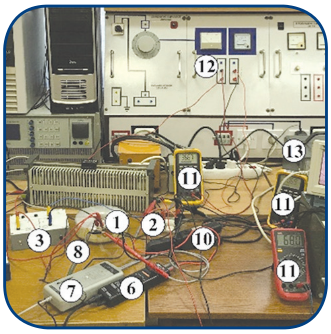
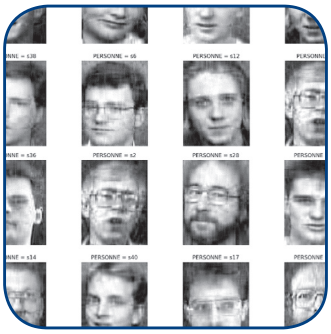
Volume 14, Number 9, 2023 ISSN 1847-6996

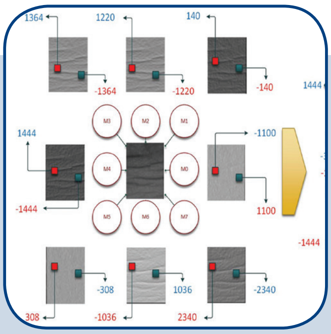
FERIT
FACULTY OF ELECTRICAL ENGINEERING, COMPUTER SCIENCE AND INFORMATION TECHNOLOGY **OSIJEK**

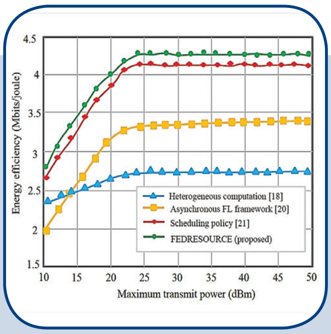
IJECES International Journal of Electrical and Computer Engineering Systems

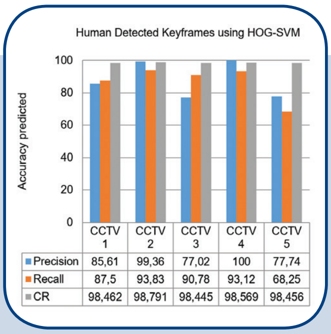
International Journal of Electrical and Computer Engineering Systems











	CCTV 1	CCTV 2	CCTV 3	CCTV 4	CCTV 5
Precision	85,61	99,36	77,02	100	77,74
Recall	87,5	93,83	90,78	93,12	68,25
CR	98,462	98,791	98,445	98,569	98,456

ISSN: 1847-6996

INTERNATIONAL JOURNAL OF ELECTRICAL AND COMPUTER ENGINEERING SYSTEMS

Published by Faculty of Electrical Engineering, Computer Science and Information Technology Osijek,
Josip Juraj Strossmayer University of Osijek, Croatia

Osijek, Croatia | Volume 14, Number 9, 2023 | Pages 959 - 1085

The International Journal of Electrical and Computer Engineering Systems is published with the financial support of the Ministry of Science and Education of the Republic of Croatia

CONTACT

International Journal of Electrical
and Computer Engineering Systems
(IJECES)

Faculty of Electrical Engineering, Computer
Science and Information Technology Osijek,
Josip Juraj Strossmayer University of Osijek, Croatia
Kneza Trpimira 2b, 31000 Osijek, Croatia
Phone: +38531224600, Fax: +38531224605
e-mail: ijeces@ferit.hr

Subscription Information

The annual subscription rate is 50€ for individuals, 25€ for students and 150€ for libraries.
Giro account: 2390001 - 1100016777,
Croatian Postal Bank

Journal is referred in:

- Scopus
- Web of Science Core Collection (Emerging Sources Citation Index - ESCI)
- Google Scholar
- CiteFactor
- Genamics
- Hrčak
- Ulrichweb
- Reaxys
- Embase
- Engineering Village

Bibliographic Information

Commenced in 2010.
ISSN: 1847-6996
e-ISSN: 1847-7003
Published: quarterly
Circulation: 300

IJECES online

https://ijeces.ferit.hr

TABLE OF CONTENTS

Research Trend Topic Area on Mobile Anchor Localization: A Systematic Mapping Study

Review Paper

Gita Indah Hapsari
Telkom university, Faculty of Informatics,
Department of Cyber-Physical System
Telekomunikasi Street, Bandung, Indonesia
gitaindahhapsari@telkomuniversity.ac.id

Bayu Erfianto
Telkom university, Faculty of Informatics,
Department of Cyber Physical System
Telekomunikasi Street, Bandung, Indonesia
erfianto@telkomuniversity.ac.id

Rendy Munadi
Telkom university, Faculty of Informatics,
Department of Cyber-Physical System
Telekomunikasi Street, Bandung, Indonesia
rendymunadi@telkomuniversity.ac.id

Indrarini Dyah Irawati
Telkom university, Faculty of Informatics,
Department of Cyber Physical System
Telekomunikasi Street, Bandung, Indonesia
indrarini@telkomuniversity.ac.id

Abstract *– Localization in a dynamic environment is one of the challenges in WSN localization involving dynamic sensor nodes or anchor nodes. Mobile anchors can be an efficient solution for the number of anchors in a 3-dimensional environment requiring more local anchors. The reliability of a localization system using mobile anchors is determined by various parameters such as energy efficiency, coverage, computational complexity, and cost. Various methods have been proposed by researchers to build a reliable mobile anchor localization system. This certainly shows the many research opportunities that can be carried out in mobile anchor localization. The many opportunities in this topic will be very confusing for researchers who want to research in this field in choosing a topic area early. However, until now there is still no paper that discusses systematic mapping studies that can provide information on topic areas and trends in the field of mobile anchor localization. A systematic Mapping Study (SMS) was conducted to determine the topic area and its trends, influential authors, and produce modeling topics and trends from the resulting modeling topics. This SMS can be a solution for researchers who are interested in research in the field of mobile anchor localization in determining the research topics they are interested in for further research. This paper gives information on the mobile anchor research area, the author who has influenced mobile anchor localization research, and the topic modeling and trend that potentially promissing research in the future. The SMS includes a chronology of publications from 2017-2022, bibliometric co-occurrence, co-author analysis, topic modeling, and trends. The results show that the development of mobile anchor localization publications is still developing until 2022. There are 10 topic models with 6 of them included in the promising topic. The results of this SMS can be used as preliminary research from the literacy stage, namely Systematic Literature Review (SLR).*

Keywords*: Mobile anchor, Localization, Systematic Mapping Study, Trend Topic Area*

1. INTRODUCTION

Wireless sensor network (WSN) localization is still being developed by researchers recently. WSN is one of the low-cost solutions to build smart environments such as smart manufacturing, smart cities, transportation, comprehensive, and real-time health monitoring [1]. In addition, localization in WSN is an important aspect in determining network coverage, routes in location-based routing protocols, sending messages to neighboring nodes, and playing a role in increasing energy efficiency [2].

Some WSN implementations must include location/ position information (localization) so that the measured data will be meaningless if it is not accompanied by accurate location/position data. Examples of WSN localization implementations are environmental monitoring systems, monitoring of animal habitats [3], forest fire surveillance [4], monitoring of natural disasters [5], tracking and navigation of robots [6], and underwater wireless sensor network/ UWSN [7].

The localization technique in WSN estimates the location of the unknown node/sensor node on the network using position knowledge from several sensors on the network which is called the Anchor Node. An anchor Node is a node that knows its position by installing a position sensor such as a GPS sensor or installing position knowledge on the node. Recently localization research on WSN refers to improving accuracy, minimizing computation, reducing the number of anchors, reducing costs, overcoming obstacles from unusual landscapes, and network security [8]. In the case of a large-scale WSN involving many sensor nodes, a very large number of anchors is required as well, this is certainly against the cost and energy performance. The localization technique from 2-dimensional to 3-dimensional coordinates is also a challenge for researchers because it requires more anchors to determine the position of the sensor node.

Mobile anchors are one solution to reduce the number of anchors [9], but this is also a challenge for researchers to produce accurate localization involving moving anchors. The dynamic environment involves moving objects (mobile sensor node/mobile anchor node) and it becomes a challenge for researchers to find the right method solution to produce accurate position calculations. One of the challenges in the dynamic environment localization technique is the calculation of distance and positioning which is influenced by the relative position of the anchor (doppler shift).

Another challenge is the trajectory of the mobile anchor to the sensor nodes and the coverage area adjusted to the deployment of sensor nodes in each case. Until now, the challenge in mobile anchor localization is to propose a localization with high accuracy, high energy efficiency, computation, communication with low cost, and the minimum number of anchors [10]. The challenge in localization using moving anchors is to produce a path planning algorithm for mobile anchors' movement to optimize minimum error localization, energy efficiency, and coverage.

This shows that the field of mobile anchor localization in WSN still has a lot of opportunities to be researched. It is very important to have a helicopter view of the latest research areas and topic areas in that field through a literature review of many papers. Researchers often have difficulty in collecting papers, conducting literacy, mapping them into research topic areas, and classifying them based on trends. In addition, researchers also sometimes have difficulty determining topics that are promising and have great research potential.

Systematic Literature Study (SMS) is a literature review methodology quantitatively and systematically to identify the appropriate research topic areas to be researched based on the classification and calculation of the contribution of the classified categories [11]. SMS is used by many researchers in several research areas to map the research area of the research field that they will research. SMS is a solution for researchers to identify the right research area and find research potential and its trends. The goal of this study is to classify, identify, and evaluate the domain of mobile anchor localization research and extract information about the topic area, the latest method, author contribution, topic modeling, and its trend.

The main contribution of this paper is given:

1. The information of the anchor localization research area in a global view of WSN through bibliometrics, so we can know the position of mobile anchor research in the field of WSN. This paper also presents the helicopter view of the current state of mobile anchor localization research.
2. The information on the authors who contributed and influenced the research topic of mobile anchor localization can direct the new researcher to follow or discuss this topic.
3. The topic modeling and trend of mobile anchor localization indicate the promising topic research that has the potential to be promising research to be further developed.

SMS methodology is applied which aims to provide an objective and systematic approach that answers a series of research questions about the state of the art of the topic. The PRISMA protocol is used to search papers, study selections, and data extraction. We involved 511 papers to map and analyze. VOS Viewer is used as a tool to get the domain of the topic area and the latest research on mobile anchor localization. Meanwhile, Orange Data Mining is a tool for processing LDA data mining that produces research domains and trends in mobile anchor localization.

This paper presented the introduction to section one. Section 2 is focused on materials and methods which are explained starting from designing the methodology, conducting, and documenting the study as well as presenting the demographics of the paper. In section 3, the results of SMS are presented by presenting bibliometrics, research domains, and trends of each domain. This section also explains which domains have the potential to be further developed.

2. RELATED WORK

Before doing SMS, it would be better to do research first on paper reviews and surveys that have been done by other researchers regarding mobile anchor localization. This is done to provide appropriate contributions and insights from systematic mapping studies so that they can be a guide for other researchers, especially in the field of mobile anchor localization. When researchers explore a topic, they must carry out literacy papers with a large enough number of papers so that there is rarely a literacy process that is less focused on the problem to be solved. The literacy system that is carried out is limited in time and cost, so it is necessary to use the right method to explore knowledge in a field so that the process becomes more efficient.

The SLR method is applied systematically and explicitly to collect, select, and analyze research literature [12, 13]. Qualitative and quantitative observations were made on the subjects to be studied to answer research questions. Meanwhile, SMS is a method that tends to be more quantitative in organizing research areas. SMS is carried out using the same protocol as SLR to search and find the required literature. However, the SMS method focuses on identifying and classifying sub-fields in the research area based on author, keyword, publication type, publication date, and publication source [14]. SMS can be used as preliminary research as an initial stage to map the topic area so that it can produce a more valuable method design [15, 16].

Table 1. The methodology used in the mobile anchor localization review paper

Methodology	2015	2016	2017	2018	2019	2020	2021	2022
Review/survey	[17] [18]	[19]	-	-	[20] [21]	-	[22]	[23]
Systematic Mapping Study (SMS)	-	-	-	-	-	-	-	-
Systematic Literature Review (SLR)	-	-	-	-	-	-	-	-

A comparison of paper reviews is carried out to get a maximum understanding of the published paper reviews and the SMS paper that will be produced more precisely and can be used as a guide to proceed to the SLR paper review. The paper review search was carried out from 2012 to 2022. The search was carried out by searching for papers using the query 'Mobile Anchor Localization review' and 'Mobile Anchor Localization Survey from the Scopus, IEEE, and Science Direct databases. The search results obtained as many papers as possible with the composition of the paper shown in Table 1.

The data in Table 1 shows paper reviews and survey papers on mobile anchor localization from 2015-2022. A survey paper by Liang Yue et al [17] in 2015 directly explained the path-planning method on mobile anchors that had been carried out by researchers. Guang Jie Han [18] describes localization theory in general including path planning in it. However, these two papers do not explain the process of collecting and classifying paper sources. Direct presentation is carried out based on the analysis of papers related to the topic. Both papers also do not explain the opportunities for research trends, topic models, and related researchers who have contributed to research topics in this field.

Likewise, the papers proposed by Subir Halder [19] Hala Abukhalaf [20] Yuxuan Long [21] and Ketan Sabale[22, 23] explained mobile anchor localization and path planning with a very brief review. In these five papers, there is no information about how the helicopter view is related to this topic with other fields. These papers also did not explain the relation among existing methods which can give insight and have opportunities for other solutions more broadly. In addition, objective information is not given on whether this topic is promising research or a good trend.

Of all the existing paper reviews, there is still no paper that explains the research area, modeling topics, and researchers and investigates the promising topic of mobile anchor localization research. No one has yet explored using the SMS method in the process of collecting, mapping, and classifying their papers to dig deeper into information about trends of topics, influential researchers, and research areas.

Weaknesses in existing paper reviews can make new researchers unable to see whether the topic still has an increasing trend and whether there are still many researchers who are interested in research in that field. Researchers also cannot know the latest methods that are being or are being proposed by many researchers related to mobile anchor localization, so there is a risk that it is difficult to get updated methods.

Mapping papers using SMS is a new thing in mobile anchor localization. With topic mapping in this paper, researchers can focus more on getting the intended paper, knowing the trend of the topic, and having knowledge about contributing and influential researchers to make it easier to find papers and do correspondence. Topic mapping using the SLR can also be used as a starting point for conducting SLR to conduct paper literacy more precisely on target to get the right research gap in SLR.

3. METHODOLOGY

This process was conducted by combining quantitative and qualitative approaches. The method is described in Fig. 1. The initial stage is determining the subject of interest and then identifying research questions that lead to the research objectives and reviewing the scope, identifying sources of research literature, conducting literature selection, collecting literature, and extracting data according to the RQ as shown in Table 2. Bibliometric analysis is conducted to answer RQ2 and 3 while paper distribution analysis is conducted to answer RQ1. Research topic analysis is making modeling topics from the collected papers and then analyzing trends from each modeling topic to answer RQ4.

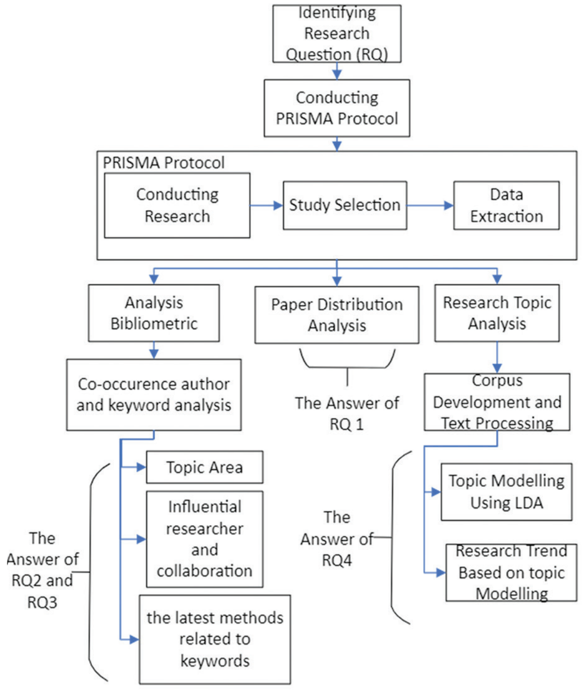

Identifying Research Question (RQ)
Conducting PRISMA Protocol
PRISMA Protocol
Conducting Research
Study Selection
Data Extraction
Analysis Bibliometric
Paper Distribution Analysis
Research Topic Analysis
Co-occurence author and keyword analysis
The Answer of RQ 1
Corpus Development and Text Processing
Topic Area
Influential researcher and collaboration
the latest methods related to keywords
The Answer of RQ2 and RQ3
The Answer of RQ4
Topic Modelling Using LDA
Research Trend Based on topic Modelling

Fig. 1. Methodology of systematic mapping study

The PRISMA protocol is paper searching related to the research topic to be analyzed. The research database resources that will be used to search the paper are Scopus, Science Direct, and IEEE research databases as shown in Fig. 2. Study selection is the stage of selecting a paper based on the query keywords used and the year of publication. The keyword queries from each research database are shown in Table 3. Meanwhile, the selected year is the publication of the last 5 years, namely 2017 to 2022.

As shown in Fig. 2, 479 papers were obtained from Scopus, 317 papers from IEEE, and 29 papers from Science Direct in the process of searching and gathering papers. After all the papers are combined, the total number of papers becomes 825 papers. Study selection is continued by excluding empty keywords and duplicate papers by title. From the excluded process, there are 511 papers left. The next stage is the data extraction of 511 papers to answer the research questions such as bibliometric analysis, paper distribution analysis, and research topic analysis.

Table 2. Research Question of SMS

Research Question (RQ)	Question
RQ1	What is the population of publications on mobile anchor localization research in 2017-2022?
RQ2	What is included in the research topic areas of mobile anchor localization?
RQ3	Who are the most influential and contributing researchers on the topic of mobile anchor localization research?
RQ4	What are topic modeling and the trend of each modeling topic in mobile anchor localization?

Bibliometric analysis determines the progress of studies that have been carried out related to the topic of mobile anchor localization research. Bibliometric mapping helps researchers to get a visualization of publication metadata so that it is easier to manage and analyze to identify research topics and clusters in certain disciplines, map authors, and map author collaboration as part of a framework to identify emerging technologies [24].

Bibliometric analysis was performed using the Vos Viewer tool. Vos Viewer is an open-source tool created by Nees Jan van Eck and Ludo Waltman at The Center for Science and Technology Studies (CWTS), Leiden University, The Netherlands. Vos Viewer features co-authorship mapping, keyword-based co-occurrence, and citation mapping [25].

As shown in Fig. 3, the bibliometric analysis covers topic areas based on keywords related to the topic and its trends, author collaboration, and author based on the number of publications and citations. The type of analysis and calculation method used in VOS Viewer is co-occurrence analysis based on author keywords and co-authorship analysis. Both are carried out using the full counting method, all have the same weight [14]. From all the documents obtained from all research databases, they are combined in a Scopus template with CSV (merged document) format then filtering the document and extracting data using Vos Viewer to obtain bibliometric visualization.

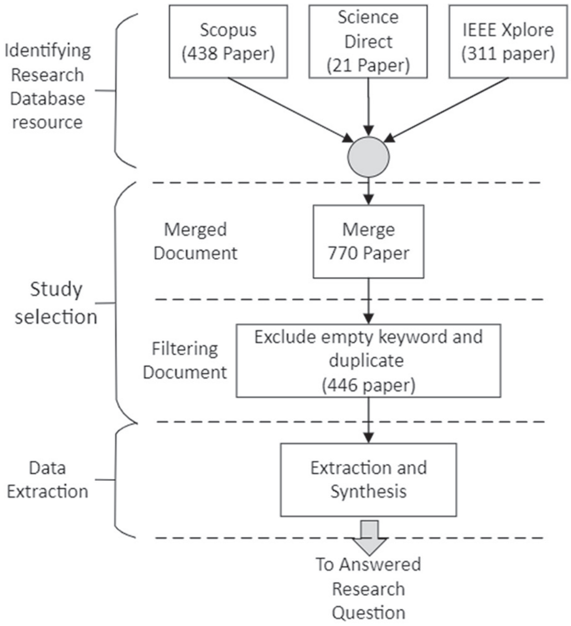


Fig. 2. PRISMA process

Research topic analysis which consists of corpus development and text processing, topic modeling, and research trend of topic modeling. The tool used is Orange Data Mining with the process structure as shown in Fig. 4. The data mining process is carried out using a corpus consisting of keywords and years. Corpus uses the Scopus template because of the merge document which only consists of the year and keywords.

The corpus is processed using preprocess text which will separate the text into smaller units (tokens) then filter and normalize (stemming, lemmatization), create n-grams, and tag tokens with part of speech labels. The text will house everything first to lowercase and remove the URL if there is a URL in the text. The tokenization selected is Regexp which will separate the text with the provided regex and remove punctuation. Filtering is done to delete or save word choices. A stop word is selected to remove stop words from words such as or, and, and in. Regex removes words that match the regular expression set to remove punctuation.

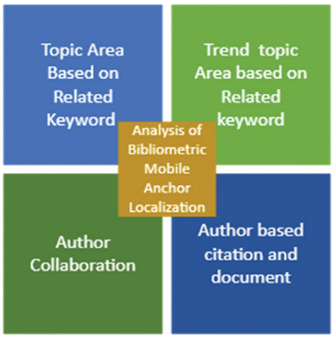

Topic Area Based on Related Keyword
Trend topic Area based on Related keyword
Analysis of Bibliometric Mobile Anchor Localization
Author Collaboration
Author based citation and document

Fig. 3. Bibliometric analysis

Table 3. Database research and query

Database Research	Query
Scopus	TITLE-ABS-KEY (mobile AND anchor AND localization) AND (LIMIT-TO (PUBYEAR , 2023) OR LIMIT-TO (PUBYEAR , 2022) OR LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017)) AND (LIMIT-TO (DOCTYPE , "cp") OR LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cr") OR LIMIT-TO (DOCTYPE , "ch") OR LIMIT-TO (DOCTYPE , "re")) AND (LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "ENGI"))
Science Direct	Title-abstract-keyword : Mobile Anchor Localization Year : 2017-2022 Subject Area : Computer Science and Engineering
IEEE Xplore	Keyword : Mobile Anchor Localization Year 2017-2022 Article Tipe : Research Article

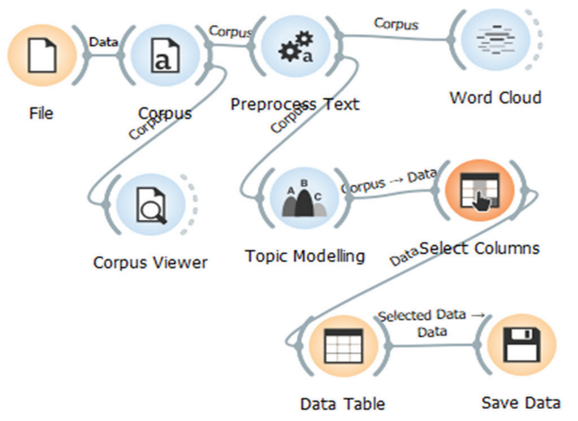


Fig. 4. Research topic analysis framework

The word cloud displays tokens in the corpus, the size of which represents the frequency of occurrence in the corpus or the average word count. Words will be listed based on weights that represent frequencies. We can find out what keywords are contained in the topic of mobile anchor localization and which keywords are often used in research from this word cloud. Fig. 5 shows the word cloud of keywords generated on the topic of mobile anchor localization. Some keywords that have a high-frequency weight of emergence from the topic of mobile anchor localization are localization, wireless, sensor, network, mobile, node, indoor, path planning, positioning, algorithm, and optimization.

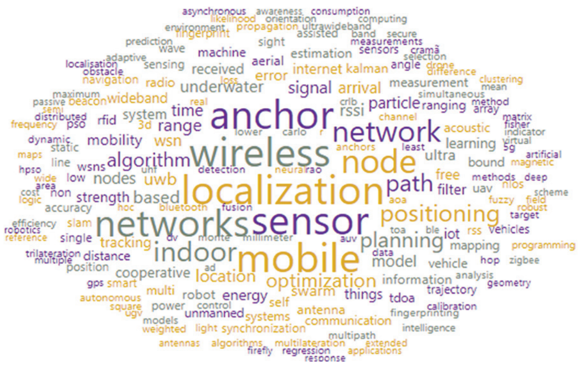

asynchronous awareness consumption
likelihood orientation computing
environment fingerprint propagation ultrawideband
prediction assisted band secure
wave sight measurements
localisation adaptive machine estimation sensors crama
obstacle sensing aerial selection
navigation received error internet kalman angle drone difference
maximum radio underwater signal arrival measurement clustering
passive beacon wideband loss simultaneous mean
semi time crlb particle ranging method
frequency distributed rfid system 3d range anchor rssi channel array
dynamic static mobility wsn lower carlo network acoustic matrix fisher indicator
maps line wsns algorithm wireless anchors least learning 5g virtual
hpso uwb node ultra bound artificial magnetic
wide low nodes detection neural rao free methods deep
area non logic cost strength based localization path filter uav nlos scheme
accuracy hoc bluetooth fusion aoa fuzzy field robust
efficiency slam networks sensor positioning target
robotics reference single tracking dv monte millimeter toa ble iot rss vehicles
trilateration distance indoor mobile auv planning mapping programming
multiple position data model hop zigbee
cooperative location ad vehicle analysis
gps smart multi optimization information geometry
autonomous robot energy self swarm things tdoa trajectory
square power control unmanned antenna fingerprinting calibration
ugv models systems communication intelligence
weighted light synchronization multipath
antennas algorithms multilateration extended
firefly regression applications
response

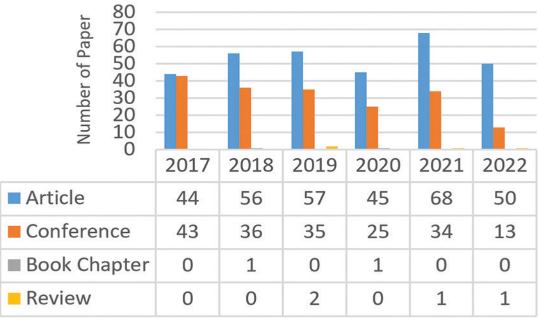
Fig. 5. Word cloud of mobile anchor localization

The function of topic modeling is to find abstract topics in the corpus based on the word groups found in each document and their respective frequencies. A document usually contains several topics in different proportions, so the widget also reports topic weights per document. LDA (Latent Dirichlet Allocation) is a probabilistic generative model of a corpus where the document is represented as a random mixture of latent subjects, each of which is defined by the word distribution as its core premise [26]. LDA is used to analyze text/word patterns and their interrelationships. LDA is an effective topic modeling technique that can be used for classification, feature selection, and information retrieval [27].

4. RESULT AND DISCUSSION

4.1. PAPER DISTRIBUTION ANALYSIS

RQ1 refers to the number of publications per year by article type. Data collection was carried out on 511 papers from filtering results based on the type of publication of articles, conferences, book chapters, and reviews every year. Fig. 6 shows a graph of publications on the topic of mobile anchor localization per year based on the type of publication. As previously explained, there are still few paper reviews and surveys in the field of mobile anchor localization. The graph shows that the topic of mobile anchor localization is mostly documented in the form of journals, followed by conferences.



	2017	2018	2019	2020	2021	2022
🟦 Article	44	56	57	45	68	50
🟧 Conference	43	36	35	25	34	13
⬜ Book Chapter	0	1	0	1	0	0
🟨 Review	0	0	2	0	1	1

Fig. 6. Publication of mobile anchor localization per year

4.2. KEYWORD CO-OCCURRENCE AND CO-AUTHOR ANALYSIS

RQ2 refers to the topic area of mobile anchor localization. Bibliometric analysis was performed to answer RQ 2. Related keywords are analyzed to find mobile anchor localization topic areas based on keywords. The results of the bibliometric visualization of the co-occurrence analysis based on author keywords in the Mobile Anchor Localization topic area are shown in Fig. 7. As a result, there are 50 keywords identified as meeting the threshold by the Vos Viewer. The link thickness in the figure shows the weight of each link, while the size of the nodes shows the accuracy of the related keyword, and the color shows the clusters on each node and link.

In bibliometrics, 9 clusters are formed with each cluster consisting of several keywords. As shown in Table 4, the cluster contains keyword data. The keyword of Cluster 1 consists of Angle of Arrival (AOA), cooperative localization, fingerprinting, GPS, indoor localization, machine learning, Receive Signal Strength (RSS), Time Difference of Arrival (TDoA), Time of Arrival (ToA), and ultra-wideband. This cluster represents the application of machine learning to indoor localization using several ranging techniques such as TDoA, ToA, RSS, and AOA. Implementation can use ultra-wideband (UWB) to be applied indoors as proposed by the following papers [28-31]. Cluster 2 represents the keywords localization error, mobile wireless sensor network, optimization, positioning, PSO, range free, received signal strength, and trilateration. The conclusions from cluster 2 are related to the research area for optimizing WSN dynamic environments using PSO with ranging range-free or RSS techniques in the positioning process using trilateration. Several papers describe research on this cluster [32, 33].

Keywords in cluster 3 are anchor, DV hop, IoT, localization, mobile beacon, mobile node RSSI, and wins. Research related to these keywords is the localization of mobile nodes using mobile anchors and emitting beacons on a WSN/IoT with RSSI or DV-hop ranging techniques. Paper [34] applies RSSI to the ranging technique and paper [35] uses RSSI which is converted into distance using machine learning. RSSI was also used in a paper on [7] underwater localization.

Cluster 4 is a topic of interest discussed in this paper, which is related to localization using mobile anchors in wireless sensor networks. Path planning is one part of the method for estimating the location of a sensor using a mobile anchor. One of which is proposing a localization algorithm based on ELPMA path planning [22]. A survey paper on path planning was also put forward by Sabale which discussed various path-planning techniques for mobile anchor localization [21]. Path planning is also applied to disaster management [36].

The keywords in the 5th cluster are also related to the mobile localization algorithm on the wireless sensor network using the Kalman filter or particle filter. The filter on the mobile anchor localization is used to overcome measurement noise. As in the paper [36], EKF (Extended Kalman Filter) is used to obtain accurate sensor node locations around the mobile anchor. Paper [34] proposes a Kalman Filter based on the Bounding Box Localization Algorithm to minimize energy, hardware costs, and computational complexity. The 6th cluster is related to location awareness in IoT, and position estimation using UAV anchors [34]. The 7th and 8th clusters are related to the WSN dynamic environment and accuracy [18, 19, 32, 37].

We try to highlight the nodes on the mobile anchor in cluster 4 to get a more focused keyword area. Figure 8 shows the topic of mobile anchor localization by average year which shows that many publications regarding mobile anchors were made in 2019. Meanwhile, the latest issues on mobile anchors are estimation and location awareness. Research on path planning and the use of UAVs to bring moving anchors was also widely used around 2019-2020.

The results in Figure 9 show that studies related to mobile anchor nodes are localization, wireless sensor network, localization algorithm, path planning, Receive Signal Strength Indicator (RSSI), range free, TDoA, connectivity, machine learning, mobile beacons, and anchor nodes. Bibliometric visualization shows that the closer the distance, the stronger the relationship between the keywords. In addition, the larger the node formed, the more often these keywords are used on the topic of mobile anchor localization.

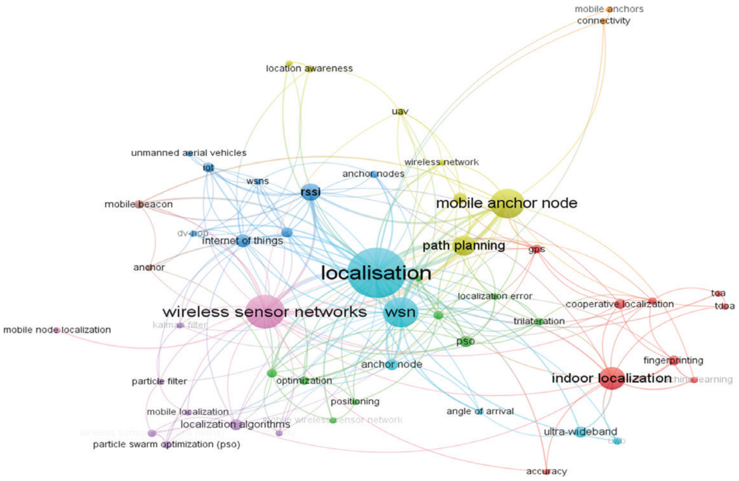


Fig. 7. Visualisasi Bibliometric

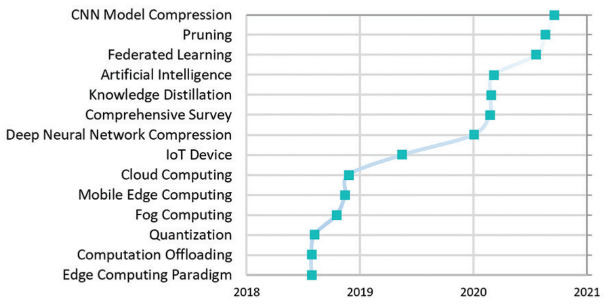


Fig. 8. Mobile anchor localization topic by an average year

The visualization results also show that the nodes most often associated with mobile anchor nodes are localization and wireless sensor networks. This shows that most of the mobile anchor publications are used for localization schemes related to wireless sensor networks. Another node is path planning, which is one of the keywords that often appears in localization publications. This shows that path planning is one part of the mobile anchor localization topic studied by researchers.

The trend topic area based on keywords can be seen in the bibliometric visualization in Fig. 9. Based on the color, we can see that machine learning is one of the methods proposed to solve the localization problem in the latest research on mobile anchor localization. UAV is also a node that is still in the current year and shows that UAV is one of the tools used as a mobile anchor in the current study.

RQ3 refers to the author's collaboration and the researchers who have most contributed to and influenced the research on mobile anchor localization. RQ3 can be answered by conducting a bibliometric analysis of the author's collaboration. The results of the bibliometric visualization of author collaboration using network visualization, overlay visualization, and density visualization where each node represents the author are shown in Fig. 10.

There are 6 research clusters formed in Fig. 10(a). Researchers with nodes close to each other have a stronger relationship based on bibliographic coupling. Researchers who are close together tend to cite the same publication or collaborate, whereas researchers with nodes that are far apart usually do not cite or collaborate on the same publication. For example, Liu Y collaborated with Yang C in research on the dynamic path planning method on mobile anchors[37]. Liu Y also collaborated with Shen Y in his research on single anchor passive localization[38]. From this co-author's bibliometrics, we can search for journals further based on the research clusters formed and the research area of each research cluster. In addition, we can also find out the author by our research topic area.

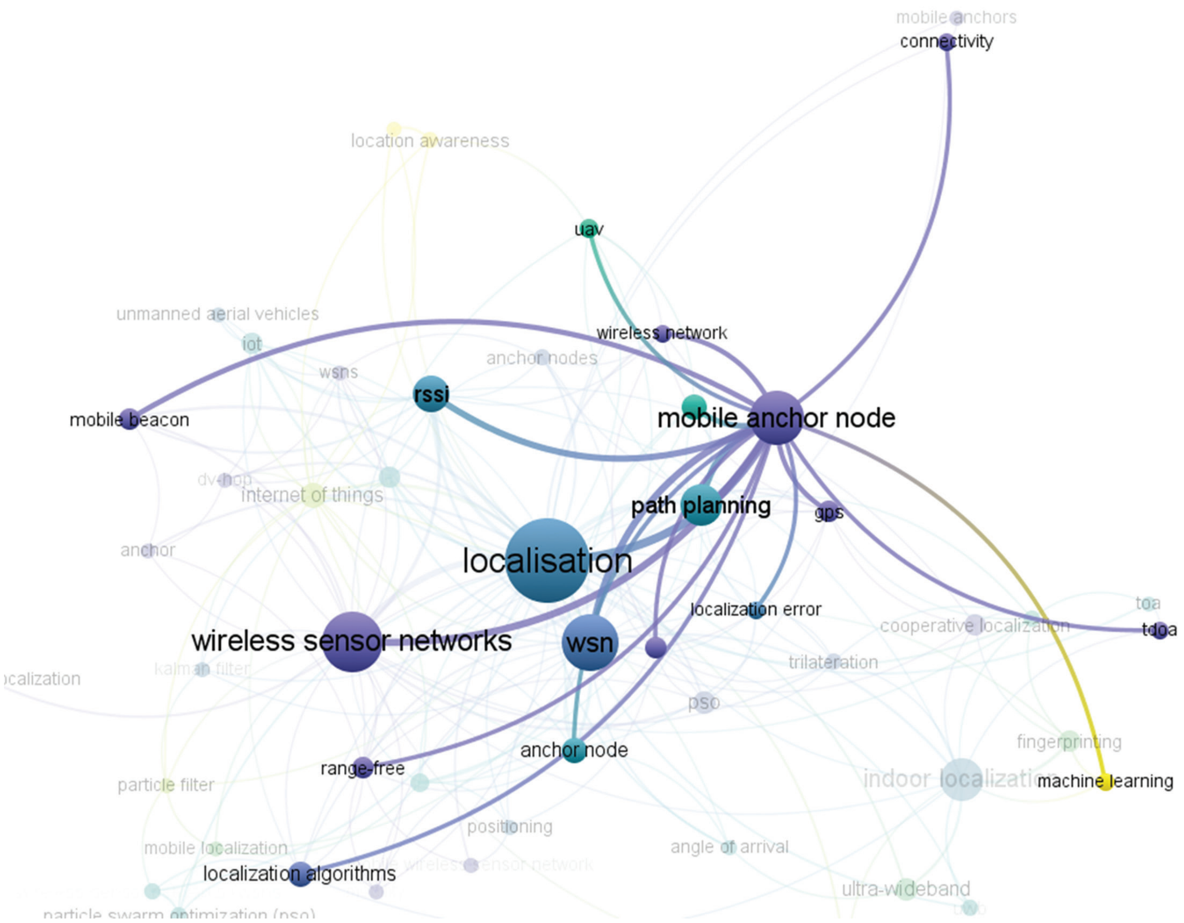
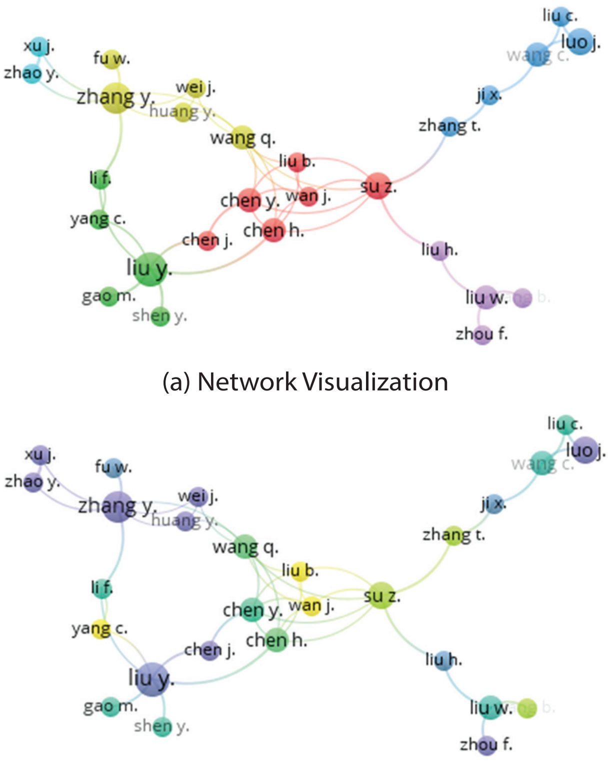


Fig. 9. The bibliometric trend of mobile anchor node topic area based on a related keyword.



(a) Network Visualization

(b) Overlay Visualization

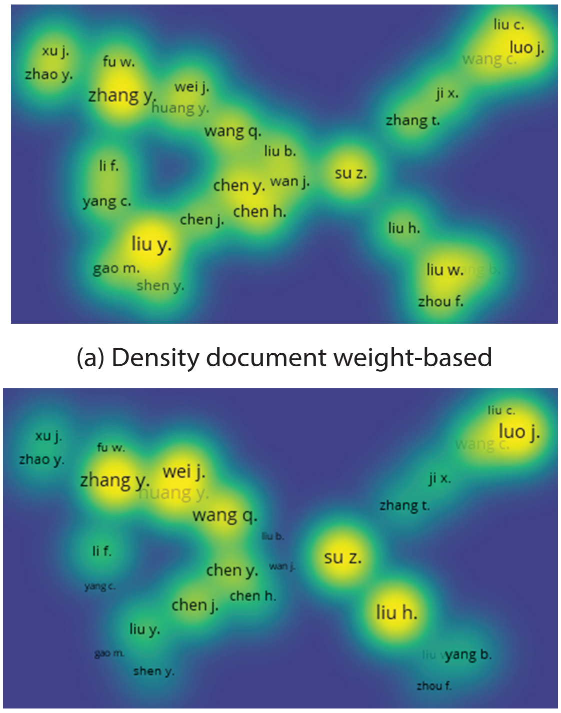
Fig. 10. Bibliometric co-author based on network and overlay visualization

In Fig. 10 (b) Liu B., Yang C., and Wan J., have yellow nodes so that they are the authors who are researching in the latest year, 2021. Liu B and Wan J proposed research on node localization algorithms in 3D environments using mobile anchors. While Yang C researched node localization using dynamic path planning in a 3D environment [37].

In Fig. 11 (a) with weights based on documents, Liu Y and Zhang Y is the author who has the most research on the topic of Mobile Anchor Localization. While Liu Y is the initials of several researchers including Yuan Liu, Liu Ying, Yan Liu, and Yiwen Liu who all researched mobile anchor localization. Yuan Liu conducted research on localization algorithms with collaborative and predictive algorithms, Liu Ying proposed localization using moving anchors based on network connectivity, and Yiwen Liu researched error analysis on positioning nodes on underwater acoustic sensors. Yan Liu proposed 2 studies consisting of node localization in a 3-dimensional environment using intelligent dynamic path planning and proposing the use of a single anchor. [38-40].

Zhang Y is the initials of several researchers who conducted research in localization including Yijin Zhang, Yuan Zhang, Yueyue Zhang, Yunzhou Zhang, Yuexia Zhang, Yuli Zhang, Yu Zhang, and Ying Zhang. However, from the listed paper, only Yuli Zhang who conducted the research on mobile anchors proposed a sensor node localization method, namely real-time localization based on neighbor nodes and real-time positioning based on mobile anchor nodes.

Fig. 11 (b) shows the initials of the researcher whose paper was cited by other researchers. The figure shows that papers from researchers with the initials Zhang Y were cited the most by other researchers, reaching a total of 58 citations. Furthermore, the researcher with the initials Luo J represents the researcher Juan Luo with 29 citations and Luo Junhai with 28 citations.



(a) Density document weight-based

(b) density citation web-based

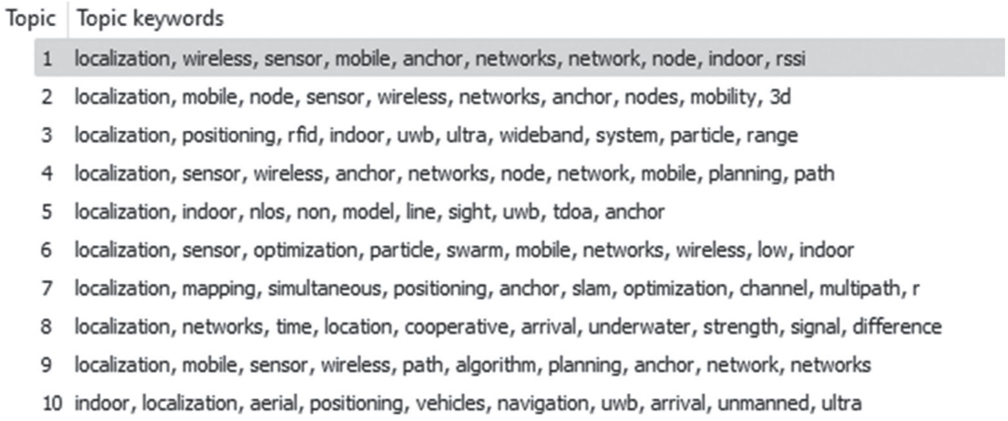
Fig. 11. Bibliometric of influential and contribute co-author

Juan Luo produced 3 papers proposing the ELPMA (Efficient Localization Algorithm based Path Planning for Mobile Anchor) method [41], GTMA (Group Tri-Mobile Anchor) [42], and indoor localization based on extracting trusted fingerprints [43]. Meanwhile, Luo Junhai produced a paper on a localization algorithm using a two-phase time synchronization-free for the localization of underwater sensor networks [44].

4.3. TOPIC MODELLING AND TREND

Topic modeling will be generated using Orange Data Mining. Fig. 11 shows the 10 topic modeling results generated from the year and keyword databases from 511 papers. The topic area on the ten topics of modeling orange data mining shows compatibility with the topic area generated using the Vos Viewer. As in topic 3, which relates to the keywords localization, positioning, RFID, indoor, UWB, ultra, wideband, system, particle, and range which are related to cluster 1 on Vos Viewer. Likewise, the 9th topic model related to cluster 4 on the Vos Viewer.

A Trend analysis is created after generating the topic modeling. Trend analysis is performed by calculating the average weight value of each keyword on each topic model based on the year. The results of the trend graph for each modeling topic are shown in Fig. 13. The results show that several topics experienced a decreasing trend, and several other topics experienced an increasing trend. If the topic model has an upward trend line, then the topic is included in the promising topic. From Fig. 11, an increase in the topic models 2, 3, 7, 8, 9, and 10. The 2nd topic model relates to the topic of sensor node localization using a 3-dimensional environment.



Topic	Topic keywords
1	localization, wireless, sensor, mobile, anchor, networks, network, node, indoor, rssi
2	localization, mobile, node, sensor, wireless, networks, anchor, nodes, mobility, 3d
3	localization, positioning, rfid, indoor, uwb, ultra, wideband, system, particle, range
4	localization, sensor, wireless, anchor, networks, node, network, mobile, planning, path
5	localization, indoor, nlos, non, model, line, sight, uwb, tdoa, anchor
6	localization, sensor, optimization, particle, swarm, mobile, networks, wireless, low, indoor
7	localization, mapping, simultaneous, positioning, anchor, slam, optimization, channel, multipath, r
8	localization, networks, time, location, cooperative, arrival, underwater, strength, signal, difference
9	localization, mobile, sensor, wireless, path, algorithm, planning, anchor, network, networks
10	indoor, localization, aerial, positioning, vehicles, navigation, uwb, arrival, unmanned, ultra

Fig. 12. Topic Modelling

Several studies have implemented mobile anchors in 3-dimensional environments such as underwater WSN (UWSN) [35, 45, 46]. and aerial anchors [47, 48]. Model topic 3 relates to UWB, indoor, localization, positioning, RFID, ultra-wideband, particle, and system. Research with mobile anchors related to UWB among them is [29-31].

Furthermore, the topic of the 7th model is related to the keyword's localization, mapping, simultaneous, positioning, anchor, slam, optimization, channel, and multipath. WSN is one of the infrastructures used to produce SLAM (Simultaneous Localization and Mapping). Optimization is also carried out to produce reliable mapping and localization such as the use of filters and computational intelligence [49]. Another topic model that has an increasing trend is the 8th topic model which relates to localization, network, time, location, cooperative, arrival, underwater, strength, signal, and difference and is related to mobile anchor research underwater [30, 40].

The 9th topic model has an upward trend and relates to localization, mobile, sensors, wireless, path, algorithm, planning, anchor, and network. So, this topic relates to path planning algorithms and localization algorithms which propose various static or dynamic path planning algorithms to optimize accuracy, energy, and coverage [22, 36, 41, 51]. The last topic is the 10th topic model related to the keywords indoor, localization, aerial, positioning, vehicles, navigation, UWB, arrival, and unmanned. UWB is used to position the robot which can be a UAV that is carried out indoors as a navigation function as proposed in papers [52-55].

4.4 MOBILE ANCHOR LOCALIZATION PAPER IN 2023

The search for papers in 2023 was carried out to find the latest research information on mobile anchor localization and obtained 4 journals and 2 conferences.

A paper at the end of 2022 was put forward by Fei Tong *et. al.* [56] who proposed the single anchor mobile localization (TSAL) method. In 2023 Huimin Chen et al [57] proposed a mobile anchor using LORA to reduce deployment costs. Gauss Markov-based mobile anchor localization (GM-MAL) was proposed by Song Xinchao et al which can improve localization precision [58].

Vaishali R Kulkarni [59] conducted a comparative analysis of the use of static and mobile anchors on localization sensors. Rinkesh Mittal et al [60] proposed a study aimed at reducing localization errors by developing moving paths and anchors. Oumaima Liouane proposed an analytical probabilistic model for estimating the multi-hop distance between the mobile anchor and the unknown nodes. The relationship between hop count and distance estimation is represented as a nonlinear function and using the recursive least square algorithm can present new formulas from the DV-Hop localization algorithm on mobile anchor localization [61].

5. CONCLUSION

A systematic SMS process can enable us to dig up information about a broad topic area regarding mobile anchor localization and the position of the research focus aimed at that topic area. Bibliometrics can show influential research topic areas, and recent methods are based on keywords to provide a helicopter view of mobile anchor localization. Research topic analysis allows us to get topic modeling and see the current trends. Previous paper reviews have not applied the SMS method in the classification process and are still focused on explaining the method from existing papers.

SMS using the PRISMA method is recommended to use the right keywords according to the terms used in related papers. By using the right keywords, the collected paper will focus on the intended field. We also find out the chronology of published papers from 2017 to 2022. This can assure us that research in this field is still being carried out so that it is still possible to conduct research in this field.

Bibliometric analysis of the topic area provides a helicopter perspective on the research area and provides insight for other researchers to determine more specific research topic areas and novelty opportunities.

The results show that the mobile anchor localization field is related to localization, wireless sensor networks, and indoor localization research areas. Focusing on mobile anchors, we can find out more specifically what is related to mobile anchors like path planning, machine learning, localization algorithms, and others. The bibliometric results also show machine learning as the latest method such as in path planning.

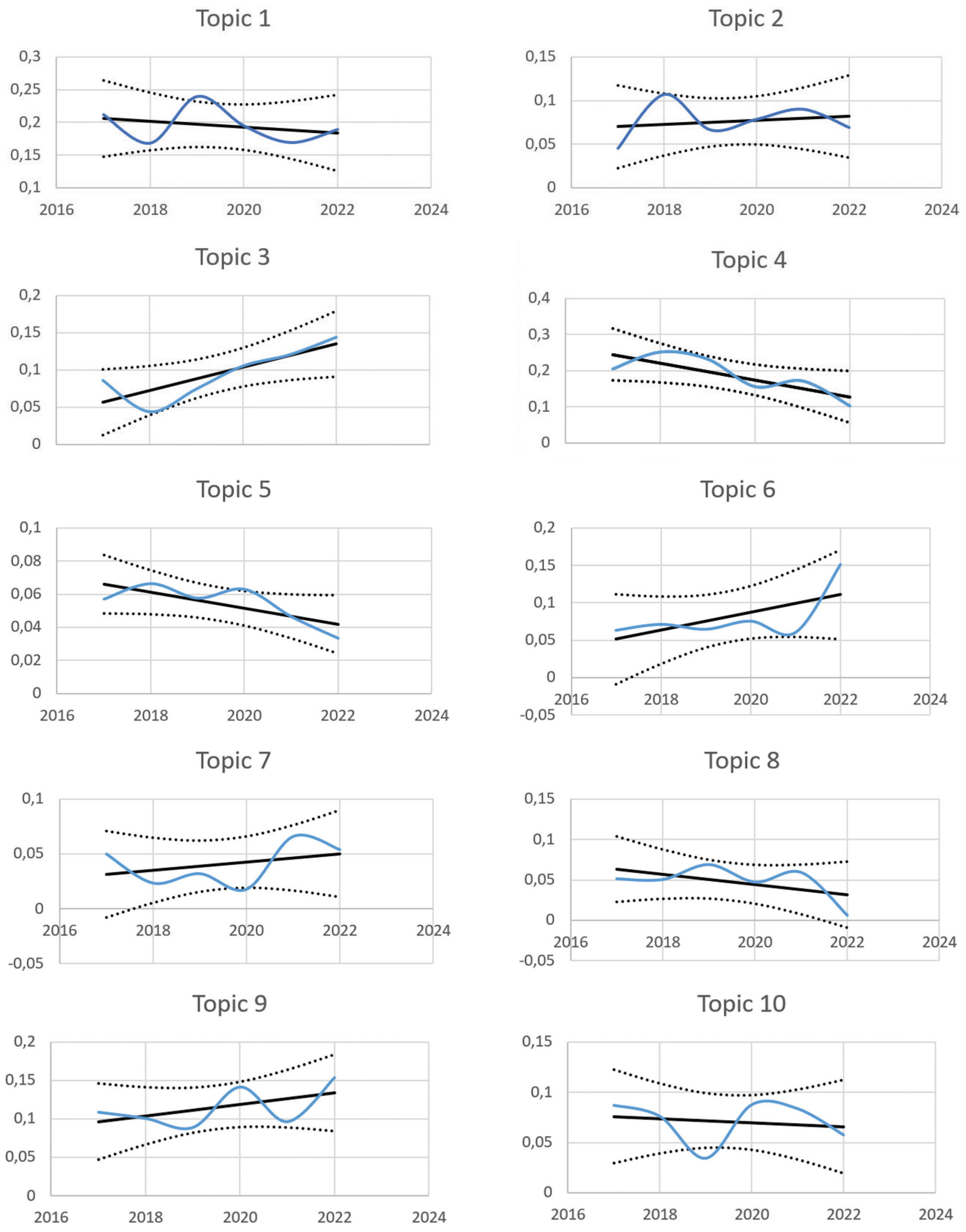


Fig. 13. The trend of each topic modeling

Co-author bibliometric analysis can find out which authors have an influence on the research field from the number of citations to their papers, find out which authors cite each other, and the contributor which allows the same topic area. It can be continued by looking for a community of these researchers. Joining the community is one solution to find a solution by holding discussions or correspondence with other researchers. Topic modeling is a way to find out the topics we can take as research. From the results of topic modeling, 6 topic models are included in the promising topic, which means that these 6 topics still have the potential to be investigated further. For example, in the 9th topic related to path planning, the trend tends to rise so that it can be chosen as the research topic to be studied.

6. REFERENCES

[1] W. Osamy, A. M. Khedr, A. Salim, A. I. Al Ali, A. El-sawy, "Coverage, Deployment and Localization Challenges in Wireless Sensor Networks Based on Artificial Intelligence Techniques: A Review", IEEE Access, Vol. 10, 2022, pp. 30232-30257.

[2] V. Sneha, M. Nagarajan, "Localization in Wireless Sensor Networks: A Review", Cybernetics and Information Technologies, Vol. 20, No. 4, 2020, pp. 3-26.

[3] R. H. Hussain, S. R. Saleh, "Hybrid Wireless Sensors Network for Tracking Animals", Journal of Engineering Science and Technology, Vol. 16, No. 6, 2021, pp. 4958-4974.

[4] P. Dasari, G. Krishna, J. Reddy, A. Gudipalli, "Forest fire detection using wireless sensor networks", International Journal on Smart Sensing and Intelligent Systems, Vol. 13, No. 1, 2020, pp. 1-8.

[5] D. Oh, J. Han, "Smart search system of autonomous flight UAVs for disaster rescue", Sensors, Vol. 21, No. 20, 2021.

[6] H. Huang, A. V. Savkin, M. Ding, C. Huang, "Mobile robots in wireless sensor networks: A survey on tasks", Computer Networks, Vol. 148, 2019, pp. 1-19.

[7] Y. Sun, Y. Yuan, Q. Xu, C. Hua, X. Guan, "A mobile anchor node assisted RSSI localization scheme in underwater wireless sensor networks", Sensors, Vol. 19, No. 20, 2019.

[8] S. Sivasakthiselvan, V. Nagarajan, "Localization Techniques of Wireless Sensor Networks: A Review", Proceedings of the IEEE International Conference on Communication and Signal Processing, Chennai, India, 28-30 July 2020, pp. 1643-1648.

[9] J. Kumari, P. Kumar, S. K. Singh, "Localization in three-dimensional wireless sensor networks: a survey", The Journal of Supercomputing, Vol. 75, No. 8, 2019.

[10] L. Chelouah, F. Semchedine, L. Bouallouche-Medjkoune, "Localization protocols for mobile wireless sensor networks: A survey", Computers and Electrical Engineering, Vol. 71, 2018, pp. 733-751.

[11] K. Petersen, S. Vakkalanka, L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update", Information and Software Technology, Vol. 64, 2015, pp. 1-18.

