative velocity		
		N	Z	P
Sprung velocity	N	C_{max}	C_{d2}	C_{min}
	Z	C_4	C_{d3}	C_{d1}
	P	C_{min}	C_{d4}	C_{max}

3.1 Particle Swarm Optimization

Particle swarm optimization (PSO) is an intelligent algorithm that was firstly introduced by Kennedy and Eberhart [10]. The working principle of the PSO is based on the behavior shown by birds' flocks and fish schools to find their food. In PSO, it utilizes the use of swarm of particles that act as the potential solutions. The particles will fly in the search space to explore the best position in order to optimize a required objective function. In every iteration, the particles will be updated with two best values namely *pbest* and *gbest*. *pbest* represents the value of the best position explored by the particle, while *gbest* is the global best value achieved by any particle in the swarm. Equations (3) and (4) will be used to update the position and velocity of the particle, respectively, based on the previous two best values.

$$v_i(k + 1) = wv_i(k) + c_1r_1(pbest - x_i(k)) + c_2r_2(gbest - x_i(k)) \tag{3}$$

$$x_i(k) = x_i^k + v_i^{k+1} \tag{4}$$

Term v_i and x_i represent particle i velocity and position in iteration of k th, respectively. r_1 and r_2 denote random number from 0 to 1. w is the constant for inertia weight. c_1 stands for cognitive acceleration while c_2 is social acceleration. Both of these were indicated as correction factor.

3.2 Particle Swarm Optimization Discoverer

The disadvantage of PSO as described by Marianne et al. [11] is that it tends to be trapped into local optima. In order to overcome this issue, various studies have been conducted to improve the original PSO. One of the improvement made on the PSO such as the one that has been developed by Yatim et al. [12] named as particle swarm optimization discoverer (PSOD) was selected for this study. The improved algorithm was done by integrating another algorithm such as artificial bee colony (ABC) into the original PSO. The scouting behavior of bees is the main inspiration for the ABC algorithm.

Originally, while the particles from the PSO are exploring for the solutions, some of it tend to fly within a small territory as they have stopped exploring globally. This will cause the entire swarm to stop exploring as well due to getting trapped into the local minimum. To overcome this problem, an explorer is initiated among the particles that will restart the exploring capability of the swarm before they went into static state. This will provide more opportunity for the swarm to find another possible position and thus solving the local minimum trapping. In order to prevent the loss of good particles, the particles' fitness values were kept under a specific iteration limit. Once they have achieved the limit, their current personal best position will be assigned to another particle. Then, by using Eq. (3), the result of the new velocity will be allocated to the particles so that they will continue the exploration in a new possible area. Figure 4 indicates the pseudo code of the PSOD. The bolded box in the flow chart indicates the integration of explorer into the original PSO process.

Based on the pseudo code, function of error between the particle fitness of i th and $gbest$ fitness is indicated by $\Delta Fitness$. Precision requirement constant which is denoted by ξ sets the benchmark for the particle to act as an explorer. $count(i)$ is an array that acts a counter for each particle's iteration of exploring. The i th particle that has achieved or exceed the iteration limit ($limit$) will be given new velocity and best position (*reinitialized*). The particle is less likely to become an explorer once the fitness error yield greater result than the required precision.

The integration of PSOD into the semi-active suspension system is exactly the same as PSO such as in the block diagram as shown in Fig. 5 except that the PSO is replaced by PSOD to tune the gains for the FL controller. The parameter values in Table 3 were used for both PSO and PSOD. The gains GX and GV at the input and GOut at the output of the FL controller will be tuned by the PSO and PSOD. The objective function for both the PSO and PSOD were based on the mean squared error (MSE) of the sprung acceleration.

```

FOR  $i = 1: swarmsize$  ; for each particle
    IF  $|\Delta Fitness(i)| < \xi$ 
        THEN  $count(i) = count(i) + 1$  ; add 1
        ELSE  $count(i) = 0$  ; reset
    END
; Memorize the particles' position that has most count
IF  $count(i) \geq limit$ 
THEN  $reinitialized(\text{that } i\text{th particle})$  ; execute new position to the particle
ELSE  $count(i) = 0$  ; reset
END
END

```

Fig. 4 Pseudo code of PSOD

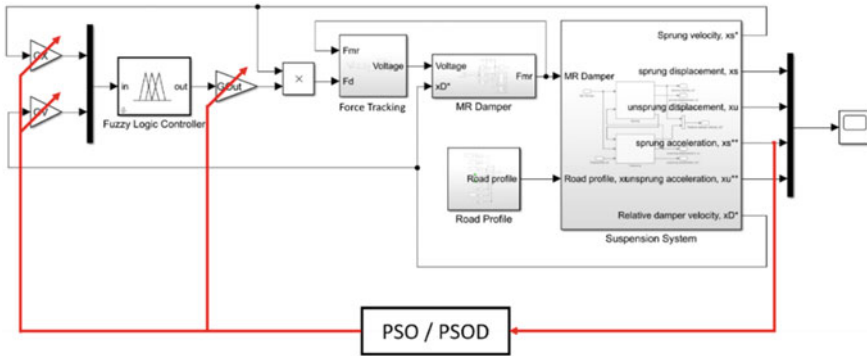


Fig. 5 Block diagram of FL controller tuned by PSO and PSOD

Table 3 Parameter value for PSO and PSOD

Parameter	Value
Swarm size	20
Number of iterations	100
Correction factor, c_1 and c_2	2
Inertia weight, w	1

4 Result and Discussion

The sinusoidal road profile input is created based on the yellow cross line implemented on the road in Malaysia with height of 0.005 m which is in compliance with the specification. This sinusoidal is set at 4 Hz of frequency which is between the natural frequency of the sprung mass and unsprung mass. Next, the round top hump is a single hump disturbance to the suspension and the maximum height of the hump is set at 0.05 m. Taken the hump to be 1 m of width and travel duration of 0.05 s, average speed of user crossing the hump is at 72 km/h.

Based on the tuned parameters that have been finalized in the 100 iterations range for both PSO and PSOD in both road profile of sinusoidal and round top hump, the finalized values for the parameter of GX, GV and GOut are listed out in Table 4. These values are used for the performance evaluation in the respective suspension configuration.

In overall, PSOD tuned suspension system shows massive improvement over the passive system in both road profiles especially for the sprung displacement and sprung acceleration as shown in Tables 5 and 6. This is followed by PSO in the same criteria. However, different trend can be seen for the unsprung acceleration criteria. In sinusoidal road profile, both PSO and PSOD show deterioration where PSOD has the worst performance. While in the round top hump road profile, these two tuning algorithms provide slight improvement over the passive system. These low results behaviors were due to the selection of the objective function for both the PSO and

Table 4 FL parameters tuned by PSO and PSOD

	Sinusoidal		Round top hump	
	PSO	PSOD	PSO	PSOD
GX	0.1426	1.0000	3.6475	5.0000
GV	3.6069	4.9699	0.2406	0.4289
GOut	0.1000	1.0000	0.2906	0.2369

Table 5 Suspension performance in sinusoidal road profile

Sinusoidal					
Performance criteria	Passive	PSO		PSOD	
	MSE	MSE	Percentage	MSE	Percentage
Sprung displacement	5.800×10^{-5}	2.500×10^{-5}	56.90	2.100×10^{-5}	63.79
Sprung acceleration	22.963	10.499	54.28	9.226	59.82
Unsprung acceleration	6.243	9.408	-50.70	10.877	-74.23

Table 6 Suspension performance in round top hump road profile

Round top hump					
Performance criteria	Passive	PSO		PSOD	
	MSE	MSE	Percentage	MSE	Percentage
Sprung displacement	1.440×10^{-4}	8.700×10^{-5}	39.58	6.300×10^{-5}	56.25
Sprung acceleration	87.933	45.921	47.78	32.970	62.51
Unsprung acceleration	598.014	595.274	0.46	518.343	13.32

PSOD. Since sprung acceleration is the only selected for the objective function, the unsprung acceleration performance was affected. Even so, selection of the sprung acceleration does provide benefits to the sprung displacement as well. Thus, in this case, it can clearly be seen that PSOD provides the best solution for suspension control as opposed to PSO and passive system specially to provide ride comfort.

From Figs. 6 and 7, it can be seen that PSO and PSOD has lower overshoot and faster settling time in the given five second simulation time. Hence, it should be noted that both intelligent controllers are superior than the conventional suspension solution.

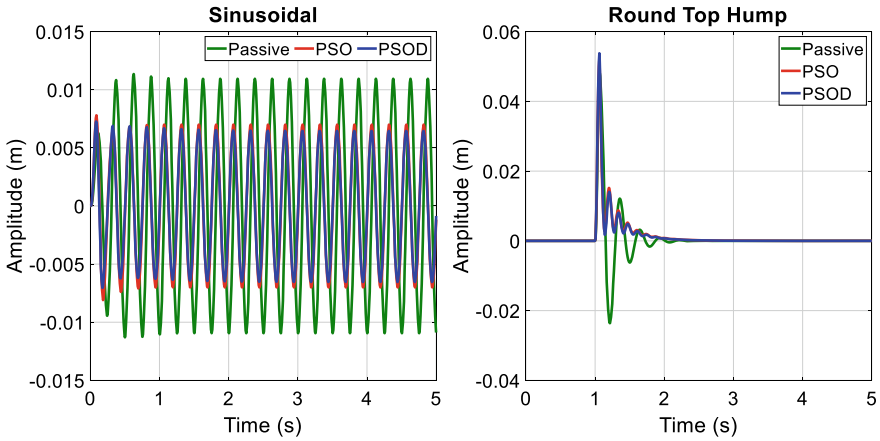


Fig. 6 Vertical body displacement performance in time domain

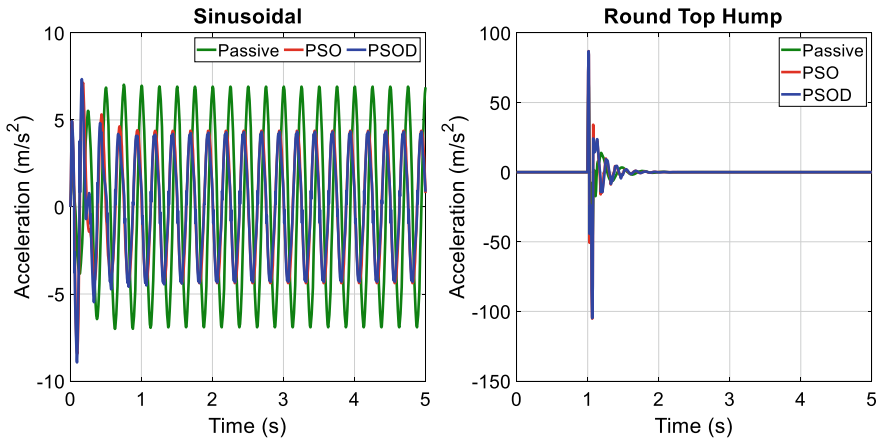


Fig.7 Vertical body acceleration performance in time domain

5 Conclusion

The performance of the PSOD based on the conducted simulation in MATLAB Simulink proved that PSOD provides better performance than PSO and passive system. This can be seen based on the result that shows an improvement as high as 63.79% and 59.82% in term of sprung displacement and sprung acceleration reduction, respectively. Hence, the implementation of PSOD in tuning the FL controller for semi-active suspension system has shown an improvement that could offer a better option to achieve vehicle comfort as opposed to PSO and passive suspension system. It is also should be highlighted that PSOD algorithm is as simple as PSO in order to be implemented as it did not need too much variables that need to be figured out.

Acknowledgements The authors would like to express their gratitude to Minister of Education Malaysia (MOE) and Universiti Teknologi Malaysia (UTM) for funding and providing facilities to conduct this research.

References

1. Soliman AMA, Kaldas MMS (2021) Semi-active suspension systems from research to mass-market—a review. *J Low Freq Noise Vibr Active Control* 40(2):1005–1023
2. Elsaady W, Oyadiji SO, Nasser A (2020) Magnetic circuit analysis and fluid flow modeling of an MR damper with enhanced magnetic characteristics. *IEEE Trans Magnet* 56(9)
3. Houzhong Z, Jiasheng L, Chaochun Y, Xiaoqiang S, Yingfeng C (2020) Application of explicit model predictive control to a vehicle semi-active suspension system. *J Low Freq Noise Vib Active Control* 39(3):772–786
4. Ab Talib MH, Mat Darus IZ, Mohd Samin P (2019) Fuzzy logic with a novel advanced firefly algorithm and sensitivity analysis for semi-active suspension system using magneto-rheological damper. *J Ambient Intell Humaniz Comput* 10(8):3263–3278
5. Ab Talib MH et al (2021) Vibration control of semi-active suspension system using PID controller with advanced firefly algorithm and particle swarm optimization. *J Ambient Intell Human Comput* 12(1):1119–1137
6. Houssein EH, Gad AG, Hussain K, Suganthan PN (2021) Major advances in particle swarm optimization: theory, analysis, and application. *Swarm Evol Comput* 63(3):100868
7. Zhang XW, Liu H, Tu LP (2020) A modified particle swarm optimization for multimodal multi-objective optimization. *Eng Appl Artif Intell* 95(8):103905
8. Gopal A, Sultani MM, Bansal JC (2020) On stability analysis of particle swarm optimization algorithm. *Arab J Sci Eng* 45(4):2385–2394
9. Mohd Yamin AH, Mat Darus IZ, Mohd Nor NS, Ab Talib MH (2021) Intelligent cuckoo search algorithm of pid and skyhook controller for semi-active suspension system using magneto-rheological damper. *Malays J Fund Appl Sci* 17(4):402–415
10. Jame K, Russell E (1995) Particle swarm optimization. In: *Proceedings of ICNN'95—International conference on neural networks*, vol 1, no 1, pp 1942–1948
11. Cherrington M, Airehrour D, Lu J, Thabtah F, Xu Q, Madanian S (2019) Particle swarm optimization for feature selection: a review of filter-based classification to identify challenges and opportunities. In: *2019 IEEE 10th annual information technology, electronics and mobile communication conference, IEMCON 2019*, pp 523–529
12. Yatim HM et al (2022) Intelligent optimization of novel particle swarm optimization with explorer (PSOE) for identification of flexible manipulator system. *Enabling Ind 4.0 Through Adv Mechatron* 900:361–373

Optimized Machine Learning Model with Modified Particle Swarm Optimization for Data Classification



Kah Sheng Lim, Koon Meng Ang, Nor Ashidi Mat Isa, Sew Sun Tiang, Hameedur Rahman, Balaji Chandrasekar, Eryana Eiyada Hussin, and Wei Hong Lim

Abstract Metaheuristic search algorithms (MSAs) receive increasing popularity in recent year due to its excellent capability of solving complex real-world optimization problems without depending on gradient information. Particle swarm optimization (PSO), as one of MSAs, is widely used in optimization task due to its simple framework and quick convergence speed toward global optimum. However, conventional PSO suffers from premature convergence and quick diversity loss of population when the population is poorly initialized due to its random characteristics. In this paper, a new variant of PSO namely PSO with multi-chaotic scheme (PSOMCS) is introduced to train artificial neural network (ANN) by optimizing its neuron weights, biases and selection of suitable activation function based on the datasets obtained from UCI machine learning repository. Initial population generated using multi-chaotic system and oppositional-based learning ensure broader search space coverage, enabling PSOMCS to solve complex optimization problems effectively. Classification performances of ANN trained with PSOMCS are compared with other existing PSO variants. Based on simulation results, ANN optimized by PSOMCS outperformed its competitors in terms of classification performance for both training and testing datasets.

Keywords Artificial neural network · Chaotic system · Classification · Particle swarm optimization · Opposition-based learning

K. S. Lim · K. M. Ang · S. S. Tiang · E. E. Hussin · W. H. Lim (✉)
Technology and Built Environment, UCSI University, 56000 Kuala Lumpur, Malaysia
e-mail: limwh@ucsiuniversity.edu.my

N. A. M. Isa
School of Electrical and Electronics Engineering, Universiti Sains Malaysia, Engineering Campus, 14300 Nibong Tebal, Pulau Pinang, Malaysia

H. Rahman
Faculty of Computing and Artificial Intelligence, Air University, Islamabad Capital Territory, Islamabad 44000, Pakistan

B. Chandrasekar
Department of Electrical and Electronics Engineering, SRM Institute of Science and Technology, Chennai 603203, Tamil Nadu, India

1 Introduction

Optimization is an essential tool used for decision-making, and it involves a set of procedures to determine the best combinations of decision variables, aiming to maximize or minimize the predefined goals expressed as objective functions [1]. In contrary to conventional optimization algorithms, metaheuristic search algorithms (MSAs) emerge as more the promising approaches to solve various real-world optimization problems [2–9] without depending on the gradient information. These complex optimization problems include black box problems where information such as objective function and constraint functions remain unknown [10]. Generally, MSAs are equipped with search mechanisms to improve candidate solutions in every generation and the best solution obtained during termination stage is utilized to solve a given problem.

Particle swarm optimization (PSO) was developed by Kennedy and Eberhart in 1995, and it was inspired by the synchronized flight pattern of birds as well as their optimal formation and grouping to search for foods [11]. This notion subsequently evolved into a population-method search mechanism in which the candidates or particles are moving together as a swarm to seek for the optimal solution [12, 13]. PSO has been used to handle a variety of optimization problems [14–18] due to its high convergence rate and promising global search ability. Although basic PSO can provide adequate performance, notable performance degradations can be observed when handling complex optimization problems, such as optimizing artificial neural network (ANN) model. The imbalanced trade-off between exploration and exploitation is main reason that causes the premature convergence of PSO toward inferior regions. Excessive explorative behavior prolongs convergence time toward global optimum, whereas excessive behavior results in rapid diversity loss of population. Appropriate modifications are required to improve PSO's performance in dealing with complex optimization problems.

Backpropagation (BP) algorithm is a conventional approach used for training ANN, but it is extremely reliant on the initial values assigned for ANN parameters, i.e., weights and biases. BP with the improper initialization of parameters tends to have higher possibility of being stuck into local optima and fail to search for the optimal ANN parameters during the training process [19]. To mitigate the drawbacks of BP, various MSAs with better global search ability were proposed as the alternative approaches used for training the ANN classifier [20]. In this study, a new variant of PSO known as particle swarm optimization with multi-chaotic scheme (PSOMCS) is first proposed and then employed as the training algorithm of ANN to enhance the latter's performance in solving classification tasks. The following are the primary contributions of this paper:

- The training of ANN model is framed as an optimization problem, aiming to determine the optimal combinations of neuron weight, bias values and activation function required for solving classification tasks.
- A new variant of PSO known as PSOMCS is designed and used as a training algorithm for ANN classifier to solve classification problems with better accuracy.

- A modified initialization scheme is introduced into PSOMCS, where multiple numbers of chaotic maps and oppositional-based learning (OBL) concept were used for developing initial population with better quality and ensure broader coverage in solution space.
- Classification performances of ANN model trained by the proposed PSOMCS are then analyzed, evaluated, and compared with other PSO variants using 16 datasets obtained from UCI machine learning repository.

The remaining portion of this work are structured as follows. Section 2 describes the mechanism of conventional PSO, and Sect. 3 presented the formulation of ANN training as an optimization problem along with the description of PSOMCS framework. In Sect. 4, the datasets used for performance evaluation as well as the classification performances of ANN trained using PSOMCS, and other PSO variants are reported. In the last section, Sect. 5 presented the conclusion and future works.

2 Conventional PSO

Conventional PSO was proposed by Kennedy and Eberhart in 1995, and it was inspired from the natural behavior of animal species such as fish schooling or bird flocking to look for food source [11]. For a given optimization problem, the dimensional size, D (i.e., number of parameters to be optimized) and population size, N are firstly defined. Indicates the i -th particle's location and velocity as $X_i = [X_{i,1}, \dots, X_{i,d}, \dots, X_{i,D}]$ and $V_i = [V_{i,1}, \dots, V_{i,d}, \dots, V_{i,D}]$, respectively, where $i = 1, \dots, N$ and $d = 1, \dots, D$. Each particle's personal best experience is expressed as $P_{best,i} = [P_{best,i,1}, \dots, P_{best,i,d}, \dots, P_{best,i,D}]$, and the global best experience found by all the particles in the population is denoted as $G_{best} = [G_{best,1}, \dots, G_{best,d}, \dots, G_{best,D}]$. During the searching process, the search trajectory of each particle is adjusted based its personal best experience, $P_{best,i}$ and global best experience, G_{best} . Given the newly updated velocity, the position vector of each particle is updated. Particularly, the new velocity $V_{i,d}(t+1)$ and position vectors $P_{i,d}(t+1)$ of each particle are calculated as follows:

$$V_{i,d}(t+1) = \omega V_{i,d}(t) + c_1 r_1 (P_{best,i,d}(t) - X_{i,d}(t)) + c_2 r_2 (G_{best,d}(t) - X_{i,d}(t)) \quad (1)$$

$$X_{i,d}(t+1) = X_{i,d}(t) + V_{i,d}(t+1) \quad (2)$$

where ω is the inertia weight; c_1 and c_2 are the coefficients for acceleration; r_1 and r_2 are the random numbers generated from uniform distribution within the interval $[0, 1]$. Once the updated position of each i -th particle is obtained, the fitness evaluation is performed on each i -th particle using objective function. The fitness value $f(X_i(t+1))$ obtained by i -th particle's new position is compared with fitness value $f(X_i(t))$ of its current position. If the new position vector is fitter than the current

position vector in terms of its fitness value, then the current personal and global best positions are replaced. This searching process is repeated and stopped when the predefined termination criteria are satisfied. Finally, the $G_{best}(t)$ are returned as optimal solution for the given optimization task.

3 Optimized ANN Model Using PSOMCS

3.1 Formulation of ANN Training as an Optimization Problem

FNN is one of the most basic ANN models for solving real-world classification tasks. In FNN, information goes in one path from input to hidden layers, then to the output layer. A three-layer structure of FNN consists of n input neurons, m hidden neurons, and l output neurons, is illustrated in Fig. 1. Suppose that x_i , h_j , and y_k indicate the i -th input neuron, j -th hidden neuron, and k -th output neuron, respectively, where $i = 1, \dots, n$, $j = 1, \dots, m$, and $k = 1, \dots, l$. The weight between i -th input neuron and j -th hidden neuron is represented as $w_{i,j}^H$; the weight between j -th hidden neuron and k -th output neuron is denoted as $w_{j,k}^O$. The biases of j -th hidden neuron and l -th output neuron are indicated as B_j^H and B_l^O , respectively. The sum of weight and biases of each x_j and h_k , denoted as S_j and S_k , respectively, are computed, followed by nonlinearization process performed by an activation function $\Phi(\cdot)$ to calculate the values of h_k and y_l as follows:

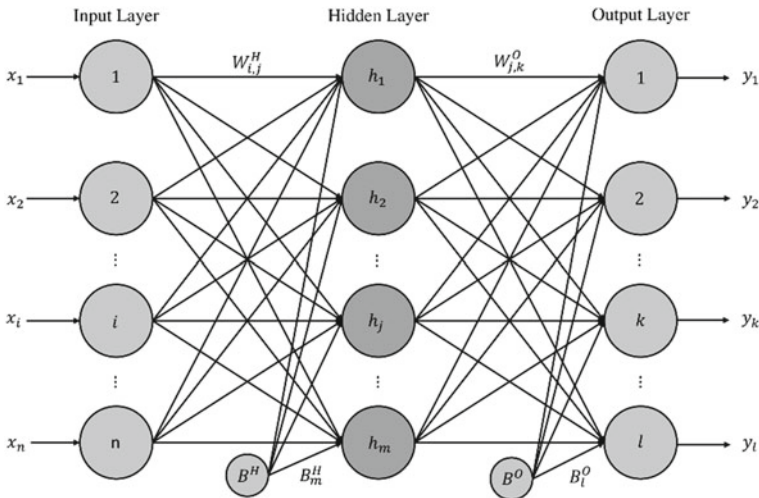


Fig. 1 Feedforward neural network

$$h_k = \Phi(S_j), \text{ where } S_j = \sum_{i=1}^n w_{i,j} x_i + B_j^H \quad (3)$$

$$y_l = \Phi(S_k), \text{ where } S_k = \sum_{j=1}^m w_{j,k} x_j + B_k^O \quad (4)$$

In this paper, the variables required to be optimized during training of ANN are (a) weight value for $w_{i,j}^H, w_{j,k}^O \in [-1, 1]$ (b) bias value for $B_j^H, B_k^O \in [-1, 1]$, and (c) index number $Q = \{1, 2, 3, 4, 5\}$ which represents the five candidates of activation functions known as binary step, sigmoid, hyperbolic tangent, inverse tangent, and rectified linear unit (ReLU) activation functions, respectively. Dimensional components for position vector of each particle can be encoded as a vector form represented as follow:

$$X = [w_{1,1}^H, \dots, w_{i,j}^H, \dots, w_{n,m}^H, w_{1,1}^O, \dots, w_{j,k}^O, \dots, w_{m,l}^O, \beta_1^H, \dots, \beta_j^H, \dots, \beta_m^H, \beta_1^O, \dots, \beta_k^O, \dots, \beta_l^O, Q] \quad (5)$$

Total dimensional size, D required for ANN training and testing is computed as follow:

$$D = (n \times m) + (m \times l) + m + l + 1 \quad (6)$$

In this paper, mean square error (MSE) is formulated as the objective function that need to be minimized. MSE is measured between the calculated output from ANN model and desired output given in the dataset, and computed as below

$$f(X) = MSE = \frac{1}{R} \sum_{r=1}^R (Y_r(X) - T_r(X))^2 \quad (7)$$

where R denotes the total quantity of data samples given in dataset; $Y_r(X)$ represents the output predicted by ANN for r -th data sample; $T_r(X)$ represents the real output of r -th data sample provided. The particle that can produce the smallest MSE value is considered as global best particle, and the corresponding optimal parameters can be extracted to generate the optimized ANN model with high classification accuracy.

3.2 Particle Swarm Optimization with Multi-chaotic Scheme

Random initialization scheme is adopted in conventional PSO to initialize the population. However, some particles might be generated at a region where the global optimum is too far away. During the searching process, the particles are more likely

to be trapped within the undesirable search region such as local optima, resulting in premature convergence [21]. To overcome this drawback, both concepts of multi-chaotic scheme (MCS) and opposition-based learning (OBL) are incorporated into PSOMCS, intending to enhance its search accuracy.

At the early stage of initialization process of PSOMCS, the population is generated based on the chaotic map sequence. Denotes γ_0 as the initial value of a chaotic variable that is randomly produced in each separate simulation. Suppose that γ_z is chaotic variable generated by the chaotic map in z -th sequence, where $z = 1, 2, \dots, Z$ and Z refers to the total number of sequences. Chaotic population is then generated by using chaotic map selected based on four factors: (i) Circle map if $\gamma_0 \leq 0.25$, (ii) Gauss map if $0.25 < \gamma_0 \leq 0.5$, (iii) Singer map if $0.5 < \gamma_0 \leq 0.75$, and (iv) Sinusoidal map if $0.75 < \gamma_0 \leq 1$, where their respective mathematical models are provided as follows:

Circle:

$$\gamma_{z+1} = \left| \gamma_z + b - \left(\frac{a}{2\pi} \right) \sin(2\pi \gamma_z), 1 \right| \tag{8}$$

Gauss:

$$\gamma_{z+1} = \begin{cases} 1, & \gamma_z = 0 \\ \frac{1}{|\gamma_z|}, & \text{otherwise} \end{cases} \tag{9}$$

Singer:

$$\gamma_{z+1} = \mu(7086\gamma_z - 23.32\gamma_z^2 + 28.75\gamma_z^3 - 13.302875\gamma_z^4), \mu = 1.07 \tag{10}$$

Sinusoidal:

$$\gamma_{z+1} = a\gamma_z \sin(\pi \gamma_z), a = 2.3 \tag{11}$$

Define $X_{i,d}^{max}$ as the upper limit and $X_{i,d}^{min}$ as lower limits for each i -th chaotic particle at d -th dimension where $i = 1, \dots, I$ and $d = 1, \dots, D$. Given final chaotic value γ_z , the d -th dimension of position for each i -th chaotic particle, $X_{i,d}^{CS}$ is generated as

$$X_{i,d}^{CS} = X_{i,d}^{min} + \gamma_z(X_{i,d}^{max} - X_{i,d}^{min}) \tag{12}$$

After $X_{i,d}^{CS}$ of i -th chaotic particle is calculated, the corresponding opposite position vector $X_{i,d}^{OBL}$ is then calculated using OBL scheme [22] to ensure wider coverage of solutions in search space. The position vector of $X_{i,d}^{OBL}$ is computed as follow:

$$X_{i,d}^{OBL} = X_{i,d}^{max} + X_{i,d}^{min} - X_{i,d}^{CS} \quad (13)$$

The chaotic and opposite particles are stored in the chaotic population $\mathbf{P}^{CS} = [X_1^{CS}, \dots, X_i^{CS}, \dots, X_I^{CS}]$ and opposite population $\mathbf{P}^{OBL} = [X_1^{OBL}, \dots, X_i^{OBL}, \dots, X_I^{OBL}]$, respectively, with population size of I . Then, a newly combine population set \mathbf{P}^{Merge} is created with a population size of $2I$ by combining two populations as follow:

$$\mathbf{P}^{Merge} = \mathbf{P}^{CS} \cup \mathbf{P}^{OBL} \quad (14)$$

All particles in \mathbf{P}^{Merge} are then sorted from the best to worst according to fitness values (i.e., mean square errors in this case). Then, the first best I particles are selected from \mathbf{P}^{Merge} to be the initial population of PSOMCS, i.e., $\mathbf{P}^{Initial} = [X_1, \dots, X_i, \dots, X_I]$.

Referring to the initial solutions of $\mathbf{P}^{Initial}$, the personal best position of each particle and global best position are generated. The velocity and position vectors of each particle are continuously updated during searching, based on Eqs. (1) and (2), respectively. The fitness value (i.e., mean square error) of the new position vector of each i -th particle is calculated and compared with those of personal best position and global best position. The latter two positions and their mean square error values are replaced by the new position if it is more superior. The optimization process is iterated and terminated when the stopping criterion $\zeta > \zeta^{max}$ is met, where ζ is the fitness evaluation counter and ζ^{max} is predefined maximum number of fitness evaluation. A pseudocode is illustrated in Fig. 2 to explain the optimization framework of PSOMCS.

4 Performance Evaluation of Optimized ANN Model

In this study, sixteen datasets are extracted from University of California Irvine (UCI) machine learning repository to evaluate the classification performance of ANN optimized by PSOMCS. Each selected dataset is split into 70% and 30% for the training and testing of ANN, respectively. The characteristics of each selected datasets are presented in Table 1. Classification accuracy rate (CAR) is a metric used to assess the performance of ANN model. Denotes \check{Z} as the number of data samples properly classified by the ANN model, where Z is the sum of data samples. An ANN model with higher CAR value indicates that the predicted results are more accurate and capable of providing better classification performance due to its exceptional capacity to classify unknown data in different classes based on its comprehension of the existing data. The CAR value is computed as

Algorithm 1: PSOMCS	
Inputs: $l, D, X_i^{max}, X_i^{min}, Z, \zeta^{max}$	
01:	Initialize $P^{CS} = \emptyset$ and $P^{OBL} = \emptyset$;
02:	for each i -th particle do
03:	for each d -th dimension do
04:	Randomly generate $\gamma_0 \in [0, 1]$ and set $z = 1$;
05:	if $\gamma_0 \leq 0.25$ then choose the Circle map as Eq. (8);
06:	else if $0.25 < \gamma_0 \leq 0.5$ then choose Gauss map as Eq. (9);
07:	else if $0.5 < \gamma_0 \leq 0.75$ then choose Singer map as Eq. (10);
08:	else if $0.75 < \gamma_0 \leq 1$ then choose Sinusoidal map as Eq. (11);
09:	end if
10:	while $z \leq Z$ do
11:	Update the chaotic variable, γ_0 with the chaotic map chosen;
12:	Update chaotic sequence with $z \leftarrow z + 1$;
13:	end while
14:	Calculate $X_{i,d}^{CS}$ and $X_{i,d}^{OBL}$ using the Eqs. (12) and (13), respectively;
15:	end for
16:	Update $P^{CS} \leftarrow P^{CS} \cup X_i^{CS}$ and $P^{OBL} \leftarrow P^{OBL} \cup X_i^{OBL}$;
17:	end for
18:	Merge the P^{CS} and P^{OBL} to form P^{Merge} using Eq. (14);
19:	Perform evaluation for all the members in P^{Merge} and sort from best to worst based on their fitness values (i.e., mean square error);
20:	Extract the first best l members from sorted P^{Merge} to form $P^{Initial}$;
21:	for each i -th particle do
22:	Initialize the velocity, V_i to be zero in all dimensions;
23:	Update $P_{best,i}, G_{best}, f(P_{best,i})$ and $f(G_{best})$;
24:	end for
25:	while $\zeta \leq \zeta^{max}$ do
26:	for each i -th particle do
27:	Update V_i and X_i using Eqs. (1) and (2), respectively;
28:	Perform evaluation on new X_i to obtain $f(X_{i,d})$;
29:	Update $P_{best,i}, G_{best}, f(P_{best,i})$ and $f(G_{best})$;
30:	$\zeta \leftarrow \zeta + 1$;
31:	end for
21:	end while
	Output: $G_{best}, f(G_{best})$

Fig. 2 Pseudocode of PSOMCS framework

$$CAR = \frac{\tilde{z}}{Z} \times 100\% \quad (15)$$

The classification performance of ANN optimized with PSOMCS is compared with those optimized by another three PSO variants known as conventional PSO [11], PSO with modified initialization scheme (PSOMIS) [23], and accelerated PSO

Table 1 Characteristics of the selected datasets

Dataset	#Attributes	#Classes	#Samples
Iris	4	3	150
Liver disorder	6	2	345
Blood transfusion	4	2	748
Statlog heart	13	2	270
Hepatitis	19	2	80
Wine	13	3	178
Breast cancer	9	2	277
Seeds	7	3	210
Australian credit	14	2	690
Haberman	3	2	306
New thyroid	5	3	215
Glass	9	6	214
Balance	4	3	625
Dermatology	34	6	338
Landsat	36	6	4435
BankNote	6	2	1376

(APSO) [24]. Every PSO variants' parameter settings are set based on the recommendations of their respective literatures. The population sizes of all PSO variants are set as $I = 100$, and each of them are simulated independently for 30 runs with $\zeta^{max} = 10000 \times D$. All simulations are run using MATLAB R2021b on a personal PC equipped with Intel® Core i7-8750H CPU @ 2.20 GHz.

The CAR values generated using PSOMCS and other PSO's variant in optimizing the ANN's parameters based on training and testing datasets are showed in Tables 2 and 3, respectively. The best CAR values generated among the compared variants in training and testing model are implied with boldface while the second-best CAR values are underlined. #BCAR showed the number of best CAR values produced by every PSO's variant from 16 datasets given. $w/t/l$ represents that the CAR values produced by PSOMCS in optimizing the ANN models are better than each compared PSO variant in w datasets, tie in t datasets and worse in l datasets, respectively.

Based on Table 2, the ANN classifier optimized by PSOMCS are reported to have the leading performances which can generate 11 best training CAR out of all 16 datasets. Meanwhile, ANN classifier optimized by PSOMIS and PSO can occasionally deliver good performances for producing three best training CAR values. ANN models trained by both algorithms are also having the most second-best CAR values comparing to APSO.

In comparison with others, the ANN classifier optimized using PSOMCS is stated to have 14 best testing CAR values out of all datasets. It is also worth noting that PSOMCS is the only one that successfully classified Iris datasets with 100% of CAR

Table 2 Training performance comparisons by all PSO variants

Dataset	PSOMCS	PSO	APSO	PSOMIS
Iris	98.08	94.58	68.75	<u>98.03</u>
Liver disorder	<u>73.11</u>	73.55	60.51	70.29
Blood transfusion	80.99	<u>79.35</u>	78.93	78.93
Statlog heart	95.71	<u>93.52</u>	84.26	92.59
Hepatitis	97.19	89.84	86.72	<u>93.75</u>
Wine	98.80	96.48	73.24	<u>97.18</u>
Breast cancer	83.60	<u>81.76</u>	76.80	80.63
Seeds	<u>94.11</u>	93.15	82.44	97.02
Australian credit	<u>89.38</u>	89.67	81.34	89.67
Haberman	75.55	74.49	73.06	<u>75.10</u>
New thyroid	97.46	93.31	85.17	<u>95.06</u>
Glass	61.77	<u>56.73</u>	23.68	40.35
Balance	<u>89.07</u>	89.40	80.80	87.70
Dermatology	29.02	<u>26.40</u>	23.79	<u>26.40</u>
Landsat	<u>75.26</u>	<u>75.26</u>	61.10	75.39
BankNote	99.71	97.81	95.67	<u>98.91</u>
# BCAR	11	3	0	3
w/t/l	–	13/1/2	16/0/0	13/0/3

value. The ANN classifier trained by PSOMIS produced three best CAR values in classifying all testing datasets, therefore showing a good classification performance.

Notably, all compared methods suffer with the overfitting issues when dealing with Liver Disorder, Blood Transfusion, and Hepatitis datasets because these three datasets might have more challenging fitness landscapes for all compared methods to search ideal combinations for the parameter that can ensure good generalization capabilities of classifiers. Nevertheless, the ANN optimized by proposed PSOMIS still delivers the best performances in terms of testing accuracy when dealing with these datasets.

5 Conclusion

In this study, a novel PSO variants called PSOMCS is designed to replace the conventional BP algorithm for training ANN model to solve classification problems. Both of multi-chaotic scheme and OBL techniques are utilized to modify the initial population of PSOMCS, enabling the generation of initial solutions with better fitness and wider coverage of search space. The simulation result reports that the ANN model optimized by PSOMCS can outperform those optimized by other PSO variants, when

Table 3 Testing performance comparisons by all PSO variants

Dataset	PSOMCS	PSO	APSO	PSOMIS
Iris	100.00	93.33	<u>99.33</u>	90.00
Liver disorder	49.63	41.30	<u>47.10</u>	44.93
Blood transfusion	60.27	<u>61.00</u>	51.67	68.00
Statlog heart	80.30	<u>78.70</u>	75.00	77.78
Hepatitis	65.23	55.38	<u>56.25</u>	50.00
Wine	<u>85.61</u>	43.06	81.94	94.44
Breast cancer	75.39	70.00	<u>72.73</u>	70.91
Seeds	99.52	69.05	<u>96.43</u>	88.10
Australian credit	83.55	68.48	<u>82.61</u>	81.88
Haberman	78.69	<u>71.31</u>	67.21	78.69
New thyroid	64.58	<u>45.35</u>	36.05	40.70
Glass	46.09	30.23	<u>39.53</u>	32.56
Balance	87.36	76.00	<u>86.80</u>	79.60
Dermatology	62.93	56.94	<u>61.11</u>	<u>61.11</u>
Landsat	77.52	40.53	<u>74.63</u>	70.52
BankNote	91.48	81.39	<u>91.06</u>	88.87
# BCAR	14	0	0	3
w/t/l	–	16/0/0	16/0/0	13/1/2

performing classification on the datasets given. The incorporation of multiple chaotic maps and OBL for population initialization is proven able to enhance the robustness of PSOMCS to handle difficult optimization problems such as ANN training. As future works, the potential of PSOMCS can be explored further especially in optimizing different machine learning models such as extreme learning machine and support vector machine.

Acknowledgements This work was supported by the Ministry of Higher Education Malaysia under the Fundamental Research Schemes with project codes of FRGS/1/2019/TK04/UCSI/02/1 and FRGS/1/2020/TK0/UCSI/02/4. This work is also supported by the UCSI University Research Excellence & Innovation Grant (REIG) with project code of REIG-FETBE-2022/038.

References

1. Ahmad MF, Isa NAM, Lim WH, Ang KM (2022) Differential evolution: a recent review based on state-of-the-art works. Alex Eng J 61:3831–3872
2. Yao L, Lim WH (2018) Optimal purchase strategy for demand bidding. IEEE Trans Power Syst 33:2754–2762

3. Yao L, Lai C-C, Lim WH (2015) Home energy management system based on photovoltaic system. In: 2015 IEEE International conference on data science and data intensive systems, pp 644–650
4. Yao L, Chen Y-Q, Lim WH (2015) Internet of things for electric vehicle: an improved decentralized charging scheme. In: 2015 IEEE International conference on data science and data intensive systems, pp 651–658
5. Natarajan E, Kaviarasan V, Ang KM, Lim WH, Elango S, Tiang SS (2022) Production wastage avoidance using modified multi-objective teaching learning based optimization embedded with refined learning scheme. *IEEE Access*. 10:19186–19214
6. Yu L-J, Rengasamy K, Lim K-Y, Tan L-S, Tarawneh M, Zulkoffli ZB, Se Yong EN (2019) Comparison of activated carbon and zeolites' filtering efficiency in freshwater. *J Environ Chem Eng* 7:103223
7. Yu L, Ahmad S, Appadu S, Kong I, Tarawneh M, Flaifel M (2014) Comparison of magnetic and microwave absorbing properties between multiwalled carbon nanotubes nanocomposite, nickel zinc ferrite nanocomposite and hybrid nanocomposite. *World J Eng* 11:317–322
8. Jamaludin FA, Ab-Kadir MZA, Izadi M, Azis N, Jasni J, Abd Rahman MS (2016) Considering the effects of a RTV coating to improve electrical insulation against lightning. In: 2016 33rd International conference on lightning protection (ICLP), pp 1–5
9. Jamaludin FA, Ab-Kadir MZA, Izadi M, Azis N, Jasni J, Rahman MSA, Osman M (2018) Effect of RTV coating material on electric field distribution and voltage profiles on polymer insulator under lightning impulse. In: 2018 34th International conference on lightning protection (ICLP), pp 1–6
10. Audet C, Hare W (2017) Derivative-free and blackbox optimization
11. Kennedy J, Eberhart R (1995) Particle swarm optimization. In: *Proceedings of ICNN'95—International Conference On Neural Networks*, vol 4, pp 1942–1948
12. Karim AA, Mat Isa NA, Lim WH (2020) Modified particle swarm optimization with effective guides. *IEEE Access*. 8:188699–188725
13. Lim WH, Isa NAM, Tiang SS, Tan TH, Natarajan E, Wong CH, Tang JR (2018) A self-adaptive topologically connected-based particle swarm optimization. *IEEE Access*. 6:65347–65366
14. Solihin MI, Lim WH, Tiang SS, Ang CK (2021) Modified particle swarm optimization for robust anti-swing gantry crane controller tuning. In: Md Zain Z, Ahmad H, Pebrianti D, Mustafa M, Abdullah NRH, Samad R, Mat Noh M (eds) *Proceedings of the 11th national technical seminar on unmanned system technology 2019*, pp 1173–1192. Springer, Singapore
15. Meng Ang K, Bin Mohamed Juhari MR, Hong Lim W, Sun Tiang S, Kit Ang C, Eiyda Hussin E, Pan L, Hui Chong T (2022) New hybridization algorithm of differential evolution and particle swarm optimization for efficient feature selection. *Proc Int Conf Artif Life Robot* 27:148–152
16. Priyadarshi N, Padmanaban S, Hiran KK, Holm-Nielson JB, Bansal RC (2021) Artificial intelligence and internet of things for renewable energy systems. *Walter de Gruyter GmbH & Co KG*
17. Ang KM, Lim WH, Isa NAM, Tiang SS, Ang CK, Chow CE, Yeap ZS (2022) Modified particle swarm optimization with unique self-cognitive learning for global optimization problems. In: Ab Nasir AF, Ibrahim AN, Ishak I, Mat Yahya N, Zakaria MA, Abdul Majeed PPA (eds) *Recent trends in mechatronics towards industry 4.0*. Springer, Singapore, pp 263–274
18. Koh W, Lim WH, Ang KM, Mat Isa NA, Tiang S, Ang CK, Solihin MI (2022) Multi-objective particle swarm optimization with alternate learning strategies. Presented at the January 1
19. Mirjalili S, Mohd Hashim SZ, Moradian Sardroudi H (2012) Training feedforward neural networks using hybrid particle swarm optimization and gravitational search algorithm. *Appl Math Comput* 218:11125–11137
20. Ang KM, Lim WH, Tiang SS, Ang CK, Natarajan E, Ahamed Khan MKA (2022) Optimal training of feedforward neural networks using teaching-learning-based optimization with modified learning phases. In: Isa K, Md Zain Z, Mohd-Mokhtar R, Mat Noh M, Ismail ZH, Yusof AA, Mohamad Ayob AF, Azhar Ali SS, Abdul Kadir H (eds) *Proceedings of the 12th National technical seminar on unmanned system technology 2020*. Springer, Singapore, pp 867–887

21. Ahmad MF, Isa NAM, Lim WH, Ang KM (2022) Differential evolution with modified initialization scheme using chaotic oppositional based learning strategy. *Alex Eng J* 61:11835–11858
22. Mahdavi S, Rahnamayan S, Deb K (2018) Opposition based learning: a literature review. *Swarm Evol Comput* 39:1–23
23. Cheng W-L, Ang KM, Choi ZC, Lim WH, Tiang SS, Natarajan E, Ang CK, Khan MKAA (2022) Particle swarm optimization with modified initialization scheme for numerical optimization. In: Md Zain Z, Sulaiman Mohd H, Mohamed AI, Bakar Mohd S, Ramli Mohd S (eds) *Proceedings of the 6th International conference on electrical, control and computer engineering*. Springer, Singapore, pp 497–509
24. Zhang H, Yang Z (2018) Accelerated particle swarm optimization to solve large-scale network plan optimization of resource-leveling with a fixed duration. *Math Probl Eng* 2018:1–11

Performance Comparison of Kalman Filter and Extended Kalman Filter for Human Tracking and Prediction with Particle Swarm Optimisation



Abiodun Afis Ajasa and Nawawi Sophan Wahyudi

Abstract In the past two decades, object tracking has progressively advanced in computer vision and image processing. Tracking is a collection of algorithms that detect and track objects in a video sequence. This has resulted in a broad variety of applications, including surveillance, biometric identification or biological imaging, human-machine interactions, traffic control, and intelligent vehicles, all of which have benefited from tracking applications. This paper presents an IoT-based human tracking system that employs the Kalman filter (KF) and extended Kalman filter (EKF) algorithms. However, the performance of the filters is impacted by noise. Therefore, instead of manually tuning the KF and EKF's process noise and measurement error, particle swarm optimisation (PSO) is utilised to optimise them. Four 2-dimensional models (namely conventional KF and EKF, optimised KF-PSO, and EKF-PSO) were developed and evaluated using ten distinct human sets of data containing 100 samples each. The object's tracked positions are estimated in horizontal and vertical directions. Accuracy analysis was used to compare the four models' quality performance. With an average mean square error of 5.99 mm (or 0.599%), the EKF-PSO model outperformed the conventional EKF model, which had an error of 7.18 mm (or 0.718%). A 17.2 mm (or 1.72%) error was found in the KF-PSO model. The last one is the traditional KF model, yielding an error of 22.4 mm (or 2.24%). As a result of its higher accuracy, the EKF-PSO model outperforms the other three models.

Keywords Kalman filter (KF) · Extended Kalman filter (EKF) · Particle swarm optimisation (PSO) · Internet of Things (IoT) · Accuracy analysis

A. A. Ajasa (✉) · N. S. Wahyudi
Department of Control and Mechatronics Engineering, School of Electrical Engineering,
Universiti Teknologi Malaysia, Johor, Malaysia
e-mail: abiodunafis@graduate.utm.my

A. A. Ajasa
Department of Electronics and Computer Engineering, Lagos State University, Epe Campus,
Lagos, Nigeria

1 Introduction

Object tracking is a deep learning application that automatically identifies objects in video and accurately interprets their trajectory [1]. More specifically, human motion tracking (HMT) entails tracking a human's movement or trajectory, i.e. the path taken by an object moving under the influence of forces. Motion tracking began in the 1970s and 1980s as a tool for photogrammetric analysis in biomechanics research. However, as technology progressed, it quickly spread to other fields such as education, training, sports, and, more recently, computer animation for television, cinema, and video games [2].

Various crimes have necessitated the use of security and protection systems to keep people and their property safe. Unfortunately, since they are so easily broken, locks and bolts do little to deter intruders and burglars [3]. However, IoT schemes, biometric identification technologies, and tracking technologies all aid in ensuring the protection of people and properties. Due to a wide range of potential applications, IoT schemes have recently aroused the interest of many researchers.

The proper estimation or prediction of the state of a linear dynamic system can be achieved by using the Kalman filter, which is regarded as one of the most efficient object tracking approaches. Prior knowledge of the process and measurement model is required for the Kalman filter. The Kalman filter can estimate the target's position, velocity, or acceleration. However, knowing the model's initial statistics with moving targets and noisy environments is challenging. In such situations, meta-heuristic optimisation techniques can improve the KF's performance [4]. The KF problem typically involves estimating a state vector for a dynamic system represented by a linear model. Kalman filtering is strengthened further if the model is nonlinear and performed through a linearisation procedure. This other filter is identified as extended Kalman filter (EKF). The EKF is the nonlinear counterpart to the Kalman filter, linearising the mean and covariance estimate [5].

Estimating the covariance matrix is an essential aspect of tuning the Kalman filter. However, the intended value usually differs from the actual value due to poor tuning. As a result, error optimisation can be utilised to fine-tune the KF. PSO is employed to optimise errors in this study. KF and EKF were employed to examine and investigate the sets of data gathered from the IoT-based monitoring system set up as the testbed for this study via MATLAB and PSO to tune the process noise and measurement error involved in the filters. The swarm particles use their individual and collective experience to determine the ideal covariance's matrix value, which leads to the optimal global value. Compared to previous algorithms, this one is more efficient and accurate. Still, performing the task takes less time [6].

The remaining part of the study is structured in the following manner: Sect. 2 investigates and explores past similar studies, and Sect. 3 outlines the study's proposed approach. Section 4 discusses the results and findings of the MATLAB analyses. Finally, Sect. 5 concludes by summarising the findings.

2 Related Works

Mathematically, the KF is a fascinating and iterative technique. It uses a series of formulae and a sequence of inputs in succession to estimate the object's actual/true value, position, velocity, etc., each of which may be subject to unexpected or random error, ambiguity, or variation. While T. N. Thiele and P. Swerling had previously devised a comparable method, Rudolf E. Kalman is thought to have invented the KF in 1960. Swerling had previously created a nearly equivalent algorithm [7–9]. In honour of its inventor, R. S. Bucy, the method is sometimes addressed as Kalman-Bucy filtering. Schmidt is generally praised for developing the first Kalman filter implementation. From radar to estimating macroeconomic models, it is employed in a wide range of engineering and economic applications. The KF has continuously found applications in the field of control theory and control systems engineering. The KF is perhaps one of the most famous algorithms of the twentieth century, and it has long been thought to be the optimal solution for several tracking and data prediction applications [10]. In addition, it has been extensively documented for its application in visual motion analysis.

Some past studies that have used IoT schemes to deploy either the KF or EKF algorithm are presented here. An inexpensive air quality monitoring and prediction system based on IoT and edge computing were proposed by Lai et al. (2019). They used a powerful Raspberry Pi as an edge device to perform the KF technique, which increased the accuracy of inexpensive sensors by 27% [11]. Wang et al. (2018) proposed a plan for protecting user privacy on IoT cloud platforms based on unscented Kalman filters. They achieved a noteworthy feat by ensuring that the offered data was helpful while maintaining individual privacy data [12]. Huang et al. (2019) implemented a KF-based technique to solve two IoT challenges [13].

According to the study carried out by Adardour et al. (2020), using a KF-based on IoT, Alzheimer's patients can now be monitored in real time through the use of a computer, Android/iOS smartphone, or other IoT-enabled devices to improve their quality of life and alleviate the burden on caregivers and nurses. Alzheimer's is a slow-developing degenerative illness. This sickness keeps patients homebound [14]. In another related study, Kulkarni et al. (2021) developed a neural network with an extended Kalman filter to identify intrusions in IoT networks. Accuracy, detection rate, and false-negative rates were utilised to assess their system. Their work yielded overwhelmingly positive results [15].

3 Proposed Approach

The proposed methodology for HMT is discussed here. Following are the key stages: IoT design implementation, Kalman filter formulation, extended Kalman filter formulation, and the concept of PSO and its formulation.

3.1 Implementation of IoT-Based Design

The IoT-based testbed for monitoring the HTM and subsequent data collection in this study (still ongoing) was discussed in detail and published in [16]. Ten distinct individuals' data was gathered, totalling 100 samples for each. The IoT architecture entails interfacing both the software's communication module and the hardware's interface module.

3.2 Kalman Filter Formulation

The computation of the position of the targeted object can be represented by Eq. 1 using Newton's motion equations

$$x = x_0 + v_0\Delta t + \frac{1}{2}a\Delta t^2 \quad (1)$$

where $x = \text{targeted object's position}$, $x_0 = \text{targeted object's initial position}$, $v_0 = \text{targeted object's initial velocity}$, $a = \text{targeted object's acceleration}$, $\Delta t = \text{time interval}$. This paper considers a human as the targeted object.

It is common practice to use the KF approach to determine the actual value of an object measured or tracked using a multi-dimensional model depicted in Fig. 1. The initial state contains the state matrix, $X_k(k = 0)$, which is a function of position or velocity, and the process covariance matrix, P_k which represents the error in the estimate or process. A one-dimensional or multi-dimensional vector matrix could be used to depict the state matrix.

Table 1 summarises the Kalman filter formulation, which had been discussed in [3].

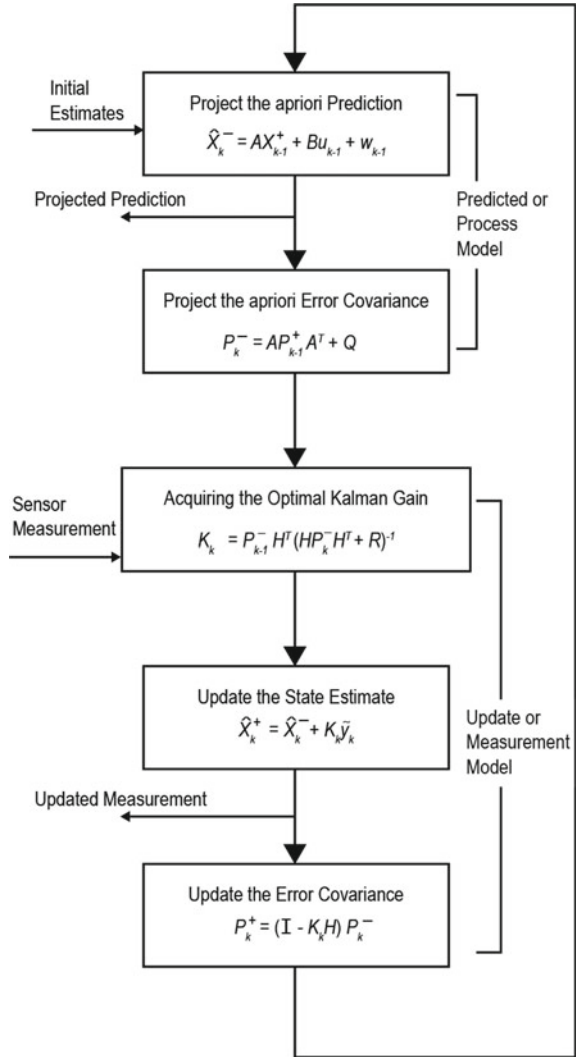
The Kalman filter estimates X_k at time k using the input information from A , B , H , Q , and R . Q and R are often tuning parameters to attain the optimum results. KF is typically portrayed in two phases or stages: the prediction stage or Propagation and the update stage or Correction, which are also identified as Propagation and Correction in different contexts. The hat operator, $\hat{\cdot}$, and superscripts (i.e. $-$ and $+$) denote the estimate of a variable, predicted and updated, respectively.

3.3 Extended Kalman Filter Formulation

Unlike the linear process and measurement models of the KF, the EKF consists of nonlinear models that are functions of other parameters. The following state transition and measurement models can be considered for EKF:

$$X_k = f(X_{k-1}, u_{k-1}) + w_{k-1} \quad (2)$$

Fig. 1 Kalman filter flowchart for multi-dimensional model



$$y_k = h(X_k) + v_k \tag{3}$$

where $X_k \in \mathcal{R}^{n \times 1}$ is the state vector, $u_k \in \mathcal{R}^{r \times 1}$ is the control vector, $y_k \in \mathcal{R}^{l \times 1}$ is the measurement vector, $w_k \in \mathcal{R}^{n \times 1}$ and $v_k \in \mathcal{R}^{l \times 1}$ are the zero-mean white Gaussian noises with covariances Q and R for the process and measurement models. The previous state, X_{k-1} and the control input, u_{k-1} are related or connected by the nonlinear process function f to generate the current state, X_k . On the other hand, the nonlinear measurement function h is a function that connects or relates the current state, X_k to the measurement, y_k . The next step is to compute, for each model in each

Table 1 Summary of computations in the prediction and update phases of KF

Prediction or propagation stage	Predicted state estimate	$\hat{X}_k^- = A\hat{X}_{k-1}^+ + Bu_{k-1} + w_{k-1}$
	Predicted process error covariance	$P_k^- = AP_{k-1}^+A^T + Q$
Update or correction stage	Measurement residual	$\tilde{y}_k = z_k - H\hat{X}_k^-$
	Kalman gain	$K_k = P_k^-H^T(H P_k^-H^T + R)^{-1}$
	Updated state estimate	$\hat{X}_k^+ = \hat{X}_k^- + K_k\tilde{y}_k$
	Updated process error covariance	$P_k^+ = (I - K_kH)P_k^-$

time step, the Jacobian matrix, which is the first-order derivative of a vector function with respect to a vector as follows:

$$F_{k-1} = \left. \frac{\partial f}{\partial x} \right|_{\hat{X}_{k-1}^+, u_{k-1}} \tag{4}$$

$$H_k = \left. \frac{\partial h}{\partial x} \right|_{\hat{X}_k^-} \tag{5}$$

The models become linearised about the current estimate due to the procedures stated in Eqs. (4) and (5). With this, the EKF algorithm becomes identical to that of the Kalman filter. The hat operator, $\hat{}$, and superscripts $-$, and $+$ follow the same definition as in KF. The EKF only differs from the KF in that it uses nonlinear functions $f(X_{k-1}, u_{k-1})$ and $h(X_k)$ to determine the predicted state estimate and predicted measurement. Figure 2 shows the extended Kalman filter flowchart.

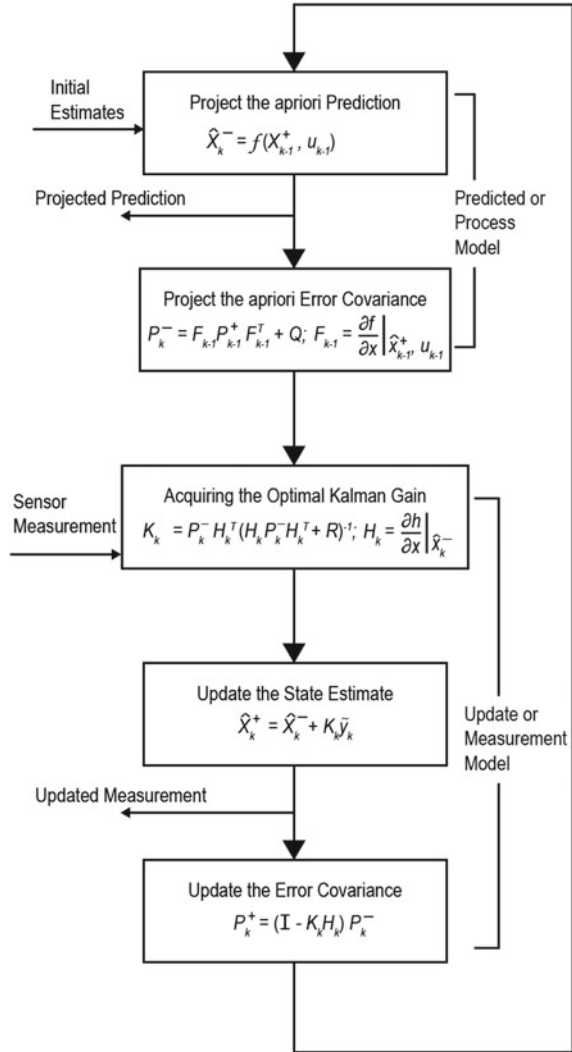
3.4 Particle Swarm Optimisation (PSO)

PSO is a stochastic population-based method. It is derived from natural occurrences such as the flocking of birds and the schooling of fish. PSO employs a collection of constraints and potential solutions to tackle the problem and achieve the best outcomes. PSO has been previously discussed extensively and published in [3]. However, only its formulation will be highlighted.

In order to find the best solution using PSO, it is necessary to use velocity and position updates to explore and exploit the search space. Therefore, Eqs. (6) and (7) are used to update the i^{th} particle's position and corresponding velocity, respectively:

$$x_i^{(t+1)} = x_i^{(t)} + v_i^{(t+1)} \tag{6}$$

Fig.2 Extended Kalman filter flowchart

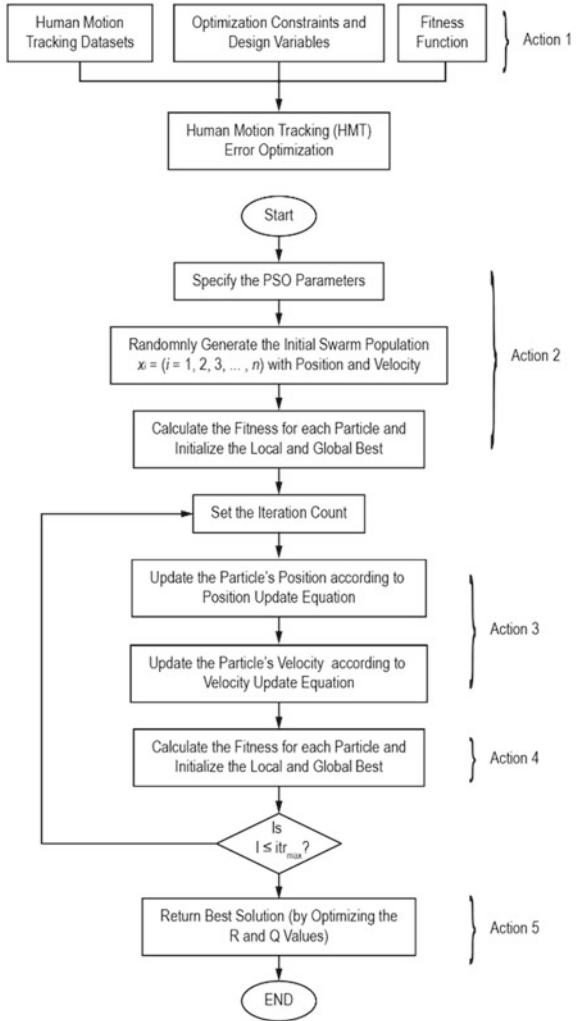


$$v_i^{(t+1)} = wv_i^{(t)} + C_1 r_1 (P_{(i,lb)}^{(t)} - x_i^{(t)}) + C_2 r_2 (P_{gb}^{(t)} - x_i^{(t)}) \tag{7}$$

where $i = \text{ith particle}$, $t = \text{generation counter}$, $v_i^{(0)}$ is randomly set, $w = \text{inertia weight}$, C_1 and $C_2 = \text{acceleration coefficients}$, $P_{(i,lb)}^{(t)} = \text{local best}$, $P_{gb}^{(t)} = \text{global best}$, and r_1 and $r_2 = \text{random numbers within } [0, 1]$.

The PSO's objective function was the Euclidean distance function of Eq. (10), in which u and v represent the reference coordinates that can be selected arbitrarily from the coordinate field (0:25). The flowchart of the PSO is depicted in Fig. 3.

Fig. 3 Flowchart of the PSO



$$f(u, v) = (u - 20)^2 + (v - 10)^2 \tag{8}$$

The proposed KF algorithm optimised with PSO (KF-PSO) and the extended Kalman filter algorithm optimised with PSO (EKF-PSO) is shown in Fig. 4a, b, respectively. The performance comparison was carried out by evaluating the implemented algorithms (i.e. models) regarding the position error using the accuracy analysis of average mean square error (Avg. MSErr).

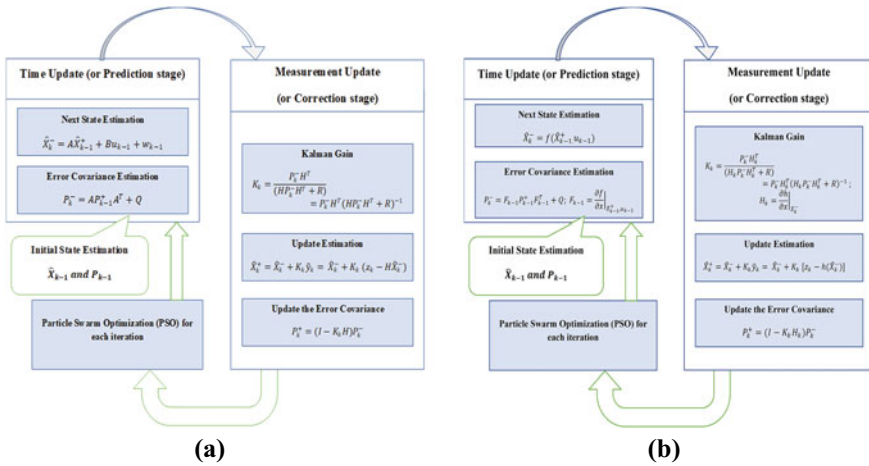


Fig. 4 Flowchart of **a** Kalman filter (KF) with PSO (KF-PSO) model and **b** extended Kalman filter with PSO (EKF-PSO)

4 Results and Discussion

This section examines the MATLAB simulation results and compares the quality performance and accuracy of the four developed models, namely conventional Kalman filter (KF only), extended Kalman filter (EKF only), proposed KF incorporated with PSO (i.e. KF-PSO), and EKF incorporated with PSO (i.e. EKF-PSO) models. For simplicity and clarity, just the results for the first human from the ten human datasets examined are shown.

Figure 5a, b depicts graphs of the actual (or measured) position vs the estimated (or predicted) position for KF tracking alone without PSO for Human 1 motion along the horizontal and vertical directions, respectively. Again, actual positions are displayed in black, whereas predicted positions are represented in blue. In addition, the per cent error (in red) is depicted using a different scale on the right side of these graphs. In contrast, Fig. 6a, b illustrates the actual position vs the predicted position for the optimised KF-PSO model along the horizontal and vertical directions, respectively. Here, the actual positions are depicted in black, while the predicted positions are in green. The left-hand side scale represents this. Furthermore, the per cent error (in red) is depicted with a separate scale on the right side. Similarly, Fig. 7a, b shows the graph of the actual position vs the predicted position for solely EKF tracking without PSO for Human 1 motion along the horizontal and vertical directions, respectively. Also, the actual positions are displayed in black, the predicted positions are in green, and the per cent error remains depicted in red.

tracking with KF-PSO prediction in the vertical direction.

In Fig. 6a, b, the positional motion of Human 1 along the horizontal and vertical axes is depicted using EKF-PSO tracking. In each of Figs. 5a, b, 6a, b, 7a, b and 8a, b, it was seen that the graphs of the actual position vs the predicted position overlapped

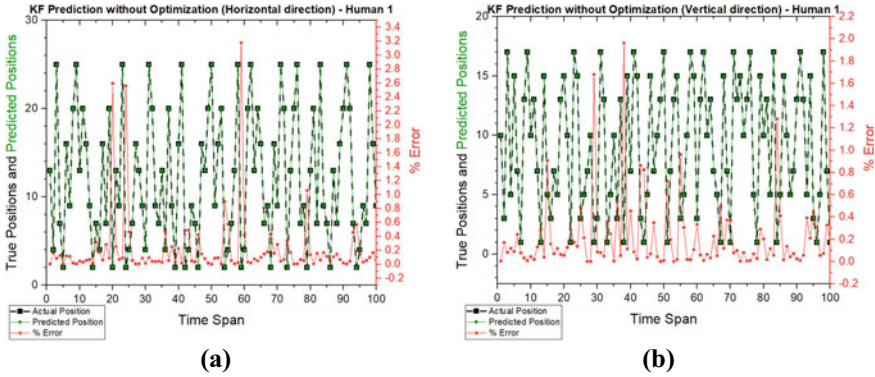


Fig. 5 **a** Actual/predicted positions of Human 1 tracking with KF prediction in the horizontal direction. **b** Actual/predicted positions of Human 1 tracking with KF prediction in the vertical direction

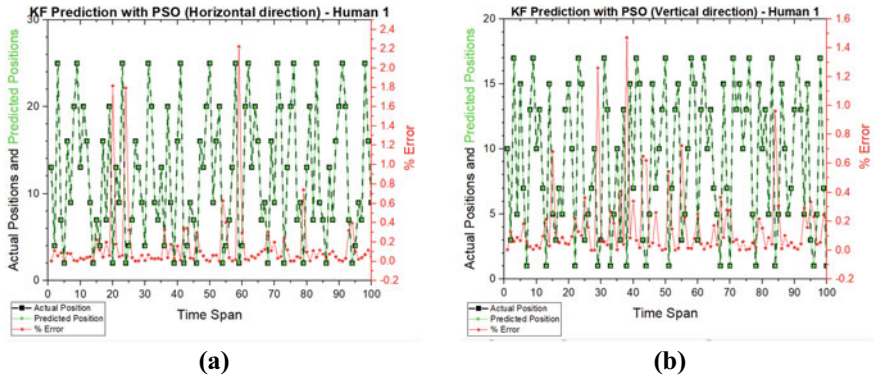


Fig. 6 **a** Actual/predicted positions of Human 1 tracking with KF-PSO prediction in the horizontal direction. **b** Actual/predicted positions of Human 1

because the curves of the two variables were in close proximity to one another. Consequently, the per cent error curve displays the per cent difference between the exact and predicted positions. This demonstrates that the four models can accurately predict the tracked person’s subsequent position estimates. Finally, mean square error (MSErr) analysis was used to compare the four models’ performances, as depicted in Figs. 9a, b.

The average MSErr for the KF model was 22.4 mm or 22.4×10^{-3} m (i.e. 2.24% error), while the average MSErr for the KF-PSO model was 17.2 mm or 17.2×10^{-3} m (i.e. 1.72% error). Equally, the average MSErr for the EKF model was 7.18 mm or 7.18×10^{-3} m (i.e. 0.718% error), while the average MSErr for the EKF-PSO model was 5.99 mm or 5.99×10^{-3} m (i.e. 0.599% error). This means that the EKF-PSO model is the most accurate (in terms of error) of the four models studied. The

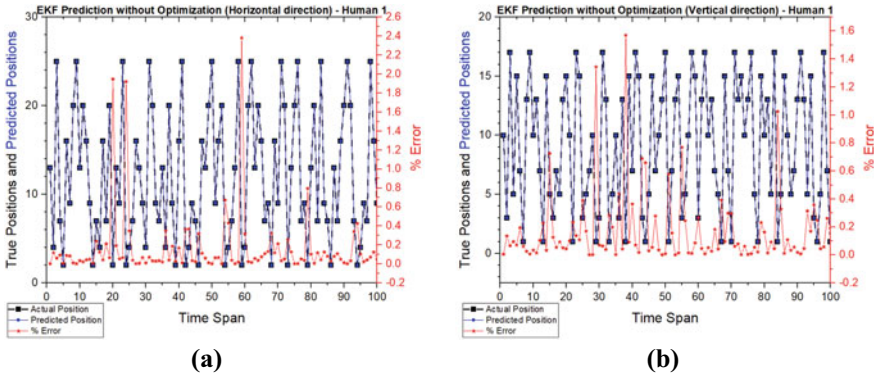


Fig. 7 **a** Actual/predicted positions of Human 1 tracking with EKF prediction in the horizontal direction. **b** Actual/predicted positions of Human 1 tracking with KF-PSO prediction in the vertical direction

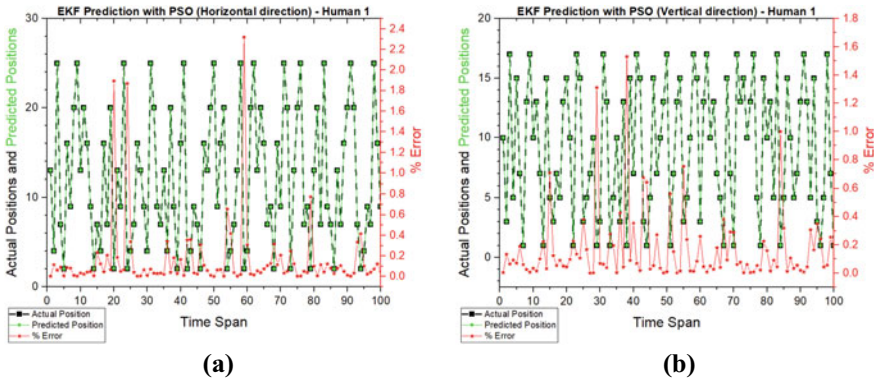


Fig. 8 **a** Actual/predicted positions of Human 1 tracking with EKF-PSO prediction in the horizontal direction. **b** Actual/predicted positions of Human 1 tracking with EKF-PSO prediction in the vertical direction

conventional EKF model with no optimisation is placed next to this one. Conversely, the conventional KF model is the most inaccurate. It was, therefore, possible to optimise both process noise and measurement error using PSO.

5 Conclusion

In this paper, four human tracking motion models of KF, EKF, KF-PSO, and KF-PSO have been developed and investigated. The data gathered from the IoT-based human motion tracking system were analysed using the four developed models. In

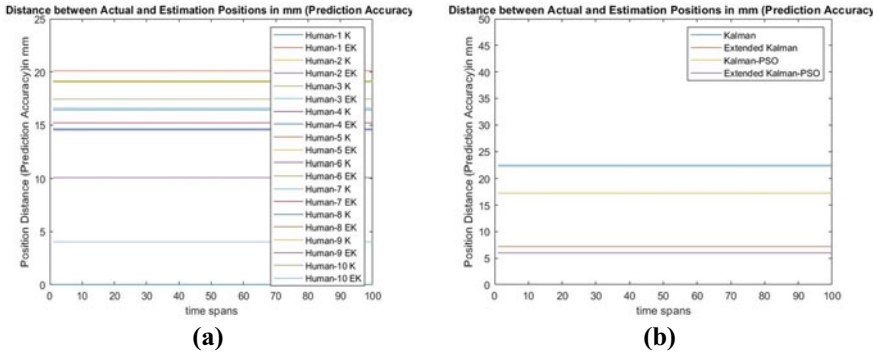


Fig. 9 The performance analysis of **a** all ten humans using KF and EKF models, **b** Avg. MSERR in EKF-PSO model (5.99×10^{-3} m), EKF model (7.18×10^{-3} m), KF-PSO model (17.2×10^{-3} m), and KF model (22.4×10^{-3} m)

addition, a performance analysis was carried out on the four models, and the EKF-PSO model showed the most remarkable accuracy, followed by the conventional EKF model. Next is the KF model, and the least accurate model is the conventional KF model. Subsequently, cuckoo search optimisation (CSO) will be implemented as the optimisation technique to benchmark other models already developed in our future work.

References

1. Athanesious JJ, Suresh P (2012) Systematic survey on object tracking methods in video. *Int J Adv Res Comput Eng Technol* 1(8):242–247
2. Sturm J, Engelhard N, Endres F, Burgard W, Cremers D (2012) A benchmark for the evaluation of RGB-D SLAM systems. In: *IEEE International conference on intelligent robots and systems*, pp 573–580. <https://doi.org/10.1109/IROS.2012.6385773>
3. Ajasa AA, Wahyudi NS (2022) Data-driven model for human tracking and prediction using Kalman filter with particle swarm optimization. In: Wahab NA, Mohamed Z (eds) *Control, instrumentation and mechatronics: theory and practice. Lecture notes in electrical engineering*, vol 921. Springer, Singapore, pp 478–490. https://doi.org/10.1007/978-981-19-3923-5_41
4. Chand BK, Khashirunnisa S, Kumari BL (2016) Performance analysis of Kalman filter with particle swarm optimization and fuzzy logic. In: *International conference on electrical, electronics, and optimisation techniques (ICEEOT)*, vol 3, no 8, pp 3204–3208
5. Akram MA, Liu P, Tahir MO, Ali W, Wang Y (2019) A state optimization model based on Kalman filtering and robust estimation theory for fusion of multi-source information in highly non-linear systems. *Sensors* MDIP 19(7). <https://doi.org/10.3390/s19071687>
6. Ramakoti N, Vinay A, Jatoh RK (2009) Particle swarm optimization aided Kalman filter for object tracking. In: *International conference on advances in computing, control and telecommunication technologies*, pp 531–533. <https://doi.org/10.1109/ACT.2009.135>
7. Kalman RE (1960) A new approach to linear filtering and prediction problems. *J Basic Eng* 35–45. <https://doi.org/10.4271/490207>
8. Lauritzen SL (1981) Time series analysis in 1880: a discussion of contributions made by T. N. Thiele. *Int Stat Rev* 49(3):319–331

9. Urrea C, Agramonte R (2021) Kalman filter: historical overview and review of its use in robotics 60 years after its creation. *J Sensors* 2021(1):1–21. <https://doi.org/10.1155/2021/9674015>
10. Humpherys J, West J (2010) Kalman filtering with Newton's method. *IEEE Control Syst* 30(6):101–106. <https://doi.org/10.1109/MCS.2010.938485>
11. Lai X, Yang T, Wang Z, Chen P (2019) IoT implementation of Kalman filter to improve accuracy of air quality monitoring and prediction. *Appl Sci* 9(9):1–23. <https://doi.org/10.3390/app9091831>
12. Wang J, Zhu R, Liu S (2018) A differentially private unscented Kalman filter for streaming data in IoT. *IEEE Access* 6:6487–6495. <https://doi.org/10.1109/ACCESS.2018.2797159>
13. Huang Y, Yu W, Ding E, Garcia-Ortiz A (2019) EPKF: energy efficient communication schemes based on Kalman filter for IoT. *IEEE Internet Things J* 6(4):6201–6211. <https://doi.org/10.1109/JIOT.2019.2900853>
14. Adardour HE, Hadjila M, Irid SMH, Baouch T, Belkhiter SE (2021) Outdoor Alzheimer's patients tracking using an IoT system and a Kalman filter estimator. *Wirel Pers Commun* 116(1):249–265. <https://doi.org/10.1007/s11277-020-07713-4>
15. Kulkarni DD, Rathore S, Jaiswal RK (2021) Intrusion detection system for IoT networks using neural networks with extended Kalman filter. In: *International conference on computer communications and networks, ICCCN*. <https://doi.org/10.1109/ICCCN52240.2021.9522335>
16. Ajasa AA, Nawawi SW, Abioye AE (2021) Design and development of IoT-based tracking for humans using Arduino. *J Electr Eng Elektr* 20(2):63–69

Stability and Bifurcation Analysis of Rössler System in Fractional Order



Ibrahim Mohammed Sulaiman, Abiodun Ezekiel Owoyemi,
Mohamad Arif Awang Nawi, Sadiya Salisu Muhammad, U. R. Muhammad,
Ali Fareed Jameel, and Mohd Kamal Mohd Nawawi

Abstract Rössler systems are introduced as prototype equations with the minimum ingredients for continuous time chaos. These systems are made up of three nonlinear ordinary differential equations that define a continuous-time dynamical system with chaotic dynamics due to the attractor's fractal features. Recently, the study on dynamics of fractional-order Rössler systems is attracting a lot of attention. In this study, the Rössler system of fractional order was numerically investigated. The existence, equilibrium points and their stability are also studied. The system is considered in the sense of Caputo fractional derivatives. In addition, an Adams-type predictor-corrector (ATPC) procedure is applied to the solutions of the system. The numerical result of the experiment shows that the system undergoes Hopf bifurcation for certain values. The result shows that selecting an appropriate value for the parameter can determine the stability of the region of our model. In conclusion, the study shows that the fractional order is very much stable than the integer order.

I. M. Sulaiman (✉) · M. K. M. Nawawi
Institute of Strategic Industrial Decision Modelling (ISIDM), School of Quantitative Sciences,
Universiti Utara Malaysia, UUM Sintok, 06010 Changlun, Kedah, Malaysia
e-mail: i.mohammed.sulaiman@uum.edu.my

A. E. Owoyemi
Department of General Studies, Federal College of Agricultural Produce Technology, Kano,
Nigeria

S. S. Muhammad
Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia

U. R. Muhammad
Department of Animal Health and Production, Federal College of Agricultural Produce
Technology, Kano, Nigeria

M. A. A. Nawi
School of Dental Sciences, Universiti Sains Malaysia, Health Campus, 16150 Kubang Kerian,
Kelantan, Malaysia

A. F. Jameel
Faculty of education and arts, Sohar university, 311 Sohar, Oman

Keywords Rössler system · Stability analysis · Bifurcation analysis · Fractional order

1 Introduction

Fractional calculus is an area of mathematics that has been in existence since 1695. Numerous investigation has shown that the fractional calculus has the superiority accuracy when describing several non-classical incidents in industrial applications and basic science, such as the biology [1] and finance [2] systems, compare to integer order. On contrary, the classical Riemann-Liouville fractional integral and famous Caputo fractional derivative (CFD) are the focus of numerous investigation in fractional calculus [3–6]. Over time, this subject has gained a lot of attention with many studies focusing on investigating the fractional calculus theory [7, 8], effective numerical techniques [9–11] and physical phenomena application [12]. In addition to that, the system stability region is increased using the fractional derivative, and this approach is more appropriate compared to the integer order [13–16]. One significant property of fractional derivatives is their nonlocal opposition of the integer derivatives local behavior [17].

In recent years, modeling the fractional-order differential equations (FODEs) has been considered by in many literatures compared to system of ordinary differential equations (ODEs). FODEs are generally associated to systems with memory which can be traced to several biological systems. Similarly, these equations are often related to fractals, relatively copious in biological systems [18]. According to a study by Cole [19], biological organism the cell membranes possess an electrical conductance in fractional order which are categorized in classes of non-integer-order models [20]. Fractional derivatives represent dynamic features of rheological cell behavior which has continuously gained utmost success in the area of rheology [17].

The above literature has motivated several researchers who studied different models for investigating fractional-order models dynamics. For instance, [17] investigated the fractional order of Zika virus infection and deduce that combining time delays and the fractional order in the epidemic model excellently enhances stability condition and strengthens dynamics of the model. The authors in [21, 22] studied the fractional-order immune systems of cancer and [23] investigated a HIV model in fractional order with nonlinear incidence and present a discussion on the stability the various equilibrium points. The dynamics of Ebola virus in fractional order was studied by [24] where their investigation shows that fractional order when combines with time delay can efficiently enhance and strengthen the stability criterion of the contagion model.

It would be interesting to note that fractional-order application of differential equations is justifiable in many cases because they often provide an efficient model when compare to the integer-order derivative models [25]. Also, differential equation in fractional-order sense provides a commanding instrument for integrating hereditary conditions and the memory of the systems in contrast to the integer-order models [17].