[12] "Guidelines for performing Systematic Literature Reviews in Software Engineering", EBSE Technical Report, EBSE-2007-01, https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf (accessed: 2023)

[13] B. Kitchenham, P. Brereton, "A Systematic Review of Systematic Review Process Research in Software Engineering", Information and Software Technology, Vol. 55, No. 12, 2013.

[14] B. A. Kitchenham, D. Budgen, O. Pearl Brereton, "Using mapping studies as the basis for further research - A participant-observer case study", Information and Software Technology, Vol. 53, No. 6, 2011, pp. 638-651.

[15] K. Petersen, H. Flensburg, R. Feldt, M. Mattsson, S. Mujtaba, "Systematic Mapping Studies in Software Engineering Understanding Agile Practice Interdepedencies - A Survey View project A Case Study on Integrating UX Practices into Software Development Organizations: A Socio-technical Perspective on Challenges in Practice View project Systematic Mapping Studies in Software Engineering", https://www.researchgate.net/publication/228350426 (accessed: 2023)

[16] C. Marshall, P. Brereton, "Tools to support systematic literature reviews in software engineering: A mapping study", Proceedings of the ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, Baltimore, MD, USA, 10-11 October 2013, pp. 296-299.

[17] K. Mondal, A. Karmakar, P. S. Mandal, "Path planning algorithms for mobile anchors towards range-free localization", Journal of Parallel and Distributed Computing, Vol. 97, 2016, pp. 35-46.

[18] G. Han, J. Jiang, C. Zhang, T. Q. Duong, M. Guizani, G. K. Karagiannidis, "A Survey on Mobile Anchor Node Assisted Localization in Wireless Sensor Networks", IEEE Communications Surveys and Tutorials, Vol. 18, No. 3, 2016, pp. 2220-2243.

[19] S. Halder, A. Ghosal, "A survey on mobile anchor assisted localization techniques in wireless sensor networks", Wireless Networks, Vol. 22, No. 7, 2016, pp. 2317-2336.

[20] H. Abukhalaf, "A Review of Mobile-Assisted Localization Algorithms in Wireless Sensor Networks", International Journal of Innovations in Engineering and Technology, Vol. 12, No. 2, 2019.

[21] Y. Long, J. Liang, "Mobile anchor assisted localization and path-planning techniques in wireless sensor networks: Challenges and Solutions", Journal of Physics: Conference Series, Institute of Physics Publishing, Vol. 1176, 2019.

[22] K. Sabale, S. Mini, "Localization in Wireless Sensor Networks with Mobile Anchor Node Path Planning Mechanism", Information Sciences, Vol. 579, 2021, pp. 648-666.

[23] K. Sabale, S. Mini, "Path planning mechanism for mobile anchor-assisted localization in wireless sensor networks", Journal of Parallel and Distributed Computing, Vol. 165, 2022, pp. 52-65.

[24] I. Tanudjaja, G. Y. Kow, "Exploring Bibliometric Mapping in NUS using BibExcel and VOSviewer", IFLA WLIC 2018 - Kuala Lumpur, Malaysia - Transform Libraries, Transform Societies, 2018.

[25] N. J. van Eck, L. Waltman, "Visualizing Bibliometric Networks", Measuring Scholarly Impact, Springer International Publishing, 2014, pp. 285-320.

[26] D. M. Blei, A. Y. Ng, M. B. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research, Vol. 3, 2003, pp. 993-1022.

[27] P. Kherwa, P. Bansal, "Topic Modeling: A Comprehensive Review", EAI Endorsed Transactions on Scalable Information Systems, Vol. 7, No. 24, 2020, pp. 1-16.

[28] B. Großwindhager et al. "SALMA: UWB-based single-anchor localization system using multipath assistance", Proceedings of the 16th Conference on Embedded Networked Sensor Systems, Association for Computing Machinery, 2018, pp. 132-144.

[29] A. Schjørring, A. L. Cretu-Sircu, I. Rodriguez, P. Cederholm, G. Berardinelli, P. Mogensen, "Performance Evaluation of a UWB Positioning System Applied to Static and Mobile Use Cases in Industrial Scenarios", Electronics, Vol. 11, No. 20, 2022.

[30] J. Peña Queralta, C. M. Almansa, F. Schiano, D. Floreano, T. Westerlund, "UWB-based System for UAV Localization in GNSS-Denied Environments: Characterization and Dataset", https://github.com/TIERS/UWB (accessed: 2023)

[31] M. Dong, "A Low-Cost NLOS Identification and Mitigation Method for UWB Ranging in Static and Dynamic Environments", IEEE Communications Letters, Vol. 25, No. 7, 2021, pp. 2420-2424.

[32] P. Singh, A. Khosla, A. Kumar, M. Khosla, "Optimized localization of target nodes using single mobile anchor node in wireless sensor network", AEU - International Journal of Electronics and Communications, Vol. 91, 2018, pp. 55-65.

[33] H. Bao, B. Zhang, C. Li, Z. Yao, "Mobile anchor assisted particle swarm optimization (PSO) based localization algorithms for wireless sensor networks", Wireless Communications and Mobile Computing, Vol. 12, No. 15, 2012, pp. 1313-1325.

[34] Y. Zhao, J. Xu, J. Jiang, "RSSI Based Localization with Mobile Anchor for Wireless Sensor Networks", Communications in Computer and Information Science, Springer Verlag, 2018, pp. 176-187.

[35] M. J. Hazar, B. N. Shaker, L. R. Ali, E. R. Alzaidi, "Using Received Strength Signal Indication for Indoor Mobile Localization Based on Machine Learning Technique", http://www.webology.org/2020/v17n1/a206.pdfhttp://www.webology.org/2020/v17n1/a206.pdf (accessed: 2023)

[36] G. Han, X. Yang, L. Liu, W. Zhang, M. Guizani, "A Disaster Management-Oriented Path Planning for Mobile Anchor Node-Based Localization in Wireless Sensor Networks", IEEE Transactions on Emerging Topics in Computing, Vol. 8, No. 1, 2020, pp. 115-125.

[37] S. Jia, C. Yang, X. Chen, Y. Liu, F. Li, "Intelligent Three-dimensional Node Localization Algorithm Using Dynamic Path Planning", Recent Advances in Electrical & Electronic Engineering (Formerly Recent Patents on Electrical & Electronic Engineering), Vol. 14, No. 5, 2021, pp. 586-596.

[38] Y. Liu, Y. Shen, M. Z. Win, "Single-Anchor Passive Localizationof Full-Duplex Agents", Proceedings of the IEEE International Conference on Communications, Kansas City, MO, USA, 20-24 May 2018.

[39] Y. Liu, Z. H. Qian, "Moving anchor node localization algorithm based on network connectivity", Tongxin Xuebao/Journal on Communications, Vol. 38, No. 4, 2017, pp. 149-157.

[40] Y. Liu, J. Chen, "A collaborative and predictive localization algorithm for wireless sensor networks", KSII Transactions on Internet and Information Systems, Vol. 11, No. 7, 2017, pp. 3480-3500.

[41] B. F. Gumaida, J. Luo, "ELPMA: Efficient Localization Algorithm Based Path Planning for Mobile Anchor in Wireless Sensor Network", Wireless Personal Communications, Vol. 100, No. 3, 2018, pp. 721-744.

[42] B. Gumaida, C. Liu, J. Luo, "GTMA: Localization in wireless sensor network based a group of tri-mobile anchors", Journal of Computational and Theoretical Nanoscience, Vol. 14, No. 1, 2017, pp. 847-857.

[43] J. Luo, X. Yin, Y. Zheng, C. Wang, "Secure indoor localization based on extracting trusted fingerprint", Sensors, Vol. 18, No. 2, 2018.

[44] J. Luo, L. Fan, "A two-phase time synchronization-free localization algorithm for underwater sensor networks", Sensors, Vol. 17, No. 4, 2017.

[45] Y. Sun, Y. Yuan, Q. Xu, C. Hua, X. Guan, "A mobile anchor node assisted RSSI localization scheme in underwater wireless sensor networks", Sensors, Vol. 19, No. 20, 2019.

[46] H. L. Yuan, J. Y. Lu, X. M. Zhang, "Research on mobile anchor node localization method based on hierarchical", Applied Mechanics and Materials, 2014, pp. 362-365.

[47] B. Yuan et al. "A UAV-Assisted Search and Localization Strategy in Non-Line-of-Sight Scenarios", IEEE Internet of Things Journal, Vol. 9, No. 23, 2022, pp. 23841-23851.

[48] V. Annepu, R. Anbazhagan, "Implementation of an efficient extreme learning machine for node localization in unmanned aerial vehicle assisted wireless sensor networks", International Journal of Communication Systems, Vol. 33, No. 10, 2020.

[49] Y. Yang, J. Liu, W. Wang, Y. Cao, H. Li, "Incorporating SLAM and mobile sensing for indoor CO2 monitoring and source position estimation", Journal of Cleaner Production, Vol. 291, 2021, p. 125780.

[50] Y. Lin, H. Tao, Y. Tu, T. Liu, "A Node Self-Localization Algorithm with a Mobile Anchor Node in Underwater Acoustic Sensor Networks", IEEE Access, Vol. 7, 2019, pp. 43773-43780.

[51] G. Yu, H. Ma, D. Witarsyah, "Optimal path selection algorithm for mobile beacons in sensor network under non-dense distribution", Open Physics, Vol. 16, No. 1, 2018, pp. 1066-1075.

[52] S. Naghdi, K. O'Keefe, "Improving Bluetooth-based Indoor Positioning Using Artificial Networks", Proceedings of the International Conference on Indoor Positioning and Indoor Navigation, Banff, Alberta, Canada, 13-16 October 2015.

[53] B. Yuan et al. "A UAV-Assisted Search and Localization Strategy in Non-Line-of-Sight Scenarios", IEEE Internet of Things Journal, Vol. 9, No. 23, 2022.

[54] J. Tiemann, C. Wietfeld, "Scalable and Precise Multi-UAV Indoor Navigation using TDOA-based UWB Localization", Proceedings of the International Conference on Indoor Positioning and Indoor Navigation, Sapporo, Japan, 18-21 September 2017.

[55] A. Benini, A. Mancini, S. Longhi, "An IMU/UWB/vision-based extended kalman filter for mini-UAV localization in indoor environment using 802.15.4a wireless sensor network", Journal of Intelligent and Robotic Systems: Theory and Applications, Vol. 70, No. 1-4, 2013, pp. 461-476.

[56] F. Tong, B. Ding, Y. Zhang, S. He, Y. Peng, "A Single-Anchor Mobile Localization Scheme", IEEE Transactions on Mobile Computing, 2022. (in press)

[57] H. Chen et al. "A Lightweight Mobile-Anchor-based Multi-Target Outdoor Localization Scheme using LoRa Communication", IEEE Transactions on Green Communications and Networking, 2023. (in press)

[58] S. Xinchao, Z. Yongsheng, W. Lizhi, "Gauss-Markov-based mobile anchor localization (GM-MAL) algorithm based on local linear embedding optimization in internet of sensor networks", Cognitive Systems Research, Vol. 52, 2018, pp. 138-143.

[59] V. R. Kulkarni, "Comparative Analysis of Static and Mobile Anchors in Sensor Localization", Proceedings of the International Conference on Device Intelligence, Computing and Communication Technologies, Dehradun, India, 2023.

[60] R. Mittal, S. Sadiq, P. Singla, "Improved Localization Algorithm to Optimizing the Trajectory of Anchor Node for Wireless Body Area Network", Proceedings of the 13th International Conference on Cloud Computing, Data Science & Engineering, Noida, India: 2023.

[61] O. Liouane, S. Femmam, T. Bakir, A. Ben Abdelali, "New Online DV-Hop Algorithm via Mobile Anchor for Wireless Sensor Network Localization", Tsinghua Science and Technology, Vol. 28, No. 5, 2023, pp. 940-951.

A robust speech enhancement method in noisy environments

Original Scientific Paper

Nesrine Abajaddi
IMMII Laboratory,
Faculty of Sciences & Technics,
Hassan First University, Settat, Morocco
n.abajaddi@uhp.ac.ma

Badia Mounir
LAPSSII Laboratory,
High School of Technology,
Cadi Ayyad University, Safi, Morocco
mounirbadia2014@gmail.com

Youssef Elfahm
IMMII Laboratory,
Faculty of Sciences & Technics,
Hassan First University, Settat, Morocco
y.elfahm@uhp.ac.ma

Abdelmajid Farchi
IMMII Laboratory,
Faculty of Sciences & Technics,
Hassan First University, Settat, Morocco
abdelmajid.farchi1@gmail.com

Abstract *– Speech enhancement aims to eliminate or reduce undesirable noises and distortions, this processing should keep features of the speech to enhance the quality and intelligibility of degraded speech signals. In this study, we investigated a combined approach using single-frequency filtering (SFF) and a modified spectral subtraction method to enhance single-channel speech. The SFF method involves dividing the speech signal into uniform subband envelopes, and then performing spectral over-subtraction on each envelope. A smoothing parameter, determined by the a-posteriori signal-to-noise ratio (SNR), is used to estimate and update the noise without the need for explicitly detecting silence. To evaluate the performance of our algorithm, we employed objective measures such as segmental SNR (segSNR), extended short-term objective intelligibility (ESTOI), and perceptual evaluation of speech quality (PESQ). We tested our algorithm with various types of noise at different SNR levels and achieved results ranging from 4.24 to 15.41 for segSNR, 0.57 to 0.97 for ESTOI, and 2.18 to 4.45 for PESQ. Compared to other standard and existing speech enhancement methods, our algorithm produces better results and performs well in reducing undesirable noises.*

Keywords*: speech enhancement, single frequency filtering, spectral subtraction, envelopes*

1. INTRODUCTION

Speech enhancement is an active area of research that aims to improve the quality of degraded speech and preferably its intelligibility. One of the most significant tasks of speech enhancement is to reduce or remove noise that has degraded speech quality, and this is an active area of research [1-3]. Noise reduction techniques are used in many applications such as mobile phones [4], speech recognition [5], teleconferencing systems [6], voice over internet protocol (VoIP) [7], and hearing aids. Most speech enhancement systems use a single microphone for economic reasons, even though better results can be obtained using multiple microphones [8]. The field of single-channel speech enhancement continues to be a significant area of research due to its simplicity and computational efficiency. These systems, which are based on a single microphone, employ adaptive filtering techniques to reduce the impact of noisy regions in speech signals that have a low SNR while preserving those with a high SNR. Within this context, the short-term spectral amplitude plays a crucial role in preserving speech quality and intelligibility, compared to phase information [9]. Furthermore, the speech enhancement algorithms can be classified as follows [10]: i) spectral subtraction algorithms, ii) statistical model algorithms, iii) subspace algorithms, and iv) machine learning algorithms.

Spectral subtraction algorithms, developed in the late 1970s, are effective and widely used in the spectral domain [11]. This technique involves subtracting the estimated noise from the noisy speech, assuming that the noise is additive and uncorrelated with the clean speech signal. However, this approach is susceptible to generating musical noise, which can be bothersome to listeners. To reduce this side effect, several techniques have been proposed, including applying a half-wave rectifier and setting all negative values to zero [12-14]. Another approach is to use an over-reduction factor and a spectral floor factor, but speech may be distorted and consequently lead to a loss of intelligibility [15, 16]. Due to the varying spectral distribution of the signal and noise across different frequency ranges, researchers have directed their attention toward subband processing methods. A widely recognized technique in the field of audio signal enhancement is the multiband spectral subtraction (MBSS) method [17]. This approach involves decomposing the signal into subbands to exploit the spectral information and apply different noise treatments in each subband, considering the varying impact of noise on the speech spectrum. To further enhance the performance of the multiband spectral subtraction approach, the authors in [18] propose a technique called MBSS_CBRS (multiband spectral subtraction with critical band rate scaling). This technique is designed to align with the characteristics of the auditory system, aiming to approximate the benefits of human perception and to improve the effectiveness of speech enhancement algorithms. Other studies concentrate on speech production features. In the study described in [19], the authors specifically investigate the harmonic properties of vowels (SS_HP) and utilize the sigmoid function to empirically determine the values for over-subtraction and the spectral reservation factor. Furthermore, researchers have recognized that speech can be considered as an amplitude-modulated signal, leading them to explore speech enhancement techniques in the modulation domain. In [20] the authors specifically employed the coherent harmonic demodulation technique (SE_CHD) to get the subband signals. This approach relies on a prior signal-to-noise ratio (SNR) and utilizes a gain function derived using the minimum mean square error (MMSE) approach.

Traditionally, speech enhancement algorithms commonly rely on the short-term Fourier transform (STFT) to estimate the short-term spectrum of a signal [17, 19]. This involves dividing the signal into subband signals through consecutive windowing or filter bank operations [18, 20]. A novel approach known as single-frequency filtering (SFF) has recently been introduced. [21]. This technique offers high spectral and temporal resolution and eliminates the effects of windowing through filtering. By employing filtering at the maximum frequency of $f_s/2$, SFF can capture both amplitude and phase information at each frequency. SFF has been explored in various applications, including voice activity detection [21], epoch extraction [22], hyper-nasality assessment [23], and dysarthria evaluation [24, 25]. This approach is gaining popularity for segmenting speech into multiple frequency bands due to its exceptional time-frequency resolution.

In this work, our contribution includes the development of a new algorithm combining the recent approach called SFF and the modified spectral subtraction method to enhance the quality and intelligibility of degraded speech signals. The integration of the SFF approach with noise estimation using the SNR is based on the identification of segments with high SNRs. In practice, the power of the noise tends to be lower near zero-bandwidth resonator of the single frequency, while the power of the speech signal, particularly if present, is relatively high. As a result, windows exhibiting high SNR will appear at different times for various frequencies. Therefore, we used the SFF for calculating the envelopes which help to reduce or eliminate unwanted noise concentrated around one or more specific frequencies. For each envelope, we estimated the noise from previous speech frames and applied a smoothing parameter to balance noise reduction and speech quality preservation. The experiments were conducted on various types of real-world noise, including car, train, restaurant, airport, and street noises, as well as machine-generated white Gaussian noise, to evaluate the performance of our proposed algorithm. The results demonstrate that our method outperforms the previously mentioned existing methods [17-20].

The paper is organized into three main sections. In Section 2, an analysis and synthesis of the single-frequency approach, envelope manipulation, modified spectral subtraction, and noise estimation are covered. Section 3 provides detailed outcomes of the conducted experiments, highlighting the performance of the proposed method. Finally, Section 4 serves as the conclusion, summarizing the key findings and implications of the research.

2. PROPOSED SPEECH ENHANCEMENT MODEL

This section aims to explore the potential of employing the single-frequency filtering (SFF) approach and modified spectral subtraction to enhance the degraded speech. The proposed method involves three main steps (Fig.1). First, the SFF analysis is performed to generate spectral envelopes at specific frequencies of interest. Then, the modified spectral subtraction is applied at each frame for each envelope. Finally, the SFF synthesis combines the processed envelopes to reconstruct the original speech. Our objective is to develop an algorithm that utilizes validated blocks to ensure the high quality and intelligibility of the speech signal. In order to achieve this, we have verified that the SFF analysis-synthesis method does not have any negative impact on the quality and/or intelligibility of the speech signal. This step aims to ensure that the SFF processing does not introduce any undesired effects or degradation to the processed signal. The second step involves determining whether the enhancement should be focused on the envelope or the phase of the signal. Choosing the optimal element is important for improving the quality and intelligibility of the speech signal. Finally, it is essential to define the suitable parameters for the spectral subtraction method to reduce the noise in the speech signal and achieve the best results in terms of improving its quality and intelligibility.

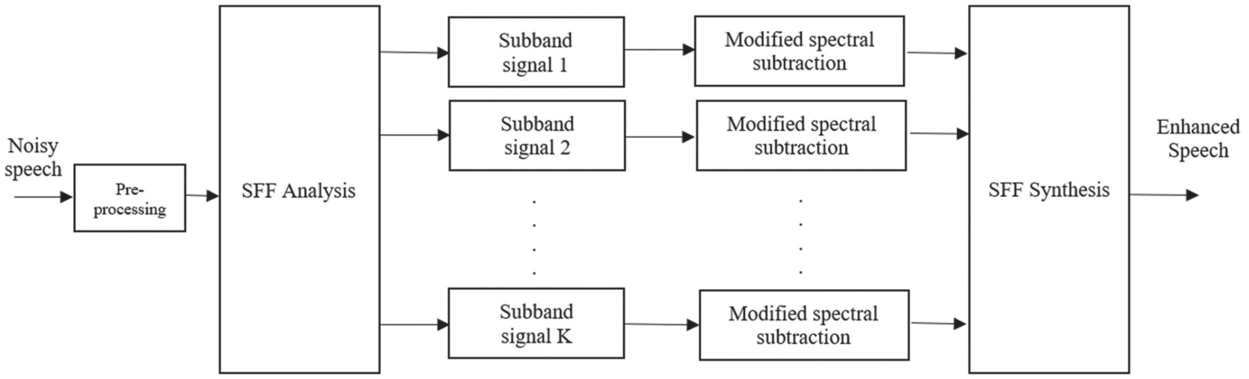


Fig. 1. Block diagram of the proposed speech enhancement algorithm

2.1. SFF ANALYSIS-SYNTHESIS APPROACH

The SFF approach is employed to achieve subband decomposition of the speech signal, allowing us to obtain the amplitude of the envelopes at a selected frequency for each instant. This technique helps to eliminate the block processing effect which can reduce the performance of the algorithms when using the Short-Time Fourier Transform (STFT) for speech enhancement. The analysis and synthesis blocks of the SFF technique are represented in Fig. 2. The amplitude envelope of the signal at any desired frequency is obtained by:

- shifting the frequency signal of the signal x[n] at frequency f_k using Eq. (1) for n=1.2...N and k=0.1…. K-1, where N and K represent the number of samples and the number of frequencies, respectively.

$$\tilde{x}[k,n] = x[n]e^{-j\frac{2\pi\bar{f}_k}{f_s}n} \tag{1}$$

- The shifted signal is then passed through a single pole resonator at the highest frequency $f_s/2$, where f_s is the sampling frequency. The filter function transfer is given by Eq. (2) and the filtered output is given by Eq. (3) where $y[k, n]$ is a complex number that can be expressed in polar form as given by Eq. (4).

$$H[z] = \frac{1}{1+rz^{-1}} \tag{2}$$

$$y[k,n] = -ry[k,n-1] + \tilde{x}[k,n] \tag{3}$$

$$y[k,n] = \ v[k,n]e^{j\phi[k,n]} \tag{4}$$

where the amplitude envelope $v[k, n]$ and the phase $\phi[k, n]$ of the subband signal $y[k, n]$ are defined by Eq. (5) and (6), respectively.

$$v[k,n] = \sqrt{y_r^2[k,n] + y_i^2[k,n]} \tag{5}$$

$$\phi[k,n] = tan^{-1}\left(\frac{y_i[k,n]}{y_r[k,n]}\right) \tag{6}$$

- The subband signal $y[k, n]$ can be reconstructed from the amplitude envelope $v[k, n]$ and phase $\phi[k, n]$ by applying Eq. (4) for the single frequency filtering synthesis. The shifted output $y[k, n]$ is shifted back to the original frequency using Eq. (7):

$$z[k,n] = y[k,n]e^{j\frac{2\pi\bar{f}_k}{f_s}n} \tag{7}$$

- The reconstructed signal is obtained by summing the outputs $z[k, n]$ and dividing by the number of frequencies K using Eq. (8).

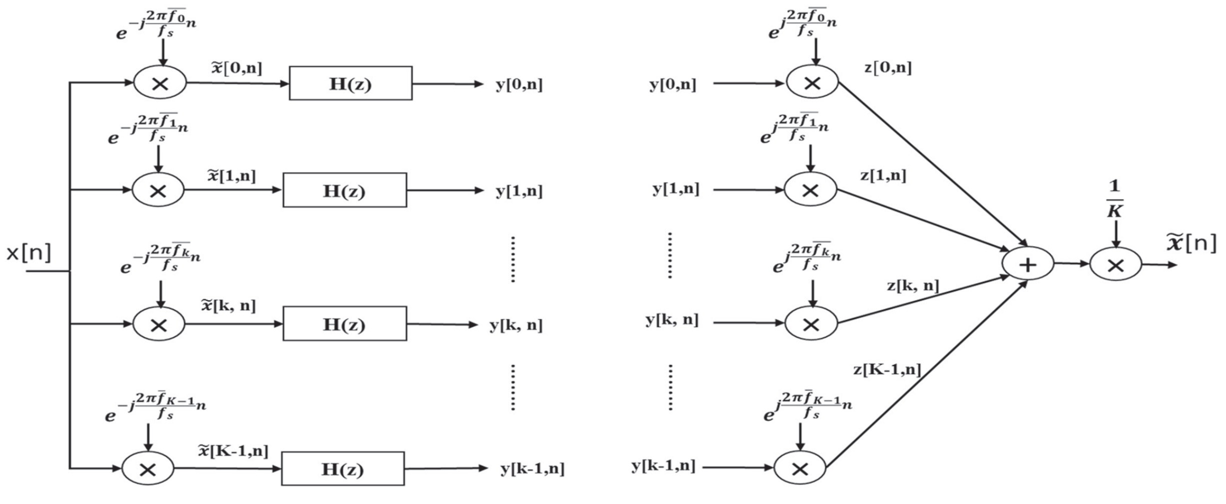
$$\hat{x}[k,n] = \frac{1}{K}\Re\left\{\sum_{k=0}^{K-1} z[k,n]\right\} \tag{8}$$

- The reconstructed signal and the original signal are combined using the following equation [26]:

$$\hat{x}[k,n] = \sum_{a=0}^{\infty}(r)^{aK}x[n-aK] \tag{9}$$

Where $K=(f_s/2)/\Delta f$, and r is less than 1 to ensure the stability of the filter.

The performance of the proposed algorithm is influenced by the two major parameters r and K. The primary objective of this research is to enhance degraded speech. To achieve this goal, a value of r=0.99 was selected, which offers higher temporal and spectral resolution. This resolution property enables a more precise and accurate analysis of the degraded speech, facilitating effective application of enhancement techniques [26]. Additionally, the impact of the number of frequencies K on speech quality and intelligibility was evaluated in this study. Clean speech signals were analyzed and synthesized at different K values while maintaining a sampling rate of 16 kHz. The resulting speech was then assessed using PESQ and ESTOI metrics. The findings presented in Table 1 show that reducing the frequency step size significantly improves speech quality and intelligibility. However, when using a frequency interval (Δf) of 50 Hz, we obtain 160 envelopes, whereas using a smaller Δf of 20 Hz results in 400 envelopes. Therefore, reducing Δf to 20 Hz can lead to improved quality and intelligibility of the processed speech. However, this reduction also increases the value of K and the number of envelopes, resulting in longer computation times. It is worth noting that utilizing the SFF method with Δf=20 Hz allows for enhancing speech without introducing any distortion.



$e^{-j\frac{2\pi \bar{f}_0}{f_s}n}$
$\tilde{x}[0,n]$
H(z)
y[0,n]
$e^{-j\frac{2\pi \bar{f}_1}{f_s}n}$
$\tilde{x}[1,n]$
H(z)
y[1,n]
x[n]
$e^{-j\frac{2\pi \bar{f}_k}{f_s}n}$
$\tilde{x}[k, n]$
H(z)
y[k, n]
$e^{-j\frac{2\pi \bar{f}_{K-1}}{f_s}n}$
$\tilde{x}[K-1,n]$
H(z)
y[k-1,n]
$e^{j\frac{2\pi \bar{f}_0}{f_s}n}$
y[0,n]
z[0,n]
$e^{j\frac{2\pi \bar{f}_1}{f_s}n}$
y[1,n]
z[1,n]
$e^{j\frac{2\pi \bar{f}_k}{f_s}n}$
y[k, n]
z[k, n]
z[K-1,n]
$e^{j\frac{2\pi \bar{f}_{K-1}}{f_s}n}$
y[k-1,n]
+
$\frac{1}{K}$
$\tilde{x}[n]$

(a) (b)

Fig. 2. (a) SFF analysis, and (b) SFF synthesis

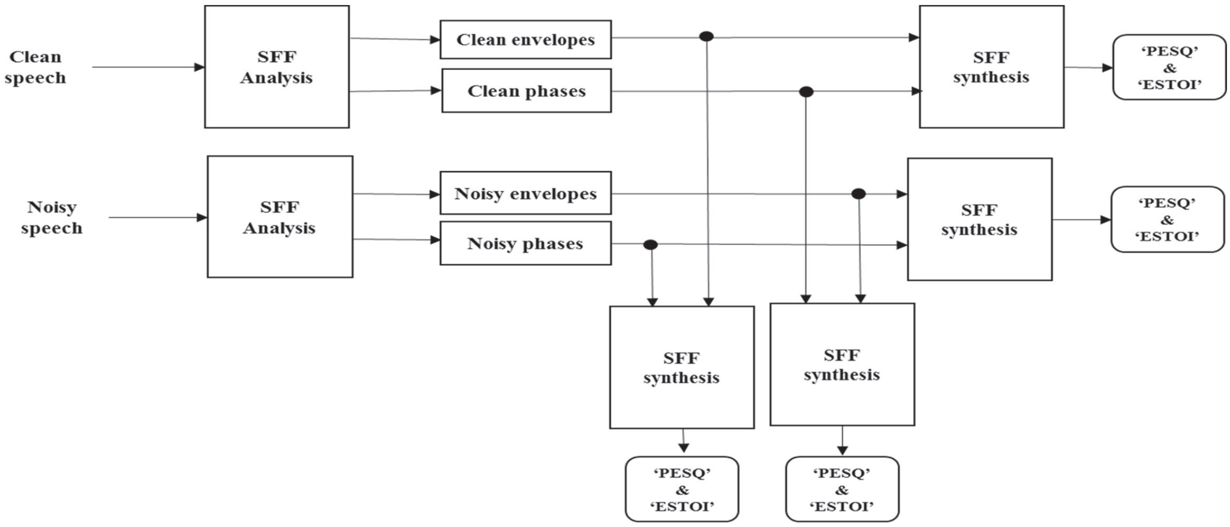


Fig. 3. Concept block diagram for the combinations of the clean and noisy temporal envelope with noisy and clean phase

Table 1. ESTOI and PESQ between original and SFF synthesized signals for different values of *Δf*.

Δf	100	50	20	10	5
ESTOI	0.987	0.999	1	1	1
PESQ	4.17	4.359	4.64	4.64	4.64

The following step is to determine whether the SFF amplitude envelope or the SFF phase should be enhanced to improve speech quality and intelligibility.

2.2. THE TEMPORAL ENVELOPE

The significance of the temporal envelope and its application in speech enhancement has been extensively explored in various studies [27-29]. The objective of this step is to verify and validate the hypothesis that "enhancing the temporal envelope of noisy speech can considerably enhance speech quality". To achieve this, we combined the clean envelope with the noisy phase and vice versa using samples from the TIMIT database, at 10 dB for white noise, with the envelope and phase calculated using the SFF technique. Fig.3 presents the adopted scheme for constructing all the combinations, while Table 2 presents the corresponding values of ESTOI and PESQ. The results indicate that the envelope plays a significant role in improving the quality and intelligibility of speech. Therefore, we recommend modifying the amplitude of the SFF envelope to enhance speech in noisy environments.

Table 2. The average PESQ and ESTOI values were computed for all combinations of TIMIT speech degraded by 10 dB white noise with *Δf*=20 Hz.

Combinations	ESTOI	PESQ
Clean envelopes - Clean phases	1	4,64
Clean envelopes - Noisy phases	0.91	1.89
Noisy envelopes - Clean phases	0.86	1.76
Noisy envelopes - Noisy phases	0.85	1.25

2.3. MODIFIED ENVELOPE SPECTRAL SUBTRACTION

Based on the results obtained from PESQ and ESTOI, we have evaluated the effectiveness of SFF analysis-synthesis and its associated envelope in enhancing speech quality and intelligibility. The findings indicate also that enhancing the temporal envelope, which is calculated using the SFF approach, can be utilized for speech enhancement, as the phase component has a negligible impact on human intelligibility. Considering its favorable performance across different noise conditions and its low computational complexity, we propose implementing the spectral subtraction method for each frame of each envelope to improve both quality and intelligibility of the speech signal. The noise is estimated using an adaptive technique, and the over-reduction factor is adjusted for each frame of each envelope Eq. (5). The equation provided below illustrates the spectrum of the enhanced k^{th} envelope.

$$|\hat{s}_k(m,n)|^2 = \begin{cases} |v_k(m,n)|^2 - \eta(m,n)\left|\hat{d}_k(m,n)\right|^2, if & \left[\frac{\hat{d}_k(m,n)}{v_k(m,n)}\right]^2 < \frac{1}{\eta(m,n)} \\ \beta|v_k(m,n)|^2 & , else \end{cases} \quad (10)$$

Where v_k (m, n), $\hat{d}_k$ (m, n), and $\hat{s}_k$ (m, n) represent the noisy envelope, estimated noise, and estimated enhanced envelope, respectively, for the n^{th} FFT transform of the m^{th} frame in the k^{th} envelope. The parameter β is the spectral floor factor that typically ranges between 0 and 1, and it is used to prevent the estimated negative speech spectrum in each envelope. To determine the over-reduction factor $\eta(m, n)$ for the envelope k at frame m, we use the segmental signal-to-noise ratio (segSNR) as follows:

$$\eta(m,n) = \begin{cases} 5 & , segSNR(m,n) < -5dB \\ \eta_0 - \frac{3}{20} segSNR(m,n), & -5dB \leq segSNR(m,n) \leq 20dB \\ 1 & , segSNR(m,n) > 20dB \end{cases} \quad (11)$$

The segSNR is calculated using the formula:

$$segSNR(m,n) = \frac{\sum_{m=0}^{N_k-1}|v_k(m,n)|^2}{\sum_{m=0}^{N_k-1}\left|\hat{d}_k(m,n)\right|^2} \quad (12)$$

where N_k represents the number of frames of the k^{th} envelope. The over-reduction parameter $\eta(m, n)$ is determined based on η_0 which represents the value of *segSNR* at 0 dB and controls the noise subtraction level for each envelope at every frame.

Accurate noise estimation is crucial for improving speech, as an incorrect estimate can result in residual noise or distorted speech. Traditional methods use voice activity detection (VAD) to estimate and adapt the noise spectra during speech silent periods, but this approach is not effective in real-time and noisy environments [30, 31]. A common technique for estimating noise is recursive averaging, where the noise spectrum is calculated by taking a weighted average of previous noise estimates and the present noisy envelope spectrum, as described in [32, 33]. The weight assigned to each estimate varies based on the a-posteriori signal-to-noise ratio (SNR) of each frequency. In our proposed technique, we independently estimate and update the noise spectrum for each frame in each envelope. Therefore, the noise spectrum estimation for each envelope is given by,

$$\left|\hat{d}_k(m,n)\right|^2 = \delta(m,n)\left|\hat{d}_k(m-1,n)\right|^2 + \{1-\delta(m,n)\}\,|v_k(m,n)|^2 \quad (13)$$

where δ is the smoothing parameter. In the recursive averaging technique, δ is chosen as a sigmoid function that depends on the a-posteriori SNR:

$$\delta(m,n) = \frac{1}{1+e^{[-a\{SNR(m,n)-T\}]}} \quad (14)$$

In this case, the variation of the noise is determined by the parameter "a" in the sigmoid function Eq. (14). The value of "a" ranges from 1 to 6 while keeping the parameter "T" constant. The parameter "T" in Eq. (14) represents the center offset of the transition curve in the sigmoid function, typically falling within the range of 3 to 5.

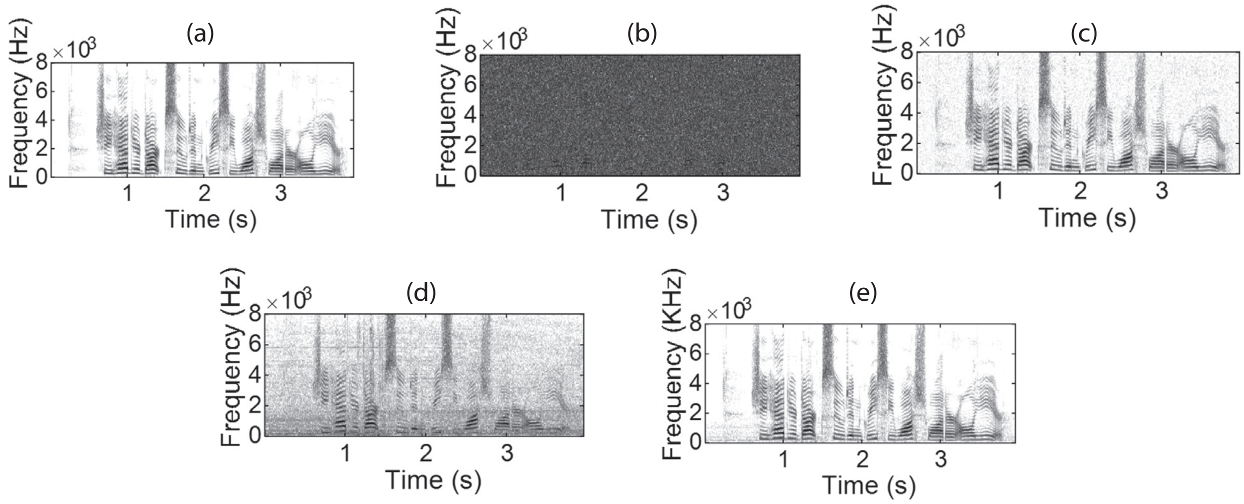


Fig. 4. Speech spectrograms for: (**a**) clean speech; (**b-d**) speech degraded by white and car noises, respectively, at 5 dB SNR; (**c-e**) the associated enhanced speech

3. EXPERIMENTAL RESULTS

In this section, we will present the results and performance evaluation of our proposed speech enhancement algorithm. Additionally, we will compare the results obtained using MATLAB software with other existing methods [17-20]. The evaluation was conducted on various types of noise, including real-world noises commonly encountered in daily life such as car, train, restaurant, airport, and street noises, as well as machine-generated white Gaussian noise. It is important to note that each type of noise exhibits a unique time-frequency distribution in speech.

To evaluate the performance of our speech enhancement algorithm, the noisy speech was obtained from the TIMIT corpus and downsampled to 8 kHz. The envelopes were calculated using the SFF with an envelope spacing of Δf=20 Hz, which provided good performance of speech quality and intelligibility as shown in Table 1. For each envelope, we applied a hamming window with a frame duration of 20 ms and a 70% overlap. The noise was estimated continuously and adaptively using Eq. (13) and the sigmoid function Eq. (14), with values of 'a' and 'T' set to 4 and 5, respectively. The over-subtraction factor $\eta(m, n)$ was computed for each envelope. Moreover, we fixed the spectral floor parameter β at a value of 0.03 [17, 18].

3.1. CORPUS

The TIMIT database, developed by the Massachusetts Institute of Technology with support from the US government, is a valuable resource for automatic speech recognition and other speech processing applications [34]. It comprises a vast collection of speech sounds and associated data. This extensive database includes recordings from 630 speakers representing different regions of the United States, each uttering 10 different phrases. It also provides transcriptions and annotations corresponding to the recorded speech. TIMIT has become a fundamental tool in the study of speech recognition and other related technologies, contributing significantly to the advancement of the field of speech processing.

3.2. PERFORMANCE EVALUATION

To assess the effectiveness of our proposed speech enhancement algorithm, we utilized objective quality and intelligibility measurement tests, including segmental signal-to-noise ratio (segSNR), extended short-term objective intelligibility (ESTOI) [35], and perceptual evaluation of speech quality (PESQ) [36]. segSNR is frequently used to detect speech distortion, as it is more precise in identifying speech distortion compared to overall SNR. Higher values for segSNR indicate lower levels of speech distortion. ESTOI is a measure of speech intelligibility that considers the accuracy and timing of phoneme recognition. PESQ is a commonly used objective measure of speech quality that compares and predicts the perceived quality of speech signals using a reference signal. Higher scores for both PESQ and ESTOI typically suggest better speech quality and intelligibility. These measures have strong correlations with subjective listening tests, making them valuable tools for assessing speech enhancement algorithms.

3.3. RESULTS AND DISCUSSIONS

We evaluated the performance of our proposed method (PM) on degraded speech corrupted by different types of noise such as white Gaussian, car, restaurant, train, street, and airport noises. We tested the PM at various SNR levels including -5, 0, 5, and 10 dB. To assess the effectiveness of the PM, we measured three parameters: segSNR, ESTOI, and PESQ. The average segSNR values obtained by the PM for different types of noise and SNR levels are presented in Table 3. For the -5 dB, 0 dB, 5 dB, and 10 dB SNR levels, the segSNR values were 5.30, 8.76, 10.07, and 13.64, respectively. It is worth noting that these segSNR values are consistently positive and exhibit an increasing trend as the SNR improves. These findings demonstrate that the proposed method performs well in terms of distortion across various noise types. In terms of intelligibility, the PM achieved average ESTOI values of 0.66, 0.76, 0.86, and 0.93 for -5 dB, 0 dB, 5 dB, and 10 dB SNR levels, respectively (Table 4). Specifically, ESTOI ranges from 0.57 to 0.74 for negative SNR values and from 0.65 to 0.97 for positive SNR values. These results provide that the PM algorithm significantly enhances the intelligibility of speech. To evaluate the speech quality, we calculated the average PESQ values, which ranged from 2.18 to 4.45 across all SNR levels (Table 5). These results confirm that our PM performs well in terms of speech quality. Fig. 4, 5, and 6 illustrate the speech spectrograms of the noisy speech at 5 dB SNR alongside the enhanced speech obtained using the PM. These figures provide a visual representation of the improvements achieved by our method.

The results of the score comparison in terms of segSNR, ESTOI, and PESQ for various types of noise at different levels of SNR, between our PM and MBSS [17], MBSS_CBRS [18], SS_HP [19], and SS_CHD [20], are presented in Tables 3, 4, and 5, respectively. The findings clearly demonstrate that our PM outperforms the other methods in terms of segSNR, indicating its effectiveness in removing background noise while preserving speech components, regardless of whether the SNR is negative or positive. The PM achieves also higher ESTOI scores, indicating improved speech intelligibility across all SNR levels. Furthermore, the PM obtains the highest PESQ score in all conditions, indicating the preservation of speech quality.

Based on these results, it can be concluded that our PM significantly enhances speech quality and intelligibility with minimal distortion compared to the methods defined previously. The effectiveness of the PM is supported by the utilization of the SFF to calculate the temporal envelopes with high-frequency resolution at

20 Hz, guided by PESQ and ESTOI scores. Moreover, our proposed method estimates the noise recursively from previous speech frames for each envelope and applies a smoothing parameter to achieve a balance between noise reduction and preservation of speech quality.

Furthermore, our PM algorithm was compared to recent algorithms that utilize deep learning techniques for tasks such as estimating parameters (e.g., tuning factor of the Wiener filter [37]) or extracting features (e.g., multi-frequency cepstral coefficients [38]).The comparative analysis revealed that our PM algorithm outperforms these approaches in terms of both speech quality and intelligibility, as depicted in Fig. 7 and 8. Moreover, an added advantage of our PM algorithm is that it does not necessitate training data. This characteristic reduces its complexity and simplifies its implementation, making it a more practical and accessible solution for noise reduction and speech enhancement tasks.

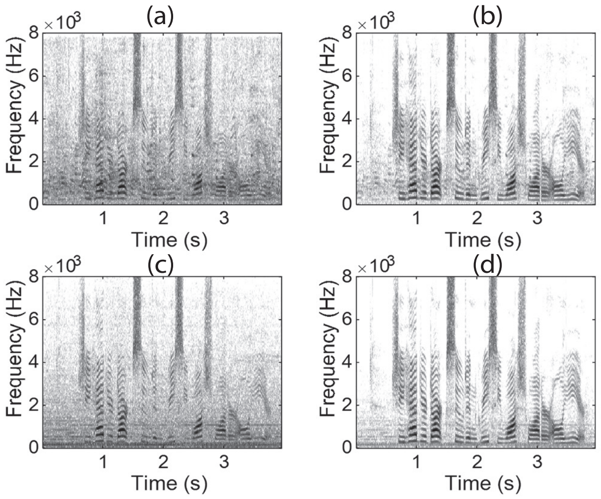


Fig. 5. Speech spectrograms for: (**a-c**) speech degraded by restaurant and train noises, respectively, at 5 dB SNR; (**b-d**) the associated enhanced speech

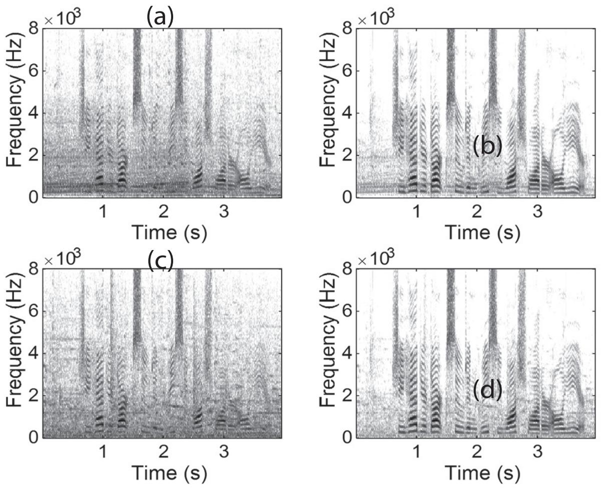


Fig. 6. Speech spectrograms for: (**a-c**) speech degraded by street and airport noises, respectively, at 5 dB SNR; (**b-d**) the associated enhanced speech

Table 3. Average segmental signal-to-noise ratio ($segSNR$) of enhanced speech signals from the TIMIT database at -5,0, 5, 10 dB

Noise type	Enhancement methods	segSNR			
		-5	0	5	10
White	Noisy	-7.59	-5.16	-1.92	1.72
	MBSS	1.64	4.07	10	13.40
	MBSS_CBRS	2.83	5.9	8.63	11.77
	SS_HP	4.57	7.66	5.97	4.31
	SE_CHD	0.90	1.09	1.59	1.81
	PM	**6.58**	**9.67**	**12.01**	**15.41**
Car	Noisy	-7.50	-5.02	-1.78	1.80
	MBSS	0.64	2.68	6.86	9.39
	MBSS_CBRS	2.72	4.23	9.05	12.05
	SS_HP	3.90	6.45	4.86	3.73
	SE_CHD	0.26	0.58	1.01	1.90
	PM	**5.91**	**8.46**	**11.06**	**14.06**
Restaurant	Noisy	-7.09	-4.63	-1.43	2.2
	MBSS	0.43	2.52	5.84	9.29
	MBSS_CBRS	1.79	2.14	7.25	9.96
	SS_HP	3.69	6.17	4.86	3.48
	SE_CHD	0.37	0.60	0.83	2.29
	PM	**5.70**	**8.18**	**9.26**	**11.97**
Train	Noisy	-7.46	-5.06	-1.77	1.85
	MBSS	0.21	3.88	5.53	9.35
	MBSS_CBRS	2.38	5.12	7.6	11.59
	SS_HP	1.69	3.92	2.96	2.16
	SE_CHD	0.28	0.41	0.53	2.10
	PM	**4.39**	**7.13**	**9.61**	**13.6**
Street	Noisy	-6.51	-3.9	-0.73	2.86
	MBSS	0.43	1.71	5.39	9.22
	MBSS_CBRS	2.27	9.81	7.02	11.36
	SS_HP	1.98	6.98	4.65	3.29
	SE_CHD	0.19	0.31	0.43	2.9
	PM	**4.28**	**11.82**	**9.03**	**13.37**
Airport	Noisy	-7.45	-5.02	-1.81	1.8
	MBSS	1.49	3.78	6.53	8.81
	MBSS_CBRS	2.91	5.26	7.43	11.44
	SS_HP	2.08	3.75	3.04	2.64
	SE_CHD	0.31	0.64	0.8	1.8
	PM	**4.92**	**7.27**	**9.44**	**13.45**

Table 4. Average extended short-term objective intelligibility (ESTOI) results of enhanced speech signals from the TIMIT database at -5,0, 5, 10 dB

Noise type	Enhancement methods	ESTOI			
		-5	0	5	10
White	Noisy	0.25	0.39	0.55	0.68
	MBSS	0.39	0.45	0.6	0.68
	MBSS_CBRS	0.4	0.58	0.69	0.71
	SS_HP	0.49	0.6	0.7	0.75
	SE_CHD	0.18	0.15	0.21	0.24
	PM	**0.57**	**0.65**	**0.79**	**0.8**
Car	Noisy	0.2	0.39	0.52	0.64
	MBSS	0.39	0.48	0.57	0.71
	MBSS_CBRS	0.43	0.51	0.62	0.78
	SS_HP	0.59	0.63	0.79	0.82
	SE_CHD	0.18	0.12	0.15	0.21
	PM	**0.69**	**0.73**	**0.89**	**0.92**
Restaurant	Noisy	0.21	0.37	0.53	0.67
	MBSS	0.38	0.53	0.69	0.71
	MBSS_CBRS	0.43	0.67	0.72	0.77
	SS_HP	0.59	0.74	0.78	0.86
	SE_CHD	0.11	0.12	0.15	0.23
	PM	**0.69**	**0.84**	**0.88**	**0.96**
Train	Noisy	0.34	0.46	0.57	0.66
	MBSS	0.49	0.53	0.61	0.78
	MBSS_CBRS	0.56	0.61	0.69	0.83
	SS_HP	0.66	0.71	0.76	0.89
	SE_CHD	0.13	0.15	0.17	0.28
	PM	**0.74**	**0.79**	**0.84**	**0.97**
Street	Noisy	0.27	0.4	0.54	0.66
	MBSS	0.35	0.49	0.67	0.75
	MBSS_CBRS	0.41	0.54	0.73	0.79
	SS_HP	0.52	0.69	0.79	0.86
	SE_CHD	0.14	0.15	0.16	0.27
	PM	**0.6**	**0.77**	**0.87**	**0.94**
Airport	Noisy	0.24	0.38	0.52	0.64
	MBSS	0.36	0.47	0.61	0.73
	MBSS_CBRS	0.41	0.59	0.75	0.81
	SS_HP	0.59	0.68	0.8	0.89
	SE_CHD	0.18	0.2	0.24	0.28
	PM	**0.67**	**0.76**	**0.88**	**0.97**

Table 5. Average perceptual evaluation of speech quality (PESQ) results of enhanced speech signals from the TIMIT database at -5,0, 5, 10 dB

Noise type	Enhancement methods	PESQ			
		-5	0	5	10
White	Noisy	1.11	1.21	1.39	1.67
	MBSS	1.31	1.53	1.82	2.22
	MBSS_CBRS	1.49	1.65	1.97	2.3
	SS_HP	1.87	2.19	2.58	2.95
	SE_CHD	0.45	0.49	0.59	0.68
	PM	**2.18**	**3.69**	**4.08**	**4.45**
Car	Noisy	1.1	1.18	1.32	1.55
	MBSS	1.31	1.49	1.98	2.25
	MBSS_CBRS	1.29	1.43	1.62	2.43
	SS_HP	2.19	2.33	2.72	3.1
	SE_CHD	0.47	0.59	0.7	0.7
	PM	**2.35**	**3.34**	**4.73**	**4.11**
Restaurant	Noisy	1.12	1.19	1.3	1.5
	MBSS	1.61	1.84	2.06	2.32
	MBSS_CBRS	1.47	1.63	1.95	2.32
	SS_HP	2.34	2.25	2.64	3.21
	SE_CHD	0.56	0.67	0.69	0.7
	PM	**2.94**	**3.85**	**4.02**	**4.18**
Train	Noisy	1.15	1.27	1.47	1.77
	MBSS	1.31	1.51	1.69	2.14
	MBSS_CBRS	1.4	1.52	1.61	2.14
	SS_HP	1.98	2.05	2.43	2.82
	SE_CHD	0.51	0.5	0.65	0.42
	PM	**2.58**	**3.65**	**4.03**	**4.42**
Street	Noisy	1.17	1.27	1.45	1.72
	MBSS	1.321	1.59	1.93	2.24
	MBSS_CBRS	1.227	1.4	2.01	2.3
	SS_HP	2.314	2.56	2.46	3.62
	SE_CHD	0.47	0.48	0.52	0.61
	PM	**3.09**	**3.76**	**4.06**	**4.22**
Airport	Noisy	1.12	1.2	1.34	1.57
	MBSS	1.31	1.79	2.1	2.42
	MBSS_CBRS	1.21	1.46	2.11	2.42
	SS_HP	2.09	2.34	2.83	3.68
	SE_CHD	0.4	0.35	0.5	0.63
	PM	**2.89**	**3.94**	**3.43**	**4.28**

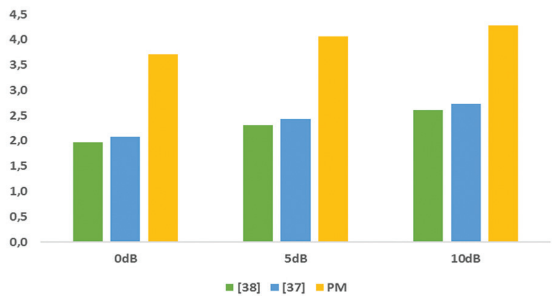


Fig. 7. Average performance comparison between the proposed method and methods used in [37] and [38] for all noises using PESQ.

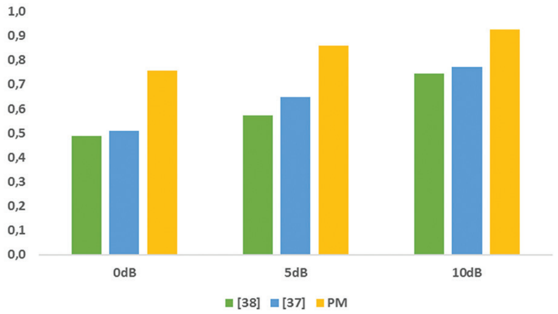


Fig. 8. Average performance comparison between the proposed method and methods used in [37] and [38] for all noises using ESTOI.

4. CONCLUSION

The proposed method aims to enhance the quality and intelligibility of speech degraded by noises. It utilizes the single-frequency filtering approach and modified spectral subtraction to effectively eliminate unwanted noise while minimizing distortion and preserving essential speech characteristics. The research demonstrates the effectiveness of this approach in improving speech quality and intelligibility under various noise types and different signal-to-noise ratio (SNR) levels. The performance of our algorithm is encouraging, and it can be suitable to meet the requirements and challenges of complex environments, by adjusting only the over-subtraction factor. It is important to note that our proposed method has a limitation related to the over-subtraction process. This aspect highlights an area for improvement in our approach. In our future work, we will primarily concentrate on addressing this limitation by incorporating voice characteristics into the algorithm. By considering the specific features of the speech segments (voiced or unvoiced), we aim to enhance the noise reduction's accuracy and effectiveness, thereby improving our method's overall performance.

5. REFERENCES:

[1] R. C. Hendriks, T. Gerkmann, J. Jensen, "DFt-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art", Synthesis Lectures on Speech and Audio Processing, Vol. 11, Springer, 2013.

[2] K. K. Wójcicki, P. C. Loizou, "Channel selection in the modulation domain for improved speech intelligibility in noise", The Journal of the Acoustical Society of America, Vol. 131, No. 4, 2012, pp. 2904-2913.

[3] M. I. Khattak, N. Saleem, J. Gao, E. Verdu, J. P. Fuente, "Regularized sparse features for noisy speech enhancement using deep neural networks", Computers and Electrical Engineering, Vol. 100, 2022.

[4] E. Mabande, F. Kuech, A. Niederleitner, A. Lombard, "Towards robust close-talking microphone arrays for noise reduction in mobile phones", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, 20-25 March 2016.

[5] D. Li, Y. Gao, C. Zhu, Q. Wang, R. Wang, "Improving Speech Recognition Performance in Noisy Environments by Enhancing Lip Reading Accuracy", Sensors, Vol. 23, No. 4, 2023.

[6] R. Bendoumia, M. T. Betina, A. Oulahcene, A. Guessoum, "Extended subband decorrelation version of feedback normalized adaptive filtering algorithm for acoustic noise reduction", Applied Acoustics, Vol. 179, 2021.

[7] S. H. Han, S. Jeong, H. Yang, J. Kim, W. Ryu, M. Hahn, "Noise reduction for VoIP speech codecs using modified wiener filter", Advances and Innovations in Systems, Computing Sciences and Software Engineering, Springer, 2007, pp. 393-397.

[8] D. O'Shaughnessy, "Speech communications: Human and machine", Second Edition, 1999.

[9] J. S. Lim, A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech", Proceedings of IEEE, Vol. 67, No. 12, 1979, pp. 1586-1604.

[10] P. C. Loizou, "Speech Enhancement: Theory and Practice", CRC Press, 2013.