More so, unlike the integer-order model, fractional-order models possess a certain degree of freedom when fitting data [25]. Over time, they have been numerous studies on fractional-order models and systems [26, 27]. In a broader context, the compartmental models including the in-host virus dynamics, pharmacokinetics, epidemic models and Rossler systems were discussed in [28].

This study is more interested in the Rossler systems, which are the minimum components for chaos continuous time proposed in late seventeenth century [29]. These systems were first presented as prototype equations [30]. Some of the Rossler systems properties were reported to be deduced through linear technique including eigenvectors. However, the major characteristic of these systems involves nonlinear phenomena such as bifurcation diagram and the Poincaré maps. These dynamics phenomena have been continuously acknowledged as important for the standard biological system function. The systems are kept far from thermodynamics equilibrium. A study by [31] shows the presence of numerous variants of rhythmic phenomena in addition to the instances of independent biological oscillations including breath, circadian rhythms, ovarian cycle and heartbeat. The period of these phenomena ranges from seconds to certain hours or even days. The rhythmic processes indicating interval of 2–3 h such as growth hormone and luteinizing hormone, regulation of hormone and the function of neuron with insulin secretion are common and very important to life. In addition, enzyme production in some bacteria covers similar period [29, 31].

Understanding the function of biological rhythms requires the knowledge mathematical models and simulations. This would aid in comprehending the change from simple to more complex behavior. Based on the connection between biological rhythms and therapeutic perturbations of radiation and drugs that are linked to biological systems delivery, [29] investigated the theoretical model of Rössler system to chronotherapy and show that the systems are more vigorous to the perturbations whenever the Rössler is more dissipative. In [32], the authors generate a Hopf limit circle using a modify projective synchronization (MPS) control method whose shape and size are adjustable via selecting various scaling features in the classical Rössler systems. The MPS method has fewer computations. For more details on the Rossler systems, (see [31–33]). Currently, the subject of Hopf bifurcation is one of the active research areas in nonlinear systems and chaotic systems. Several researchers have been investigating the Hopf bifurcation of classical chaotic systems.

In the present paper, we investigated fractional-order Rössler system model in the sense of Caputo fractional derivatives. The paper further studied the existence, points of equilibrium and the asymptotic stability of the system.

Definition 1 [3] Fractional integral of the function $x(t)$, $t > 0$ with fractional-order $\beta \in \mathfrak{R}^+$ is defined as

$$I^\beta x(t) = \int_0^t \frac{(t-s)^{\beta-1}}{\Gamma(\beta)} x(s) ds \quad (1)$$

with the function of Euler's gamma denoted by $\Gamma(\cdot)$ and $t = t_0$ is the initial time.

Definition 2 For an order $\alpha \in (n - 1), n$ of $x(t), t > 0$, the Caputo fractional derivative is defined as.

$${}_c D_*^\alpha x(t) = I^{n-\alpha} D^n x(t), D_* = \frac{d}{dt}. \tag{2}$$

2 Rössler System

This system is very simple then but powerful. The simplicity of this model has led other scholars to advance the system in many areas. The application of the model is very relevant in modeling equilibrium in chemical reactions. The system is given as.

$$\frac{dx}{dt} = -(y + z) \tag{3a}$$

$$\frac{dy}{dt} = x + ay \tag{3b}$$

$$\frac{dz}{dt} = b + xz - cz \tag{3c}$$

where the state variables are represented as x, y, z , why a, b, c are the parameters.

Next, we replace the derivatives of Rössler system in Eqs. (3a)–(3c) by a fractional order by applying the Caputo derivatives of first order with $\alpha \in (0 - 1]$ as is given below

$${}_c D_t^\alpha x(t) = -(y + z) \tag{4a}$$

$${}_c D_t^\alpha y(t) = x + ay \tag{4b}$$

$${}_c D_t^\alpha z(t) = b + xz - cz \tag{4c}$$

where with the initial values

$$x(0) = x_0, y(0) = y_0, z(0) = z_0 \tag{5}$$

where $0 < \alpha \leq 1$, at any time, t .

3 Existence, Points of Equilibrium Points and Asymptotic Stability

For evaluation of the equilibrium points (EP), let

$${}_c D_t^\alpha x(t) = 0 \tag{6a}$$

$${}_c D_t^\alpha y(t) = 0 \tag{6b}$$

$${}_c D_t^\alpha z(t) = 0 \tag{6c}$$

The following theorem would be used in investigating the stability of the EP of system (6a–6c).

Theorem 1 [6] *For a commensurate fraction-order system of order*

$${}_c D_*^\alpha y(t) = f(t, y(t)), \tag{7a}$$

$$y(t_0) = y_0, \tag{7b}$$

where ${}_c D_*^\alpha$ is the CFD with order $\alpha \in (0, 1]$. Then, for fractional-order Caputo derivative to be local asymptotically stable, the necessary and sufficient conditions from (4a–4c) with $\alpha \in (0, 1]$ is, if and only if $|\arg \lambda_i| > \alpha \frac{\pi}{2}, i = 1, 2, 3$.

In order to evaluate EP, let put

$${}_c D_t^\alpha x(t) = f(t, y_i(t)), , y_i(t_0) = y_0 \tag{8}$$

For every $i = 1, 2, 3$, which can produce the EP $y_1^{eqn}, y_2^{eqn}, y_3^{eqn}$.

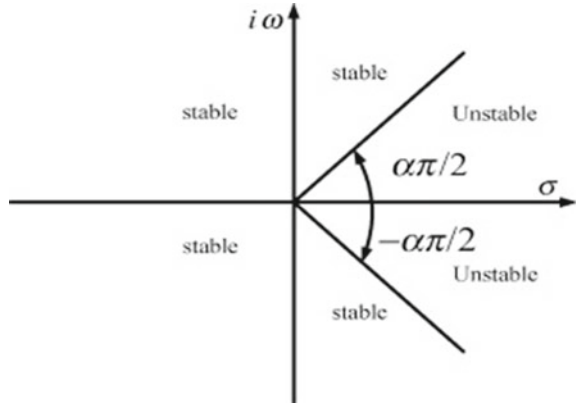
Next is evaluating the asymptotic stability, by considering the system ${}_c D_*^\alpha f(x) = f(x, y)$ in the Caputo sense, let $y_i(t) = y_i^{eqn} \varepsilon_i(t)$. The local asymptotic stability of equilibrium point $(y_1^{eqn}, y_2^{eqn}, y_3^{eqn})$ holds if the Jacobian eigenvalues

$$\left[\begin{matrix} m_{1,1} & m_{1,2} & m_{1,3} \\ m_{2,1} & m_{2,2} & m_{2,3} \\ m_{3,1} & m_{3,2} & m_{3,1} \end{matrix} \right], \text{ matrix A computed at EP is satisfy by } |\arg(\lambda_{1,2,3})| > \alpha \frac{\pi}{2}.$$

Figure 1 displays the fractional-order system stability region, illustrating that the efficiency of the region in the case of integer order compare to that of fractional order for $0 < \alpha \leq 1$. Thus, ω and σ represent the eigenvalue’s imaginary and real parts respectively where $i = \sqrt{-1}$.

Definition 3 The discriminant D(f) of a polynomial.

Fig. 1 Stability the region of the stability for fractional-order system



$f(x) = x^n + a_1x^{n-1} + a_2x^{n-2} + \dots + a_n$ is define by $D(f(x)) = (-1)^{\frac{1}{2}n(n-1)}R\left(f(x), \frac{dy}{dx}f(x)\right)$, where f^l is the derivative of f and where $g(x) = x^n + b_1x^{l-1} + b_2x^{l-2} + \dots + b_l$ and $R(f, g)$ is $(n+l)(n+1)$.

The idea of this paper is based on Routh-Hurwitz conditions in fractional order from [30], which was also considered in [34], [2]. The local asymptotic stability is said to hold if the following necessary and sufficient and necessary conditions are satisfied;

- (a) Assume $D(P) > 0$, the system (8) is satisfied for $a_1 > 0, a_3 > 0, a_1a_2 > a_3$,
- (b) Suppose $D(P) < 0$, (8) is satisfied for $a_1 > 0, a_2 > 0, a_1a_2 > a_3$
- (c) Suppose $D(P) < 0$, then (8) is satisfied for $0 \leq a_1, 0 \leq a_2, 0 \leq a_3$ and $\alpha < \frac{2}{3}$.

4 Numerical Stability Analysis and the Existence of Equilibrium Point

The numerical stability and the existence of the EP is performed in this section. The parameters; $a = 0.6, b = 0.2, c = 10$ are considered for model (4a)–(4c). The equilibrium points E , are

$$E_1(x_1, y_1, z_1) = 0.004001601281, -0.02000800641, 0.02000800641$$

and

$$E_2(x_2, y_2, z_2) = 9.995998399, -49.97999199, 49.97999199$$

The Jacobian matrix for the model is given as

$$J = \begin{vmatrix} 0 & -1 & -1 \\ 0 & 0.2 & 0 \\ z & 0 & x - 10 \end{vmatrix}$$

eigenvalues are $\lambda_{1,2} = 0.0990087305230765 \pm 0.9949850471999731i$, $\lambda_3 = -9.99401586004615 + 0.1i$ while fractional system characteristic equation based on (4a–4c) is $P(\lambda) = \lambda^3 + 9.79\lambda^2 - 0.9480\lambda + 9.9800$. Therefore, the argument $|\arg(\lambda_{1,2,3})|$ of matrix J at $\alpha = 0.8$ lies in the values range, 3.141592654. The values of λ_1 for $E_1(x_1, y_1, z_1)$ points are now stable, while the system is said to give the asymptotically stable if all the eigenvalues fulfill $|\arg(\lambda_i)| > \frac{\alpha\pi}{2}$. Whereas, after our evaluation, it was discovered that $|\arg(\lambda_i)| = 1.471615067$. Therefore, $|\arg(\lambda_i)| < \frac{\alpha\pi}{2}, 1.471615067 < 3.141592654 = \frac{\alpha\pi}{2}$, which make system (4a–4c) unstable.

It easy to show that $D(P) = -41724.37637 < 0$. As earlier mentioned, using the condition of Routh–Hurwitz, the stability properties of the nonlinear system (4a–4c) can be obtained. If $D(P) < 0$, then system (4a–4c) is satisfied for $a_1 > 0, a_2 > 0, a_1a_2 = a_3$. But in this case, though, $D(P) < 0$ but did not satisfy the second condition in our Definition 3, which says $a_1 > 0, a_2 > 0, a_1a_2 > a_3$. Here, $a_1a_2 < a_3$ and $a_1a_2 = a_3$, hence the point of equilibrium for the system is unstable. Figure 2 displays the unstable wave over a time, t , solution of the system (4a–4c) with the initial value $x = 0.01, y = 0.04, z = 0.04$.

We realize that by choosing appropriate value for the parameter can determine the stability of the region of a system. This is a great influence on the order of fractional-order system on the running state. We shall discuss more of that in the next section (Hopf bifurcation).

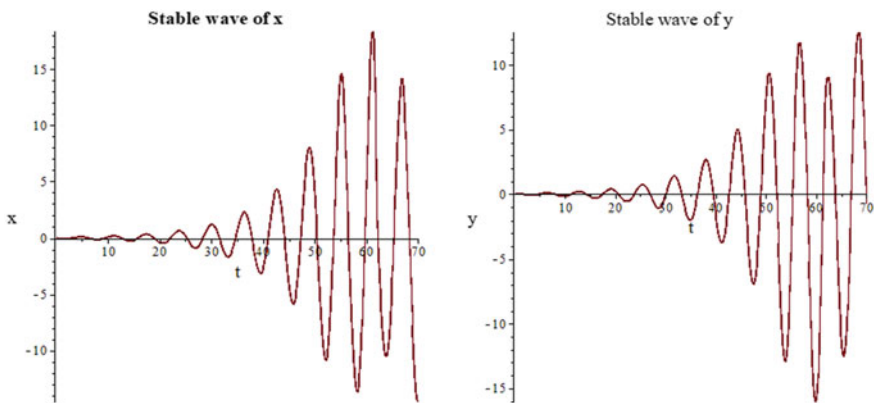


Fig. 2 Unstable wave of the system

5 Existence of Hopf Bifurcation in Fractional-Order α

In Sect. 3, we shown that our model was not stable. Therefore, in this section, we shall demonstrate how the unstable model can become stable via a Hopf bifurcation analysis.

This section begins by demonstrating the condition for which α is considered as a bifurcation parameter which is an important condition of the fractional-order system compare to integer order. The two major technique for processing the Hopf bifurcation are the successor function and Poincare Birkhoff canonical form methods [2]. This study adopts the successor function approach considered in [2, 34].

Theorem 1 Assume α^* is the system order parameter passing the Hopf bifurcation where $\alpha^* \in [0, 1]$. The existence of Hopf bifurcation surrounding the point of equilibrium system (4a–4c) holds if the conditions defined below are fulfilled.

- (a) Characteristic equations of the main system have a set of complex conjugate roots $\lambda_i = a + bi$ (where $a > 0$) whereas the rest of the eigenvalues are negative roots.
- (b) $m(\alpha^*) = \frac{\alpha^*\pi}{2} - \min|\arg(\lambda_i)| = 0$
- (c) $\frac{dm(\alpha^*)}{d\alpha}|\alpha = \alpha^* \neq 0$ (Transversality condition). The cycles appeared at $\alpha = \alpha^*$, $[\arg \lambda](\alpha^*) = \frac{\alpha^*\pi}{2}$.

The proof of this Theorem 1 is by successor function approach. The proof begins by assuming the equilibrium point of the approximation (linear) equations makes the trigonometric-substitution to the systems at origin with the expansion of the power series along particular direction based on the condition. Thereafter, we assess the successor function, then define the function in relation to periodic and origin solution.

Assuming a function $f^\alpha(x, \lambda)$ is a real analytic on an open set w such that $x \in w : \lambda(-\lambda_0, \lambda_0)$, so for $\frac{d^\alpha x}{dt^\alpha} = f(x, \lambda)$. The EP is $O(0, 0)$ for any λ , while the derivatives operator $D^\alpha f(0, \lambda)$ on $x = 0$ is represented as $A^\alpha(\lambda)$. Hence, conjugate complex set of numbers $a(\lambda) \pm bi(\lambda)$, $b(\lambda) > 0$ represents eigenvalues of $A^\alpha(\lambda)$.

Therefore, when $m_i(\alpha) < 0$, the equilibrium point is asymptotically stable, on the other hand, if $m_i(\alpha) > 0$, then, it is not stable anymore. This shows that the $m_i(\alpha)$ has similar effect with the eigenvalues real part of integer-order system. This implies that analytical function for sufficiently small x , $m_i(\alpha) = \frac{\alpha\pi}{2} - \min|\arg(\lambda_i)| = 0$ is one only. Suppose the system trajectory pass through $(x, 0)$ as a close orbit, which forms stable period.

Obviously, Theorem 1 (a) is fulfilled using the fractional-order property. The eigenvalues of fractional-order system and that of the integer-order system are similar. However, the existence of α^* fulfills the condition (b). The critical value α^* of bifurcation parameter interchanging the stability point order is now

$$\alpha^* = 2 \frac{\arctan \left| \frac{b(w)}{a(w)} \right|}{\pi}, \alpha^* \in (0, 1)$$

For the transversally condition given in Theorem 1(c), it is assumed that running state of the system $m(\alpha)$ can be affected by passing Hopf bifurcation parameter alpha through the critical value α^* . This demonstrates the dissimilarity of the equilibrium points of the system state in $(0, \alpha^*)$ and $(\alpha^*, 1)$, i.e., $O(0,0,0)$ is asymptotically stable of $\alpha \in (0, \alpha^*)$ and unstable when $\alpha^* < \alpha < 1$. Therefore, it can be deduced that Hopf bifurcation in (4a)–(4c) exist at $\alpha = \alpha^*$.

Based on the Hopf bifurcation condition of integer-order system and the discriminant criteria of the point of equilibrium stability of system of fractional-order system, we can obviously state that the fractional Hopf bifurcation existence criteria have been improved.

Next is to analysis the order satisfying the three properties presented by [35] for the fractional-order prey-predator system.

6 Existence of Hopf Bifurcation in Fractional-Order α

In this section, similar to previous Section, the ATPC method was employed for the simulation experiment. The parameters; $a = 0.6, b = 0.2, c = 10$ were considered as an example for model (4a)–(4c). The following eigenvalues were obtained by choosing α as the bifurcation parameter,

$$\lambda_{1,2} = 0.0990087305230765 \pm 0.9949850471999731$$

$$\lambda_3 = -9.99401586004615 + 0.1$$

Equally, we obtained the $D(P) = -41724.37637 < 0$ using Definition 3. From the above result, it is obvious there exist a pair imaginary conjugate eigenvalue, whereas the rest are negative real values. The characteristic equations of the main system have a set of complex conjugate roots $\lambda_{1,2} = a + bi$ (where $a > 0$) despite the fact that the rest eigenvalue, λ_3 is negative roots. Thus, Theorem 1(a) is fulfilled. From condition (b) in Theorem 1, critical value parameter of bifurcation is obtained as follows

$$\begin{aligned} \alpha^* &= 2 \frac{\arctan \left| \frac{b(w)}{a(w)} \right|}{\pi} = 2 \frac{\arctan \left| \frac{0.0990087305230765}{0.9949850471999731} \right|}{\pi} \\ &= 0.9368592487. \end{aligned}$$

We therefore have $\frac{dm(\alpha^*)}{d\alpha} |_{\alpha = \alpha^*} \neq 0$, which implies that the transversality condition Theorem 1(c) holds. For the fractional-order system, we say Hopf bifurcation occurs if α is selected as the bifurcation parameter after passing through the critical value $\alpha^* = 0.9$. So, the equilibrium of (8) attains its stability, whereas Hopf bifurcation ensues when α increases past $\alpha^* = 0.9$. This illustrates that an increase in the enrichment of a resource may cause a destabilization effect in prey predator.

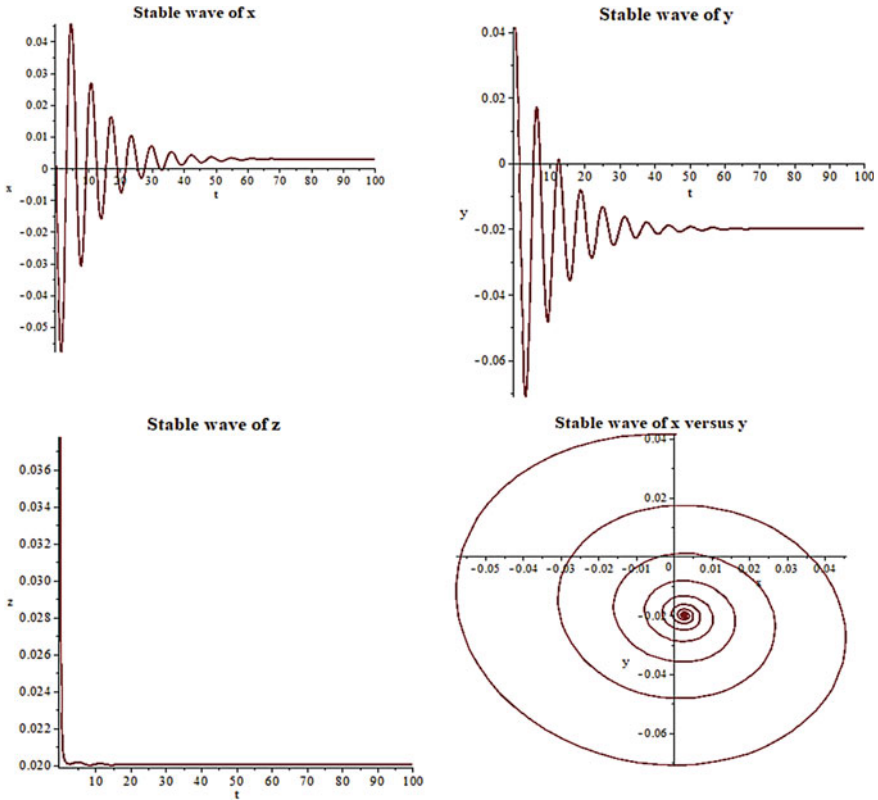


Fig. 3 Bifurcation occurs at when $\alpha = 0, 9$

Figure 3 shows the stability graph solution for (4a)–(4c) against t when α passes through critical value $\alpha^* = 0.9$. When α passes through the critical value $\alpha^* = 0.9$, it further presents the prey-predator phase portrait graph when α passes via critical value $\alpha^* = 0.9$.

7 Conclusion

In this paper, the fractional-order Rössler system has been presented with the dynamical behaviors analyzed. The system of the model was considered in Caputo fractional derivatives sense. By choosing an appropriate value for the parameter, we can determine the stability of the region of a model. By choosing appropriate parameter values, the region stability, existence, and equilibrium points of the model can be determined. The study also employs the Adams-type predictor-corrector scheme for the computational solutions of the models. The result obtained show that a Hopf bifurcation

occurs for some values at the equilibrium point in the model. Also, the simulation result shows that the fractional order is very much stable than the integer order.

Though this study investigated the fractional-order Rössler system, there are still many unknown dynamical behaviors of the biological systems that deserve to be further investigated. This can be achieved by modifying the biological models to produce Rossler type bands. Researchers in biological field can also consider the autonomous Rössler model as the dynamic of a rhythmic disease such as arthritis or asthma. More works are needed to extend the Rössler system model of fractional order to other areas, for example, the cases of perturbing frequency where the dominant frequency might diminish.

References

1. Rostamy D, Mottaghi E (2016) Stability analysis of a fractional-order epidemics model with multiple equilibriums. *Adv Differ Eq* 170. <https://doi.org/10.1186/s13662-016-0905-4>
2. Ma J, Ren W (2016) Complexity and Hopf bifurcation analysis on a kind of fractional-order IS-LM macroeconomic system. *Int J Bifurc Chaos*
3. Podlubny I (1999) Fractional differential equations. An introduction to fractional derivatives, fractional differential equations, some methods of their solution and some of their applications
4. Baleanu D, Tenreiro Machado JA, Cattani C, Baleanu MC, Yang XJ (2014) Local fractional variational iteration and decomposition methods for wave equation on cantor sets within local fractional operators. *Abstr Appl Anal*
5. Herrmann R (2011) Fractional calculus: an introduction for physicists
6. Carroll JE (2012) Fractional calculus: an introduction for physicists, by Richard Herrmann. *Contemp Phys*
7. Diethelm K, Ford NJ (2002) Analysis of fractional differential equations. *J Math Anal Appl*
8. Baleanu D, Rezapour S, Mohammadi H, Baleanu D, Rezapour S (2013) Some existence results on nonlinear fractional differential equations Some existence results on nonlinear fractional differential equations
9. Bhrawy AH, Taha TM, Machado JAT (2015) A review of operational matrices and spectral techniques for fractional calculus. *Nonlinear Dyn*
10. Al-Khaled K (2015) Numerical solution of time-fractional partial differential equations using sumudu decomposition method. *Rom J Phys*
11. Rahimkhani P, Ordokhani Y, Babolian E (2017) A new operational matrix based on Bernoulli wavelets for solving fractional delay differential equations. *Numer Algorithms*
12. Caputo M, Fabrizio M (2015) A new definition of fractional derivative without singular Kernel. *Progr Fract Differ Appl*
13. El-Saka HAA (2013) The fractional-order SIR and SIRS epidemic models with variable population size. *Math Sci Lett Math Sci Lett Int J* 2(3):195–200
14. Al-Salti N, Karimov E, Sadarangani K (2016) On a differential equation with Caputo-Fabrizio fractional derivative of order $1 < \beta \leq 2$ and application to mass-spring-damper system. *Progr Fract Differ Appl* 2(4):257–263
15. Abiodun EA, Sulaiman IM, Mamat M, Olowo SE, Adebisi OA (2020) Analytic numeric solution for coronavirus (covid–19) pandemic model in fractional-order. *Commun Math Biol Neurosci* 10(61):1–18
16. Owoyemi AE, Ibrahim SM, Muhammad SS (2021) Stability and Hopf bifurcation analysis of a biotic resource enrichment on a prey predator population in fractional-order system. *AIP Conf Proc* 2355(1):1–11. <https://doi.org/10.1063/5.0053307>

17. Rakkiyappan R, Latha VP, Fathalla AR (2019) A fractional-order model for Zika virus infection with multiple delays. *Complexity* 2019, Article ID 4178073, 20 p. <https://doi.org/10.1155/2019/4178073>
18. Rocco A, West BJ (1999) Fractional calculus and the evolution of fractal phenomena. *Phys A* 265(3):535–546
19. Cole KS (1993) Electric conductance of biological systems. *Cold spring harbor symposia on quantitative biology* pp 107–116
20. Rihan FA (2013) Numerical modeling of fractional-order biological systems. *Abstr Appl Anal* Article ID 816803, 11 p. <https://doi.org/10.1155/2013/816803>
21. Ahmed E, Hashish A, Rihan FA (2012) On fractional order cancer model. *J Fract Calcul Appl Anal* 3(2):1–6
22. Rihan FA, Lakshmanan S, Maurer H (2019) Optimal control of tumour-immune model with time-delay and immuno-chemotherapy. *Appl Math Comput* 353(7):147–165
23. Zhang L, Huang G, Liu A, Fan R (2015) Stability analysis for a fractional HIV infection model with nonlinear incidence. *Discr Dyn Nat Soc* 2015, Article ID 563127, 11 p
24. Latha VP, Rihan FA, Rakkiyappan R, Velmurugan G (2017) A fractional-order delay differential model for ebola infection and CD8⁺ T-cells response: stability analysis and hopf bifurcation. *Int J Biomath* 10(8), Article ID 1750111
25. Tateishi AA, Ribeiro HV, Lenzi EK (2017) The role of fractional time-derivative operators on anomalous diffusion. *Front Phys* 5(1–9)
26. Atangana A, Botha JF (2013) A generalized groundwater flow equation using the concept of variable-order derivative. *Bound Value Probl*
27. Ameen I, Novati P (2017) The solution of fractional order epidemic model by implicit Adams methods. *Appl Math Model*
28. Angstmann CN, Erickson AM, Henry BI, Mcgann AV, Murray JM, Nichols JA (2017) Fractional order compartment models. *SIAM J Appl Math*
29. Betancourt-Mar JA, Alarcón-Montelongo IS, Nieto-Villar JM (2005) The Rössler system as a model for chronotherapy. *J Phys Conf Ser* 23(1). IOP Publishing
30. Ahmed E, El-Sayed AMA, El-Saka HAA (2006) On some Routh-Hurwitz conditions for fractional order differential equations and their applications in Lorenz, Rössler, Chua and Chen systems. *Phys Lett Sect A Gen At Solid State Phys* 358(1):1–4
31. Hald BG, Laugesen CN, Nielsen C, Mosekilde E, Larsen ER, Engelbrecht J (1989) Rössler bands in economic and biological systems. In: Milling PM, Zahn EOK (eds) *Computer-based management of complex systems*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-74946-9_55
32. Wu R, Li X (2012) Hopf bifurcation analysis and anticontrol of Holf circles of the Rossler-like system. *Abstr Appl Anal* 2012, Article ID 341870, 16 p. <https://doi.org/10.1155/2012/341870>
33. Ibrahim KM, Jamal RK, Ali FH (2018) Chaotic behaviour of the Rossler model and its analysis by using bifurcations of limit cycles and chaotic attractors. *J Phys Conf Ser* 1003:012099
34. Li X, Wu R (2014) Hopf bifurcation analysis of a new commensurate fractional-order hyperchaotic system. *Nonlinear Dyn*
35. El-Saka HA, Ahmed E, Shehata MI, El-Sayed AMA (2009) On stability, persistence, and Hopf bifurcation in fractional order dynamical systems. *Nonlinear Dyn*

SUAS-Based NDVI and RGB Image for Remote Landscape and Environmental Monitoring on University Campus



Ahmad Anas Yusof, Mohd Faid Yahya, Mohd Khairi Mohamed Nor, and Muhammad Fahmi Miskon

Abstract This project measures NIR and visible spectrum reflectance on the landscape of the Universiti Teknikal Malaysia Melaka Main Campus in Durian Tunggal, Melaka. A small fixed-wing drone with a GoPro and a Multispectral Camera is utilized to provide an overview of the University's terrain as well as for in-situ environmental monitoring. The measurement is taken from a dedicated flight plan, focusing on the Chancellery building, Faculty of Electrical Engineering, Chancellor Hall, Mosque, Lectures Hall Complex and Co-curriculum Center. The results are described through detailed NDVI imagery, with comparison to RGB images. Reflection and absorption of light spectrum are the main influences that can modify the NDVI readings. This paper reviews and highlights various landscape properties with varying NDVI values, as well as discusses the implications of such differences.

Keywords SUAS · NIR spectrum · NDVI

1 Introduction

SUAS-based near-infrared (NIR) spectrum imaging is gaining popularity in landscape and environmental monitoring applications, and it is used extensively in measuring the Normalized Difference Vegetation Index (NDVI) for plant canopy and health assessment, as well as drought and rock slope monitoring. NDVI is one of the best indices for assessing plant health. It is a measurement based on how the object reflects radiation at different frequencies, with some waves are absorbed and others are reflected. In plant, the cellular structure of the leaves greatly reflects

A. A. Yusof (✉) · M. F. Yahya

Robotics and Industrial Automation Research Group, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
e-mail: anas@utem.edu.my

M. K. Mohamed Nor · M. F. Miskon

Rehabilitation Engineering and Assistive Technology Research Group, Faculty of Electrical Engineering, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

near-infrared spectrum, while the chlorophyll, a common health indicator, absorbs visible spectrum. A case study in Vietnam, on the landscape-ecological information for management of land use in the Mekong delta has been presented with the integration of NDVI map. The approach is highly effective and accurate for classifying land use in heterogeneous and highly intensive crop area [1]. Meanwhile, in Sudan, there have been a study in determining the relationship between rainfall and the NDVI value for drought monitoring [2]. SUAS-based NDVI image has also been conducted in United Kingdom and Malaysia, whereby a time-series NDVI data have been retrieved over a highly heterogeneous ecosystem during a period of spring green-up, and the latter is used for individual crown for oil palm trees [3, 4]. The same measurement concept has been used to produce an efficient technique for monitoring the stability of rock slopes and soil in Malaysia, with a focus on the use of remote sensing compact sensors from the visible, near-infrared and infrared thermography spectrum, respectively, [5, 6]. This paper provides an in-depth analysis using NDVI classification in measuring greeneries for remote landscape and environmental monitoring for Universiti Teknikal Malaysia Melaka Main Campus.

2 Methodology

Universiti Teknikal Malaysia Melaka is a public university located in 766 acres of land in Melaka, Malaysia, (2.3138° N, 102.3211° E). The university's main campus is located within the lush green landscape of Durian Tunggal, approximately 15 km from the Heritage City of Melaka. The SUAS-based NDVI data measurement is part of an integrated environmental monitoring effort by using aerial robotics application, coordinated by the Robotics and Industrial Automation Research Group, Faculty of Electrical Engineering. Figure 1 shows the designated flight plan and test locations.

The flights are carried out by using a modified Parrot Disco Drone from Parrot Inc. France that is equipped with a Go Pro Hero 5 and Mapir camera, as shown in Fig. 2. The cameras are placed underneath the drone and programmed to take multiple geo-tagged image during flight, allowing for the creation of orthographic and NDVI maps for later analysis. NIR and visible spectrum reflectance has been measured by using the SUAS-equipped NIR Camera.

NDVI compares red and NIR spectrum mathematically to help distinguish plant from non-plant and healthy plant from sick plant. It is the ratio of the difference between the near-infrared and the red spectrum and the sum of these two spectrums [7].

$$NDVI = \frac{NIR - RED}{NIR + RED} \quad (1)$$

Fig. 1 Test locations

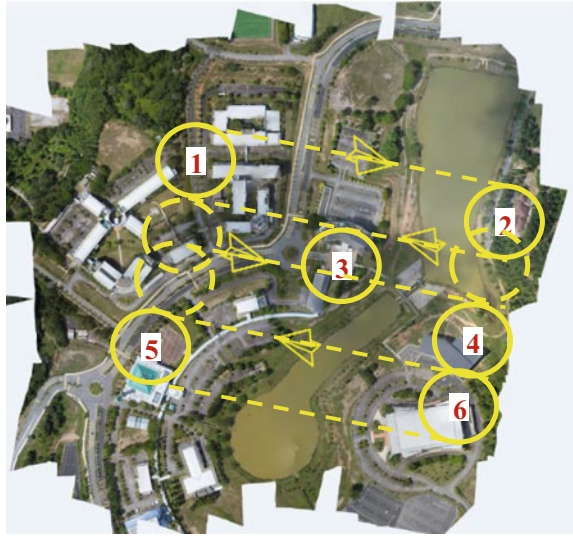
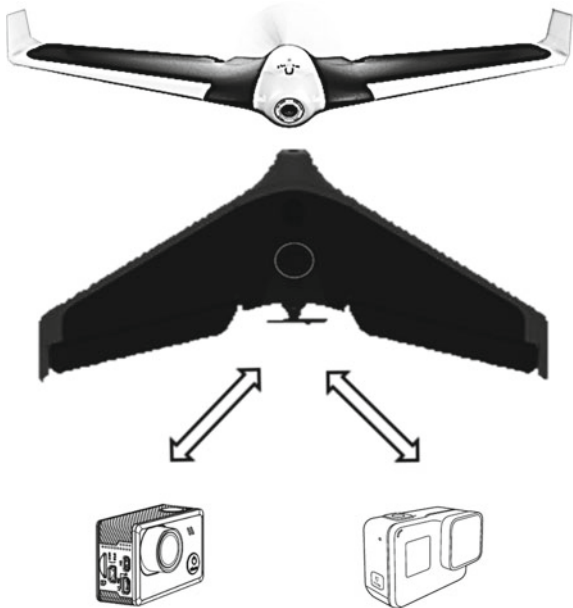


Fig. 2 Modified parrot Disco Drone with a GoPro and a Mapir camera



The value of the NDVI will always fall between -1 and $+1$. The reflectance of water is normally observed within the visible spectrum as incident radiation is completely absorbed within the NIR spectrum. The reflection of soil on the other hand, is influenced by factors such as moisture, texture and mineral composition. Soil that contains moisture will has low reflectance because it absorbs much of the incident radiation.

The moisture is strongly related to soil texture. Thus, fine soil particles that typically retain water will have low reflection compared to soil composed of coarse materials. At the same time, soil made up of iron oxide and organic matter will also reflect low radiation.

3 Results and Discussion

3.1 Test Site 1 (*Faculty of Electrical Engineering*)

NDVI measurements of the site provide NDVI value for the cloud above the horizon ranging from -0.67 to -1 , as shown in Fig. 3. The roof index for the Faculty of Electronic and Computer Engineering and the Faculty of Electrical Engineering is within the same range as well. The hill closest to the Residential College has the highest greenery, with values ranging from 0.67 to 1 . The NDVI for roads along the faculties ranges from -0.33 to -0.67 , with a few smaller roads falling between 0 and -0.33 .

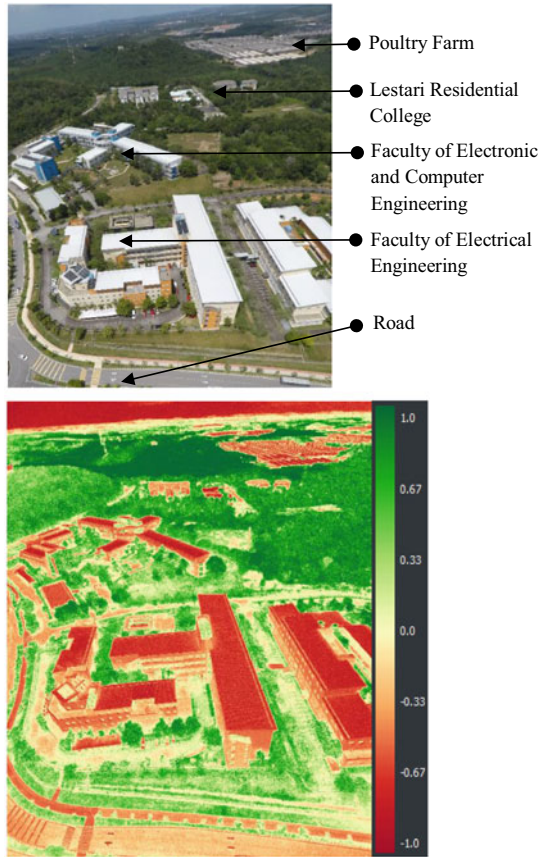
3.2 Test Site 2 (*Chancellor Hall*)

Figure 4 depicts the NDVI value for test site 2, which focuses on the Chancellor Hall and its surroundings. The road index around the hall ranged from -0.33 to -0.67 , while the roof index of the hall ranges from -0.67 to -1.0 , which coincides with the value of inorganic objects such as stones and roads. The yard is the darker green area on the right and left side of the hall, with values ranging from 0.67 to 1.0 , representing the most greenery and live green vegetation. Orange-yellow areas with values ranging from 0.0 to 0.33 , indicate bare soil or dead vegetation.

3.3 Test Site 3 (*The Chancellery*)

The NDVI value for test site 3, which includes the roundabout, the Chancellery, the Student Affairs Building, and a portion of the slope and lake, is depicted in Fig. 5. The top-of-the-image roundabout index ranged from 0.33 to 0.67 , while the surrounding roads index ranged from -0.33 to -0.67 , which corresponded to the value of inorganic items. The roof index for the Chancellery and Student Affairs Building is in the same range, from -0.67 to -1.0 . Greenery areas ranging from 0.67 to 1.0 may be found on the left side of the slope facing the lake, which is known as Lake *Terbilang*, whereas orange-yellowish areas with values ranging from

Fig. 3 RGB image (above) and NDVI data (below) for test site 1



0 to -0.67 can be found on the right side of the slope facing Lake *Gemilang*. The color of both lakes is red, ranging from -0.67 to -1.0 .

3.4 Test Site 4 (Lecture Halls Complex)

NDVI measurements of the test site 4 provide value for both Lake *Gemilang* and Lake *Terbilang*, ranging from -0.67 to -1.0 ., as shown in Fig. 6. The roof index for the Lecture Halls Complex ranges from 0.0 to -0.67 . The sediment going into the Lake *Gemilang* ranges from 0.0 to -0.33 , while the hill closest to the Lecture Halls Complex has the highest greenery, with values ranging from 0.67 to 1.

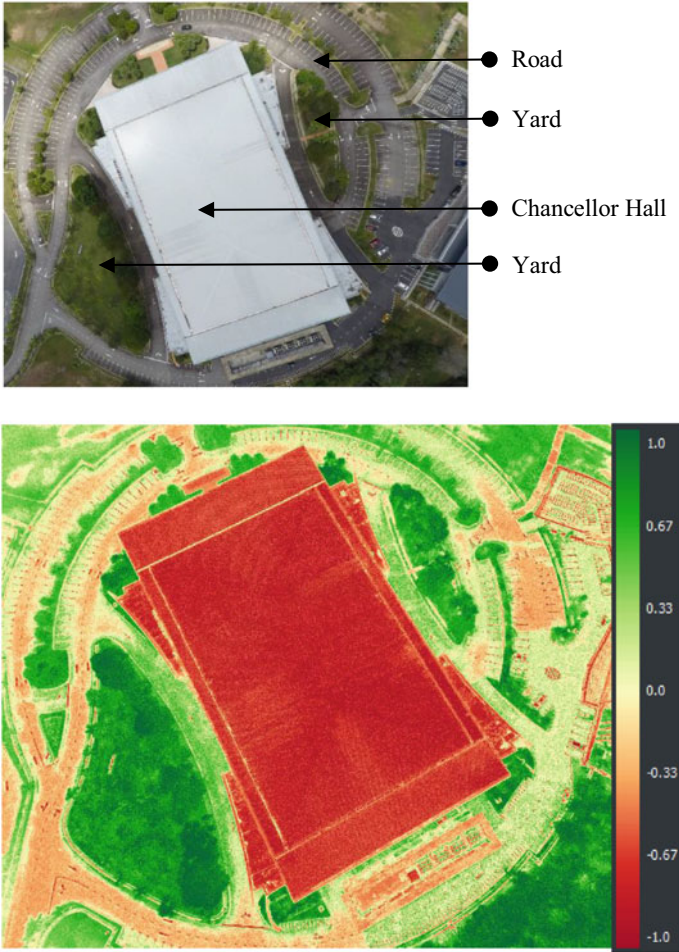


Fig. 4 RGB image (above) and NDVI data (below) for Test Site 2

3.5 Test Site 5 (Mosque)

Figure 7 depicts the NDVI value for Test Site 5, which includes the Mosque, Satria Residential College, UTeM Hill, Walkway and Lake *Gemilang*. The cloud on the image's horizon is represented by a red color that ranges from -0.67 to -1.0 . The roof of the Mosque, the staff cafeteria, the walkway and Lake *Terbilang* all have the same range. Greenery areas with values ranging from 0.67 to 1.0 can be found on a portion of the UTeM hill, the slope facing the lake and the hill surrounding the residential college. On the left side of the hill, the stabilized slope is an orange-yellowish area with values ranging from 0 to -0.67 .