[11] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 27, No. 2, 1979, pp. 113-120.

[12] M. M. Sondhi, C. E. Schmidt, L. R. Rabiner, "Improving the Quality of a Noisy Speech Signal", The Bell System Technical Journal, Vol. 60, No. 8, 1981, pp. 1847-1859.

[13] J. H. L. Hansen, M. A. Clements, "Use of objective speech quality measures in selecting effective spectral estimation techniques for speech enhancement", Proceedings of the Midwest Symposium on Circuits and Systems, Champaign, IL, USA, 14-16 August 1989.

[14] H. Xu, Z. H. Tan, P. Dalsgaard, B. Lindberg, "Spectral subtraction with full-wave rectification and likelihood controlled instantaneous noise estimation for Robust speech recognition", Proceedings of the 8th International Conference on Spoken Language Processing, 2004.

[15] M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of speech corrupted by acoustic noise", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Washington, DC, USA, 2-4 April 1979.

[16] N. Abajaddi, B. Mounir, L. Elmaazouzi, I. Mounir, A. Farchi, "Speech Spectral Subtraction in Modulation Domain", Lecture Notes in Networks and Systems, Springer, 2022.

[17] S. Kamath, P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, USA, 13-17 May 2002.

[18] N. Upadhyay, A. Karmakar, "Single-Channel Speech Enhancement Using Critical-Band Rate Scale Based Improved Multi-Band Spectral Subtraction", Journal of Signal and Information Processing, Vol. 04, No. 03, 2013.

[19] C. T. Lu, K. F. Tseng, Y. Y. Chen, L. L. Wang, C. L. Lei, "Speech enhancement using spectral subtraction algorithm with over-subtraction and reservation factors adapted by harmonic properties", Proceedings of the International Conference on Applied System Innovation, Okinawa, Japan, 26-30 May 2016.

[20] S. Samui, I. Chakrabarti, S. K. Ghosh, "Speech enhancement based on modulation domain processing using coherent harmonic demodulation technique", Electronics Letters, Vol. 53, No. 24, 2017.

[21] G. Aneeja, B. Yegnanarayana, "Single Frequency Filtering Approach for Discriminating Speech and Nonspeech", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 23, No. 4, 2015, pp. 705–717.

[22] S. R. Kadiri, B. Yegnanarayana, "Epoch extraction from emotional speech using single frequency filtering approach", Speech Communication, Vol. 86, 2017, pp. 52-63.

[23] M. H. Javid, K. Gurugubelli, A. K. Vuppala, "Single frequency filter bank based long-term average spectra for hypernasality detection and assessment in cleft lip and palate speech", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4-8 May 2020.

[24] K. Gurugubelli, A. K. Vuppala, "Perceptually Enhanced Single Frequency Filtering for Dysarthric Speech Detection and Intelligibility Assessment", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12-17 May 2019.

[25] K. Gurugubelli, A. K. Vuppala, "Analytic phase features for dysarthric speech detection and intelligibility assessment", Speech Communication, Vol. 121, 2020. pp. 1-15.

[26] N. Chennupati, S. R. Kadiri, B. Yegnanarayana, "Spectral and temporal manipulations of SFF envelopes for enhancement of speech intelligibility in noise", Computer Speech & Language, Vol. 54, 2019, pp. 86-105.

[27] A. Wiinberg, J. Zaar, T. Dau, "Effects of Expanding Envelope Fluctuations on Consonant Perception in Hearing-Impaired Listeners", Trends in Hearing, Vol. 22, 2018.

[28] T. Langhans, H. W. Strube, "Speech enhancement by nonlinear multiband envelope filtering", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Paris, France, 3-5 May 1982.

[29] M. Koutsogiannaki, H. Francois, K. Choo, E. Oh, "Real-time modulation enhancement of temporal envelopes for increasing speech intelligibility", Proceedings of the Annual Conference of the International Speech Communication Association, 2017.

[30] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics", IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 5, 2001, pp. 504-512.

[31] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging", IEEE Transactions on Speech and Audio Processing, Vol. 11, No. 5, 2003, pp. 466-475.

[32] L. Lin, W. H. Holmes, E. Ambikairajah, "Speech denoising using perceptual modification of Wiener filtering", Electronics Letters, Vol. 38, No. 23, 2002.

[33] L. Lin, W. H. Holmes, E. Ambikairajah, "Adaptive noise estimation algorithm for speech enhancement", Electronics Letters, Vol. 39, No. 9, 2003.

[34] J. Garofolo et al. "TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download", Philadelphia Linguistic Data Consortium, 1993.

[35] J. Jensen, C. H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 24, No. 11, 2016, pp. 2009-2022.

[36] ITU, "ITU-T Recommendation P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", 2000.

[37] A. Garg, "Speech enhancement using long short term memory with trained speech features and adaptive wiener filter", Multimedia Tools and Applications, Vol. 82, No. 3, 2023, pp. 3647-3675.

[38] A. Garg, O. P. Sahu, "Deep Convolutional Neural Network-based Speech Signal Enhancement Using Extensive Speech Features", International Journal of Computational Methods, Vol. 19, No. 8, 2022.

An Enhanced Spatio-Temporal Human Detected Keyframe Extraction

Original Scientific Paper

Rajeshwari D
Research Department of Computer Science
Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women,
Affiliated to University of Madras, Chennai, India.
rajeshwari.d@sdnbvc.edu.in

Victoria Priscilla C
PG Department of Computer Science,
Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women,
Affiliated to University of Madras, Chennai, India.
aprofvictoria@gmail.com

Abstract *– Due to the immense availability of Closed-Circuit Television surveillance, it is quite difficult for crime investigation due to its huge storage and complex background. Content-based video retrieval is an excellent method to identify the best Keyframes from these surveillance videos. As the crime surveillance reports numerous action scenes, the existing keyframe extraction is not exemplary. At this point, the Spatio-temporal Histogram of Oriented Gradients - Support Vector Machine feature method with the combination of Background Subtraction is appended over the recovered crime video to highlight the human presence in surveillance frames. Additionally, the Visual Geometry Group trains these frames for the classification report of human-detected frames. These detected frames are processed to extract the keyframe by manipulating an inter-frame difference with its threshold value to favor the requisite human-detected keyframes. Thus, the experimental results of HOG-SVM illustrate a compression ratio of 98.54%, which is preferable to the proposed work's compression ratio of 98.71%, which supports the criminal investigation.*

Keywords: *Histogram of Oriented Gradients-Support Vector Machine, Keyframe Extraction, Spatio-temporal feature Extraction, Content-Based Video Retrieval*

1. INTRODUCTION

The use of Closed-circuit television (CCTV) surveillance for specific safety measures has increased incrementally in the majority of public areas in recent years. Surveillance plays an essential part in crime scene investigation by actively monitoring the circumstances inside a specific, stationary region. In the field of investigation, many investigators still struggle to identify the victim in the cases. Here are some of the most common issues, such as (1). Videos of poor quality (2) Videos with low frame rates lose detail between frames. (3). Analyzing and evaluating larger datasets in videos requires a significant amount of time and effort by the investigators.

Content-Based Video Retrieval (CBVR) is widely regarded as a crucial step in video analysis and Key frame extraction. It retrieves the desired video from a massive video storage database. Keyframe Extraction is the process of extracting a significant segment of a video by exploring its content to generate a condensed and semantically rich summary.

Moreover, if these keyframes for crime investigation reports are highlighted with humans, it is easier to suspect those responsible for the crime. To efficiently quote the sequences, it is necessary to determine an algorithm for human-detected keyframes in particular.

The Histogram of Gradients-Support Vector Machine (HOG-SVM) approach can identify people in the surveillance footage, although it occasionally fails in certain frames. As a result, the suggested work uses HOG-SVM with background subtraction to report human detection in all pertinent frames. Additionally, a Visual Geometry Graph (VGG-16) pre-trained these frames for the categorization report of human-detected frames. Finally, the frames (images) are pre-processed with the Canny-Edge detection method for enhanced structural information, and the desired Keyframes are extracted using the inter-frame difference method with its threshold value. These Keyframes play a crucial role in the investigation of crimes by substantially reducing the temporal and spatial complexities of the process.

The documentation is systematically structured as outlined below: Section 2 provides a concise summary of the current study on CBVR with human motion recognition and keyframe extraction techniques. In Section 3, the recommended approach of employing the HOG-SVM technique along with background subtraction is discussed in detail. The details regarding the implementation and the experimental findings can be found in Section 4, while the conclusion is provided in Section 5.

2. RELATED WORK

This literature probes the study of detecting humans through various algorithms. Consigning humans to other existing objects is very complicated in CCTV surveillance. Also, a person prolongs a long or short stay in a place to represent a certain action [1].

In most instances, humans are identified by their motion. Currently, the frame subtraction method, the background subtraction method, and the optical flow method are the most frequently utilized techniques for motion detection.

Optical Flow: This method observes the moving object based on its maximal frame-to-frame deviation. Identifying human motion in a video stream using the optical flow method requires a great deal of computational time [2].

Background Subtraction: This method attempts to encapsulate information regarding background scene changes concerning the video frame sequence [3]. There are various methods for performing background subtraction. The most common approaches are (a) Adaptive Gaussian mixture, which uses motion analysis to distinguish the foreground from the complex background [4], (b) Kalman filter, which is used to enhance image quality through background elimination [5], (c) Temporal differencing, which uses pixels to calibrate the motion detection on the foreground [6], and (d) Clustering techniques, which look at groups of pixels that are similar [7]. This method merits high accuracy but demerits to have a static background. Frame Difference: The moving object is identified efficiently at a complex background by taking the difference between the two frames [8, 9] but reports with less accuracy.

The motion detection phase in the video can also be detected using combination approaches such as background subtraction with the optical flow. This reduces the noise effect and eliminates the shadow present in the frames [10, 11]. Another combination is background subtraction with the frame difference method. This combination's main advantage results in the fast elimination of shadows [12] and inexpensive detection of frames [13]. Thus, this combination supports speculating the appropriate motion detection phase to extract keyframes from the surveillance video.

Keyframe Extraction refers to the video's summary because it removes redundant frames and provides only the video's essential content. It is a probabilistic task to extract keyframes from video footage containing massive amounts of data [8]. Many scholars have classified keyframe extraction techniques using Shot boundary, Motion Analysis, Visually Segmented, and cluster-based analysis as depicted in Table 1

Table 1. Existing Methods & Techniques for Keyframe Extraction

METHODS & TECHNIQUES	ACCURACY & MERITS	DEMERITS
Shot-boundary Detection		
SIFT-point distribution Histogram [14]	94.36% accuracy with less computation	Selecting only the Salient segment from each segmented shot
Middle Range Binary Local Pattern (MRLBP) [15]	96.34% with high entropy measures	Only Abrupt shot boundary detection is performed
Adaptive Threshold [16]	91.93% with less computation	Less performance due to blurred frames.
SVD Pattern Matching [17]	85.5% with high detection speed	Less precision value for gradual detection
Hadamard Transform [18]	88.7% based on significant feature	Less accuracy level at gradual transition
Genetic Algorithm and fuzzy logic [19]	86.8% with increased iterations	Time Complexity is high when iteration increases
Multimodal techniques [20]	88.7% with the selection of candidate segment	Speed is not detected and gradual detection has to be improved
Motion-Analysis		
Discrete cosine coefficients and rough sets theory [21]	82% for visual representation	Enormous Space Complexity
Thresholding technique [22]	81% based on threshold value	Using Key-object to analyze with less precision
Color and Structure Based [23]	86% with high computation	Poor performance on complex transitions
Perceived Motion Energy Mode [24]	80% with motion and color based	Requires improvement in color variation
Convolutional Neural Network [25]	92% with improved frame difference method	High computation time
Visually Segmented		
Region of Interest-KNN, SVM [26]	90% of motion detection by pixel change	Concentrated more on noise reduction
Multiple Feature Analysis [27]	80% of motion detection by pixel-level classification	Performance at the static background
Region Of Interest-FCN with CNN [28]	97% of detecting multiple objects	Less Performance in crowded areas
Cluster-Based		
Weighted Multi-View Cluster [29]	81.53% for medoid frames	Fails to report the number of clusters
Dynamic Spatio-Temporal Slice Clustering [30]	92.68% with high accuracy	Proposed only on human action video dataset

The primary result of these methodologies clarifies the applicability of distinguishing objects and identifying events in keyframes with an appropriate level of complexity. This research proposes a faster, more accurate, and computationally efficient strategy for Video Keyframe Extraction.

3. PROPOSED METHODOLOGY

The proposed approach's general framework (Figure 1) consists of four sequential steps: (1) Pre-processing the video. (2) Human motion detection using Background Subtraction and Frame Difference method. (3) Spatio-temporal feature extraction - HOG-SVM. (4) VGG-16 pre-trained CNN and (5) the Keyframe extraction using Threshold value along with Canny Edge Detection Method.

3.1. VIDEO PRE-PROCESSING

The recorded CCTV surveillance footage is in the initial stage of pre-processing. In this phase, the video footage endures a conversion to gray scale and is also resized to 640*480 for faster detection.

3.2. BACKGROUND SUBTRACTION TECHNIQUE

Background subtraction (Equation 1) is widely used for motion (Human) detection in video surveillance of static cameras. The detection of motion is achieved by calculating the disparity between the present frame and the reference frame [32].

$$|Frame_i\text{-}Background_i|>Threshold \quad (1)$$

Whereas the frame difference method (Equation 2) calibrates the difference between the two frames by the pixel variation.

$$|Frame_{i-1}\text{-}Frame_i|>Threshold \quad (2)$$

The background subtraction and frame difference algorithms' discrete performance are subject to false detection. To overcome it, the combination of background subtraction and frame difference assists surveillance video to detect motion more accurately.

Converting the video to gray scale frames simplifies the background subtraction process and facilitates the detection of humans. Calibration is performed by capturing the non-moving pixel specks in the first frame. If the pixel has changed in the subsequent frame, motion is detected. Then, these frames are subjected to the frame difference method, which identifies differencing structures to eradicate redundant frames. Still, the researchers have some limitations, as mentioned in Table 2.

Table 2. Merits and Demerits of Combination Factors

METHODS	ADVANTAGES	DISADVANTAGES
Background subtraction with frame difference using Running Gaussian Average [13]	Shadows are more efficiently removed	Once motionless, the whole part is considered background
Background with frame difference [33]	Eliminate the noise efficiently	Represent video with static background
Background and consecutive frame difference method [34]	Efficient method for surveillance datasets	The Dynamic background is not supported
Background subtraction with frame difference [11]	Rectangular contour for moving objects with noise elimination	Too many detections of moving objects
Background subtraction and frame difference using correlation coefficient [35]	Highly correlated with background image for speed and detection accuracy	The Shape and edge on each frame have to be concentrated more.
Background subtraction using pixel intensity [36]	Deduction of the person by pixel change	The speed of the process is slightly slow

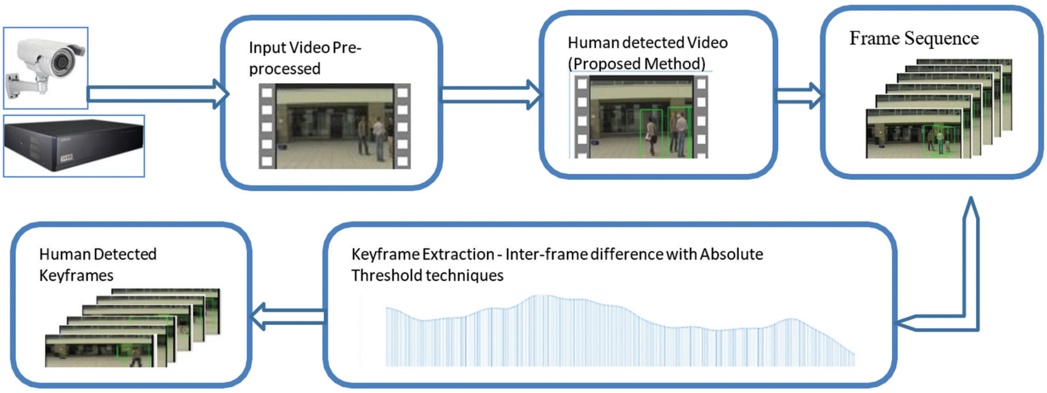

Input Video Pre-processed
Human detected Video (Proposed Method)
Frame Sequence
Human Detected Keyframes
Keyframe Extraction - Inter-frame difference with Absolute Threshold techniques

Fig. 1. The Overall Framework of Proposed Approach

3.3. SPATIO-TEMPORAL FEATURE EXTRACTION

3.3.1. Histogram Of Oriented Gradients

The Spatio-temporal feature extraction method supports human detection techniques using HOG, which was developed by Dalal and Trigg [31]. HOG represents the human shape and regional appearance based on the local histograms of image gradients in a dense grid. Here, the selected frame is partitioned into a small, connected area called cells. These cells contain several pixels, which unite to make a histogram of gradients. The computed gradients from the detector window are tiled like a grid of overlapped blocks, in which the HOG is extracted with normalized cells. The normalized cells give better accuracy on the variation through illumination and intensity.

3.3.2. Support Vector Machine

Support Vector Machine (SVM), is a supervised Machine Learning Algorithm that represents the most accurate image classification. Here, the resultant descriptors are fed into the linear SVM [37] for human/non-human classification.

3.4. VGG-16 CONVOLUTIONAL NEURAL NETWORK

In this proposal, the human-detected frames are trained by the VGG-16 pre-trained CNN (convolutional neural network) [38] model. Using a multi-class classification problem, the frames are categorized into three classes: 0 for False human predicted (FHP), 1 for True human predicted (THP), and 2 for Without human identification (WH) frames. All of these frames are resized so that the input image has dimensions of 224*224*3 and then sent to the input layer. The concealed layer is then convoluted three times with a dropout of 0.5, and the output layer is established using Softmax. By compiling the model with Adam optimizer, the accuracy reported for human-detected HOG-SVM in Table 3 and for the proposed work in Table 4 is significantly improved.

Table 3. VGG-16 trained HOG-SVM Human Detected Frames

Human detection using HOG-SVM						
Surveillance dataset	Detection	Total frames	Precision	Recall	F1-score	Accuracy
CCTV1	FHP	520	0.94	0.94	0.94	98.33
	THP		0.99	0.99	0.99	
	WH		1.00	1.00	1.00	
CCTV2	FHP	579	1.00	0.53	0.69	97.38
	THP		0.95	1.00	0.97	
	WH		0.00	0.00	0.00	
CCTV3	FHP	643	1.00	0.79	0.88	98.50
	THP		0.98	0.99	0.99	
	WH		0.98	1.00	0.99	
CCTV4	FHP	629	0.58	0.54	0.56	98.08
	THP		0.97	0.97	0.97	
	WH		0.00	0.00	0.00	
CCTV5	FHP	584	0.91	1.00	0.95	99.18
	THP		1.00	0.93	0.96	
	WH		0.99	1.00	1.00	

Table 4. VGG-16 trained Human Detected Frames for Proposed Work

Human detection using HOG-SVM						
Surveillance dataset	Detection	Total frames	Precision	Recall	F1-score	Accuracy
CCTV1	FHP	435	0.95	0.83	0.88	98.33
	THP		0.99	0.99	0.99	
	WH		0.73	1.00	0.84	
CCTV2	FHP	537	0.83	0.56	0.67	98.80
	THP		0.97	0.99	0.98	
	WH		0.00	0.00	0.00	
CCTV3	FHP	580	0.96	1.00	0.98	98.61
	THP		1.00	0.98	0.99	
	WH		1.00	1.00	1.00	
CCTV4	FHP	629	0.58	0.54	0.56	98.08
	THP		0.97	0.97	0.97	
	WH		0.00	0.00	0.00	
CCTV5	FHP	534	0.95	0.95	0.95	99.21
	THP		0.97	0.97	0.97	
	WH		1.00	1.00	1.00	

3.5. KEYFRAME EXTRACTION

After the preceding stages have been completed, the frames are fine-tuned using a Canny-edge detector to obtain a clear image. Now, the keyframe must be extracted from frames that differ significantly from one another. The average inter-frame difference greater than the threshold value, as calculated by equation (3), yields the keyframes.

$$|Average\ Inter\text{ - }frame\ difference|=0.6 \qquad (3)$$

Here, the proposed method is combined with the threshold range to identify the ideal human-detected keyframes, as shown in Fig. 2.

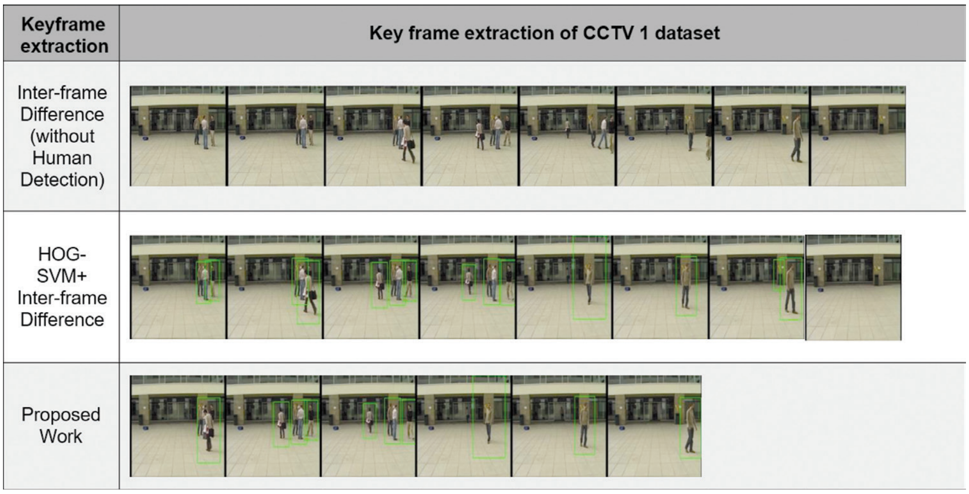

Keyframe extraction
Key frame extraction of CCTV 1 dataset
Inter-frame Difference (without Human Detection)
HOG-SVM+ Inter-frame Difference
Proposed Work

Fig. 2. Performance Analysis of Keyframe Extraction

For instance, the first CCTV1 surveillance system proposal included 435 video frames, from which 6 keyframes were extracted. The keyframes keyframe_99, keyframe_144, keyframe_178, keyframe_284, keyframe-336, and keyframe_375 are chosen based on their abrupt pixel change and difference from the overall frames. As depicted in Fig. 2, the performance evaluation of HOG-SVM with background subtraction reveals a reduction from 8 to 6 keyframes. The proposed task, in contrast, extracts only the required keyframes with perfect human detection. Consequently, the proposed result of HOG- SVM, along with background subtraction and inter-frame differences with the necessary threshold values, demonstrates the most accurate detection.

4. RESULTS AND DISCUSSION

The proposed task is carried out using Python Open CV image processing. The performance measurement derived from the obtained frames resulting in the keyframes is processed for the evaluation of metrics such as average frame per second (Equation 4) and frame per second (Equation 5).

4.1. AVERAGE FRAME PER SECOND

$$Avg\ FPS= (Total\ frames\ per\ second)/(Current\ frame) \quad (4)$$

The average frame per second is determined by comparing the total frames per second to the current frame's frame rate. The frames per second are the unit of measurement for the video's performance.

4.2. FRAMES PER SECOND

$$FPS=1/((end\ time\text{-}start\ time)) \quad (5)$$

The frame rate is the number of frames displayed every second. In this instance, the average frame rate of CCTV 4 and CCTV 5 in the study under consideration exhibits an increase in Table 6 relative to Table 5. This may be attributed to the utilization of densely annotated surveillance footage, which facilitates the discovery of optimal keyframes that accurately show human activity.

4.3. COMPRESSION RATIO

This is used to determine the compression level achieved by the keyframes depicted in the video sequence. (Equation 6).

$$CR=1\text{-}\{N_k / N_f\}*100 \quad (6)$$

Where N_k represents the number of extracted keyframes and N_f represents the total number of frames obtained.

4.4. PRECISION

This reveals the extraction accuracy, which is used to analyze the actual keyframe extracted (Equation 7).

$$Precision=N_c/(N_c+N_f)*100\% \quad (7)$$

Here N_c refers number of human-detected frames and N_f with total frames obtained.

4.5. RECALL

A sensitivity producer reveals the relationship between the obtained keyframe extractions to that of the actual number of required keyframes (Equation 8).

$$Recall=N_c/(N_c+ N_m)*100\% \quad (8)$$

Here N_c is the number of human-detected frames, and N_m is the number of human-detected frames that were not detected.

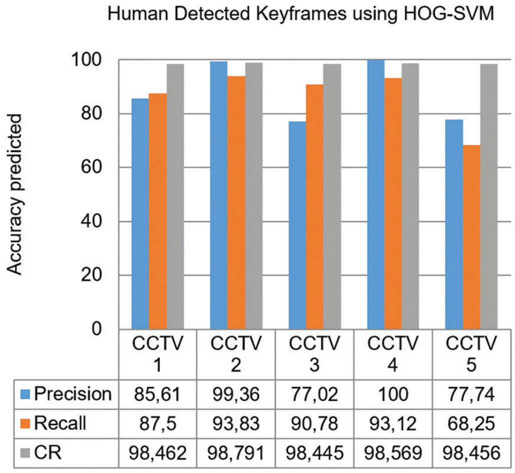
Table 5. Accuracy Determination for Human Detected Keyframes using HOG-SVM

Surveillance dataset	Avg. Fps	Frames	HOG- SVM key frame extraction	Precision	Recall	CR
CCTV1	6.892	520	8	85.61	87.50	98.462
CCTV2	6.806	579	7	99.36	93.83	98.791
CCTV3	7.043	643	10	77.02	90.78	98.445
CCTV4	6.718	629	9	100	93.12	98.569
CCTV5	6.473	583	9	77.74	68.25	98.456
					Average	**98.54**

Table 6. Accuracy Determination for Human Detected Keyframes Using Proposed Method

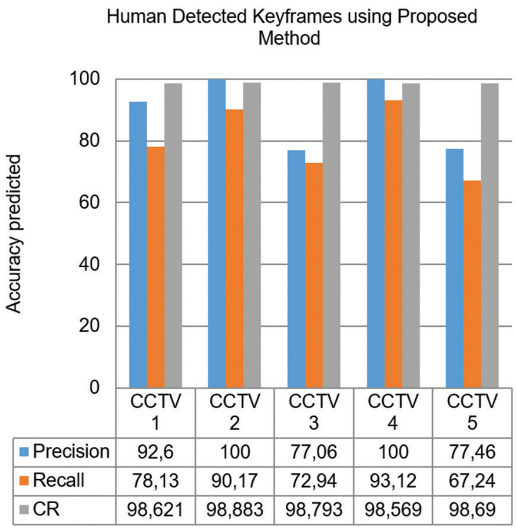
Surveillance dataset	Avg. Fps	Frames	HOG- SVM key frame extraction	Precision	Recall	CR
CCTV1	6.809	435	6	92.60	78.13	98.621
CCTV2	6.698	537	5	100	90.17	98.883
CCTV3	6.922	580	7	77.06	72.94	98.793
CCTV4	6.918	629	9	100	93.12	98.569
CCTV5	6.749	534	5	77.46	67.24	98.699
					Average	**98.71**

Therefore, reports of average frames per second with a smaller time deduction achieve the demonstration of time complexity. Additionally, the attainment of space complexity is determined by comparing the keyframe obtained in Table 6 to that of Table 5, where the former is found to be superior.



	CCTV 1	CCTV 2	CCTV 3	CCTV 4	CCTV 5
Precision	85,61	99,36	77,02	100	77,74
Recall	87,5	93,83	90,78	93,12	68,25
CR	98,462	98,791	98,445	98,569	98,456

Fig. 3. Accuracy Metrics of Human Detected Keyframes using HOG-SVM



	CCTV 1	CCTV 2	CCTV 3	CCTV 4	CCTV 5
■ Precision	92,6	100	77,06	100	77,46
■ Recall	78,13	90,17	72,94	93,12	67,24
■ CR	98,621	98,883	98,793	98,569	98,69

Fig. 4. Accuracy Metrics of Human Detected Keyframes using the proposed method

Figs. 3 and 4 illustrate the precision and recall calibrations, demonstrating that the proposed work obtains the highest level of differentiation in comparison to prior work. The accuracy metrics of the proposed method yield an average compression ratio of 98.71%, which is superior to the prior method's maximum compression ratio of 98.54%, as shown in Tables 5 and 6 for human-detected keyframes. Consequently, the complexity of performance analysis reports is reduced in terms of both time and space.

5. CONCLUSION AND FUTURE WORK

Human-detected keyframes are categorized in this paper based on the progression of research in content-based video retrieval of surveillance video. The background subtraction method and frame difference facilitate the classification of human motion via pixel change. The human is highlighted as a rectangular segment by the Spatio-temporal feature extraction using HOG-SVM. Experiments utilizing the aforementioned combination algorithm demonstrate that the proposed work enhances the human detection accuracy of keyframe extraction, thereby reducing the time complexity of criminal investigations. The proposed method eliminates the maximal redundancy of frames and demonstrates the space complexity. In future work, the video footage will be fine-tuned under all circumstances to explicitly report human detection at crime scenes.

6. REFERENCES:

[1] M. P. J. Ashby, "The Value of CCTV Surveillance Cameras as an Investigative Tool: An Empirical Analysis", European Journal on Criminal Policy and Research, Vol. 23, No. 3, 2017, pp. 441-459.

[2] A. Ranjan, D. T. Hoffmann, D. Tzionas, S. Tang, J. Romero, M. J. Black, "Learning Multi-human Optical Flow", International Journal of Computer Vision, Vol. 128, No. 4, 2020, pp. 873-890.

[3] B. Garcia-Garcia, T. Bouwmans, A. J. R. Silva, "Background subtraction in real applications: Challenges, current models and future directions", Computer Science Review, Vol.35,2020.

[4] P. Karpagavalli, A. V. Ramprasad, "An adaptive hybrid GMM for multiple human detection in crowd scenario", Multimedia Tools & Applications., Vol.76, No. 12, 2017, pp. 14129-14149.

[5] S. Abdul, R. Shaikh, L. R. Wadekar, E. Engineering, T. Engineering, "Object Detection And Classification Using Sparsity Regularized Pruning On Low-Quality Image / Video", International Journal of Creative Research Thoughts, Vol. 10, No. 6, 2022, pp. 977-985,

[6] N. Paul, A. Singh, A. Midya, P. P. Roy, D. P. Dogra, "Moving object detection using modified temporal differencing and local fuzzy thresholding", Journal of Supercomputing, Vol. 73, No. 3, 2017, pp. 1120-1139.

[7] H. S. G. Supreeth, C. M. Patil, "Efficient multiple moving object detection and tracking using combined background subtraction and clustering", Signal, Image Video Processing, Vol. 12, No. 6, 2018, pp. 1097-1105.

[8] M. S. Zaharin, N. Ibrahim, T. M. A. T. Dir, "Comparison of human detection using background subtraction and frame difference", Bulletin of Electrical Engineering and Informatics, Vol. 9, No. 1, 2020, pp. 345-353.

[9] L. Maddalena, A. Petrosino, "Background subtraction for moving object detection in RGBD data: A survey", Journal of Imaging, Vol. 4, No. 5, 2018.

[10] M. N. Chapel, T. Bouwmans, "Moving objects detection with a moving camera: A comprehensive review", Computer Science Review, Vol. 38, 2020, p. 100310.

[11] S. K. Singh, "A Comparative Study Of Different Motion Detection Algorithms In Computer Vision Applications", International Research Journal of Modernization in Engineering Technology and Science, Vol. 4, No. 6, 2022, pp. 4173-4184.

[12] R. Zhong, R. Hu, S. Member, Z. Wang, S. Wang, "Using Compressed Video", IEEE Signal Processing Letters, Vol. 21, No. 7, 2014, pp. 834-838.

[13] V. Ghait, S. Karekar, N. Lagad, K. Mohare, M. Thorat, "Survey On Key Frame Extraction And Object Detection", International Research Journal of Engineering and Technology, Vol. 7, No. 12, 2020, pp. 2002-2005.

[14] R. Hannane, A. Elboushaki, K. Afdel, P. Naghabhushan, M. Javed, "An efficient method for video shot boundary detection and keyframe extraction using SIFT-point distribution histogram", International Journal of Multimedia Information Retrieval, Vol. 5, No. 2, 2016, pp. 89-104.

[15] B. S. Rashmi, H. S. Nagendraswamy, "Effective Video Shot Boundary Detection and Keyframe Selection using Soft Computing Techniques", International Journal of Computer Vision and Image Processing, Vol. 8, No. 2, 2018, pp. 27-48.

[16] S. Chakraborty, D. M. Thounaojam, "SBD-Duo: a dual-stage shot boundary detection technique robust to motion and illumination effect", Multimedia Tools and Applications, Vol. 80, No. 2, 2021, pp. 3071-3087.

[17] Z. M. Lu, Y. Shi, "Fast video shot boundary detection based on SVD and pattern matching", IEEE Transactions on Image Processing, Vol. 22, No. 12, 2013, pp. 5136-5145.

[18] P. G. G. Lakshmi, S. Dominic, "Walsh-Hadamard transform kernel-based feature vector for shot boundary detection", IEEE Transactions on Image Processing, Vol. 23, No. 12, 2014, pp.5187-5197.

[19] D. M. Thounaojam, T. Khelchandra, K. M. Singh, S. Roy, "A Genetic Algorithm and Fuzzy Logic Approach for Video Shot Boundary Detection", Computer Intelligence and NeuroScience, Vol. 2016, 2016.

[20] S. Tippaya, S. Sitjongsataporn, T. Tan, M. M. Khan, K. Chamnongthai, "Multi-Modal Visual Features-Based Video Shot Boundary Detection", IEEE Access, Vol. 5, No. C, 2017, pp. 12563-12575.

[21] P. Aigrain, H. Zhang, D. Petkovic, "Content-based representation and retrieval of visual media: A state-of-the-art review", Multimedia Tools and Applications, Vol. 3, No. 3, 1996, pp. 179-202.

[22] H. J. Zhang, J. Y. A. Wang, Y. Altunbasak, "Content-based video retrieval and compression: A unified solution", Proceedings of International Conference on Image Processing, Santa Barbara, CA, USA, 26-29 October 1997, pp. 13-16.

[23] U. Gargi, R. Kasturi, S. H. Strayer, "Performance characterization of video-shot-change detection methods", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 10, No. 1, 2000, pp. 1-13.

[24] T. Liu, H. J. Zhang, F. Qi, "A Novel Video Key-Frame-Extraction Algorithm Based on Perceived Motion Energy Model", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 13, No. 10, pp. 1006-1013, 2003.

[25] U. Gawande, K. Hajari, Y. Golhar, "Deep Learning Approach to Key Frame Detection in Human Action Videos", Recent Trends in Computational Intelligence, 2020 pp. 1-16.

[26] S. S Gornale, A. K. Babaleshwar, P. L. Yannawar, "Detection and Classification of Signage's from Random Mobile Videos Using Local Binary Patterns", International Journal of Image, Graphics and Signal Processing, Vol. 10, No. 2, 2018, pp. 52-59.

[27] D. Asha, Y. Madhavee Latha, V. S. K. Reddy, "Content-Based Video Retrieval System Using Multiple Features", International Journal of Pure and Applied Mathematics, Vol. 118, No. 14, 2018, pp. 287-294,

[28] M. Jian, S. Zhang, L. Wu, S. Zhang, X. Wang, Y. He, "Deep key frame extraction for sports training", Neurocomputing, Vol. 328, 2019, pp. 147-156.

[29] A. Ioannidis, V. Chasanis, A. Likas, "Weighted multi-view key-frame extraction", Pattern Recognition Letters, Vol. 72, 2016, pp. 52-61,

[30] M. Sima, "Key frame extraction for human action videos in dynamic spatio-temporal slice clustering", Journal of Physics: Conference Series, Vol. 2010, No. 1, 2021.

[31] N. Dalal, B. Triggs, "Histograms of Oriented Gradients for Human Detection", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 7509, 2005, pp. 94-102.

[32] A. J. Lipton, H. Fujiyoshi, R. S. Patil, "Moving target classification and tracking from real-time video", Proceedings Fourth IEEE Workshop on Applications of Computer Vision, Princeton, NJ, USA, 19-21 October 1998, pp. 8-14.

[33] J. Guo, J. Wang, R. Bai, Y. Zhang, Y. Li, "A New Moving Object Detection Method Based on Frame-difference and Background Subtraction", IOP Conference Series: Materials Science and Engineering, Vol. 242, No. 1, 2017.

[34] D. S. Suresh, M. P. Lavanya, "Motion Detection and Tracking using Background Subtraction and Consecutive Frames Difference Method", International Journal of Research Studies in Science, Engineering and Technology, Vol. 1, No. 5, 2014, pp. 16-22.

[35] P. Ramya, R. Rajeswari, "A Modified Frame Difference Method Using Correlation Coefficient for Background Subtraction", Procedia Computer Science, Vol. 93, 2016, pp. 478-485.

[36] T. Mahalingam, M. Subramoniam, "A robust single and multiple moving object detection, tracking and classification", Applied Computing and Informatics, Vol. 17, No. 1, 2017, pp. 2-18.

[37] V. N. Vapnik, "An overview of statistical learning theory", IEEE Transactions on Neural Networks, Vol. 10, No. 5, 1999, pp. 988-999.

[38] H. Aung, A. V. Bobkov, N. L. Tun, "Face detection in real-time live video using yolo algorithm based on VGG16 convolutional neural network", Proceedings of the International Conference on Industrial Engineering, Applications and Manufacturing, Sochi, Russia, 17-21 May 2021, pp. 697-702.

[39] H. Yang, Q. Tian, Q. Zhuang, L. Li, Q. Liang, "Fast and robust key frame extraction method for gesture video based on high-level feature representation", Signal, Image and Video Processing, Vol. 15, No. 3, 2021, pp. 617-626.

A combined method based on CNN architecture for variation-resistant facial recognition

Original Scientific Paper

Hicham Benradi
University Mohammed V,
High School of Technology Salé, Mohammadia School of Engineering,
Systems Analysis and Information Processing and Industrial Management LaboratoryRabat, Morocco
benradi.hicham@gmail.com

Ahmed Chater
University Mohammed V,
High School of Technology Salé, Mohammadia School of Engineering, Systems Analysis and Information Processing and Industrial Management LaboratoryRabat, Morocco
ahmedchater11@gmail.com

Abdelali Lasfar
University Mohammed V,
High School of Technology Salé, Mohammadia School of Engineering, Systems Analysis and Information Processing and Industrial Management LaboratoryRabat, Morocco
ali.lasfar@gmail.com

Abstract *– Identifying individuals from a facial image is a technique that forms part of computer vision and is used in various fields such as security, digital biometrics, smartphones, and banking. However, it can prove difficult due to the complexity of facial structure and the presence of variations that can affect the results. To overcome this difficulty, in this paper, we propose a combined approach that aims to improve the accuracy and robustness of facial recognition in the presence of variations. To this end, two datasets (ORL and UMIST) are used to train our model. We then began with the image pre-processing phase, which consists in applying a histogram equalization operation to adjust the gray levels over the entire image surface to improve quality and enhance the detection of features in each image. Next, the least important features are eliminated from the images using the Principal Component Analysis (PCA) method. Finally, the pre-processed images are subjected to a neural network architecture (CNN) consisting of multiple convolution layers and fully connected layers. Our simulation results show a high performance of our approach, with accuracy rates of up to 99.50% for the ORL dataset and 100% for the UMIST dataset.*

Keywords*: Histogram equalization, PCA, CNN, facial recognition, variations*

1. INTRODUCTION

Facial recognition is a technology that allows the identification of a person by analyzing and comparing unique features of the face, such as the shape of the nose, the distance between the eyes, or the facial lines. This technology is increasingly used in various fields such as security [1, 2], Human face recognition and age estimation [3], video surveillance [4], gender identification from an image [5], biometric identification [6] or individual identification [7-9]. However, the presence of variance that can occur in several forms (lighting, orientation, pose, accessories, etc.) in an image can affect facial recognition, since facial recognition algorithms need a clear, sharp image of the face to identify unique features and compare them with a database of recorded faces [10]. This is why it is important to take variance into account when designing and evaluating facial recognition algorithms, and to ensure that they are capable of handling a wide variety of situations and image conditions to guarantee accurate and reliable identification.

Recently convolutional neural networks (CNN) have been very successful in many computer vision applications such as medicine [11, 12], agriculture [13], and environment [14, 15]. And also, among these applications, several works dealing with facial recognition are based on the use of CNNs due to their robustness in feature extraction and classification such as [16-18]. CNNs are a category of deep neural networks used mainly in the field of computer vision. Inspired by the structure and functioning of the human visual system, CNNs are mainly used for classification tasks. Their architecture consists of several layers including convolution layers which are responsible for extracting features from the image using convolution filters applied on different parts of the image, then pooling layers which are used to reduce the dimensionality of the dataset by selecting the most important features and finally fully connected layers which are responsible for the final classification. Thanks to these layers the CNNs can automatically learn the discriminating features of a facial image.

In this paper, we propose a combined approach to improve the accuracy and robustness of our facial recognition model when multiple variance shapes are present in an image. This approach first uses histogram equalization technique pre-processing to improve the quality of the images as it adjusts the distribution of gray levels in an image to improve the visibility of details and increase the contrast this is beneficial as it enhances the contours and details of the face, thus facilitating the detection and identification of the unique facial features. Then the principal component analysis (PCA) method [19] is applied to extract the most important features from the images by reducing the dimensionality of the data set. PCA also simplifies and reduces the complexity of the face data by extracting the most important and discriminating features. And at the end, the set of processed images is transmitted to a CNN architecture for training our model. Our approach has been evaluated using two image databases which are ORL [20] and UMIST [21] which represent multiple variations in pose and lighting and accessories such as glasses, scarf, and beard... The performance of our model is evaluated using an accuracy metric.

The results of our simulations show that our method performed satisfactorily, with accuracy rates of up to 99.50% for the ENT dataset and 100% for the UMIST dataset. These results are competitive with those of other research studies in the same field.

2. RELATED WORK

Facial recognition is a fast-growing research area that presents many challenges, including variation under different types such as occlusion, pose, and the presence of accessories (glasses, cap, scarf ...). To remedy this problem several relevant works have addressed this problem. The authors in [22] proposed an efficient face recognition system incorporating genetic algorithms. Their model is based on two steps: the first one consists in extracting the face features and the second one consists in matching the face models. The results of the simulations have allowed us to obtain quality results with an accuracy that reaches a value of 88.9%. In [23] the authors proposed a face recognition algorithm based on depth map transfer learning to efficiently recognize face images taken in an unrestricted environment. This method was able to record an accuracy rate that reached a value of 98.31%. On the other hand, another model based on transfer learning has been designed [24]. Its goal is to design a facial recognition model invariant to the activated age. For this a preprocessing is performed on the facial images to improve their quality, then a BES-DTL-AIFR model which is based on the Inception V3 model is also used to learn deep features and at the end, the features are passed to the optimal deep belief network model DBN. The model recorded an accuracy rate of a value of 99.14%. The authors in [25] proposed penalized competitive deep rival learning RPCL for deep face recognition in a low-resolution image. Their model achieved an accuracy value of 95.13%. Another hybrid biometric system for face recognition considering uncontrollable environmental conditions was developed in [26]. The proposed system uses two features which are discrete wavelet transform based on the Gaussian Laplace filter and Log Gabor filter. Both features were used by a multi-class support vector machine. The system was able to achieve an accuracy rate of a value of 94.68%.

3. METHODOLOGY

The figure below presents the steps of our approach. It consists of four steps which are:

- Application of histogram equalization on the set of images used
- Application of dimension reduction by the PCA method
- Labeling and data preparation
- Feature extraction and classification by CNN architecture

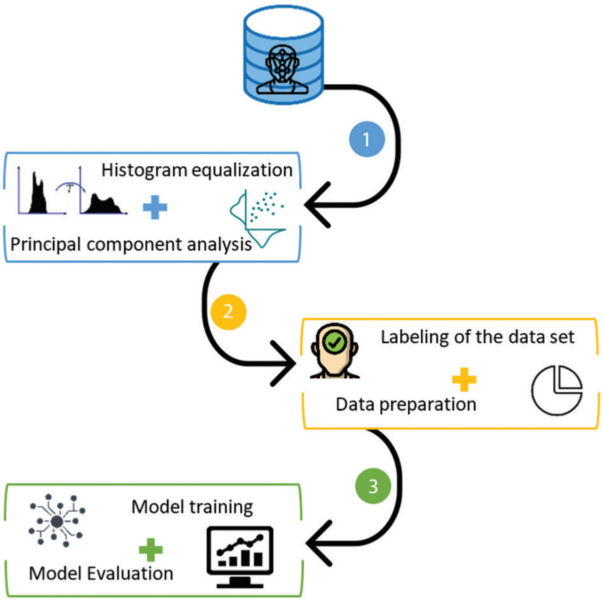


Fig. 1. Graphic representation of the different steps of our method

3.1. HISTOGRAM EQUALIZATION

The equalization of the histogram is a technique of image processing made by Gonzalez and Woods in ref which allows the uniform adjustment of the distribution of gray levels on the entire surface of an image. For this, we begin by calculating the histogram of an image (X) which is a graphical representation of the distribution of gray levels (L) in the form of a curve that indicates the number of pixels that have a particular level of gray in an image. Then the cumulative distribution function CDF of the histogram is calculated to determine the transformation to be applied to the image to equalize its histogram. We define the number NK which is the number of occurrences of the level XK it gives that the probability of occurrence of a pixel of level XK in an image is presented by the following equation

$$p_x(x_k) = p(x = x_k) = \frac{n_k}{n}, 0 \le k \ < L \quad (1)$$

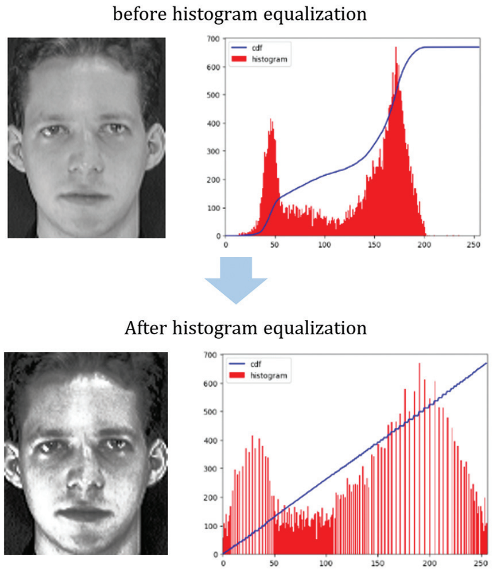
With: n presents the total number of pixels of an image and p_x presents the histogram normalized on [0,1].

Finally, an equalization transformation (T) is applied to the image using the CDF. It aims to replace each gray level of the image by its equivalent value in the CDF. For this we associate a new value SK=T(Xk) has this transformation T on each pixel of value XK as shown in the following equation:

$$T(x_k) = (L-1)\sum_{j=0}^{k} p_x(x_j) \quad (2)$$

With $\sum_{j=0}^{k} p_x(x_j)$ showing the cumulative histogram of an image.

The resulting image will have a uniform histogram which will allow the increase of contrast and brightness of the image. Fig. 2 below shows an example of the use of this technique on an image belonging to the ORL dataset and as can be seen the cumulative histogram distributed over the entire surface so that it becomes uniform.


before histogram equalization

After histogram equalization

Fig. 2. Application of histogram equalization on an image belonging to the ORL database

3.2. PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis PCA [19] is a dimensionality reduction technique that decomposes a data matrix into principal components while retaining the maximum amount of information contained in the data. It aims to improve the performance of Deep Learning algorithms because it keeps just the uncorrelated variables and eliminates the correlated ones that do not contribute to a decision.

For this purpose, each image is represented by a vector Ii containing pixels, which will later be treated as a one-dimensional array were

$$\mathrm{I_i} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \quad (3)$$

where $1 \le a_{ij} \le 255$

a represents the pixels of an image

Then the matrix will be converted to a one-dimensional array as follows:

$$\tau_i = \begin{pmatrix} a_{11} \\ \cdots \\ a_{mn} \end{pmatrix} \quad (4)$$

Where τ_i is an array that will contain many pixels.

By the soot, the average of all the images is determined by the elimination of all that is in common with the individuals, as shown by the following equation:

$$\psi_{\mathrm{moy}} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{I_i} \quad (5)$$

Where $\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{I_i}$ is the sum of the values for each image.

Then the matrix $\emptyset_i$ constructed by performing a subtraction between the one-dimensional array of pixels τ_i the average ψ_{moy} as shown in the equation below:

$$\emptyset_i = \tau_i - \psi_{moy} \quad (6)$$

Then another modified image matrix this time of covariance representing the interaction between the images of a single individual as shown in equation (7) Below:

$$C = \frac{1}{M}\sum_{n=1}^{M} \emptyset_n \emptyset_n^T = AA^T \quad (7)$$

where A=($\emptyset_1$, $\emptyset_2$,....,$\emptyset_M$)(N^2×M)

Where C is the covariance matrix, M is a set of vectors and $\emptyset_n$ $\emptyset_n{}^T$ represents the tensor product of the feature vectors $\emptyset_n$ and $\emptyset_n{}^T$.

Subsequently, the eigenvectors U_i, are calculated from the covariance matrix C, then a sorting of these vector verticals is formed based on their eigenvalue where $||U_i||$=1 which corresponds to the importance of the direction of the data set. A selection of the first K eigenvectors that contain the most information for the projected of the dataset on the selected K eigenvectors to reduce the dimensionality of the dataset. At the end a reconstruction of the image using the K-selected eigenvectors.

As shown in Fig. 3 below, even if we apply the PCA method, we can see that the appearance of the image is preserved, but the dimensions are reduced.

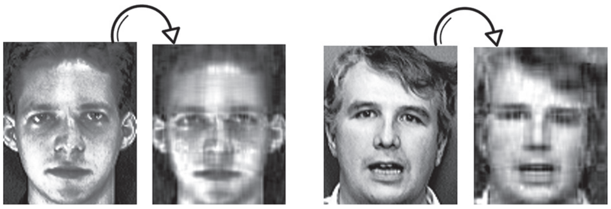


Fig. 3. Application of PCA on images belonging to the ORL database

3.3. DATA PREPARATION

Once the dataset is processed a preparation is performed on it aiming to unify the size of the images (48x48) then all the images will be labeled where each image will be labeled by the class that corresponds to it, then we proceed to the formation of two subsets of data. Each of these two will be used in a learning phase the first part will be used in the training phase of our model will contain 80% of the images is will be called TRAIN, while the second subset called TEST will contain 20% of the images is will be used in the validation phase of our model.

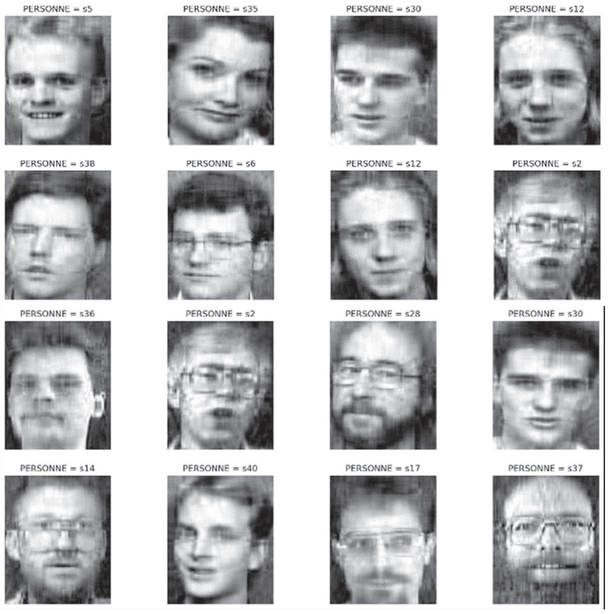


Fig. 4. Labeling of the ORL data set

3.4. CONVOLUTIONAL NEURAL NETWORK ARCHITECTURE CNN

The final phase aims at building an architecture of convolutional neural networks (CNN) which are a class of deep neural networks mainly used in the field of computer vision. Their structure and operation are inspired by the human visual system. They are used for classification tasks as they are designed for automatic extraction of image features. Their complex architecture consists of several layers in our case we used convolution layers that are responsible for the extraction of features from an image by applying convolution filters on different parts of the image. Also, the POOLING layer's role is to reduce the dimensionality of the dataset by selecting the most important features, so a FLATTEN layer has a very important role in our architecture because it allows us to convert the output of the last convolution layer into a 1D vector so that they can be used as input data by the layer of fully connected neurons DENSE to perform classification based on the extracted features. The DENSE layer consists of neurons that are connected to all the neurons of the previous layers which will allow us to learn complex relations between the features extracted by the previous convolution layers by applying linear and non-linear operations to transform the input vector into an output that can be interpreted as a prediction with the help of an activation function. In our case, we used the Rectified Linear Unit (RELU) activation function which is one of the most popular activation functions used in CNN. Its role is to introduce nonlinearity into the neural network because the convolution layers perform linear operations, which means that the output of the layer is a linear combination of the inputs. Adding a nonlinear function will allow the neural networks to learn more complex nonlinear representations of the input data. This is done using the following equation:

$$f(x) = \max(0, x) \begin{cases} if\ x > 0\ , & f(x) = x, \\ if\ x < 0, & f(x) = 0 \end{cases} \tag{8}$$

x represents the output of the layer.

One of the advantages of using this activation function is that it allows us to deal with complex input data because it avoids the disappearance of gradients and also it helps in a good regularization of our model. Also, we used the SOFTMAX optimization algorithm in the output layer to perform a multi-class classification to classify the data into several categories. Its main role is to normalize the scores of each output class into a probability distribution that represents the probability of each class being the correct prediction using the following equation:

$$\sigma(z)_j = \frac{e^{zj}}{\sum_{k\text{à}1}^{k} e^{zj}} \text{ where } j \in \{1, \ldots\ldots, k\} \tag{9}$$

Z is a vector of real numbers that represents a score for a particular class j and k is the class number.

Fig. 5 below summarizes our adopted architecture. It is composed of three convolution layers, three pooling layers, and a flattened layer. All these layers will be responsible for the extraction of features from the images. Then a fully connected layer is used to perform a classification.

Our model was compiled using a categorical_crossentropy loss function which is widely used in machine learning and in particular to solve multi-class classification problems. Its role is to measure the divergence between the probability distribution predicted by a model and the actual distribution of classes by computing the output probability of a model and the labels that correspond to the classes. This operation is performed using the following equation:

$$j(\theta) = -\sum_i y_i \times \log(\hat{y}_i) \qquad (10)$$

Where $j(\theta)$ presents the total loss calculated from the predictions of the model, y_i represents the true value of class i(0 or 1) and $\hat{y}_i$ represents the probability predicted for class i by a model.

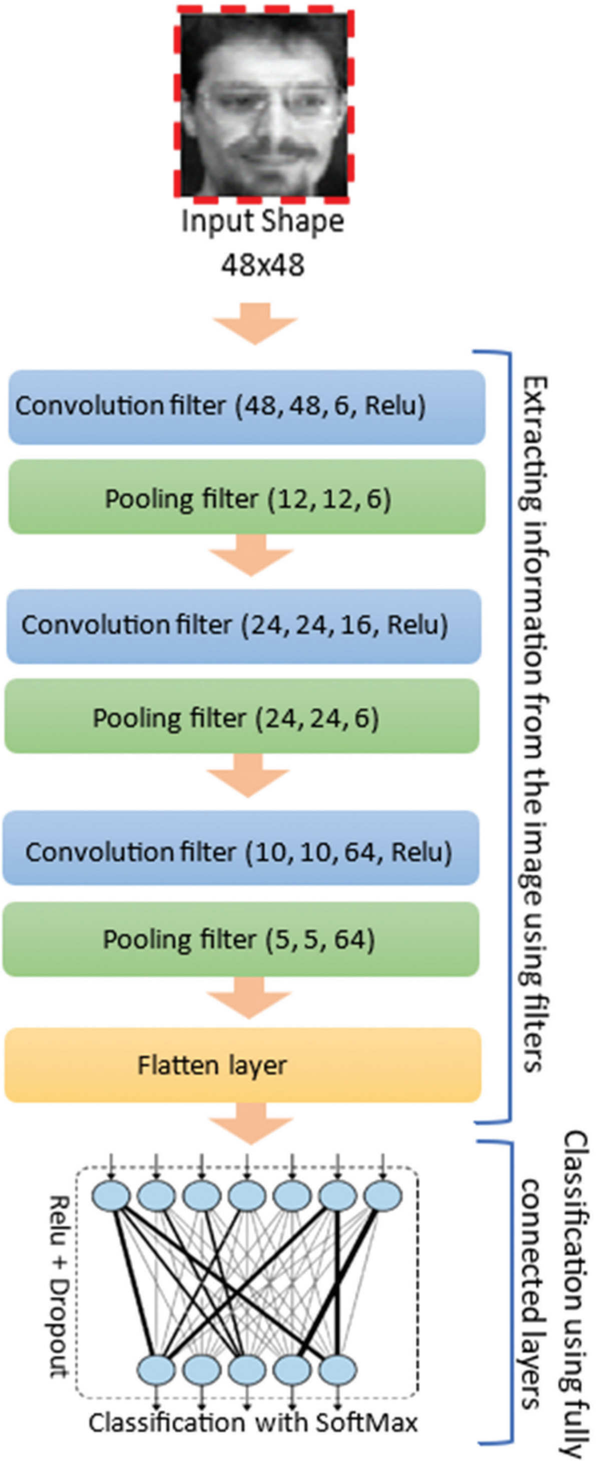

Input Shape
48x48
Convolution filter (48, 48, 6, Relu)
Pooling filter (12, 12, 6)
Convolution filter (24, 24, 16, Relu)
Pooling filter (24, 24, 6)
Convolution filter (10, 10, 64, Relu)
Pooling filter (5, 5, 64)
Flatten layer
Extracting information from the image using filters
Relu + Dropout
Classification with SoftMax
Classification using fully connected layers

Fig. 5. CNN architecture adopted for our approach

The main objective of using this loss function is to minimize the value of the loss by adjusting the weights and biases of our model to obtain predictions that are as close as possible to the actual labels.