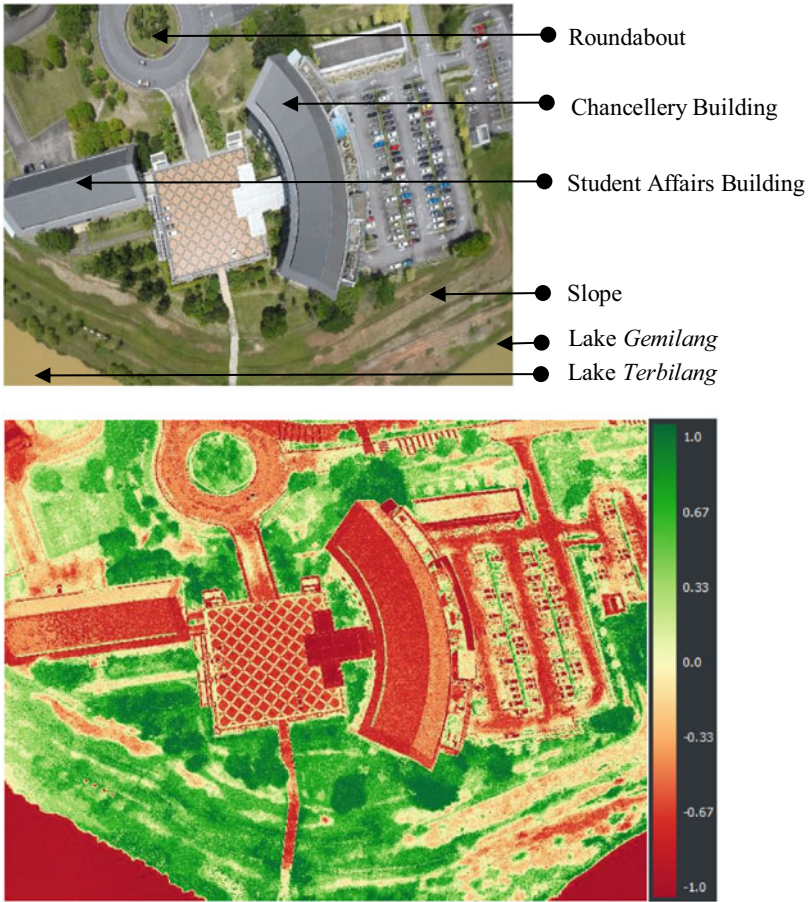


Fig. 5 RGB image (above) and NDVI data (below) for Test Site 3

3.6 Test Site 6 (Co-Curriculum Center)

The NDVI value for Test Site 6 is shown in Fig. 8, which includes the Co-curriculum Center and surrounding greenery, the trail behind the center, Lake *Gemilang* and sediment from a nearby drain. The greeneries area index ranges from 0.33 to 1.0, whereas the trail and the center’s roof have values ranging from 0 to -0.33 . The lake, on the other hand, represented with red color, that varies from -0.67 to -1.00 , while the sediment going into the lake ranges from 0.0 to -0.33 .

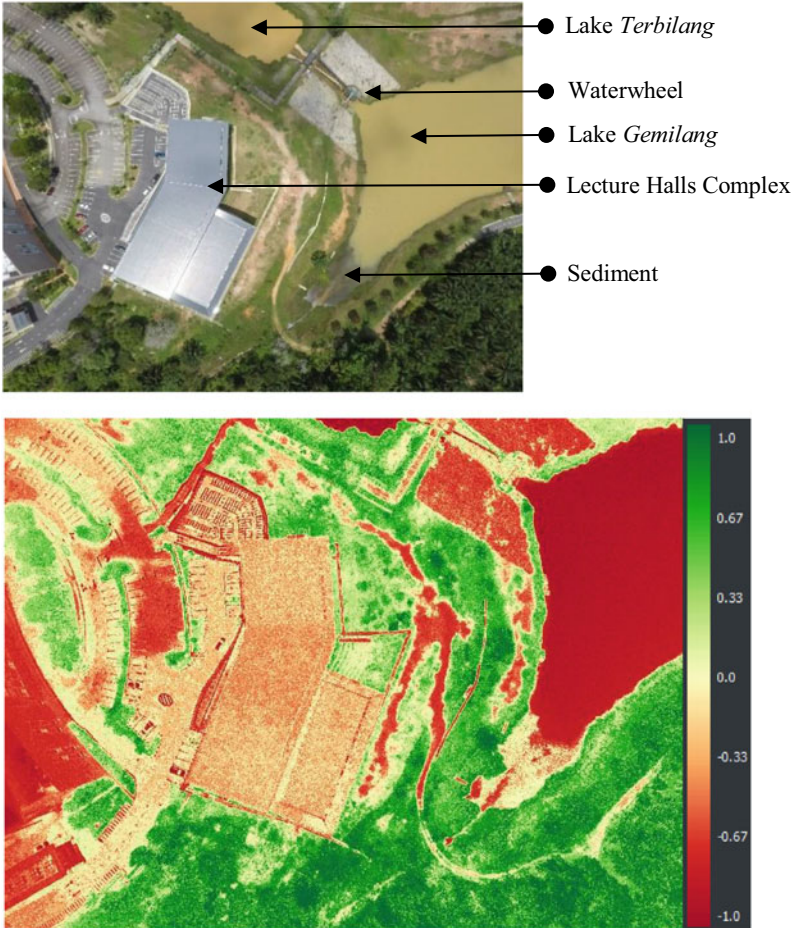


Fig. 6 RGB image (above) and NDVI data (below) for Test Site 4

4 Conclusion

This project demonstrates the concept of measuring NIR and visible spectrum reflectance by using Small Unmanned Aerial System (SUAS). The measurements are carried out by using modified flying wing drone, equipped with a GoPro and multispectral camera. The NDVI index has been used to identify the green areas within the campus. The thick tree canopy typically has positive NDVI values (0.67 to 1.0). Non-vegetative objects, on the other hand, have low NDVI values. Metallic roofing materials and water bodies such as lake exhibit poor NIR reflectance and as

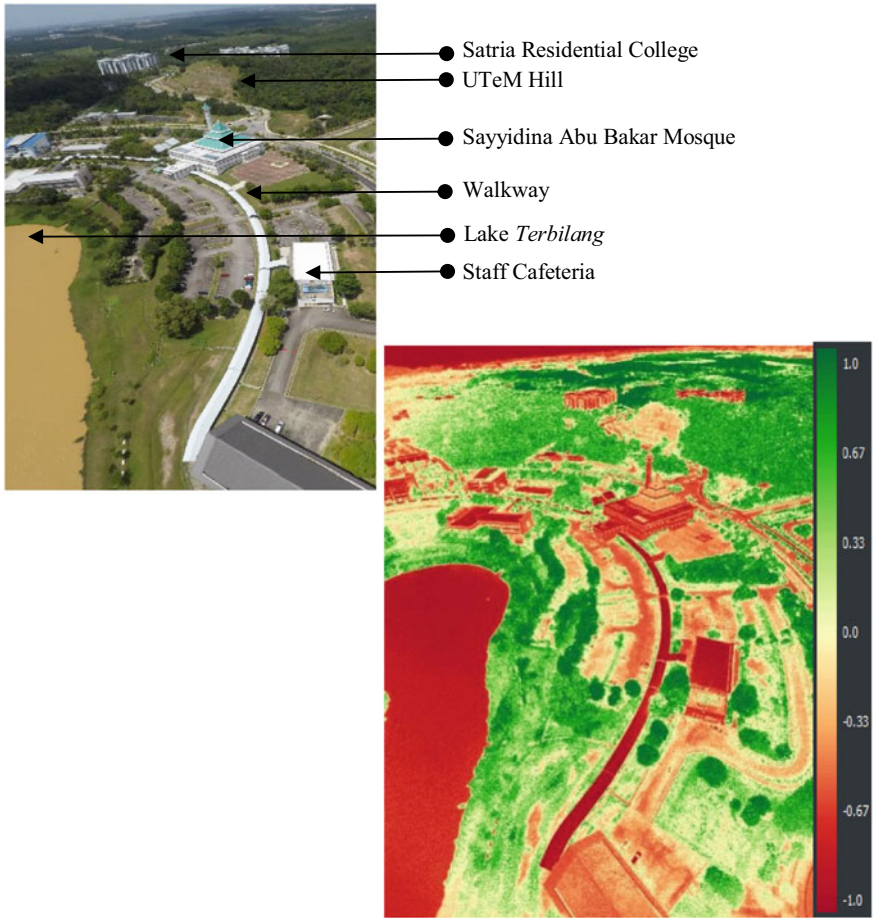


Fig. 7 RGB image (above) and NDVI data (below) for Test Site 5

a result, their NDVI readings are very negative, (-0.67 to -1.0), while water that is contaminated by sediment reflects a brownish color, (0.0 to -0.33). Rocks can be represented by orange-yellowish color (0 to -0.67), while bare soils frequently have tiny positive NDVI values (0.0 to 0.33).

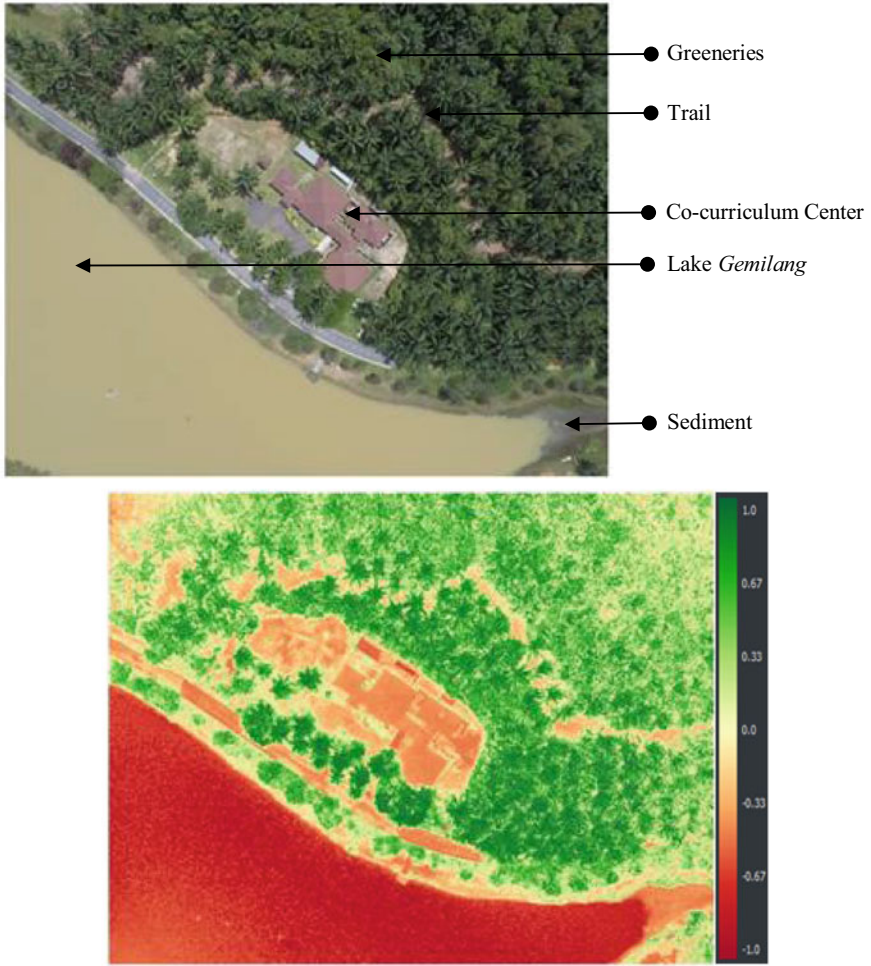


Fig. 8 RGB image (above) and NDVI data (below) for Test Site 6

Acknowledgements The authors wish to thank Universiti Teknikal Malaysia Melaka for their support in the research (PJP/2021/FKE/S01817).

References

1. Amjad A (2009) Comparison of strengths and weaknesses of NDVI and landscape-ecological mapping techniques for developing an integrated land use mapping approach. Master Thesis. International Institute for Geo-Information Science and Earth Observation (ITC)
2. Alemayehu K (1999) Drought risk monitoring for the Sudan using NDVI. Master Thesis. University College London
3. Fawcett D, Bennie J, Anderson K (2021) Monitoring spring phenology of individual tree crowns using drone-acquired NDVI data. *Remote Sens Ecol Conserv* 7(2):227–244
4. Suab SA, Syukur MS, Avtar R, Korom A (2019) The international archives of the photogrammetry, remote sensing and spatial information sciences, Volume XLII-4/W16. In: 6th international conference on geomatics and geospatial technology (GGT 2019), 1–3 October 2019, Kuala Lumpur, Malaysia
5. Yaacob MLM (2020) Rock slope monitoring using drone based multispectral and thermal images. *IOP Conf Series: Earth Environ Sci* 540(1):012024. <https://doi.org/10.1088/1755-1315/540/1/012024>
6. Yusof AA, Nor MKM, Mohd Shaari Azyze1 NLA, Kassim AM, Shamsudin SA, Sulaiman H, Hanafi MA (2022) Land clearing, preparation and drone monitoring using Red-Green-Blue (RGB) and thermal imagery for Smart Durian Orchard Management project. *J Adv Res Fluid Mech Thermal Sci* 91(1):115–128
7. Rouse J (1974) Monitoring vegetation systems in the great plains with ERTS. NASA Special Publication, 351, 309

A Mathematical Model of PD Controller-Based DC Motor System Using System Identification Approach



Nur Naajihah Ab Rahman and Nafrizuan Mat Yahya

Abstract A mathematical model is a crucial element of a system. This is to ensure the system obtains outstanding performance, particularly when there is a controller included. Thus, in this study, a comparison between DC motor PD controllers with and without system identification will be made with the concept of poles and zeros. Furthermore, the Cohen-Coon tuning method will be applied to tune the parameters of the proposed controller by using the MATLAB/Simulink software. Then, some tests were performed by varying the number of poles and zeros. After that, the performance of the DC motor with the proposed controller will be assessed in terms of transient response aspects. Throughout the study, it can be guaranteed that the process of system identification is needed to ensure that the performance of the DC motor can be enhanced. With that justification, the performance of the DC motor PD controller with two poles and no zero is better compared to the others. It had the shortest rise time of 0.052 s, the shortest settling time of 1.906 s, the shortest peak time of 1.142 s, and the lowest overshoot of 56.56 percent with no steady-state error.

Keywords Cohen-Coon tuning method · DC motor system · PD controller · Poles and zero · System identification

1 Introduction

Generally speaking, an electric motor is a device that uses electricity to create mechanical energy. Alternating current motors (AC motors) and direct current (DC motors) are the two main types of motors [1]. A wide variety of industrial and robotics applications are possible with a DC motor. It can be found in a wide range of control systems, including domestic electrical systems, automobiles, and process control

N. N. Ab Rahman (✉) · N. Mat Yahya
Faculty of Manufacturing and Mechatronic Engineering Technology, Universiti Malaysia Pahang,
26600 Pekan, Pahang, Malaysia
e-mail: nurnaajihahabraham@gmail.com

N. Mat Yahya
e-mail: nafrizuanmy@ump.edu.my

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
M. A. Abdullah et al. (eds.), *Advances in Intelligent Manufacturing and Mechatronics*,
Lecture Notes in Electrical Engineering 988,
https://doi.org/10.1007/978-981-19-8703-8_22

263

systems [2, 3]. It is widely used due to some reasons which are the maintenance ease to handle, reasonable price when it comes to changing a DC motor with a new one, and simple in terms of speed control [4]. Due to the mentioned benefits that have been provided by the DC motors, they can be used for many things that need variable speed as well as constant or low-speed torque.

Furthermore, a mathematical model of the DC motor system with precise parameters is important for generating an outstanding system response. Because, referring to [2], the authors stated that without correct parameters in the model of the DC motor itself, a mathematical model cannot portray how something should behave and it may not be the best mathematical model for the system. Thus, for the DC motor with a controller to have a good system response, an approach to deciding the transfer function of the DC motor may be implemented. Before that, a controller that has been chosen for this project is a proportional-derivative (PD) controller. It is a controller which combined proportional control with derivative control. According to [5], the trajectory of robotic manipulators as well as other mechatronic devices is frequently tracked by using the PD controllers. In addition, it is also has been used for electronics systems, measurement and sensor systems, biomechanics, and others. Surya and Singh in [6] declared that, without altering the steady-state parameters, the transient behavior may be enhanced by using the proposed controller.

The system identification toolbox together with the concept of poles and zeros which have been supplied in the MATLAB/Simulink software can be utilized for this project. As mentioned by [7], parameters and the properties of dynamic systems can be found by using specific identification methods which are very popular right now. It is possible to employ a variety of system identification approaches with the help of the transfer function models or process model functions in the system identification toolkit.

According to several papers in [8–10], the art and science of creating mathematical models of dynamic systems or calculating system transfer from observed input–output data are known as system identification (SI). Several phases need to be passed for the SI process to be successful. The data from real plants need to be collected then, formatted (using a selection model), processed (using a model), and lastly identified using the validation model [11]. The goal of making mathematical models of dynamical systems based on how the system is seen can be met with the recommended method. Other than that, the toolbox itself is useful for showing how dynamic systems work and makes sure that some black-box model structures are linear and nonlinear [12].

Therefore, the inaccurate mathematical model of DC motor with PD controller will affect the system response itself, which leads to the main purpose of this paper that focuses on the implementation of the mentioned technique, i.e., system identification in obtaining a near-exact model that will improve the system response. In addition, the controller will be used to have excellent performance compared to the DC motors alone. Thereafter, the performance of both DC motors with PD controllers undergoing or not undergoing the system identification process will be evaluated in terms of transient response aspects. By having an accurate mathematical model of the DC motor with the PD controller, itself, the best performance can be achieved from it.

2 Methodology

2.1 Mathematical Model of DC Motor

DC motor circuit which is divided into two parts: electrical and mechanical can be applied to construct the transfer function of the DC motor itself. The DC motor circuit can be illustrated in Fig. 1 where Kirchoff's voltage law (KVL) will be employed for the electrical element.

The electrical elements supplied as in the following figure are voltage source, armature resistance, armature inductance, armature current as well as the back emf voltage.

By implementing the KVL and mentioned parameters, the equation can be expressed as below.

$$E_a = R_a \cdot I_a + L_a \cdot (dI_a/dt) + E_b \tag{1}$$

where

I_a = armature current (A)

E_b = back emf voltage (V)

E_a = voltage source (V)

R_a = armature resistance (Ω)

L_a = armature inductance (H)

Then, the back emf, E_b directly proportional to the angular velocity of the shaft, θ by a constant factor, K_b may be stated as follows.

$$E_b = K_b(d\theta/dt) \tag{2}$$

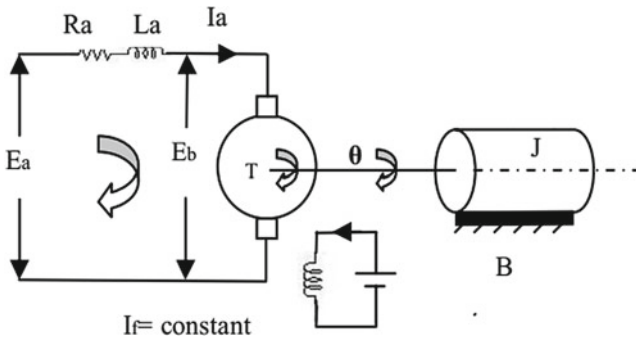


Fig. 1 DC motor circuit

For the mechanical part of the DC motor, it can be derived as below where it consists of the rotor moment of inertia, J_m , frictional coefficient known as B_m , and torque of motor which is T_m .

$$J_m(d^2\theta/dt^2) + B_m(d\theta/dt) = T_m \quad (3)$$

As for the motor's torque, it is proportional to the armature current, I_a by a constant ratio, K_t .

$$T_m = K_t \cdot I_a \quad (4)$$

After that, a mathematical method that is a substitution approach will be carried out to (1) and (2) generated from the preceding (3). By adding (2) into (1), a new equation will be generated, (5).

$$E_a = R_a \cdot I_a + L_a \cdot (d_{i_a}/d_t) + K_b(d\theta/dt) \quad (5)$$

Substitute (4) into (3) will build (6).

$$J_m(d^2\theta/dt^2) + B_m(d\theta/dt) = K_t \cdot I_a \quad (6)$$

Later, both Eqs. (5) and (6) may be delivered using the Laplace transform.

$$E_a(s) - K_b s \theta(s) = (R_a + sL_a) I_a(s) \quad (7)$$

$$J_m s^2 \theta(s) + B_m s \theta(s) = K_t I_a(s) \quad (8)$$

Lastly, put (8) in place of (7). The mathematical model from the voltage source, $E_a(s)$ to the output angle $\theta(s)$ then follows directly.

$$\theta(s)/E_a(s) = K_t / (J_m L_a s^2 + (J_m R_a + B_m L_a) s^2 + R_a B_m s + K_t K_b s) \quad (9)$$

Before proceeding to the stage of system identification, the parameters of the DC motor which have been used from [13] are as below, where transfer function (10) is expressed.

$$J_m = 0.093 \text{ kg.m}^2$$

$$B_m = 0.008 \text{ Nms}$$

$$K_b = 0.6 \text{ V/rad s}^{-1}$$

$$K_t = 0.7274 \text{ Nm/A}$$

$$R_a = 0.6\Omega$$

$$L_a = 0.006\text{H}$$

$$\theta(s)/E_a(s) = 0.7274 / (0.000558s^3 + 0.055848s^2 + 0.44124s) \quad (10)$$

The aforementioned (10) will be utilized to evaluate the performance of the DC motor. After conducting a simulation with the MATLAB/Simulink program, it will provide an output response. Changing the parameter input values will produce distinct system responses, such as a stable or unstable system. Simply expressed, a mathematical model developed from the DC motor circuit will impact system responsiveness.

2.2 System Identification

As referring to the objective of this work which is to utilize the use of system identification in obtaining the best mathematical model, the transfer function (10) that has been developed using the circuit of the DC motor will be used. In (10), the transfer function will be combined with the proposed controller. To tune the controllers' parameters, Cohen-Coon tuning method will be applied.

The first step in system identification is to measure the output signals when the system is loaded with an input signal. This can be done by using the block diagram of the DC motor with the PD controller in the MATLAB/Simulink software. Figure 2 is the input–output signals that will be exported to the workspace of MATLAB.

To export the data from the Simulink software, it can be done by using the 'To Workspace' function where it will be set as the input and output variables. The system identification is recognized by using the (10) in Fig. 3 and followed by the input and output signals in Fig. 4. This signal is produced from the procedure of SI in the figure below using the time plot tools, where u_1 is set to be the input variable and y_1 is the output variable. In the SI process, a variety of choices of the model to be estimated is provided such as transfer function model, nonlinear model, polynomial model, and others. Estimation in terms of transfer function will be chosen in this project.

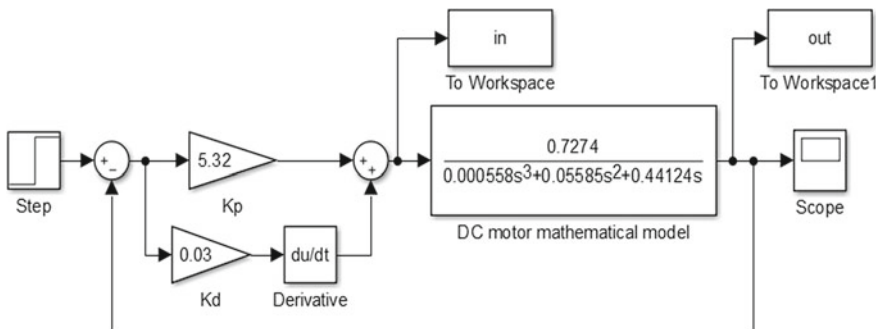


Fig. 2 Block diagram of DC motor with PD controller for system identification

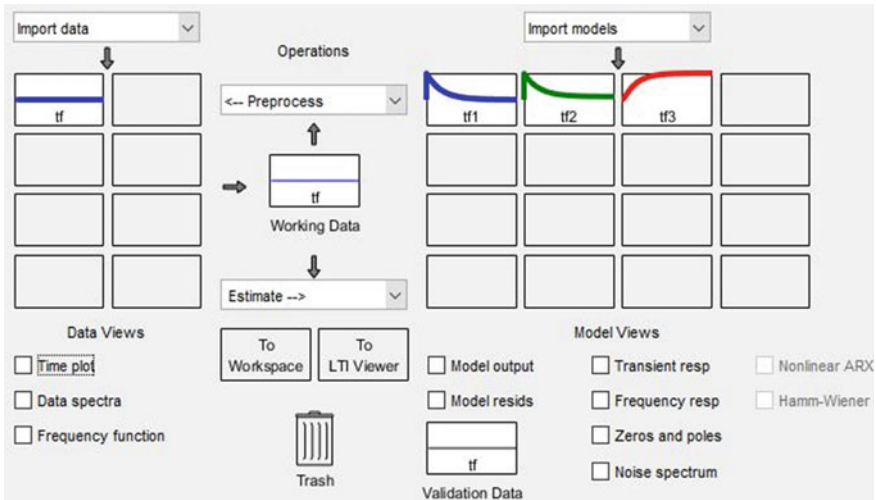


Fig. 3 System identification user interface

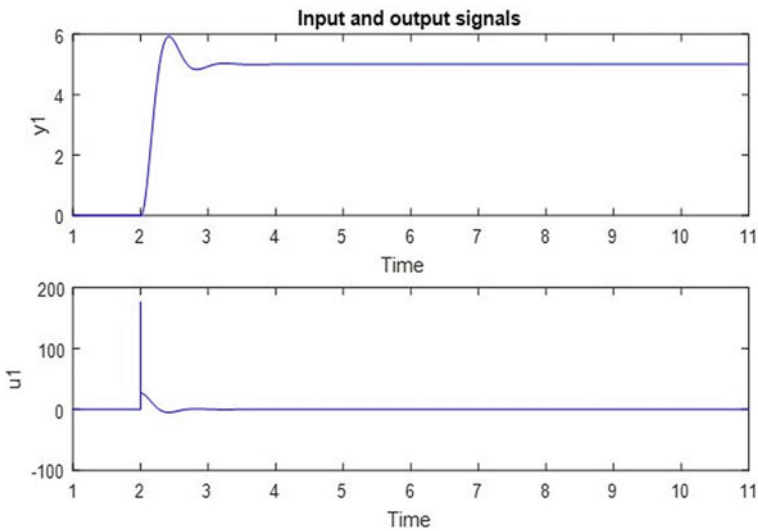


Fig. 4 Input and output signals

Using the idea of poles and zero, three different experiments were done to find the best mathematical model that could be used by a DC motor. According to [14], poles can be interpreted as the value of s in the transfer function that equals zero in the denominator while zero is the value of s that equals zero in the numerator. With poles and zero, users can figure out how stable and well a system works.

In Fig. 3, there are three tests. The first is a transfer function with two poles and no zeros. The second transfer function has two poles and one zero, whereas the third one has three poles and no zeros. Based on these three different assessments, the identification method will give different estimates of how accurate the model is. In the next section, the method of how the Cohen-Coon tuning technique works will be discussed followed by the result of these several tests will be reviewed in-depth.

2.3 Tuning Method for K_p and K_d

Parameter for a controller can be tuned in many ways nowadays. The Ziegler-Nichols method, Cohen-Coon method, bio-inspired computation, and swarm intelligence, which is a type of artificial intelligence are the most common ways to tune something [15]. As for this article, the Cohen-Coon technique will be implemented to analyze how well a DC motor and a suggested controller that has a different pole and zero values work.

The Cohen-Coon technique is an offline tuning strategy, which implies that once the input has attained a steady-state, a step change can be introduced. The output may then be monitored by using the time constant and time delay, and the response is used to calculate the initial control settings. Closed-loop systems can benefit from this tuning strategy since their reaction time is faster [16].

In the Cohen-Coon technique, three process characteristics have been used: process gain, dead time, as well as the time constant. These three traits may be evaluated using a step test and a result analysis [17]. Figure 5 explains how to locate the process variable and the controller output to obtain the three items described above. In addition, to compute the process gain, divide the change in process variable (PV) by the change in controller output (CO) in percentage form.

A tangent line must be drawn from the highest slope of the PV graph to connect with the initial level of PV (before the step-change in CO). As a result, the dead time, t_d characteristics may be computed. The time gap between the change in CO and the intersections of the tangent line with the initial PV level is described as dead time.

Following the calculation of t_d , the value of the time constant, τ required the computation of 63 percent of the entire changes in PV. The time required for the PV to reach the level may then be calculated. Finally, with these parameters, gains for each of the controllers, K_p , and K_d may be constructed as the tuning rules shown in Table 1.

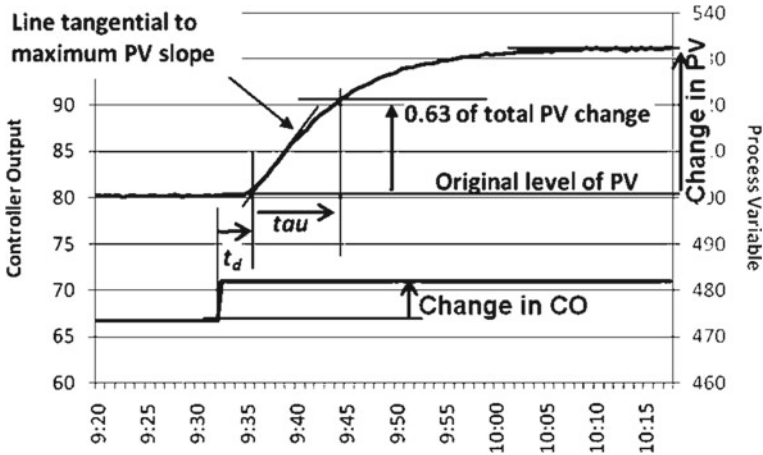


Fig. 5 Step test for Cohen-Coon method

Table 1 Cohen-Coon tuning rules

Controller	Gain's value	
	K_p	K_d
PD	$1.24/g_p ((\tau/t_d) + 0.129)$	$0.27t_d((\tau / 0.324t_d)/(\tau+0.129t_d))$

2.4 Development of DC Motor with PD Controller

By referring to the user interface of system identification in Fig. 3, those different mathematical models give differing degrees of precision that correspond to the estimation data. These accuracy findings, together with their mathematical model, will be given in Table 2.

Table 2 shows that each of the mathematical models generated by estimating the values of poles and zeros illustrates that the identification is good since it is approximately 100% correct, resulting in a better response for the DC motor with suggested controller performance.

These mathematical models, which contain the original, created from the DC motor circuit, will then be developed using PD controller in MATLAB/Simulink

Table 2 Mathematical model and accuracy of system identification

	System identification	
	Mathematical model	Accuracy (%)
1 (P = 2, Z = 0)	$12.48/(s^2 + 7.91 s + 0.003235)$	98.18
2 (P = 2, Z = 1)	$(-0.207 s + 13.98)/(s^2 + 8.94 s + 0.000344)$	99.78
3 (P = 3, Z = 0)	$834.7/(s^3 + 66.95s^2 + 533.6 s + 2.11 \times 10^{-15})$	99.94

software. Although the third model with three poles and no zeros provide an excellent accuracy that is 99.94%, the first and second model still need to be tuned with the proposed controller. This is due to some characteristics that we need to see in detail to decide which is the best mathematical model with the best controller's gain value that can be implemented. The Cohen-Coon tuning approach will be used to modify the settings of the proposed controller for each mathematical model. It is critical to have the required output or performance of the DC motor in the presence of specified controller settings.

This controller will feature proportional gain, K_p , and derivative gain, K_d . Each of these controllers' gains has its advantage and will influences how the system responds. K_p is a system stiffness measurement that shows how much restoring force is needed to compensate for position inaccuracy. After that is K_d which illustrates the damping effects of the system and acts to reduce the overshoot as well as the oscillations with K_p [18].

As it will influence the system response, Jenkins [19] had stated that K_p will assist in minimizing the rising time and steady-state error while boosting the overshoot in the system. K_i 's rising time will be shorter than K_p 's, but its overshoot and settling time will be longer. Finally, K_d 's rising time will be slightly altered to help minimize overshoot and settling time.

Only a few elements can influence the dynamic structure of a closed-loop step response system. T_r is the length of time required for the system to grow from 10 to 90% of its ultimate value. T_s is a settling time that refers to the amount of time it takes for the system to stabilize. T_p is the length of time it takes for the output to reach its maximum level. Because the highest point exceeds the steady-state, a percentage overshoot is calculated based on the disparity.

The steady-state error, e_{ss} is defined as the difference between the beginning step value and the end value. Figure 6 represents the block schematics of each mathematical model with a PD controller.

2.5 Performance Evaluation

Using the characteristics stated above, an evaluation will be performed to identify which of the four mathematical models of DC motor with PD controller works best in terms of positioning accuracy and speed of reaction. The main features that will be compared are T_r , T_s , and e_{ss} , with T_p and percentage overshoot as an extra aspect. The result from studies carried out by Akpama et al. [20] and Chong et al. [21] that is also regarding the DC motor will be made used to measure the transient response characteristics.

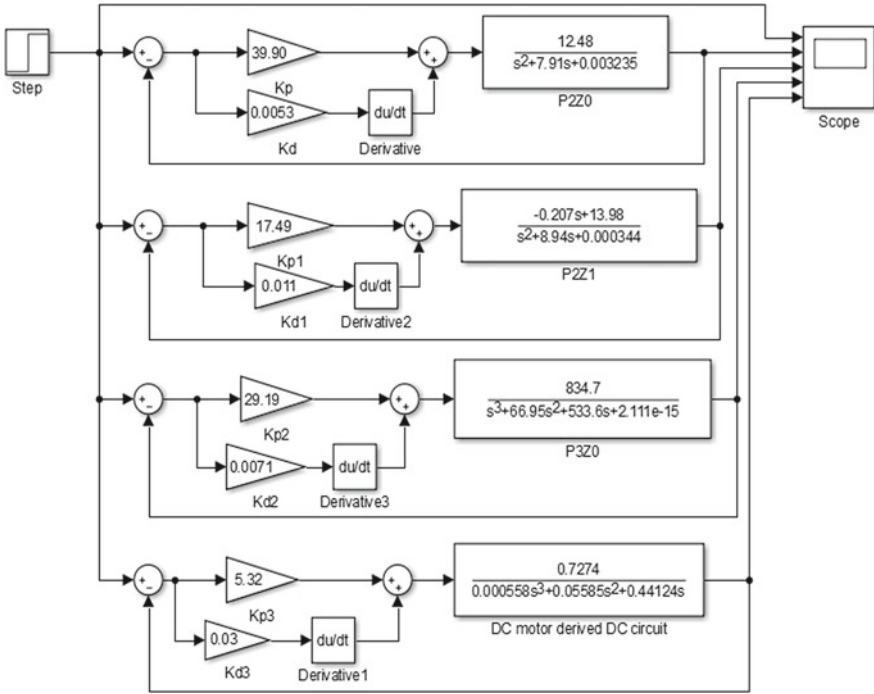


Fig. 6 Block diagram of DC motor with PD controller

3 Result and Discussion

3.1 Simulink Result of System Identification

Based on the simulation in the Simulink program of the system identification procedures, the findings obtained between four various mathematical models of DC motor with PD controller will be examined. During the simulation, the ultimate value or step input is 5 m. Figure 7 displays a combination graph of mathematical models of DC motors with PD controllers generated by the identification technique and mathematical models of DC motors with PD controllers obtained from the DC motor circuit and DC motor parameter assumptions.

By referring to Fig. 7, the output response for the system portrayed different behavior as each of the models has different values of rising time, percentage overshoot as well as settling time. Overall, all of the models are in good condition as it goes back to the steady-state value that is 5 m although some of the models do have few oscillations.

The performance of the DC motor with the proposed controller, which has two poles and no zeros, is good and steady, as shown in the figure above. Even though the accuracy of the mathematical model is not as high as in others, this performance was

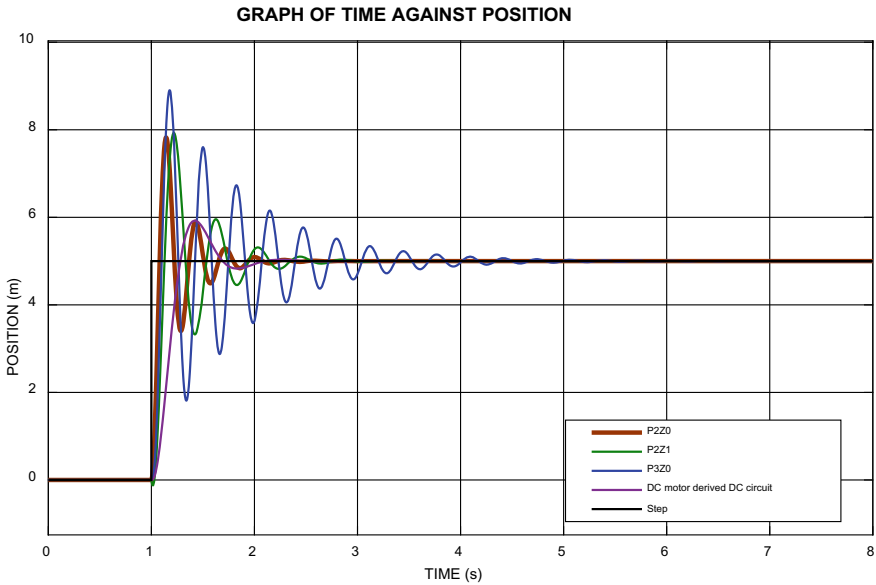


Fig. 7 Simulation result of the combined block diagram of DC motor with PD controller

achieved due to the precision achieved in the identification method with the provided values of poles and zeros. When all four mathematical models are analyzed, the first transfer function responds faster than the others.

The evaluation that has been done in detail for each of the mathematical models of DC motor with PD controller will be explained in the next topic.

4 Transient Response of DC Motor with PD Controller

The system response in Fig. 7 above was used to obtain the transient response for each mathematical model of the DC motor PD controller. Then, the transient response of the DC motor is measured using the PD controller which is tabulated in Table 3.

Table 3 Transient response of DC motor with PD controller for each mathematical model

	Transient response				
	T_r (s)	T_s (s)	T_p (s)	e_{ss}	%OS
1 ($P = 2, Z = 0$)	0.052	1.906	1.142	0.000	56.560
2 ($P = 2, Z = 1$)	0.072	2.473	1.219	0.000	58.660
3 ($P = 3, Z = 0$)	0.060	4.110	1.176	0.000	78.380
4 (From DC motor circuit)	0.184	1.983	1.423	0.000	18.460

Step response Simulink simulation demonstrates that the mathematical model developed from the DC motor circuit and identification procedure with two poles and no zero is the most stable output response when compared to the other mathematical models (Fig. 7).

As in Table 3, by referring to the first transfer function that is two poles and no zero, it has the shortest rising time, T_r , which is 0.052 s for the system response to reach from 10 to 90% of steady-state response. While second, third, and fourth transfer functions obtained 0.072 s, 0.060 s, and 0.184 s accordingly. Due to the high value of proportional gains, it will somehow affect the rise time aspect where it will lower its value of it although the derivative's gain does not affect the characteristic.

While it took longer for other mathematical models to reach and stay within 2 percent of the steady-state value, T_s , again, the first mathematical model needed only 1.906 s to reach and stay within 2 percent of the steady-state value. This happened due to the simulation using very small amounts of derivatives gains. As for the peak time, T_p , of the mathematical model with two poles and no zero obtained 1.142 s, while the other models needed 1.219 s, 1.176 s, and 1.423 s.

Additionally, overshoot is the response of a system that is bigger than its final value. When compared to others, a small amount of overshoot happened for the first transfer function which is 56.560% while, other transfer functions of DC motor with PD controller came out with 58.660%, 78.380%, and 18.46%. The overshoot characteristic was influenced by the proportional gain, where if the value of the mentioned gain is higher, the value of the overshoot will be higher whereas the derivative gain will decrease the value.

The last feature evaluated is that a system's steady-state error, defined as the difference between predicted and actual values, is zero for all previously stated transfer functions. This occurred because of the precision of the identification procedure as well as the creation of a mathematical model of a DC motor from the DC motor circuit itself.

From the Simulink simulation graph, it is plausible to conclude that a precise mathematical model of a DC motor is essential for delivering a dependable response. In comparison to other DC motor mathematical models, the PD controller model obtained from the identification process with two poles and no zero outperforms in performance because it requires a short time from an initial value of 10% to hit 90% steady-state response, to reach and remain in 2% steady-state response, and a small amount of overshoot.

5 Conclusion

Using Kirchoff's voltage law, Laplace transform, and other mathematical tools, system identification based on poles and zeros have been established. To improve the DC motor performance, system identification requires an accurate mathematical model and a PD controller calibrated using the Cohen-Coon technique.

The DC motor with PD controller, which has two poles and zero zeros, is good and steady, according to the simulation findings. Four mathematical models of DC motor with PD controller were compared for the transient response. For the rise time, it takes 0.052 s to travel from 10 to 90% of the final and 1.906 s to maintain 2% of the steady-state. The initial peak overrun took 1.142 s. The model with two poles and no zero has a small amount of overshoot compared to the transfer function obtained from the system identification and no steady-state error when it reaches its ultimate value.

To summarize, a precise derivation for the mathematical model and some identification technique to find the ideal transfer function to be applied are critical for having an efficient output response for a DC motor. This is because a good reaction may be achieved and is impacted by a perfect mathematical model, especially when a controller is utilized.

Acknowledgements The authors would like to thank Universiti Malaysia Pahang for providing financial support under Post Graduate Research Scheme (PGRS) (Grant No. PGRS210365) and the Faculty of Manufacturing and Mechatronic Engineering Technology, Universiti Malaysia Pahang for laboratory facilities.