The classification performance evaluation of our model is performed using an accuracy metric which has as its main objective to measure the proportion of correct predictions about the set of predictions resulting from our model. This measure represented in percentage (%) is performed by comparing the predictions of our model with the real labels belonging to the data set used as shown in the following formula:

$$Accuracy = \frac{N_c}{N_t} \times 100 \qquad (11)$$

Where N_c represents the correct number of predictions predicted by our model and N_t is the total number of items in the data set used.

We also note the use of the Adam optimization algorithm. Its role is to update the weights of our model by calculating the gradients of our model for a training batch of data, then updates the first-moment (M_t) and the second moment (V_t) using the following two formulas:

$$M_t = \beta_1 \times m_{(t-1)} \times (1 - \beta_1) \times g \qquad (12)$$

$$V_t = \beta_2 \times v_{(t-1)} \times (1 - \beta_2) \times g^2 \qquad (13)$$

Where g represents the gradients, β_1 and β_2 represents the exponential decay parameters.

Then the moment biases M_t and V_t are corrected using the following two formulas:

$$M_{t\_corr} = \frac{M_t}{(1-\beta_1^t)} \qquad (14)$$

$$V_{t\_corr} = \frac{V_t}{(1-\beta_2^t)} \qquad (15)$$

Namely, t is the iteration number.

And finally, the Adam algorithm will update all the weights of our model using the following formula:

$$\theta_{(t+1)} = \frac{\theta_t - \alpha \times M_{t\_corr}}{\sqrt{V_{t\_corr} + \varepsilon}} \qquad (16)$$

In the formula, θ represents the weights of our model, α corresponds to the learning rate and ε is a small value added to avoid division by 0 (in our case we used 10-8).

All these operations will allow our model to adjust the learning rates adaptively based on the estimates of the first and second moments, which guarantees rapid convergence and stability during the training phase of our model.

4. RESULT AND DISCUSSION

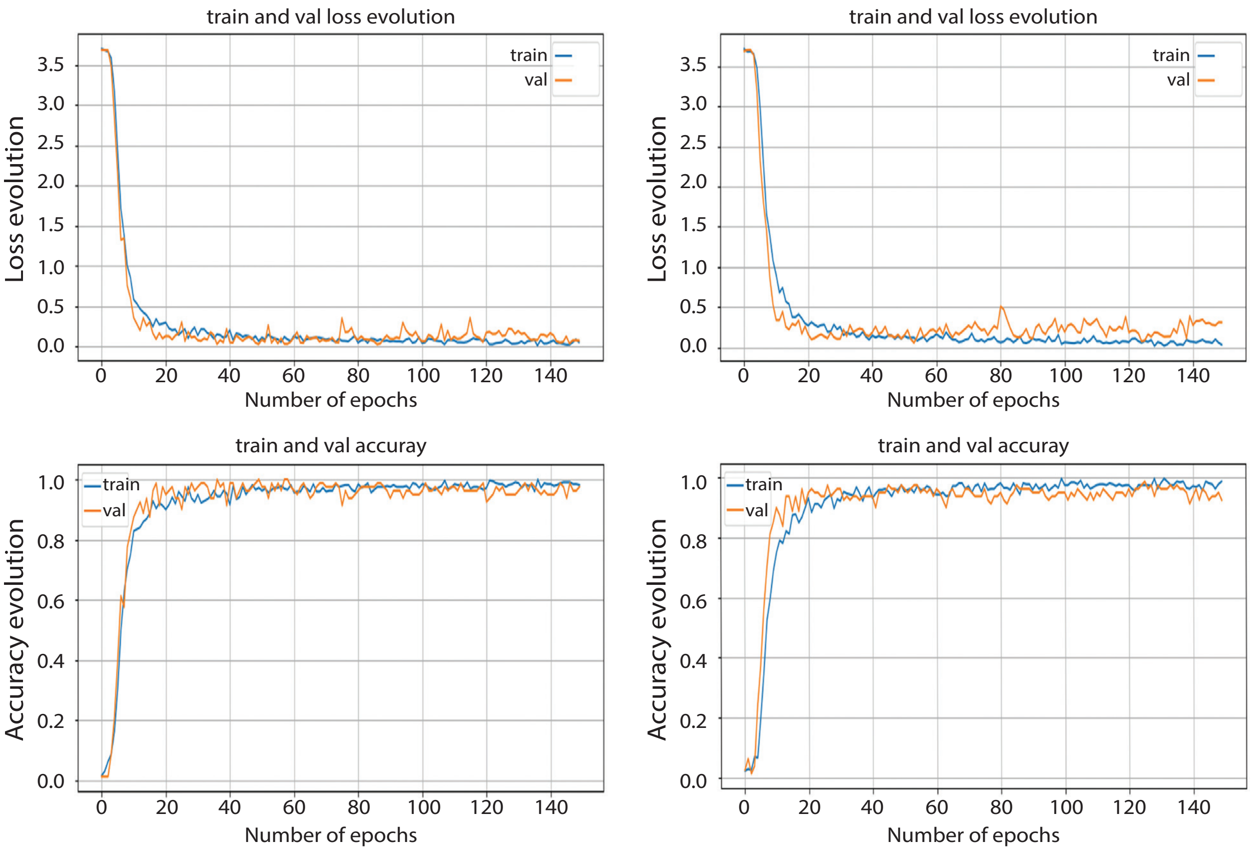
Our model was trained using two databases ORL and UMIST which each represent the presence of variations of different types, over a total of 150 epochs, where each epoch corresponds to an iteration on the TRAIN training data set. Also, the batch_Size parameter was set to 8 which means that the model is trained using mini-batches of a sample of 8 mages at a time. We also note that the TEST dataset is used for the validation of our model during the training phase. The loss function and evaluation metrics are calculated for this dataset after each epoch.

The results of our simulations shown in Table 1 below have demonstrated a remarkable performance recorded in both cases (ORL and UMIST) as shown in Table 1 below. In the case of the ORL database which represents the presence of variations in contrast, lighting, and occlusion (glasses, sling ...) our approach recorded an accuracy rate of up to 99.50% which is the best score recorded among the other techniques while a value of 0.07 was recorded by our model as the precision of the loss function which indicates that our model has succeeded in minimizing the value of the loss function. In the case of the UMIST database which represents a variation at the pose level (from profile to front view), our model was able to achieve 100% value in accuracy and a value of 0.0000093 was recorded by our model as the accuracy of the loss function.

In the case of the ORL database, as illustrated in Figure 6 below, the analysis of the graph representing the evolution of the two values Accuracy and Loss during the two training and validation phases shows that the accuracy value increases constantly and that the loss value also decreases constantly, which indicates that our model is gradually improving until it reaches optimal values (99.50% for accuracy and 0.07 for loss value) on the two data sets Train and Test. On the other hand, in the case of using the histogram equalization technique combined with the CNN, the two values (precision and loss) do not reach interesting values (92.50% as the maximum precision value and 0.31 as the minimum loss value) which makes the model resulting from this technique weak compared to our method.

Table 1. Simulation results on both databases

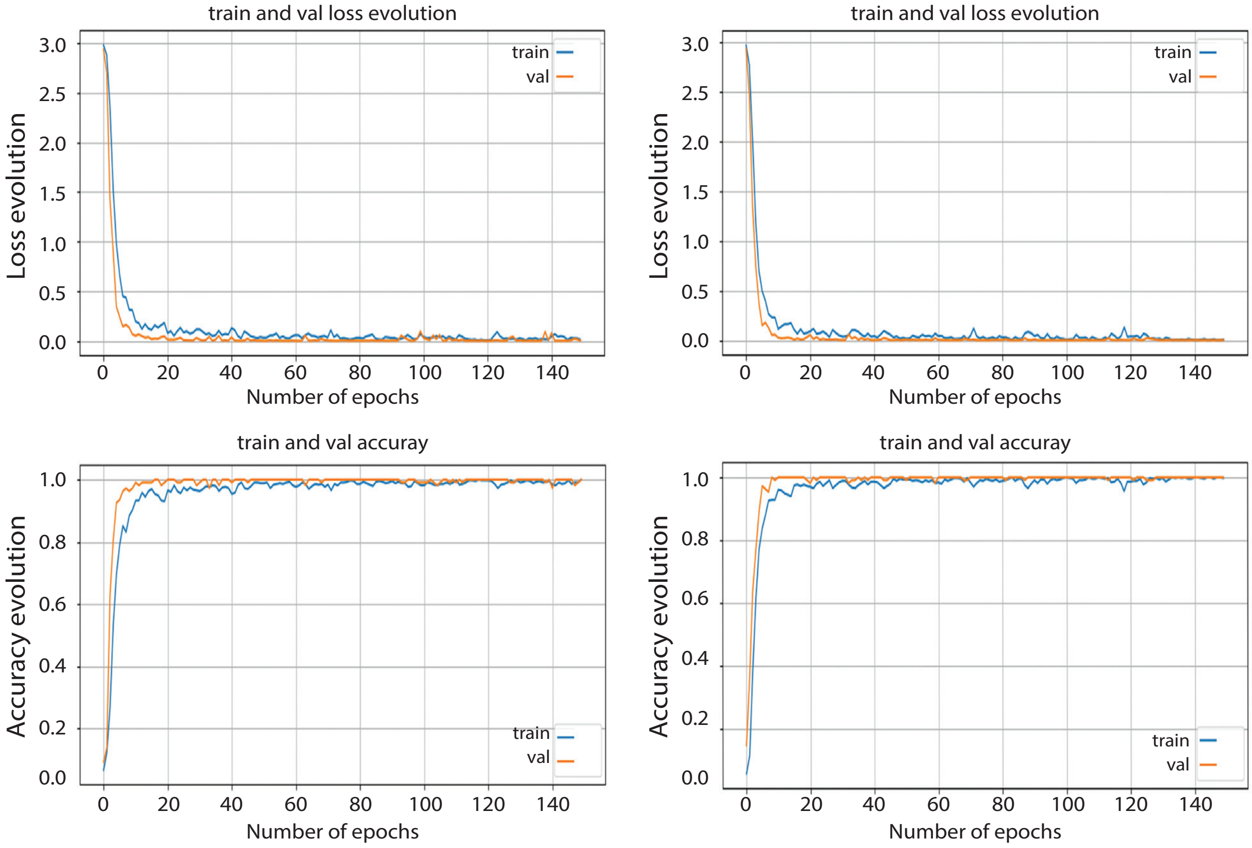
Database	ORL		UMIST	
Technique	Accuracy	Loss	Accuracy	Loss
CNN	93,75%	0.25	99,13%	0.15
ACP	93,75%	0.15	100%	0.00065
Equalization + CNN	92,50%	0.31	99,13%	0.081
Method	99,50%	0.07	100%	0.0000093



(a) (b)

Fig. 6. Graph representing the evolution of precision and loss values a: in the case of our approach and b: in the case of using histogram equalization + CNN

In the case of the UMIST database as illustrated in Fig. 7 below the analysis of the graph representing the evolution of the two values Accuracy and Loss during the two phases Training and Validation we note the stability of the curves over the periods and also that the value of accuracy increases in a constant way and that the value of loss also decreases constantly until reaching a final value of accuracy of 100% and 0.0000093 as the final value of a loss on the two sets of data Train and Test. It is also noted that in the case of using the PCA technique combined with the CNN, the model also reaches a final accuracy value of 100% but the final value recorded by this model 0.00065 shows that it is not effective as the case of our method.



(a) (b)

Fig. 7. Graph representing the evolution of the precision and loss values a: in the case of our approach and b: in the case of using the PCA + CNN method

A comparison aimed at evaluating our approach with other research works using the same databases (ORL and UMIST) has been carried out as illustrated in Table 2 below. The analysis of this comparison shows that our approach is competitive with other results of various research works done in the same direction and achieves convincing results in terms of the accuracy rate of facial recognition with the presence of variations and occlusion of different types.

Table 2. Comparison of the accuracy rates of others research works with our approach

ORL		UMIST	
Method	**Accuracy**	**Method**	**Accuracy**
LBP+CNN [8]	100%	Genetic algorithm [27]	100 %
HSL [28]	96.67 %	SESRC&LDF [29]	99.13 %
GABOR [30]	100%	Fusion Local& Glob [31]	99.4 %
Modified PSO [32]	99 %	CRHM [33]	99.51%
LTP-Deep CNN [34]	98.75 %	SIFT+SVM [35]	99.44%
Method	99.50 %	Method	100%

5. CONCLUSION

Facial recognition, part of the artificial intelligence sector, is a technique that aims to identify an individual from an image. This identification can be ineffective in the presence of variation or occlusion. In this paper, we propose a new approach aimed at improving the performance of face recognition in the presence of different types of variation or occlusion. To this end, we have used two image datasets (ENT and UMIST) that present the presence of several variations and occlusions. Our method begins with a pre-processing phase performed on the images used, which consists in applying histogram equalization to increase the contrast and visibility of facial image details, thereby improving recognition accuracy. Next, the PCA method was also applied to all the images used, reducing the dimensionality of all the facial image data while retaining the most important information. The second phase consists of passing all the pre-processed images from the first phase to our own CNN architecture, which consists of several convolution layers for extracting image features and also fully connected layers for image classification. The results of our simulations demonstrated the effectiveness of our approach, recording an accuracy value of 100% when using the UMIST dataset and 99.50% when using the ENT dataset. These results make our approach competitive with others developed by other researchers. This combination can be used in biometric face recognition systems, as it has demonstrated high performance. It should also be noted that in the future, we plan to use other techniques to improve the performance of our model in face recognition in the presence of variance or occlusion, such as the use of reinforcement learning techniques to improve CNN efficiency and reduce dependence on training data, as well as exploring other deeper and more complex CNN architectures.

6. REFERENCES

[1] T. Bagchi et al. "Intelligent security system based on face recognition and IoT", Materials Today: Proceedings, Vol. 62, 2022, pp. 2133-2137.

[2] F. Majeed et al. "Investigating the efficiency of deep learning based security system in a real-time environment using YOLOv5", Sustainable Energy Technologies and Assessments, Vol. 53, 2022, p. 102603.

[3] Kvita, R. S. Chhillar, "Human Face Recognition and Age Estimation with Machine Learning: A Critical Review and Future Perspective", International Journal of Electrical and Computer Engineering Systems, Vol. 13, No. 10, 2022.

[4] R. M. Alairaji, I. A. Aljazaery, H. T. S. Alrikabi, A. H. M. Alaidi, "Automated Cheating Detection based on Video Surveillance in the Examination Classes", International Journal of Interactive Mobile Technologies, Vol. 16, No. 8, 2022, pp. 124-137.

[5] M. J. Al Dujaili, H. T. H. S. Al Rikabi, N. K. Abed, I. R. N. Al Rubeei, "Gender Recognition of Human from Face Images Using Multi-Class Support Vector Machine (SVM) Classifiers", International Journal of Interactive Mobile Technologies, Vol. 17, No. 8, 2023, pp. 113-134.

[6] A. Thapliyal, O. P. Verma, A. Kumar, "Multimodal Behavioral Biometric Authentication in Smartphones for Covid-19 Pandemic", International Journal of Electrical and Computer Engineering Systems, Vol. 13, No. 9, 2022, pp. 777-790.

[7] K. Romic, C. Livada, A. Glavas, "Single and Multi-Person Face Recognition Using the Enhanced Eigenfaces Method", International Journal of Electrical and Computer Engineering Systems, Vol. 7, No. 1, 2016.

[8] J. Tang, Q. Su, B. Su, S. Fong, W. Cao, X. Gong, "Parallel ensemble learning of convolutional neural networks and local binary patterns for face recognition", Computer Methods and Programs in Biomedicine, Vol. 197, 2020, p. 105622.

[9] S. B. R. Prasad, B. S. Chandana, "Human Face Emotions Recognition from Thermal Images Using DenseNet", International Journal of Electrical and Computer Engineering Systems, Vol. 14, No. 2, 2023, pp. 155-167.

[10] H. Benradi, A. Chater, A. Lasfar, "A hybrid approach for face recognition using a convolutional neural network combined with feature extraction techniques", International Journal of Artificial Intelligence, Vol. 12, No. 2, 2023, p. 627.

[11] A. Chattopadhyay, M. Maitra, "MRI-based brain tumour image detection using CNN based deep learning method", Neuroscience Informatics, Vol. 2, No. 4, 2022, p. 100060.

[12] C. B. Gonçalves, J. R. Souza, H. Fernandes, "CNN architecture optimization using bio-inspired algorithms for breast cancer detection in infrared images", Computers in Biology and Medicine, Vol. 142, 2022, p. 105205.

[13] V. Singh, A. Chug, A. P. Singh, "Classification of Beans Leaf Diseases using Fine Tuned CNN Model", Procedia Computer Science, Vol. 218, 2023, pp. 348-356.

[14] P. Mei, M. Li, Q. Zhang, G. Li, L. Song, "Prediction model of drinking water source quality with potential industrial-agricultural pollution based on CNN-GRU-Attention", Journal of Hydrology, Vol. 610, 2022, p. 127934.

[15] M. Mentet, N. Hongkarnjanakul, C. Schwob, L. Mezeix, "Method to apply and visualize physical models associated to a land cover performed by CNN: A case study of vegetation and water cooling effect in Bangkok Thailand", Remote Sensing Applications: Society and Environment, Vol. 28, 2022, p. 100856.

[16] M. L. Prasetyo et al. "Face Recognition Using the Convolutional Neural Network for Barrier Gate System", International Journal of Interactive Mobile Technologies, Vol. 15, No. 10, 2021, p. 138.

[17] G. Revathy, K. Bhavana Raj, A. Kumar, S. Adibatti, P. Dahiya, T. M. Latha, “Investigation of E-voting system using face recognition using convolutional neural network (CNN)”, Theoretical Computer Science, Vol. 925, 2022, pp. 61-67.

[18] A. Chater, H. Benradi, A. Lasfar, “Method of optimization of the fundamental matrix by technique speeded up robust features application of different stress images”, International Journal of Electrical and Computer Engineering, Vol. 12, No. 2, 2022, p. 1429.

[19] M. Turk, A. Pentland, “Eigenfaces for Recognition”, Journal of Cognitive Neuroscience, Vol. 3, No. 1, 1991, pp. 71-86.

[20] F. S. Samaria, A. C. Harter, “Parameterisation of a stochastic model for human face identification”, in Proceedings of 1994 IEEE Workshop on Applications of Computer Vision, Sarasota, FL, USA, 5-7 December 1994, pp. 138-142.

[21] D. B. Graham, N. M. Allinson, “Characterising Virtual Eigensignatures for General Purpose Face Recognition”, Face Recognition, Springer Berlin Heidelberg, 1998, pp. 446-456.

[22] N. Asha, A. S. S. Fiaz, J. Jayashree, J. Vijayashree, J. Indumathi, “Principal component analysis on face recognition using artificial firefirefly swarm optimization algorithm”, Advances in Engineering Software, Vol. 174, 2022, p. 103296.

[23] D. Tang, J. Hao, “A deep map transfer learning method for face recognition in an unrestricted smart city environment”, Sustainable Energy Technologies and Assessments, Vol. 52, 2022, p. 102207.

[24] S. Alsubai, M. Hamdi, S. Abdel-Khalek, A. Alqahtani, A. Binbusayyis, R. F. Mansour, “Bald eagle search optimization with deep transfer learning enabled age-invariant face recognition model”, Image and Vision Computing, Vol. 126, 2022, p. 104545.

[25] P. Li, S. Tu, L. Xu, “Deep Rival Penalized Competitive Learning for low-resolution face recognition”, Neural Networks, Vol. 148, 2022, pp. 183-193.

[26] Vijaya K. H. R., Mathivanan M., “A novel hybrid biometric software application for facial recognition considering uncontrollable environmental conditions”, Healthcare Analytics, Vol. 3, 2023, p. 100156.

[27] P. Sukhija, S. Behal, P. Singh, “Face Recognition System Using Genetic Algorithm”, Procedia Computer Science, Vol. 85, 2016, pp. 410-417.

[28] S. Zhao, W. Liu, S. Liu, J. Ge, X. Liang, “A hybrid-supervision learning algorithm for real-time uncompleted face recognition”, Computers and Electrical Engineering, Vol. 101, 2022, p. 108090.

[29] M. Liao, X. Gu, “Face recognition approach by subspace extended sparse representation and discriminative feature learning”, Neurocomputing, Vol. 373, 2020, pp. 35-49.

[30] R. Hammouche, A. Attia, S. Akhrouf, Z. Akhtar, “Gabor filter bank with deep autoencoder based face recognition system”, Expert Systems with Applications, Vol. 197, 2022, p. 116743.

[31] A. M. Sahan, A. S. Al-Itbi, “The Fusion of Local and Global Descriptors in Face Recognition Application”, Advances in Communication and Computational Technology, Lecture Notes in Electrical Engineering, Vol. 668, Springer Nature Singapore, 2021, pp. 1397-1408.

[32] Y. Zhang, L. Yan, “Face recognition algorithm based on particle swarm optimization and image feature compensation”, SoftwareX, Vol. 22, 2023, p. 101305.

[33] H. Li, Z. Zhou, C. Li, C. Y. Suen, “A near effective and efficient model in recognition”, Pattern Recognition, Vol. 122, 2022, p. 108173.

[34] A. Zeroual, M. Amroune, M. Derdour, A. Bentahar, “Lightweight deep learning model to secure authentication in Mobile Cloud Computing”, Journal of King Saud University - Computer and Information Sciences, Vol. 34, No. 9, 2022, pp. 6938-6948.

[35] B. Hicham, C. Ahmed, L. Abdelali, “Face recognition method combining SVM machine learning and scale invariant feature transform”, Proceedings of the 10th International Conference on Innovation, Modern Applied Science & Environmental Studies, 2022, p. 01033.

ICU Patients' Pattern Recognition and Correlation Identification of Vital Parameters Using Optimized Machine Learning Models

Original Scientific Paper

Ganesh Yallabandi
Department of Nephrology,
Apollo Hospitals, Jubilee Hills,
Hyderabad, 500033, Telangana, India

Veena Mayya*
Department of Information & Communication Technology, Manipal Institute of Technology,
Manipal, Manipal Academy of Higher Education,
Manipal, 576104, Karnataka, India.
veena.mayya@manipal.edu *Corresponding author

Jayakumar Jeganathan
Department of General Medicine,
Kasturba Medical College,
Manipal Academy of Higher Education,
Mangalore - 575001, Karnataka, India

Sowmya Kamath S.
Healthcare Analytics and Language Engineering (HALE) Lab, Department of Information Technology,
National Institute of Technology Karnataka,
Surathkal, Mangalore - 575025, Karnataka, India

Abstract *– Early detection of patient deterioration in the Intensive Care Unit (ICU) can play a crucial role in improving patient outcomes. Conventional severity scales currently used to predict patient deterioration are based on a number of factors, the majority of which consist of multiple investigations. Recent advancements in machine learning (ML) within the healthcare domain offer the potential to alleviate the burden of continuous patient monitoring. In this study, we propose an optimized ML model designed to leverage variations in vital signs observed during the final 24 hours of an ICU stay for outcome predictions. Further, we elucidate the relative contributions of distinct vital parameters to these outcomes The dataset compiled in real-time encompasses six pivotal vital parameters: systolic (0) and diastolic (1) blood pressure, pulse rate (2), respiratory rate (3), oxygen saturation (SpO2) (4), and temperature (5). Of these vital parameters, systolic blood pressure emerges as the most significant predictor associated with mortality prediction. Using a fivefold cross-validation method, several ML classifiers are used to categorize the last 24 hours of time series data after ICU admission into three groups: recovery, death, and intubation. Notably, the optimized Gradient Boosting classifier exhibited the highest performance in detecting mortality, achieving an area under the receiver-operator curve (AUC) of 0.95. Through the integration of electronic health records with this ML software, there is the promise of early notifications regarding adverse outcomes, potentially several hours before the onset of hemodynamic instability.*

Keywords*: Mortality Prediction, Clinical Decision Support Systems, Healthcare Informatics*

1. INTRODUCTION

In critical care applications, the process of taking practical decisions on managing the care of intensive care patients can help augment the efficiency of caregivers, through the use of predictive data analysis on the large amounts of data generated while monitoring these patients. The most important aspect of a clinical decision support system (CDSS) in the ICU is, undoubtedly, its ability to accurately predict in advance the mortality or severity risk of a patient so that doctors and other healthcare personnel can be prepared to intervene in time with the resources available in the ICU. Apart from measuring the severity of illness, mortality prediction can also play a crucial role in the assessment of treatment and critical care policies in a hospital. Hence, ICU mortality prediction has remained a well-researched problem over the years. Detecting the deterioration of patients in the ICU at an early stage has the potential to enhance patient outcomes. In ICUs, conventional severity scores, including the Acute Physiology and Chronic Health Evaluation (APACHE) score and the Simplified Acute Physiology Score [1], have become essential tools for assessing mortality risk. Globally, APACHE-II, SAPS-II, SOFA [2-6] remain the most widely utilized techniques for gauging mortality risk. However, the factors considered and the severity level assigned can vary significantly based on the chosen severity scale. The computation of severity scores relies on laboratory findings and a patient's medical history, and this process is both time-consuming and complex.

Given the limitations of traditional scoring systems, there is a growing interest among researchers in leveraging machine learning (ML) techniques to predict mortality [7]. Various studies, such as those conducted by Wong et al. [8], Johnson et al. [9], and Schuetz et al. [10], have demonstrated the superior performance of ML models compared to conventional severity scores. The mortality prediction algorithm put forth by Pirracchio et al. [11] utilized a set of 17 variables that are present in the SAPS- II score. In their study, Nemati et al. [12] utilized a set of 65 variables computed on an hourly basis and subsequently given to a ML algorithm to forecast the initiation of sepsis. In a different study, Zahid et al. [13] employed a self-normalizing neural network, leveraging over 20 parameters, to foresee the mortality outcomes for patients within the ICU. Another recent investigation by Camacho-Cogollo et al. [14] adopted a distinct approach by employing 31 medically relevant features (MRF) to predict sepsis. These features were meticulously chosen from a pool of 145 potential features, guided by the expert medical insights of a proficient physician. Subsequently, a variety of ML models were tested using this refined set of features. Weissman et al. [15] found that the inclusion of clinical notes along with structured clinical data dramatically improved the ability of ML models to predict ICU mortality. Payrovnaziri et al. [16] used both unstructured (discharge summaries) and structured patient data for performing myocardial infarction based mortality prediction. In a recent study by Huang et al. [17], a novel stacking ensemble model was devised to address the challenge of mortality risk assessment in patients with cerebrovascular conditions. This innovative model made use of multimodal data, integrating various sources, including laboratory test data, structured information, and textual radiology reports. However, the incorporation of these predictive systems into the healthcare realm encounters noteworthy challenges. This is primarily due to the need for a substantial number of features, including intricate laboratory findings, measurements of urine output, evaluations based on the Glasgow Coma Scale (GCS), and even clinical notes. These intricacies create impediments for the practical implementation of these systems, as they necessitate medical personnel to manually input a multitude of parameters to ensure the precision of predictions. Often, this process demands repeated investigations, further contributing to the intricacy and potentially hindering the effective application of these predictive models.

Conducting statistical analysis through bivariable trend models, Churpek et al. [18] determined that vital sign trends play a pivotal role in the detection of critical illness. In the study conducted by Bloch et al. [19], the authors manually selected four important features. These features were determined by analyzing their significance across a range of tested models. The selected features include the median change in heart rate, the number of trend changes in respiratory rate, the minimal change in respiratory rate, and arterial pressure. Recent studies by Baker et al. [20] emphasize the significance of vital signs as influential factors. In their work, they fused convolutional (CNN) layers with bidirectional long short-term memory (BiLSTM) networks to anticipate mortality using statistics characterizing variations in heart rate, blood pressure, respiratory rate, blood oxygen levels, and temperature. They derived a total of 49 statistical features for each of the seven vital signs. It's apparent from the study that prior to employing ML models, it's imperative to compute the statistical properties of the vital signs. The need to perform these computations prior to implementing the ML models introduces complexities that could hinder the seamless application of this approach in a clinical setting.

A substantial amount of time and effort is dedicated to recording vital signs in the ICU. However, there is a scarcity of studies focused solely on recognizing trends derived from these fundamental and straightforward parameters. The importance of vial sign trends in relation to patient outcomes remains relatively underexplored. The primary objective of this study is to identify discernible patterns within essential vital signs, namely blood pressure, respiratory rate, pulse rate, and SpO2. These vital signs have been observed to display correlations with the eventual outcomes of patients in the ICU. This investigation employs optimized ML techniques to achieve this objective and further examines the individual contributions of each vital parameter to these outcomes.

The rest of the paper is organized as follows: Section 2 provides a comprehensive review of the most relevant and effective mortality prediction systems reported in the literature. Section 3 details the data collation process and the methodology used for patients' pattern recognition and correlation identification of vital parameters. Section 4 documents the evaluation of the ML models and details the extensive experiments conducted using ML models trained under different timelines and hyperparameter optimization. Section 5 concludes the proposed experimental study and presents future work.

2. LITERATURE REVIEW

Mortality prediction plays a crucial role in assessing the severity of an illness and aiding in the enhancement of patients' prognoses. In recent years, researchers have focused on designing non-parametric CDSSs built using data mining, ML, and deep learning (DL) techniques to enable higher accuracy for ICU mortality prediction. The majority of these studies have made use of publicly accessible datasets such as the multiparameter intelligent monitoring in intensive care (MIMIC) dataset [21-23], the PhysioNet computing in cardiology challenge dataset [24], the high time-resolution ICU dataset (HiRID) [25], and the women in data science (WiDS) challenge dataset [26]. Through these studies, the versatility and effectiveness of ML in the critical care domain have been convincingly demonstrated. Nevertheless, a notable challenge arises from the complexity of these datasets, each containing more than 20 parameters. This complexity poses a hindrance to the practical implementation of these ML systems, as it requires medical personnel to manually input a multitude of parameters to ensure the accuracy and precision of the predictions. This manual input process can be time-consuming and prone to errors, potentially undermining the overall utility of these predictive systems.

As the field of ML continues to evolve, addressing this challenge is pivotal for achieving seamless integration of these predictive models into real-world clinical settings. Streamlining the data input process, reducing the number of required parameters, or developing automated methods for data extraction could all contribute to enhancing the feasibility and effectiveness of ML applications in critical care scenarios. This ongoing effort to bridge the gap between complex datasets and practical implementation holds the potential to revolutionize the way critical care is managed and optimized. Recent studies utilized automated feature selection [27-30] and reduction [31, 32] techniques to select the important parameters from the recorded ICU datasets. As the number of ICU patients increases, accumulating a large number of parameters becomes increasingly difficult. In this investigation, we conducted a pilot study with only six vital signs that are routinely monitored during ICU stays. According to recent studies, ML models were found to have superior predictive capabilities with structured data input than DL models [17]. Therefore, this study focuses on optimizing ML models for predicting mortality using crucial vital signs.

3. MATERIALS AND METHODS

3.1. PARTICIPANTS AND DATA

The study was conducted in ICUs attached to Kasturba Medical College, Mangalore, and Manipal Academy of Higher Education, Manipal, India. Patients with an age greater than 18 years who are admitted to the ICU from August 2019 to November 2020 and who provide their consent in a written informed form are included in this study. Patients who stayed less than 24 hours in the ICU or who were admitted before August 2019 and those who denied consent are excluded from the study. A total of 285 patients' data were considered for the study. Each patient record includes age, gender, length of ICU stay, outcome (recovered, death, intubated), and the last 24 hours of time series data. Time series data includes six vital parameters: systolic (0), diastolic (1), pulse rate (2), respiratory rate (3), SpO2 (4), and temperature (5). Fig. 1 illustrates the comprehensive procedure employed during data collection, with the utilization of the Philips mp20 monitor for recording vital signs. The distribution of the ICU patient data is depicted in Fig. 2. It can be observed that the data is highly imbalanced in terms of recorded patient outcomes.

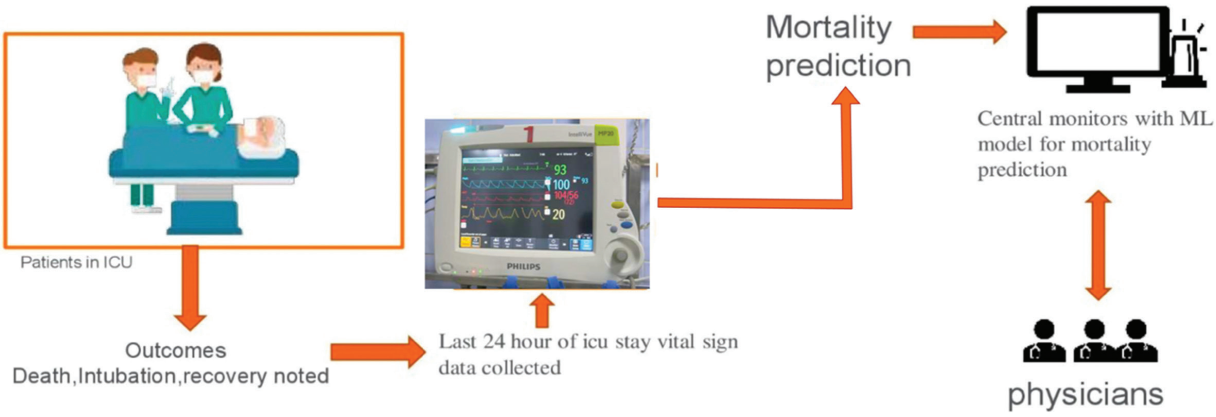

Patients in ICU
Outcomes
Death,Intubation,recovery noted
Last 24 hour of icu stay vital sign data collected
Mortality prediction
Central monitors with ML model for mortality prediction
physicians

Fig 1. Structure of the Proposed CVD Prediction System

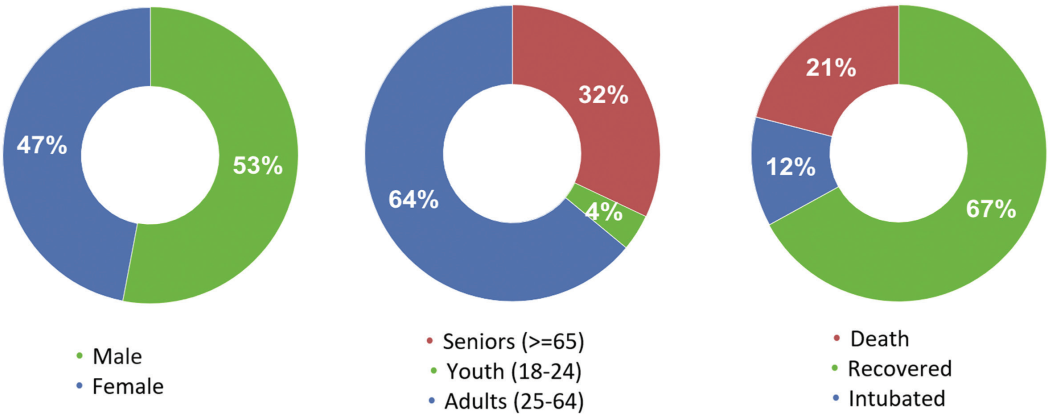


Fig 2. Patient Data Analysis: (a)Gender-wise distribution (b)Age-wise distribution (c)Outcome

3.2. METHODOLOGY

Due to the fact that a subset of rows (about 250) were marked as ”not recorded” (NR), these cases were imputed using the ”backward filling” method, which used the set of recorded patient data that came before it. The dataset encompassed a total of 285 patient records, each encompassing six vital parameters, and spanned a continuous 24-hour period. This data was organized in a structured format (285, 24, 6). Because ML models work best with one-dimensional (1D) data, the data for each patient record, which was a sequence of 24 values for each of six vitals, was put into a linear format with a row major layout, which led to 144 features. As a result of this reformatting, a distinct 1D dataset was generated for each individual patient, facilitating compatibility with the ML algorithms.

Ten different ML classifiers, including K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), AdaBoost, RUSBoost, Random Forest (RF), Decision Tree (DT), Gradient Boosting (GB), XGBoost classifier, and textRNN, were put through a thorough evaluation and fine-tuning process. KNN is a simple and intuitive algorithm that classifies data records based on the majority class among their k-nearest neighbours in the feature space. It measures distances between data points and assigns labels based on the neighbours' labels. SVM seeks to find a hyperplane that best separates different classes of data. SVM aims to maximize the margin between classes and can handle both linear and non-linear separation. MLP consists of multiple layers of interconnected nodes (neurons). It is capable of learning complex relationships in data through forward and backward propagation of signals, making it suitable for a wide range of tasks. AdaBoost (Adaptive Boosting) is an ensemble learning technique that combines the outputs of multiple weak classifiers to create a stronger overall classifier. It assigns higher weights to misclassified instances in each iteration to improve classification performance. RUSBoost (Random Under- Sampling Boosting) is a variant of AdaBoost that incorporates random undersampling of the majority class. This helps address the class imbalance in the dataset, making it particularly useful for imbalanced data scenarios. RF is an ensemble method that constructs multiple DTs during training and combines their outputs to make predictions. It improves accuracy and reduces overfitting by introducing randomness in the tree-building process. A DT is a hierarchical structure that recursively splits data into subsets based on the values of input features. It makes decisions by traversing the tree from the root node to a leaf node, where the final classification is determined. Gradient Boosting is another ensemble method that builds a strong model by sequentially adding weak learners (usually DTs) and focusing on instances that were previously misclassified. It aims to minimize the prediction error iteratively. XG Boost (Extreme GB) is an optimized and highly efficient implementation of GB. It includes regularization techniques, handling missing values, and parallel processing to enhance performance and predictive accuracy. In order to configure the TextRNN model’s hyperparameters, a combination of empirical observations and systematic experimentation was utilized. Through careful calibration of multiple variables, it was determined that implementing GRU (Gated Recurrent Unit) units in a two-layer configuration produced significantly enhanced outcomes.

The primary objective was to effectively categorize the input ICU time series dataset into 3 discrete classes: recovery, mortality, and intubation. Initially, each individual vital parameter is used to predict the outcomes. Further, all possible combinations of vital signs were experimented to determine the most crucial vital parameters. Since the data is highly imbalanced, combining the weak classifiers would improve the performance of the model. So the experiment was conducted using boosting classifiers.

3.3. EXPERIMENTS AND RESULTS

We evaluated the ML classifiers using a five-fold cross-validation method. Within this approach, a single fold was dedicated to testing, while the remaining folds were employed for training the classification model. This process was reiterated across all folds to ensure a uniform and stable performance evaluation. The Python open-source ML packages [33] were used for carrying out the experiments. Initially, the default values as set by the Python packages were set for network parameters. The results obtained are listed in Table 1. In our efforts to enhance performance, we endeavored to ensemble the top three performing ML models (RF, AdaBoost, and Gradient Boost) using voting and stacking algorithms. However, the process of ensembling did not yield a substantial improvement in overall performance. This could potentially indicate that the inherent ensembling nature of GB already integrated the advantages offered by ensembling with different classifiers.

Table 1. Mortality prediction results

Classifier	Precision	Recall	F1-score
TextCNN	0.51	0.57	0.54
RUSBoost	0.61	0.64	0.62
MLP	0.68	0.67	0.67
KNN	0.68	0.73	0.66
DT	0.68	0.69	0.69
AdaBoost	0.69	0.72	0.70
RF	0.73	0.75	0.73
XGBoost	0.74	0.76	0.74
SVM	0.65	0.74	0.68
GB	0.75	0.77	0.75

Given the superior performance of tree-based algorithms in comparison to other ML models, we meticulously refined their hyperparameters through the utilization of the particle swarm optimizaiton (PSO) algorithm [34–36]. Notably, we observed that certain hyperparameters—namely, n_estimators, max_depth, min_samples_leaf, max_features, and min_samples_split held significant importance across the spectrum of tree-based ML models. After achieving refined parameters using the PSO algorithm for the RF model, these fine-tuned hyperparameters were applied as initializations for the remaining ML models. Remarkably, it was discerned that the optimized hyperparameters from the initial RF tuning yielded the best performance across the other ML models as well. The hyperparameter ranges and the corresponding optimal values determined by the PSO approach have been detailed in Table 2. The outcomes achieved through this optimization process, encompassing the refined hyperparameters, have been documented in Table 3. Notably, to ensure a fair comparison, the same random state was upheld throughout the experiments. Evidently, the proposed optimization methodology led to a significant performance enhancement, with improvements of up to 10%.

Table 2. Hyper-parameters Range

Hyperparameter	Lower bound	Upper bound	PSO chosen value
n_estimators	10	200	168
max_features	1	20	8
max_depth	2	20	10
min_samples_split	2	20	10
min_samples_leaf	1	20	1

Table 3. Optimized classifiers adopted for mortality prediction

Classifier	Precision	Recall	F1-score
RUSBoost	0.68	0.68	0.68
DT	0.69	0.71	0.70
AdaBoost	0.76	0.78	0.77
RF	0.80	0.81	0.79
XGBoost	0.75	0.77	0.76
GB	0.81	0.82	0.80

Due to the minimal variance observed in body temperature among the studied cases within the last 24 hours after admission, its contribution to predictive modelling was limited. Consequently, the focus shifted to the remaining vital parameters for subsequent experimentation, aimed at comprehending the individual significance of each parameter on performance. In this context, a minimum of three vital parameters were selected at a time for classification using the GB classifier. A comprehensive set of 16 combinations was tested, and the corresponding accuracies for each combination are detailed in Table 4. Notably, the optimized models showcased enhancements in the classifier's performance. While the combination of systolic blood pressure, pulse rate, and SpO2 demonstrated potential for achieving higher accuracy on its own, it was observed that incorporating all vital parameters led to improvements not only in accuracy but also in precision.

Table 4. Mortality prediction with Adaboost classifier using combinations of vitals

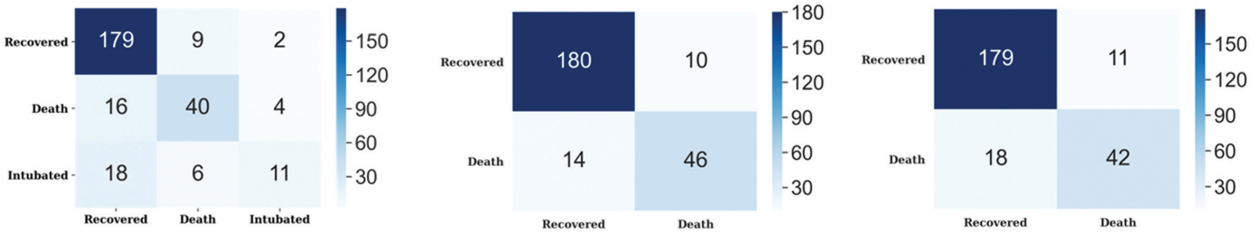
Combination of vitals	Accuracy	Precision	Recall	F1-score
systolic(0)	0.72	0.67	0.72	0.68
diastolic(1)	0.68	0.61	0.68	0.63
pulse rate(2)	0.71	0.66	0.71	0.68
respiratory rate(3)	0.69	0.64	0.69	0.65
SpO2(4)	0.71	0.67	0.71	0.68
(0,1,2)	0.74	0.67	0.74	0.69
(0,1,3)	0.74	0.69	0.74	0.70
(0,1,4)	0.76	0.73	0.76	0.74
(0,2,3)	0.75	0.70	0.75	0.71
(0,2,4)	0.82	0.80	0.80	0.80
(0,3,4)	0.78	0.75	0.78	0.76
(1,2,3)	0.74	0.69	0.74	0.70
(1,2,4)	0.78	0.76	0.78	0.77
(1,3,4)	0.75	0.72	0.75	0.73
(2,3,4)	0.79	0.77	0.79	0.77
(0,1,2,3)	0.75	0.71	0.75	0.71
(0,1,2,4)	0.79	0.77	0.79	0.78
(0,1,3,4)	0.76	0.73	0.76	0.74
(0,2,3,4)	0.82	0.80	0.81	0.80
(1,2,3,4)	0.79	0.77	0.79	0.77
(0,1,2,3,4)	0.82	0.81	0.82	0.80

Fig. 3(a) illustrates the confusion matrix derived from the predictions generated by employing the combination of all vital parameters with the optimized GB classifier. Notably, the model demonstrates a substantial capacity to accurately distinguish between recovered patients and those at risk of mortality. However, due to the study's limited inclusion of only 13% of intubated patients, the model struggled to discern the intricate patterns necessary for precise classification. To address this limitation, further analysis was undertaken with a focus on binary classification into two distinct categories: the possibility of death or survival. In pursuit of this, all data pertaining to intubated patients was excluded. The outcome of this refined classification strategy is depicted in Fig. 3(b), showcasing an impressive 90% accuracy achieved through the utilization of the proposed optimized GB classifier.

The study achieved remarkable results in binary classification using an optimized GB classifier. Precision, recall, accuracy, and F1 score were all recorded at 0.90, indicating a high level of performance. In the context of tree-based ML models, the process of identifying mortality instances can be understood by examining the DTs employed. The visualization of these DTs can be seen in Fig. 4. The data for each individual record, denoted as X in Fig. 4, is organized in a specific order: systolic blood pressure (0), diastolic blood pressure (1), pulse rate (2), respiratory rate (3), and SpO2 (4). Each parameter has associated data collected over a 24-hour period. By analyzing the DTs, it is possible to pinpoint the exact hour and parameter that contributed to a specific decision. This provides valuable insights into the factors influencing the classification outcome.

The ROC curve, presented in Fig. 5(a), showcases the performance of the optimized GB binary classifier. This curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR), with TPR represented on the y-axis and FPR on the x-axis. Remarkably, the area under the ROC curve, which amounts to 0.95, signifies the model's ability to make accurate predictions approximately 95% of the time based on the last 24 hours of vital sign data.

For a more detailed analysis, we conducted an assessment by excluding vital sign data from the 12 hours immediately preceding the outcome within the last 24 hours. Fig. 3(c) represents the confusion matrix derived from predictions made by the optimized GB binary classifier utilizing combinations of all vital parameters. Impressively, the model continues to exhibit substantial accuracy in distinguishing between recovered patients and those potentially facing mortality. Illustrated in Fig. 5(b), the area under the ROC curve is presented, plotting the TPR against the FPR.



(a) (b) (c)

Fig. 3. Confusion matrix obtained for (a)all classes (b)two classes (c)two classes for 12 hours data

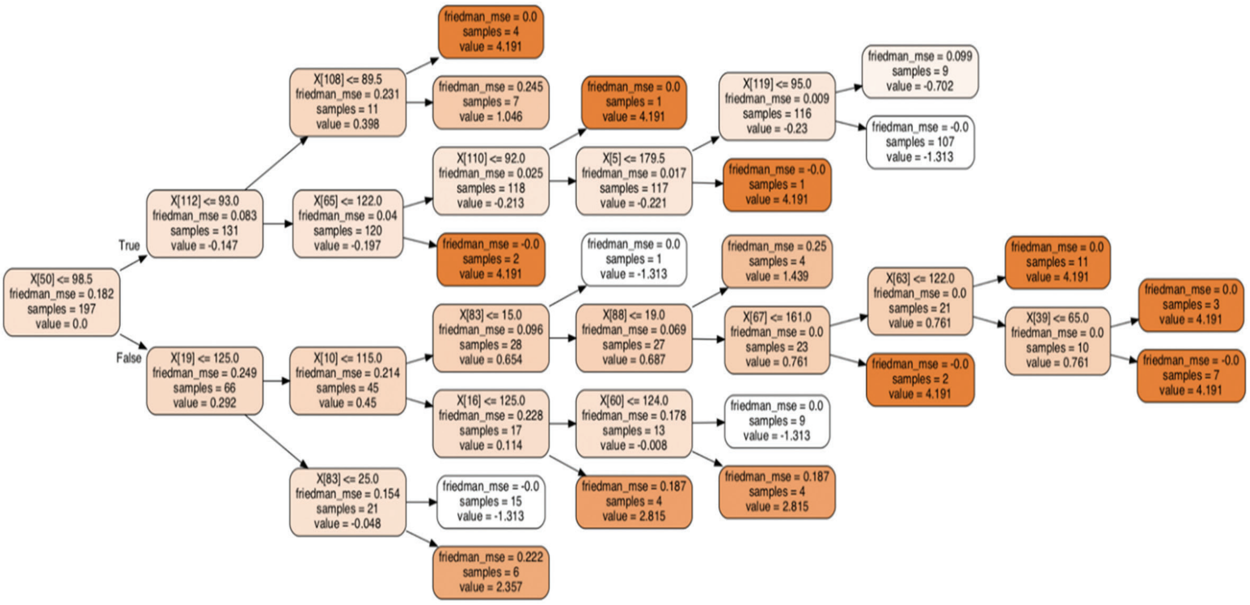
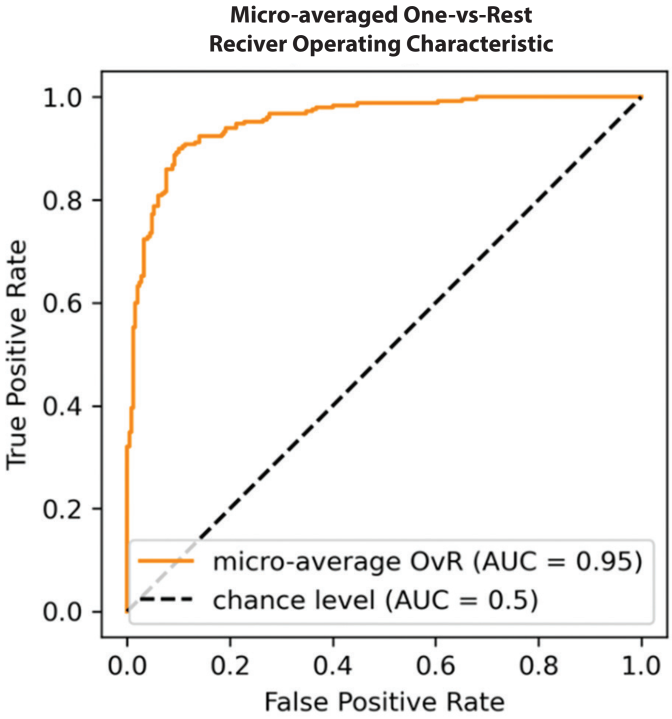
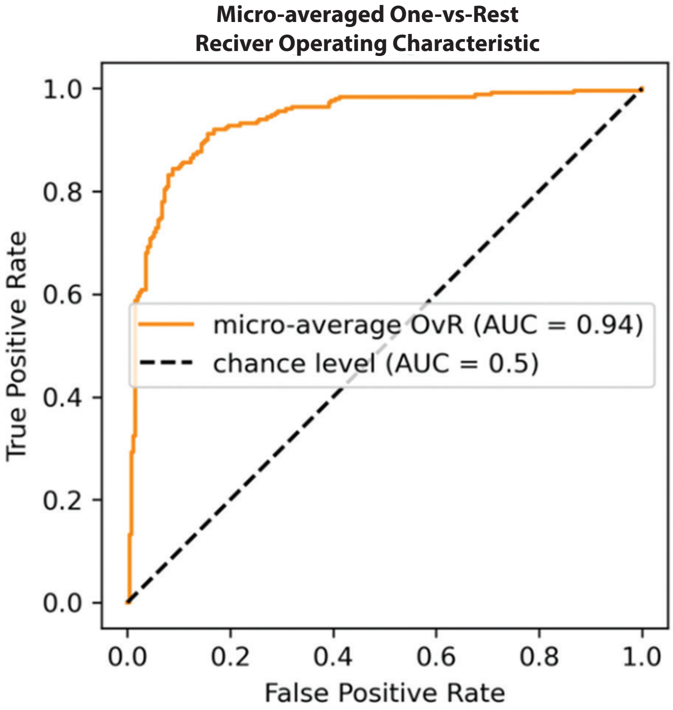


Fig. 4. The decision tree for mortality detection



(a)



(b)

Fig. 5. ROC curve obtained for mortality detection (a) for 24 hours input (b) for 12 hours input

The performance mirrors the outcomes achieved with the 24-hour data window, showcasing the classifier's consistency and ability to predict mortality effectively. In Table 5, a comprehensive summary of precision, recall, F1 score, and accuracy for the initial 12-hour window preceding the actual outcome is provided. The average precision, recall, and F1 score are detailed in the same table, reflecting the model's performance. Notably, the optimized GB classifier achieves 87% accuracy in predicting mortality, even when utilizing data collected during the first 12 hours before the actual outcome. This result underscores the classifier's robust performance across this critical early time frame. In essence, the proposed optimized GB classifier showcases commendable performance across both 12-hour and 24-hour time windows, attesting to its proficiency in forecasting outcomes.

Table 5. Performance obtained for 12 hours data

Metric	Recovered	Death	Weighted average
Precision	0.90	0.76	0.87
Recall	0.93	0.68	0.87
F1 Score	0.92	0.72	0.87
Accuracy	-	-	0.87

4. DISCUSSION

In this work, we have proposed an optimized model for predicting hospital mortality in ICU patients, designed to be especially applicable and beneficial in low- and middle-income countries. Our approach centers on utilizing a streamlined set of variables that are both readily accessible and straightforward to collect. Notably, the model that exhibited the highest performance in our investigation was a GB classifier, which harnesses the strength of an ensemble of weak classifiers, requiring solely vital sign data. These are routinely measured and effortlessly acquired within an ICU setting. Significantly, these variables do not necessitate knowledge of the patient's diagnosis or laboratory results.

Our findings closely align with those of a study conducted by Alistair et al. [37]. In their work, Alistair et al. [37] developed a model relying on a staggering 148 distinct variables, the majority of which emanate from intricate laboratory results, urine output, and Glasgow Coma Scale (GCS) measurements—a collection of vital parameters. This approach, however, introduces complexities that can hinder the model's practical utility. Medical practitioners would need to measure and input an extensive array of factors for an accurate prognosis, which not only consumes time but also necessitates repeated investigations. In contrast, our study offers a user-friendly solution devoid of labour-intensive data entry or cumbersome investigations during the ICU admission process. The optimized GB classifier showcased notable performance even when using vital sign data gathered 12 and 24 hours before the outcome. While our findings are encouraging, further research is required to explore the impact of more frequent intervals for vital sign data collection, aiming to enhance the accuracy of ML models in predicting outcomes.

4.1. LIMITATIONS

The performance of the ML model in identifying intubation outcomes among patients in our study was limited, possibly due to inadequate training on a large dataset. Additionally, our study did not encompass data from post-intubated patients. Furthermore, patients who were recovered and discharged from the ICU were not monitored until their discharge, potentially impacting the model's predictive capability. The low variance in temperature observed within the last 24 hours may have hindered its contribution to prediction, and a more frequent temperature monitoring interval would be necessary for a comprehensive generalization of findings. Notably, within our study, systolic blood pressure emerged as the most influential vital parameter for mortality prediction. However, it's important to acknowledge that blood pressure fluctuations are often evident in patients facing terminal cardiovascular collapse. The unique attribution of critical parameters by the ML model to each feature might be specific to our institution. Despite accounting for patient variability in our hospital and study, external and independent evaluations are imperative to validate the findings. It's crucial to underscore that our results are reflective of a single hospital population, whereas other studies have drawn insights from patient data collected across multiple hospitals. Therefore, external validation using data from a distinct institution is pivotal before broad conclusions can be drawn. Our research employed a smaller database compared to previous studies that aimed to make extensive population-level generalizations. Being retrospective in nature, our data gathering process could benefit from oversight post-results. While our dataset primarily captures the last 24 hours of a patient's ICU stay, an approach involving data collected throughout the entire ICU stay and analyzed across different time windows could enhance model accuracy. The exploration of DL techniques [38, 39], often deemed superior to supervised learning techniques, is an avenue worth exploring in future studies. However, despite experimenting with state-of-the-art DL models, the limitations posed by our data's scope led to relatively lower performance with DL models. To realize practical applicability, the developed ML model must undergo testing in a real-world environment.