References

1. Rahman NNA, Yahya NM (2021) A mathematical model of a brushed DC motor system. *Data Anal Appl Math* 2(2):60–68
2. Emhemed AAA, Bin Mamat R (2012) Modelling and simulation for Industrial DC motor using Intelligent control. *Procedia Eng* 41:420–425
3. Chotai J, Narwekar K (2017) Modelling and position control of brushed DC motor. In: 2017 international conference on advances in computing, communication and control (ICAC3), pp 1–5
4. Kimbrell J (2016) The DC motor advantage. <https://www.processingmagazine.com/pumps-motors-drives/article/15586862/the-dc-motor-advantage>. Accessed 01 Jun 2022
5. Boskovic MC, Rapaic MR, Sekara TB, Ponjavic M, Barjaktarovic M, Lutovac B (2019) Novel tuning rules of PD controller for industrial processes. In: 2019 8th mediterranean conference on embedded computing (MECO), June, pp 1–5
6. Surya S, Singh DB (2019) Comparative study of P, PI, PD and PID controllers for operation of a pressure regulating valve in a blow-down wind tunnel. In: 2019 IEEE international conference on distributed computing, VLSI, electrical circuits and robotics (DISCOVER), August, pp 1–3
7. Valousek L, Jalovecky R (2021) Use of the MATLAB® system identification toolbox® for the creation of specialized software for parameters identification. In: 2021 international conference on military technologies (ICMT), June, pp 1–5
8. Donjarenon N, Nuchkum S, Leeton U (2021) Mathematical model construction of DC Motor by closed-loop system Identification technique using Matlab/Simulink. In: 2021 9th international electrical engineering congress (iEECON), March, pp 289–292
9. Ljung L (2010) Perspectives on system identification. *Annu Rev Control* 34(1):1–12
10. Kefal A, Maruccio C, Quaranta G, Oterkus E (2019) Modelling and parameter identification of electromechanical systems for energy harvesting and sensing. *Mech Syst Signal Process* 121:890–912

11. Arifin B, Nugroho AA, Suprpto B, Prasetyowati SAD, Nawawi Z (2021) Review of method for system identification on motors. In: 2021 8th international conference on electrical engineering, computer science and informatics (EECSI), October, pp 257–262
12. 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: 2018 IEEE conference of Russian young researchers in electrical and electronic engineering (EIConRus), January, pp 1801–1804
13. Sharma K, Palwalia DK (2017) A modified proportional integral derivative control with adaptive fuzzy controller applied to direct current motor. In: 2017 international conference on information, communication, instrumentation and control (ICICIC), August, pp 1–6
14. Craig K (2012) The significance of poles and zeros. <https://www.edn.com/the-significance-of-poles-and-zeros-2/>. Accessed 05 Jun 2022
15. Ribeiro JMS, Santos MF, Carmo MJ, Silva MF (2017) Comparison of PID controller tuning methods: analytical/classical techniques versus optimization algorithms. In: 2017 18th International Carpathian control conference, ICC 2017, pp 533–538
16. Bennett J, Bhasin A, Grant J, Lim WC (2022) 9.3 : PID tuning via classical methods. Engineering LibreTexts. [https://eng.libretexts.org/Bookshelves/Industrial_and_Systems_Engineering/Book%3A_Chemical_Process_Dynamics_and_Controls_\(Woolf\)/09%3A_Proportional-Integral-Derivative_\(PID\)_Control/9.03%3A_PID_Tuning_via_Classical_Methods#:~:text=Removeintegralandderi](https://eng.libretexts.org/Bookshelves/Industrial_and_Systems_Engineering/Book%3A_Chemical_Process_Dynamics_and_Controls_(Woolf)/09%3A_Proportional-Integral-Derivative_(PID)_Control/9.03%3A_PID_Tuning_via_Classical_Methods#:~:text=Removeintegralandderi). Accessed 05 Jun 2022
17. Smuts J (2011) Cohen-Coon Tuning Rules. <https://blog.opticontrols.com/archives/383>. Accessed 05 Jun 2022
18. Collins D (2022) What are PID gains and feed-forward gains?. <https://www.motioncontroltips.com/faq-what-are-pid-gains-and-feed-forward-gains/>. Accessed 05 Jun 2022
19. Jenkins (2014) Tuning for PID controllers. http://faculty.mercer.edu/jenkins_he/documents/TuningforPIDControllers.pdf. Accessed 06 Jun 2022
20. Akpama EE, Ezenwosu R (2021) Simulink design of a DC motor control for water pump using fuzzy logic. In: 2nd international conference on electrical power engineering (ICEPENG 2021), July, pp 16–19
21. Chong SH, Tze Ter T, Sakthivelu V (2015) Positioning control of ball screw system driven by DC motor. *Appl Mech Mater* 761:142–147

The Classification of Wafer Defects: An Evaluation of Different Feature-Based ResNet Transfer Learning Models with Support Vector Machine



Lim Shi Xuen, Ismail Mohd Khairuddin, Mohd Azraai Mohd Razman, Jessnor Arif Mat Jizat, Edmund Yuen, Eng Hwa Yap, Andrew Huey Ping Tan, and Anwar P. P. Abdul Majeed

Abstract Wafer defect detection is a non-trivial issue in the semiconductor industry. Conventional means of defect detection are often labour-intensive based that is prone to error owing to a myriad of issue. Hence, there is push towards automatic defect detection in the industry. This work shall investigate the efficacy of a transfer learning pipeline that consists of a different pre-trained ResNet convolutional neural network models in which its fully connected layer is swapped with different support vector machine (SVM) models in classifying the defect state of a wafer whether it pass or fail. The optimal hyperparameters are identified via the grid-search technique. It was shown from the present investigation that the features extracted via the ResNet101v2 transfer learning model with a linear-based SVM model with a C and gamma parameter of 0.01, respectively, could yield a validation and test classification accuracy

L. S. Xuen · I. Mohd Khairuddin · M. A. Mohd Razman · J. A. Mat Jizat ·

A. P. P. Abdul Majeed (✉)

Innovative Manufacturing, Mechatronics and Sports (iMAMS) Laboratory, Faculty of Manufacturing and Mechatronic Engineering Technology (FTKPM), Universiti Malaysia Pahang, 26600 Pekan, Pahang, Malaysia

e-mail: Anwar.Majeed@xjtlu.edu.cn

L. S. Xuen · E. Yuen

Ideal Vision Integration Sdn Bhd, 02-25, Level 2, Setia Spice Canopy, Jln Tun Dr Awang, 11900 Bayan Lepas, Penang, Malaysia

E. H. Yap · A. H. P. Tan

School of Intelligent Manufacturing Ecosystem, XJTLU Entrepreneur College (Taicang), Xi'an Jiaotong-Liverpool University, 215127 Taicang, PR China

E. H. Yap · A. P. P. Abdul Majeed

School of Robotics, XJTLU Entrepreneur College (Taicang), Xi'an Jiaotong-Liverpool University, 215127 Taicang, PR China

A. P. P. Abdul Majeed

Faculty of Engineering, Technology and Built Environment, UCSI University. Kuala Lumpur Campus, 56000 Cheras, Kuala Lumpur, Malaysia

EUREKA Robotics Centre, Cardiff School of Technologies, Cardiff Metropolitan University, Cardiff CF5 2YB, UK

of 96% and 94%, respectively, on a stratified 60:20:20 data split ratio. The result from the present study demonstrates that the proposed pipeline is able to classify the defect level of the wafer well.

Keywords Transfer learning · Wafer inspection · DenseNet

1 Introduction

Wafer defect is a common undesirable issue in the semiconductor industry. It affects the production yield as well as the overall manufacturing process pipeline that consequently incurs untoward associated costs. Conventionally, manual inspection by trained workers is used to evaluate the quality of the wafers. It is worth noting that it takes between six to nine months to train human workers to be able to achieve a detection of accuracy to up to 90% [1]. Nonetheless, the efficacy could drop as low as 70% within 15 months owing to myriad of factors. Therefore, there is a shift amongst industry players in adopting automated optical inspection (AOI) systems [2].

However, it is worth mentioning that the electronics industry has evolved to the nanoscale production that demands for a higher quality of wafers. The detection of such defects through conventional rule-based AOI means is no longer competitive and hence the need for the employment of more advanced technologies, for instance deep learning techniques. Owing to the advancement of computing technology, the use of convolutional neural networks (CNN) has gained traction in a myriad of fields including defect detection [3–6]. Instead of training the CNN models from scratch, researchers have attempted to use pre-trained CNN (also known as transfer learning) models to further expedite the training process and have demonstrated appreciable performance.

Ghosh et al. [7] proposed the use of a transfer learning approach to classify printed circuit board (PCB) defects. The PCB images taken consist of 4655 grayscale images of true defects class and 2888 grayscale images of pseudo-defects class. A pre-trained CNN model, i.e. Inception V3, was used to extract the features, whilst the SVM model was used to classify classes. It was shown in the study that the proposed method was able to achieve an accuracy of 91.125%. Abdulkadis Seker et al. [8] proposed the use of a pre-trained CNN model, viz. AlexNet for fabric defect detection. A total of 3275 images, which includes 936 fabric defect images were used in the study. The AlexNet model was used for both extracting the features as well as classifying the defect images. It was shown from the study that an accuracy of 98.75% is attainable via the proposed method.

In a recent study, Pan et al. [9] proposed the use of a modified transfer learning model in the classification of welding defects. A total of 6208 images that consist of five types of welding defects were investigated in the study. A modified version of the MobileNet by adding an additional fully connected layer with Softmax classifier could achieve a classification accuracy of 96.88%. It is evident that from the related literature on defect detection that different iterations of transfer learning models could provide reasonable classification in distinguishing the evaluated defects. Therefore, this study aims at investigating the efficacy of different feature-based DenseNet transfer learning models for feature extraction purpose, whilst the optimized SVM model shall classify the quality of the wafer, i.e. pass or fail.

2 Methodology

2.1 Data Collection

A total of 395 wafer images obtained from a multi-national manufacturing company located in Penang, Malaysia, were used in the present study. The images were taken from an industrial machine vision platform, i.e. Jäger Vision provided by Ideal Vision Integration Sdn Bhd that consists of a $5\times$ telecentric camera. The dataset consists of both pass (non-defect) and fail (defect) types of wafer images was split into a stratified ratio of 60:20:20 for training, testing and validation, respectively. Figure 1 depicts the sample of the images used in the present study.

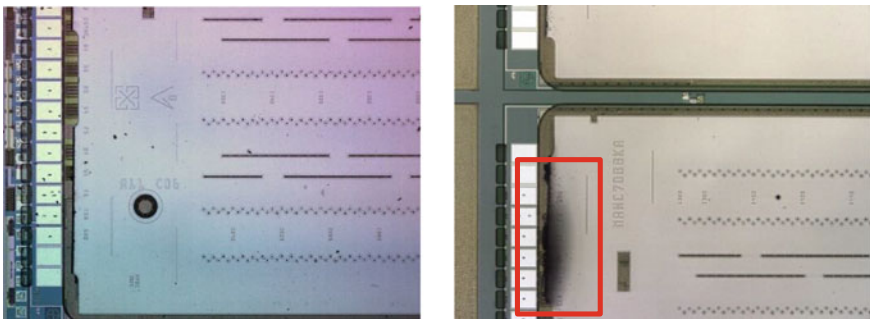


Fig. 1 Left: Pass image. Right: Fail image (Defect in red box)

2.2 Models and Evaluation

In the present study, six types of ResNet models were employed in extracting the features from the images in the proposed transfer learning pipeline. The fully connected layer in the present study is swapped with the support vector machine (SVM) classifier. The images were rescaled to 224×224 prior feeding it to the ResNet models and flattened to $7 \times 7 \times 2048$. Once the features were extracted from the images, it is then fed into the SVM model. The hyperparameters of the SVM model, i.e. kernel, regularization, gamma and degree for the polynomial kernel, were optimized through the exhaustive grid-search technique via five-fold cross-validation on the training dataset [10, 11]. The regularization parameter, C and the gamma parameters were varied between 0.01 and 100 with a multiplication of 10 interval. The kernels investigated were linear, polynomial, radial basis function (RBF) and sigmoid. Whilst for the polynomial kernel, the quadratic and cubic kernels were investigated. Subsequently, the performance of the pipelines was evaluated through the classification accuracy, precision, recall and F1-score. The models were developed via Spyder, a Python programming IDE with its associative Scikit-learn, Keras and TensorFlow libraries.

3 Results and Discussions

From Fig. 2, it is apparent that the best performing pipeline is the ResNet101 v2-optimized SVM pipeline which has an average accuracy of 95% for validation and testing dataset. By only referring to the training dataset accuracy, all of the ResNet transfer learning models with it is associated optimized SVM models are able to reach a classification accuracy (CA) of 99%. Nevertheless, the CA for the test and validation dataset varies between the pipelines where the ResNet101v2-optimized SVM pipeline was demonstrated to be the best at 95%. The optimized SVM classifier hyperparameters for this particular pipeline are 0.01 for the regularized, C and gamma parameters, respectively of along with a linear kernel. The other performance indicators of the ResNet101 v2-SVM pipeline on the test dataset are shown in Table 1.

Based on the confusion matrix of the best pipeline for the different dataset evaluated as illustrated in Fig. 3, where 0 denoted as pass, whilst 1 as fail. It could be seen that 2 failed images (1% of total images) were misclassified as pass (under-reject) in the training dataset. Whilst for the validation dataset, 3 pass images (4% of total images) were failed (over-reject), whilst for the testing dataset, 2 pass images and 3 fail images were misclassified via the developed pipeline. It could be observed that the ResNet101v2 pre-trained CNN model has a better ability in extracting significant features amongst the other ResNet models of the evaluated wafers.

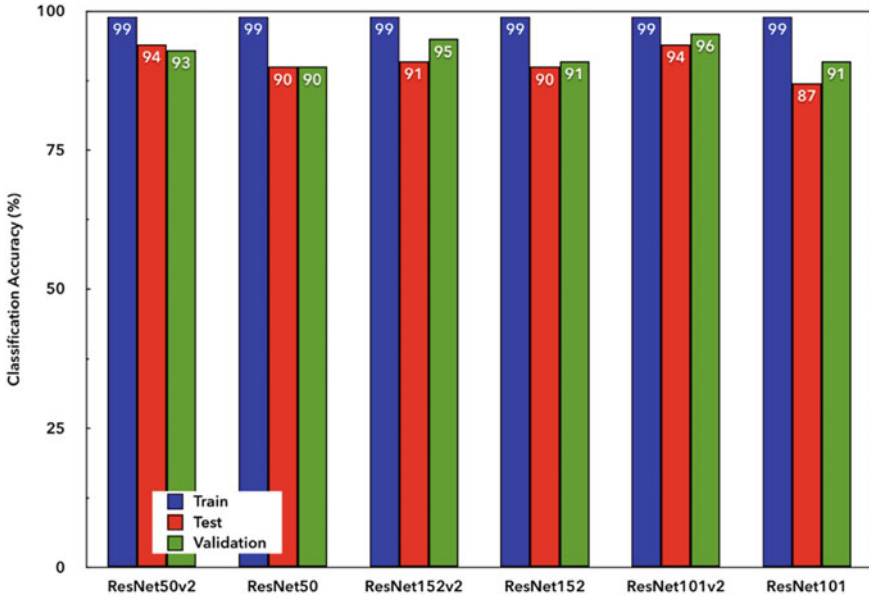


Fig. 2 Comparison between evaluated ResNet-optimized SVM pipelines

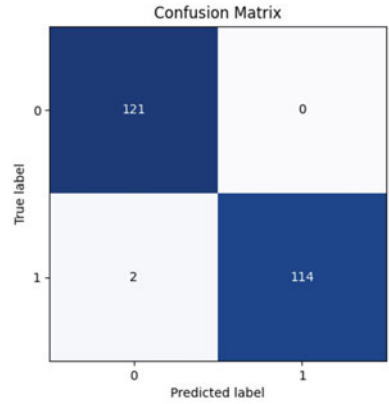
Table 1 Performance measure of ResNet101v2-optimized SVM on the test dataset

Category	Class	Precision	Recall	F1-Score	CA
Pass	0	0.93	0.95	0.94	0.94
Fail	1	0.95	0.92	0.90	

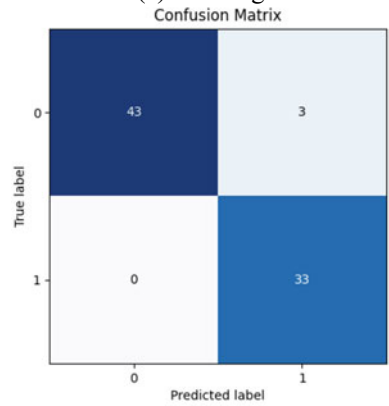
4 Conclusion

The present study has investigated the efficacy of different feature-based ResNet transfer learning family with optimized SVM model in classifying the state of the wafer defect investigated. It was shown that the identified pipeline, i.e. the ResNet101v2-optimized SVM pipeline is able to distinguish well the defects. Future works shall investigate multi-class defects within the fail class as well as investigate other transfer learning families and classifiers towards a robust identification of wafer defects.

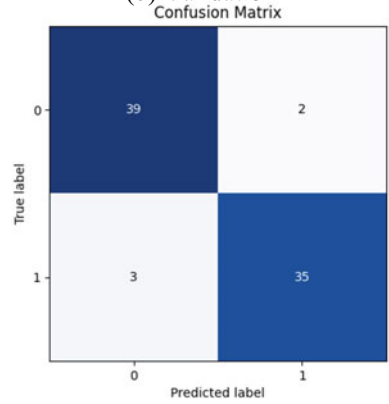
Fig. 3 Comparison between the evaluated pipelines



(a) Training



(b) Validation



(c) Testing

Acknowledgements The authors would like to thank IdealVision Sdn Bhd for providing the image dataset as well as Universiti Malaysia Pahang for funding the study via UIC200815 and RDU202404.

References

1. Mat Jizat JA, Abdul Majeed PPA, Ahmad AF, Taha Z, Yuen E (2021) Evaluation of the machine learning classifier in wafer defects classification. *ICT Express* 0–4. <https://doi.org/10.1016/j.ict.2021.04.007>
2. Huang SH, Pan YC (2015) Automated visual inspection in the semiconductor industry: a survey. *Comput Ind* 66:1–10. <https://doi.org/10.1016/j.COMPIND.2014.10.006>
3. Kumar JLM, Rashid M, Musa RM, Razman MAM, Sulaiman N, Jailani R, Abdul Majeed APP (2021) The classification of EEG-based wink signals: a CWT-Transfer Learning pipeline. *ICT Express*. <https://doi.org/10.1016/j.ict.2021.01.004>
4. Mahendra Kumar JL, Rashid M, Muazu Musa R, Mohd Razman MA, Sulaiman N, Jailani R, Abdul Majeed PPA (2021) The classification of EEG-based winking signals: a transfer learning and random forest pipeline. *PeerJ* 9:e11182. <https://doi.org/10.7717/peerj.11182>
5. Abdullah MA, Ibrahim MAR, Shapiee, Bin MNA, Mohd Razman MA, Musa RM, Abdul Majeed APP (2020) The classification of skateboarding trick manoeuvres through the integration of IMU and machine learning. In: *Lecture notes in mechanical engineering*. Springer, pp 67–74. https://doi.org/10.1007/978-981-13-9539-0_7
6. Rangasamy K, As'ari MA, Rahmad NA, Ghazali NF (2020) Hockey activity recognition using pre-trained deep learning model. *ICT Express*
7. Ghosh B, Bhuyan MK, Sasmal P, Iwahori Y, Gadde P (2018) Defect classification of printed circuit boards based on transfer learning. In: *Proceedings of 2018 IEEE application signal process conference ASPCON*, pp 245–248. <https://doi.org/10.1109/ASPCON.2018.8748670>
8. Seker A (2018) Evaluation of fabric defect detection based on transfer learning with pre-trained AlexNet 9–12
9. Pan H, Pang Z, Wang Y, Wang Y, Chen L (2020) A New image recognition and classification method combining transfer learning algorithm and MobileNet model for welding defects. *IEEE Access* 8:119951–119960. <https://doi.org/10.1109/ACCESS.2020.3005450>
10. Shapiee MNA, Ibrahim MAR, Mohd Razman MA, Abdullah MA, Musa RM, Abdul Majeed APP (2020) The Classification of skateboarding tricks by means of the integration of transfer learning and machine learning models. In: Mohd Razman M, Mat Jizat J, Mat Yahya N, Myung H, Zainal Abidin A, AKM (ed) *Embracing industry 4.0. Lecture Notes in Electrical Engineering*, vol 678. Springer, Singapore, pp 219–226
11. Almanifi ORA, Mohd Khairuddin I, Mohd Razman MA, Musa RM, Abdul Majeed PPA (2022) Human activity recognition based on wrist PPG via the ensemble method. *ICT Express*. <https://doi.org/10.1016/j.ict.2022.03.006>

The Correlation Between Peltier Module, Solution Volume and Temperature in IoT-Controlled Hydroponic Nutrient Solution Management



Hamdan Sulaiman, Ahmad Anas Yusof, and Mohd Khairi Mohamed Nor

Abstract This paper presents the study to determine the correlation of factors considered in developing a temperature control system that remotely controls the heating and cooling process of a hydroponic nutrient solution (HNS). The system incorporates the use of the Internet of Things (IoT), to remotely monitor, control, and acquire data. The ESP32 is used to provide IoT connectivity through the Blynk application. The temperature measurements were collected using the DS18B20 temperature sensor. The number of Peltier modules utilized for both heating and cooling cores varies from 1 to 4 units, as well as the volume of solution employed, varies between 1000 ml, 1500 ml, and 2000 ml, respectively. The heating and cooling performances were evaluated by determining the thermal equilibrium temperature meanwhile the correlation was analyzed using a fit regression model. The highest and lowest temperature recorded are 67.36 °C and 11.88 °C, respectively, with the uses of 4 Peltier modules as the temperature regulation core. Furthermore, the result reveals that the number of Peltier modules factor has significant affect ($P < 0.05$) on both of the heating and cooling capabilities of the system meanwhile the volume of the HNS factor is only significant for cooling performance ($P = 0.023$) compared to heating performance ($P = 0.205$). The correlation of the factors is represented by the regression model.

Keywords Temperature control · Thermoelectric Peltier · Regression analysis · Internet of Things · Hydroponic nutrient management

H. Sulaiman (✉)

Faculty of Plantation and Agrotechnology, Universiti Teknologi MARA, Jasin Campus, 77300 Merlimau, Melaka, Malaysia

e-mail: Hamdan91821@uitm.edu.my

A. A. Yusof

Robotics and Industrial Automation Research Group, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

A. A. Yusof · M. K. Mohamed Nor

Faculty of Electrical Engineering, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

1 Introduction

In both educational and industrial works, high accuracy reading is vital to ensure a precise and reliable result. Temperature variation has a very high potential to influence some properties and behavior of an aqueous solution, which can lead to erroneous results. The temperature has a major effect on an aqueous solution according to many types of research. Meticulous research regarding the effect of temperature on certain properties of an aqueous solution necessitates a temperature regulation process to undergo either heating or cooling process. This study presents a temperature control system that utilize Peltier module as the temperature regulation core and the Internet of Things (IoT) as the platform for temperature monitoring, control, and data acquisition. Furthermore, the performances of the system were evaluated by determining the maximum and minimum temperatures in unit of Celsius ($^{\circ}\text{C}$). The factors considered in the development of the system were analyzed using regression analysis to identify the significance of the factors on heating and cooling capabilities.

A simple way to describe a Thermoelectric (TEC) Peltier module is that it resembles a flat-square plate with two sides, the cold side, and the hot side. The Peltier effect occurs when electricity is forced to flow through a circuit, generating a temperature difference between the junctions of electric conductors made of two different types of materials [1, 2]. An array of p- and n-type semiconductor components highly doped with electrical carriers makes up a thermoelectric module. Electrically connected in series, yet thermally coupled in parallel, the elements are organized in an array. Figure 1 shows the array is subsequently mounted to two ceramic substrates, one on either side of the elements. It has been used for a variety of applications, including refrigeration, clinical cooling, and electrical equipment cooling. The Peltier module is regarded as the system's temperature-regulating core.

Research had been done and the results show that the temperature changes are known to alter alcohol aggregation behavior and water structure [3]. Others have demonstrated that varying water temperatures result in different pH effects on copper toxicity in green microalgae [4]. Temperature is also important in determining the lubricity of a solution [5]. All scholars agree that the changes of temperature are able to alter some properties or behavior of an aqueous solution. This effect could potentially change some properties of a Hydroponic Nutrient Solution (HNS). The HNS is one of the most important factors to ensure the successfulness of hydroponics horticulture, and two fundamental parameters of the nutrient solution that must be managed and monitored regularly are electrical conductivity (EC) and hydrogen potential (pH) to ensure that plant growth is at an optimal level. As a result, more research regarding temperature compensation can be done to see how temperature affects the readings of EC and pH in the hydroponic nutrient solution [6–8]. However, a temperature control system needs to be developed in the first place.

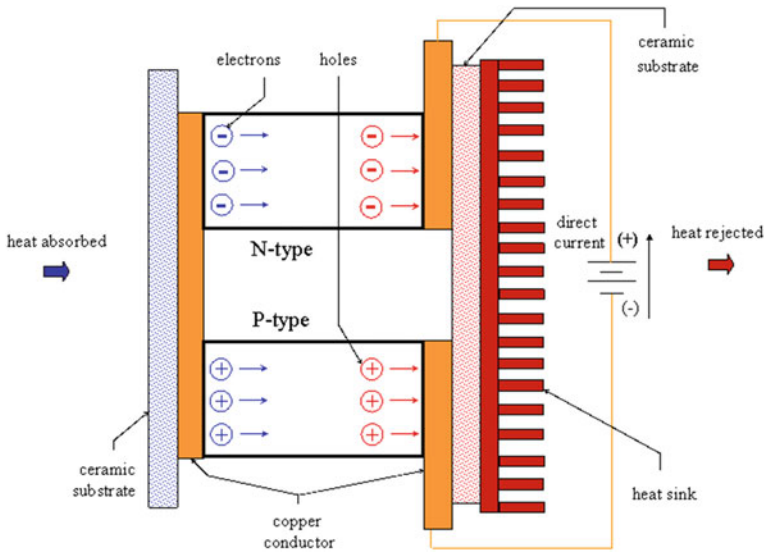


Fig. 1 Illustration of a TEC Peltier module

2 Methodology

Figure 2 shows a schematic diagram of a temperature control system utilizing 4 Peltier modules as one of the factors considered in the experiment's conditions.

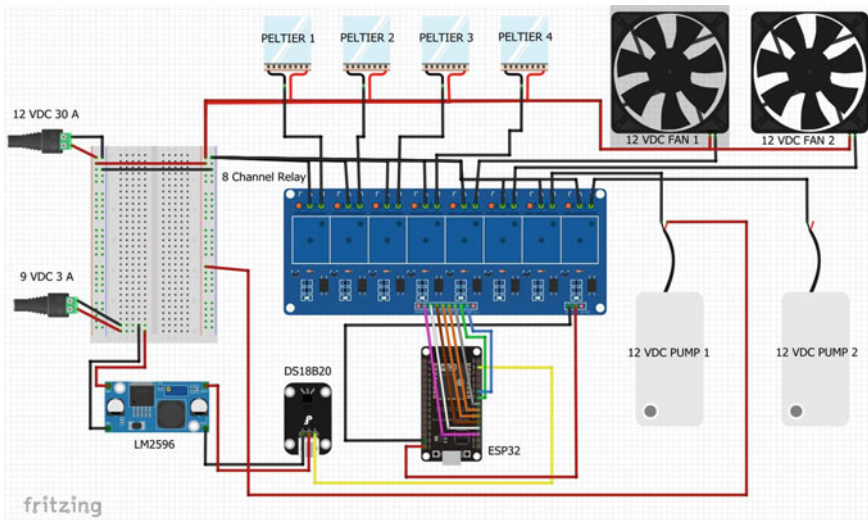


Fig. 2 Schematic diagram of temperature control system

By referring to Fig. 2, a 12 VDC 30 A power supply was used to energize 12 V components used in the system which are four Peltier modules (TEC12706) as the temperature regulation cores and, two 12 VDC submersible pump which is used to convey the HNS from a container into the temperature regulation system. Besides that, 12 VDC fan is also used to enhance the convection process of the system to increase the heating and cooling capabilities. All of these components were controlled by ESP32 microcontroller through 8-channel relay. ESP32 is used as the microcontroller due to its capability to provide a Wi-Fi connection so that the IoT can be implemented in the system. Therefore, the process of triggering on and off of the components are done remotely through Internet. Another power supply (9 VDC 3A) is used to power up the temperature sensor (DFRobot DS18B20), ESP32 microcontroller and the 8-channel relay. However, the voltage of 9 VDC is stepped down to 5 VDC on the first place before powering up these components as the voltage requirement of the components is 5 VDC.

To describe the connections of the components, the negative wire (GND) of the Peltier module 1, 2, 3, and 4 were connected to the normally open (NO) pins of the relay channel 1, 2, 3, and 4, respectively. Channel 5 and 6 of the relays are used to control the fan of the heating and cooling system, respectively. Channel 7 and 8 of the relays are used to control the pumps for heating and cooling system, respectively. Note that these open circuit is controlled by the ESP32. Channel 1, 2, 3, 4, 5, 6, 7, and 8 of the relays is controlled by pin 15, 2, 4, 16, 17, 5, 18, and 19 of the ESP32 and the relay can be triggered on by writing “*digitalWrite(relayNumber, LOW);*” in the program. Furthermore, the data pin of the temperature sensor is connected to digital pin 22. The program of the system is written by using the Arduino IDE. The HNS is filled manually according to the experiment, and the Peltier module is also activated manually using a smart device through Blynk Apps IoT platform. Temperature readings were taken every minute and sent to the cloud. Each set of experiments lasted three hours. Table 1 represents the experimental design, which is a multilevel full factorial design with 24 experiments containing a combination of factors.

To evaluate the system's performance, the experiments used a multilevel full factorial design. The number of Peltier modules used as heating and cooling cores, as well as the volume of the test solution, were considered as the continuous factors. The level denotes the number of points considered to be the value of an independent variable. There are 4 levels of Peltier module factor and 3 levels of volume of water factor. The controlled variables are the experiment period and test solution, which are 3 h, and HNS, respectively. The experiments are divided into two categories: heating experiments and cooling experiments. Each category contains 12 sets of experiments, for a total of 24 experiments to be carried out.

Table 1 Design of experiment for system performance test

Run	Peltier numbers (level)	Volume of water (level)
1	4	1
2	3	3
3	3	1
4	2	3
5	1	3
6	2	1
7	1	1
8	4	2
9	4	3
10	1	2
11	2	2
12	3	2

3 Results and Discussion

The results acquired through the experiments are tabulated in Table 2. The highest temperature recorded is 67.36 °C whereas the minimum temperature is 11.88 °C with the temperature regulation cores of 4 Peltier modules and 1000 ml of water. Besides that, the maximum temperature is noted to increase, with the increase of the number of Peltier and the decrease in the volume of water. On the other hand, the minimum temperature decreases with the same conditions as the maximum temperature.

Minitab Statistical Analysis software is used to analyze the result by using fit regression model to determine the correlation between the number of Peltier modules, volume of water, and maximum and minimum temperature as shown in Fig. 3. The maximum and minimum temperature has been set as the responses whereas the number of Peltier module and the volume of water has been set as the continuous predictor. The regression analysis is run automatically by the Minitab software and results is tabulated in Table 3.

By referring to Table 3, the P-value of the regression analysis represents the significance of the factor toward either heating or cooling process. A P-value which is less than 0.05 indicates that the factor has significant effect on the respective process whereas a P-value more than 0.05, vice-versa. The number of Peltier module used for heating process has a greater influence on the achievable maximum temperature ($P = 0.000$) compared to the water volume factor ($P = 0.205$). This can be validated when observing the maximum temperature according to the manipulation of the variable. The experiments (run 1, 3, 6, and 7) which manipulate the number of Peltier module (1 to 4 units) but having the same volume of water (1000 ml), shows a difference of temperature up to 22.74 °C (run 1–run7). On the other hand, the difference of temperature with the setting of fixed number of Peltier module (1 unit)

Table 2 Results of temperature control experiment

Run	Factor 1 Peltier module (unit)	Factor 2 Volume of water (ml)	Response 1 Max Temperature (°C)	Response 2 Min Temperature (°C)
1	4	1000	67.36	11.88
2	3	2000	60.47	16.04
3	3	1000	62.69	14.81
4	2	2000	53.46	16.81
5	1	2000	43.23	20.94
6	2	1000	55.00	16.45
7	1	1000	44.62	18.38
8	4	1500	67.31	12.98
9	4	2000	67.28	13.05
10	1	1500	43.65	19.94
11	2	1500	53.67	16.50
12	3	1500	61.54	15.94

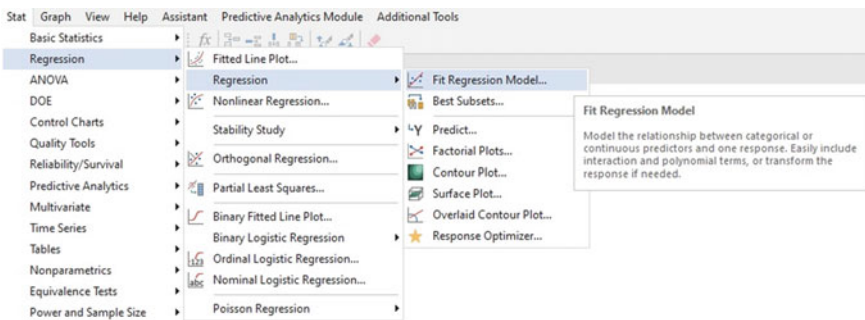


Fig. 3 Fit regression model using Minitab

and manipulated volume of water (run 5, 7, and 10) is only up to 1.39 °C (run 7–run 5).

As for the cooling process, the number of Peltier module ($P = 0.000$) has a greater effect on the achievable minimum temperature compared to the factor of the volume of water ($P = 0.023$). However, the volume of water is considered as a significant factor as the P-value is lower than 0.05. The R-sq of the heating and cooling process indicates 98.25% and 92.65% of confidence, respectively, that the variance in the maximum and minimum temperature contributed by the factors included in the analysis. The high percentage of R-sq(pred), 96.68% and 87.00%, also represents that the model has adequate predictive ability to predict the value of the maximum

Table 3 Regression analysis for temperature control system

Term	Coef	SE Coef	T-Value	P-Value	VIF
<i>Heating</i>					
Constant	39.16	1.73	22.68	0.000	
Peltier module	7.797	0.350	22.30	0.000	1.00
Volume of water	-0.001307	0.000958	-1.37	0.205	1.00
Regression model	Min Temp = 19.733-2.234 Peltier Module + 0.001330 Volume of Water				
Model summary	S	R-sq	R-sq(adj)	R-sq(pred)	
	1.35419	98.23%	97.84%	96.68%	
<i>Cooling</i>					
Constant	19.733	0.878	22.47	0.000	
Peltier module	-2.234	0.178	-12.56	0.000	1.00
Volume of water	0.001330	0.000487	2.73	0.023	1.00
Regression model	Min Temp = 19.733-2.234 Peltier Module + 0.001330 Volume of Water				
Model summary	S	R-sq	R-sq(adj)	R-sq(pred)	
	0.689051	92.65%	91.02%	87.00%	

and minimum temperature using the regression model correspond to the heating or cooling process, respectively.

4 Conclusion

The IoT-based temperature regulation system has been assessed to determine the heating and cooling performance by identifying the thermal equilibrium temperature. The highest and lowest temperature recorded is 67.36 °C and 11.88 °C, respectively, with use of 4 Peltier module as the temperature regulation core. The results show that the heating and cooling capability significantly increases with the increase in the number of Peltier modules used. To be specific, the regression models represent the correlation between the factors and the responses and it indicates that the performance of the system significantly depends on the number of Peltier modules used as the P-Value is <0.05. Therefore, the desired temperatures can be determined by using the regression model with an accuracy of 96.68% and 87.00% for the heating and cooling processes, respectively. With regards to hydroponic nutrient management, the developed temperature regulation system can be applied to regulate the temperature of the hydroponic nutrient solution to the desired temperature by referring to the regression model to predict the number of Peltier module needed with respect to the volume of water. On the other hand, the developed system can also be used by the reader as a test rig for further analysis with regards to the effect of temperature onto

an aqueous solution such as determining the effect of temperature on EC and pH of the hydroponic nutrient solution.

Acknowledgements The authors wish to thank the Ministry of Higher Education, Universiti Teknologi MARA, and Universiti Teknikal Malaysia Melaka for their support.

References

1. Romd Singh A (2019). Air conditioner using Peltier module. *Int J Res Appl Sci Eng Technol* 7:383–386. <https://doi.org/10.22214/ijraset.2019.2047>
2. Kudva N, Veerasha RK, Muralidhara (2020) A review on thermoelectric (Peltier) module. *Int J Progress Res Sci Eng* 1(4)
3. Parameswaran S, Choi S, Choi JH (2022) Temperature effects on alcohol aggregation phenomena and phase behavior in n-butanol aqueous solution. *J Mol Liq* 347. <https://doi.org/10.1016/j.molliq.2021.118339>
4. Pascual G, Sano D, Sakamaki T, Akiba M, Nishimura O (2022) The water temperature changes the effect of pH on copper toxicity to the green microalgae *Raphidocelis subcapitata*. *Chemosphere* 291. <https://doi.org/10.1016/j.chemosphere.2021.133110>
5. Kreivaitis R, Kupčinskas A, Žunda A, Ta T-N, Horng J-H (2021) Effect of temperature on the lubrication ability of two ammonium ionic liquids. *Wear* 492–493:204217. <https://doi.org/10.1016/j.wear.2021.204217>
6. Song J, Xu L, He D, Tuskagoshi S, Kozai T, Shinohara Y (2019) Estimating EC and ionic EC contribution percentage of nutrient solution based on ionic activity. *Int J Agric Biol Eng* 12:42–48. <https://doi.org/10.25165/j.ijabe.20191202.4399>
7. Singh H, Dunn B, Payton M (2019) Hydroponic pH modifiers affect plant growth and nutrient content in leafy greens. *J Hortic Res* 27:31–36. <https://doi.org/10.2478/johr-2019-0004>
8. Sublett W, Barickman C, Sams C (2018) The effect of environment and nutrients on hydroponic lettuce yield, quality, and phytonutrients. *Horticulturae* 4:48. <https://doi.org/10.3390/horticulturae4040048>

The Statistical Impact of Artificial Intelligence Towards the Price Change of Financial Instrument



Lim Guo Huang, Choong Kah Wei, Nor Aziyatul Izni, Loh Yue Fang, Tan Sher Lyn, and Sarah Atifah Saruchi

Abstract Forecasting future price is not an easy task due to wide impact variables, nonlinear, and complexity in financial market. Fundamental analysis plays an essential role in price valuation when financial analyst, investors, and institutional traders were on their ways to evaluate current stock prices. Industrial Revolution 4.0 (IR4.0) creates a trend of evolution of artificial intelligence (AI) in financial technology field. The machine learning (ML) become one of the famous techniques used by the investors to forecast the movement of financial derivative's price. Therefore, this study aims to determine the optimal model in forecasting the S&P 500 market index price. The models used are Artificial Neural Network (ANN), Long Short-Term Memory (LSTM), and Random Forest (RF). The duration of data used is 10 years which is from January 1st 2011 to December 31st 2021, which included the weekends and public holidays. The data was split into training set and testing set. Training set is fitted into the models to obtain the optimal combination of parameter to gain the accuracy and reliability of result. After getting the appropriate parameters for each model, testing phase was carried out to determine the optimal model by using the error metrics which are mean absolute percentage error (MAPE), root mean squared error (RMSE), and mean absolute error (MAE). As a results, ANN is the best optimal model in predicting the S&P 500 adjusted close process with the lowest error metrics and highest accuracy compared to LSTM and RF. The findings of this study can be used by individual and institution to forecast the future price changes of stock market

L. G. Huang · C. K. Wei · T. S. Lyn

Institute of Actuarial Science and Data Analytics, UCSI University, 56000 Cheras, Kuala Lumpur, Malaysia

N. A. Izni (✉)

Centre of Foundation Studies, Universiti Teknologi MARA, Cawangan Selangor, Kampus Dengkil, 43800, Dengkil Selangor, Malaysia

e-mail: naizni@uitm.edu.my

L. Y. Fang

Faculty of Business and Management, UCSI University, 56000 Cheras, Kuala Lumpur, Malaysia

S. A. Saruchi

Faculty of Manufacturing and Mechatronics Engineering Technology, Universiti Malaysia Pahang, 26600 Pekan, Pahang, Malaysia

and indexes. This can help the users to have a better prediction on future price so that they make an appropriate decision in the investment process.

Keywords Stock price prediction · Decision support · Artificial neural network · Financial instrument

1 Introduction

The fundamental analysis plays an important role in price valuation when financial analyst, investors and institutional traders were on their ways to evaluate current stock prices. In twentieth century, individual has lack of resources in access to relevant information in financial market. Most of them are pursuing information through learning from experienced, receiving information from economic newspaper and telegram. From a theoretical point of view, an efficient valuation of a firm should reflect on the firm's development and crucially dependent on the available information for investors. The financial news was leading to the adjustment of investor's expectation towards the price prediction [1].

A dramatic shift in the way of trading happens after the development of AI. With the help of AI, investors can collect data in a short period and execute trade in microseconds [2]. Other than that, self-learning of AI improves the accuracy the result obtained. Elimination of mental processes of human experts during a financial decision improve the development of AI in financial market.

In twentieth century, investors could purchase for certain financial instrument such as stocks, bonds, trust funds through phone calling, walk-in service and trusteeship. The inconvenience does not affect the prompt of investment trend in the era. As the concept of investment is getting wider in the era, more citizens are willing to turn themselves from saving money into investing money. In the study of financial market, there are various of investment products such as Initial Public Offering (IPO) stocks, treasury-bills, and corporate bonds, which meets public's preferences.