5. CONCLUSIONS

Among the vital parameters studied, systolic blood pressure emerged as the most significant predictor linked to mortality prediction, underscoring its pivotal role. Additionally, SpO2 and pulse rate exhibited notable associations with predictive outcomes. Conversely, temperature variance exhibited a limited contribution to predicting outcomes in this study, potentially due to its low variability. An intriguing observation is the potential of a combination of systolic blood pressure, pulse rate, and SpO2 to yield enhanced accuracy. Moreover, incorporating all vital parameters not only enhances accuracy but also improves precision. To further refine the model's accuracy in predicting intubated patients, additional training with larger databases is essential.

Remarkably, the proposed optimized GB classifier achieves 90% accuracy in predicting recovery or mortality, utilizing a mere six recorded vital parameters. These findings suggest the viability of employing ML techniques for routine monitoring of ICU patients. This presents an opportunity to evolve beyond traditional prognostic scores like APACHE, SAPS, and SOFA and integrate more accessible and less intricate ML techniques that rely solely on essential vital parameters. Regular implementation of such techniques can serve as a valuable supplement to conventional ICU scoring systems. As ICU technologies progress towards greater automation, integrating central monitors with ML software could provide an early warning system, preemptively alerting healthcare providers to potential adverse outcomes hours before hemodynamic instability manifests.

ETHICAL APPROVAL

Ethical approval for this study was obtained from Institutional Ethics Committee Kasturba Medical College, Mangaluru (ID IEC KMC MLR 0919/445).

DATA AVAILABILITY

The data are not publicly available due to privacy and ethical concerns. The data that support the findings of this study are available from the first author, Ganesh Y., upon reasonable request.

6. REFERENCES

[1] J.-L. Vincent et al. "The SOFA score to describe organ dysfunction/failure", Intensive Care Medicine, Vol. 22, No. 7, 1996, pp. 707-710.

[2] W. Knaus, E. Draper, D. Wagner, J. Zimmerman, "Apache ii: A severity of disease classification system", Critical Care Medicine, Vol. 13, No. 10, 1985, p. 818-829.

[3] W. Knaus et al. "The APACHE III prognostic system: Risk prediction of hospital mortality for critically ill hospitalized adults", Chest, Vol. 100, No. 6, 1991.

[4] J. E. Zimmerman, A. A. Kramer, D. S. McNair, F. M. Malila, "Acute physiology and chronic health evaluation (Apache) iv: Hospital mortality assessment for today's critically ill patients", Critical Care Medicine, Vol. 34, No. 5, 2006.

[5] J.-R. Le Gall, P. Loirat, A. Alperovitch, P. Glaser, C. Granthil, D. Mathieu, P. Mercier, R. Thomas, D. Villers, "A simplified acute physiology score for ICU patients", Critical Care Medicine, Vol. 12, No. 11, 1984, pp. 975-977.

[6] J.-R. Gall, S. Lemeshow, F. Saulnier, "A new simplified acute physiology score (SAPS II) based on a European/North American multi-center study", Journal of the American Medical Association, Vol. 270, No. 24, 1993.

[7] A. Naemi et al. "Machine learning techniques for mortality prediction in emergency departments: a systematic review", BMJ Open, Vol. 11, No. 11, 2021.

[8] L. Wong, J. Young, "A comparison of ICU mortality prediction using the Apache II scoring system and artificial neural networks", Anaesthesia, Vol. 54, No. 11, 1999, pp. 1048-1054.

[9] A. Johnson, A. Kramer, G. Clifford, "A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy", Critical Care Medicine, Vol. 41, 2013.

[10] P. Schuetz et al. "Effect of procalcitonin-guided antibiotic treatment on mortality in acute respiratory infections: a patient level meta-analysis", The Lancet Infectious Diseases, Vol. 18, 2018.

[11] R. Pirracchio, M. L. Petersen, M. Carone, M. R. Rigon, S. Chevret, M. J. van der Laan, "Mortality prediction in intensive care units with the super ICU learner algorithm (sicula): a population-based study." The Lancet. Respiratory medicine, Vol. 31, 2015, pp. 42-52.

[12] S. Nemati, A. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, T. G. Buchman, "An interpretable machine learning model for accurate prediction of sepsis in the ICU", Critical Care Medicine, Vol. 46, No. 4, 2018, pp. 547-553.

[13] M. Zahid, J. Lee, "Mortality prediction with self normalizing neural networks in intensive care unit patients", Proceedings of the IEEE EMBS International Conference on Biomedical and Health Informatics, Las Vegas, NV, USA, 4-7 March 2018, pp. 226-229.

[14] J. E. Camacho-Cogollo, I. Bonet, B. Gil, E. Iadanza, "Machine learning models for early prediction of sepsis on large healthcare datasets", Electronics, Vol. 11, No. 9, 2022.

[15] G. Weissman, R. Hubbard, L. Ungar, M. Harhay, C. Greene, B. Himes, S. Halpern, "Inclusion of

unstructured clinical text improves early prediction of death or prolonged ICU stay", Critical Care Medicine, Vol. 46, No. 7, 2018, pp. 1125–1132.

[16] S. N. Payrovnaziri, L. A. Barrett, D. Bis, J. Bian, Z. He, "Enhancing prediction models for one-year mortality in patients with acute myocardial infarction and post myocardial infarction syndrome", Studies in health technology and informatics, Vol. 264, 2019, pp. 273–277.

[17] R. Huang et al. "Stroke mortality prediction based on ensemble learning and the combination of structured and textual data", Computers in Biology and Medicine, Vol. 155, 2023.

[18] M. M. Churpek, R. Adhikari, D. P. Edelson, "The value of vital sign trends for detecting clinical deterioration on the wards." Resuscitation, Vol. 102, 2016, pp. 1-5.

[19] E. Bloch, T. Rotem, J. Cohen, P. Singer, Y. Aperstein, "Machine learning models analysis of vital signs dynamics: A case for sepsis onset prediction", Journal of Healthcare Engineering, Vol. 2019, 2019.

[20] S. B. Baker, W. Xiang, I. M. Atkinson, "Continuous and automatic mortality risk prediction using vital signs in the intensive care unit: a hybrid neural network approach", Scientific Reports, Vol. 10, No. 1, 2020.

[21] M. Saeed et al. "Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database", Critical care medicine, Vol. 39, 2011.

[22] A. E. Johnson et al. "Mimic-iii, a freely accessible critical care database", Scientific Data, Vol. 3, 2016.

[23] A. E. Johnson et al. "Mimic-iv, a freely accessible electronic health record dataset", Scientific Data, Vol. 10, 2023.

[24] I. Silva, G. Moody, D. J. Scott, L. A. Celi, R. G. Mark, "Predicting in-hospital mortality of ICU patients: The physionet/computing in cardiology challenge 2012", in Computing in Cardiology, Vol. 39, 2012, pp. 245-248.

[25] S. L. Hyland et al. "Early prediction of circulatory failure in the intensive care unit using machine learning", Nature Medicine, Vol. 26, 2020.

[26] M. Lee et al. "WIDS datathon 2020: ICU mortality prediction", PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals, 2020.

[27] C. Steinmeyer, L. Wiese, "Sampling methods and feature selection for mortality prediction with neural networks", Journal of Biomedical Informatics, Vol. 111, 2020.

[28] G. S. Krishnan, S. Kamath, "A novel GA-ELM model for patient-specific mortality prediction over large-scale lab event data", Applied Soft Computing, Vol. 80, 2019, pp. 525-533.

[29] Y.-C. Huang, K.-Y. Chen, S.-J. Li, C.-K. Liu, Y.-C. Lin, M. Chen, "Implementing an ensemble learning model with feature selection to predict mortality among patients who underwent three-vessel percutaneous coronary intervention", Applied Sciences, Vol. 12, No. 16, 2022.

[30] Q. Wang, G. Chen, X. Jin, S. Ren, G. Wang, L. Cao, Y. Xia, "Bit-mac: Mortality prediction by bidirectional time and multi-feature attention coupled network on multivariate irregular time series", Computers in Biology and Medicine, Vol. 155, 2023, p. 106586.

[31] A. Kline, T. Kline, Z. S. Hossein Abad, J. Lee, "Novel feature selection for artificial intelligence using item response theory for mortality prediction", Proceedings of the 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, Montreal, QC, Canada, 20-24 July 2020, pp. 5729-5732.

[32] B. Srimedha, R. Naveen Raj, V. Mayya, "A comprehensive machine learning based pipeline for an accurate early prediction of sepsis in ICU", IEEE Access, Vol. 10, 2022, pp. 105120-105132.

[33] F. Pedregosa et al. "Scikit-learn: Machine learning in python", Journal of Machine Learning Research, Vol. 12, 2011, pp. 2825-2830.

[34] J. Kennedy, "Particle Swarm Optimization", Springer US, 2010, pp. 760-766.

[35] M. Clerc, J. Kennedy, "The particle swarm - explosion, stability, and convergence in a multi-dimensional complex space", IEEE Transactions on Evolutionary Computation, Vol. 6, 2002, pp. 58-73.

[36] P. R. Lorenzo, J. Nalepa, M. Kawulok, L. S. Ramos, J. R. Pastor, “Particle swarm optimization for hyperparameter selection in deep neural networks”, Proceedings of the Genetic and Evolutionary Computation Conference, July 2017, p. 481-488.

[37] A. E. W. Johnson, R. G. Mark, “Real-time mortality prediction in the intensive care unit”, AMIA Annual Symposium Proceedings, 2017, pp. 994-1003.

[38] S. Purushotham, C. Meng, Z. Che, Y. Liu, “Benchmarking deep learning models on large healthcare datasets”, Journal of Biomedical Informatics, Vol. 83, 2018, pp. 112-134.

[39] Z. Che, S. Purushotham, K. Cho, D. Sontag, Y. Liu, “Recurrent neural networks for multivariate time series with missing values”, Scientific Reports, Vol. 8, No. 1, 2018, p. 6085.

Trust And Energy-Aware Routing Protocol for Wireless Sensor Networks Based on Secure Routing

Original Scientific Paper

Muneeswari G
School of Computer Science and Engineering,
VIT-AP University, Amaravati,
Andhra Pradesh, India
muneeswari.g@vitap.ac.in

Ahilan A
Department of Electronics and Communication Engineering, PSN College of Engineering and Technology, Tirunelveli, Tamil Nadu, India
listentoahil@gmail.com

Rajeshwari R
Department of Information technology,
Panimalar Engineering College,
Chennai, Tamil Nadu, India
rajeshwariit@gmail.com

Kannan K
Electronics and Communication Engineering,
R.M.K. College Of Engineering And Technology,
Puduvoyal,Chennai Tamil Nadu, India.
Kannan@rmkcet.ac.in

John Clement Singh C
Department of Electronics and Communication Engineering,
Kings Engineering College,
Sriperumbudur, Chennai, Tamil Nadu, India
Johnclement12@gmail.com

Abstract – *Wireless Sensor Network (WSN) is a network area that includes a large number of nodes and the ability of wireless transmission. WSNs are frequently employed for vital applications in which security and dependability are of utmost concern. The main objective of the proposed method is to design a WSN to maximize network longevity while minimizing power usage. In a WSN, trust management is employed to encourage node collaboration, which is crucial for achieving dependable transmission. In this research, a novel Trust and Energy Aware Routing Protocol (TEARP) in wireless sensors networks is proposed, which use blockchain technology to maintain the identity of the Sensor Nodes (SNs) and Aggregator Nodes (ANs). The proposed TEARP technique provides a thorough trust value for nodes based on their direct trust values and the filtering mechanisms generate the indirect trust values. Further, an enhanced threshold technique is employed to identify the most appropriate clustering heads based on dynamic changes in the extensive trust values and residual energy of the networks. Lastly, cluster heads should be routed in a secure manner using a Sand Cat Swarm Optimization Algorithm (SCSOA). The proposed method has been evaluated using specific parameters such as Network Lifetime, Residual Energy, Throughpu,t Packet Delivery Ratio, and Detection Accuracy respectively. The proposed TEARP method improves the network lifetime by 39.64%, 33.05%, and 27.16%, compared with Energy-efficient and Secure Routing (ESR), Multi-Objective nature-inspired algorithm based on Shuffled frog-leaping algorithm and Firefly Algorithm (MOSFA) , and Optimal Support Vector Machine (OSVM).*

Keywords*: Wireless Sensor Network, Routing, Sensor Nodes, Aggregator Nodes, sand cat swarm optimization algorithm*

1. INTRODUCTION

Wireless Sensor Network (WSN) consists of a few co-operative sensor nodes that are spread out geographically. As a result of technological advances in wireless networking techniques and the availability of inexpensive, intelligent, and small-sized sensors, ubiquitous computing has been made possible [1]. The goal of a WSN implementation is to gather data about objects found in the monitoring area, transform that data into electrical signals, and transmit those signals to the base station through wireless multi-channel communication [2]. The sensor nodes join together to create a network in order to gather information from their immediate surroundings and then communicate with one another to carry out specific tasks [3]. During Mobility Wireless Sensor Networks (MWSN), sensors are mobile and can link to a variety of providers, such as robotic systems and intelligent modes of transportation, to detect and collect data that can then be transmitted to the BS via direct or multi-hop communication models [4]. Unstructured WSNs consist of ad-hoc deployments of dense sensor nodes. The network is called a structured WSN, depending on the other extreme, when all nodes are placed simultaneously [5]. In Sensor nodes' the computing capacity, power, and the battery life are all constrained in WSN. The topology of the network changes when certain network nodes lose power. The network may even become paralyzed and cease to operate correctly if there are too many dead nodes. Due to the inability to detect malicious sites, attacks and energy usage are two issues wireless sensor networks face. They use a particular routing method, they use effort effectively, they choose cluster heads, and the technology they use to create a wireless sensor network are all important considerations.

WSN is one of the most contemporary communication-related technologies. Due to its open architecture and limited resource availability, WSN is challenging to secure and utilize energy effectively. Routing and clustering are just two of the many technologies that have been introduced to secure WSNs. Many reasons, including shortened sensor node usage time, increased power consumption due to larger number of hops, distribution fewer packet distribution, and decreased throughput, may result in improper data transmission from one node to another. This research proposes a novel Trust and Energy Aware Routing Protocol (TEARP) technique, which enhances the security of routes using wireless sensor networks. The major contributions of the proposed TEARP techniques are given as follows.

- Initialization, registration, and authentication are accomplished during the authentication of ANs and SNs on public and private blockchains, respectively.
- The proposed technique provides a thorough trust value for nodes based on their direct trust values while taking volatility and adaptable penalty elements into concern. Filtering mechanisms also generate indirect trust values.
- Further, an enhanced threshold technique is employed to identify the most appropriate clustering heads based on dynamic changes in the extensive trust values and residual energy of the networks.
- Lastly, cluster heads should be routed in a secure manner is determined using a sand cat swarm optimization algorithm.

The remainder of the research is organized as follows. In Section II, a summary of the literature is provided. In Section III, the proposed TEARP methodology is thoroughly explained. The experimental findings are presented in Section IV, and conclusions and future scope is presented in Section V.

1.1. BACKGROUND STUDY

An improved Artificial Bee Colony (iABC) metaheuristic is presented in [6] to maintain a solid balance between mining and exploration abilities while using the least amount of RAM possible. The suggested metadata's abilities to produce ideal cluster heads and increase WSN energy efficiency are inherited by an energy-efficient bee clustering algorithm based on iABC information.

An improved version of the firefly algorithm is presented in [7] which is applied to improve the network lifetimeWhen LEACH, the basic Firefly set of rules, and particle swarm optimization are applied to the same community infrastructure model, the performance of the improved Firefly set of rules is compared to them. In terms of performance and stability, the enhanced Firefly approach is superior than existing algorithms.

A Whale Moth Flame Optimization (WMFO) and Improved African Buffalo Optimization (IABO) is presented in [8] which is applied for effective clustering and routing. The WMFO method can be utilized for effective clustering by employing a fitness function connected to the distance within the cluster, the distance between clusters, the energy, and the equilibrium coefficient. The WMFO algorithm creates a tuning function that contains certain factors like residual energy and distance coefficient in order to choose the best routes in the WSN.

A Particle Distance Updated Sea Lion Optimization (PDU-SLnO) is presented in [9] which is developed to consume less energy consumption and increases the network lifetime. For the WSN, a new hierarchical routing energy-sensitive CH selection architecture is provided using the hybrid optimization technique. When selecting a CH, capacity, distance, latency, and quality of service (QoS) are taken into account. To choose the optimal CH, the Sea Lion Optimization (SLnO) and Particle Swarm Optimization (PSO) algorithm principles are integrated in the new matching method known as the PDU-SLnO algorithm.

2. LITERATURE SURVEY:

The WSN performance, including energy use, network lifetime, etc., has been the subject of many researches. One of the most important characteristics of WSNs is secure routing. Among those, some of the techniques have been reviewed in this section.

In 2019, Haseeb K., et al [10] presenting an energy-efficient and secure routing (ESR) protocol for intrusion defence in IoT based on wireless sensor networks. The proposed solution utilized greedy algorithms to construct routing paths and overlooked intrusions in an infrastructure-less and unattended environment. As a result, there are a great No. of route discovery and re-transmissions, especially when there are attacker networks present and there is a great deal of internet traffic.

In 2020, Barzin et al. [11] Presented a Multi-objective nature-inspired algorithm (MOSFA) is developed from fireflies and shuffling frog-leaping algorithms and is a successful protocol for WSNs. Using this technique, both fireflies and shuffled frog-leaping algorithms are utilized simultaneously. SIF, ERA, FSFLA, and LEACH have common lifespan development of 68%, 82%, 30%, and 28%, respectively, according to simulation data.

In 2021, Amaran S., et al [12] presented a novel optimal Support Vector Machine (OSVM) based IDS in WSN. The suggested technique's OSVM model has an accuracy of more than 94.09% and a detection rate of 95.02%.In 2021, Reddy D.L. et al [13] Presented a hybrid Ant Colony Optimization (ACO) approach that integrates Glow Worm Swarm Optimization. According to experimental results, the suggested solution keeps more nodes alive and uses less network energy than standard techniques.

From the aforementioned analyses, it's clear that those solutions have several hazards, including the nodes' consumption of electricity and steady routing when transferring data packets to their destinations.To overcome these drawbacks, novel Trust and Energy Aware Routing Protocol (TEARP) techniques are recommended.

3. PROPOSED METHOD

In this paper, a Trust and Energy Aware Routing Protocol (TEARP) in WSN is proposed, which use blockchain technology to maintain the identity of the Sensor Nodes (SNs)and Aggregator Nodes (ANs). Initialization, registration, and authentication are accomplished during the authentication of ANs and SNs on public and private blockchains, respectively. Fig. 1 illustrates the overall structure of the proposed method.

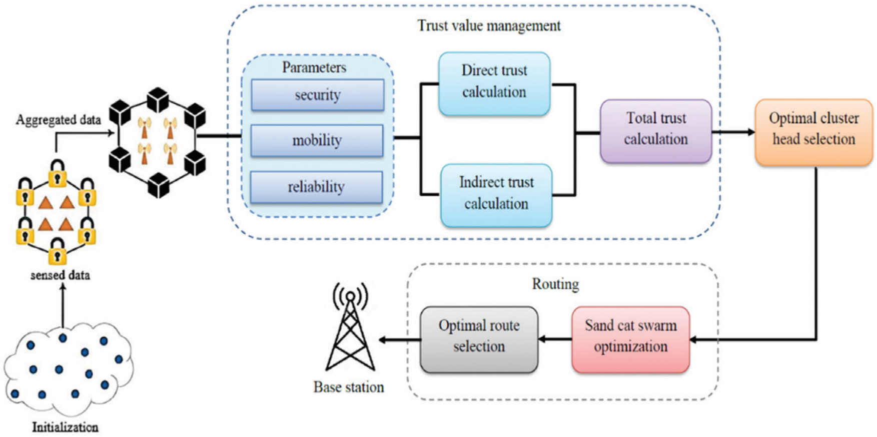


Fig. 1. Overall block diagram for the proposed TEA method

3.1. BLOCKCHAIN TECHNOLOGY

In the proposed blockchain-based routing and reliability evaluation method, BSs transmit encrypted data about routing and trust values to other network nodes. All node-to-node transactions are also verified by the blockchain's. The AN authenticates and authorizes the SN each time they communicate, allowing the SN to send packets to the AN. Additionally, these BSs authenticate the ANs before allowing them to communicate with other ANs or BSs. The blockchain is updated with transactions when the node's identification has been verified. The blockchain cannot be used to delete the transaction data. The transparency and traceability of the blockchain enable the proposed methodology to identify rogue nodes. In this approach, the blockchain offers secure routing and a productive technique for evaluating trust to find malicious nodes.

For SN and AN authentication, blockchains can be either private or public. In this architecture, two different kinds of blockchains are utilized to lessen the stress placed on the NAs. ANs died in the initial rounds of the prior authentication process because they had to register and validate other ANs. However, in our suggested model, the BS, which has powerful computational capabilities, registers and authenticates the AN. The NAs' workload is lightened in this way. Therefore, the coexistence of the two blockchains helps to lower the computational expense of the proposed paradigm. Because the AN is directly connected to the public blockchain, and the identification of every node is uploaded to the blockchain.

3.2. TRUST CALCULATION

Trust value has been calculated for two parameters, such as Direct Trust and Indirect Trust, which are described as follows. The TEARP version contains three inputs, namely security, portability, and dependability, to determine the cost of trust.

3.2.1. Indirect Direct Trust (IDT)

DT displays a node that contains the opinion variable. To use the IDT, which is defined below, a node must have a witness variable, which is not possible without one.

$$\mathbf{MFX_m^f(\tau)} = \frac{\mathbf{1}}{\mathbf{s}}\sum_{\mathbf{m=1}}^{\mathbf{s}} \mathbf{FX_m^f\,(f)} \tag{1}$$

For the purpose of preventing attacks and enhancing the security of the trust mechanism, formula 5 is used to calculate fraud ratings.

$$\boldsymbol{d_k^t} = \sqrt{\frac{\sum_{B_x \in B}\left(\bar{D} - D_{kB_x}^t\right)^2}{l}} \tag{2}$$

3.2.2. Direct Trust (DT)

A link between the m^{th} source node and the f^{th} endpoint node takes an estimated time to form, which is called the direct trust (DT). Therefore, Direct trust involving the use of the m^{th} source node and f^{th} endpoint has been described as,

$$FX_m^f(\tau) = \frac{1}{3}\left[FX_m^f(\tau-1) - \left(\frac{\tau_{appx}-\tau_{appx}}{\tau_{appx}}\right) + \omega\right] \quad (3)$$

Where τ_{appx} defines the anticipated duration, and τ_{est} specifies the estimated duration. This indicates that it takes time to τ_{appx} acquire and τ_{est} transfer the public key between the destination and the node. ω denotes the nodes' opinion variable.

$$R_f = \frac{\gamma * re_f - rf_f}{me_f} \quad (4)$$

$$S_f = \frac{\gamma * se_f - us_f}{me_f} \quad (5)$$

where re_f and se_f represent number of packets f has transmitted and received, respectively. The amount of information that f has discarded to be received and delivered, respectively, is represented by rf_f and us_f. The total number of packets that node f has received and transmitted is shown in the message. The adaptive penalty coefficient is written as γ.

3.3. CLUSTERING AND OPTIMAL CLUSTER HEAD SELECTION

The SCSO approach maximizes the network's lifespan. If damaged nodes are unable to send data due to damage, collaborate with nearby nodes to replace them. By swapping out the node, the SCSO version of the Cluster Head presented in this study performs better than the prior SCSO. The challenge of keeping them in a small space led to the development of the Sand Cat Swarm Optimization (SCSO) approach. Equation 10 offers an algebraic representation of SCSO.

$$T_m^n = \begin{cases} ET_m + p_1[(VC_m - NC_m)p_2 + NC_m]p_3 \geq 0 \\ ET_m - p_1[(VC_m - NC_m)p_2 + VC_m]p_3 < 0 \end{cases} \quad (6)$$

Where, $T_m{}^n$ is the First cluster head position in m^{th} dimension, ET_m is Food Source's position in m^{th} dimension, VC_{mis} upper bound in m^{th} dimension NC_m is lower bound in m^{th} dimension and p_1, p_2 is random numbers based on the interval [0,1]. The significant coefficient r_1, which is employed in Equation 11 to balance the processes of food acquisition and consumption, is the most crucial factor.

$$p_1 = 2f^{-\left(\frac{4x}{M}\right)^2} \quad (7)$$

The number L denotes the recent round, and M is the extreme number of rounds, where p_1 is a significant coefficient of SCSO.

3.4. ROUTING USING SAND CAT SWARM OPTIMIZATION

The performance of sand cats in nature served as the basis for a metaheuristic algorithm known as sand cat swarm optimization (SCSO). Sand cats, as opposed to domestic cats, survive in stony and sandy deserts. Sand cats have a 2 KHz hearing threshold. They resemble domestic cats and other cat species in regards to appearance. Sand cats only have fur on their hands and soles because of the intense conditions they endure. This protects them from heat and cold at home. This trait makes it challenging to follow a cat's trace. A sand cat's unique physical characteristic is their ability to hear low-frequency disturbancesThe Sand cat swarm optimisation algorithm (SCSO) replicates this characteristic to provide a close to optimal result, enabling them to immediately and accurately determine their targets.

3.4.1. Objective function for Routing

The cluster-based WSN will be able to maximize network lifetime by selecting the optimum path. To achieve this, a four-factor adaptive function is created, accounting for the nodes' remaining energy, their size, their location within the cluster, and their coverage rate. These parameters' definitions and derivatives are as follows:

Node Degree (N_D): It is the quantity of non-CH members that belong to each CH. Thus, it is recommended for CH to have the lowest node degree.

$$N_D = \sum_{x=1}^{p} |C_{m^x}| \quad (8)$$

Here, | C_{m^x} | is the x^{th} cluster head's number of cluster members.

Residual Energy(R_E): It represents the node's present energy level. It is calculated as the difference between the total amount of energy utilised over a period of time and the initial energy level.

$$R_E = \sum_{x=1}^{p} \frac{1}{CH_x} \quad (9)$$

where CH_x is the x^{th} cluster head's remaining energy.

Distance to neighbour (D_N): It specifies the distance between its own CH and its neighbour. The distance between a normal sensor and CH is given by equation (14).

$$D_N = \sum_{x=1}^{p}\left(\sum_{x=1}^{L_y} dis(n_x, CH_y)/L_y\right) \quad (10)$$

Node Centrality (N_C): It is described as the distance between a node's centre location and its neighbours, and it is written in equation (11).

$$N_C = \sum_{x=1}^{p} \frac{\sqrt{(\sum_{y \in n} dis^2(x,y))/n(x)}}{Network\ dimension} \quad (11)$$

where $n(x)$ is the number of nodes that are neighbours to CH_y.

The weighted values are ϑ_1, ϑ_2, ϑ_3,and ϑ_4. The equation (12) displays the single objective function.

$$Fitness = \vartheta_1 N_D + \vartheta_2 R_E + \vartheta_3 D_N + \vartheta_4 N_C, where\ \sum_{x=1}^{4} v_x = 1, v_x \in (0,1) \quad (12)$$

A metaheuristic algorithm leads the method to satisfy the problem objective, such as minimization or maximization. Every strategy's fitness (cost) for the search agent determines the subsequent repetition, and so on until optimal outcomes are obtainedThe most effective outcome is typically determined by the hunting mechanism. SCSO search agents look for targets after initiation to identify the most efficient approach. The Sand Cat's capacity to make low-frequency sounds is used to achieve this goal. Every search agent has a predefined sensitive range starting at 2 kHz. In SCSOA, the population size is 500, number of iteration is 1000 and the number of independent runs is 10.

Equation 13 shows the SCSO algorithm $\vec{P_N}$ variable drops gradually from 2 to 0. In this case, the T_D parameter was supposed to be 2. Iteration count is iterc, while iteration maximum is itermax. The sand cat's behaviour becomes sophisticated after half of the repetitions and is swift in the first iteration. Similar to this, the SCSO balances exploration and exploitation processes using T_D variables.

$$\overrightarrow{p_N} = T_D - \left(\frac{T_D \times iter_m}{iter_{Max}}\right) \tag{13}$$

$$\vec{P} = 2 \times \overrightarrow{p_N} \times rand(0,1) - \overrightarrow{p_N} \tag{14}$$

According to Equation 14, phase transformations are balanced. Equation 15 also prevents trapping in the local optimum. A $\vec{p}$ parameter controls evolutionary algorithms' efficiency. SCSO updates each agent's location.

$$\vec{p} = \overrightarrow{p_N} \times rand(0,1) \tag{15}$$

Equation 16 guarantees that the most suitable location of applicants for a search agent ($\overrightarrow{Pos_{im}}$) is updated after each algorithm iteration. Along with the agent's current location ($\overrightarrow{Pos_{im}}$) and sensitivity area ($\vec{p}$), this information is obtained. The SCSO continues with the subsequent step of its procedure, which is the exploiting of the target discovered after looking for it (exploration).

$$\overrightarrow{pos}(s+1) = \vec{p}.\left(\overrightarrow{Pos_{\iota m}}(s) - rand(0,1).\overrightarrow{Pos_m}(s)\right) \tag{16}$$

$$\overrightarrow{pos_{puv}} = \left|rand(0,1).\overrightarrow{Pos_\iota}(s) - \overrightarrow{Pos_m}(s)\right| \tag{17}$$

$$\overrightarrow{Pos}(s+1) = \overrightarrow{Pos_\iota}(s) - \vec{p}.\overrightarrow{Pos_{puv}}.\cos(\theta) \tag{18}$$

The direction between the optimum ideal position and the present position of each search agent is determined by Equation 19. The most optimal (balanced) results locations in Equation 22, the ($\overrightarrow{Pos_i}$) and ($\overrightarrow{Pos_{rnd}}$) are as well as the randomly selected locations, appropriately.

$$Y(s+1) = \begin{cases} \overrightarrow{Pos_\iota} - \vec{p}.\overrightarrow{Pos_{puv}}.\cos(\theta) & |P| \leq; exploitation \\ \vec{p}.\left(\overrightarrow{Pos_{\iota m}}(s) - rand(0,1).\overrightarrow{Pos_m}(s)\right) & |P| >; exploration \end{cases} \tag{19}$$

Pseudocode of SCSOA

```
Initializing Population
Compute the fitness function dependent upon the main function
Initializing the $r$, $r_G$, $R$
While( $t \leq t_{Max}$)
        For all the SCs
    Obtain an arbitrary angle$\theta$ ($0 \leq \theta \leq 360°$ )
                If($|R| \leq 1$)
    Upgrade the searching agent dependent upon the exploitation phase of equation (23); $\overrightarrow{Pos_i}$ - $\vec{p}.\overrightarrow{Pos_{puv}}$.cos($\theta$)
                Else
    Upgrade the searching agent dependent upon the exploration phase of equation (23); $\vec{p}.(\overrightarrow{Pos_{im}}(s)$-$rand(0,1).(\overrightarrow{Pos_m})(s))$
                End
        End
    $t = t+1$
    End
```

4. RESULT

This segment presents the experimental analysis of the suggested approach to Trust and Energy Aware Routing Protocol (TEARP) techniques.

Table 1. Stimulation parameters

Parameters	Units
Frequency	30khz
Queue size	50 packets
Simulation time	50 s
Number of nodes	500 nodes
Packets size	500 bytes
Data rate	2 Mbps
Length of data packet	500 bytes

4.1. COMPARISON ANALYSIS

A comparison is conducted between the proposed Trust and Energy Aware Routing Protocol (TEARP) technique and existing methods ESR [13], MOSFA [16], and OSVM [18] in terms of the No. of nodes, the Packet Delivery Ratio, the Residual Energy, and the Throughput the Network Lifetime.

In Fig. 2, the proposed method strategy shows the Network lifetime. TEARP outperforms other techniques with a lower fraction of nodes, almost doubling the network lifetime in the process.The proposed method achieves a better network lifetime of 61.40 %, 39.64 %, and 52.24 %, than ESR, MOSFA, and OSVM.

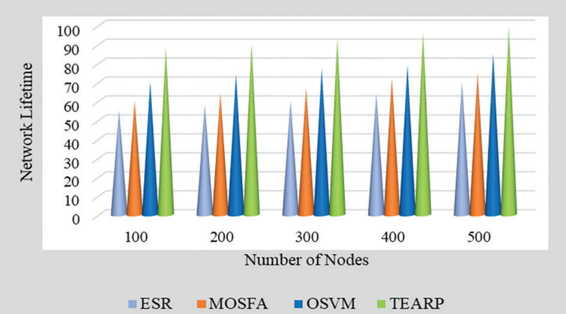


Fig. 2. Comparison of Network Lifetime

Fig. 3 presents the equivalence of the packet delivery ratio of the suggested technique in comparison with existing techniques. TEARP performs better than other existing techniques and the ratio appears to be large. The proposed method achieves a better Packet Delivery Ratio of 45.54 %, 27.16 %, and 38.48 % than ESR, MOSFA, and OSVM.

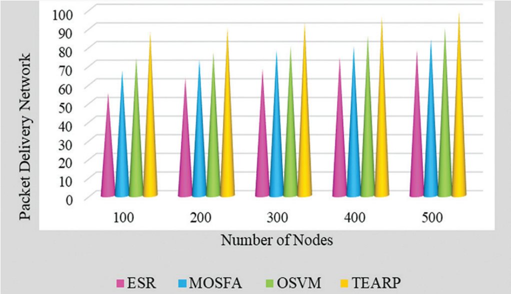
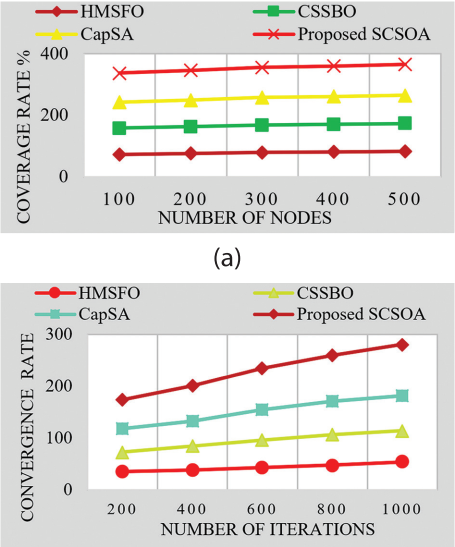


Fig. 3. Comparison of Packet Delivery Network

Fig. 4(a) and (b) examine the four algorithms' rates of coverage, and rates of convergence. The link between coverage and population size when using the SCOA, HMSFO, CSSBO, and CapSA algorithms is depicted in Figure 4(a). Four algorithms will enhance network coverage as the population grows. The HMSFO, CSSBO, and CapSA algorithms cannot compete with the proposed SCSO algorithm. The relationship between convergence rate and iterations for the SCSOA, HMSFO, CSSBO, and CapSA algorithms is depicted in Figure 4(b).



(b)

Fig. 4. Performance comparison of different algorithms

The SCSO algorithm peaks and converges quickly in terms of growth scope at the number of iterations, whereas the other three algorithms continue to increase quickly after that point. As a result, the SCSO algorithm's convergence speed and time to optimal value are both faster. The SCSO algorithm exhibits a better simulation effect in the algorithm's convergence area. In conclusion, the SCSO method outperforms the other three algorithms in terms of convergence speed and coverage ratio.

Fig. 5 displays a comparison of latency with various options. Due to how long it takes to choose the starting path, current solutions cannot reduce latency. Additionally, a safe and effective path is selected for data transfer. As a result, the delay time will be reduced by the proposed TEA RP approach.

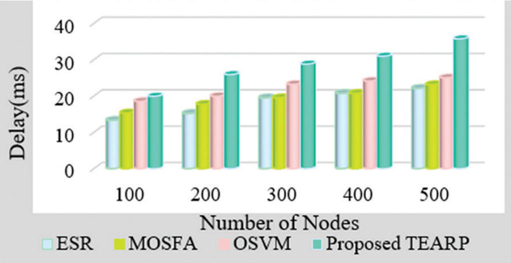


Fig. 5. Comparison of Residual Energy

The proposed technique's throughput equivalent in relation to existing techniques is depicted in Fig. 6. The proposed technique achieves higher throughput than other existing techniques. ESR, MOSFA, and OSVM, and the proposed Trust and Energy Aware Routing Protocol (TEARP) are achieving better than throughput is 50.15 %, 32.45 %, and 29.64 %.

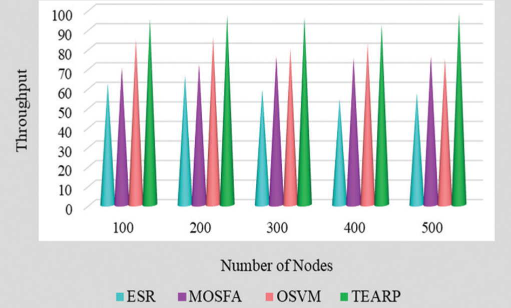


Fig. 6. Comparison of Throughput

4.3. DETECTION ACCURACY

The detection accuracy indicator shows the proportion of correct detections made by the suggested technique with the minimum possible false reports. TEARP detection accuracy is 28.35% and 67.43%, respectively. Detection Accuracy is shown in Fig. 7.

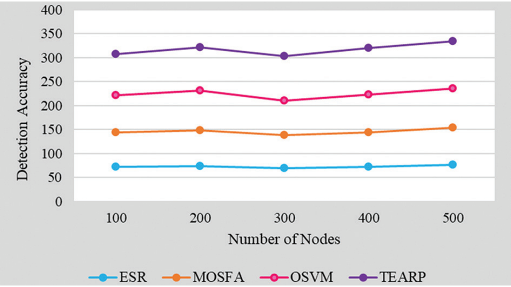


Fig. 7. Detection Accuracy

5. CONCLUSIONS

In this paper, a Trust and Energy Aware Routing Protocol (TEARP) in WSNs is proposed, which use blockchain technology to maintain the identity of the SNs and ANs. The proposed TEARP has been simulated using MATLAB. The simulation outcomes demonstrate that the proposed TEARP framework outperforms more established methods like ESR, MOSFA, and OSVM.The proposed TEARP method improves the network lifetime by 39.64%, 33.05%, 29.64% and 27.16%, respectively, and has better detection accuracy of 28, 35% and 67.43%. compared with ESR, MOSFA and OSVM techniques. The TEARP method is not applicable in large-scale situations. The TEARP technique must be used in a large-scale context in future to overcome such constraints.Additionally the proposed TEARP approach might include algorithmic tests with an extensive network that employs agent-based communication for trust modeling.

Acknowledgment

The author would like to express his heartfelt gratitude to the supervisor for his guidance and unwavering support during this research for his guidance and support.

6. REFERENCES:

[1] M. Rathee, S. Kumar, A. H. Gandomi, K. Dilip, B. Balusamy, R. Patan, “Ant colony optimization-based quality of service aware energy balancing secure routing algorithm for wireless sensor networks”, IEEE Transactions on Engineering Management, Vol. 68, No. 1, pp. 170-182.

[2] Z. Wang, H. Ding, B. Li, L. Bao, Z. Yang, Q. Liu, “Energy efficient cluster-based routing protocol for WSN using firefly algorithm and ant colony optimization”, Wireless Personal Communications, Vol. 125, No. 3, 2022, pp. 2167-2200.

[3] K. Haseeb, N. Islam, A. Almogren, I. U. Din, “Intrusion prevention framework for secure routing in WSN-based mobile Internet of Things”, IEEE Access, Vol. 7, 2019, pp. 185496-185505.

[4] Y. Xiong, G. Chen, M. Lu, X. Wan, M. Wu, J. She, “A two-phase lifetime-enhancing method for hybrid energy-harvesting wireless sensor network”, IEEE Sensors Journal, Vol. 20, No. 4, 2019, pp. 1934-1946.

[5] M. Selvi, K. Thangaramya, S. Ganapathy, K. Kulothungan, H. Khannah Nehemiah, A. Kannan, “An energy aware trust based secure routing algorithm for effective communication in wireless sensor networks”, Wireless Personal Communications, Vol. 105, 2019, pp. 1475-1490.

[6] P. S. Mann, S. Singh, “Improved artificial bee colony metaheuristic for energy-efficient clustering in wireless sensor networks”, Artificial Intelligence Review, Vol. 51, 2019, pp. 329-354.

[7] M. Zivkovic, N. Bacanin, E. Tuba, I. Strumberger, T. Bezdan, M. Tuba, "Wireless Sensor Networks Life Time Optimization Based on the Improved Firefly Algorithm", Proceedings of the International Wireless Communications and Mobile Computing, Limassol, Cyprus, 15-19 June 2020, pp. 1176-1181.

[8] S. K. Barnwal, A. Prakash, D. K. Yadav, “Improved African Buffalo Optimization-Based Energy Efficient Clustering Wireless Sensor Networks using Metaheuristic Routing Technique”, Wireless Personal Communications, Vol. 130, No. 3, 2023, pp. 1575-1596.

[9] R. K. Yadav, R. P. Mahapatra, “Hybrid metaheuristic algorithm for optimal cluster head selection in wireless sensor network”, Pervasive and Mobile Computing, Vol. 79, 2022, p. 101504.

[10] S. Banerjee, R. B. Karennavar, P. Sirigeri, R. Jayashree, “Multimedia Text Summary Generator for Visually Impaired”, Proceedings of the 6th International Conference on Communication and Electronics Systems, Coimbatre, India, 8-10 July 2021, pp. 1166-1173.

[11] K. Haseeb, A. Almogren, N. Islam, I. Ud Din, Z. Jan, “An energy-efficient and secure routing protocol for intrusion avoidance in IoT-based WSN”, Energies, Vol. 12, No. 21, 2019, p. 4174.

[12] A. Barzin, A. Sadegheih, H. K. Zare, M. Honarvar, “A hybrid swarm intelligence algorithm for clustering-based routing in wireless sensor networks”, Journal of Circuits, Systems and Computers, Vol. 29, No. 10, 2020, p. 2050163.

[13] S. Amaran, R. M. Mohan, “Intrusion detection system using optimal support vector machine for wireless sensor networks”, Proceedings of the International Conference on Artificial Intelligence and Smart Systems, Coimbatore, India, 25-27 March 2021, pp. 1100-1104.

[14] D. L. Reddy, C. Puttamadappa, H. N. Suresh, “Merged glowworm swarm with ant colony optimization for energy efficient clustering and routing in wireless sensor network”, Pervasive and Mobile Computing, Vol. 71, 2021, p. 101338.

[15] G. Manoharan, A. Sumathi, "Efficient routing and performance amelioration using Hybrid Diffusion Clustering Scheme in heterogeneous wireless sensor network", International Journal of Communication Systems, Vol. 35. No. 15, 2022, p. e5281.

[16] M. Rizwanullah, H. K. Alsolai, M. Nour, A. S. A. Aziz, M. I. Eldesouki, A. A. Abdelmageed, "Hybrid Muddy Soil Fish Optimization-Based Energy Aware Routing in IoT-Assisted Wireless Sensor Networks", Sustainability, Vol. 15, No. 10, 2023, p. 8273.

[17] A. Sharma, H. Babbar, S. Rani, D. K. Sah, S. Sehar, G. Gianini, "MHSEER: A Meta-Heuristic Secure and Energy-Efficient Routing Protocol for Wireless Sensor Network-Based Industrial IoT", Energies, Vol. 16, No. 10, 2023, p. 4198.

[18] J. Paruvathavardhini, B. Sargunam, "Stochastic Bat Optimization Model for Secured WSN with Energy-Aware Quantized Indexive Clustering", Journal of Sensors, Vol. 2023, 2023.

FEDRESOURCE: Federated Learning Based Resource Allocation in Modern Wireless Networks

Original Scientific Paper

P. G. Satheesh
Research Scholar,
Sathyabama Institute of Science and Technology,
Chennai, Tamil Nadu, India
reachpgs@gmail.com

T. Sasikala
Prof. & Dean, School of Computing,
Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology,
Chennai, Tamil Nadu, India
dean.computing@sathyabama.ac.in / sasi.madhu2k2@yahoo.co.in

Abstract *– Deep reinforcement learning can effectively deal with resource allocation (RA) in wireless networks. However, more complex networks can have slower learning speeds, and a lack of network adaptability requires new policies to be learned for newly introduced systems. To address these issues, a novel federated learning-based resource allocation (FEDRESOURCE) has been proposed in this paper which efficiently performs RA in wireless networks. The proposed FEDRESOURCE technique uses federated learning (FL) which is a ML technique that shares the DRL-based RA model between distributed systems and a cloud server to describe a policy. The regularized local loss that occurs in the network will be reduced by using a butterfly optimization technique, which increases the convergence of the FL algorithm. The suggested FL framework speeds up policy learning and allows for adoption by employing deep learning and the optimization technique. Experiments were conducted using a Python-based simulator and detailed numerical results for the wireless RA sub-problems. The theoretical results of the novel FEDRESOURCE algorithm have been validated in terms of transmission power, convergence of algorithm, throughput, and cost. The proposed FEDRESOURCE technique achieves maximum transmit power up to 27%, 55%, and 68% energy efficiency compared to Scheduling policy, Asynchronous FL framework, and Heterogeneous computation schemes respectively. The proposed FEDRESOURCE technique can increase discrimination accuracy by 1.7%, 1.2%, and 0.78% compared to the scheduling policy framework, Asynchronous FL framework, and Heterogeneous computation schemes respectively.*

Keywords*: Deep reinforcement learning, federated learning, resource allocation, butterfly optimization technique*

1. INTRODUCTION

Modern wireless networks and mobile devices frequently come with sophisticated sensors and powerful computers, enabling them to acquire and interpret enormous amounts of data produced at the network edge [1]. The 5th generation (5G) wireless networks have strengthened the traditional connection service and supported many vertical industries [2]. A cloud and edge computing system [3, 4] that intelligently uploads user tasks to a cloud data center layer and an edge computing layer can provide computation and data storage services. Implementing energy-efficient node setup and RA in the course of cooperative operations is a major difficulty in wireless networks due to the high quality of service (QoS) requirements of IoT applications.

The efficient RA scheme can extend sensors' lifetime and play a major role in maximizing system performance along with better scheduling [3-5]. Utilizing machine learning (ML) techniques [6, 7], with a variety of RA strategies have recently been investigated which reduces the wireless networks becoming increasingly complex [8]. Particularly for difficult decision-making issues, deep reinforcement learning (DRL) has been applied extensively [9]. They can be used to train a DL model with a large representation capacity to develop a RA strategy for complicated networks. However, such DRL-based approaches still face significant obstacles in practice [10, 11].

An important problem is the policy's inability to respond to changes in network needs [12]. Wireless networks often introduce new systems with the same goals as current ones [13, 14]. It is therefore possible to apply policies to newly arrived systems with no additional learning if they are network adaptable [15, 16]. By using DRL-based wireless network approaches, we can deploy them more effectively than before. In this paper, a novel federated learning-based resource allocation (FEDRESOURCE) has been proposed, which efficiently performs RA in wireless networks. The main contributions of the FEDRESOURCE framework are as follows.

- The proposed FEDRESOURCE technique uses federated learning (FL) optimized using butterfly optimization for resource allocation in wireless networks, that share the DRL-based RA model between distributed systems and a cloud server to describe a policy.
- The policy for the RA in wireless networks can be learned collaboratively while taking use of the FL technique.
- The regularized local loss that occurs in the network will be reduced by using a butterfly optimization technique, which increases the convergence of the FL algorithm.
- The suggested FL framework speeds up policy learning and allows for adoption by employing deep learning and the optimization technique.

The remainder of the paper is organized as follows: Section 2 presents a thorough survey of current efforts on federated learning-based RA techniques. Section 3, presents the system model and problem formulation for resource allocation. Section 4 presents the design of the FEDRESOURCE framework in detail. Section 5 presents the experimental data and its analysis. Section 6 presents the conclusions and suggestions for future research.

2. LITERATURE REVIEW

Due to the high Quality-of-Service (QoS) requirements of IoT applications, resource scheduling wireless networks are becoming more important for better service. Several techniques have been developed by many experts for RA in wireless networks. Among these, we have discussed a few algorithms here. In [16] authors introduced a circumstance-independent policy that can successfully address the various network scenarios even with a single policy. Based on the outcomes of the simulation, a single suggested policy can be applied in a range of circumstances with results that are comparable to those of a situation-based policy, which chooses the appropriate course of action for each circumstance on its own.

In [17] authors presented a heterogeneous computation and RA approach based on heterogeneous mobile architectures. Using simulation data, the proposed scheme improves the energy efficiency of the wireless-powered FL system more than the baseline systems, according to the simulation data.

In [18] authors recommended using communication pipelining to enable FL in mobile edge computing applications to become more efficient at utilizing wireless spectrum and to become more concurrent. They also provide numerical findings that highlight the benefits of the suggested technique for various datasets and deep learning architectures.

In [19] authors proposed a new asynchronous FL framework that considers time-varying local training data, wireless link conditions, and computing capability. The framework also uses a dynamic scheduling algorithm to optimize learning performance under long-term energy constraints and per-round latency requirements. The proposed architecture has been demonstrated to improve learning performance and system efficiency over other approaches through numerical simulations.

In [20] authors proposed a scheduling policy that took user device training data representation and channel quality into consideration simultaneously. Based on simulations, the channel-aware data importance-based scheduling policy is shown to be more efficient than cutting-edge FL methods. In an asynchronous FL environment, an "age-aware" aggregation weighting approach can also improve learning performance.

In [21] authors established an efficient integration of common edge intelligence nodes based on research on energy-efficient bandwidth allocation, CPU frequency calculation, optimized transmission performance, and required level of learning accuracy. Based on the simulation results, the proposed Alternative Direction Algorithm (ADA) can reduce energy consumption while slightly increasing FL time in the central processing unit. There have been few studies that examine FL design in wireless networks for RA. However, FL structures used in the literature are not very effective, and model updates derived from old global models may have limited meaningful information about the current version, resulting in slow convergence. To the best of our knowledge, no work has specifically addressed how to use FL to resolve a wireless network RA issue. Instead, to efficiently run FL on wireless networks, current studies concentrate on finding a solution to the RA problem in wireless networks.

2.1. DIFFERENCES BETWEEN THE EXISTING AND PROPOSED WORK

The important findings from their research as well as the differences between the proposed study and the existing work are given below.

a. Unlike the proposed method many of the algorithms did not consider FL to resolve a wireless network RA issue

b. A new FEDRESOURCE technique that uses federated learning (FL) that shares the DRL-based RA model between distributed systems is presented in the proposed work that is not included in any other method so far, which makes it unique and significant from existing methods.

c. Proposed a federated learning architecture that incorporates policy chosen from DRL for resource allocation in wireless networks.

d. The existing techniques used in the literature are not very effective, and model updates derived from old global models may have limited meaningful information about the current version, resulting in slow convergence. However, it is discovered that the suggested method increases the convergence rate.

3. SYSTEM MODEL AND PROBLEM FORMULATION

We take into account a downlink of numerous TDMA networks in wireless networks, where each network has a local model and a global model, as shown in Fig. 1. R=1,... ,S, where S is the total number of systems, defines the set of systems. $\mho_s$=1,…,U_s, where U_s is the total number of sensing nodes (SN) in system s, defines the set of SN in system s. In the system $s \in R$, its AP provides services to U_s SN across discrete time intervals of $k \in \{1,2,\dots\}$. We make the commonly recognized assumption that each system's wireless channels between the global model and local model are time-varying but constant during a timeslot and satisfy the Markov property. The global model schedules one user and transmission power in timeslot k of system s, where T is the set of potential transmission power levels. P_s is the definition of a state of the system s in timeslot k, and $p_s{}^k$ is the state space of the system s. Each user's feature information, including channel gain and QoS satisfaction levels, is represented by the state.

We use a tuple to represent SN m in system s in simple notation (s, m). We use $f^k{}_{s,m,i}$ to denote the ith feature information of the user (s, m) in timeslot k. $c_s{}^k$=($m_s{}^k$,$T_s{}^k$)$\in C_s$ is the definition of an action of system s in timeslot k, which denotes a scheduling choice. A system s policy is indicated by the notation π_s:$P_s \to C_s$. Then, l($p_s{}^k$, π_s,($p_s{}^k$)), where l(.,.) is a utility function typically utilized in the systems and is used to express the instantaneous utility of system s in timeslot k. In wireless networks with numerous systems, we can construct a general RA issue to maximize overall utility as follow state represents each individual's feature information.

$$maximize_{\pi} \sum_{s \epsilon \mathcal{R}} \mathbb{E}[\sum_{k=0}^{\infty} (\gamma)^k \, l(p_s^k, \pi_s, (p_s^k))] \qquad (1)$$

where the discount factor is 0 < γ < 1 and the policy for all systems is π:$\prod_{s \in R} P_s \to \prod_{s \in R} C_s$. For instance, one could use the formula l($p_s{}^k$, π_s , ($p_s{}^k$)) = d($p_s{}^k$, π_s ,($p_s{}^k$)) to frame the problem of maximizing the total average data rate. where the function d(.,.) calculates the current data rate.

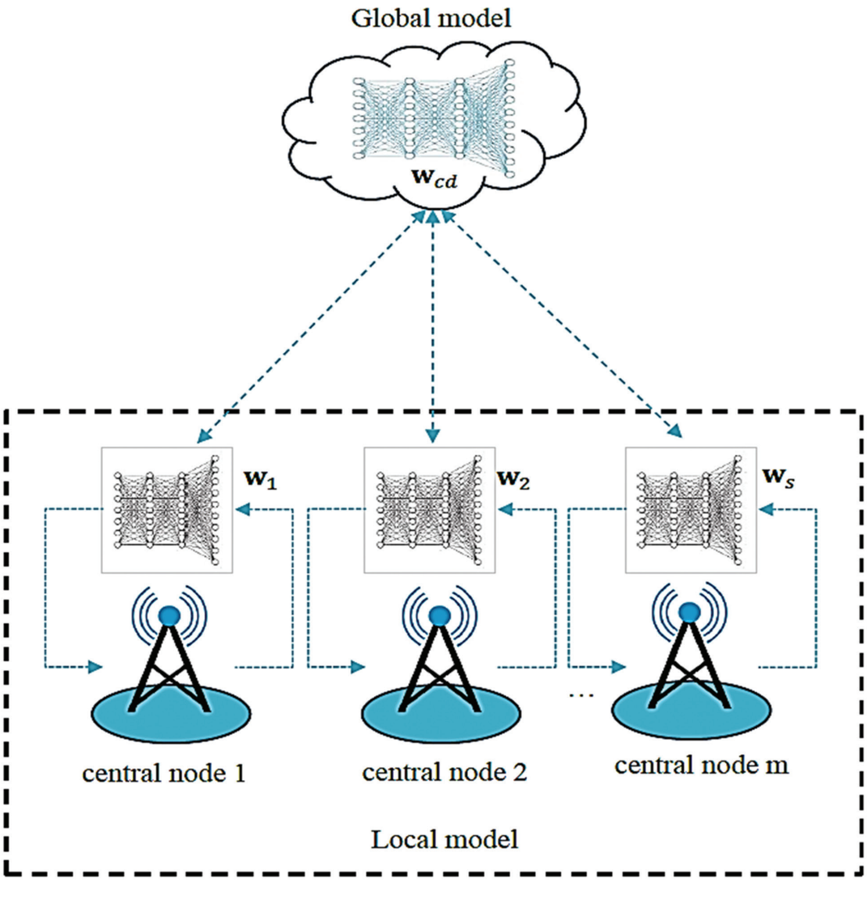


Fig. 1. System model

3.1. FEDERATED LEARNING FOR DYNAMIC RESOURCE ALLOCATION

When S, the number of systems, is high, the complexity of the problem is too great to tackle. To fix this problem, we break it down into its parts according to each system s as

$$\text{maximize}_{\pi_s} \; \mathbb{E}\left[\sum_{k=0}^{\infty} (\gamma)^k l\left(p_s^k, \pi_s(P_s^k)\right)\right] \tag{2}$$

A (sub)optimal rule for the distributed system s deconstructed problem may then be identified, and the problem can be resolved. A Markov decision process is used to solve the decomposed problem, and the reward function and environment are the utility function l(.,.) and transition probabilities over the state and action spaces, respectively. The well-known DRL can therefore be utilized to resolve the system-wise deconstructed problem, like the most recent research on RA in wireless networks. In particular, every system s participates in the DRL as an agent to learn the best policy $\pi{*}s$ for the problem's decomposition. Decomposed problems fall under the same class as other problems that use the same utility functions to accomplish the same objective (in this case, RA). FL can therefore be used to solve issues more effectively. Fig. 1 illustrates how FL can be used and provided if there is a common policy π^{-} that can be employed in any system.