As entering to twenty-first century, there are more access for investors to get relevant information, through Internet browsers, financial education, and even pursuing for degree major in economic or finance. Investors now could do self-learning through online or offline same as in twentieth century. The evolution of technology creates a brand-new trend of finance investment. Investors interested in using machine learning techniques to develop an automated agriculture commodities price forecast system [3]. Popular machine learning algorithms such as Artificial Neural Networks (ANN), Long Short-Term Memory (LSTM), and Random Forest (RF) have been investigated and implemented with large historical datasets and creates a model with small mean-square error for the selection of price prediction engine of a proposed system.

The financial market is influenced by a variety of economic factors. The prediction of future prices will be more accurate when using a wider sample of economic elements. Therefore, the objective of the research is to investigate the relationship between the financial factors and future price change. This study used economic data

(Open, High, Low, and Close (OHLC) of Standard and Poor's 500 (S&P 500)), US 10Y bond yield, USD Dollar Index, crude oil price, and gold price) as independent variables to predict the future price of S&P 500. There is no combination yet of mentioned variables in computing a model for assisting investors in price prediction. Hence, the impact of independent variables using machine learning is implemented through machine learning and leads to the next objective which is to examine the usage of ANN, RF, and LSTM and examine the accuracy of models in prices change prediction to obtain the best performance model.

2 Literature Review

In the article of Jareño and Negrut [4], the stock market in the United States had a positive and statistically significant association with gross domestic product (GDP) and industrial production index (IPI) variables, but a negative and statistically significant relationship with unemployment and interest rates [4]. Gokmenoglu and Fazlollahia [5] conclude that gold is an excellent stock substitute. In short-term, the volatility of gold has no impact on S&P 500 [5]. In addition, gold shows a negative correlation with S&P 500 in financial crisis [6]. Hsing and Hsieh [7] found that the Poland's stock index rises due to a higher real GDP [7]. Hsing [8] also concludes that the US stock market index will grow as real GDP rises, real Treasury bill rates fall, and inflation falls [8]. In the study of Danso [9], they discovered a negative relationship between unemployment and GDP growth rate and stock market performance, but a direct relationship between inflation and stock market performance [9]. Besides, Balcilar and Ozdemir [10] conclude a result of the switching model demonstrate that oil futures return has a high predictive capacity for each of the S&P 500 sub-index returns evaluated over various sub-periods in the sample, but each of the S&P 500 sub-index returns has a low predictive capacity for the oil futures price [10].

Similar interpretation of both models, ANNs using epoch approach as the weighted inputs in computing neurons by calculating the standard error of the variables [11]. The proposed system would benefit from close collaboration between the AI and neuroscience fields, which would lead to a number of practical breakthroughs [12]. Back Propagation Algorithm applied in ANN models in making complex calculation become more user-friendly and more accurate in determining the standard error [13]. The hybrid method explains ANN prediction models, and to get the best accuracy, the Classification Error Percentage (CEP) is utilized to measure accuracy by altering the weightage of inputs [14].

According to the research report from von Mettenheim and Breitner [15], OHLC data can be used to compute price predictions because it does not require real intraday tick data for back testing [15]. In the experimental analysis, LSTM model had the lowest result on mean square error (MSE), but it had the worst prediction result coming up with lowest accuracy. In all the comparisons, associated neural network worked on the best result as it can predict multiple types of price data at the same data and had more than 95% accuracy of predicted value to the real value [16]. LSTM

has a better performance than other models such as random forest and DNN model in non-stationary data. LSTM also has good performance in time-series forecasting, text recognizing, and sentiment analyzing [17].

Among the categorization strategies, Random Forest produced the best level of prediction accuracy, while Decision Tree-based regression models produced the least error in stock price modeling and prediction [18]. The empirical data show that the best results are obtained when the RF is utilized for both stock selection and stock price trend forecasting. Since the winning rate is the greatest among the models utilized, the RF-RF model has a very good performance in long-term stock price prediction [19]. The performance of the RF and LSTM models is superior to that of the PCA and LSTM models [20]. Random Forest and Support Vector Machines have a return of 2-times higher than S&P 500 index. The result shows that Random Forest has the highest cumulative return among these three models. It has a return of 87.36% [21].

3 Methodology

This section focuses on the details of the datasets. There will be several stages to be conducted in this chapter. First, Yahoo Finance will be used to extract historical daily prices for the S&P 500, US10Y bond yield, DXY, crude oil price, and gold price. Besides, the methodology related to the implementation of the proposed models to forecast future price of S&P 500 which included ANN, LSTM, and RF will also be discussed in this section. Lastly, this section explains the error metrics that are used to evaluate the forecasting performance and decide the best model in this study. The root mean squared error (RMSE), mean absolute percentage error (MAPE), and mean absolute error (MAE) are the error metrics used in this investigation. The extracted dataset spans ten years, from January 1, 2011, to December 31, 2021. Each independent variable and dependent variable have a sample size of 2769.

The training data comprises 80% of the dataset and is used to train the model and determine the appropriate parameters, while the testing data comprises 20% of the dataset and is used to assess the models' performance. The goal of having a testing set is to find the best model for predicting future prices. The normalization method is used to further process the training and testing sets. This procedure is used to normalize data to improve the model's accuracy. The normalization method is as below:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

Table 1 shows the correlation coefficient between the dependent variable and independent variables.

The OHLC of S&P 500 shows a strong positive correlation with the adjusted close of S&P 500. The correlation coefficient of the OHLC is close to 1. DXY and gold

Table 1 Correlation coefficient between the independent variables and dependent variable

Correlation coefficient	S&P 500 adjusted close	Correlation coefficient	S&P 500 Adjusted close
S&P 500 Open	0.9996	US 10Y Bond Yield	-0.4266
S&P 500 High	0.9998	DXY	0.5983
S&P 500 Low	0.9998	Crude Oil	-0.4704
S&P 500 Close	1.0000	Gold	0.3327

price have positive correlation with adjusted close of S&P 500. DXY has a correlation coefficient of 0.5983 which are higher than gold price of 0.3327. Besides, US 10Y bond yield and crude oil price have a negative correlation with adjusted close of S&P 500. The adjusted close of S&P 500 has a -0.4266 and -0.4704 correlation coefficient with US 10Y bond yield and gold price, respectively. When the variables are positive correlated, their value will move in the same direction and vice versa. The closer the value of correlation coefficient towards 1 or -1, the stronger the relationship between the variables.

Throughout the experiments, the optimal number of epochs in constructing ANN model is 500. This is because the error metric is lowest and it provides a highest accuracy, compared with 10, 100 and 1000 epochs number. When the number increases, the value of error metrics will become lower, and accuracy will become higher. However, there is a marginal effect when the number of epochs increases. Figure 1 shows the relationship between the number of epochs and the training loss. When the number of epochs becomes larger, the training loss of the model becomes lower.

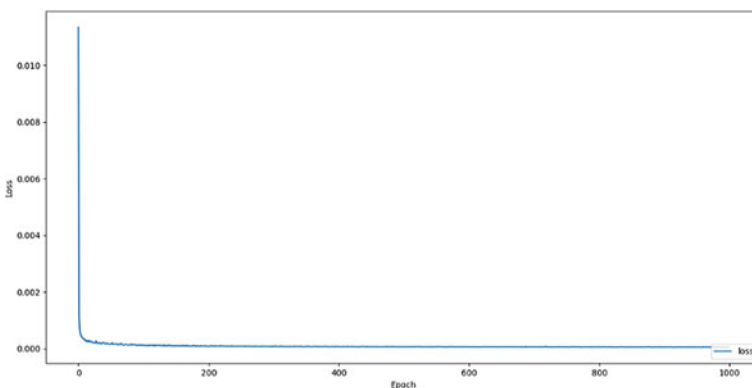


Fig. 1 Relationship between the number of epochs and the training loss

The optimal number of trees to construct RF model is 1000 trees. This is because a large number of trees can increase the accuracy of prediction and prevent overfitting. When the number of trees increases, the error metrics becomes smaller, and the accuracy becomes higher.

4 Results and Discussion

In this section, the performance of three models will be compared and the optimal model is selected. The number of epochs used to examine the accuracy of the model is 500 because the ANN model with 500 epochs number shows the best performance in training process. The parameters used in ANN test prediction is shown in Table 2. The visualization of prediction value and actual value is shown in Fig. 2.

Table 2 Parameters used in ANN test prediction

Parameters		2Y test prediction error metrics	
Number of hidden layers	3	MAPE	4.3430
Number of neurons in each hidden layer	500, 500, 300	RMSE	187.3319
Activation function	ReLU	MAE	160.9507
Kernel initializer	Uniform	Accuracy	95.6600
Batch size	20		
Number of epochs	500		
Optimizer	Adam		
Loss function	MSE		

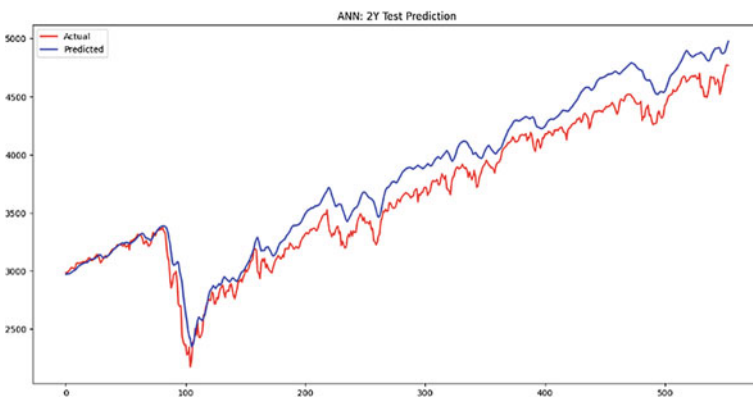


Fig. 2 ANN test prediction with 500 epochs

Table 3 Parameters used in LSTM test prediction

Parameters		2Y test prediction error metrics	
Epoch	500	MAPE	10.6732
Batch size	20	RMSE	583.5753
Activity regularizer	Regularizers	MAE	440.4756
Activation function	ReLU	Accuracy	89.3300
Loss function	MSE		
Optimizer	Adam		

In the testing process, y test is used to compare with prediction value to generate the chart and the error metrics value. In this process, the error metrics calculated is not lower as the error metrics in training process. In Fig. 2, the pattern of the predicted value is close to the actual value. Most of the prediction value are higher than the actual value. Therefore, when real time analysis is carried on, the predicted value should be expected lower. The average dispersion of prediction from the actual values is 4.3430.

Like ANN model, the optimal number of epochs used in LSTM model is 500 epochs. This is because the LSTM model with 500 epochs number shows a lower error metrics value and higher accuracy than that of 1000 epochs number. The parameters used in LSTM test prediction is shown in Table 3. The visualization of prediction value and actual value is shown in Fig. 3.

In Fig. 3, the prediction value from 0 to 200 have a good prediction pattern compared with the actual value. However, start from 270, the prediction pattern starts to diverge from the actual value. The slope of the prediction value becomes flat. The predicted direction is same as the actual, but the degree of movement is much lower than the actual value. This causes the prediction trend starts to move



Fig. 3 LSTM test prediction with 500 epochs

Table 4 Parameters used of RF test prediction

Parameters		2Y test prediction error metrics	
Number of estimators	1000	MAPE	18.8758
Random state	42	RMSE	955.0128
Bootstrap	True	MAE	760.6988
Max features	8	Accuracy	81.1200
Min samples split	2		
Min samples leaf	1		
Max depth	20		
Max leaf nodes	None		

away from the actual trend. The average dispersion of prediction from the actual values is 440.4756.

The number of estimators used in test prediction of RF model is 1000. The parameters used in RF test prediction is shown in Table 4. The visualization of prediction value and actual value is shown in Fig. 4.

In Fig. 4, most of the value of the prediction value is maintained in a same value. This scenario is called extrapolation. To solve this problem, the time-series components of data should be ignored while training the RF model. However, this method is not applied as the methodology of building three models in this research are standardized. The data from 200 to 550 do not show a clear relationship between the prediction value and actual value. The mean dispersion of the prediction value with the actual value is very huge, it is 760.6988. Table 5 shows the error metrics and accuracy of each model in 2 years test prediction.

ANN model has the lowest MAPE, RMSE, and MAE in the 2 years test prediction. Although in the 8 years train prediction, LSTM has a better performance than ANN, but ANN shows a better performance than LSTM in the 2 years test prediction. ANN

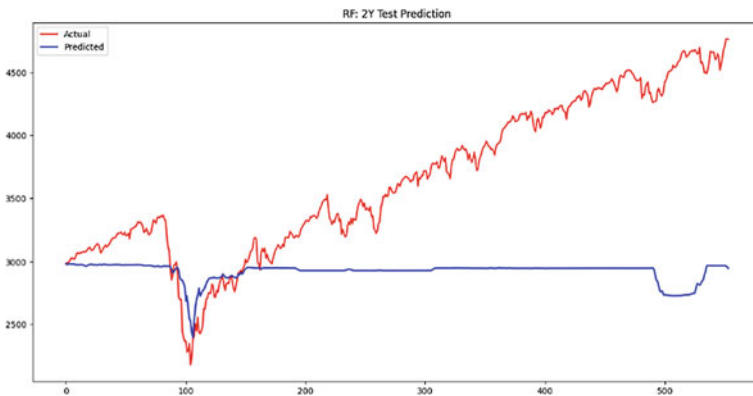


Fig. 4 RF test prediction with 1000 trees

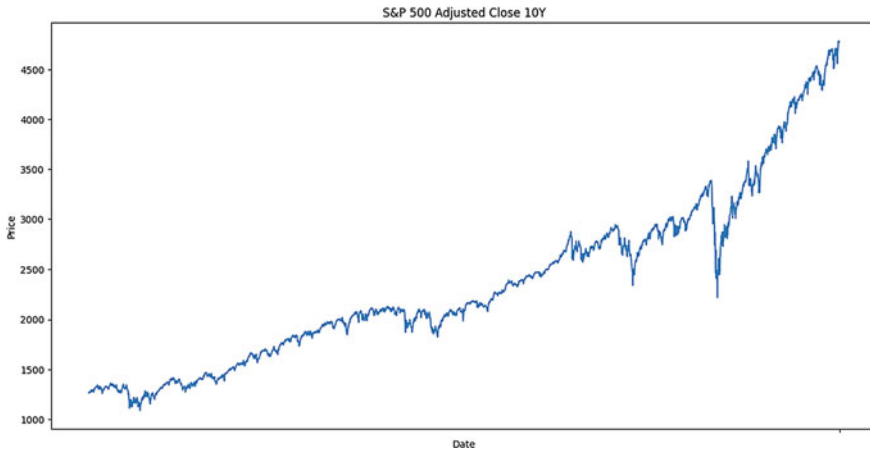


Fig. 5 S&P 500 adjusted close price for 10 years

Table 5 Error metrics and accuracy of three model in 2 years test prediction

	MAPE	RMSE	MAE	Accuracy
ANN	4.3430	187.3319	160.9507	95.6600
LSTM	10.6732	583.5753	440.4756	89.3300
RF	18.8758	955.0128	760.6988	81.1200

has an accuracy of 95.66% in the test prediction. The prediction trend and value are closer with the actual value compare with that of LSTM. In addition, ANN model does not show extrapolation, which happens in the RF. Therefore, ANN is the optimal model in predicting the future price movement of S&P 500.

Figure 5 shown that the price of S&P 500 is in an increasing trend for 10 years. From the visualization of the graph, there is a huge plunge of S&P 500 Price from February 19th 2020 to March 23rd 2020. The plunge of 34.45% for S&P 500 Price is due the global occurrence of the Covid-19 pandemic. US Government policy to overcome the serious issue of Covid-19 pandemic in nation was challenging the market psychology which places nationwide restriction on the financing benefits of public. S&P 500 is a market index which consists of technology, travel, leisure, transportation, and manufacturing services, and most of the services was forced to shut down temporary.

In general, the ANN model results in the best performance among the other two models in price prediction approach. As the result shown in Table 5, ANN model has the smallest error of MAPE of 4.3430 where consistent result of error can also been found in RMSE of 187.3319 which indicates that it has a lower range of errors with higher forecasting accuracy compared to other two models. The accuracy of the model is 95.66%, which is the highest among the three models. The illustration of the test prediction results from Table 2 to Table 4 displays that the error metrics and the

accuracy has the best performance. From Fig. 2 to Fig. 4, the ANN model has the best fit with slight error between the predicted value and the actual value, which could be visualized. On the other hand, the trendline of predicted value is approximately fit to the trendline of actual value, with the gaps of difference in value, due to the fundamental impact of the US Market. On the other hand, both RF and LSTM model show poor performance which are reflected in the plotted graph in Figs. 3 and 4.

5 Conclusion

Forecasting future price is not an easy task due to wide impact variables, nonlinear and complexity in financial market. However, with the evolution of computer science and artificial intelligence, machine learning become one of the famous techniques used by the investors to forecast the movement of financial derivatives' price. ANN is the best optimal model in predicting the S&P 500 adjusted close prices with the lowest error metrics and highest accuracy, compared with LSTM and RF models. ANN model shows an accurate price movement. To sum up, deep learning models are better in the prediction of stock index.

References

1. Hagenau M, Liebmann M, Neumann D (2013) Automated news reading: stock price prediction based on financial news using context-capturing features. *Decis Support Syst* 55(3):685–697
2. McGowan MJ (2010) The rise of computerized high frequency trading: use and controversy. *Duke Law and Technology Review*, 1–25
3. Chen Z, Goh HS, Sin KL, Lim K, Chung NKH, Liew XY (2021) Automated agriculture commodity price prediction system with machine learning techniques. *Adv Sci, Technol Eng Syst J* 6(2):1–8
4. Jareño F, Negrut L (2015) US stock market and macroeconomic factors. *J Appl Bus Res* 32(1):325–240
5. Gokmenoglu KK, Fazlollahi N (2015) The interactions among gold, oil, and stock market: evidence from S&P500. *Procedia Econ Financ* 25:478–488
6. Tuysuz S (2013) Conditional correlations between stock index, investment grade yield, high yield and commodities (gold and oil) during stable and crisis periods. *Int J Econ Financ* 5(9):28–44
7. Hsing Y, Hsieh WJ (2012) Impacts of macroeconomic variables on the stock market index in Poland: new evidence. *J Bus Econ Manag* 13(2):334–343
8. Hsing Y (2011) Impacts of macroeconomic variables on the U.S. stock market index and policy implications. *Econ Bull* 31(1):883–892
9. Danso EI (2020) Assessing the impact of macroeconomic variables on the performance of the U.S. stock market. *Res J Financ Account* 11(14):64–69
10. Balcilar M, Ozdemir ZA (2013) The causal nexus between oil prices and equity market in the U.S.: a regime switching model. *Energy Econ* 39:271–282
11. Ariyo AA, Adewumi AO, Ayo CK (2014) Stock price prediction using the ARIMA model. In: 2014 UKSim-AMSS 16th international conference on computer modelling and simulation. IEEE, Cambridge, UK, pp 106–112

12. Van Gerven M, Bohte S (2017) Artificial neural networks as models of neural information processing. *Front Comput Neurosci* 11:114
13. Sun W, Huang C (2020) A carbon price prediction model based on secondary decomposition algorithm and optimized back propagation neural network. *J Clean Prod* 243:118671
14. Yaghini M, Khoshraftar MM, Fallahi M (2013) A hybrid algorithm for artificial neural network training. *Eng Appl Artif Intell* 26(1):293–301
15. von Mettenheim HJ, Breitner MH (2012) Forecasting and trading the high-low range of stocks and ETFs with neural networks. In: Jayne C, Yue S, Iliadis L (eds) *Engineering applications of neural networks. EANN 2012. Communications in Computer and Information Science*, vol 311. Springer, Heidelberg, pp 423–432
16. Ding G, Qin L (2020) Study on the prediction of stock price based on the associated network model of LSTM. *Int J Mach Learn Cybern* 11:1307–1317
17. Van Houdt G, Mosquera C, Napoles G (2020) A review on the long short-term memory model. *Artif Intell Rev* 53:5929–5955
18. Sen J, Chaudhuri T (2017) A robust predictive model for stock price forecasting. In: *Proceedings of the 5th international conference on business analytics and intelligence (ICBAI 2017)*. Indian Institute of Management, Bangalore, India, pp 11–13
19. Yuan X, Yuan J, Jiang T, Ain QU (2020) Integrated long-term stock selection models based on feature selection and machine learning algorithms for China stock market. *IEEE Access* 8:22672–22685
20. Ma Y, Han R, Fu X (2019) Stock prediction based on random forest and LSTM neural network. In: *2019 19th international conference on control, automation and systems (ICCAS)*. IEEE, Jeju, Korea, pp. 126–130
21. Chen CC, Chen CH, Liu TY (2020) Investment performance of machine learning: analysis of S&P 500 Index. *Int J Econ Financ Issues* 10(1):59–66

Total Harmonic Distortion Study for Improvement of AC-AC Converter Under Buck-Type



Mohd. Shafie Bakar, Nurul Amira Ibrahim, and Abu Zaharin Ahmad

Abstract An AC-AC converter is a converter that is widely used in the industrial sector today because of its ability to convert AC power at high frequency. Although the AC-AC converter is commonly used in the industry, it also has problems producing smooth output. The presence of high THD_v at the output of AC-AC topology has a detrimental effect on the output of the system. The single-stage and double-stage of DC-modulated methodology were implemented, studied, compared, and analyzed under THD_v-based. It demonstrates an effective method for dealing with significant THD_v in AC-AC topologies and achieving low THD_v at the output voltage according to IEEE-519-1992 standard while keeping the high performance of efficiency at less than 5%.

Keywords Total Harmonics Distortion (THD) · Silicon Controlled Rectifier (SCR)

1 Introduction

The semiconductor switching device is a common condition that considerably impacts generating harmonics because of the switching behaviors mainly in AC-AC converters. The heat losses in the power system may be increased by the harmonics created by the AC-AC converter. These harmonics-related losses impair system efficiency, induce apparatus overheating, raise power coats, and harmonics current that impacts power electronics equipment. In traditional AC-AC topologies, as shown in Fig. 1 where two SCRs are being used. To perform a full AC voltage waveform, the PWM must trigger the SCR and generates unwanted harmonics [1]. To achieve power quality requirements, passive components are highly needed on the AC side of the power converter to reduce switching power frequency-related harmonics [2].

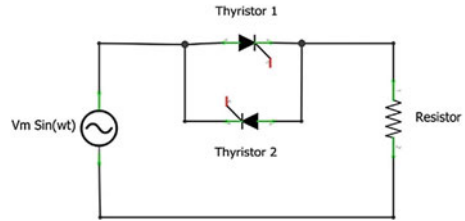
Thus, this paper is organized in the following steps. In Sect. 2, represent the circuit design of two different scenarios to evaluate the total harmonic distortion voltage (THD_v) and efficiency of the system that are separately being simulated. The first

Mohd. S. Bakar (✉) · N. A. Ibrahim · A. Z. Ahmad
FTKKEE, Universiti Malaysia Pahang, Pekan, Malaysia
e-mail: shafie@ump.edu.my

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
M. A. Abdullah et al. (eds.), *Advances in Intelligent Manufacturing and Mechatronics*,
Lecture Notes in Electrical Engineering 988,
https://doi.org/10.1007/978-981-19-8703-8_26

305

Fig. 1 Conventional AC-AC topologies



scenario covers single-stage DC-modulated AC-AC converters, whereas the second scenario covers double-stage DC-modulated AC-AC converters [3]. In Sect. 3, the results of both scenarios are presented, and an experiment is carried out to verify the suggested system. This section examines into the comparisons for both scenarios, and the obtained experiment result coincides well according to IEEE Std 519-1992 [4].

2 Methodology

This study comes out with the idea of applying bidirectional switches namely master switch (S_m) and slave switch (S_s) [5]. MOSFETs are being used for replacing the design of the switches. The simulation development is carried out from AC conventional topology as DC-modulated in AC-AC converter. The design parameter in this study are duty cycle and phase angle. There will be two scenarios in designing a circuit of DC-modulated AC-AC converter, which are, a single-stage and double-stage BUCK-Type AC-AC converter. Both scenarios will use the same value of parameters setting includes: $L = 10$ mH, $R = 150$ Ω , $C = 3$ μ F, $V_s = 200$, and V_{rms} with frequency = 50 Hz.

2.1 Single-Stage DC-Modulated BUCK-Type AC-AC Converter

Basically, in the traditional topology of AC-AC converters SCRs are being used as switching devices where user controlled the gate input terminal. SCRs work on two conditions; forwarding conduction mode and reverse blocking diode, which depends on the positive and negative cycle of the applied voltage. Figure 2 presents the schematic diagram that has been constructed in MATLAB/Simulink. First scenario depicts a single-stage DC-modulated AC-AC converter where the bidirectional switches are connected as illustrated in Fig. 3.

This sudden switching of a voltage yields a large number of harmonics or electrical noise. Such disturbance may disrupt any electrical operation [6]. So, replacing the SCRs with a better switching component can reduce the harmonics or any unpleasant

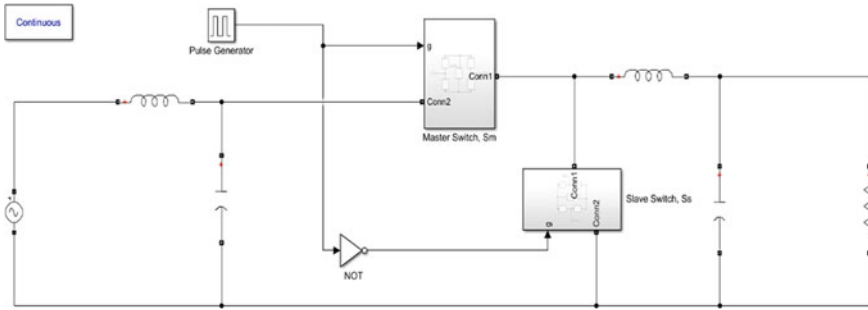
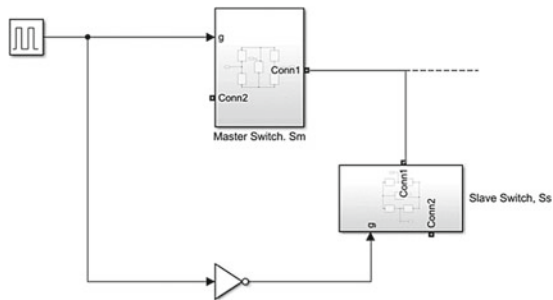


Fig. 2 Single-stage DC-modulated AC-AC converter

Fig. 3 Single-stage bidirectional switches connection



noise in the output system while maintaining the good efficiency of the system. In this scenario, there will be a pair of bidirectional switches connected in parallel.

2.2 Double-Stage BUCK-Type Converter

A DC-modulated double-stage Buck-Type AC/AC converter as shown in Fig. 4. The bidirectional switches will have double unit of Sm and Ss where Sm1 are series with Sm2 and Ss1 are series with Ss2 as can be seen in Fig. 5.

With the idea of applying bidirectional switches to replace the SCRs, the switching device can be MOSFETs or IGBTs. In this study, MOSFETs have been chosen for the design as shown in bidirectional switches for single-stage and double-stage. The term bidirectional means that there will be two units of a switch always connected. The operation of a bidirectional switch is that the master switch is controlled by a PWM that has an adjustable conduction duty cycle and it conducts the input current to forwardly flow in the positive input voltage. Next, it can conduct in the opposite direction where the input current is reversely flowing in the negative input voltage. Finally, the slave switch is conducted when the master switch is in off mode. A slave switch can be defined as a free-wheeling device to conduct the current flow [7].

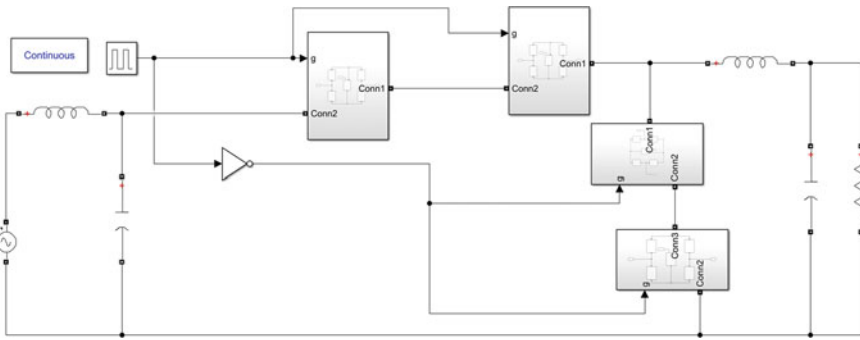


Fig. 4 Double-stage DC-modulated BUCK-Type

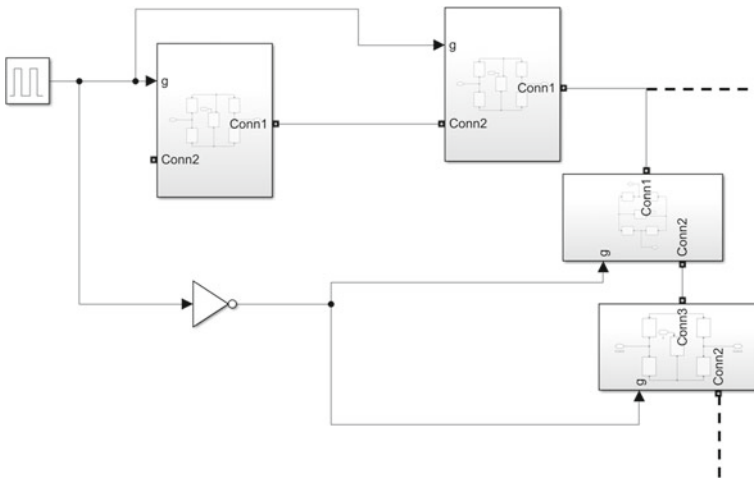


Fig. 5 Double-stage bidirectional switches connection

In addition, if certain converters require several bidirectional exclusive slave switches, the additional S_s just need to copy or repeat. If more than one bidirectional slave switch and one synchronously bidirectional slave switch are required by some converters, the synchronously bidirectional slave switch S_s can be built by simply copying or repeating the master switch S_m .

3 Result and Analysis

This section depicts the THD_v and efficiency of each design parameter for single-stage and double-stage at the selected duty cycle, which is 30% under phase delay of 45° for single-stage and 75% under phase delay of 45° for double-stage.

3.1 Single-Stage DC-Modulated BUCK-Type AC-AC Converter

Figure 6 shows a THDv under two duty cycles at 30 and 70%, respectively. In single-stage scenarios, the THDv is greater by 30% at the duty cycle, compared to 75%. The duty cycle 75% as illustrated in Fig. 7 of this scenario has high efficiency as compared to the duty cycle of 30%. The result confirms the association between duty cycle and output voltage when the duty cycle is low in a BUCK converter, the output voltage is low, and when the duty cycle is high, and the output voltage is high.

Fig. 6 THDv for single-stage DC-modulated BUCK-Type AC-AC converter

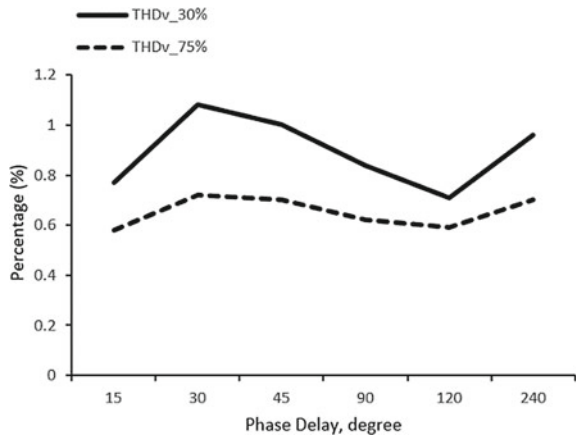
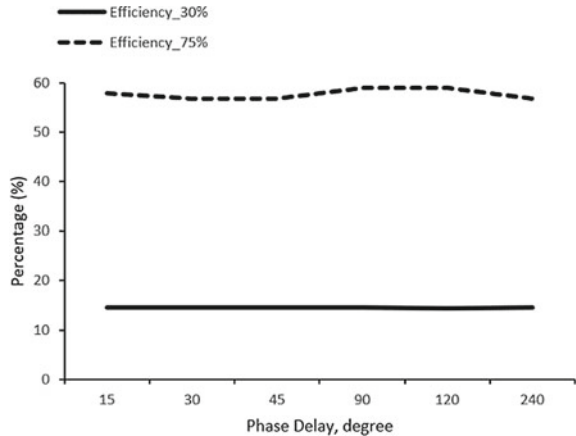


Fig. 7 Efficiency of single-stage DC-modulated BUCK-Type AC-AC converter



3.2 Double-Stage DC-Modulated BUCK-Type AC-AC Converter

In this part, the THDv and efficiency of a double-stage DC-modulated BUCK-Type AC-AC converter are observed. The effect of a double-stage DC-modulated BUCK-Type AC-AC converter on THDv and efficiency is shown in Figs. 8 and 9.

For double-stage DC-modulated AC-AC converter, the THDv have a constant reading for duty cycle 30% but at the duty cycle 75% there are small changes of THDv at phase delay 240°. The efficiency of double-stage DC-modulated is more stable compared with single-stage DC-modulated. It is because in double-stage it will process the bidirectional switches twice and it give more efficient and accurate reading for both THDv and efficiency.

Other studies [8] have found a link between duty cycle and THDv. As there is no reactive power, no capacitor, and no inductor in this suggested design, the power

Fig. 8 THDv for double-stage DC-modulated BUCK-Type AC-AC converter

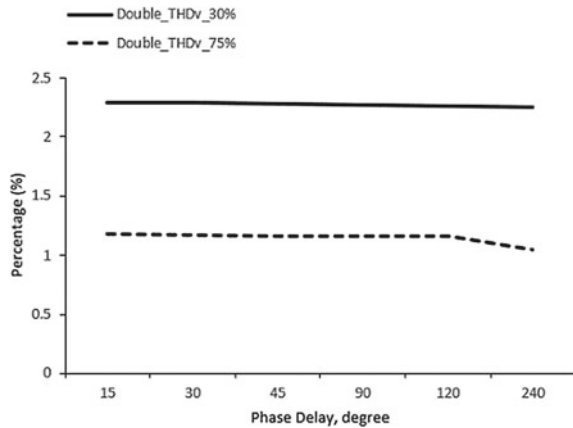
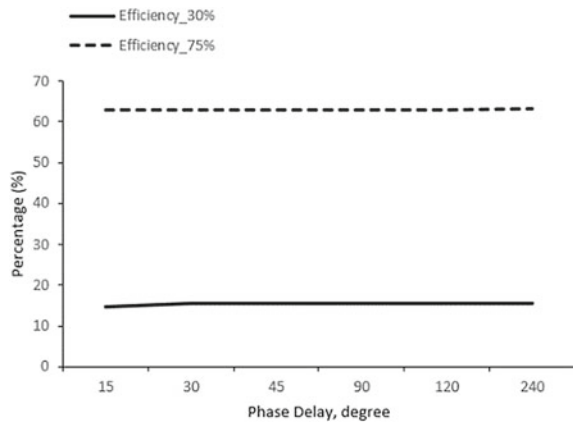


Fig. 9 Efficiency of double-stage DC-modulated BUCK-Type AC-AC converter



factor has no effect on the power system because the load is simply resistive where the power factor is extremely high.

4 Conclusion

The main idea of this study to resolve the un-smooth output voltage by viewing the THD_v as main indicator. This study successfully investigated and developed a novel DC-modulated method for single-stage and double-stage AC-AC converters. It has the ability to overcome the drawbacks of standard AC-AC converters, such as high THD_v. By using this approach, the THD_v can be reduced by adhering to IEEE STD 519–1992, which states that the THD_v for voltage supplies less than 69 kV should be less than 5%.

References

1. Schleicher M (2016) Thyristor power controller basic principles and tips for practitioners. JUMO GmbH & Co. KG, Germany
2. Suriadi (2006) Analysis of harmonics current minimization on power by Suriadi Thesis submitted in fulfillment of the requirements for the degree of Master of Science June 2006, no. June, 2006
3. Luo FL, Ye H (2007) Research on DC-Modulated power factor correction AC/AC converters, vol 00, pp 1478–1483
4. IEEE Recommended Practices and Requirements for Harmonic Control in Electric Power Systems, IEEE std 519–1992, Ieee, pp 1–9, 1992, [Online]. Available: <https://www.eaton.com/content/dam/eaton/products/backup-power-ups-surge-it-power-distribution/power-conditions/harmonic-correction-unit/IEEE-std-519-1992-harmonic-limits.pdf>.
5. Jiang H, Lu Y, Wu Y, Wang Y (2018) The strategy of inverter seamless mode switching in master-slave independent micro-grid. In: MATEC Web Conference, vol 160. <https://doi.org/10.1051/mateconf/201816004003>
6. Gulez K (2008) Neural network based switching control of AC-AC converter with DC-AC inverter for voltage sags, harmonics and EMI reduction using hybrid filter topology. *Simul Model Pract Theory* 16(6):597–612. <https://doi.org/10.1016/j.simpat.2008.03.001>
7. Ghate S, Juneja R, Tutakane DR, Debre PD (2018) DC-modulated buck type AC/AC converter for single phase induction motor drive. In: Proceedings of 2017 international conference on innovations in information, embedded and communication systems ICII ECS 2017, vol 2018-Janua, pp 1–5, 2018. <https://doi.org/10.1109/ICII ECS.2017.8276112>
8. Elgack W, Huang Shen C, Albert FYC, Mohd Fuad MN (2015) Analysis of total harmonic distortion and power consumption in AC to AC voltage convertor using integral cycle PWM control for fast pyrolysis. In: 2015 international conference on computer, communications, and control technology (I4CT), no. I4ct, pp 437–440. <https://doi.org/10.1109/I4CT.2015.7219614>

Training Feedforward Neural Networks Using Arithmetic Optimization Algorithm for Medical Classification



Koon Meng Ang, Wei Hong Lim, Sew Sun Tiang, Hameedur Rahman, Chun Kit Ang, Elango Natarajan, Mohamed Khan Afthab Ahamed Khan, and Li Pan

Abstract Feedforward neural network (FNN) is popular machine learning technique widely implemented for image classification, data clustering, object recognition, etc. due to its outstanding capability in processing data. Backpropagation is commonly employed as a conventional method to adjust the weights and biases of FNNs. As a gradient-based algorithm, backpropagation tends to have slow convergence rate and highly dependent on the initial solution generated. Arithmetic optimization algorithm (AOA) emerges as a promising metaheuristic search algorithm (MSA) to replace conventional method in training FNNs due to its outstanding global search ability. In this paper, AOA is designed to optimize the weights, biases and selection of activation functions of FNN for image classification. Medical datasets are extracted from UCI Machine Learning Repository to assess the capability of AOA in training FNN. Comparative studies report that the promising performance AOA in training FNN for medical classification.

Keywords Feedforward neural network · Machine learning · Arithmetic optimization algorithm · Classification

1 Introduction

Artificial neural network (ANN) is a machine learning method inspired by human nervous system in simulating the cerebral cortex of brain structure. ANNs that were proposed by researchers can be categorized into several types known as feedforward neural network (FNN), convolutional neural network (CNN), deep neural network (DNN), recurrent neural network (RNN), etc. FNN is one of the most commonly

K. M. Ang · W. H. Lim (✉) · S. S. Tiang · C. K. Ang · E. Natarajan · M. K. A. A. Khan · L. Pan
Faculty of Engineering, Technology and Built Environment, UCSI University, 56000 Kuala Lumpur, Malaysia
e-mail: limwh@ucsiuniversity.edu.my

H. Rahman
Faculty of Computing and Artificial Intelligence, Air University, Islamabad Capital Territory, Islamabad 44000, Pakistan

known ANNs due to its outstanding performance in performing machine learning tasks and simplicity of network architecture [1]. Generally, FNN structure can be separated into three main parts, known as input, hidden, and output layers. The information extracted from the input data is processed in each layer and transferred in one-way from the input layer to hidden layer, followed by the output layer [1]. In each neuron, an activation function is implemented to transform the summed weighted input received by the neuron into nonlinear output. This nonlinear characteristic plays a crucial role for a FNN model to have competitive performance in tackling machine learning tasks with different complexity level such as functions approximations, data predictions [2]. However, a training process is required to obtain the best combination of biases and weights of each neuron in a FNN model by optimizing an objective function that measures the mean differences between the obtained and expected results [3].

Backpropagation (BP) is a conventional gradient-based algorithm employed to train FNN models [1]. However, several studies reported that BP tends to restrict the performance of FNN by producing suboptimal solutions of weights and biases in the training process [1, 3, 4]. Metaheuristic search algorithms (MSAs) are envisioned as potential solutions in optimizing the weights and biases of FNN due to its excellent global search ability in tackling various optimization problems [5–23] with high convergence rate. Arithmetic optimization algorithm (AOA) [24] is a state-of-the-art MSA motivated by the distribution behavior of the arithmetic operators, known as addition, subtraction, multiplication, and division. Different strengths of explorative and exploitative behaviors adopted in AOA tend to prevent the loss of solution diversity in the early stage and promote the convergence of possible solutions toward the global optima in the late stage of searching process, respectively. Nevertheless, AOA was initially proposed to solve only benchmark functions and its capability to identify the optimal combination of weights and biases of FNN for medical classification remain questionable.