Through her DRL technique, each system in FL learns common policies on its own. The cloud server then collects the common policies for the system and delivers the combined policies. As the experience of all systems is used, this accelerates the learning of common policies. By adding the new global model to the wireless network and using the cloud server's common policy, it is also possible to adapt to newly introduced systems.

4. DESIGN OF FEDERATED LEARNING-BASED RESOURCE ALLOCATION

In this section, a novel FEDRESOURCE has been proposed which efficiently performs RA in wireless networks. Federated learning is designed to minimize training loss while handling distributed neural network training across many devices with their local training data. The proposed FEDRESOURCE technique shares the DRL-based RA model between distributed systems and a center node to describe a policy for the FL framework. The weights of the DL models at the center node and the SN are indicated in the figure by the notations W^K_{s+1} and W_m, respectively. We refer to the deep learning models as policy models since they display the policies. A DRL approach is used by each distributed system to individually learn its local policy model. The overall block diagram for the proposed FEDRESOURCE model is given in Fig. 2.

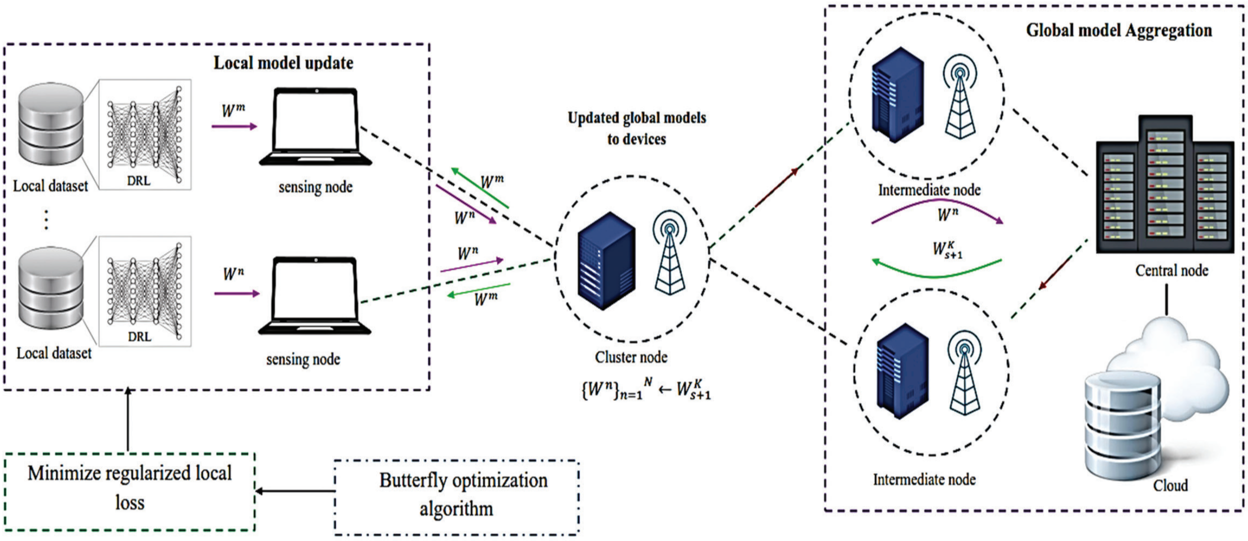


Fig. 2. Proposed FEDRESOURCE model

Similar to the traditional FL approach, the cloud server combines policy models learned from the system to update the central policy model. Then, as shown in Fig. 2, every system replaces its local policy model with the updated central policy model that was redistributed by the cloud server to the systems. To exploit all local experiences in the distributed system for learning, even if local experiences are not broadcast, FL can quicken the learning of the policy model by performing this iteratively. The central policy architecture at the center node also offers adaptation to recently arrived systems.

FEDRESOURCE employs an iterative method that calls for S_g global rounds for global model changes. The SN and cloud server interact in the following ways during each global round. SNs modify regional models: Each SN m first receives the feedback data from the server, to generate the local model $w_m{}^r$ at a global round r. It then minimizes the surrogate function.

$$min_{w\in\mathbb{R}^k} R_m^r(w) := E_m(w) + \langle \eta\nabla\bar{E}^{r-1} -, w\rangle \tag{3}$$

One of the fundamental principles of FEDRESOURCE is that a sensing device can roughly solve the problem to provide an approximation solution $w_m^{\ r}$ satisfying $||\nabla R^r_{\ m}\ (w_m^{\ r})||\leq\theta||\nabla R^r_{\ m}\ (w^{r-1})\ )||,\forall m$, which is parametrized by a local accuracy (0, 1) that is shared by all sensing devices. Here, $\theta = 0$ indicates that the local problem must be handled optimally, and $\theta = 1$ indicates that no progress has been made, for example, by setting $w_m^{\ r} = w^{\ r-1}$. FEDRESOURCE avoids employing proximal terms to restrict an extra controlling parameter (i.e.,), uses the global gradient estimate $\nabla E^{\ r-1}$ which the server can measure from the SN's data—instead of the exact but unrealistic $\nabla E(w^{r-1})$ and flexibly resolves local problems roughly by controlling will have $R_m^{\ r}\ (w)=E_m\ (w)+\eta\nabla E^{r-1}-\nabla E_m\ (w^{\ r-1})$, that contains both local gradient estimate $E_m\ (w)$ and global gradient estimate weighted by a programmable parameter η. Later, we'll discover how influences FEDRESOURCE convergence. To attain the advantages of a) theoretical linear convergence and b) experimentally fast convergence which will be discussed in later sections, FEDRESOURCE needs more information than currently accepted standard approaches.

4.1. DYNAMIC LEARNING FOR RESOURCE ALLOCATION USING FL

In this section, the RA strategy for numerous systems with a center node has been done by maintaining the policy. For the DRL-based policy that has been frequently utilized, we here assume a typical DQN method, but any alternative DRL-based techniques can also be applied. The optimal action-value function $A^*\ (st,\ ac)$, that denotes the maximum return which can be realized in state st with action ac, is approximated using a deep neural network (DNN) trained to perform the DQN method. DQN, as a result, is the name given to the DNN, and it is employed to construct policy by identifying the action that maximizes return for a given state. We use $\bar{\pi}(\overline{st};\ w)$ to represent the common policy based on DNN.

We designate w_{cd} and $w_{s'}$ respectively, as the weights of the DNN at the center node and system s. Each system s initializes its DNNs w_s and $w_s^{\ pr}$ as w_{cd} once the center node initializes its DNN w_{cd}. System s uses the DNN w_s and the translation functions $t_s^{\ st}$ and $t_s^{\ ac}$ to select the action $ac_s^{\ k}$ in timeslot k after observing its state $st_s^{\ k}$. The chosen action can be simply identified by the formula $ac_s^{\ k}=t_s^{\ ac}(\bar{\pi}(t_s^{\ st}\ (st_s^{\ k});\ w_s))$. The system provides services to the user following the selected action $(ac_s^{\ k})$, and it monitors the utility as $l_s^{\ k}=l(st_s^{\ k},\ ac_s^{\ k})$. When training the DNN, the experience of system s in timeslot k is described as $(st_s^{\ -k},\ ac_s^{\ -k},\ l_s^{\ k},\ st_s^{\ -k+1})$ and stored in the buffer. $st_s^{\ -k+1}= t_s^{\ st}\ (st_s^{\ k+1})$. The DNN w_s is trained to utilize the experiences using a variety of training methods, including experience replay and fixed target-Q. Each system s determines its local gradients for each FL interval by deducting its previous aggregated DNN, $w_s^{\ pr}$ from its present DNN. The cloud server then updates its DNN w_{cd} by combining the local gradients from all systems. The cloud server broadcasts w_{cd} to all systems after aggregation, and each system substitutes w_s and $w_s^{\ pr}$ with w_{cd}. Algorithm 1 provides a summary of the process.

Algorithm 1 FEDRESOURCE

```
The cloud server initializes $w_{cd}$
Each system s initializes $w_s$ and $w_s^{\ pr}$ as $w_{cd}$
for $k \in\{0,1,\ldots\}\ d0$
    for each system s do ▹ DQN Algorithm
        Observe $st_s^{\ k}$ and translate it as $\overline{st}_s^{\ k}\leftarrow t_s^{\ k}\ (st_s^{\ k})$
        Choose $\overline{\overline{ac}}_s^{\ k}\leftarrow \bar{\pi}(4w_s)$ and translate it as $ac_s^{\ k}\leftarrow t_s^{\ st}\ (\overline{\overline{ac}}_s^{\ k})$
        Do action $ac_s^{\ k}$ and observe $l_s^{\ k}$ and $st_s^{\ k+1}$
        Translate $st_s^{\ k+1}$ as $\overline{st}_s^{\ k+1} \leftarrow t_s^{\ st}\ (st_s^{\ k+1})$
        Store experience $(\overline{st}_s^{\ k},\ \bar{a}_s^{\ k},\ l_s^{\ k},\ \overline{st}_s^{\ k+1})$
        Update $w_s$ using its experiences by a DQN algorithm
    end for
    if mod($t$, $T_{FL}$)==0 then ▹ FL
        All systems calculate their local gradients $\nabla E_s$'s from their previous DNN $w_s$ To the current DNNs $w_s$'s
        The cloud server updates $w_{Cd}$ by aggregating the local gradients from all system
        All system replaces their DNNs $w_s$'s and $W_s^{\ pr}$'s to $w_{Cd}$
    endif
end for
```

4.2. MINIMIZE REGULARIZED LOCAL LOSS

Using the DRL algorithm may introduce regularized loss. To reduce the loss, the butterfly optimization algorithm (BOA) has been used in this paper. Butterfly reproductive behavior and its attraction to pheromones have been modeled by the BOA, a meta-heuristic algorithm with an emphasis on group and swarm behavior. To attract the opposite gender or to advertise where the best blooms are in the environment, butterflies release pheromones into their surroundings. Pheromones are not only employed by butterflies; other insects, such as ants, also release this chemical into the environment and use it to guide or lead other creatures. The more pheromones a butterfly produces, the more likely it is to attract additional butterflies, as butterflies like to travel to pheromone-rich environments. This algorithm makes the following assumptions:

- Every butterfly offers a different approach to the issue.
- The objective function decides which butterflies are eligible for pheromone release.
- The more pheromones present, the better the butterfly's ability to draw in additional butterflies and the better the problem's resolution.

The butterfly optimization process is considered to have a population member called a feature vector, each of whose components reflects the choice of the desired quality. Equation (4) demonstrates how the BOA algorithm produces a butterfly.

$$\text{F} =\ll F_i^1, F_i^2 \dots \dots F_i^D \gg \tag{4}$$

F_i is a D-dimensional feature vector, and F_i^j denotes the j^{th} component of the I^{th} feature vector. These feature vectors F can be generated randomly and used as the initial BOA population, as shown by Eq (5).

$$\text{F} =\ll F_1, F_2 \dots \dots F_n \gg \tag{5}$$

The initial population of feature vectors utilized for intrusion detection is denoted by the letter F, and the total number of feature vectors employed in the BOA is represented by the number n. The feature vector must be optimized by minimizing these two components of the objective function.

$$\text{fitness} = \alpha . mse + \beta \frac{||R||}{||N||} \tag{6}$$

In Eq. (6), Network intrusion detection uses a total of $||N||$ features, whereas $||R||$ is the number of features chosen to identify unauthorized traffic. The mean absolute error of approved network traffic is known as mse, and A number at random between 0 and 1 is "α", assuming β=1-α. In BOA, appropriate features can be selected to minimize the objective function.

$$F_i = F_i + (r^2 \times F^* - F_i) \times f_i \tag{7}$$

$$F_i = F_i + \left(r^2 \times F_j - F_i\right) \times f_i \tag{8}$$

In these equations, the feature vector, F_i can be updated by the vector F_j and F_k, as well as by optimized feature vectors like $F*$. F_i is the quantity of pheromone or attraction that a member of the BOA population produces, and r is a random number in the range of [0, 1].

$$F_i^j = \begin{cases} 0 & rand < |\frac{2}{\pi}\arctan\left(\frac{2}{\pi}F_j^i\right)| \\ 1 & rand < |\frac{2}{\pi}\arctan\left(\frac{2}{\pi}F_j^i\right)| \end{cases} \tag{9}$$

As a result, as indicated in Eq (9), extract the absolute or binary values using a transition function, such as a Gaussian or V-shaped function.

5. RESULT AND DISCUSSION

We create a special Python-based simulator just for the experiments, in which the following system is implemented. On a machine with 64 GB memory and an Intel Core i7-10700 processor, the simulation is run. We take into account various systems, each of which has a 5 MHz bandwidth. The noise spectral density is set to 106 dBm/Hz, the path loss exponent is set to 3.76, and a log-normal shadowing with a 6 dB standard deviation is taken into consideration. Each system's maximum transmission power will be 1 W. The Shannon capacity is used to determine the instantaneous data rate. we take into account a RA issue that seeks to satisfy average data requirements while minimizing average transmission power. The simulation setup for the proposed system is given in Table 1.

Table 1. Network simulation setup

Cellular network parameters	Values
Channel bandwidth	11.34
Noise power	2.76
The base station transmits power	2.03
Path loss between the base station and the user	3.76
Lognormal distribution shadow fading	6DB

5.1. PERFORMANCE ANALYSIS

The proposed method has been compared with existing techniques such as Heterogenous computation [18], Asynchronous FL framework [20], and scheduling policy [21] in terms of transmission power, convergence of algorithm, throughput, and cost.

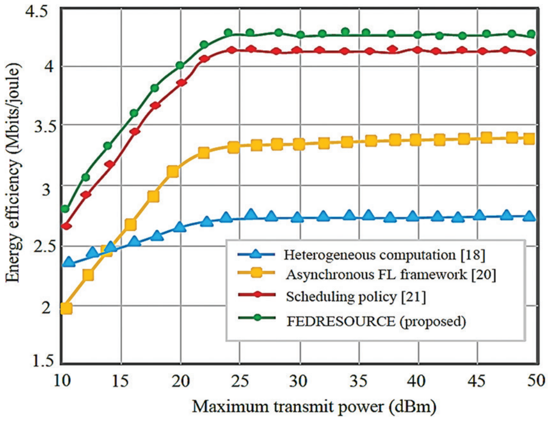


Fig. 3. Energy efficiency

When the maximum transmission power is altered, the energy efficiency is indicated in Fig. 3. This simulation demonstrates that as the maximum transmit power of the center node rises, all systems' energy efficiency initially rises and eventually stabilizes. This is because power efficiency does not increase monotonically with transmit power. The extra transmit power is not used since it is power-efficient if the maximum transmit power is 25 dBm or higher. Fig. 3 also indicates that the suggested FEDRESOURCE approach performs better than the Scheduling policy [21], Asynchronous FL framework [20], and Heterogeneous computation [18] schemes. For high maximum transmit power, FEDRESOURCE can increase up to 27%, 55%, and 68% energy efficiency when compared with the Scheduling policy [21], Asynchronous FL framework [20], and Heterogeneous computation [18] schemes respectively.

The convergence of the suggested FL algorithm and the fundamental method is shown in Figure 4. This figure demonstrates that the suggested FL algorithm, when compared to the scheduling policy framework, Asynchronous FL framework, and Heterogeneous computation schemes can increase discrimination accuracy by roughly 1.7%, 1.2%, and 0.78%. This is because the proposed FL algorithm updates the policy in the local model and is monitored by the DRL technique.

The model loss is reduced by butterfly optimization which also helps to increase the identification accuracy and the convergence rate. From the figure, it is clear that the proposed method achieves higher identification accuracy than existing techniques.

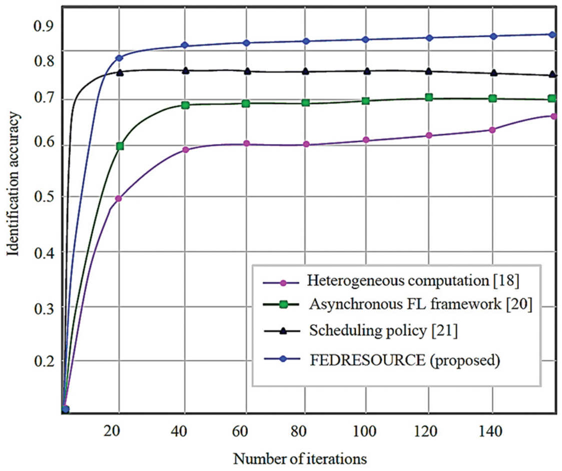


Fig.4. Convergence of FL algorithms

Fig. 5 displays the FL cost variation for the proposed approach with iteration for each number of devices. The term "iteration" refers to the execution of both the device allocation algorithm and the RA algorithm in a single process. The FL cost values show quick convergence (up to 6 iterations) for different device counts in the suggested design. With more SNs, the cost will display lower figures. The probability of being allocated to the closest SN grows as the number of sensor devices increases, which accounts for this pattern. Throughput is increased and FL costs are subsequently decreased by mapping the device to the nearest node.

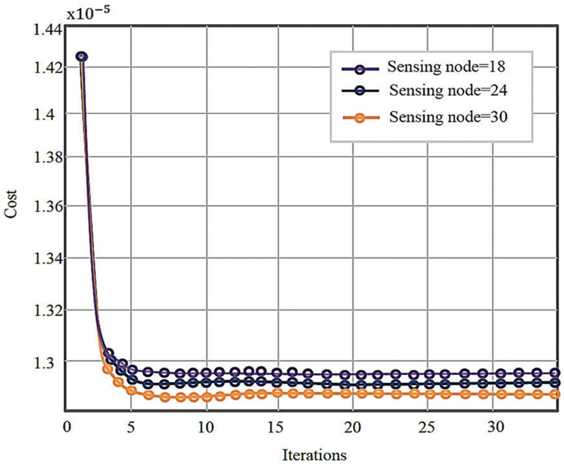


Fig. 5. Cost of FL

It can be seen from Fig. 6 that the suggested FEDRESOURCE framework can achieve almost the same throughput performance as the Scheduling policy algorithm. The reason is that the FEDRESOURCE algorithm considers the butterfly optimization algorithm, which can avoid loss when allocating resources for sensing devices. The throughput performance obtained by the Asynchronous FL algorithm is lower than that of the FEDRESOURCE algorithm. The uplink throughput performance of the Heterogenous computation algorithm is the lowest among the four algorithms.

6. CONCLUSIONS

In this paper, a novel FEDRESOURCE framework has been proposed which efficiently performs RA in modern wireless networks. We used experiments to show that the suggested FL framework may speed up RA policy learning and offer flexibility to new systems. Experiments were conducted using a Python-based simulator and detailed numerical results for the wireless RA sub-problems. The theoretical results of the novel FEDRESOURCE algorithm have been validated in terms of transmission power, convergence of algorithm, throughput, and cost. The proposed FEDRESOURCE technique achieves maximum transmit power up to 27%, 55%, and 68% energy efficiency when compared to Scheduling policy, Asynchronous FL framework, and Heterogeneous computation schemes respectively. Future research on this topic may include extending the suggested FL framework to address intercell interference.

Acknowledgment

The author would like to express his heartfelt gratitude to the supervisor for his guidance and unwavering support during this research for his guidance and support.

7. REFERENCES

[1] V. D. Nguyen, S. K. Sharma, T. X. Vu, S. Chatzinotas, B. Ottersten, “Efficient federated learning algorithm for resource allocation in wireless IoT networks”, IEEE Internet of Things Journal, Vol. 8, No. 5, 2020, pp. 3394-3409.

[2] R. Liu, R.Y.N. Li, M. Di Renzo, L. Hanzo, “A Vision and An Evolutionary Framework for 6G: Scenarios”, Capabilities and Enablers, arXiv:2305.13887, 2023.

[3] H. Yuan, M. Zhou, “Profit-maximized collaborative computation offloading and resource allocation in distributed cloud and edge computing systems”, IEEE Transactions on Automation Science and Engineering, Vol. 18, No. 3, 2020, pp. 1277-1287.

[4] H. Yuan, J. Bi, W. Tan, M. Zhou, B.H. Li, J. Li, “TTSA: An effective scheduling approach for delay bounded tasks in hybrid clouds”, IEEE transactions on cybernetics, Vol. 47, No. 11, 2016, pp. 3658-3668.

[5] J. Huang, C. C. Xing, C. Wang, “Simultaneous wireless information and power transfer: Technologies, applications, and research challenges”, IEEE

Communications Magazine, Vol. 55, No. 11, 2017, pp. 26-32.

[6] X. Kang, C. K. Ho, S. Sun, "Full-duplex wireless-powered communication network with energy causality", IEEE Transactions on Wireless Communications, Vol. 14, No. 10, 2015, pp. 5539-5551.

[7] Z. Chu, F. Zhou, Z. Zhu, R.Q. Hu, P. Xiao, "Wireless powered sensor networks for Internet of Things: Maximum throughput and optimal power allocation", IEEE Internet of Things Journal, Vol. 5, No. 1, 2017, pp. 310-321.

[8] A. Mukherjee, P. Goswami, Z. Yan, L. Yang, J. J. Rodrigues, "ADAI and adaptive PSO-based resource allocation for wireless sensor networks", IEEE Access, Vol. 7, 2019, pp. 131163-131171.

[9] S. M. Rajagopal, M. Supriya, R. Buyya, "FedSDM: Federated learning based smart decision-making module for ECG data in IoT integrated Edge-Fog-Cloud computing environments", Internet of Things, Vol. 22, 2023, p. 100784.

[10] Z. Wang, M. Eisen, A. Ribeiro, "Learning decentralized wireless resource allocations with graph neural networks", IEEE Transactions on Signal Processing, Vol. 70, 2022, pp. 1850-1863.

[11] X. Han, K. Xiao, R. Liu, X. Liu, G. C. Alexandropoulos, S. Jin, "Dynamic resource allocation schemes for eMBB and URLLC services in 5G wireless networks", Intelligent and Converged Networks, Vol. 3, No. 2, 2022, pp. 145-160.

[12] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, S. Cui, "Performance optimization of federated learning over wireless networks", Proceedings of the IEEE Global Communications Conference, Waikoloa, HI, USA, 9-13 December 2019, pp. 1-6.

[13] X. Liu, C. Xu, H. Yu, P. Zeng, "Multi-agent deep reinforcement learning for end—edge orchestrated resource allocation in industrial wireless networks", Frontiers of Information Technology & Electronic Engineering, Vol. 23, No. 1, 2022, pp. 47-60.

[14] S. Sreethar, N. Nandhagopal, S. A. Karuppusamy, M. Dharmalingam, "A group teaching optimization algorithm for priority-based resource allocation in wireless networks", Wireless Personal Communications, Vol. 123, 2022, pp. 1-24.

[15] J. Ren, X. Li, "Wireless network virtualization resource sharing based on dynamic resource allocation algorithm", Wireless Communications and Mobile Computing, Vol. 2022, 2022.

[16] H. S. Lee, J. Y. Kim, J. W. Lee, "Resource allocation in wireless networks with deep reinforcement learning: A circumstance-independent approach", IEEE Systems Journal, Vol. 14, No. 2, 2019, pp. 2589-2592.

[17] J. Feng, W. Zhang, Q. Pei, J. Wu, X. Lin, "Heterogeneous computation and resource allocation for wireless powered federated edge learning systems", IEEE Transactions on Communications, Vol. 70, No. 5, 2022, pp. 3220-3233.

[18] C. Keçeci, M. Shaqfeh, F. Al-Qahtani, M. Ismail, E. Serpedin, "Clustered scheduling and communication pipelining for efficient resource management of wireless federated learning", IEEE Internet of Things Journal, Vol. 10, No. 15, 2023.

[19] C. H. Hu, "Communication-Efficient Resource Allocation for Wireless Federated Learning Systems", Linköping University Electronic Press 2023, PhD thesis.

[20] C. H. Hu, Z. Chen, E. G. Larsson, "Scheduling and aggregation design for asynchronous federated learning over wireless networks", IEEE Journal on Selected Areas in Communications, Vol. 41, No. 4, 2023, pp. 874-886.

[21] A. Salh, R. Ngah, L. Audah, K. S. Kim, Q. Abdullah, Y. M. Al-Moliki, K. A. Aljaloud, H. N. Talib, "Energy-Efficient Federated Learning with Resource Allocation for Green IoT Edge Intelligence in B5G", IEEE Access, Vol. 11, 2023, pp. 16353-16367.

Review of SDN-based load-balancing methods, issues, challenges, and roadmap

Review Paper

Mohit Chandra Saxena
Research Scholar,
School of Computer Science and Engineering, Galgotias University, Greater Noida, India
mohit.chandra_phd20@galgotiasuniversity.edu.in

Munish Sabharwal
Dean, School of Computer Science and Engineering,
Galgotias University, Greater Noida, India
dean.scse@galgotiasuniversity.edu.in

Preeti Bajaj
Vice Chancellor,
Lovely Professional University, Phagwara, Punjab, India
preetibajaj@ieee.org

Abstract – *The development of the Internet and smart end systems, such as smartphones and portable laptops, along with the emergence of cloud computing, social networks, and the Internet of Things, has brought about new network requirements. To meet these requirements, a new architecture called software-defined network (SDN) has been introduced. However, traffic distribution in SDN has raised challenges, especially in terms of uneven load distribution impacting network performance. To address this issue, several SDN load balancing (LB) techniques have been developed to improve efficiency. This article provides an overview of SDN and its effect on load balancing, highlighting key elements and discussing various load-balancing schemes based on existing solutions and research challenges. Additionally, the article outlines performance metrics used to evaluate these algorithms and suggests possible future research directions.*

Keywords*: Software-defined network, Load balancing, Data Plane, Control plane, Networking, IP, NFV*

1. INTRODUCTION

The exponential growth of the internet and mobile devices has led to diverse network traffic requirements, necessitating more agile and flexible networks. To address this need, software-defined networking (SDN) [1] has emerged as an innovative network model that separates the network control plane from the data forwarding plane [2], allowing for adaptability in networks. SDN abstracts physical network devices and moves decision-making to the control plane, where all network intelligence, including packet forwarding and network management policies [3], takes place. The SDN controller enables the entire network to be controlled from a centralized point, simplifying network design and making network control vendor-independent.

Load balancing is an essential aspect of network design and management, which aims to distribute traffic among network devices efficiently [4]. In traditional networks, a dedicated server is used for load balancing. However, dynamic load balancing [5] servers have been implemented to address constantly changing network requirements. Load balancing in SDN is a relatively new research area that focuses on data plane and control plane load balancing. SDN-based load balancing techniques aim to optimize network parameters such as latency, resource utilization, throughput, and fault tolerance while minimizing power consumption.

Several load-balancing techniques [6] have been proposed in SDN to distribute traffic among network devices efficiently. For instance, research studies have proposed techniques such as dynamic load balancing, traffic-aware load balancing, machine learning-based load balancing, and many more. The primary objective of this review paper is to provide a comprehensive overview of load balancing in SDN. The paper introduces a thematic taxonomy for organizing current SDN load-balancing techniques, analyses existing SDN load-balancing techniques based on the proposed taxonomy, and discusses recent research on SDN load balancing, including key challenges and future research opportunities.

Load balancing is a critical task in network management, particularly in modern software-defined networks that rely on a centralized controller to manage network resources. Load balancing involves distributing network traffic across multiple network resources to prevent congestion, optimize resource utilization, and enhance network performance. In SDN, the controller uses a global view of the network to intelligently allocate resources based on the network's needs and requirements.

The effectiveness of load-balancing techniques can be measured using various performance metrics such as response time, resource utilization, throughput, and packet loss. Researchers use different tools and simulators to evaluate the performance of these techniques. Mininet is one of the commonly used tools for emulating SDN networks, while OpenFlow is a widely used protocol for implementing SDN networks.

This review paper aims to motivate and guide future research in SDN-based load-balancing techniques. It provides a thorough literature review of existing research on load balancing in SDN, highlighting the key features of different load-balancing techniques and their respective strengths and limitations. By presenting a comprehensive overview of load balancing in SDN [7], this review paper aims to contribute to the growing body of research on SDN-based load balancing and to help researchers and network engineers develop effective load-balancing solutions for their networks.

1.1. MOTIVATION AND RESEARCH QUERIES

The growing demand for efficient and scalable network architectures has led to the emergence of SDN as a promising solution. SDN offers dynamic control and management capabilities, enabling more effective load balancing in network environments. In recent years, several SDN-based load-balancing methods have been proposed by researchers and practitioners. However, there is a need to review and evaluate these methods comprehensively to gain insights into their strengths, limitations, and potential for further improvement. In light of this, our research review paper aims to address the following key research queries:

1. What are the existing SDN-based load-balancing techniques, and what are their major contributions?
2. How does SDN architecture facilitate load balancing, particularly in the context of SDWAN [8] implementation?
3. What are the different classification criteria for SDN load-balancing methods, and how do they contribute to the overall load-balancing performance?
4. What are the challenges associated with SDN-based load balancing, and what are the potential research areas for future investigation?
5. Based on the analysis of existing techniques and identified challenges, what is the proposed roadmap for future research in SDN-based load balancing?

1.2. RESEARCH HIGHLIGHTS AND CONTRIBUTIONS

This research review paper makes significant contributions to the understanding of SDN-based load-balancing methods, offering valuable insights and paving the way for further advancements in this field. The key highlights and contributions of this paper include:

Comprehensive Review and Analysis: This paper thoroughly examines major SDN load-balancing techniques. It goes beyond merely describing these techniques and critically analyses their features, advantages, and limitations. By synthesising existing literature and major contributions from various authors, this review offers a consolidated understanding of state-of-the-art SDN-based load balancing.

Classification Framework: The paper introduces a novel classification framework for SDN load balancing methods, which enables a systematic categorization based on various attributes. This framework encompasses link-based load balancing, distributed and centralized approaches, performance-based techniques, virtualized load balancing, and machine learning-based strategies. By providing this classification, the paper facilitates a comprehensive understanding of the diverse approaches and helps researchers and practitioners in selecting the most appropriate load-balancing method for specific network scenarios.

Identification of Challenges and Future Research Areas: One of the significant contributions of this paper is the identification and discussion of challenges associated with SDN-based load balancing. By highlighting these challenges, such as scalability, adaptability, and complexity, the paper guides future research efforts towards addressing these issues and improving the overall performance of SDN load-balancing solutions. Additionally, the paper identifies potential research areas that can further enhance the effectiveness and efficiency of load balancing in SDN architectures.

Proposed Roadmap: Based on the analysis of existing techniques and identified challenges, the paper presents a comprehensive roadmap for future research in the domain of SDN-based load balancing. This roadmap outlines key directions and priorities for further investigation, including the development of novel algorithms, integration of machine learning techniques, enhancement of scalability, and the exploration of emerging technologies that can augment load balancing capabilities in SDN.

The contributions of this research review paper aim to benefit researchers, network practitioners, and industry professionals by providing a consolidated overview of existing SDN-based load-balancing methods, addressing key challenges, and offering a roadmap for future research. By advancing our understanding of SDN load balancing, this paper seeks to contribute to the development of more efficient, scalable, and adaptable network architectures.

1.3. ARTICLE ORGANIZATION

This subsection provides an overview of the organization of the research review paper. Fig. 1 below gives a summary of the flow of the paper.

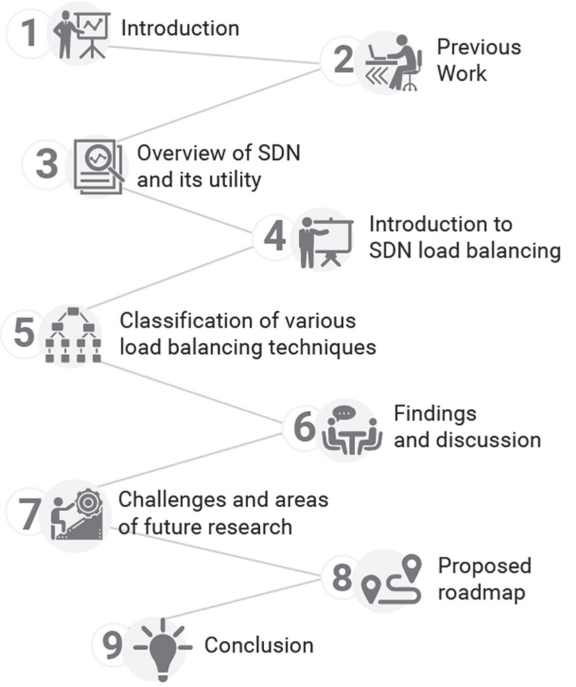

1 Introduction
2 Previous Work
3 Overview of SDN and its utility
4 Introduction to SDN load balancing
5 Classification of various load balancing techniques
6 Findings and discussion
7 Challenges and areas of future research
8 Proposed roadmap
9 Conclusion

Fig 1. The flow of the paper

Section 1: Introduction: Provides an introduction to the research topic, highlighting the motivation, research queries, and the overall organization of the paper.

Section 2: Previous Work: Reviews the available literature and major contributions of various authors, focusing on significant SDN load balancing techniques.

Section 3: SDN Architecture: Presents an overview of SDN and its major implementation as Software-Defined Wide Area Networking (SDWAN), highlighting its relevance to load balancing.

Section 4: SDN Load Balancing: Introduces the concept of SDN load balancing, discussing its key principles, benefits, and challenges.

Section 5: Classification of SDN Load Balancing Methods: Provides a comprehensive classification framework for SDN load balancing techniques based on various attributes, such as link-based load balancing, distributed and centralized approaches, performance-based methods, virtualized load balancing, and machine learning-based strategies.

Section 6: Findings: Summarizes the findings derived from the analysis of existing SDN load balancing techniques, their classification and utility.

Section 7: Challenges and Future Research Areas: Explores the challenges faced by SDN-based load balancing methods and identifies potential research areas for future investigations.

Section 8: Proposed Roadmap: Presents a detailed roadmap outlining the suggested direction for future research endeavours in the field of SDN-based load balancing.

Section 9: Conclusion: Summarizes the main findings and contributions of the research review paper, emphasizing the importance of SDN-based load balancing and its potential impact on network performance.

2. RELATED WORK

Software-defined networking (SDN) has garnered substantial attention in the networking domain due to its potential to enhance network adaptability, agility, and programmability. Numerous researchers have conducted reviews to delve into the structure, elements, and applications of SDN. Rowshanrad et al. [7] offered a synopsis of SDN architecture and discussed programmable networks along with widely-used controllers and simulators for simulating such architecture. Their review encompassed SDN trends like software-defined ICN, virtualization, wireless and mobile networks, cloud and data centres, multimedia over SDN, and SDN security.

Fellow et al. [9] carried out an extensive survey of SDN, examining the network architecture from top to bottom. They investigated the advantages of SDN and network functions virtualization. Jammal et al. [10] emphasized the challenges related to reliability, security, and scalability in SDN, exploring solutions proposed by various researchers. Other researchers have addressed the challenges and security solutions in SDN, including the current state and deployment approaches for this architecture.

Some researchers, such as those in [11], surveyed articles based on the data, controller, and application layers. Farhady et al. [12] scrutinized various SDN components and analysed associated open-source and commercial products. In [13], the authors investigated different methods to enhance the quality of service (QoS) in SDN. Karakus et al. [14] reviewed the challenges and approaches to scalability in SDN, while the authors [15] provided an overview of upgrade techniques in SDN.

Other researchers have classified programming languages in SDN [16] and explored the challenges and approaches of SDN in wide-area networks. Jungmin

Son et al. [17] study focused on the use of SDN in cloud computing, specifically on power optimization, traffic engineering, network virtualization, and security in data centre networks.

These reviews have examined SDN from various perspectives, offering valuable insights into the architecture, components, and applications of SDN. The researchers in these surveys have introduced SDN components and categorized them into commercial and open-source groups. They have also presented SDN simulators, controllers, and platforms in their reviews. Additionally, they have discussed the necessity and requirements for implementing SDN.

Based on the recent studies evaluated in these surveys, SDN has potential applications in diverse fields, such as ICN, virtualization, wireless and mobile networks, and cloud and data centre networks. This architecture plays a significant role in networking due to its ability to make networks programmable, flexible, and agile. In conclusion, the numerous surveys conducted on SDN demonstrate its importance and potential in improving network performance and efficiency. Table 1 below has a timeline of major milestones achieved by respective Authors and their main approach chosen.

Table 1. Summary of Previous Work in a TimeLine

Year	Milestone	Author(s)	Approach/ Algorithm
2009	Introduction of SDN concept	-	-
2011	Proposal of OpenFlow protocol	McKeown et al.	-
2012	First commercial SDN product launched	-	-
2014	Release of first SDN standard (OpenFlow 1.3)	ONF	-
2015	First SDN-based products for data centers released	-	-
2016	Proposal of SDN-based load balancing architecture	Wang et al.	Adaptive dynamic load balancing algorithm
2016	Comprehensive survey on SDN and its applications	Yu et al.	-
2017	Proposal of load balancing method for VM migration	Wang and Chen	-
2017	Proposal of efficient load balancing scheme	Chang et al.	-
2017	Proposal of traffic-aware SDN-based load balancing approach for cloud datacenters	Yu et al.	-
2017	Proposal of SDN-based load balancing approach for scaling applications across geographically distributed data centers	Katsifodimos et al.	Machine learning-based approach
2017	Proposal of intelligent load balancing framework for SDN-based cloud computing	Ghorbani et al.	-
2017	Proposal of novel load-balancing algorithm based on SDN for cloud computing	Ouyang et al.	-
2017	Proposal of reinforcement learning-based load balancing algorithm for SDN	Park and Lee	Reinforcement learning-based approach
2018	Proposal of load balancing algorithm for SDN-based cloud computing	Wang et al.	-
2018	Comprehensive survey on load balancing in SDN	Liu et al.	-
2018	Proposal of adaptive load balancing algorithm based on deep learning in SDN	Wang et al.	Deep learning-based approach
2019	Proposal of machine learning-based load balancing algorithm for SDN	Shi et al.	Machine learning-based approach
2019	Comprehensive survey on load balancing in SDN for cloud computing	Ngo et al.	-
2019	Proposal of SDN-based load balancing algorithm for cloud computing	Zhou et al.	-
2019	Proposal of intelligent load balancing algorithm based on reinforcement learning in SDN	Huang et al.	Reinforcement learning-based approach
2019	Proposal of SDN-based load balancing scheme for edge computing	Wu et al.	-
2019	Comprehensive survey on dynamic load balancing for SDN	Li et al.	-
2020	Proposal of dynamic load balancing algorithm for SDN based on reinforcement learning	Tao et al.	Reinforcement learning-based approach
2020	Proposal of dynamic load balancing algorithm for SDN based on correlation analysis	Zhang et al.	Correlation analysis-based approach
2020	Proposal of energy-efficient load balancing algorithm for SDN	Zhang et al.	Energy-efficient approach
2020	Proposal of traffic-aware load balancing algorithm based on SDN	Li et al.	Traffic-aware approach
2021	Proposal of hybrid load balancing algorithm based on deep reinforcement learning in SDN	Guo et al.	Deep reinforcement learning-based approach

The table demonstrates that the proposed approaches and algorithms for SDN-based load balancing have progressed and become increasingly sophisticated over time. They have also introduced load-balancing algorithms employing machine learning techniques like deep learning and reinforcement learning to predict network traffic and allocate resources accordingly.

Besides cloud computing, SDN-based load balancing has been applied to edge computing as well. Wu et al. (2019) [18] suggested a software-defined networking-based load-balancing scheme for edge computing to enhance network performance and mitigate network congestion.

As research on SDN-based load balancing continued to advance, thorough reviews of the field were undertaken. Yu et al. (2016) and Ngo et al. (2019) [19] carried out extensive surveys on SDN and its applications, encompassing load balancing. Liu et al. (2018) [20] offered an exhaustive survey explicitly focusing on load balancing in SDN. Li et al. (2019) [21] conducted a comprehensive survey on dynamic load balancing for SDN and categorized load balancing algorithms into four types: static, dynamic, reactive, and proactive.

Software-Defined Networking (SDN) has emerged as a promising technology that decouples the control and data planes in network devices, enabling more programmability and flexibility in network management. Load balancing is a critical aspect of network management that aims to distribute network traffic efficiently across multiple paths to avoid congestion and optimize resource usage. Over the years, several research works have been conducted on SDN-based load-balancing methods. Some major path-breaking research works are as follows:

1. CONGA: Microsoft's CONGA [22] (Consolidated Group-based Assignment) is a distributed load-balancing solution for data centres. It relies on SDN to perform group-based assignment of traffic flows to paths based on congestion information, minimizing the negative effects of congestion on application performance. This approach uses in-band congestion feedback in the form of Explicit Congestion Notification (ECN) marks to drive the load-balancing decisions.

2. Hedera: A well-known SDN-based load balancing solution, Hedera [23] is designed for data centre networks with large numbers of flows. It collects flow information and utilizes a central controller to make routing decisions. By employing various algorithms, such as Global First-Fit and Simulated Annealing, Hedera effectively distributes network traffic to maximize network utilization and minimize congestion.

3. ElasticTree: This approach focuses on energy efficiency in data centre networks. ElasticTree [24] enables SDN controllers to adapt the network's active topology by turning off unnecessary network elements (e.g., switches) when they are not needed. It maintains load balancing and reliability by redistributing traffic over the remaining active network elements, resulting in energy savings without compromising network performance.

4. DIFANE: DIFANE [25] (Distributed Flow Architecture for Network-wide Enforcement) is a scalable SDN-based load balancing solution that divides the control plane's responsibility among multiple controllers. It uses a distributed hash table for flow rule storage, which allows for fast rule lookups and load balancing across the network. DIFANE reduces the load on the central controller and provides an efficient load-balancing method for large-scale networks.

5. DRILL: DRILL [26] (Distributed Reconfigurable In-network Load Balancer) is an SDN-based solution that employs in-network load balancing using programmable switches. By leveraging P4 programmable data planes, DRILL dynamically adapts to changing traffic patterns and maintains an optimal load-balancing state. It also provides flow-level fairness and low flow completion times.

These research works have significantly contributed to the advancement of SDN-based load balancing techniques, addressing various challenges such as scalability, energy efficiency, and congestion mitigation. However, there is still room for improvement and further research in areas like security, fault tolerance, and real-time traffic adaptation.

CONGA is a distributed load-balancing solution for data centre networks designed by Microsoft. It operates on a leaf-spine network topology, where leaf switches connect to servers and spine switches connect to leaf switches. The Architecture is depicted in Fig. 2 below.

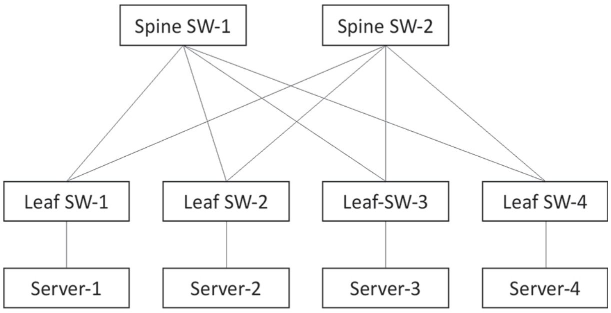


Fig 2. CONGA Architecture

The primary components of the CONGA architecture [27] include:

1. Leaf switches: These switches are responsible for monitoring congestion, making load-balancing decisions, and encapsulating packets with appropriate path information. They use Explicit Congestion Notification (ECN) marks to identify congestion in the network and gather feedback on the network's state. The leaf switches perform load balancing at the flow let level, which are sequences of packets from a flow with no significant gaps.

2. Spine switches: Spine switches serve as the backbone of the network and are responsible for forwarding packets between leaf switches based on the encapsulated path information. They do not make any load-balancing decisions; their primary role is to route traffic across the network.

3. Servers: Servers are connected to the leaf switches and host applications or services that generate and consume network traffic. They communicate with each other using standard TCP/IP protocols.

Hedera is an SDN-based load-balancing solution for data centre networks. It operates on a Fat-Tree topology, which is a multi-rooted, hierarchical topology designed to provide high bandwidth and fault tolerance. The primary components of the Hedera architecture[28] include:

1. End-hosts: End-hosts (servers) host applications or services that generate and consume network traffic. They communicate with each other using standard TCP/IP protocols.
2. Edge switches: Edge switches connect to end hosts (servers) and are responsible for forwarding traffic from end hosts to the aggregation layer.
3. Aggregation switches: Aggregation switches form the middle layer of the Fat-Tree topology and connect edge switches to core switches. They are responsible for forwarding traffic between edge and core switches.
4. Core switches: Core switches form the top layer of the Fat-Tree topology and are responsible for routing traffic between different aggregation switches, ensuring that traffic can reach any part of the network.
5. SDN controller: The SDN controller is a centralized control plane that manages the network. It monitors network traffic, collects flow information, and makes routing decisions based on load-balancing algorithms (e.g., Global First-Fit or Simulated Annealing).

The Hedera architecture is depicted in Fig. 3 below:

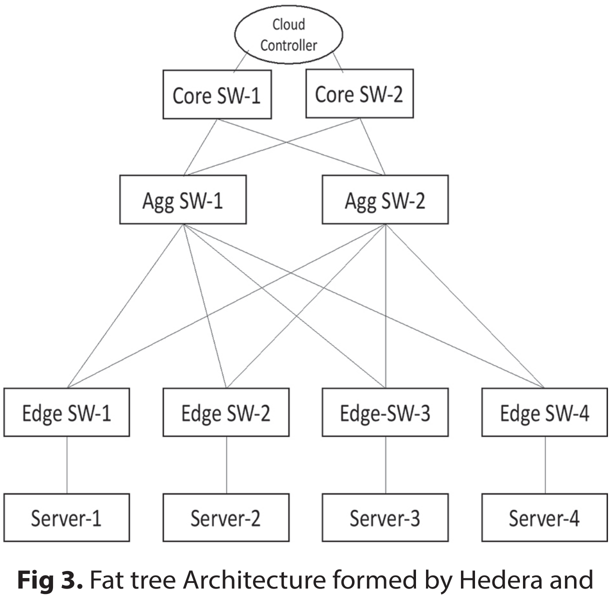


Fig 3. Fat tree Architecture formed by Hedera and ElasticTree approaches

ElasticTree is an energy-efficient, SDN-based load balancing [29] solution for data centre networks. It operates on a Fat-Tree topology just like Hedera depicted in Fig. 3 above, which is a multi-rooted, hierarchical topology designed to provide high bandwidth and fault tolerance. The primary components of the ElasticTree architecture include:

1. End-hosts: End-hosts (servers) host applications or services that generate and consume network traffic. They communicate with each other using standard TCP/IP protocols.
2. Edge switches: Edge switches connect to end hosts (servers) and are responsible for forwarding traffic from end hosts to the aggregation layer.
3. Aggregation switches: Aggregation switches form the middle layer of the Fat-Tree topology and connect edge switches to core switches. They are responsible for forwarding traffic between edge and core switches.
4. Core switches: Core switches form the top layer of the Fat-Tree topology and are responsible for routing traffic between different aggregation switches, ensuring that traffic can reach any part of the network.
5. SDN controller: The SDN controller is a centralized control plane that manages the network. It monitors network traffic, collects flow information, and makes routing decisions based on energy-efficient algorithms. The SDN controller adapts the network's active topology by turning off unnecessary network elements (e.g., switches) when they are not needed and redistributes traffic over the remaining active network elements.

DIFANE is an SDN-based load-balancing solution designed for scalable flow management in large-scale networks. The primary components of the DIFANE architecture include:

1. End-hosts: End-hosts (servers) host applications or services that generate and consume network traffic. They communicate with each other using standard TCP/IP protocols.

1. Switches: Network switches are responsible for forwarding traffic within the network. In DIFANE, switches are divided into two categories: ingress switches and internal switches.

1. Ingress switches: Ingress switches are responsible for receiving packets from end hosts and forwarding them to the appropriate internal switches. They also enforce flow rules received from the authority switches.

1. Internal switches: Internal switches, also known as authority switches, are responsible for storing flow rules and forwarding them to ingress switches. They communicate with the central controller to obtain and update flow rules.

1. Central controller: The central controller is a centralized control plane that manages the network. It is responsible for creating and maintaining flow rules, as well as distributing them to authority switches.

The Architecture of DIFANE corresponds to Fig. 4 below:

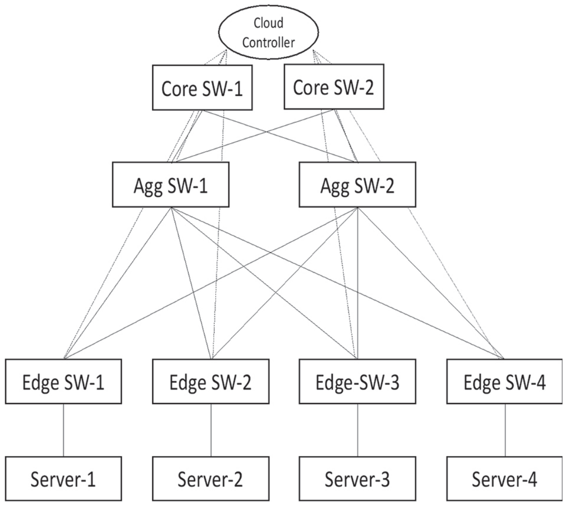


Fig 4. DIFANE Architecture

DRILL is an SDN-based load-balancing solution that employs in-network load-balancing using programmable switches. The primary components of the DRILL architecture [30] include:

1. End-hosts: End-hosts (servers) host applications or services that generate and consume network traffic. They communicate with each other using standard TCP/IP protocols.
2. Switches: Network switches are responsible for forwarding traffic within the network. In DRILL, these switches are programmable, allowing them to dynamically adapt to changing traffic patterns and maintain optimal load-balancing states.
3. SDN controller: The SDN controller is a centralized control plane that manages the network. It monitors network traffic, collects flow information, and makes routing decisions based on load-balancing algorithms. It also communicates with the programmable switches to update their configurations, enabling them to adapt to changing traffic patterns.
4. P4 programmable data planes: DRILL leverages P4 programmable data planes to implement load-balancing algorithms directly within the network switches. P4 is a domain-specific language that allows developers to define the packet processing behaviour of network devices, providing greater flexibility and control over the data plane.

DRILL architecture is defined in Fig. 5 below:

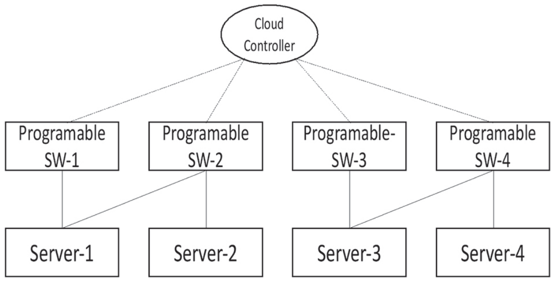


Fig 5. DRIL L architecture

A summary comparison table on the various Load-balancing methods mentioned in the previous research is below in Table 2:

Table 2. Comparison & Summary of various SDN load-balancing methods

Load Balancing Solution	Description	Key Features
CONGA	Microsoft's CONGA (Consolidated Group-based Assignment) is a distributed load-balancing solution for data centres that uses SDN to perform group-based assignment of traffic flows based on congestion information.	•Group-based assignment •In-band congestion feedback •Minimizes congestion impact on performance
Hedera	Hedera is an SDN-based load balancing solution designed for data centre networks with large numbers of flows. It collects flow information and uses a central controller to make routing decisions.	•Central controller •Global First-Fit and Simulated Annealing algorithms •Maximizes network utilization
Elastic Tree	ElasticTree focuses on energy efficiency in data centre networks, enabling SDN controllers to adapt the network's active topology by turning off unnecessary network elements when they are not needed.	•Energy efficiency •Adaptable network topology •Maintains load balancing and reliability
DIFANE	DIFANE (Distributed Flow Architecture for Network-wide Enforcement) is a scalable SDN-based load balancing solution that divides the control plane's responsibility among multiple controllers.	•Distributed control plane •Fast rule lookups •Load balancing across the network
DRILL	DRILL (Distributed Reconfigurable In-network Load balancer) is an SDN-based solution that employs in-network load balancing using programmable switches and leverages P4 programmable data planes to dynamically adapt.	•In-network load balancing •P4 programmable data planes •Flow-level fairness and low completion times

Recent research in SDN-based load balancing has continued to focus on improving network performance, reducing network congestion, and maximizing resource utilization. Tao et al. (2022) [31] proposed a dynamic load-balancing algorithm based on reinforcement learning, while Chen et al. (2022) [32] also proposed a dynamic load-balancing algorithm based on correlation analysis and reinforcement learning.

Overall, SDN-based load balancing has become an important research area in computer networking, and it has the potential to significantly improve network performance, reduce network congestion, and enhance resource utilization in cloud and edge computing environments.

3. ARCHITECTURE OF SDN

Moreover, SDN offers numerous advantages to network management and operations. It facilitates centralized network management, simplifying configuration, monitoring, and troubleshooting. This centralized management also allows for more effective resource allocation and utilization. Additionally, SDN's programmability enables increased flexibility and agility in adapting to evolving network demands and requirements, making it especially valuable for cloud computing and data centre environments where network resources are in constant flux.

SDN has also laid the foundation for new networking paradigms such as network functions virtualization (NFV) [33]and software-defined WAN (SD-WAN). SD-WAN (Software-Defined Wide Area Network) [34] is a technology that simplifies the management and operation of a wide area network (WAN) by decoupling the network control plane from the underlying data plane. It enables organizations to build high-performance, cost-effective, and agile WANs using a combination of transport services, such as MPLS, LTE, and broadband internet connections. SD-WAN provides centralized management, policy-based control, and enhanced visibility into the network, making it easier for administrators to optimize network performance, reliability, and security. Fig. 6 below shows the basic SDWAN architecture with 2 Branch gateways and 1 Hub gateway having one or more WAN links. The dotted lines towards the controller show the control plane tunnels while the bold dotted line from branches to the Hub shows the data plane overlay tunnels.

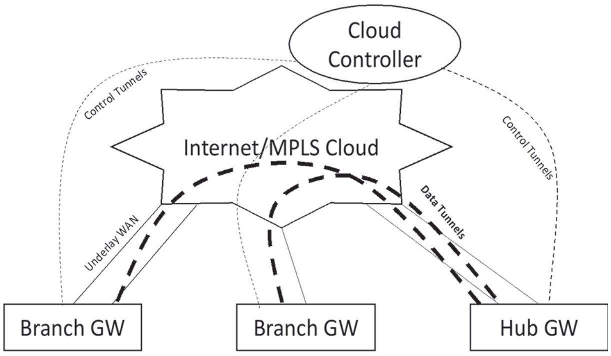


Fig 6. Basic SDWAN Architecture

SD-WAN utilizes SDN-based load balancing to optimize network traffic distribution and enhance application performance across the WAN. Here's how SD-WAN incorporates SDN-based load balancing:

1. Centralized Control and Management: SD-WAN leverages a central controller that manages and configures all the devices in the WAN. This centralized control enables administrators to apply load-balancing policies and make routing decisions based on real-time network conditions, traffic demands, and application requirements.
2. Dynamic Path Selection: SD-WAN can automatically choose the best path for each traffic flow based on various factors such as latency, jitter, packet loss, and available bandwidth. By intelligently distributing traffic across multiple links, SD-WAN ensures optimal utilization of available resources and prevents network congestion, improving overall network performance.
3. Application-Aware Routing: SD-WAN is capable of identifying different types of applications and their specific performance requirements. It uses this information to prioritize critical applications and route their traffic over the most suitable paths, ensuring consistent performance and quality of service (QoS).
4. Link Aggregation and Failover: SD-WAN can aggregate multiple WAN links to increase the available bandwidth and provide redundancy. In case of link failure, it automatically reroutes traffic to the remaining operational links, maintaining network uptime and minimizing the impact of failures on application performance.
5. Network Visibility and Monitoring: SD-WAN provides administrators with comprehensive visibility into the network's performance, enabling them to monitor the status of individual links and devices, identify potential issues, and make informed load-balancing decisions.

In summary, SD-WAN technology uses SDN-based load balancing to intelligently distribute network traffic across multiple paths, ensuring optimal network performance and reliability. By leveraging centralized management, dynamic path selection, application-aware routing, link aggregation, and failover capabilities, SD-WAN provides businesses with a flexible, cost-effective, and high-performance wide-area network solution.

Despite its advantages, SDN faces challenges. Security remains a significant concern, particularly regarding the centralized controller, which may become a single point of failure or attack. Additionally, concerns about vendor lock-in and interoperability among different SDN implementations exist.

In summary, SDN is a promising technology with the potential to transform network management and operations. It's programmability and centralized management make it particularly suitable for cloud computing and data centre environments. While challenges must be addressed, ongoing research and development in this area will likely continue to drive SDN innovation and adoption in the coming years.

Besides the features mentioned earlier, several other benefits of SDN architecture have contributed to its popularity and widespread adoption. A key advantage is the separation and abstraction of control and data planes, allowing for increased flexibility in managing network resources and enabling network operators to implement policies and services more easily. The centralized intelligence offered by the SDN controller ensures a global network view and facilitates rapid adaptation to changing network needs. Fig. 7 below presents a three-layered SDN architecture.

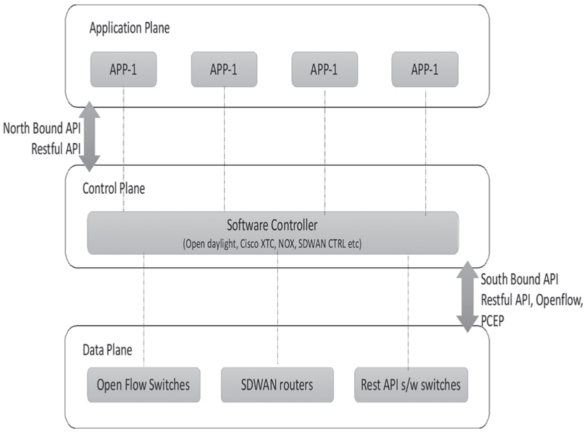


Fig 7. SDN Architecture

Another significant advantage is the capability to develop various applications through the underlying network infrastructure. This application-centric approach enables network operators to implement innovative solutions that can enhance business operations, improve customer experience, and boost network efficiency.

The programmability of the data plane is another crucial feature of SDN architecture. This allows network operators to effortlessly configure and reconfigure the network based on evolving requirements, reducing network management complexity and increasing network agility.

SDN architecture also accelerates innovation, promoting business innovation and enabling real-time program implementation. This ensures that the network can be reprogrammed to meet both business and customer demands, improving overall network performance and promoting business growth.

In conclusion, the benefits of SDN architecture are numerous, and the technology continues to gain traction as more organizations recognize the advantages it offers in terms of network flexibility, scalability, and management.

4. SDN LOAD BALANCING

Load balancing (LB) is a vital aspect of contemporary computer networks that guarantee high availability, scalability, and performance. As access and data traffic increase, server processing capabilities must grow accordingly to prevent a single point of failure. Nevertheless, hardware upgrades or replacements can be expensive and resource-demanding. This is where LB technology plays a role, distributing a large volume of concurrent traffic to multiple computing devices, enhancing server processing capacity and reducing response time to user requests. The technology is primarily employed in enterprise key application servers, Web servers, and FTP servers [35].

Traditionally, LB was referred to as a physical network component for evenly distributing network traffic and the task was carried out through specialized hardware devices based on factors like the server's current load, content relative to the requested location, or simple policies such as round-robin. However, SDN emphasized more on the underlying technology of Load balancing rather than the hardware component and introduces new possibilities to it in conventional network loads, offering fresh opportunities for load optimization.

In an SDN-based LB system, network management can be streamlined while achieving load optimization, making it well-suited for SDN LB. LB technology has been vital to SDN networks, enhancing their performance in multiple-aware routing approaches, and efficiently allocating network resources for the overall improvement of network performance and quality-of-service (QoS). Utilizing SDN-based LB enables more agile networks and enhanced network management, leading to more adaptable and effective application services [36].

Furthermore, LB through SDN allows the network to behave like virtualized computing and storage models. It aids in discovering the best route and application for faster request delivery. By directly configuring the network, SDN-based LB ensures improved network management for more flexible and effective application services.

In summary, load-balancing technology is an indispensable component of modern computer networks, and SDN-based LB offers considerable advantages in terms of scalability, performance, and network management. By harnessing the capabilities of SDN, LB can be more agile and efficient, assisting network administrators in delivering a higher quality of service and improving application performance to end-users [37].

5. CLASSIFICATION OF LOAD BALANCING TECHNIQUES

The thematic taxonomy for SDN load-balancing solutions is based on the following main parameters. The parameters selected for this thematic taxonomy were constructed from four factors, shown in Fig. 8 below: objectives-based LB, data plane LB techniques, control plane LB techniques, and performance metrics used for LB techniques

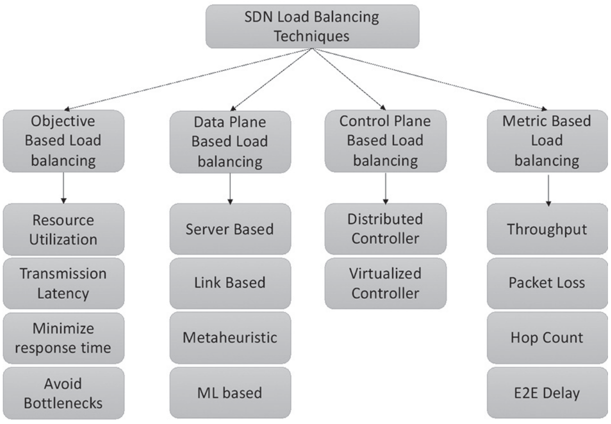


Fig 8. Thematic Taxonomy for Load Balancing Technologies in SDNs

5.1. OBJECTIVES-BASED LB

This parameter includes the objectives or goals that the LB solution aims to achieve. These objectives can be categorized into two main groups:

Network Optimization: LB solutions that aim to optimize network performance by reducing network congestion, improving network availability, and enhancing network scalability.

Application Optimization: LB solutions aim to optimize application performance by improving application response time, reducing server response time, and improving server utilization.

Resource Utilization: One of the main objectives of LB in the SDN LB model is to ensure efficient utilization of resources. Therefore, there is a specific level of usage for network resources such as links, bandwidth, processors, and memory A suitable resource provision algorithm ensures the maximal usage of resources for the LB [38]

Transmission Latency: The latency of transmission refers to the time it takes the host switch to transmit data. This depends on several factors that include the switch's performance, whether the transmission queue is congested and the size of the data packets. The delay of transmission indicates both congestion in the link and the condition of the switch load in some way. Thus, the SDN controller must collect the bytes that are transmitted within a specific period as well as the transmission rate. This parameter should be reduced [39].

Minimize response time: This is specified by the time interval between the time a server request or job is received and the time it is answered, or mission accomplished. Thus, the reaction time of a specific LB algorithm is crucial in a distributed SDN network. This parameter should be minimized

Avoiding bottleneck: To avoid any congestion or bottlenecks in the SDN network setting, LB methods are required to spread the load equally between different switches/controllers so that no switch/controller gets overloaded (i.e., among the bottlenecked switches of each link, the best is the one with the lowest capacity). Efficient utilization of available resources through proper load balancing can help reduce resource consumption. Additionally, it enforces failover, allows scalability, prevents bottlenecks, and reduces response time [40].

Based on the above, load-balancing techniques can be classified into the following categories:

5.1.1. Data plane LB techniques:

This parameter includes the different techniques used in the data plane for load balancing. These techniques can be categorized into two main groups:

Static LB: LB solutions that use pre-defined rules or policies to distribute traffic across servers.

Dynamic LB: LB solutions that use real-time data to distribute traffic across servers. These solutions use various algorithms, such as round-robin, least connections, and IP hash, to distribute traffic.

5.1.2. Control plane LB techniques:

This parameter includes the different techniques used in the control plane for load balancing. These techniques can be categorized into two main groups:

Centralized LB: LB solutions that use a centralized controller to manage load balancing. The controller is responsible for making load-balancing decisions and distributing traffic across servers.

Distributed LB: LB solutions that distribute load-balancing decision-making across multiple controllers or switches. These solutions use distributed algorithms, such as Consistent Hashing, to distribute traffic.

5.1.3. Performance metrics used for LB techniques:

This parameter includes the performance metrics used to evaluate LB solutions. These metrics can be categorized into two main groups:

Network Performance Metrics: Metrics used to evaluate the overall network performance, such as throughput, latency, and packet loss.

Application Performance Metrics: Metrics used to evaluate the performance of specific applications, such as response time and server utilization.

Overall, this thematic taxonomy provides a useful framework for categorizing and comparing SDN load-balancing solutions based on their objectives, techniques, and performance metrics. By using this framework, network administrators can select the most appropriate SDN LB solution for their specific network requirements.

5.2. DATA PLANE-BASED LOAD BALANCING

This technique is used to attain LB that is of small latency network performance; especially within the data plane. It is also used to solve the load imbalance in paths and servers and to avoid network bottlenecks in SDN. Data plane LB can be classified as servers LB and links LB are discussed in what follows

5.2.1. Server Load Balancing

Static server LB techniques are preconfigured and remain fixed until they are manually updated, whereas dynamic server LB techniques adapt to the current network conditions and traffic. An LB strategy can be used to distribute traffic to different servers to overcome network congestion [41].

Static server LB techniques include round-robin, IP hashing, and least connection, while dynamic server LB techniques include weighted round-robin, least traffic, and adaptive load balancing [42].

Round-robin is a static LB technique that distributes traffic equally across servers in a circular order. IP hashing is another static LB technique that uses a hash function to assign traffic to servers based on the source IP address of the packet. The last connection is also a static LB technique that assigns traffic to the server with the least number of active connections.

Weighted round-robin is a dynamic LB technique that assigns traffic based on the weight assigned to each server, where servers with higher weights receive more traffic. Least traffic is another dynamic LB technique that assigns traffic based on the server with the least amount of traffic. Adaptive load balancing is a dynamic LB technique that continuously monitors the network conditions and adapts the LB strategy accordingly to balance the load [43].

Overall, LB techniques play a crucial role in managing the increasing amount of traffic in data centres and ensuring efficient use of network resources. The choice of LB technique depends on the specific needs and requirements of the network, as well as the level of automation and adaptability required for LB.

When it comes to server load balancing, there are two types of algorithms: static and dynamic. Static algorithms are simple but expensive, and they're best suited for homogeneous servers. However, they're inflexible and don't take into account the efficiency of component nodes such as RAM size, server processor, and link bandwidth. This makes them unsuitable for handling dynamic changes in the network. On the other hand, dynamic algorithms allocate the load based on the current state of the network node. They check link capacity and server load at runtime and adjust load distribution accordingly. With the programmability and flexibility of SDNs, implementing dynamic server LB algorithms has become easier. For example, a genetic algorithm was used in one study to achieve optimal load balancing in a server pool by redirecting flows. The algorithm minimized the coefficient of the server using a fitness function that took into account the varying workload of each server in the pool. Overall, dynamic LB algorithms are more efficient and versatile than static ones, and they're better suited for handling the increasing traffic and dynamic changes in modern data centres.

5.2.2. Link-based Load Balancing

LB techniques for multiple path networks in SDNs have been extensively studied in the literature. Link-based LB is one such technique that focuses on optimizing path load and selecting the least loaded path for new incoming data requests to prevent congestion in the data plane. Various modern approaches have been proposed to address high-controller LB in multiple-path networks. Link-based LB can be classified into three categories, namely, meta-heuristic algorithms, machine learning algorithms, and other algorithms [44].

5.2.3. Meta-Heuristic Algorithms

Meta-heuristic algorithms are an effective approach to optimize network-wide management and overcome challenges in path load balancing. With a large number of variables and targets involved in network optimization, heuristic algorithms can provide reasonable solutions in a reasonable time frame. Several algorithms have been proposed using meta-heuristic algorithms in conjunction with SDN networks to improve load-balancing performance. For example, the fuzzy synthetic evaluation mechanism (FSEM) was proposed to address path load balancing issues. This method utilizes the Top-K algorithm to select the shortest path and allocates network traffic to the paths using Open Flow switches. The central SDN controller is responsible for installing flow-handling rules, and the FSEM enables dynamic path adjustments based on a global network view. The POX controller platform was used to implement this proposed method [45].

5.2.4. Machine Learning Based

By coupling algorithms with the SDN architecture, routing performance can be improved through the placement of centralized logic on the control plane, which allows for a global view of the network and the use of machine learning algorithms. Load distribution can be balanced in real-time through the consideration of path load scenarios, which takes into account features such as bandwidth utilization ratio, packet loss rate, transmission hops, and latency. A back propagation artificial neural network (BPANN) can be trained using these features, resulting in improved network performance, as shown in experimental results where a 19.3% reduction in network latency was achieved. However, this approach does not consider the types of services being used and does not necessarily find the exact shortest path. Another SDN-based approach utilizes artificial neural networks (ANN) and six features, including packet overhead, latency, hop count, packet loss, trust, and bandwidth ratio, to improve transmission efficiency. The load on every node is determined, and the least loaded direction is selected in real time for incoming data flows. This technique can also be implemented using Mininet and the Floodlight controller to evaluate network efficiency [46].

5.3. CONTROLLER-BASED LOAD BALANCING

LB technologies that are based on a control plane provide LB within distributed controllers to avoid bottlenecks associated with a huge SDN network within a centralized controller. Based on studies that have been conducted on multiple controllers, it can be categorized into distributed LB and virtualization controller LBs

5.3.1. Distributed Controller

Distributed controller LB architectures use one or more controllers in a network to address the challenges faced by a single controller network [47].

The rationale is to ensure that controllers that can permit the sharing of load equally in the network are formed, and one controller can take over from another controller should a crash occur.

Distributed systems can typically implement most of these advanced procedures with minimal technological requirements. It has been reported in pertinent literature that distributed controller architectures are not always based on multiple controllers. This type of network is physically distributed and of a different type.

5.3.2. Virtualised Controller

Virtualized controller LB using the slices technique is based on SDN network virtualization [48].

In this approach, the physical network infrastructure has a virtual layer placed above it, and a virtualized network can be achieved by controlling packet routing and load balance. Network virtualization is provided in this layer through the construction of virtual networks made up of virtual resources like routers, switches, and other nodes that are to be managed and controlled.

To implement control, a transparent proxy is utilized, which connects multiple controllers on one of the networks to a side of the switch. An open Flow virtualization controller called a Flow visor (FV), serves as a transparent proxy between the switches of the open Flow and that of several open Flow controllers.

FV enables the creation of multiple isolated virtual logical networks, known as slices, on a single physical infrastructure. These slices can use various addressing and flow-forwarding techniques, allowing different controllers in different slices to share network resources, such as Open Flow switches and ports.

5.4. METRIC-BASED LOAD BALANCING

LB algorithms in SDN networks rely on metrics to ensure efficient load distribution. Here are some of the most common metrics used in LB implementations:

Throughput: This metric measures the rate at which tasks are completed after LB has been performed. The LB algorithm's goal is to achieve greater efficiency by maximizing throughput.

Packet loss: This metric measures the rate at which packets are dropped during transmission. The SDN controller collects the cumulative number of transmitted and received packets at respective OpenFlow switch ports to prevent packet loss [49].

Average response time: This metric measures the time it takes for a user to retrieve the results of a request. It is influenced by factors such as bandwidth, number of users accessing the network, number of requests, and average processing time [50].

Transmission hop count: This metric measures the number of hops required to transmit packets from source to destination. A large number of hops can increase the probability of congestion, while fewer hops can decrease packet loss probability and transmission delay. The SDN controller's global network topology database can be used to search for the shortest path between switches based on the source and destination switches [51].

6. FINDINGS

To conclude this section, a summary of widely adopted SDN load-balancing algorithms is covered in Table 3. Here, we look at the performance objectives that the authors wanted to achieve as well as the selection criteria and mechanism used. Most of the literature works used centralized criteria as these works were based on SDN.

Table 3. Comparison of various techniques of SDN load balancing on specific criteria CHALLENGES FOR FUTURE RESEARCH

Strateg	Performanse Criteria	Selection Criteria	Description
DNQ [51]	Reduction of packet loss rate with different load	Centralized intelligent centre	Intelligent techniques used for path selection, important nodes, and flow forecasting
Least Connection [52]	Resource utilization optimization	Centralized and cooperative approach	Server with the least number of active connections is allocated more connections to balance traffic
SDSNLB [53]	Throughput, link load jitter	Centralized	Allocates network traffic to different flow paths for optimal and productive resource use
Dynamic Agent-based Load Balancing [54]	Efficient and adaptive	Centralized	Global visibility of SDN is used to efficiently migrate virtual machines in data centre networks
RLMD [55]	Node efficiency, node attractiveness, path quality, controller load balancing rate	Centralized and cooperative approach	Scheme for effective deployment of controllers and successful load balancing among them
Fuzzy Synthetic Dynamically Select the Evaluation Optimal Path Mechanism [56]	-	Centralized	Network flow is sent to flow paths under open flow switches for SDN controller to install flow handling
Switch Migration Based Decision-Making [57]	Response time, load distribution, and migration cost	Centralized approach	Chooses a master controller to enhance load balancing factor based on low cost
Adaptive Load Balancing Scheme [58]	Throughput and loss rate	Centralized and cooperative approach	New adaptive technique in data centres leveraging SDN for load balancing
Self-Adaptive Load Balancing [59]	Throughput testing, load balancing time, bandwidth utilization, and loss rate	Centralized approach	Ensures effective load balancing and distance between devices are considered
Double Deep Q Network Based VNF Placement Algorithm [60]	Path delay, running time of VNFIs, number of VNFIs, and utilization ratio of VNFIs	Centralized	Customized algorithm designed using gathered information to optimize network performance

7. CHALLENGES FOR FUTURE RESEARCH

Software-defined networking (SDN) has revolutionized the way network administrators manage and control their networks. One of the most significant applications of SDN is load balancing, which is the process of distributing traffic evenly across multiple servers to optimize resource utilization and ensure high availability of services. SDN-based load-balancing techniques have gained popularity in recent years due to their flexibility, scalability, and cost-effectiveness [62].

However, these techniques also face several challenges and issues that need to be addressed for their wider adoption and improved performance. In this article, we will explore these issues, challenges, and future research directions in SDN-based load balancing.

7.1. ISSUES AND CHALLENGES IN SDN-BASED LOAD BALANCING:

Controller overload: One of the most significant challenges in SDN-based load balancing is the controller's processing capacity, which can become a bottleneck when handling a large number of requests. The controller's processing power limits the number of switches and the amount of traffic that can be managed, leading to poor performance and increased latency.

Scalability: Another significant challenge is the scalability of SDN-based load balancing. As the number of switches and servers in the network increases, managing and controlling the network becomes increasingly complex and challenging. This complexity can lead to poor performance, increased latency, and even network outages.

Security: SDN-based load balancing also poses several security challenges, such as DDoS attacks, malware infections, and unauthorized access to the network. These security threats can compromise the availability and performance of the network, leading to significant financial losses.

Traffic engineering: SDN-based load balancing techniques must consider different traffic patterns and routing requirements to optimize network performance. However, designing efficient traffic engineering algorithms that can handle complex network topologies and diverse traffic patterns is a challenging task.

Service-level agreements: SDN-based load balancing must also ensure that service-level agreements (SLAs) are met, such as minimum response time, maximum latency, and minimum throughput. Meeting these SLAs can be difficult, especially when dealing with a large number of requests or unpredictable traffic patterns.

Dynamic load balancing for multiple controllers: In large-scale SDN networks with multiple controllers, there is a need for a dynamic load balancing mechanism that can handle burst traffic and adjust controller loads without compromising traffic balancing [63].

Network management for SD-IoT: With the increasing use of IoT devices, there is a need for suitable technologies to manage the massive amounts of data generated by these devices. SDN-based technologies can help distribute and monitor network traffic flows for load balancing and network delay minimization [64].

Hierarchical controller load balancing: Centralized SDN control using a single controller can result in a single point of failure and network collapse. Hierarchical load balancing using a super controller can help maintain global controller load information, but this requires time-consuming LB decisions [65].

Data plane fault tolerance and low-latency load balancing for SD-WAN: SDN provides opportunities to design custom, adaptive routing schemes for low end-to-end latency and failure recovery mechanisms. However, it remains unclear how to pick better paths in real time in the presence of connection failure or congestion [66].

Security challenges: Availability is a key security problem in SDN, and multiple controllers can cause cascade failures. Protecting the controller and building trust between controllers is a key issue, as is providing LB with improved protection against DDoS and long-awaited queues. The scalability of SDN also needs to be improved to prevent targeted attacks that can cause control plane saturation.

7.2 FUTURE RESEARCH DIRECTIONS IN SDN-BASED LOAD BALANCING:

Machine learning: Machine learning techniques can help optimize SDN-based load balancing by predicting traffic patterns and resource utilization, identifying network anomalies, and detecting security threats. These techniques can also help optimize traffic engineering algorithms and improve SLA compliance.

Blockchain: Blockchain technology can provide a secure and decentralized platform for SDN-based load balancing, allowing for greater transparency and accountability in network management. Blockchain can also enable more efficient and secure management of network resources and services.

Fog computing: Fog computing is a distributed computing paradigm that extends cloud computing to the edge of the network, where devices and sensors are located. Fog computing can help improve SDN-based load balancing by enabling faster response times, reducing latency, and improving resource utilization.

Network function virtualization: Network function virtualization (NFV) is the process of virtualizing network functions such as firewalls, load balancers, and routers. NFV can help optimize SDN-based load balancing by providing more flexible and scalable network services and reducing the complexity of network management.

Hybrid solutions: Hybrid SDN-based load-balancing solutions that combine traditional load-balancing techniques with SDN-based approaches can provide more efficient and reliable network management. Hybrid solutions can leverage the benefits of both approaches and provide better performance and scalability than either approach alone.

In conclusion, SDN-based load balancing is a promising technology that offers many benefits, including improved network performance, scalability, and flexibility. However, it also faces several challenges and issues that need to be addressed for its wider adoption and improved performance.

Future research directions in SDN-based load balancing include machine learning, blockchain, fog computing, network function virtualization, and hybrid solutions, among others.

8. PROPOSED ROADMAP

Introducing a new SDN-based load balancing technique, "Adaptive Multi-Objective Load Balancing (AMOLB)." This approach aims to optimize network performance by dynamically balancing multiple objectives, such as latency, throughput, energy efficiency, and resource utilization.

The AMOLB technique leverages the capabilities of SDN to monitor and control the network in real-time while using machine learning to adapt to changing network conditions.

Key components of the AMOLB technique:

- Centralized Controller: The centralized controller is responsible for managing the entire network, collecting real-time network statistics, and making load-balancing decisions based on the multi-objective optimization model.
- Multi-Objective Optimization Model: This model considers multiple objectives and assigns weights to each of them based on the network administrator's preferences or dynamic network requirements. The optimization model generates an optimal set of load-balancing policies that balance the objectives according to their assigned weights.
- Machine Learning Module: The machine learning module continuously analyses network traffic patterns and adjusts the weights assigned to different objectives based on real-time network conditions. It can also predict future traffic patterns and preemptively adjust the load balancing policies to maintain optimal network performance.
- Real-time Network Monitoring: The AMOLB technique relies on real-time network monitoring to collect network statistics, such as latency, throughput, and resource utilization. This data is used to update the multi-objective optimization model and make informed load-balancing decisions.
- Programmable Data Plane: The programmable data plane allows the AMOLB technique to implement dynamic load-balancing policies at the forwarding level. This ensures that traffic is optimally distributed across the network, considering the multiple objectives defined in the optimization model.

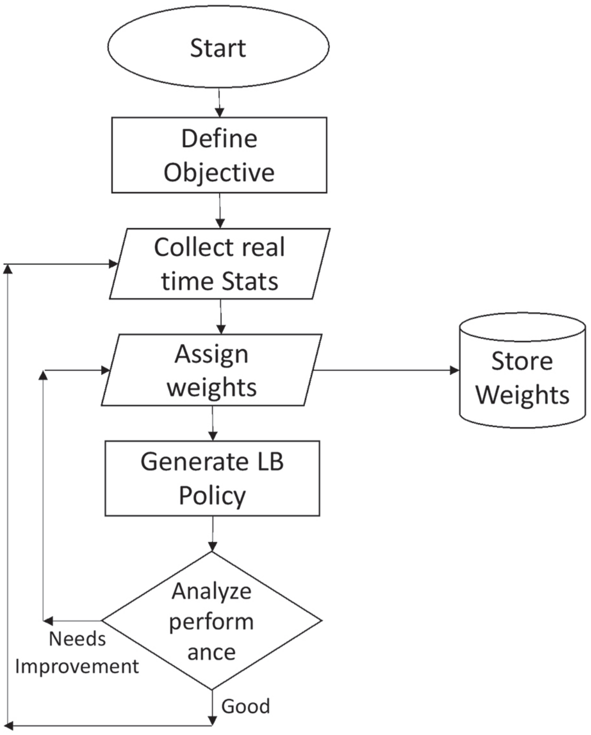
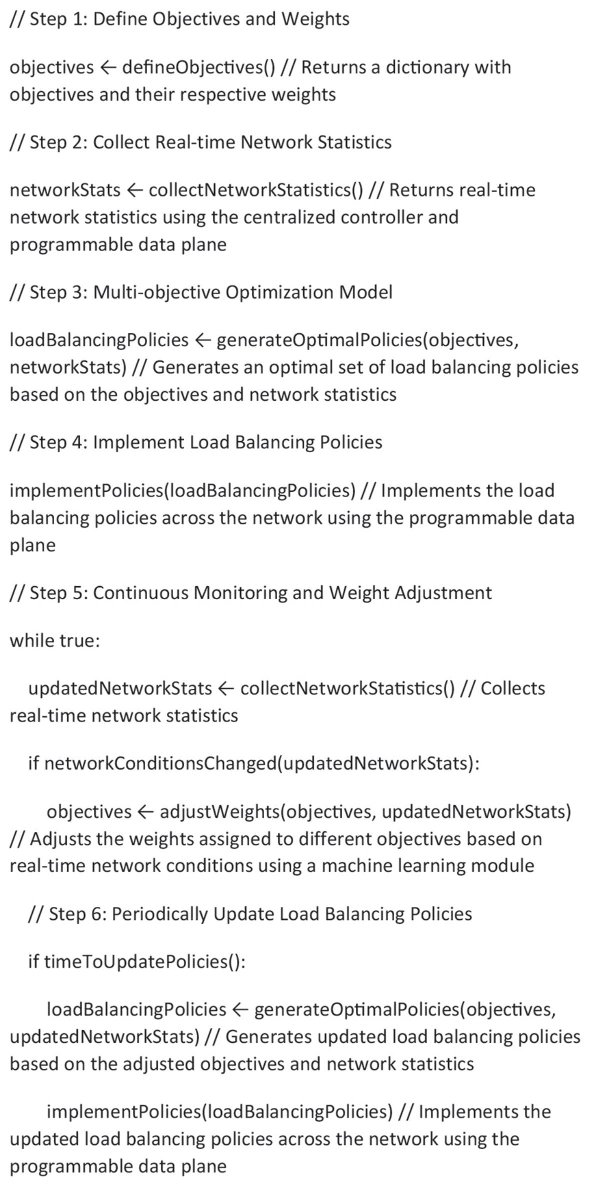


Fig. 9. Adaptive Multi-Object Load Balancing data flow

Fig. 9 above depicts the flow chart involving steps to implement the proposed AMOLB technique:

1. Define the objectives and their respective weights based on network requirements or administrator preferences.
2. Collect real-time network statistics using the centralized controller and programmable data plane.
3. Use the multi-objective optimization model to generate an optimal set of load balancing policies that consider the defined objectives and their weights.
4. Implement the load balancing policies across the network using the programmable data plane.
5. Continuously monitor the network and adjust the weights assigned to different objectives based on real-time network conditions using the machine learning module.
6. Periodically update the load balancing policies to maintain optimal network performance, considering the dynamic nature of network conditions and traffic patterns.

The pseudocode for the same is presented below in Fig. 9.:



```
// Step 1: Define Objectives and Weights

objectives ← defineObjectives() // Returns a dictionary with
objectives and their respective weights

// Step 2: Collect Real-time Network Statistics

networkStats ← collectNetworkStatistics() // Returns real-time
network statistics using the centralized controller and
programmable data plane

// Step 3: Multi-objective Optimization Model

loadBalancingPolicies ← generateOptimalPolicies(objectives,
networkStats) // Generates an optimal set of load balancing policies
based on the objectives and network statistics

// Step 4: Implement Load Balancing Policies

implementPolicies(loadBalancingPolicies) // Implements the load
balancing policies across the network using the programmable data
plane

// Step 5: Continuous Monitoring and Weight Adjustment

while true:

  updatedNetworkStats ← collectNetworkStatistics() // Collects
real-time network statistics

  if networkConditionsChanged(updatedNetworkStats):

    objectives ← adjustWeights(objectives, updatedNetworkStats)
// Adjusts the weights assigned to different objectives based on
real-time network conditions using a machine learning module

  // Step 6: Periodically Update Load Balancing Policies

  if timeToUpdatePolicies():

    loadBalancingPolicies ← generateOptimalPolicies(objectives,
updatedNetworkStats) // Generates updated load balancing policies
based on the adjusted objectives and network statistics

    implementPolicies(loadBalancingPolicies) // Implements the
updated load balancing policies across the network using the
programmable data plane
```

Fig 10. Pseudo-Code implementation for AMOLB

This technique offers a novel approach to SDN-based load balancing by considering multiple objectives and dynamically adapting to changing network conditions. By leveraging the capabilities of SDN and machine learning, the AMOLB technique can provide improved network performance, increased flexibility, and more efficient resource utilization.

9. CONCLUSION

Numerous load-balancing techniques have been proposed and implemented in SDN networks. These techniques can be classified into three main categories: proactive, reactive, and hybrid.

Despite the progress made in load-balancing research for SDN, several challenges and open issues still need to be addressed. For example, load balancing in a multi-controller environment remains a challenge, and new techniques like AMOLB are needed to address this issue. The security of SDN networks is also a concern, and load-balancing techniques need to be developed to prevent cyber-attacks that can compromise the network's availability and performance. Moreover, the Internet of Things (IoT) is expected to generate massive amounts of data, and load-balancing techniques must be developed to handle this data efficiently.

In conclusion, load balancing is an essential aspect of network management in SDN. It involves the intelligent allocation of network resources to improve network performance and avoid congestion. The proposed AMOLB technique offers a novel and efficient approach to SDN-based load balancing, considering multiple objectives and dynamically adapting to changing network conditions. This approach to SDN-based load balancing considers multiple objectives and dynamically adapts to changing network conditions. By leveraging the capabilities of SDN and machine learning, the AMOLB technique can provide improved network performance, increased flexibility, and more efficient resource utilization.

In conclusion, this paper has made a valuable contribution by thoroughly examining the challenges inherent in SDN-based load balancing. Through a comprehensive analysis of issues such as scalability, adaptability, and complexity, this research serves as a guiding resource for future investigations aimed at tackling these obstacles and optimizing the performance of SDN load-balancing solutions.

Moreover, this study has identified promising research avenues that hold the potential to augment the effectiveness and efficiency of load balancing within SDN architectures. By shedding light on these areas, the paper offers a roadmap for researchers to delve deeper and devise innovative solutions that can further enhance load-balancing techniques in SDN environments.

In summary, this research has provided essential insights into the existing challenges and future possibilities of SDN-based load balancing. By addressing these challenges head-on and exploring new research directions, we can collectively foster advancements in SDN technology and contribute to the continuous improvement of network performance and reliability. The findings of this paper lay a solid foundation for the ongoing pursuit of optimized load-balancing solutions in SDN domains, thus paving the way for more robust and efficient networking infrastructures in the future.

Researchers continue to explore new load-balancing techniques, like AMOLB, to address the challenges and open issues associated with SDN networks.

10.REFERENCES:

[1] A. Abdelaziz, A. Fong, A. Gani, U. Garba, S. Khan, A. Akhunzada, H. Talebian, K. K. R. Choo, "Distributed controller clustering in software-defined networks", PLoS One, Vol. e0174715, 2017, p. 12.

[2] N. Rowshanrad, "A survey on SDN, the future of networking", Journal of Advanced Computer Science & Technology, Vol. 3, No. 2, 2014, p. 232.

[3] P. Martinez-Julia, A. Skarmeta, "Empowering the Internet of things with software defined networking", White Paper, IoT6-FP7 European Research Project (accessed: 2020)

[4] A. A. Neghabi, N. Jafari Navimipour, M. Hosseinzadeh, A. Rezaee, "Load Balancing Mechanisms in the Software Defined Networks: A Systematic and Comprehensive Review of the Literature", IEEE Access, Vol. 6, pp. 14159-14178, 2018.

[5] W.-H. Liao, S.-C. Kuai, C.-H. Lu, "Dynamic Load-Balancing Mechanism for Software-Defined Networking", Proceedings of the International Conference on Networking and Network Applications, Hakodate, Japan, 2016, pp. 336-341.

[6] A. Neghabi, N. Navimipour, M. Hosseinzadeh, A. Rezaee, "Load balancing mechanisms in the software-defined networks: A systematic and comprehensive review of the literature", IEEE Access, Vol. 6, 2018, pp. 14159-14178.

[7] N. Rowshanrad et al. "A survey on SDN, the future of networking", Journal of Advanced Computer Science & Technology, Vol. 3, No. 2, 2014, p .232.

[8] M. Saxena, M. Sabharwal, P. Bajaj, "A novel method to enhance the reliability of transmission over secured sd wan overlay", Journal of Theoretical and Applied Information Technology, Vol. 101, No. 14, 2023.

[9] D. Kreutz et al. "Software-defined networking: a comprehensive survey", Proceedings of the IEEE, Vol. 103, No. 1, 2015, pp. 14-76.

[10] M. Jammal, T. Singh, A. Shami, R. Asal, Y. Li, "Software-defined networking: state of the art and research challenges", Computer Networks, Vol. 72, 2014, pp. 74-98.

[11] S. Scott-Hayward, S. Natarajan, S. Sezer, "A Survey of Security in Software Defined Networks", IEEE Communications Surveys & Tutorials, Vol. 18, No. 1, 2016, pp. 623-654.

[12] H. Farhady, H. Y. Lee, A. Nakao, "Software-defined networking: a survey", Computer Networks, Vol. 81, 2015, pp. 79- 95.

[13] A. Mirchev, "Survey of concepts for QoS improvement via SDN", Proceedings of the Seminars Future Internet and Innovative Internet Technologies and Mobile Communications, September 2015, pp. 33-40.

[14] D. A. Karakus, "Quality of Service (QoS) in software defined networking (SDN) a survey", Journal of Network and Computer Applications, Vol. 80, 2017, pp. 200-218.

[15] D. Li, S. Wang, K. Zhu, S. Xia, "A survey of network update in SDN", Frontiers of Computer Science, Vol. 11, No. 1, 2017, pp. 4-12.

[16] A. Mushtaq, R. Mittal, J. McCauley, M. Alizadeh, S. Ratnasamy, S. Shenker, "Datacenter congestion control: Identifying what is essential and making it practical", ACM SIGCOMM Computer Communication Review, Vol. 49, No. 3, 2019, 32-38.

[17] O. Michel, E. Keller, "SDN in wide-area networks: a survey", Proceedings of the 4th International Conference on Software Defined Systems, Valencia, Spain, 8-11 May 2017, pp. 37-42.

[18] T. Hu et al. "Controller load balancing mechanism based on distributed policy in SDN", ACTA ELECTONICA SINICA, Vol. 46, No. 10, 2018, p. 2316.

[19] D. T. T. Hien, T. D. Ngo, D. D. Le, H. Sekiya, V. H. Pham, K. Nguyen, "A software defined networking approach for guaranteeing delay in Wi-Fi networks", Proceedings of the 10th International Symposium on Information and Communication Technology, December 2010, pp. 191-196.

[20] J. Cui, Q. Lu, H. Zhong, M. Tian, L. Liu, "A load-balancing mechanism for distributed SDN control plane using response time", IEEE Transactions on Network and Service Management, Vol. 15, No. 4, 2018, pp. 1197-1206.

[21] L. Li, Q. Xu, "Load balancing research in SDN: A survey", Proceedings of the IEEE International Conference on Electronics Information and Emergency Communication, Macau, China, 21-23 July 2017, pp. 403-408.

[22] M. Alizadeh et al. "CONGA: Distributed congestion-aware load balancing for datacenters", Proceedings of the ACM conference on SIGCOMM, August 2014, pp. 503-514.

[23] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, A. Vahdat, “Hedera: Dynamic flow scheduling for data centre networks”, Proceedings of the 7th USENIX Conference on Networked Systems Design and Implementation, 2010.

[24] A. S. Thorat, S. K. Sonkar, “A Review on Energy Efficient Load Balancing Techniques for Secure and Reliable Cloud Eco-system”, International Journal of Advance Research and Innovative Ideas in Education, Vol. 2, No. 1, 2016.

[25] M. Yu, J. Rexford, M. J. Freedman, J. Wang, “Scalable flow-based networking with DIFANE”, Proceedings of the ACM SIGCOMM 2010 Conference on SIGCOMM, 2010, pp. 351-362.

[26] B. Chang, A. Akella, L. D'Antoni, K. Subramanian, “Learned Load Balancing”, Proceedings of the 24th International Conference on Distributed Computing and Networking, January 2023, pp. 177-187.

[27] N. Katta, M. Alizadeh, J. Rexford, D. Walker, “Infinite cache flow in software-defined networks”, Proceedings of the third workshop on Hot topics in Software Defined Networking, 2016.

[28] R. Pries, M. Jarschel, D. Schlosser, M. Klopf, & P. Tran-Gia, “Power consumption analysis of data center architectures”, Proceeding of Green Communications and Networking: First International Conference, Revised Selected Papers, October 2011, pp. 114-124.

[29] M. Al-Fares, M. Loukissas, A. Vahdat, “A scalable, commodity data center network architecture”, Proceedings of the ACM SIGCOMM Conference on Data Communication, 2018, pp. 63-74.

[30] C. Lim, “Enhancing robustness of per-packet load-balancing for fat-tree”, Applied Sciences, Vol. 11, No. 6, 2021, p. 2664.

[31] J. Tao, S. Liu, C. Liu, “A Traffic Scheduling Scheme for Load Balancing in SDN-Based Space-Air-Ground Integrated Networks”, Proceedings of the IEEE 23rd International Conference on High Performance Switching and Routing, Taicang, Jiangsu, China, 6-8 June 2022, pp. 95-100.

[32] J. Chen et al. “ALBRL: Automatic Load-Balancing Architecture Based on Reinforcement Learning in Software-Defined Networking”, Wireless Communications and Mobile Computing, Vol. 2022, 2022, pp. 1-17.

[33] J. Pei, P. Hong, M. Pan, J. Liu, J. Zhou, "Optimal VNF placement via deep reinforcement learning in SDN/NFV-enabled networks”, IEEE Journal on Selected Areas in Communications, Vol. 38, 2019, p. 263-278.

[34] P. B. M. C. Saxena, "Evolution of Wide Area network from Circuit Switched to Digital Software defined Network”, Proceedings of the International Conference on Technological Advancements and Innovations, Tashkent, Uzbekistan, 10-12 November 2021.

[35] Z. Shang, W. Chen, Q. Ma, B. Wu, “Design and implementation of server cluster dynamic load balancing based on OpenFlow”, Proceedings of the International Joint Conference on Awareness Science and Technology & Ubi-Media, Aizu-Wakamatsu, Japan, 2-4 November 2013, pp. 691-697.

[36] H. Babbar, S. Parthiban, G. Radhakrishnan, S. Rani, “A genetic load balancing algorithm to improve the QoS metrics for software defined networking for multimedia applications”, Multimedia Tools and Applications, Vol. 81, No. 7, 2022, pp. 9111-9129.

[37] T. Koponen et al. “Onix: A distributed control platform for large-scale production networks”, Proceedings of the 9th USENIX conference on Operating systems design and implementation, 2010.

[38] C. Trois, M. D. Del Fabro, L. C. de Bona, M. Martinello, “A survey on SDN programming languages: Toward a taxonomy”, IEEE Communications Surveys & Tutorials, Vol. 18, No. 4, pp. 2687-2712.

[39] K. E. Jungmin, "SDN in wide-area networks: a survey”, Proceedings of the 4th International Conference on Software Defined Systems, 2017, pp. 37-42.

[40] K. Benzekki, A. El Fergougui, A. E. Elalaoui, ”Software-defined networking (SDN): a survey”, Security and communication networks, Vol. 9, No. 18, 2016, pp. 5803-5833.

[41] J. Son, R. Buyya, “A taxonomy of software-defined networking (SDN)-enabled cloud computing”, ACM Computing Surveys, Vol. 51, No. 3, 2018, pp. 1-36.

[42] L. Zhuo, C. L. Wang, F. C. Lau, "Load balancing in distributed web server systems with partial document replication", Proceedings of the International Conference on Parallel Processing, Vancouver, BC, Canada, 21 August 2022, pp. 305-312.

[43] S. Abuthahir, S. C. B. Jaganathan, R. Saha, "An Efficient Enhanced Dynamic Load Balancing Weighted Round Robin Algorithm for Virtual Machine in Cloud Computing", Journal of Algebraic Statistics, Vol. 13, No. 2, 2022, pp. 2121-2128.

[44] C. Yu, Z. Zhao, Y. Zhou, H. Zhang, "Intelligent Optimizing Scheme for Load Balancing in Software Defined Networks", Proceedings of the IEEE 85th Vehicular Technology Conference, Sydney, NSW, Australia, 4-7 June 2017, pp. 1-5.

[45] S. Kaur, J. Singh, N. S. Ghumman, "Network programmability using POX controller", Proceeding of the International Conference on Communication, Computing & Systems, Vol. 138, August 2014, p. 70

[46] C. Fancy, & M. Pushpalatha, "Performance evaluation of SDN controllers POX and floodlight in mininet emulation environment", Proceedings of the International Conference on Intelligent Sustainable Systems, Palladam, India, 7-8 December 2017, pp. 695-699.

[47] F. Benamrane, R. Benaini, "An East-West interface for distributed SDN control plane: Implementation and evaluation", Computers & Electrical Engineering, Vol. 57, 2017, pp. 162-175.

[48] R. Schmidt, C. Y. Chang, N. Nikaein, "FlexVRAN: A Flexible Controller for Virtualized RAN Over Heterogeneous Deployments", Proceedings of the IEEE International Conference on Communications, Shanghai, China, 20-24 May 2019, pp. 1-7.

[49] N. L. Van Adrichem, C. Doerr, F. A. Kuipers, "Opennetmon: Network monitoring in OpenFlow software-defined networks", Proceedings of the IEEE Network Operations and Management Symposium, May 2014, pp. 1-8.

[50] S. Mohmmad, M. A. Shaik, K. Mahender, R. Kanakam, B. P. Yadav, "Average Response Time (ART): Real-Time Traffic Management in VFC Enabled Smart Cities", IOP Conference Series: Materials Science and Engineering, Vol. 981, No. 2, 2020, p. 022054.

[51] R. Jamal, L. C. Fourati, "Implementing shortest path routing mechanism using OpenFlow POX controller", Proceedings of the International Symposium on Networks, Computers and Communications, Hammamet, Tunisia, 17-19 June 2014, pp. 1-6.

[52] A. Jalili, M. Keshtgari, R. Akbari, "A new framework for reliable control placement in software-defined networks based on multi-criteria clustering approach", Soft Computing, Vol. 24, 2020, p. 2897- 2916.

[53] A. Jalili, M. Keshtgari, R. Akbari, R. Javidan, "Multi criteria analysis of controller placement problem in software defined networks", Computer Communications, Vol. 133, 2019, p. 115-128.

[54] S. Jamali, A. Badirzadeh, M. S. Siapoush, "On the use of the genetic programming for balanced load distribution in software defined networks", Digital Communications and Networks, Vol. 5, 2019, pp. 288-296.

[55] A. Javadpour, "Providing a way to create balance between reliability and delays in SDN networks by using the appropriate placement of controllers", Wireless Personal Communications, Vol. 110, 2020, pp. 1057-1071.

[56] X. Jia, Y. Jiang, Z. Guo, Z. Wu, "Reducing and balancing flow table entries in software-defined networks", Proceedings of the IEEE 41st Conference on Local Computer Networks, Dubai, United Arab Emirates, 7-10 November 2016, pp. 575-578.

[57] X. Jia, Y. Jiang, Z. Guo, J. Sun, "A low overhead flow-holding algorithm in software-defined networks", Computer Networks, Vol. 123, 2017, pp. 170-180.

[58] E. R. Jimson, K. Nisar, M. H. A. Hijazi, "The state of the art of software defined networking (SDN) issues in current network architecture and a solution for network management using the SDN", International Journal of Technology Diffusion, Vol. 10, 2019, pp. 33-48.

[59] S. Askar, "Adaptive load balancing scheme for data centre networks using software defined network", Science Journal of University of Zakho, Vol. 4, 2016, pp. 275-286.

[60] M. Priyadarshini, J. Mukherjee, P. Bera, S. Kumar, A. Jakaria and M. Rahman, "An adaptive load balancing scheme for software-defined network controllers", Computer Networks, Vol. 164, No. 106918, 2019.

[61] L. Wang, W. Mao, J. Zhao, Y. Xu, “A double deep Q-learning approach to online fault-tolerant SFC placement”, IEEE Transactions on Network and Service Management, Vol. 18, No. 1, 2021, pp. 118-132.

[62] N. Hai, D. Kim, "Efficient load balancing for multi-controller in SDN-based mission-critical networks", Proceedings of the IEEE 14th International Conference on Industrial Informatics, Poitiers, France, 19-21 July 2016, p. 420-425.

[63] M. Mathur, M. Madan, M. C. Saxena, "A Proposed Architecture for Placement of Cloud Data Centre in Software Defined Network Environment", International Journal of Engineering and Advanced Technology, Vol. 1, No. 2, 2012, pp. 104-116.

[64] J. Zhao, M. Tong, H. Qu, J. Zhao, "An Intelligent Congestion Control Method in Software Defined Networks”, Proceedings of the IEEE 11th International Conference on Communication Software and Network, Chongqing, China, 12-14 June 2019, pp. 51-56.

[65] C. Yu, J. Lan, Z. Guo, Y. Hu, “Optimizing the routing in software-defined networks with deep reinforcement learning”, IEEE Access, Vol. 6, 2018, pp. 64533-64539.

[66] P. Sun, Y. Hu, J. Lan, L. Tian, M. T. Chen, "Time-relevant deep reinforcement learning for routing optimization”, Future Generation Computer Systems, Vol. 99, 2019, pp. 401-409.

A Novel Nodesets-Based Frequent Itemset Mining Algorithm for Big Data using MapReduce

Original Scientific Paper

Borra Sivaiah
Research Scholar,
Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Kakinada, Andra Pradesh, India,
CMR College of Engineering &Technology, Hyderabad
sivabetld@gmail.com

Ramisetty Rajeswara Rao
Professor of CSE,
Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Gurajada, Andra Pradesh, India
raob4u@jntukucev.ac.in

Abstract *– Due to the rapid growth of data from different sources in organizations, the traditional tools and techniques that cannot handle such huge data are known as big data which is in a scalable fashion. Similarly, many existing frequent itemset mining algorithms have good performance but scalability problems as they cannot exploit parallel processing power available locally or in cloud infrastructure. Since big data and cloud ecosystem overcomes the barriers or limitations in computing resources, it is a natural choice to use distributed programming paradigms such as Map Reduce. In this paper, we propose a novel algorithm known as A Nodesets-based Fast and Scalable Frequent Itemset Mining (FSFIM) to extract frequent itemsets from Big Data. Here, Pre-Order Coding (POC) tree is used to represent data and improve speed in processing. Nodeset is the underlying data structure that is efficient in discovering frequent itemsets. FSFIM is found to be faster and more scalable in mining frequent itemsets. When compared with its predecessors such as Node-lists and N-lists, the Nodesets save half of the memory as they need only either pre-order or post-order coding. Cloudera's Distribution of Hadoop (CDH), a MapReduce framework, is used for empirical study. A prototype application is built to evaluate the performance of the FSFIM. Experimental results revealed that FSFIM outperforms existing algorithms such as Mahout PFP, Mlib PFP, and Big FIM. FSFIM is more scalable and found to be an ideal candidate for real-time applications that mine frequent itemsets from Big Data.*

Keywords*: Big Data, Frequent Itemset Mining (FIM), MapReduce Programming Paradigm (MRPP), Fast and Scalable Frequent Item set Mining (FSFIM)*

1. INTRODUCTION

Frequent Itemset Mining (FIM) is a phenomenon in data mining used to extract frequently occurring items that exhibit latent relationships in the data. FIM leads to the generation of association rules that provide Business Intelligence (BI) when interpreted by domain experts. Association rule mining is the process of finding patterns, associations, and correlations among sets of items in a database. The Association Rules generated have an antecedent and a consequent. An Association Rule is a pattern of the form $X \wedge Y \Longrightarrow Z$ [support, confidence], where X, Y, and Z are items in the dataset. The left-hand side of the rule $X \wedge Y$ is called the antecedent of the rule and the right-hand side Z is called the consequent of the rule. Within the dataset, confidence and support are two measures to determine the certainty or usefulness of each rule. Support is the probability that a set of items in the dataset contains both the antecedent and consequent of the rule i.e P ($X \cup Y \cup Z$). Confidence is the probability that a set of items containing the antecedent also contains the consequent. i.e., P ($Z/(X \cup Y)$

Many FIM algorithms that came into existence are classified into Apriori-based and pattern-growth-based algorithms. These algorithms cannot exploit the parallel processing power of cloud data centers.

Therefore, they suffer from the capability of dealing with big data. Big Data is data that has characteristics such as volume, variety, and velocity as shown in Fig.1. Big Data is voluminous (often measured in petabyte-scale), and has a variety of data such as structured, unstructured, and semi-structured besides having continuous growth or streaming data. The importance of Big Data processing signifies that when big data (complete data) is not considered, it results in biased conclusions. The amount of data being generated by different sources of Big Data such as social media, World Wide Web (WWW), and Internet of Things (IoT) use cases is unprecedented and essentially needs a specialized modus operandi to mine it efficiently to arrive at Business Intelligence (BI). There are two types of approaches used in frequent item sets mining of Big Data: Apriori-based and FP-Growth based. The Apriori-based algorithms consume more memory and time in the generation of frequent item sets from Big Data. The FP-Growth-based algorithms were developed for faster generation of the frequent itemsets instead of the Aprori-based algorithm. Therefore, FP-growth-based algorithms are faster than the Apriori-based approaches. However, FP-growth algorithms also suffer from storing conditional FP-trees in memory and then mining from them may require more time and memory. The drawbacks of the existing works:

1. Some of the existing frequent item sets mining algorithms consumed more memory and more mining time.
2. Some of the existing algorithms don't use distributed programming paradigms to solve the scalability problem.

To address the existing drawbacks, the proposed research is introduced. The main purpose of the research in this article is to develop an efficient MapReduce-based algorithm for faster and scalable discovery of frequent itemsets from big data. The proposed method needs distributed programming frameworks like Hadoop for mining big data as explored in Section 1.2.

1.1. MOTIVATION

Many researchers developed frequent pattern algorithms for handling Big Data. The algorithms are divided into two categories: Apriori-based and tree-based. Tree-based algorithms are faster than Apriori-based algorithms. The traditional tools and techniques cannot handle big data, due to their limited capabilities. Frequent itemset mining algorithms are faster but cannot use local or cloud-based parallel processing power. Researchers used parallel and distributed Hadoop MapReduce systems to quickly generate frequent item sets. Large data makes it difficult to extract frequent item sets from transactional databases, if extracted faster, then it could be helpful for better decision-making. Using the best data structures can reduce half of the memory and execution time of frequent item generation from big data. When a case study like healthcare is considered, there is ever-growing data size leading to big data. Unless there is faster convergence in frequent itemset mining, it takes more time for frequent itemset mining. In many applications, there is a need for quick convergence. It is the motivation behind the research carried out and presented in this paper. Our work has focused on building a new algorithm for mining frequent item sets based on the best data structure known as Nodeset and tree structure known as POC tree. Towards this end, we proposed a framework for FIM that is significantly faster than existing methods. Our contributions to this paper are as follows.

1. A novel algorithm known as Fast and Scalable Frequent item set mining (FSFIM) is proposed to extract frequent item sets from big data.
2. A Pre-Order Coding (POC) tree is used to represent data and improve speed in processing leading to improved performance.
3. A prototype is built to evaluate the FSFIM and compare it with the state-of-the-art FIM techniques.

The remainder of the paper is structured as follows. Section 2 reviews the literature on FIM techniques, especially parallelized ones. Section 3 presents the proposed algorithm and the underlying mechanisms and data structures. Section 4 presents experimental results and discussion. Section 5 concludes the work in the paper and gives suggestions for improving the research further in the future.

1.2. CLOUDERA'S DISTRIBUTION OF HADOOP

Apache Hadoop framework is widely used in the real world for processing big data. Cloudera's Distribution of Hadoop (CDH) is based on the Apache Hadoop framework. It supports the MapReduce paradigm where the number of worker nodes in the distributed environment acts as a mapper and reducer.

Hadoop Distributed File System (HDFS) plays a crucial role in the framework as it stores inputs and outputs. HDFS can access data from various servers computed distributed across the globe. The framework involves a job tracker and task tracker to keep track of a given job and underlying tasks respectively. The map phase is carried out by thousands of worker nodes and the results are given to reducers (worker nodes). On the other hand, the reducers act on the data to produce the final output.

Fig. 2 shows the MapReduce phenomenon which is essentially meant for dealing with large volumes of data. It supports parallel processing to process data faster. Input data is divided into many pieces and assigned to worker nodes that complete a given job and return the results.

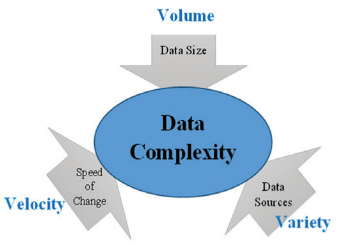


Fig. 1. Illustrates characteristics of big data

2. RELATED WORK

Literature is found rich in the research on frequent itemset mining algorithms. Many FIM algorithms that target big data came into existence which are briefly presented in this section.

Qiu *et al.* [1] proposed a parallel FIM algorithm known as Yet Another Frequent Itemset Mining (YAFIM) which is based on Apriori. They defined the YAFIM algorithm to run using a distributed programming framework known as Spark.

Yasir *et al.* [2] proposed an FIM algorithm to handle sparse big data. They named it TRimmed Transaction LattICE (TRICE). It generates trimmed subsets iteratively to leverage performance in terms of memory consumption and execution time. They intended to improve it to have other variants of frequent itemsets such as maximal frequent itemsets besides working on streaming data.

Gole and Tidke [3] proposed a MapReduce-based method for FIM on big data. It is named ClustBigFIM. It is derived from the BigFIM algorithm, for better speed and scalability.

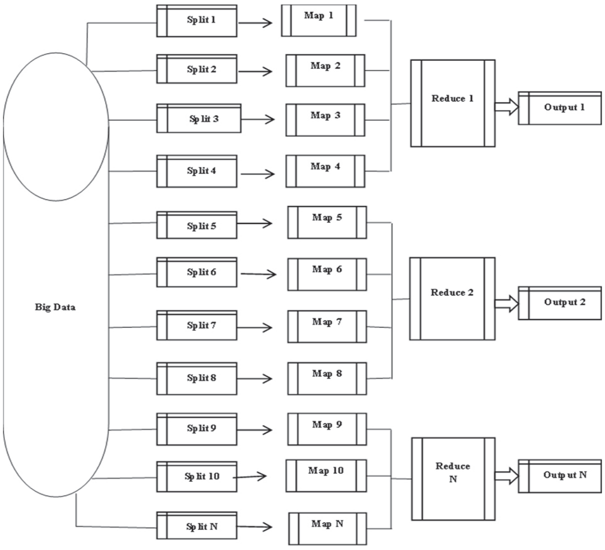


Fig. 2. Map Reduce Frame Work in Hadoop

Djenouri *et al.* [4] proposed two FIM algorithms namely the Enhanced Approach for Single Scan approach for Frequent Itemset Mining (EA-SSFIM) and MapReduce Single Scan approach for Frequent Itemset Mining (MR-SSFIM) for big data. They are designed to deal with sparse big data and big data respectively for improvements in terms of reducing execution time and improving efficiency in processing big data.

Apiletti *et al.* [5] review different FIM algorithms that are developed for big data. Fernandez-Basso *et al.* [6] proposed a distributed method for FIM which is meant for extracting frequent itemsets from streaming data. Sethi and Ramesh [7] proposed an FIM algorithm known as Hybrid Frequent Itemset Mining (HFIM) that works in two phases. In the first phase, it extracts frequent itemsets and in the second phase, it obtains frequent itemsets of k-cardinality where k is greater than or equal to 2. It could improve speedup and execution time. Chon *et al.* [8] proposed a Graphics Processing Unit (GPU) based FIM known as GMiner which is much faster as it could exploit the power of GPU.