In this paper, AOA is employed as a training algorithm to optimize the weights, biases, and selection of activation functions of FNN model. The main contributions of this study are highlighted as follows:

1. An optimization model is formulated to facilitate the searching of optimal combination of weights, biases, and activation functions of FNN model.
2. AOA is employed to train the FNN classifier, aiming to enhance its classification performance when dealing with medical classification tasks.
3. The performance of AOA and other six MSAs in training FNN classifier for solving medical classification tasks are assessed by using medical datasets extracted from UCI Machine Learning Repository.

For remaining sections of this paper, related works are summarized in Sect. 2, followed by Sect. 3 that describes the mechanisms of training FNN model with AOA. The performances of all MSAs in training FNN for medical classification are compared in Sect. 4. The conclusion and future study are summarized in Sect. 5.

2 Related Works

2.1 Inspiration of Arithmetic Optimization Algorithm

Two typical behaviors known as exploration and exploitation can be observed from MSAs during optimization. Exploration enables MSAs to have wide coverage of search space with good solution diversity, whereas exploitation refines solutions obtained around global optima. Imbalanced of exploration or exploitation strengths can result in premature convergence of algorithm due to the entrapment of solutions in local optima.

Arithmetic optimization algorithm (AOA) is a MSA inspired by the distribution behavior of various arithmetic operators known as addition, subtraction, multiplication, and division [24]. The exploration and exploitation behaviors of AOA were achieved by the arithmetic operators, where the search behavior was adaptively selected in different stages of the optimization process. Specifically, multiplication and division operators were formulated to promote the exploration strengths, while addition and subtraction operators were formulated to enhance the exploitation strengths of AOA. In order to adaptively select the search behavior in different stages of optimization process, a Math Optimizer Accelerated (*MOA*) function was designed as follow:

$$MOA_{\tau} = A^{\min} + \tau \times \left(\frac{A^{\max} - A^{\min}}{\tau^{\max}} \right) \quad (1)$$

where τ is current fitness evaluation number within the range of $[1, \tau^{\max}]$; A^{\min} and A^{\max} represent the lowest and highest values of the function, respectively.

2.2 Approaches of FNN Training Using MSAs

A hybridized particle swarm optimization (PSO) and steepest descent algorithm [25] was introduced to address overfitting issues when optimizing the FNN models. Specifically, PSO was implemented to locate the best combination of FNN structures by reducing the learning error function, while the steeped descent was incorporated to enhance the training accuracy and speed by fine-tuning the weights obtained from the global best solution. A hybridized PSO with gravitational search algorithm (GSA) [3] was introduced to optimize the FNN model in terms of its weights and biases, where PSO operators were employed to address the poor convergence issue of GSA.

Besides classification tasks, MSAs were also applied to optimize FNN structures for tackling real-world prediction problems. In [26], teaching-learning-based optimization (TLBO) was introduced to optimize FNN model to estimate the capacity of both driven and drilled shaft piles implanted in un-cemented soils. In [27], an

improved TLBO was introduced to forecast building energy usage. In [28], a TLBO variant with modified learning phases was proposed to optimize FNN model, in terms of weights, biases, and activation functions, to solve classification problems. These previous studies suggested the potential of MSAs as promising alternative to train FNN models robustly.

3 AOA-Optimized FNN Model

3.1 Formulation FNN Training Optimization Problem

A typical FNN structure consists of three layers, known as input, hidden, and output layers, with P input neurons, Q hidden neurons, and S output neurons, respectively, is illustrated in Fig. 1. Let I_p , H_q , and O_s be the p th input neuron, q -th hidden neuron, and s -th output neuron, respectively, where $p = 1, \dots, P$, $q = 1, \dots, Q$, and $s = 1, \dots, S$. The weight between I_p and H_q is indicated as $W_{p,q}^H$; the weight between H_q and O_s is indicated as $W_{q,s}^O$. Denote B_q^H and B_s^O as the biases of q -th hidden neuron and s -th output neuron, respectively. The values of each H_q and O_s are calculated by the sum of input weight and bias. Subsequently, non-linearization process of these weighted summation is performed by using an activation function $\Phi(\cdot)$ as follow:

$$H_q = \Phi \left(\sum_{p=1}^P W_{p,q}^H I_p + B_q^H \right) \tag{2}$$

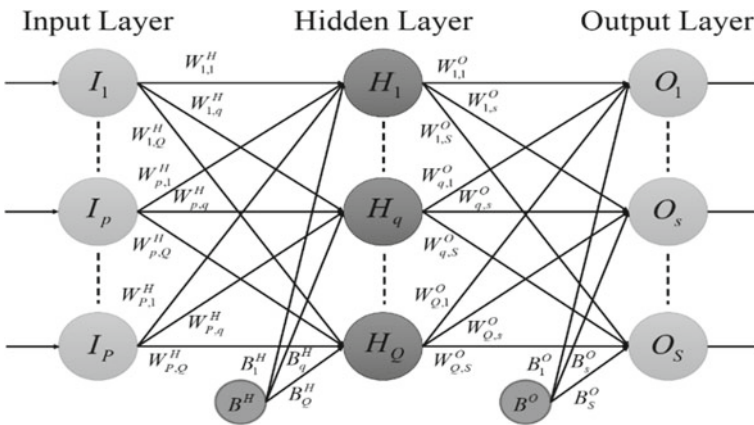


Fig. 1 FNN model with three layers

$$O_s = \Phi \left(\sum_{q=1}^Q W_{q,s}^O H_q + B_s^O \right) \tag{3}$$

The FNN training optimization problem considered in this study not only considers the optimal combination of weights and biases, but also search for best activation functions for specific datasets. Hence, the dimensional components considered by AOA in training FNN are separated into three categories, i.e., (i) weights $W_{p,q}^H, W_{q,s}^O \in [-1, 1]$, (ii) biases $B_q^H, B_s^O \in [-1, 1]$, and (iii) index of activation function $K = \{1, 2, 3, 4, 5\}$, where the values of 1, 2, 3, 4, and 5 denote the binary step, uni-polar sigmoid [2], hyperbolic tangent [2], inverse tangent, and rectified linear unit (ReLU) [29], respectively. The encoding of AOA solution (i.e., X) and its dimensional sizes are given as:

$$X = [W_{1,1}^H, \dots, W_{p,q}^H, \dots, W_{p,Q}^H, \dots, W_{1,1}^O, \dots, W_{q,s}^O, \dots, W_{Q,S}^O, B_1^H, \dots, B_q^H, \dots, B_Q^H, \dots, B_1^O, \dots, B_S^O, \dots, B_S^O, K] \tag{4}$$

$$D = P \cdot Q + Q \cdot S + Q + S + 1 \tag{5}$$

In this study, the objective function $f(X)$ is defined by measuring the mean square error ε^{ms} between the obtained and expected output values produced by AOA in training FNN model. The obtained output of g -th data sample produced by FNN structure optimized by candidate solution of AOA and the corresponding expected output of g -th data sample extracted from the dataset are \hat{Y}_g and Y_g , respectively, where $g = 1, \dots, G$. The ε^{ms} value obtained by the FNN model trained using solution X are computed as:

$$f(X) = \varepsilon^{ms}(X) = \frac{1}{G} \sum_{g=1}^G (\hat{Y}_g(X) - Y_g)^2 \tag{6}$$

Notably, the FNN training optimization is a minimization problem because it aims to reduce ε^{ms} that can lead to the increasing of classification accuracy of FNN model.

3.2 Arithmetic Optimization Algorithm (AOA)

At the beginning of the optimization process after the population is initialized randomly, the MOA function as expressed in Eq. (1) is used to select the searching phase between exploration phase (i.e., multiplication and division) and exploitation phase (i.e., addition and subtraction). If the MOA value is greater than a random number r_1 produced by uniform distribution within 0–1, the explorative operators are selected to update each candidate solution in the search space. Otherwise, the exploitative operators are selected.

In exploration phase, the division and multiplication operators share the same probability to be selected to calculate the new d -th dimension of n th solution $X_{n,d}^{new}$ as follow:

$$X_{n,d}^{new} = \begin{cases} X_d^{best} \div (MOP + \chi) \times ((X_d^U - X_d^L) \times 0.5 + X_d^L), & r_2 < 0.5 \\ X_d^{best} \times MOP \times ((X_d^U - X_d^L) \times 0.5 + X_d^L), & otherwise \end{cases} \quad (7)$$

where X_d^{best} refers to the d th dimensional component of best-found solution, where $d = 1, \dots, D$; χ indicates a small integer number; X_d^U and X_d^L represent the upper and lower boundaries of d -th dimensional component, respectively; r_2 refers to a number randomly generated by uniform distribution within 0–1. Given the current fitness evaluation number τ and the maximum fitness evaluation number τ^{\max} , the MOP value is calculated as follow:

$$MOP_{\tau} = 1 - \frac{\tau^{0.25}}{(\tau^{\max})^{0.25}} \quad (8)$$

Similar to exploration phase, the subtraction and addition operators share the same probability to be selected to calculate the new d th dimension of n th solution $X_{n,d}^{new}$ in exploitation phase as follow:

$$X_{n,d}^{new} = \begin{cases} X_d^{best} - MOP \times ((X_d^U - X_d^L) \times 0.5 + X_d^L), & r_3 < 0.5 \\ X_d^{best} + MOP \times ((X_d^U - X_d^L) \times 0.5 + X_d^L), & otherwise \end{cases} \quad (9)$$

The overall framework of AOA is described in Fig. 2. At the initialization stage, the population of AOA is randomly generated and the associated fitness values (i.e., MSE values) are calculated. For each fitness evaluation τ , the MOA value is firstly calculated to select the searching phase. Then, either exploration or exploitation phases is selected to update the candidate solution. The optimization process is repeated until the termination criterion $\tau > \tau^{\max}$ is fulfilled. The best solution obtained at the end of the optimization process is considered as the best combination of weights, biases, and activation function for the FNN model in classifying the given dataset.

4 Performance Comparison in Training FNN Models

In this study, classification accuracy rate R^C is incorporated to assess the classification performance of an FNN model. A greater value of R^C produced by an FNN model indicates that the FNN model has better performance in dealing with classification problems by producing higher accuracy rate. Assume that R^{σ} and R^T refer to the number of correctly classified data and the total number of data samples, respectively, the R^C value is computed as follow:

Algorithm 1: AOA	
Input:	N, D, X^U, X^L
01:	Initialize $\tau = 0$;
02:	for $n = 1$ to N do
03:	Randomly generated candidate solution X_n ;
04:	Evaluated $f(X_n)$ using Eq. (6);
05:	$\tau \leftarrow \tau + 1$;
06:	end
07:	while $\tau \leq \tau^{\max}$ do
08:	Calculate MOP using Eq. (1);
09:	Calculate MOA using Eq. (8);
10:	for $n = 1$ to N do
11:	for $d = 1$ to D do
12:	if $r_1 < MOA$ then /*Exploration*/
13:	Calculate $X_{n,d}^{new}$ using Eq. (7);
14:	else /*Exploitation*/
15:	Calculate $X_{n,d}^{new}$ using Eq. (9);
16:	end if
17:	end for
18:	Evaluated $f(X_n^{new})$ using Eq. (6);
19:	Update $X_n, f(X_n), X^{best}$, and $f(X^{best})$;
20:	$\tau \leftarrow \tau + 1$;
21:	end for
22:	end while
Output:	X^{best}

Fig. 2 Overall framework of AOA

$$R^C = \frac{R^\sigma}{R^T} \times 100\% \quad (10)$$

The performance of AOA and six existing MSAs, known as conventional PSO [30], conventional TLBO [31], chaotic-opposition-based hybridized differential evolution with TLBO (COHDEPSO) [32], single-objective version of improved TLBO (ITLBO) [33], gorilla troops optimizer [34], and artificial hummingbird algorithm (AHA) [35], in solving ten medical datasets extracted from UCI Machine Learning Repository, known as (a) New Thyroid, (b) Liver Disorder, (c) Breast Cancer, (d) Hepatitis, (e) Haberman's Survival, (f) Blood Transfusion, (g) Primary Tumor, (h) Lymphography, (i) Cervical Cancer, and (j) Fertility, is investigated. The information of each selected dataset, in terms of number of attributes, number of classes, and number of samples, are summarized in Table 1. All the datasets are randomly extracted into two parts for training and testing with ratio of 80 and 20%,

Table 1 Properties of ten selected datasets for FNN training

Datasets	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)
# Attributes	5	6	9	19	3	4	17	18	19	9
# Classes	3	2	2	2	2	2	2	8	2	2
# Samples	215	345	277	80	306	748	131	148	72	100

respectively. The parameters of all compared methods are configured based on the recommended settings of their corresponding literatures. The FNN models produced by each method are simulated for 30 independent runs in solving all datasets with maximum fitness evaluation numbers of $\tau^{\max} = 10,000 \times D$. The simulations are conducted on a workstation installed with Intel® Core i9-10900 K CPU @ 3.70 GHz and Matlab 2021a.

The R^C values obtained by the FNN models constructed by AOA and six other MSAs in solving training and testing samples are reported in Tables 2 and 3, respectively. In both Tables 1 and 2, the R^C values highlighted with boldface implying the best R^C values, while the second-best R^C values are underlined. Moreover, # BR^C and $w/t/l$ are used to summarize the overall performance of each algorithm. Specifically, # BR^C indicates the number of best R^C values obtained by each method. Meanwhile, $w/t/l$ implies AOA has better performance than the compared method in w function, tie in t function, and worse performance in l function.

Tables 2 and 3 reported that AOA is able to produce six best R^C in solving 10 training and testing datasets, respectively, implying its best performance among the compared methods. In Table 2, PSO also shows its competitive performance in training FNN model to classify training datasets by producing four best and four second-best R^C out of ten datasets. In Table 3, COHDEPSO is reported to produce

Table 2 Training R^C values produced by all compared algorithms

Datasets	Train	AOA	PSO	TLBO	COHDEPSO	ITLBO	GTO	AHA
(a)	R^C	99.9	<u>98.5</u>	95.2	87.9	90.9	83.9	85.4
(b)	R^C	58.8	62.2	57.7	<u>58.9</u>	57.9	57.5	57.5
(c)	R^C	71.7	73.2	68.8	<u>73.0</u>	63.8	72.3	69.8
(d)	R^C	<u>82.8</u>	87.0	75.0	81.4	65.4	73.9	69.5
(e)	R^C	73.0	<u>72.6</u>	71.5	72.5	71.8	71.9	72.1
(f)	R^C	79.1	78.8	78.9	78.7	<u>78.9</u>	78.5	78.8
(g)	R^C	<u>85.1</u>	88.4	76.3	82.3	74.4	76.8	73.8
(h)	R^C	40.3	38.5	38.8	<u>39.0</u>	38.8	38.0	36.1
(i)	R^C	96.7	<u>95.3</u>	87.9	93.4	80.5	84.6	71.2
(j)	R^C	94.8	<u>90.6</u>	88.3	88.7	88.1	87.2	87.3
	# BR^C	6	4	0	0	0	0	0
	$w/t/l$	–	6/0/4	10/0/0	8/0/2	10/0/0	9/0/1	10/0/0

Table 3 Testing R^C values produced by all compared algorithms

Datasets	Test	AOA	PSO	TLBO	COHDEPSO	ITLBO	GTO	AHA
(a)	R^C	91.4	<u>67.2</u>	56.0	45.5	44.6	42.0	45.1
(b)	R^C	49.6	44.7	<u>48.2</u>	41.0	47.9	47.5	45.0
(c)	R^C	66.6	<u>66.0</u>	64.3	63.8	65.0	61.0	61.6
(d)	R^C	56.9	<u>60.0</u>	55.0	61.2	45.0	57.5	50.0
(e)	R^C	76.7	74.9	75.7	76.0	74.1	72.6	<u>76.2</u>
(f)	R^C	54.9	48.4	59.6	64.3	<u>61.4</u>	61.4	61.4
(g)	R^C	65.4	56.1	<u>65.0</u>	<u>65.0</u>	60.0	61.9	64.2
(h)	R^C	36.7	<u>36.3</u>	<u>36.3</u>	35.3	35.6	34.6	34.0
(i)	R^C	<u>85.7</u>	90.5	<u>85.7</u>	80.7	72.8	80.0	82.8
(j)	R^C	84.5	85.5	81.5	85.5	85.5	<u>88.0</u>	90.0
	$\#BR^C$	6	1	0	2	0	0	1
	$w/t/l$	–	7/0/3	8/1/1	7/0/3	8/0/2	7/0/3	8/0/2

two best and one second-best R^C , implying its optimized-FNN model is competitive in classifying testing datasets. This is followed by PSO and AHA that produce one best R^C each. It is notable that the FNN model trained by ITLBO has the worst overall performance in both training and testing cases. It is reasonable because ITLBO was initially proposed to solve multi-objective problems.

5 Conclusions

In this paper, the searching of best combination of weights, biases, and activation functions are formulated as a single-objective optimization problems to be optimized by MSAs. AOA is employed as a training algorithm of FNN model, aiming to enhance its classification performance in dealing with medical image datasets extracted from UCI Machine Learning Repository. Performance comparisons concluded that AOA has the best performance in training FNN model to solve both training and testing datasets.

As potential future works, novel searching mechanisms can be incorporated into AOA to investigate its performance in dealing with optimization problems with higher complexity level. The convergence characteristics of AOA is worth to be studied also by employing extensive theoretical framework. Finally, the performance of AOA in conducting neural architecture search of convolutional neural networks is another research direction.

Acknowledgements This work was supported by the Ministry of Higher Education Malaysia under the Fundamental Research Schemes with project codes of FRGS/1/2019/TK04/UCSI/02/1 and FRGS/1/2020/TK0/UCSI/02/4. This work is also supported by the UCSI University Research Excellence & Innovation Grant (REIG) with project code of REIG-FETBE-2022/038.


References

1. Wu H, Zhou Y, Luo Q, Basset MA (2016) Training feedforward neural networks using symbiotic organisms search algorithm. *Comput Intell Neurosci*
2. Feng J, Lu S (2019) Performance analysis of various activation functions in artificial neural networks. *J Phys Conf Ser* 022030. IOP Publishing
3. Mirjalili S, Hashim SZM, Sardroudi HM (2012) Training feedforward neural networks using hybrid particle swarm optimization and gravitational search algorithm. *Appl Math Comput* 218:11125–11137
4. Tarkhaneh O, Shen H (2019) Training of feedforward neural networks for data classification using hybrid particle swarm optimization, Mantegna Lévy flight and neighborhood search. *Heliyon* 5:e01275
5. Ang KM, Lim WH, Isa NAM, Tiang SS, Ang CK, Natarajan E, Solihin MI (2020) A constrained teaching-learning-based optimization with modified learning phases for constrained optimization. *J Adv Res Dyn Control Syst* 12:15
6. Chong OT, Lim WH, Isa NAM, Ang KM, Tiang SS, Ang CK (2020) A teaching-learning-based optimization with modified learning phases for continuous optimization. In: *Science and information conference*. Springer, Berlin, pp 103–124
7. Koh WS, Lim WH, Ang KM, Isa NAM, Tiang SS, Ang CK, Solihin MI (2022) Multi-objective particle swarm optimization with alternate learning strategies. Springer Singapore, pp 15–25
8. Ang KM, Lim WH, Isa NAM, Tiang SS, Ang CK, Chow CE, Yeap ZS (2022) Modified particle swarm optimization with unique self-cognitive learning for global optimization problems. Springer, Singapore, pp 263–274
9. Cheng W-L, Ang KM, Choi ZC, Lim WH, Tiang SS, Natarajan E, Ang CK, Khan MKAA (2022) Particle swarm optimization with modified initialization scheme for numerical optimization. In: *Proceedings of the 6th international conference on electrical, control and computer engineering*. Springer, Berlin, pp 497–509
10. Natarajan E, Kaviarasan V, Ang KM, Lim WH, Elango S, Tiang SS (2022) Production wastage avoidance using modified multi-objective teaching learning based optimization embedded with refined learning scheme. *IEEE Access* 1–1
11. Ang KM, Juhari MRM, Cheng W-L, Lim WH, Tiang SS, Wong CH, Rahman H, Pan L (2022) New particle swarm optimization variant with modified neighborhood structure
12. Ang KM, Juhari MRM, Lim WH, Tiang SS, Ang CK, Hussin EE, Pan L, Chong TH (2022) New hybridization algorithm of differential evolution and particle swarm optimization for efficient feature selection. 27:5
13. Voon YN, Ang KM, Chong YH, Lim WH, Tiang SS (2022) Computer-vision-based integrated circuit recognition using deep learning. In: *Proceedings of the 6th international conference on electrical, control and computer engineering*. Springer, Berlin, pp 913–925
14. Suresh S, Elango N, Venkatesan K, Lim WH, Palanikumar K, Rajesh S (2020) Sustainable friction stir spot welding of 6061–T6 aluminium alloy using improved non-dominated sorting teaching learning algorithm. *J Market Res* 9:11650–11674
15. Yao L, Lim WH (2017) Optimal purchase strategy for demand bidding. *IEEE Trans Power Syst* 33:2754–2762
16. Yao L, Lai C-C, Lim WH (2015) Home energy management system based on photovoltaic system. In: *2015 IEEE international conference on data science and data intensive systems*. IEEE, pp 644–650
17. Yao L, Chen Y-Q, Lim WH (2015) Internet of things for electric vehicle: an improved decentralized charging scheme. In: *2015 IEEE international conference on data science and data intensive systems*. IEEE, pp 651–658
18. Yao L, Lim WH, Tiang SS, Tan TH, Wong CH, Pang JY (2018) Demand bidding optimization for an aggregator with a genetic algorithm. *Energies* 11:2498
19. Ahmad MF, Isa NAM, Lim WH, Ang KM (2022) Differential evolution with modified initialization scheme using chaotic oppositional based learning strategy. *Alex Eng J* 61:11835–11858

20. Muhieldeen MW, Lye LC, Kassim M, Yen TW, Teng K (2021) Effect of rockwool insulation on room temperature distribution. *J Adv Res Exp Fluid Mech Heat Transfer* 3:9–15
21. Muhieldeen M, Lim Y, Govinda S, Tey WY (2020) Investigation of the effect of awning using sunlight sensor to reduce cooling load in the room. *J Adv Res Fluid Mech Thermal Sci* 67:136–145
22. Yu L-J, Rengasamy K, Lim K-Y, Tan L-S, Mouad' AT, Zulkoffli ZB, Yong ENS (2019) Comparison of activated carbon and zeolites' filtering efficiency in freshwater. *J Environ Chem Eng* 7:103223
23. Yu L-J, Ahmad SH, Tarawneh MA, Abd Razak SBB, Natarajan E, Ang CK (2019) Magnetic, thermal stability and dynamic mechanical properties of beta isotactic polypropylene/natural rubber blends reinforced by NiZn ferrite nanoparticles. *Defence Technol* 15:958–963
24. Abualigah L, Diabat A, Mirjalili S, Abd Elaziz M, Gandomi AH (2021) The arithmetic optimization algorithm. *Comput Methods Appl Mech Eng* 376:113609
25. Kandasamy T, Rajendran R (2018) Hybrid algorithm with variants for feed forward neural network. *Int Arab J Inf Technol* 15:240–245
26. Benali A, Hachama M, Bounif A, Nechnech A, Karray M (2019) A TLBO-optimized artificial neural network for modeling axial capacity of pile foundations. *Eng Comput* 1–10
27. Li K, Xie X, Xue W, Dai X, Chen X, Yang X (2018) A hybrid teaching-learning artificial neural network for building electrical energy consumption prediction. *Energy Build* 174:323–334
28. Ang KM, Lim WH, Tiang SS, Ang CK, Natarajan E, Ahamed Khan M (2022) Optimal training of feedforward neural networks using teaching-learning-based optimization with modified learning phases. In: *Proceedings of the 12th national technical seminar on unmanned system technology*. Springer, Berlin, pp 867–887
29. Lin G, Shen W (2018) Research on convolutional neural network based on improved Relu piecewise activation function. *Procedia Comput Sci* 131:977–984
30. Kennedy J, Eberhart R (1995) Particle swarm optimization. In: *Proceedings of ICNN'95—international conference on neural networks*, vol 1944, pp 1942–1948
31. Rao RV, Savsani VJ, Vakharia DP (2011) Teaching-learning-based optimization: a novel method for constrained mechanical design optimization problems. *Comput Aided Des* 43:303–315
32. Choi ZC, Ang KM, Lim WH, Tiang SS, Ang CK, Solihin MI, Juhari MRM, Chow CE (2021) Hybridized metaheuristic search algorithm with modified initialization scheme for global optimization. In: *Advances in robotics, automation and data analytics: selected papers from ICITES 2020*, vol 1350, p 172
33. Patel VK, Savsani VJ (2016) A multi-objective improved teaching-learning based optimization algorithm (MO-ITLBO). *Inf Sci* 357:182–200
34. Abdollahzadeh B, Soleimani Gharehchopogh F, Mirjalili S (2021) Artificial gorilla troops optimizer: a new nature-inspired metaheuristic algorithm for global optimization problems. *Int J Intell Syst* 36:5887–5958
35. Zhao W, Wang L, Mirjalili S (2022) Artificial hummingbird algorithm: a new bio-inspired optimizer with its engineering applications. *Comput Methods Appl Mech Eng* 388:114194

Various Type of Crops and Trees Detection Using Clustering Technique Through Image Processing



Mohd Izzat Mohd Rahman, Mohd Azraai Mohd Razman ,
Ismail Mohd Khairuddin, Anwar P. P. Abdul Majeed,
Muhammad Amirul Abdullah, and Wan Hasbullah Mohd Isa

Abstract Based on ResNet-18 and MobileNetV1, this research created a custom identification neural network to detect chili, eggplant, and potato for the NVIDIA® Jetson Nano Developer Kit. This research will aim to implement drones, automatic machines, or autonomous systems in the agricultural sector, from the farm, orchard, or greenhouse to the consumer. Both neural networks go through two different epochs: 30 and 2000, with the same dataset quantity of 30 images for each crop selected. According to the experiment results, the ResNet-18 can identify chili with 94.79% accuracy, eggplant with 47.04 percent accuracy, and potato with 38.74% accuracy. While the MobileNetV1 can identify the chili with 99.8% accuracy, the eggplant with 99.9% accuracy, and the potato with 100% accuracy.

Keywords SSD · ResNet-18 · MobileNetV1 · Machine learning

1 Introduction

Agriculture is one of Malaysia's commodities with a high economic value and a high development potential, and this is because Malaysia is one of the countries that not only produces and consumes vegetables but also exports them to other countries, with a total value of RM114,451 million in 2018 [1]. As a result, crop and tree farming is one of the most important sources of national income.

Proper and delicate horticulture is required to maintain the quality of crops and trees if Malaysia wants to be more competitive among crop producers worldwide. Malaysia is rich in crop and tree varieties, requiring the expertise of experienced farmers and nurseries. However, relying solely on humans to maintain and care for

M. I. M. Rahman · M. A. M. Razman (✉) · I. M. Khairuddin · M. A. Abdullah · W. H. M. Isa
Faculty of Manufacturing and Mechatronics Engineering Technology, Universiti Malaysia Pahang,
26600 Pekan, Pahang, Malaysia
e-mail: mohdazraai@ump.edu.my

A. P. P. A. Majeed
School of Robotics, XJTLU Entrepreneur College (Taicang), Xi'an Jiaotong-Liverpool
University, Suzhou 215123, P. R. China

farms is insufficient; thus, technology can either take over or assist farmers and nursery specialists in maintaining crop quality.

After decades of semiconductor innovation, technology is becoming less expensive but better in output quality. This trend is also affecting camera build quality design, which is robust and rigid in addition to being waterproof, and image processing systems, which are becoming sharp and high quality. Since then, cameras have been used as surveillance devices to monitor the surrounding area, in sports, and on farms [2–4]. With the rapid advancements of artificial intelligence, the transfer learning process has transformed the camera function usually used to record video and capture picture moments for memories into a self-aware machine learning system [5].

The machine learning system can identify various crops or trees and their health status by manipulating the sensors and camera positions. Each sensor position has advantages and disadvantages that must be explored further. The image processing algorithm will process the captured video and image in the machine learning system.

2 Neural Networks

2.1 Overview

Over time, computer vision has become popular among researchers and technology users. As a result, the computer vision application becomes broader, and the objects that must be detected become more complex, increasing the object classification. Furthermore, due to current system limitations such as library storage space and processing speed, researchers and computer vision engineers innovated computer vision technology by introducing Deep Learning (DL) for object detection [6].

DL is a subset of machine learning based on an Artificial Neural Network (ANN). It is intended to be inspired by and mimic the function of the human brain. Like the human brain, the ANN is composed of multiple layers of many neurons known as neural networks, each of which can perform a simple operation and interact with one another to compute the results or make decisions.

Because of the architecture of ANN, which is made up of multiple layers of neurons, training takes more power and time. The ANN complexity contributes to the effects of overfitting. When ANN models show signs of overfitting, the detection accuracy suffers.

As a result, computer vision researchers use CNN, which is analogous to ANN, to exploit the neural network that focuses on images in the field of pattern recognition. There are two kinds of neural networks: one-stage algorithms and two-stage algorithms. The one-stage algorithm performs well in processing speed, but the two-stage algorithm performs well in object classification and positioning accuracy.

There are several CNN algorithms that researchers frequently adopt and employ in their research. Regional Convolutional Neural Network (R-CNN)-based techniques

include Mask R-CNN and Faster R-CNN, Single Shot Multibox Detector (SSD), and You Only Look Once (YOLO). Each algorithm has its own set of advantages and disadvantages.

Agriculture has become increasingly important in the global economy in the recent years, and this is because the world population is continuing to grow, putting additional strain on the agricultural system, and the cultivated land area is decreasing significantly due to urbanization. As a result, the demand for efficient and safe agricultural methods is increasing.

To accelerate an increase in agricultural productivity more accurately, innovative sensing and driving technology, including advanced information and communication technologies, must be added to current traditional agricultural management methods that promote the development of high-quality and high-yield agriculture [7]. In the recent decades, computer vision inspection systems have become important tools for farm operations. As a result, it is becoming more common for agricultural production management to adopt more expert and computer vision algorithms intelligent systems, which increase agricultural productivity and efficiency when computer vision-based agricultural automation is used.

Aside from the massive growth of artificial intelligence, the capability of computer vision technology has improved dramatically due to the rapid development of technologies in Graphics Processing Units (GPU) and Deep Belief Networks (DBNs). Many suggestions and insights for decision support and practices for farmers have been developed as a result of improved resource efficiency, ensuring agribusiness efficiency. As a result, agricultural automation is increasingly incorporating computer vision technology, ushering agriculture into the era of intelligent agriculture 4.0.

2.2 *Base Neural Networks*

The model used in this research will classify and detect crops using the SSD neural network, as shown in Fig. 1. The SSD is a feed forward convolutional network that generates fixed-size bounding boxes and scores the presence of object class instances in those boxes before performing a non-maximum suppression step to produce the final detection. The SSD is capable of detecting faster than previous state of the art for YOLO neural networks and more accurately than region proposal and pooling (Faster R-CNN) neural networks [8].

It will use a modified SSD architecture for this experiment that replaces the VGG-16 base network layers with ResNet-18 [9] and MobileNetV1 [10] layers. The trained model will be evaluated and compared after going through the same number of training epochs.

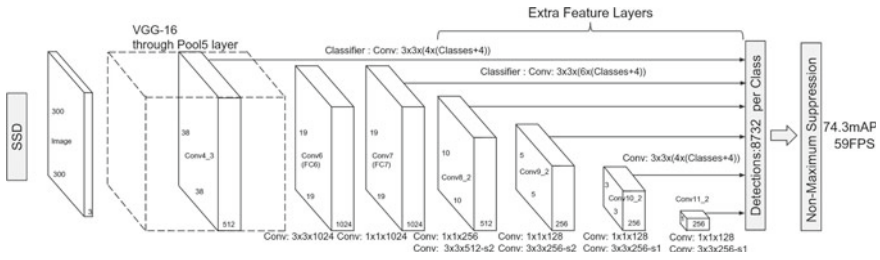


Fig. 1 VGG-16 SSD architecture [8]

3 Methodology

This experiment aims to develop a system that can be immediately applied to real-world applications. As a result, the SBCs are the best and ideal solution due to their low energy consumption and power processing compute unit that can analyze, evaluate, and produce results in a matter of seconds. Because of their small form factor, SBCs can be installed in autonomous systems such as drones and machines [11], and in this experiment, we used the NVIDIA® Jetson Nano Developer Kit for the SBCs, as shown in Fig. 2 [12].

The NVIDIA® Jetson Nano Developer Kit is equipped with a Quad-core ARM A57 running at 1.43 GHz, a 128-core Maxwell GPU, and 4 GB LPDDR4, allowing it to handle, compute, and process datasets based on the chosen neural network [13]. The NVIDIA® Jetson Nano Developer Kit can also be used to build drones and machines because it has a camera, USB, HDMI, and DisplayPort ports, a GPIO pin, and a 12 V DC barrel jack.

Next, we choose the crop objects, chili, eggplant, and potato, aside from the fact that it is difficult and limited to find in free online community datasets. The crop object chosen is unique in terms of shape, size, and color, and the system must be able to identify the crops.

Fig. 2 NVIDIA® Jetson Nano Developer Kit [12]



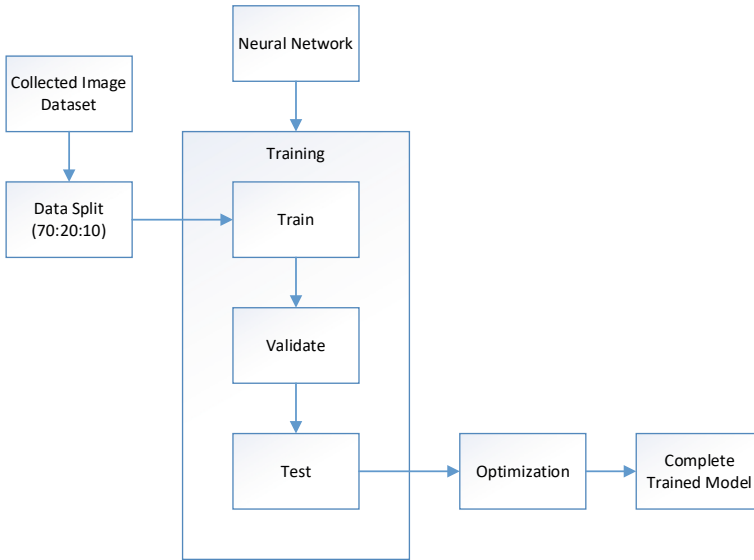


Fig. 3 Flowchart of the training methodology

According to Fig. 3 below, the dataset will be created from the selected custom crops object, and the source for the dataset is crop image capture using a standard webcam with the same number of images for each crop. 70% of the collected dataset will be used for training, 10% for testing, and 20% for validation. The default value for NVIDIA® Jetson Nano Developer Kit training epoch or iteration is 30, but we will progress from the default value to a specific value in the remaining time.

The dataset will train, test, and validate the chosen identification and neural detection networks. The entire neural network will be tested in real time and post-processing detection. The identification and detection accuracy will be monitored, and if the identification accuracy is low, the epochs training will be increased.

4 Results and Discussions

Experiments were conducted to measure actual performance for both identifications from proposed neural networks. Training with 30 and 2000 epochs was chosen for both neural networks’ significantly different performances. The trained model will then execute and identify the crops in real time. The results of classification and identification are then recorded.

After analyzing the experiment results, we can see a difference in accuracy percentage. The accuracy difference exists not only between two neural networks but also within the neural network itself. Tables 1 and 2 give the difference in accuracy

Table 1 Classification accuracy results with ResNet-18

Crops	ResNet-18 30 epochs (%)	ResNet-18 2000 epochs (%)	Difference (%)	Results
Chili	34.96	94.79	59.53	Increment
Eggplant	44.25	47.04	2.79	Increment
Potato	32.65	38.74	6.09	Increment

Table 2 Detection accuracy results with MobileNetV1

Crops	MobileNetV1 30 epochs (%)	MobileNetV1 2000 epochs (%)	Difference (%)	Results
Chili	97.1	99.8	2.7	Increment
Eggplant	89.4	99.9	10.25	Increment
Potato	99.9	100	0.1	Increment

within the neural network when the epoch cycle is changed; the higher the percentage, the higher the accuracy.

Table 1 gives the difference in classification accuracy between the selected crops. When epochs training is 30, eggplant has 44.25 percent classification accuracy, while chili has 94.79% classification accuracy. Although all crop identification shows an accuracy increase with higher epochs training, the chili shows the most significant increase with 59.53%.

Table 2 gives the difference in detection accuracy between the selected crops. When epochs training is 30, the highest detection accuracy is a potato with 99.9%, and the results remain the same when epochs training is 2000 with 100%. Although all crop identification shows an accuracy increase with higher training epochs, eggplant has the highest increase of 10.25%.

By tabulating accuracy results and directly comparing both neural networks, we can see that the MobileNetV1 neural network performs significantly better in crop identification, as shown in Table 3 and Figs. 4, and 5 shows the SSD-MobileNetV1 detection identification results in real-time detection.

Table 3 Overall identification accuracy results

Crops	Classification (ResNet-18)		Detection (MobileNetV1)	
	30 epochs (%)	2000 epochs (%)	30 epochs (%)	2000 epochs (%)
Chili	34.96	94.79	97.1	99.8
Eggplant	44.25	47.04	89.4	99.9
Potato	32.65	38.74	99.9	100

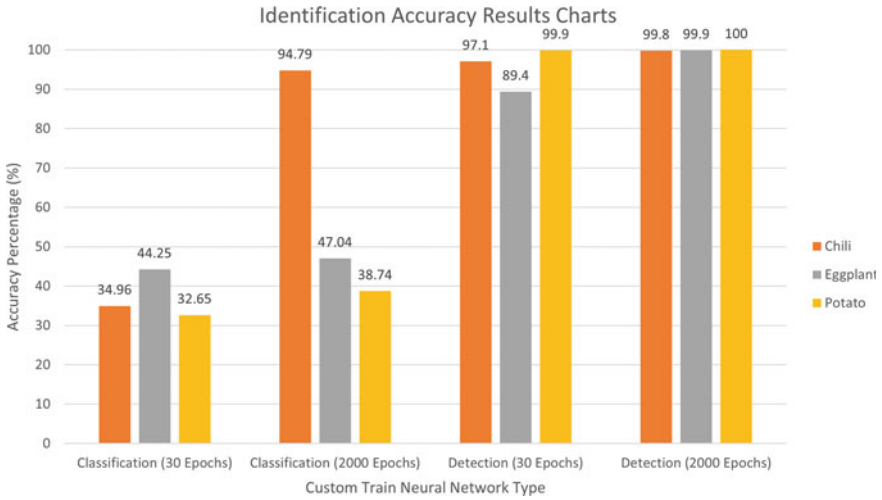


Fig. 4 Bar chart for overall identification accuracy results

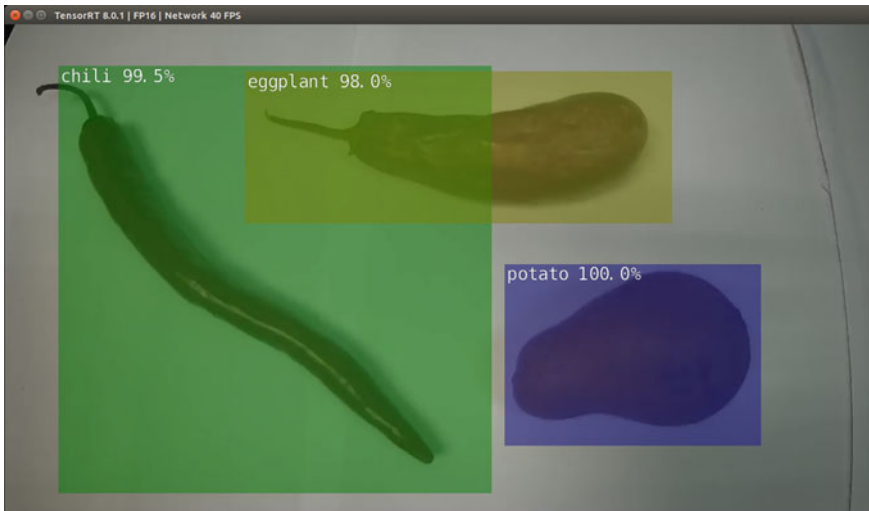


Fig. 5 Real-time detection using SSD-MobileNetV1

5 Conclusion

Based on the experiment results, we can conclude that the proposed SSD neural network, as well as the neural networks ResNet-18 and MobileNetV1, are capable of crop classification and detection. Even with the default epochs value setting for NVIDIA® Jetson Nano Developer Kit, the identification system can still identify

and differentiate the crops. This experiment also demonstrates that the NVIDIA® Jetson Nano Developer Kit can execute the trained model and compute and train a collection of image datasets. Furthermore, it has the potential to accelerate the adoption of autonomous systems in agriculture from upstream to downstream, that is, from farm to consumer.

References

1. DOSM Homepage. <https://www.dosm.gov.my/v1/index.php>
2. Putra B, Soni P, Marhaenanto B, Pujiyantim Subudi Harsono S, Fountas S, Using information from images for plantation monitoring: a review of solutions for smallholders. <https://doi.org/10.1016/j.inpa.2019.04.005>
3. Abdullah MA, Ibrahim MAR, Shapiee MNAB, Mohd Razman MA, Musa RM, Abdul Majeed APP (2020) The classification of skateboarding trick manoeuvres through the integration of IMU and machine learning. In: Jamaludin Z, Ali Mokhtar MN (eds) Intelligent manufacturing and mechatronics. SympoSIMM 2019. Lecture notes in mechanical engineering. Springer, Singapore. https://doi.org/10.1007/978-981-13-9539-0_7
4. Mohd Rudin NA, Suhaimi Puteh S, Mat Jizat JA, Mohd Khairuddin I, Abdul Majeed APP, Mohd Razman MA, Classification of Plant Health (*Capsicum Frutescens*) Normalize Differences Vegetation Index using Image Processing. <https://doi.org/10.15282/mekatronika.v3i1.7158>
5. Shapiee MNA, Abdul Manan AA, Mohd Razman MA, Mohd Khairuddin IPP, Abdul Majeed A (2022) Chili plant classification using transfer learning models through object detection. In: Enabling Industry 4.0 through advances in mechatronics. Lecture notes in electrical engineering, vol 900. Springer, Singapore. https://doi.org/10.1007/978-981-19-2095-0_46
6. O' Mahony N, Campbell S, Carvalho A, Harapanahalli S, Velasco Hernandez G, Krpalkova L, Riordan D, Walsh J, Deep learning versus traditional computer vision. <https://doi.org/10.48550/arXiv.1910.13796>
7. Tian H, Wang T, Liu Y, Qiao X, Li Y, Computer vision technology in agricultural automation—a review. <https://doi.org/10.1016/j.inpa.2019.09.006>
8. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C, Berg A, SSD: single shot multibox detector. <https://doi.org/10.48550/arXiv.1512.02325>
9. He K, Zhang X, Ren S, Sun J, Deep residual learning for image recognition. <https://doi.org/10.48550/arXiv.1512.03385>
10. Howard A, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M.: MobileNets: efficient convolutional neural networks for mobile vision applications. <https://doi.org/10.48550/arXiv.1704.04861>
11. Amini Rad P, Hoffmann D, Andres Pertuz Mendez S, Goehringer D.: Optimized deep learning object recognition for drones using embedded GPU. <https://doi.org/10.1109/ETFA45728.2021.9613590>
12. Jetson Nano Developer Kit Homepage.: <https://developer.nvidia.com/embedded/jetson-nano-developer-kit>
13. Jetson Zoo Homepage. https://elinux.org/Jetson_Zoo

Transient Pressure Analysis in Water Hydraulics Machine Using Induced Pressure Effect from the Compression of Different Materials



Ahmad Anas Yusof, Suhaimi Misha, Faizil Wasbari,
Mohamed Hafiz Bin Md Isa, Mohd Qadafie Ibrahim,
and Mohd Shahir Kasim

Abstract The purpose of this study is to investigate the effect of pressure transients on the use of a water hydraulics machine in a food processor. Water is used as a pressure medium to control the movement of double-acting cylinders within a custom-built food-processing machine. The system employs Cartesian robotics movement, with one cylinder working horizontally and the other working vertically to provide compression. The machine is converted into a simple compression machine in this experiment so that the pressure transient during the process can be measured and analyzed. As a result, this paper presents an analysis of pressure transients during continuous compression tests of malleable, viscoelastic, elastic, and brittle materials. It is noted that several pressure transients are observed during the tests.

Keywords Water hydraulics · Food processing · Pressure transients

1 Introduction

Engineers and scientists have been fascinated with transients in hydraulics pipelines, or “water hammer,” as these phenomena are more generally known, for nearly a century and a half [1–3]. They can occur in any pipe system that contains a liquid that can move. Disturbances in terms of waves of pressure, velocity, stress, and strain can propagate across the system when the steady motion of the liquid changes and are determined by the pipe network’s geometry and the physical qualities of the liquid and pipe material. These conditions can cause failures in the components, resulting in damages if precautions are not taken. Transients are then becoming an issue in modern water hydraulics technology, as simulation utilizing the Joukowski equation revealed that pressure transients in water hydraulics systems are more severe than in oil hydraulics systems, due to their higher amplitude and frequency. In the

A. A. Yusof (✉) · S. Misha · F. Wasbari · M. H. B. M. Isa · M. Q. Ibrahim · M. S. Kasim
Robotics and Industrial Automation Research Group, Faculty of Electrical Engineering, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
e-mail: anas@utem.edu.my

simulation, pressure transients occur when the valve is being closed instantly. By lowering the starting flow velocity and/or extending the valve switching time, the pressure transients will be minimized, thus reducing the risk of liquid erosion and wear [4]. Therefore, the primary goal of this research is to investigate the effect of pressure transients or fluctuations in the water hydraulics compression process, which is used in a water-powered food-processing test rig, by subjecting the rig into various types of material to induce the pressure effect.

2 Literature Review

Many users are used to exclusively utilizing mineral oil or other fluids in hydraulics devices; thus, tap water as the working medium is a completely new concept. The use of pressured water as a working medium is actually not new, as it dates back over two thousand years ago. Archimedes and other ancient Greek and Roman inventors exploited water hydraulics in their innovations. Al-Jazari used water hydraulics in his water-raising machinery, programmable robot, and famous automata and elephant clock during the Islamic Golden Age's thirteenth century [5]. It took five centuries later for water hydraulics to become a relatively accurate and cost-effective method of power transmission during the industrial revolution. In the nineteenth century, pressurized water was widely used to power machinery in Europe. Pump stations are being built near the Thames in London, providing pressurized water up to 50 bar, that power elevators and cranes through a 100 km transmission line. It was even used to operate a Tower Bridge across the river by the end of the century. Water hydraulics applications can also be found in the Paris Eiffel Tower, where it was used to power several elevators. The large pumps used to operate the elevators are still well preserved and on display at the tower. Due to the quick growth of electrical power and the creation of the first oil hydraulics system in 1906, the development of water as a pressure medium began to decline in the early twentieth century [6, 7]. Water hydraulics had resurfaced as a topic of interest by the end of the twentieth century, owing to environmental concerns and the need for sustainable development.

The development of modern water hydraulics system, such as Danfoss Nessie hydraulics, is an excellent example of combining environmental commitment with real-world benefits [8]. In the twenty-first century, various concepts of water hydraulics technology have been tested and applied. A water-based mobile hydraulics system was developed and specifically designed to operate on a Jacobsen Greens King VI mower, where water hydraulics components are used to replace the existing oil hydraulic components. Testing has revealed that the machine is fully functional and operates at a desirable speed of approximately 3 mph [9]. The concept has been applied to a novel water hydraulics variable ballast system, which is used to adjust the ballast water and is an excellent choice for ultra-deep-sea applications, such as in autonomous underwater vehicles [10, 11]. In the food-processing industry, the technology has been tested in a cookie-processing machine, burger compression machine, beef cutter, cheese manufacturing, and ice-filled machine [12-14]. The application

can also be found in the use of an environment-friendly waste packer lorry, industrial water cleaning, die castings, industrial automation, humidification devices, and firefighting systems [15]. Water hydraulics technology has even been tested inside a thermonuclear fusion reactor using various types of hydraulics manipulators, representing the most advanced application of water hydraulics to date [16-18]. Thus, the goal of this research is to look into the impact of pressure transients on the operation of a water hydraulics machine in a food processor. The pressure transients are determined by continuous compression tests performed on malleable, viscoelastic, elastic, and brittle materials to simulate various compression scenarios.

3 Methodology

Figure 1 shows the complete water hydraulics circuit of the automatic food-processing machine, which includes a pump, inverter, controller, valve, sensor, and cylinder. In this research, the machine is set to compress various sorts of materials in an attempt to produce water pressure in the system for pressure transient analysis. The machine can be operated using either relays, programmable logic controllers, and even embedded systems like the Arduino or Raspberry Pi. The spray pump employed in this study is a triplex piston pump with a maximum pressure of 40 bar, which is often used in car wash businesses. It has an integrated pressure regulator and an electric motor as its prime mover and measures 64 cm × 50 cm × 50 cm in dimensions. Custom-built water hydraulics cylinders have also been developed as the system's actuators. The cylinders have bore and stroke of 40 and 125 mm, respectively, and are based on double-acting, tie-rod cylinder design. Tap water is utilized to deliver energy and pressure from the pump to the cylinder. The focus of this work is on the analysis of pressure transients during the compressive test of malleable, viscoelastic, elastic, and brittle materials at system pressure of 6 bar. It is a typical pressure for food-processing machines that use pneumatics. [19]. A set of data loggers Hydrotechnik MultiSystem 5060 plus is used to record pressure flow data in the system over time. It has 24 channels and 2 GB of memory and can measure pressures up to 100 bar.

4 Results and Discussions

Figures 2, 3, 4 and 5 illustrate the pressure distribution during a continuous compression test at 6 bar system pressure. Figure 2 shows a compression test with flour dough, representing the viscoelastic material. After 5 s, the pressure climbs from 3 bar to an average of 5 bar, with some pressure variation. Starting at the point of contact, the dough is extruded at an average pressure of 5 bar for roughly 4.4 s. It should be noted that the dough is completely extruded in 9.4 s. The pressure climbs to the full 6 bar system pressure in 10.3 s, reaching maximum system pressure before decreasing to

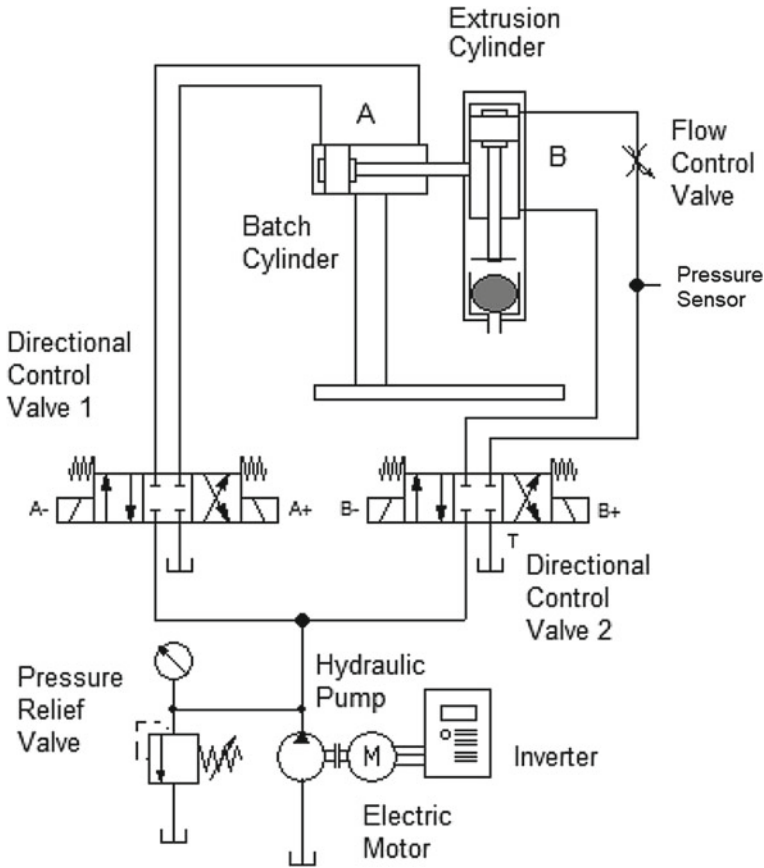


Fig. 1 Automatic traditional cookies machine

the initial 3 bar pressure during cylinder retraction. The dough is allowed to extrude through a small hole during compression, resulting in a constant average pressure of 5 bar. The viscoelastic properties of the dough are clearly displayed in the test, with the dough exhibiting both viscous and elastic properties when deformed. Viscous materials, such as water, resist shear flow and strain linearly with time when a load is applied. Elastic materials stretch and then return to their original state once the force is released.

In Fig. 3, the dough is replaced in the machine by an aluminum beverage can. The pressure begins to fluctuate after 5 s of cylinder extension, at which point the aluminum beverage can ruptures and the plastic deformation process begins, with pressure fluctuations starting at 3.1 bar, increasing to a maximum spike of 5.2 bar at time equals to 5.4 s, and continuing to fluctuate. This occurs at intervals of 5.4 s–7.4 s. The plastic deformation stops at time equals to 7.4 s and the pressure increases to a maximum pressure of 6 bar in 9.3 s before falling and settling to a pressure

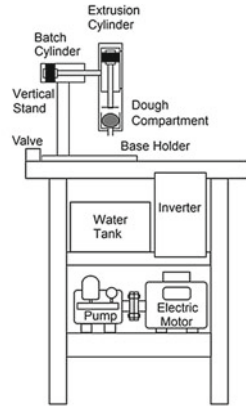
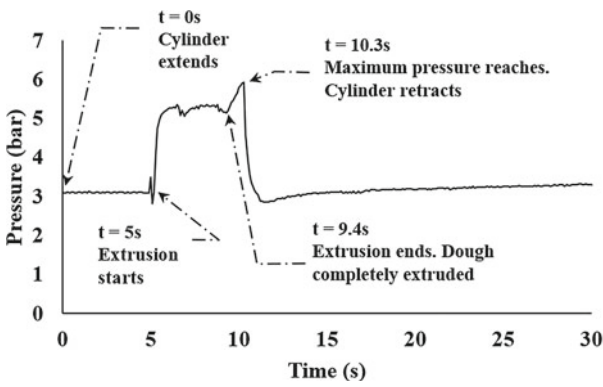


Fig. 2 Pressure distribution (viscoelastic—dough)

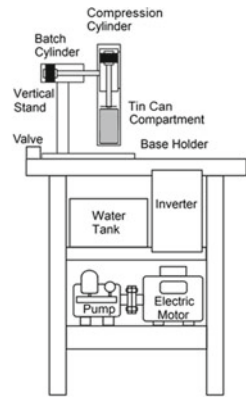
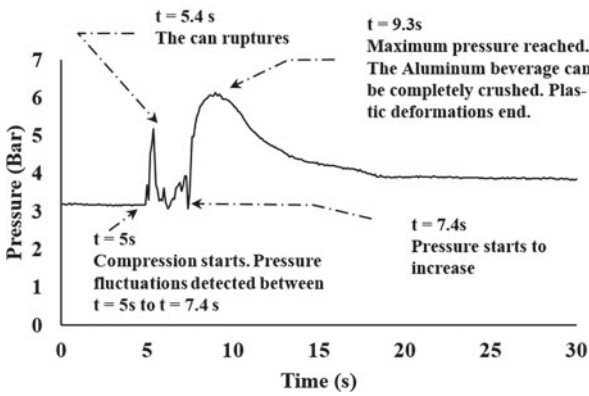


Fig. 3 Pressure distribution (malleable—aluminum can)

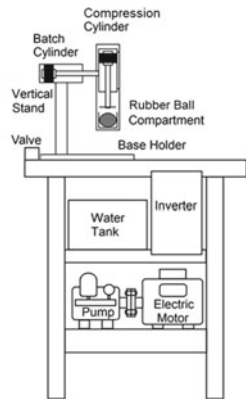
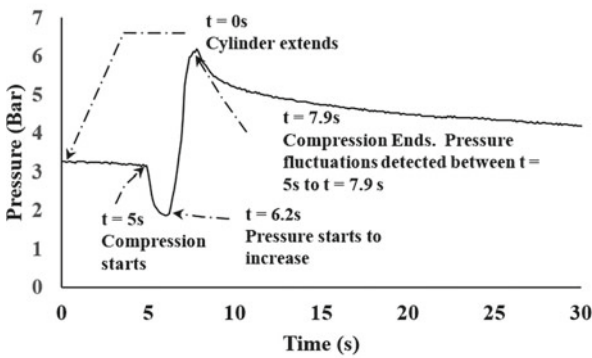


Fig. 4 Pressure distribution (elastic—rubber)

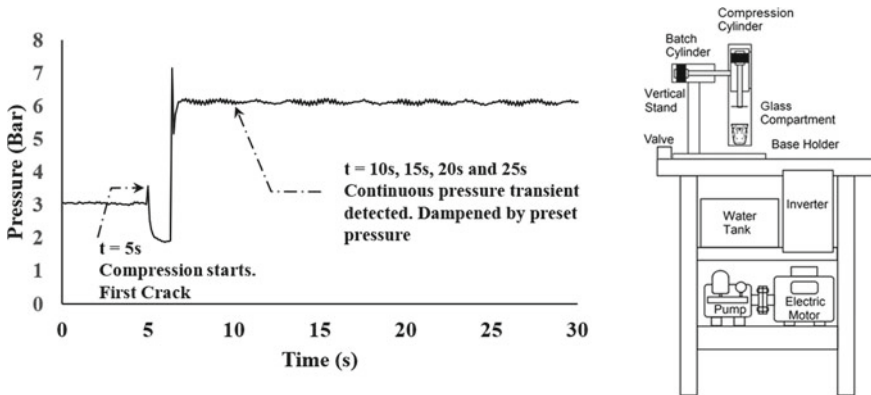


Fig. 5 Pressure distribution (brittle—glass)

average of 3.8 bar during the retraction process. It is noted that it takes about 1.3 s for the pressure to rise to the maximum after the plastic deformation of the aluminum beverage can ends.

Figure 4 shows the pressure distribution for a compression test on a rubber ball. The pressure begins to fluctuate after 5 s of cylinder extension, at which point the ball starts to fight back, which causes an increase in pressure at time equals to 6.2 s. The compression lasted for 1.7 s before it reaches the maximum preset pressure at 6 bar. No fluctuation is recorded at the peak of the compressions. The deformation comes to an end at this moment. After that, during the retraction process, it is noted that the elastic property of the rubber ball pushed back at the retracting cylinder, causing a steady decrease in retraction pressure, before settling at a pressure of 4 bar.

Figure 5 shows the pressure distribution for a compression test on a brittle material. Continuous pressure transients are detected after the increase of spike due to the broken glass during compression. The fluctuation is dampened by preset pressure of 6 bar, during time intervals ranging from 6.4 s to 30 s. It is noted that the first crack is detected at time equals to 5 s, and maximum spike up to 7 bar of pressure is detected in the measurement.

5 Conclusion

This project demonstrates the development of a water hydraulics machine in food-processing application. In this paper, pressure analysis in the water-powered food-processing machine has been presented. The relationship between stroke movement in compressing different materials and the induced pressures due to the process has been analyzed. It is noted that during the compression of the dough, aluminum can, rubber ball, and glass, the pressure transient time laps decrease with the increase of system pressure. At the same time, the pressure fluctuations are noted to be increasing

with substantial amount, during the constant compression process of the cylinder, but in smaller amount when the cylinder reaches the maximum strokes, which activates the rapid opening and closing of the pressure relief valve. In comparison to all materials, substantial amount of pressure transient is recorded in the compression of the glass, relative to other materials. The broken glass represents the sudden changes in water hydraulics loading, which influences the pressure transient at that particular time.

Acknowledgements This work is funded by the Ministry of Higher Education (MOHE) of Malaysia, under the Fundamental Research Grant Scheme (FRGS) FRGS/1/2016/TK03/FKM-CARE-F00317. The authors wish to thank the Ministry of Higher Education and Universiti Teknikal Malaysia Melaka for their support.

References

1. Korteweg DJ, de Vries G (1895) On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. *Philos Mag Ser 5*(39):240. 422–443
2. Thorle ARD (1969) Pressure transients in hydraulic pipelines. *J Basic Eng* 91(3):453–460
3. Tijsseling A, Anderson A, Moens I, Korteweg DJ (2012) On the speed of propagation of waves in elastic tubes. In: *Proceedings of 11th international conferences on pressure surges*. pp 227–245
4. Rydberg K (2002) New materials and component design—key factors for water hydraulic systems. *SAE Trans* 154–161
5. Romdhane L, Zegloul S, Al-Jazari (1136–1206) In: Ceccarelli M (ed) distinguished figures in mechanism and machine science. *History of mechanism and machine science*, vol 7. Springer, Dordrecht
6. Koskinen KT, Leino T, Riipinen H (2008) Sustainable development with water hydraulics—possibilities and challenges. In: *Proceeding of the 7th JFPS international symposium on fluid power*. Toyama Japan, pp 11–18
7. Trostmann E, Bo F, Bo HO, Bjarne H (2001) *Tap water as a hydraulic pressure medium*. Marcel Dekker, New York
8. Higgins M (1996) Water hydraulics—the real world. *Ind Robot: Int J* 23(4):13–18
9. Cassens L, Thomas M, Krutz G (2002) Modified water powered greens king VI mower. *SAE Technical Paper* 2002–01–1382
10. Haugen GK, Conrad F, Grahl-Madsen M (2005) Innovative new ROV technology utilizing water hydraulics. In: *Proceedings of the 6th JFPS international symposium on fluid power*. Tsukuba. Japan
11. Yinshui L, Xufeng Z, Defa W, Donglin L, Xiaohui L (2015) Study on the control methods of a water hydraulic variable ballast system for submersible vehicles. *Ocean Eng* 108:648–661
12. Pham PN, Ito K, Ikee S (2008) Energy saving for water hydraulic pushing cylinder in meat slicer. In: *Proceeding of the 7th jfps international symposium on fluid power*. Toyama Japan, pp 95–100
13. Yusof AA, Misha S, Isa MHM, Wasbari F, Ibrahim MQ, Kasim MS (2017) Low cost water hydraulics technology for Malaysian traditional cookies production. In: *Proceedings of the 10th JFPS international symposium on fluid power*. Fukuoka, Japan
14. Yusof AA, Misha S, Isa MHM, Wasbari F, Ibrahim MQ, Kasim MS (2018) Simulation and experimentation of water hydraulics technology for automatic traditional cookies production. *J Adv Res Fluid Mech Therm Sci* 47(1):136–150

15. Finn C (2005) Trends in design of water hydraulics—motion control and open-ended solutions. In: Proceedings of the 6th JFPS international symposium on fluid power. Tsukuba, Japan, pp 420–430
16. Siuko M et al (2003) Water hydraulic actuators for ITER maintenance devices. *Fusion Eng Des* 69:141–145
17. Gregory D, David O, Nozais F, Yvan M, Friconneau J-P, Palmer J (2008) Assessment of a water hydraulic joint for remote handling operations in the divertor region. *Fusion Eng Des* 83:1845–1849
18. Lyytikäinen V, Kinnunen P, Koivumäki J, Mattila J, Siuko M, Esque S, Palmer J (2013) Divertor cassette locking system remote handling trials with WHMAN at DTP2. *Fusion Eng Des* 88(9–10):2181–2185
19. Pneumatic Equipment for Food Industry. SMC Catalogue Shoku-P-A-C1-CS3e.indd

X-Ray Baggage Object Detection Using Neural Networks Approach for Safety Purpose



Samuel Ato Gyasi Otahir, Sew Sun Tiang, Wei Hong Lim, Hung Yang Leong, and Bo Sun

Abstract Airport security a matter of urgent attention and this calls for measures to ensure that all baggage that move within the airport contain non-harmful objects that may people at the airport in danger. Over the last decade, X-ray machines have evolved to scan baggage, while an airport officer verifies that the content of the baggage bag contains benign items; if anomaly items are identified, the owner of the baggage is called aside to have further enquiries. However, this process is time consuming and moreover lacks accuracy at times due to fatigue on the side of the officer checking the scanned X-ray images. This paper is prepared to develop a convolutional neural network (CNN) to aid in the process of X-ray baggage object detection. The inference speed of the proposed model is discussed in the paper and compared with other convolutional neural networks. The efficiency of the proposed model is evaluated by means of quantitative metrics. The proposed model leverages the YOLOv5 algorithm, which achieved an accuracy of 90%, precision of 90.4%, recall of 84.6%, and an F1-score of 87.40%. The designed and proposed model is capable of real-time anomaly object detection with a very fast inference speed in baggage to help increase the security at airports.

Keywords Convolutional neural network · Object detection · YOLOv5

1 Introduction

The transportation industry undeniably plays a vital role in every country's economy; it connects various cities around the world. Statistics show that the Sydney Airport is expected to reach 74 million by 2033 and 6 billion worldwide [1]. In order to instill security at these transportation facilities, there exists a security screening process which entails X-ray machines scanning the content of passenger's baggage so as to detect anomaly objects by the screening officers [2]. This screening process tends

S. A. G. Otahir · S. S. Tiang (✉) · W. H. Lim · H. Y. Leong · B. Sun
Faculty of Engineering, Technology and Built Environment, UCSI University, Kuala Lumpur, Malaysia
e-mail: tiangss@ucsiuniversity.edu.my

to take a longer time around half a minute, also there is a usual occurrence for lack of accuracy due to fatigue from the screening officers, and the X-ray machine itself not giving quality images due to overlap of passenger baggage and the machine been worn out [3]. This makes it very hard to identify anomaly objects in the baggage, which in return usually results in a breach of security [4]. Deep learning is a subfield of artificial intelligence that enables the computers to extract meaningful information from various input sources and perform the specific tasks such as classification [5–7], and fault detection [8, 9]. Convolutional neural networks (CNNs) a form of neural network have gained much recognition in the image processing and preprocessing domain [10]. In the early nineties, CNN was employed to solve tasks such as character recognition tasks; the recent surge in application of CNN is because CNN was employed in the ImageNet image classification challenge, and it proved second to none [11]. YOLOv5, a CNN algorithm, will be implemented to build a model that will seek to reduce the downsides of the security screening process at the transportation industry mainly at the airports. This paper mainly focuses on developing a model with the YOLOv5 algorithm which seeks to have a very high inference speed of about 0.008 s, high precision, and high accuracy as opposed to the conventional way which takes about 15 s to detect objects in an X-ray baggage by the officer.

2 Related Work

A summary of other similar works in using neural networks for X-ray baggage object detection for safety purpose is given in Table 1.

Daniel et al. [12] proposed to use YOLOv3 to detect guns, knives, razor blades, and shuriken. The best mean average precision record for each class was 96.3% for guns, 76.2% for knives, 86.9% for razor blades, and 93.7% for shuriken while having the mAP for all anomaly objects 80.0%. Similarly, Reagan et al. [13] employed the use of YOLOv3 to detect anomaly objects, and mAP of 52.40% was attained. The object of interest to be detected includes batteries, mortars, and wires from a dataset of close to a million X-ray images. However, in Yona et al. [14] work, they proposed a model known as Mask R-CNN. The process a detection stage followed by a classification stage; the object detection stage entailed identification of the respective class thereby followed by the classification into benign or anomaly object. The proposed model in [14] worked around a six-class object detection which is comprised of the following: bottle, hairdryer, iron, toaster, mobile, and a laptop. The

Table 1 Summary of previous works

Paper	Methods	Dataset size	Accuracy	Precision	Recall	F1-score
[12]	YOLOv3	24,520	80.0%	–	–	–
[13]	YOLOv3	1,281,167	52.40%	–	–	–
[14]	Mask R-CNN	3534	59.9%	57.0%	58.0%	59.25%

performance metrics recorded from the Mask R-CNN model totaled an accuracy of 59.9%, precision of 57.0%, recall of 58.0%, and a F1-score of 59.25%.

3 Methodology

This section entails the description, in which a model is proposed and discussed. The model was built on Google Colab platform. Google Colab provides a free GPU and TPU to aid in user with low computational specs to build and implement models. However, the runtime was set to use both that Google Colab and a local computer specification: 11th Gen Intel(R) Core (TM) i5-1135G7 @ 2.40 GHz (8 CPUs), ~2.4 GHz, 16 Gb RAM, Python 3.10.

3.1 Model Architecture

The proposed model employs the YOLOv5 algorithm. The algorithm is designed to have a backbone of Cross Stage Partial Network (CSPNet) to aid in achieving a faster inference speed, a neck of Spatial Pyramid Pooling (SPP), and Path Aggregation Network (PANet) to enhance model accuracy by maintaining spatial information of the input data and finally a head which uses Generalized Intersection over Union (GIoU) to calculate for the loss of the bounding box regression.

3.2 Data Acquisition

Neural networks are usually built and trained by feeding it with real or synthetic data as seen in [6]. In this paper, the dataset used here was acquired from a pre-collected airport data from [5] known as PIDray. PIDray is a public dataset which has 47,677 X-ray images. There is a total of 12 categories of hidden threat items in the dataset, the 12 categories include baton, bullet, gun, hammer, handcuffs, knife, lighter, pliers, power bank, scissors, sprayer, and a wrench. Each image in the PIDray dataset is provided with an image and instance-level annotation.

3.3 Dataset Annotation and Preprocessing

To create annotations for each of the images in the dataset, RoboFlow was employed because RoboFlow provides a safe place to label, store, and export the labeled dataset for model training. The ratio of division between training, validation, and testing dataset is in a 7:1:2 ratio as per standard dataset splitting. All images are preprocessed

Table 2 Summary of parameters used

Parameters for classification model	
Batch size	16
Learning rate	0.01
Epoch	200
Loss function	Multi-class cross-entropy loss
Optimizer	Adam

Table 3 PIDray description

Division	
Train set size	29,457
Test set size	18,220

3.6 Performance Evaluation

Robustness of the model is measured by the precision, recall, F1-score, the intersection over union (IoU), and the mean average precision (mAP). The model’s hyperparameters are adjusted based on the values obtained from the metrics.

Precision is a metric which entails the identification of the frequency at which the model correctly predicts a positive class, i.e., how often is the model predicting correctly. The formula for precision is represented in Eq. (1).

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \tag{1}$$

Recall is a metric that identifies the percentage of how many predictions did the trained model miss. Recall can be calculated by the formula seen in Eq. (2).

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{2}$$

F1-score is the harmonic mean of the recall metric and the precision metric. The higher the F1-score the better the model. The formula for F1-score is represented in Eq. (3).

$$F1 = 2 \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right) \tag{3}$$

Intersection over union (IoU) is a term that is used to describe the degree of overlap of two boundary boxes, i.e., the predicted box with respect to the ground-truth bounding box. The expected range of an IoU is usually between 0 and 1. The formula for IoU can be seen in Eq. (4).

$$IoU = \frac{\text{Area of Intersection of two boxes}}{\text{Area of Union of two boxes}} \quad (4)$$

The mean average precision (mAP) arises from the average precision (AP) which entails under the precision-recall curve. The mAP is the average of all the AP for all the classes. AP formula is seen in Eq. (5) and the mAP formula in Eq. (6).

$$AP = \frac{1}{11} \sum_{\text{Recall}_i} \text{Precision}(\text{Recall}_i) \quad (5)$$

$$mAP = \frac{1}{N} \times \sum_{i=1}^N AP_i \quad (6)$$

3.7 Deployment of Classifier Model

A few instances including using the curl command, development of a web app, usage of a webcam, deploying to NVIDIA Jetson, or deploying to Luxonis OAK come into mind when deployment of a classifier model is raised. However, after multiple trial and error of various instances, the desired deployment option would be a web app. Figure 2 shows the flowchart for the functionality of the web app designed for the classifier model.

4 Results and Discussion

4.1 Quantitative Analysis of Three Neural Networks

Results obtained from training the three models are given in Table 4.

Table 4 gives the training results obtained from YOLOv5, YOLOv3, and MobileNetSSDv2 after training each model for a period spanning between 8 and 12 h. MobileNetSSDv2 comes in with the lowest metric scores of 28.09, 44.0, 38.0, 40.78% for mAP@0.5, precision, recall, and F1-score, respectively. YOLOv3 comes in with slightly better metrics of 77.0, 76.8, 74.3, 68.5%, for mAP@0.5, precision, recall, and F1-score, respectively. It is apparent that YOLOv5 has the highest mAP@0.5 of 90.0%. Figure 3 shows a graph obtained from training the YOLOv5 model. It is deduced that the box, objectness, and classification loss are going low as the number of epochs rises, while the precision, recall, and mAP are rising steadily.

Fig. 2 Flowchart of X-ray baggage object detection model webapp

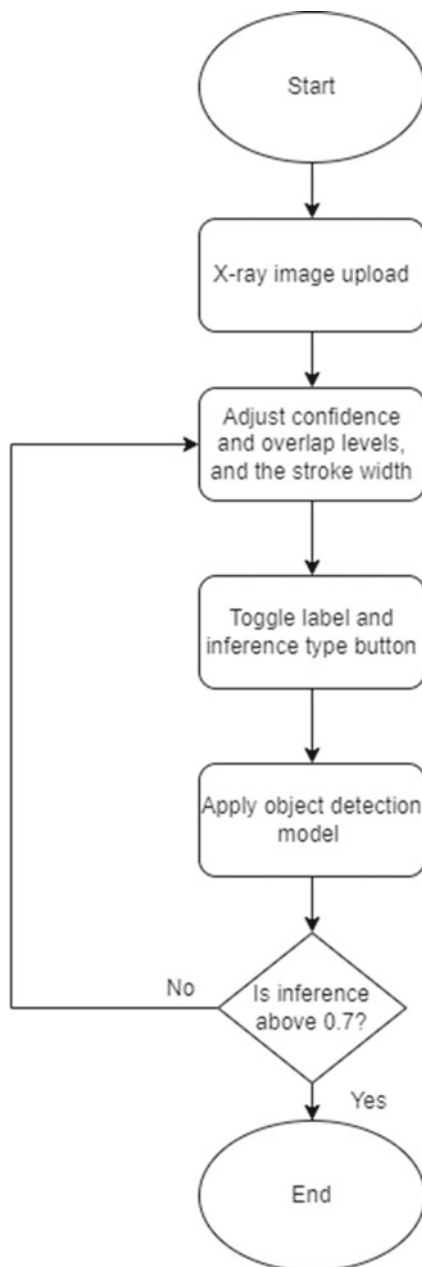


Table 4 Comparison analysis between the three models

	YOLOv5 (%)	YOLOv3 (%)	MobileNetSSDv2 (%)
mAP@0.5	90.0	77.0	28.09
Precision	90.4	76.8	44.0
Recall	84.6	74.3	38.0
F1-score	87.4	68.5	40.78

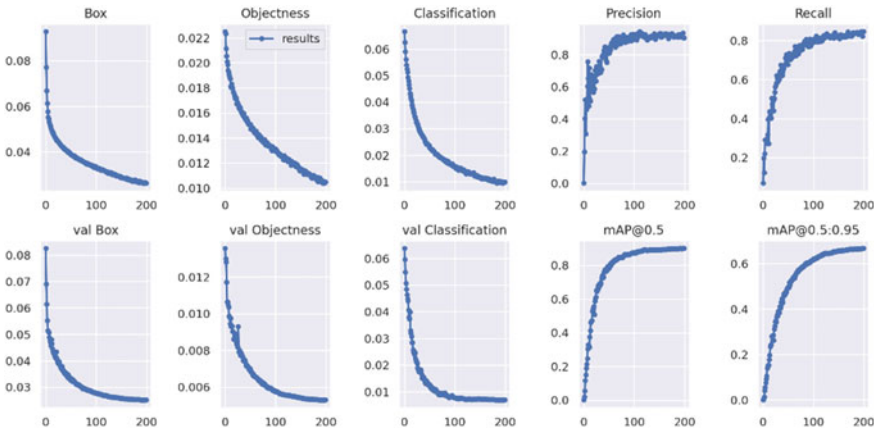


Fig. 3 Training graph of YOLOv5 model

4.2 Performance of X-Ray Object Detection Model

Figure 4a–f shows sample X-ray baggage images from the test and validation dataset after inference is run to predict objects if there may. The inference images are from the dataset split from the ratio discussed earlier. Figure 4a shows that the proposed model can predict a wrench with a very confidence level of about 0.95. The proposed model can predict a hammer, a knife, a baton, and a bullet with confidence levels of 0.86, 0.85, 0.90, and 0.94, respectively, as seen in Fig. 4b, c, e, f. However, the proposed model has a very recall for prediction of power banks, as seen in Fig. 4d; the proposed model predicts a bottle to be a power bank. Moreover, the inference speed of YOLOv5 is amazingly fast, as it takes an average of 0.008 s to detect objects in the X-ray baggage image.

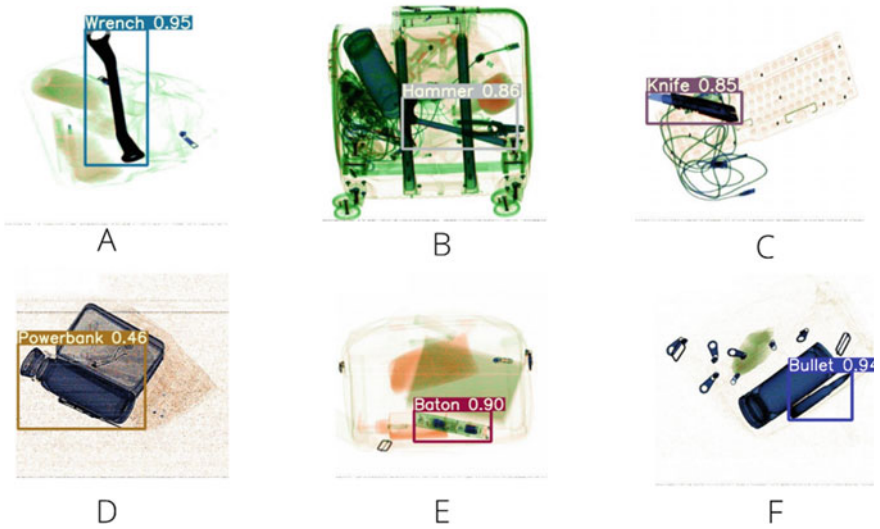


Fig. 4 Samples of predicted objects **a** Wrench, **b** Hammer, **c** Knife, **d** Power bank, **e** Baton, and **f** Bullet

5 Conclusion

The build of an X-ray baggage object detection using neural networks is discussed and a look into the performance analysis is discussed in this paper. The proposed model can detect anomaly objects including baton, bullet, gun, hammer, handcuffs, knife, lighter, pliers, power bank, scissors, sprayer, and a wrench. Metrics of 90.0% accuracy, 90.4% precision, 84.6% recall and a F1-score of 87.40% were attained after training it with the best fit hyperparameters discussed above. For the test and validation dataset, the model can detect the objects with an average of 0.008 s, which is a way faster and convenient than the conventional way of checking. In a nutshell, to build a model for X-ray object detection, it is prerequisite to ensure that the hyperparameters are optimized and best fit. However, for future works, it is advisable top venture into real-time object detection to see how well the model intercepts unknown data.

Acknowledgements This work was supported by the Ministry of Higher Education Malaysia under the Fundamental Research Schemes with project codes of Proj-FRGS/1/2019/TK04/UCSI/02/1 and the UCSI University Research Excellence & Innovation Grant (REIG) with project code of REIG-FETBE-2022/038.

References

1. Naji M, Anaissi A, Braytee A, Goyal M (2021) Anomaly detection in x-ray security imaging: a tensor-based learning approach. In: Proceedings of international joint conference on neural networks. <https://doi.org/10.1109/IJCNN52387.2021.9534034>
2. Wetter OE, Lipphardt M, Hofer F (2010) External and internal influences on the security control process at airports. In: Proceedings of international carnahan conference on security technology, pp 301–309. <https://doi.org/10.1109/CCST.2010.5678708>.
3. Wei Y, Zhu Z, Yu H, Zhang W (2021) An automated detection model of threat objects for X-ray baggage inspection based on depthwise separable convolution. *J Real-Time Image Process* 18(3):923–935. <https://doi.org/10.1007/s11554-020-01051-1>
4. Li X, Lai T, Wang S, Chen Q, Yang C, Chen R (2019) Weighted feature pyramid networks for object detection. In: Proceedings of 2019 IEEE International Conference Parallel Distributed Processing with application big data cloud computing, sustainable computing and communications, social computing and networking, *ISPA/BDCLOUD/SustainCom/SocialCom 2019*, pp 1500–1504. <https://doi.org/10.1109/ISPA-BDCLOUD-SUSTAINCOM-SOCIALCOM48970.2019.00217>
5. Voon YN, Ang KM, Chong YH, Lim WH, Tiang SS (2022) Computer-vision-based integrated circuit recognition using deep learning. In: Md. Zain Z et al (ed) Proceedings of the 6th international conference on electrical, control and computer engineering, vol 842. Springer Singapore, pp 913–925
6. Jdid B, Lim WH, Dayoub I, Hassan K, Mohamed Juhari MRB (2021) Robust automatic modulation recognition through joint contribution of hand-crafted and contextual features. *IEEE Access* 9:104530–104546
7. Low JW, Tiang SS, Lim WH, Chong YH, Voon YN (2022) Tomato leaf health monitoring system with SSD and mobileNet. In: Zain MZ et al (ed) Proceedings of the 6th international conference on electrical, control and computer engineering, LNEE, vol 842. Springer, Singapore, pp 795–804
8. Alrifay M, Lim WH, Ang CK (2021) A novel deep learning framework based RNN-SAE for fault detection of electrical gas generator. *IEEE Access* 9:21433–21442
9. Alrifay M et al (2022) Hybrid deep learning model for fault detection and classification of grid-connected photovoltaic system. *IEEE Access* 10:13852–13869
10. Morris T, Chien T, Goodman E (2019) Convolutional neural networks for automatic threat detection in security X-ray images. In: Proceedings of 17th IEEE international conference on machine learning and applications ICMLA 2018, pp 285–292. <https://doi.org/10.1109/ICMLA.2018.00049>
11. Alex K, Ilya S, HE (2007) Handbook of approximation algorithms and metaheuristics. Chapman and Hall/CRC
12. Saavedra D, Banerjee S, Mery D (2021) Detection of threat objects in baggage inspection with X-ray images using deep learning. *Neural Comput Appl* 33(13):7803–7819. <https://doi.org/10.1007/s00521-020-05521-2>
13. Galvez RL, Dadios EP, Bandala AA, Vicerra RRP (2019) YOLO-based Threat Object Detection in X-ray Images. In: 2019 IEEE 11th international conference on humanoid, nanotechnology, information technology, communication and control, environment, and management HNICEM 2019, pp 2–6. <https://doi.org/10.1109/HNICEM48295.2019.9073599>
14. Gaus YFA, Bhowmik N, Akcay S, Guillen-Garcia PM, Barker JW, Breckon TP (2019) Evaluation of a dual convolutional neural network architecture for object-wise anomaly detection in cluttered X-ray security imagery. In: Proceedings of international joint conference on neural networks. <https://doi.org/10.1109/IJCNN.2019.8851829>