Liang and Wu [9] proposed a distributed FIM algorithm known as Sequence-Growth. It makes use of a lexicographical sequence tree that follows the idea of lexicographical order. It also has a pruning strategy known as breadth-wide support-based which makes it scalable and efficient. In the future, they intend to make it an incremental FIM algorithm.

Joy and Sherly [10] proposed a parallel FIM algorithm known as Faster-IAPI that is executed using the Spark RDD environment. It is employed to have symptom correlations in patients' data in the healthcare domain for disease prediction. Djenouri *et al.* [11] proposed three versions of High-Performance Computing (HPC) that make use of a single database scan. They are known as Single Scan on GPU (GSS), Single Scan on Cluster (CSS), and Single Scan on Cluster and GPU (CGSS). Out of which CGSS was found to have better performance.

A distributed FIM algorithm known as BIGMiner has been proposed in [12] which is scalable and causes less network communication overhead. This algorithm exploits the GPU and MapReduce programming model to have significantly improved performance.

Raj *et al.* [13]. have proposed an EAFIM method for an efficient apriori-based frequent itemset mining algorithm on Spark for big transactional data. Spark is gaining attention in big data processing due to its in-memory processing capabilities. EAFIM uses parallel and distributed computing environments and introduces two novel methods to improve efficiency. Unlike apriori, candidate generation and count of support values occur simultaneously during input dataset scanning. The updated input dataset is calculated by removing useless items and transactions, reducing the size of the input dataset for higher iterations.

Moens *et al.* [14] investigate the usage of the MapReduce paradigm to execute different FIM algorithms. They studied two such algorithms such as BigFIM and Dist-Eclat thoroughly. They found that MapReduce models outperform their predecessors. Xun *et al.* [15] proposed parallel FIM based on Hadoop clusters. The algorithm is characterized by its data partitioning approach that paves the way for a locality-based approach to enhance the performance of the algorithm.

Asbern and Asha [16] explored different algorithms for FIM that operate on big data using the MapReduce paradigm. Kumar and Mohbey [17] investigated different parallel FIM algorithms that are executed in distributed environments. Different issues they identified in such algorithms include scalability, privacy, complex data types, load balancing, and gene regulation patterns.

Zitouni *et al.* [18] proposed a parallel FIM known as CloPN. It follows a prime number-based approach for FIM from big data. It is supposed to mine closed frequent itemsets (CFI). Leung *et al.* [19] on the other hand proposed an alternative data mining approach for big data. It is known as scalable vertical mining using Spark. Galetsi *et al.* [20] studied big data analytics associated with the healthcare domain. They investigated on different machine learning techniques used for data analytics including FIM algorithms.

Fernandez-Basso *et al.* [21] defined a fuzzy mining approach to have FIM on big data. They implemented a method known as the automatic fuzzification method for this purpose. It takes weather forecast data and generates association rules with energy efficiency using the Spark environment.

Fumarola and Malerba [22] proposed a method known as approximate FIM that uses parallel processing using the MapReduce paradigm. Djenouri *et al.* [23] proposed a parallel framework for FIM using a metaheuristic approach. It is known as Cluster for FIM (CFIM). At nodes in the cluster, the algorithm partitions data. The framework is integrated with different metaheuristics such as Genetic Algorithm (GA), BSO (Bees Swarm Optimization), and Particle Swarm Optimization (PSO) for better performance.

Aggarwal *et al.* [24] investigated air quality data with location and time awareness for performing FIM. First, it understands spatiotemporal dependencies in the data and then employs the FIM process to generate frequent itemsets. Luna *et al.* [25] proposed different parallel versions of Apriori to work on big data using the MapReduce paradigm. Their Hadoop-based implementation showed better performance than the existing methods. From the literature, it is understood that parallel approaches used for FIM can perform better than traditional FIM methods. However, the performance achieved due to distributed environments is insufficient as the underlying method for FIM is expected to have a better approach. Towards this end, we proposed a framework for FSFIM that is significantly faster than existing methods.

S. Nalousi *et al.* [26] introduced a novel efficient approach called weighted frequent itemset mining using weighted subtrees (WST-WFIM) to identify the average weight of frequent rules. The average weight of found rules is calculated using special trees and some novel data structures on the frequent pattern growth (FP-Growth) method. It works with the data set that each item in each transaction has a certain weight and saves them in the dedicated tree.

Fayuan Li *et al.* [27], Based on the calculation of item set fuzziness, this approach incorporates the unpredictability of potential world models to tackle the problem of mining fuzzy frequent item sets based on probability threshold. Fuzzy theory and uncertainty are based on linguistic information and have been expanded to cope with partial truth concepts. A dynamic programming-based approach is used to compute the frequent fuzzy probability.

TR-FC-GCM (Transaction Reduction - Frequency Count - Generate Combination Method) created by Ajay Sharma et al [28] discovers all significant frequent patterns by creating all potential combinations of an item with a single database search and performs better for null and full datasets. B Sivaiah et al [29], Reviewed Incremental mining, which aims to extract patterns from dynamic databases that have applications in domains such as product recommendation, text mining, market basket analysis, and web click stream analysis.

Reshu Agarwal [30] suggested a method for finding high average-utility item sets (HAUIs) that takes into account both the length of the itemsets and their utilities. HUIs are found using the standard method based on the individual utility of an item set, which is calculated as the sum of the utilities of individual items. The difficulty is that the aforementioned method of computing HUIs does not take the length of the item set into account.

Wanyong Tian *et al.* [31] suggested a technique for uncertain frequent item sets called UFP-ECIS (Uncertain Frequent Pattern Mining with Ensembled Conditional Item-wise Supports). The difficulties of information redundancy and loss caused by a single probabilistic frequent threshold can be successfully improved by assembling numerous conditional item-wise supports. Furthermore, by employing several pruning algorithms based on the sorted downward closure feature and the concept of least minimal probability frequent threshold. Many existing frequent itemset mining algorithms have good performance but scalability problems as they cannot exploit parallel processing power available locally or in cloud infrastructure. Since big data and cloud ecosystem overcomes the barriers or limitations in computing resources, it is a natural choice to use distributed programming paradigms such as Map Reduce.

3. FAST AND SCALABLE FREQUENT ITEMSET MINING (FSFIM) ALGORITHM

The proposed algorithm is known as Fast and Scalable Frequent Itemset Mining (FSFIM) used to extract frequent item sets from big data. The data representation before discovering frequent itemsets is made using POC-tree. Therefore, the construction of the POC tree is an important part of the FSFIM. However, POC construction is made after producing frequent 1-itemsets. Based on the minimum support (statistical measure to know the quality of frequent pattern), after scanning the entire database, a set of frequent 1-itemsets, denoted as F1, with corresponding support is generated. Then the items in F1 are ordered using support-descending order. The ordered items are denoted as L1 where the frequent items that have the same support are just taken in any order. Then POC tree is constructed as follows. The first root of the tree, denoted as Tr, is created but it is set to "null". Then for every transaction in the given database frequent itemsets, in the order of F1, are selected and sorted. Let [p|P] denote a sorted frequent item list where the first element is denoted as p while others are denoted as P. Then the insert tree function is carried out. If the transaction has as child node N with the same item name as that of p, N's count is increased by 1, if not new node N is created and its count is initialized to 1. Then it is added to the transaction's children list. As far as P has some items, the insert tree is invoked recursively to complete POC tree construction. Afterward, the POC tree is scanned to, following pre-order traversal, have a pre-order of each node in the tree. Once the POC tree is constructed, it is possible to find all frequent 2-itemsets and corresponding node sets. Afterward, frequent k(>2)-itemsets are discovered. The algorithm has a pruning strategy known as promotion which depends on the notion of superset equivalence property. To represent each search space, the algorithm uses a set-enumeration tree as shown in Fig. 3.

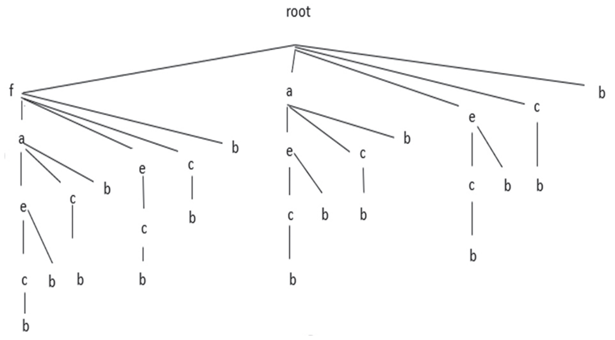


Fig. 3. Illustrates a sample set-enumeration tree

3.1. POC-TREE

The node sets data structure used in this paper is based on the POC (Pre-Order Coding) tree. Unlike its predecessor PPC-tree which needs encoding of nodes with pre-order and post-order codes causing overhead, POC-tree needs only either pre-order or post-order and not both. Since it is efficient, in this paper, we adapted it from [26]. POC-tree can be understood based on the data in Table 1 and the POC-tree representation in Fig.4. The tree has a root node labeled "null" and many subtrees where each node is prefixed by an item. Each node in the tree has different fields like item-name, children-list, count, and pre-order. Item node refers to an item the node represents. Count indicates the number of transactions denoted by the path till this node. Children-list refers to the children of the node and pre-order is the node's pre-order code. The POC tree is similar to that of the PPC tree where each node in the POC tree is encoded by its pre-order while the PPC tree has each node encoded by its pre-order and post-order. After generating the POC tree, the node sets are created. Once node sets are created, the POC tree is not required. The transactional items are provided in support-descending order in the last column of the table. The POC-tree representation helps in the faster generation of frequent itemsets. In other words, it accelerates the frequent itemset mining process. When there are large volumes of data, in the presence of cloud computing resources, it may be of use to leverage performance benefits.

Table 1. Shows a simple transactional database

ID	Ordered Frequent Items	Support-Descending Items
1	a, c, g, f	c, f, a
2	e, a, c, b	b, c, e, a
3	e, c, b, i	b, c, e
4	b, f, h	b, f
5	b, f, e, c, d	b, c, e, f

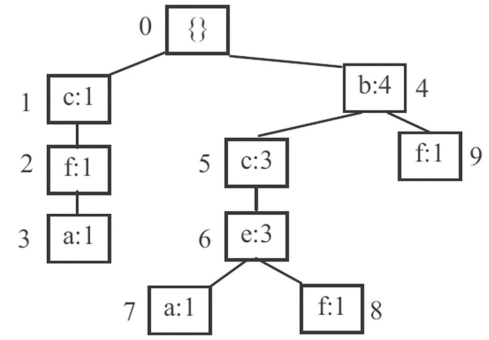


Fig. 4. POC-tree constructed for the data in Table

The FSFIM algorithm for mining Big Data works as follows: FSFIM takes transactional database T and minimum support threshold th as inputs. It produces frequent itemsets (results) R. Step 1 constructs the POC tree as discussed earlier in this section. Step 3 discovers all frequent 1-itemsets. Step 3 starts an iterative process for each node in the POC tree. Step 5 extracts the item associated with the visiting node in the POC tree. Step 5 through Step 11, there is an iterative process used to obtain all frequent 2-itemsets into F2. Steps 12 through 18, prunes infrequent 2-itemsets present in F2. Step 19 scans the POC tree and an iterative process from Step 20 through Step 29 discovers frequent (>2) itemsets. The map() function ends here by returning F representing frequent itemsets for the given portion of transactions. This way multiple map() functions operate based on the worker nodes involved in the map phase of the MapReduce paradigm. The intermediate results (F of each mapper) are given to reducers to combine results and return final frequent itemsets obtained from the given T.

Algorithm: Fast and Scalable Frequent Itemset Mining algorithm

Pseudocode: A Novel Nodesets Based Fast and Scalable Frequent Itemset Mining

Input: A transactional database T, minimum support threshold ***th***

Output: Discovered frequent itemsets R

```
Map() Function
1.      Construct POC tree
2.      $F1$ ← FindFrequent1Itemsets ()
3.      For each node $n$ in the POC tree
4.        item ← FindItem($n$)
5.        For each ancestor of $n$ and $n'$
6.          item$_{n'}$ ← FindItem($n'$)
7.          IF item and item$_{n'}$ belong to $F2$ Then
8.            support_of_item_item$_{n'}$
              support_of_item_item$_{n'}$ + $n$.acc
```

```
                $F2 \leftarrow F2 U\{item, item_n\}$
            End If
        End For
    For each itemset $Q$ in $F2$
        IF $Q.sup$<th $x$ $|T|$ Then
            $F2 \leftarrow F2-\{Q\}$
        Else
            $Q.nodeset \leftarrow$ null
        End If
    End For
    Scan POC tree
    For each node $n$ in the POC tree
        item ← FindItem($n$)
        For each ancestor of $n'$ and $n''$
            $item_{n''}$ ← FindItem($n''$)
            IF item and $item_{n'}$ belong to $F2$ Then
                $item\_item_{n'}.nodeset \leftarrow item\_item_{n''}.nodeset\ U\ item_{n''}.N\_info$
            End If
        End For
    $F$=$F$ U $F1$
    Return $F$
Reduce() Function
    For each $F$ in intermediate Frequent Itemsets
        $R$+=$F$
    End For
Return $R$
```

4. EXPERIMENTAL RESULTS AND ANALYSIS

We experimented with the Cloudera environment to evaluate the FSFIM algorithm and compare that with state-of-the-art methods such as Mahout PFP [32], Mlib PFP [33], and Big FIM [34]. For experiments, the real dataset known as Delicious explored in [35] is used. This dataset is a collection of web tags. Each record represents the tag assigned by a user to a URL and it consists of 4 attributes: date, user id (anonymized), tagged URL, and tag value. The transactional representation of the delicious dataset includes one transaction for each record, where each transaction is a set of four pairs (attribute, value), i.e., one pair for each attribute. The dataset stores more than 3 years of web tags. It is very sparse because of the huge number of different URLs and tags. A prototype application is built using Java language on top of the MapReduce paradigm to implement the proposed algorithm. The dataset has 41,949,956 transactions and 57,372,977 items. Observations are made in terms of execution time against different minimum support values and several attributes.

Table 2. Shows execution time of different algorithms against various mins up values

Minsup (%)	Execution Time (Seconds)			
	Mahout PFP	Mlib PFP	Big FIM	FSFIM
0	90000	10000	7500	5000
0.2	9500	8000	6200	3000
0.4	500	2500	500	300
0.6	400	900	400	250
0.8	400	600	400	150
1	400	400	400	100

As presented in Table 2, the execution time of the algorithms is provided for different minsup (%) values.

As presented in Fig. 5, it is evident that the minimum support values used for experiments are presented on the horizontal axis while the execution time is shown on the vertical axis. The results revealed that the proposed method FSFIM outperforms the state-of-the-art methods.

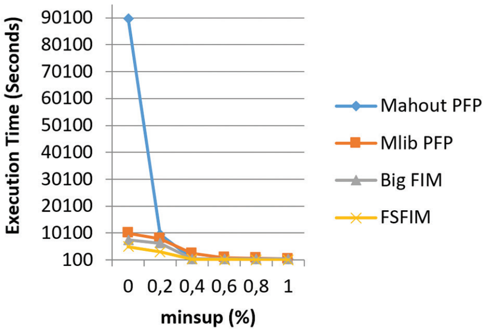


Fig. 5. Performance comparison in terms of execution time against different mins up (%)

As presented in Table 3, the execution time of the algorithms is provided for different transaction length values.

Table 3. Shows the execution time of different algorithms against various transaction lengths

Transaction Length	Execution Time (Seconds)			
	Mlib PFP	Mahout PFP	Big FIM	FSFIM
10	100	120	120	70
20	100	125	125	115
30	120	150	150	100
40	450	300	300	200
50	1250	425	425	354
60	3200	550	550	530
70	5100	600	600	585
80	7500	950	650	620
90	8600	1400	900	835
100	9800	1800	1200	1100

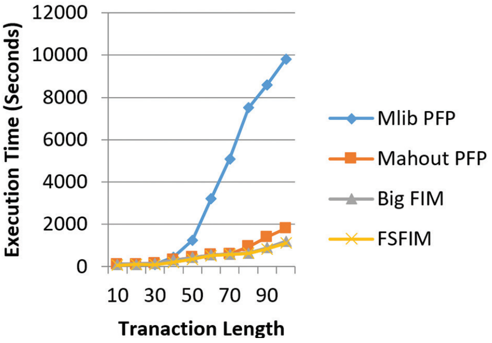


Fig. 6. Performance comparison in terms of execution time against transaction length

As presented in Fig. 6, it is evident that the transaction length values used for experiments are presented on the horizontal axis while the execution time is shown on the vertical axis. The results revealed that the proposed method FSFIM outperforms the state-of-the-art methods.

The execution time of the algorithms is provided for a different number of attributes is presented in Table 4.

Table 4. Shows the execution time of different algorithms against several attributes

Number of Attributes	Execution Time (Seconds)		
	Mahout PFP	Mlib PFP	FSFIM
0	100	400	50
10	1400	4200	600
20	1800	12100	800
30	2500	13900	1200
40	7000	18760	4500
50	18100	19850	15200

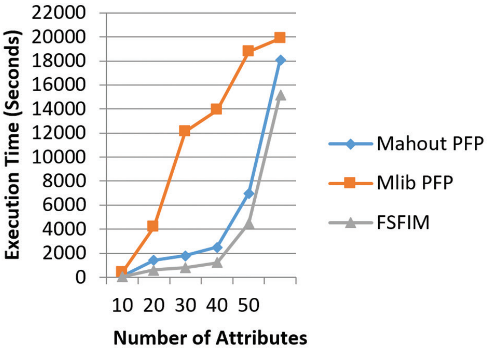


Fig. 7. Performance comparison in terms of execution time against the number of attributes

The FSMFI is better than the existing Mahout PFP, and Mlib PFP. As presented in Fig. 7, it is evident that the number of attributes used for experiments is presented in the horizontal axis while the execution time is shown in the vertical axis. The results revealed that the proposed method FSFIM outperforms the state-of-the-art methods. Experimental results revealed that FSFIM outperforms existing algorithms such as Mahout PFP, Mlib PFP, and Big FIM. FSFIM is more scalable and found to be an ideal candidate for real-time applications that mine frequent itemsets from big data.

5. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel algorithm known as A Novel Nodesets Based Fast and Scalable Frequent Itemset Mining (FSFIM) to extract frequent itemsets from big data. Pre-Order Coding (POC) tree is used to represent data and improve speed in processing. Nodeset is the underlying data structure that is efficient in discovering frequent itemsets. When compared with its predecessors such as Node-lists and N-lists, the Nodeset saves half of the memory as it needs only either a pre-order or post-order code. Cloudera's Distribution of Hadoop (CDH), a MapReduce framework, is used for empirical study. A prototype application is built to evaluate the performance of the FSFIM. Experimental results revealed that FSFIM outperforms existing algorithms such as Mahout PFP, Mlib PFP, and Big FIM. FSFIM is more scalable and found to be an ideal candidate for real-time applications that mine frequent itemsets from big data. Findings in the empirical study include faster execution and scalability. In the future, we would like to extend the FSFIM to support incremental mining of frequent itemsets to avoid scanning the entire database and reinventing wheel everything when the algorithm is executed.

6. REFERENCES

[1] H. Qiu, R. Gu, C. Yuan, Y. Huang, "YAFIM: A Parallel Frequent Itemset Mining Algorithm with Spark", Proceedings of the IEEE International Parallel & Distributed Processing Symposium Workshops, Phoenix, AZ, USA, 19-23 May 2014, pp. 1-8.

[2] M. Yasir et al. "TRICE: Mining Frequent Itemsets by Iterative TRimmed Transaction LattICE in Sparse Big Data", IEEE Access, Vol. 7, 2019, pp. 181688-181705.

[3] S. Gole, B. Tidke, "Frequent itemset mining for Big Data in social media using ClustBigFIM algorithm", Proceedings of the International Conference on Pervasive Computing, Pune, India, 8-10 January 2015, pp. 1-6.

[4] Y. Djenouri, D. Djenouri, J. C.-W. Lin, A. Belhadi, "Frequent Itemset Mining in Big Data with Effective Single Scan Algorithms", IEEE Access, Vol. 6, 2018, pp. 1-15.

[5] D. Apiletti, E. Baralis, T. Cerquitelli, P. Garza, F. Pulvirenti, L. Venturini, "Frequent Itemsets Mining for Big Data: A Comparative Analysis", Big Data Research, Vol. 9, 2017, pp. 67-83.

[6] C. Fernandez-Basso, A. J. Francisco-Agra, M. J. Martin-Bautista, M. D. Ruiz, "Finding tendencies in streaming data using Big Data frequent itemset mining", Knowledge-Based Systems, 2018, pp. 1-21.

[7] K. K. Sethi, D. Ramesh, "HFIM: a Spark-based hybrid frequent itemset mining algorithm for big data processing", The Journal of Supercomputing, Vol. 73, No. 8, 2017, pp. 3652-3668.

[8] K.-W. Chon, S.-H. Hwang, M.-S. Kim, "GMiner: A fast GPU-based frequent itemset mining method for large-scale data", Information Sciences, Vol. 439-440, 2018, pp. 19-38.

[9] Y.-H. Liang, S.-Y. Wu, "Sequence-Growth: A Scalable and Effective Frequent Itemset Mining Algorithm for Big Data Based on MapReduce Framework", Proceedings of the IEEE International Congress on Big Data, New York, NY, USA, 2015, pp. 1-8.

[10] R. Joy, K. K. Sherly, "Parallel frequent itemset mining with spark RDD framework for disease prediction", Proceedings of the International Conference on Circuit, Power and Computing Technologies, Nagercoil, India, 18-19 March 2016, pp. 1-5.

[11] Y. Djenouri, D. Djenouri, A. Belhadi, A. Cano, "Exploiting GPU and cluster parallelism in single scan frequent itemset mining", Information Sciences, Vol. 496, 2018, pp. 1-15.

[12] K.-W. Chon, M.-S. Kim, “BIGMiner: a fast and scalable distributed frequent pattern miner for big data”, Cluster Computing, Vol. 21, 2018, pp. 1-14.

[13] S. Raj et al. “EAFIM: efficient apriori-based frequent itemset mining algorithm on Spark for big transactional data”, Knowledge and Information Systems, Vol. 62, 2020, pp. 3565-3583.

[14] S. Moens, E. Aksehirli, B. Goethals, “Frequent Itemset Mining for Big Data”, Proceedings of the IEEE International Conference on Big Data, Silicon Valley, CA, USA, 6-9 October 2013, pp. 1-8.

[15] Y. Xun, J. Zhang, X. Qin, X. Zhao, “FiDoop-DP: Data Partitioning in Frequent Itemset Mining on Hadoop Clusters”, IEEE Transactions on Parallel and Distributed Systems, Vol. 28, No. 1, 2017, pp. 101-114.

[16] A. Asbern, P. Asha, “Performance evaluation of association mining in Hadoop single node cluster with Big Data”, Proceedings of the International Conference on Circuits, Power and Computing Technologies, Nagercoil, India, 19-20 March 2015, pp. 1-5.

[17] S. Kumar, K. K. Mohbey, “A review on big data parallel and distributed approaches of pattern mining”, Journal of King Saud University - Computer and Information Sciences, Vol. 34, No. 5, 2019, pp. 1-24.

[18] M. Zitouni, R. Akbarinia, S. B. Yahia, F. Masseglia, “A Prime Number Based Approach for Closed Frequent Itemset Mining in Big Data”, Proceedings of Database and Expert Systems Applications, Valencia, Spain, pp. 509-516.

[19] C. K. Leung, H. Zhang, J. Souza, W. Lee, “Scalable Vertical Mining for Big Data Analytics of Frequent Itemsets”, Proceedings of Database and Expert Systems Applications, 2018, pp. 3-17.

[20] P. Galetsi, K. Katsaliaki, S. Kumar, “Big data analytics in the health sector: Theoretical framework, techniques, and prospects”, International Journal of Information Management, Vol. 50, 2020, pp. 206-216.

[21] C. Fernandez-Basso, M. D. Ruiz, M. J. Martin-Bautista, ''A fuzzy mining approach for energy efficiency in a Big Data framework”, IEEE Transactions on Fuzzy Systems, Vol. 28, 2020, pp. 1-12.

[22] F. Fumarola, D. Malerba, ''A parallel algorithm for approximate frequent itemset mining using MapReduce”, Proceedings of the International Conference on High-Performance Computing & Simulation, Bologna, Italy, 21-25 July 2014, pp. 1-8.

[23] Y. Djenouri et al. ''A Novel Parallel Framework for Metaheuristic-based Frequent Itemset Mining“, Proceedings of the IEEE Congress on Evolutionary Computation, Wellington, New Zealand, 10-13 June 2019, pp. 1-7.

[24] A. Aggarwal, D. Toshniwal, “Frequent Pattern Mining on Time and Location Aware Air Quality Data”, IEEE Access, Vol. 7, 2019, pp. 98921-98933.

[25] J. M. Luna, F. Padillo, M. Pechenizkiy, S. Ventura, “Apriori Versions Based on MapReduce for Mining Frequent Patterns on Big Data”, IEEE Transactions on Cybernetics, Vol. 48, No. 10, 2017, pp. 2851-2865.

[26] S. Nalousi, Y. Farhang, A. B. Sangar, "Weighted Frequent Itemset Mining Using Weighted Subtrees: WST-WFIM", IEEE Canadian Journal of Electrical and Computer Engineering, Vol. 44, No. 2, 2021, pp. 206-215.

[27] F. Li, Z. Zhang, B. Cheng, P. Zhang, "Probabilistic Fuzzy Frequent Item Sets Mining (PPFIM),” Proceedings of the 7th International Conference on Cloud Computing and Big Data Analytics, Chengdu, China, 2022, pp. 127-132.

[28] A. Sharma, R. K. Singh, "An Efficient Approach to Find Frequent Item Sets in Large Database”, Proceedings of the 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology, Bhubaneswar, India, 2021, pp. 1-6.

[29] B. Sivaiah, R. R. Rao, "A Survey on Fast and Scalable Incremental Frequent Item Set Methods for Big Data”, Proceedings of the International Conference on Intelligent Controller and Computing for Smart Power, Hyderabad, India, 2022, pp. 1-5.

[30] R. Agarwal, A. Gautam, A. K. Saksena, A. Rai, S. V. Karatangi, "Method for Mining Frequent Item Sets Considering Average Utility”, Proceedings of the International Conference on Emerging Smart Computing and Informatics, Pune, India, 2021, pp. 275-278.

[31] W. Tian, F. Li, Y. Liu, Z. Wang, T. Zhang, "Depth-First Uncertain Frequent Itemsets Mining based on Ensembled Conditional Item-Wise Supports", Proceedings of the International Conference on Intelligent Supercomputing and BioPharma, Zhuhai, China, 2023, pp.121-128.

[32] S. Bagui, K. Devulapalli, J. Coffey, "A heuristic approach for load balancing the FP-growth algorithm on MapReduce", Array, Vol. 7, 2020, p. 100035.

[33] X. Meng et al. “Mllib: Machine learning in Apache spark”, Journal of Machine Learning Research, Vol. 17, No. 1, 2016, pp. 1235-1241.

[34] Y. Rochd., I. Hafidi, B. Ouartassi, “A Review of Scalable Algorithms for Frequent Itemset Mining for Big Data Using Hadoop and Spark”, Real-Time Intelligent Systems, Lecture Notes in Real-Time Intelligent Systems, Springer, 2017, pp. 90-99.

[35] R. Wetzker, C. Zimmermann, C. Bauckhage, “Analyzing social bookmarking systems: adel.icio.us cookbook”, Mining Social Data Workshop Proceedings, 2008, pp. 26-30.

Patterns Identification of Finger Outer Knuckles by Utilizing Local Directional Number

Original Scientific Paper

Raid Rafi Omar Al-Nima
Technical Engineering College of Mosul,
Northern Technical University, Mosul, Iraq
raidrafi3@ntu.edu.iq

Hasan Maher Ahmed
Software Department,
College of Computer Science and Mathematics,
University of Mosul, Mosul, Iraq
hasanmaher@uomosul.edu.iq

Nagham Tharwat Saeed
Department of Computer Science,
College of Education for Pure Science,
University of Mosul, Mosul, Iraq
nagham.th@uomosul.edu.iq

Abstract *– Finger Outer Knuckle (FOK) is a distinctive biometric that has grown in popularity recently. This results from its inborn qualities such as stability, protection, and specific anatomical patterns. Applications for the identification of FOK patterns include forensic investigations, access control systems, and personal identity. In this study, we suggest a method for identifying FOK patterns using Local Directional Number (LDN) codes produced from gradient-based compass masks. For the FOK pattern matching, the suggested method uses two asymmetric masks—Kirsch and Gaussian derivative—to compute the edge response and extract LDN codes. To calculate edge response on the pattern, an asymmetric compass mask made from the Gaussian derivative mask is created by rotating the Kirsch mask by 45 degrees to provide edge response in eight distinct directions. The edge response of each mask and the combination of dominating vector numbers are examined during the LDN code-generating process. A distance metric can be used to compare the LDN code's condensed representation of the FOK pattern to the original for matching purposes. On the Indian Institute of Technology Delhi Finger Knuckle (IITDFK) database, the efficiency of the suggested procedure is assessed. The data show that the suggested strategy is effective, with an Equal Error Rate (EER) of 10.78%. This value performs better than other EER values when compared to different approaches.*

Keywords*: Finger Outer Knuckles, Local Directional Number, Pattern Identification, Image Processing*

1. INTRODUCTION

Access control and proper personal identification technologies are necessary to create safe and efficient systems. Each person's internal and external physiological biometrics have several characteristics in common. With imaging technology, even in a touchless manner, they are more practical and simple to use. Because of the intricate architecture of the human hand, remarkable recognition abilities can be found on the dorsal surface of the fingers. [1]. Finger Outer Knuckles (FOK) have a distinctive pattern formation that makes it a promising biometric identifier for advancements in personal identification because it is both pleasant and touchless [2]. The asymmetry of the finger is caused by the anatomy's capacity for forward bending and resistance to backward movement, which causes many wrinkles on the palm.

Metacarpophalangeal joints, commonly referred to as finger outer knuckles, are important joints that attach the fingers to the hand's bones. The shape of the knuckles is important for dexterity, grip strength, and hand movements. Due to their distinctive patterns, there has been an increase in interest in employing finger outer knuckles for biometric authentication in recent years [3, 4]. However, their complicated and uneven structure makes finger outer knuckles difficult to find patterns in. Thankfully, new developments in machine learning and image processing methods have made it possible to evaluate these patterns precisely.

One such method involves identifying and categorizing the patterns in the finger's outer knuckles using local directional numbers (LDNs). LDNs are mathematical descriptors that quantify the gradient orientation at each pixel in an image. In numerous computer vision applications, such as texture analysis and object recognition, they successfully measure the local directional information in an image [5, 6]. Researchers can employ LDNs to extract directional information from patterns on finger outer knuckles and classify such patterns into various groups.

There are numerous issues that need to be resolved because using LDNs to determine finger outer knuckle patterns is still a relatively new technology. For joint fingerprinting, central joint line extraction, for instance, is a considerable challenge, necessitating the implementation of dependable components in addition to the knuckle print matching technique [7, 8].

There are numerous possible uses for identifying finger outer knuckle patterns utilizing LDNs. It might be utilized, for instance, in biometric authentication systems, where the patterns could act as distinctive personal IDs for people. Additionally, it could be applied to medical imaging to detect disorders that affect the finger joints or injuries to the hands [9, 10].

Utilizing the LDN to FOKs for identification is the paper's goal and main contribution. We intend to open up a new arena for the problem of identifying people by employing this method.

This paper is structured as follows: Section 1 presents the introduction, Section 2 reviews prior studies, Section 3 explains the proposed approach, Section 4 discusses obtained results and Section 5 provides the conclusion.

2. FOK LITERATURE REVIEW

To correctly recognize finger outer knuckle (FOK) patterns, portions of the finger or hand must be segmented. False positives or false negatives may come from misidentification caused by improper segmentation [11]. The process of segmentation involves separating regions or important objects from an image or data set. Finding the finger or hand joint regions that contain the FOK patterns requires precise segmentation [12].

The clarity of the segmentation can be affected by several variables, such as the intricacy of the hand or finger structure, lighting circumstances, and image quality. To increase segmentation accuracy, researchers have created a variety of methodologies, including deep learning-based approaches and morphological procedures [13]. Furthermore, using various imaging techniques like magnetic resonance imaging (MRI) and ultrasound can provide more thorough details on the structure of the hand or finger, which can help with precise segmentation and FOK pattern identification [14, 15].

FOK patterns are built on lines easily retrieved using a low-cost sensor, making them a promising biometric modality. FOKs are a feasible alternative to other biometric modalities like fingerprints because they are naturally protected and less likely to change due to aging or injury [16]. Finding FOK patterns, meanwhile, comes with its own set of difficulties. The FOK patterns may appear differently in each image due to rotations and translations, which makes it more difficult to match them precisely. Illumination contrast can also lead to poor image quality [17, 18].

Typically, the FOK imprint is made up of two parallel lines joined by smaller lines to create a unique design. This pattern can be extracted by a number of sensors, including contact-based sensors, capacitive sensors, and optical sensors [19]. Different directions of the FOK have rich pattern structure lines, however, they are severely discriminated against. When compared to fingerprints, the failure rates of FOK patterns are anticipated to be lower as these lines tend to fade with age [20, 21].

Calculating the number of focus edges, the amount of clutter, the focus edge distribution, the entropy sensible of the focused advantages, the reflection caused by the light source or camera flash, and the quantity of contrast can increase the quality of the knuckle print image. To ensure durability against changing illumination, features based on vertical and horizontal joint lines may be of poor quality and require refinement and conversion [22, 23].

3. PROPOSED APPROACH

3.1. DESCRIPTION

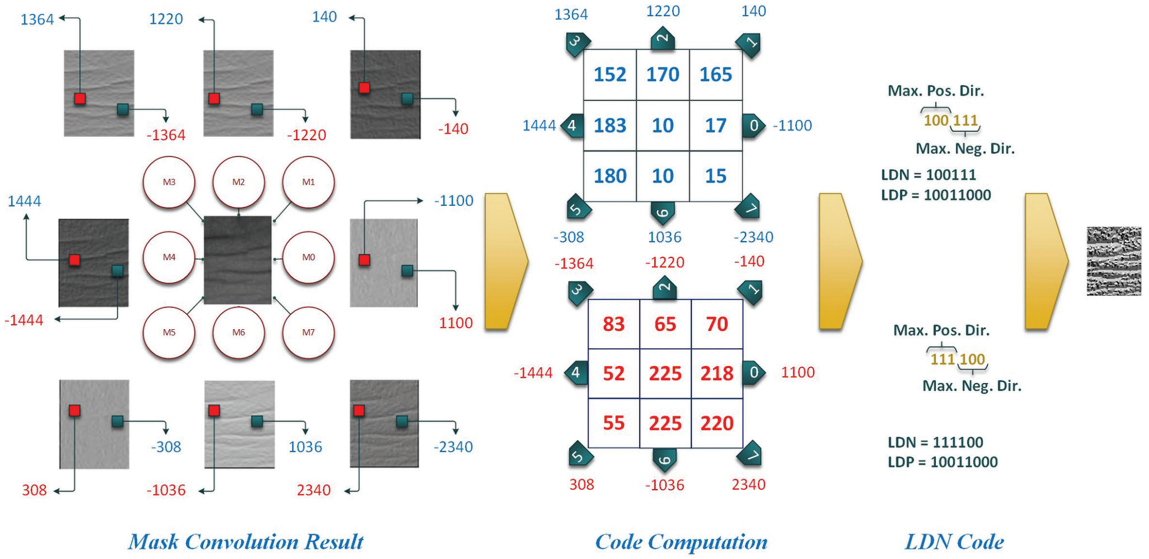
The suggested method entails examining the structural data and density variations in the tissue of the outer finger joints using image processing methods. The local area structure of the FOK is encrypted by the LDN algorithm, which then examines the data it contains. The edge responses are then calculated using a compass mask in eight different orientations. As demonstrated in Fig. 1, a relevant descriptor for the tissue's structural pattern is generated by choosing the highest positive and negative trends. This method enables the separation between texture intensity variations, such as those from bright to dark and vice versa, which can aid in differentiating between related patterns.

The suggested method has several benefits over conventional fingerprinting methods. It can be utilized when standard fingerprinting isn't an option, as when the skin is severely injured or calloused. Furthermore, LDN analysis can offer a higher level of precision when identifying people based on their FOKs. Instead of using isolated calculated points, the complete information area is utilised, and the data is converted to a six-bit code. The encoded information is expanded by applying various masks and resolutions on the mask to gain properties that a single mask could overlook. Multiple encoding levels have been shown to enhance the detecting process.

It is essential to use all available information for precise pattern recognition while studying Finger Outer Knuckles (FOKs) utilizing Local Directional Number (LDN) approaches. To do this, a six-bit code based on the structural patterns and density variations in the imaged tissue of the FOK is formed, and various masks and resolutions are applied to the mask to obtain more information. The detecting method is further enhanced by having various encoding levels since different codes can be merged to produce a more precise representation of the FOK pattern.

The direction of the gradient between bright and dark areas is revealed by the positive and negative computations in LDN, which offers important information on tissue architecture.

The ability of LDN to distinguish between blocks while reversing the positive and negative axes is crucial for correctly identifying specific texture modifications. These transformations are given a specific code by LDN, ensuring that they are appropriately distinguished.



Mask Convolution Result *Code Computation* *LDN Code*

Fig. 1. Demonstration of computing the LDN code

In comparison to conventional fingerprinting methods, the suggested LDN method offers a more thorough and effective means of recognizing patterns of finger outer knuckles. The LDN method can offer improved precision, resolution, and sensitivity to changes in texture and lighting conditions by making use of the complete information region of the FOK, applying several masks with varying resolutions, and adopting a more reliable and effective encoding scheme.

Additionally, the LDN algorithm outperforms conventional texture analysis methods since it is largely insensitive to variations in lighting. Further expanding the encoded information and enhancing the detection process are the use of several encoding levels and the application of various masks with various resolutions.

By examining edge responses in eight distinct directions and utilizing two different masks—Gaussian-derivative and Kirsch's compass mask—the LDN algorithm can give more resolution and sensitivity to changes in texture and lighting conditions than the Local Binary Pattern (LBP) technique. The LDN algorithm also filters and prioritizes local information prior to coding, minimizing the effects of low resolution and noise sensitivity.

The suggested method for recognizing patterns of finger outer knuckles using local directional number (LDN) has several advantages over conventional fingerprinting approaches, including greater accuracy, resolution, and sensitivity to changes in texture and lighting conditions. The LDN algorithm can extract more meaningful patterns from the input image by using a more thorough approach that considers the input image's information and applying multiple masks with various resolutions, making it a more robust and successful method for pattern identification of finger outer knuckles.

The proposed work (as in Fig. 2) can be summarized through the following algorithm:

1. Obtain a dataset of images of the Finger Outer Knuckles (FOK) for testing and training the algorithm, and then increase contrast and eliminate noise from the photos.
2. Employ the suggested technique to extract Local Directional Number (LDN) codes from the preprocessed images. The directional information can be extracted from the FOK patterns using the LDN codes, which quantify the gradient orientation at each pixel in the image. The following steps can be used to calculate the LDN codes:
   a. Blocks of a specified size, with no overlap, should be created from the preprocessed image.
   b. Use a filter to calculate the gradient's strength and direction at each pixel in each block.
   c. Quantize the gradient direction into a set of bins of a certain size, such 8 or 16.

d. Concatenate the quantized gradient orientations of each pixel's surrounding pixels in a circular pattern to determine the LDN code for each pixel in each block.

3. Determine the model's Equal Error Rate (EER), a widely used parameter in biometric identification. The False Acceptance Rate (FAR) and False Rejection Rate (FRR) are identical at the EER. The FRR is the likelihood of rejecting a legitimate user as an impostor, while the FAR is the likelihood of accepting an impostor as a genuine user.
4. Examine the findings and contrast the suggested method's EER with those of other FOK pattern recognition techniques already in use. The objective is to demonstrate that the proposed technique outperforms or performs on par with competing methods in terms of accuracy and resilience.
5. To attain the lowest EER feasible, optimize the proposed method's parameters, such as the sigma value. The sigma value affects the method's ability to discriminate between different data sets by regulating the size of the Gaussian kernel used to smooth the LDN codes. A grid search or a random search over a range of sigma values can be used to perform the optimization.
6. Repeat steps 2 through 5 until the suggested technique performs admirably on the test dataset.

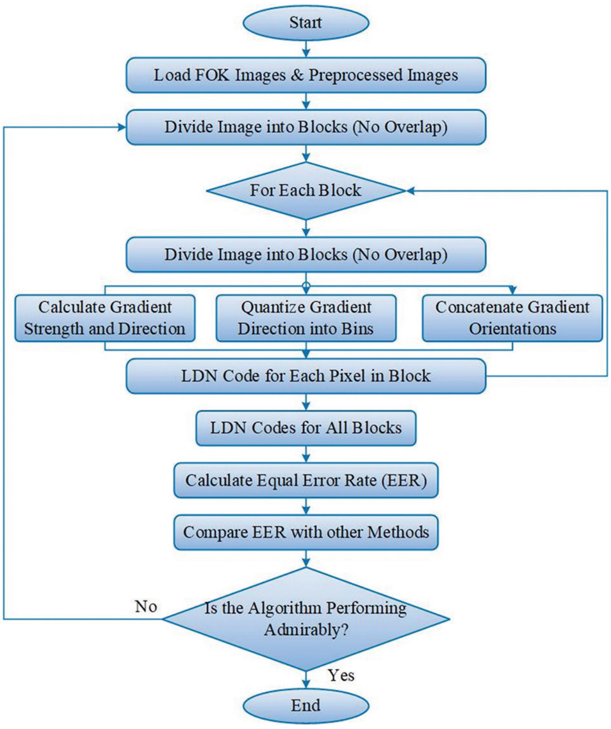

Start
Load FOK Images & Preprocessed Images
Divide Image into Blocks (No Overlap)
For Each Block
Divide Image into Blocks (No Overlap)
Calculate Gradient Strength and Direction
Quantize Gradient Direction into Bins
Concatenate Gradient Orientations
LDN Code for Each Pixel in Block
LDN Codes for All Blocks
Calculate Equal Error Rate (EER)
Compare EER with other Methods
No
Is the Algorithm Performing Admirably?
Yes
End

Fig. 2. General Flowchart for Research Method

3.2. CODING SCHEME

In the suggested work, LDN codes are created by analyzing each mask's edge response ($M0$,..., $M7$) and the combination of dominating vector numbers. An use of feature descriptors called LDN codes is in image analysis and computer vision. Each mask's edge response identifies any salient dark or bright areas in the image, and the signal data is then implicitly used to encode these regions. The three most significant bits in the code reflect the highest positive directional number, which is given a fixed position in the code [24]. The largest negative directional number is represented by the three least significant bits, as seen in Fig. 1. The following is a definition of the key LDN equation:

$$LDN(x,y) = 8i_{x,y} + j_{x,y} \tag{1}$$

where (*x, y*) is the center pixel of the region being coded, $i_{x,y}$ is the vector number of the maximum positive response, and $j_{x,y}$ is the vector number of the minimum negative response specified by:

$$\begin{aligned} i_{x,y} &= \arg\max_i \{\mathbb{I}^i(x,y) \mid 0 \le i \le 7\}, \\ j_{x,y} &= \arg\min_j \{\mathbb{I}^j(x,y) \mid 0 \le j \le 7\}, \end{aligned} \tag{2}$$

where $\mathbb{I}^i$ is the convolution of the original image I for the i^{th} mask, M^i is defined by:

$$\mathbb{I}^i = I * M^i \tag{3}$$

3.3. COMPASS MASKS

Compass masks are a sort of filter used in image processing for edge detection and feature extraction. These masks are used in the planned work for FOK pattern matching and identification. One of the main benefits of employing compass masks is that the LDN code is computed using the gradient area rather than the density feature area. Because it implicitly carries the relationships between pixels in the image, the gradient area has more information than the density feature area. This makes it possible to describe the image structure more accurately and identify important tissue features more effectively [25].

In order to increase the stability of the gradient computation, Gaussian smoothing is additionally applied to the image before gradient calculation. Gaussian smoothing optimizes the accuracy of the LDN code calculation by lowering image noise and enhancing image clarity. The proposed method is strengthened and made more trustworthy for matching and identifying FOK patterns as a result of these procedures. Even in the face of noise and other image abnormalities, the employment of compass masks and Gaussian smoothing enables a more accurate depiction of the image structure and a more effective identification of significant tissue characteristics.

A compass mask must be supplied to compute the image's edge responses and generate the LDN code. Two asymmetric masks—Kirsch and Gaussian derivative masks—were examined for FOK pattern detection in the research that was proposed. Compass masks, such as Kirsch masks, comprise eight distinct 3x3 kernels. The edge response is determined by the convolution of the kernel with the image, with each kernel being orientated differently. The Kirsch mask is a flexible option for edge detection in image processing since it can identify edges in all directions.

On the other hand, the Gaussian derivative mask is a sort of filter that determines the edge response using the gradient space of the picture. In order to stabilize the code in the presence of noise, it additionally employs Gaussian smoothing. Gaussian smoothing optimizes the accuracy of the LDN code calculation by lowering image noise and enhancing image clarity.

Both masks work in the image's gradient space, which displays the image's fundamental structure and improves the method's discrimination when identifying relevant tissue characteristics. The LDN code calculation's stability is further enhanced by using Gaussian smoothing, which increases the method's resistance to noise and other image distortions. To acquire the edge response in eight different directions, the Kirsch mask employed in the proposed work is rotated 45 degrees from the horizontal and vertical directions, as illustrated in Fig. 1. The Kirsch mask inspired the LDNK method, which is used to identify FOK patterns using this mask.

The suggested technique uses the Kirsch mask and an oblique Gaussian derivative mask to produce an asymmetric compass mask for computing the edge response on texturing. This mask is intended to produce strong edge responses while resisting noise and brightness variations.

A Gaussian function is convolved with the derivative of the picture to produce the Gaussian derivative mask that is employed in the proposed work. The standard deviation, which establishes how much smoothing is applied to the image, defines the Gaussian function. The derivative is computed in a certain direction, which shows the mask's orientation. Gaussian mask selection is based on:

$$\begin{aligned} G_\sigma(x,y) &= \tfrac{1}{2\pi\sigma^2}\exp\left(-\tfrac{x^2+y^2}{2\sigma^2}\right) \\ M_\sigma(x,y) &= G'_\sigma(x+k,y) * G_\sigma(x,y) \end{aligned} \tag{4}$$

where σ is the width of the Gaussian bell, G is the derivative of G_σ concerning x, $*$ is the convolution process and k is the Gaussian displacement concerning its center as a quarter of the mask diameter used for this displacement. Next, an $\{M_{0\sigma}$ compass mask is created. . ., $M_{7\sigma}\}$ alternating M_σ, 45 each separately in eight different directions. Thus, a set of masks similar to that shown in Fig. 1 is obtained. Because M_σ rotates the mask, there is no need to compute the derivative concerning y (since it is equivalent to 90 degrees rotating mask) or some other combination of these variables.

4. RESULTS AND DISCUSSIONS

4.1. EMPLOYED DATABASE

The Indian Institute of Technology Delhi Finger Knuckle (IITDFK) database was used in this investigation [26]. This database includes 500 segmented pictures of 100 subjects' finger outer knuckle (FOK) patterns. There are five grayscale photos of each subject, each with an 80x100 pixel resolution.

The subjects for the photographs in the IITDFK database placed their fingers on a level surface with a black background, and the photographs were taken with a typical digital camera under controlled lighting settings. The subjects were standing about 30 cm away from the camera when the pictures were taken.

Several earlier studies on identifying and recognizing FOK patterns utilized the IITDFK database. It has been extensively used to assess the efficacy of several methods for identifying FOK patterns, such as texture analysis, feature extraction, and machine learning techniques [26].

A standardized database, like the IITDFK database, can compare and evaluate various methods for consistently identifying FOK patterns. It also makes it easier to create and refine FOK pattern recognition systems for a variety of practical uses, such as forensic analysis, access control, and personal identity.

4.2. RESULTS

On the Indian Institute of Technology Delhi Finger Knuckle (IITDFK) database, the proposed method for FOK pattern recognition utilizing Local Directional Number (LDN) codes produced from gradient-based compass masks was assessed. The database includes 100 participants' left and right index, middle, and ring fingers in nine photographs for each of the collection's 500 images of FOK patterns.

The LDN codes were created for each image in the database using Kirsch and Gaussian derivative masks in order to assess the effectiveness of the suggested method. The LDN codes of each pair of photos were then compared using a distance metric, and the results were examined. In this investigation, 100 participants were taken into account. As said, there are 5 photos for each theme. These result in 123750 attempts for impostor comparisons and 1000 attempts for real comparisons during identification.

First of all, experiments are implemented by using the Matlab software (version R2020a) and a computer with the following hardware specifications: hp laptop, Intel core i7 processor, 2.70GHz processor speed and 8GB Random Access Memory (RAM). The performance of the approach is assessed using the Equal Error Rate (EER), which is one of the most considered essential metrics in biometrics. Its lower numbers indicate greater performance and vice versa. The outcomes show that the suggested approach identified FOK patterns successfully with an EER of 10.78%. It can be seen that the proposed technique outperforms the other ways by contrasting the EER values of the two methods. This shows that the suggested strategy is more accurate and effective for identifying FOK patterns. The table also demonstrates that altering the sigma value significantly affects the EER, with the lowest EER being attained at a sigma value of 0.85. This shows that choosing the right sigma value is essential for using the proposed method to identify FOK patterns with high accuracy.

The suggested method's low loss rates show that LDN codes created using gradient-based compass masks are a dependable and effective for identifying FOK patterns. The FOK pattern is uniquely represented by the asymmetric Kirsch and Gaussian derivative masks used to create the LDN codes, enabling accurate matching even in the presence of noise and distortion.

The proposed approach is suited for usage in various real-world settings, including forensic investigations, access control systems, and personal identification, according to the low loss rates obtained for each finger. However, the right ring finger's somewhat greater loss rate suggests that additional research may be required to optimize the strategy for this finger.

4.3. DISCUSSIONS

We conducted experiments using various mask sizes while maintaining the default values of mask type ('kirsch') and sigma value (0.5) to assess the impact of modifying the mask size on identification outcomes. The identification outcomes for varying the mask size when using the standard mask type of "kirsch" and a sigma value of 0.5 are shown in Table 1. EER, a popular statistic used in biometric identification to assess a system's performance, is used to measure the outcomes.

Table 1. Identification results for changing the mask size and using the default values of mask = 'kirsch' and sigma = 0.5

Ind.	Mask Size	EER (%)	Accuracy (%)
1	3x3	15.09	84.91
2	5x5	12.52	87.48
3	7x7	11.92	88.08
4	9x9	11.71	88.29
5	11x11*	12.91	87.09

* a problem with sizing appeared, so, it has been solved by resizing.

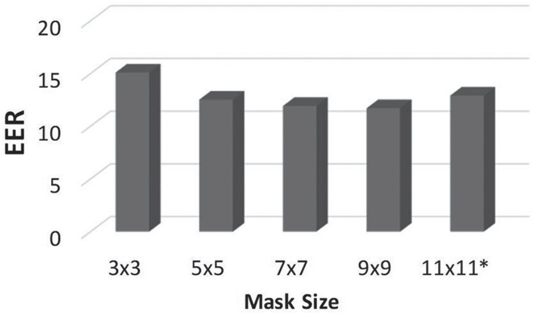


Fig. 3. bar chart to identify results for changing the mask size

AS INDICATED IN THE TABLE, the EER reduces with increasing mask size, with a mask size of 99 producing the lowest EER of 11.71%. This implies that increasing the mask size can increase the FOK pattern identification method's accuracy when employing LDN codes.

The results do, however, also suggest that there is a limit beyond which raising the mask size can reduce the system's performance. For instance, a resizing issue occurred when a mask size of 1111 was employed, leading to a higher EER of 12.91%.

When employing the LDN code technique for identifying FOK patterns, the findings in this table emphasize how crucial it is to choose the right mask size carefully. High levels of accuracy and improved system performance can be obtained by selecting a suitable mask size.

Table 2. Identification results for changing the mask type and using the values of mask size = 9x9 and sigma = 0.5

Ind.	Mask Type	EER (%)	Accuracy (%)
1	Prewitt	13.98	86.02
2	Sobel	13.21	86.79
3	Kirsch	11.71	88.29
4	Gaussian	11.50	88.5

With a fixed mask size of 99 and a sigma value of 0.5, Table 2 shows the identification results for varying the mask type. EER, a popular metric for assessing system performance in biometric identification, is used to measure the outcomes.

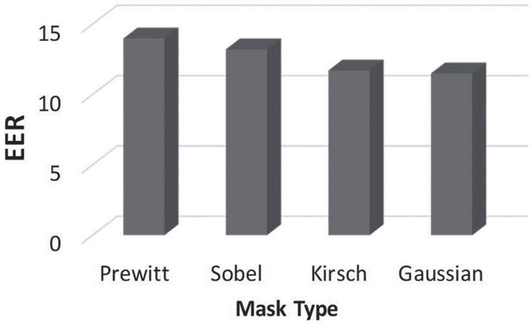


Fig. 4. bar chart to identify results for changing the mask type

With an EER of 11.50%, the table shown demonstrates that the FOK pattern identification strategy employing LDN codes performs best when using the Gaussian mask type. With an EER of 11.71, the Kirsch mask type also performs well. Prewitt and Sobel masks, on the other hand, had greater EERs, at 13.98% and 13.21%, respectively.

These findings emphasize the significance of choosing the proper mask type when employing the LDN code technique for identifying FOK patterns, as the choice of mask type can significantly affect the system's performance. The dataset in this scenario responds best to the Gaussian mask type, closely followed by the Kirsch mask type. It is crucial to keep in mind, nevertheless, that the ideal mask type could change based on the application in question and the features of the FOK patterns being examined.

It is significant to note that the Gaussian mask type has no impact on the values of mask size. The parameter that is impacted in this scenario is the sigma value. To identify the ideal sigma value for the Gaussian mask type, more tests must be done.

The identification outcomes for varying the sigma value while utilizing the Gaussian mask type, with the EER as the generally employed performance metric in biometric identification, are shown in Table 3.

The EER first drops as the sigma value rises, reaching a minimum value, as indicated in the table. After that, the EER increases once more as the sigma value grows. This dataset's ideal sigma value appears to be 0.85, which produced the lowest EER of 10.78%.

The EERs shown in this table are often greater than the accuracy numbers reported in the previous tables, which is important to note. This is so because accuracy simply considers the quantity of correctly identified patterns, whereas EER also considers incorrect acceptance and false rejection rates.

Table 3. Identification results for changing the sigma value and using the mask type of Gaussian

Ind.	Sigma Value	EER (%)	Accuracy (%)
1	0.05	15.36	84.64
2	0.10	15.36	84.64
3	0.15	15.22	84.78
4	0.20	15.50	84.5
5	0.25	15.20	84.8
6	0.30	15.02	84.98
7	0.35	11.89	88.11
8	0.40	11.83	88.17
9	0.45	11.66	88.34
10	0.50	11.50	88.5
11	0.55	11.63	88.37
12	0.60	11.69	88.31
13	0.65	11.25	88.75
14	0.70	11.51	88.49
15	0.75	10.99	89.01
16	0.80	10.88	89.12
17	0.85	10.78	10.78
18	0.90	10.89	89.11
19	0.95	11.04	88.96
20	1.00	11.42	88.58

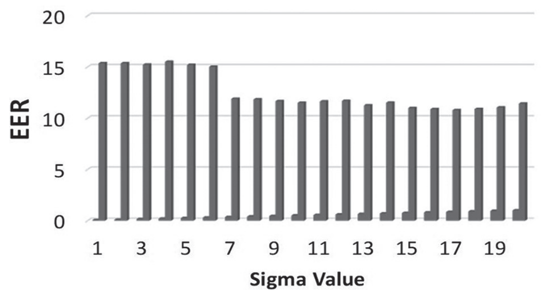


Fig. 5. Bar chart to identification results for changing the sigma value and using the mask type of Gaussian

These findings imply that selecting a sigma value can significantly influence the effectiveness of the FOK pattern detection approach employing LDN codes. In this instance, a sigma value of 0.85 seems to work well for this dataset. However as was already established, the ideal sigma value may change based on the particular application and the properties of the FOK patterns being examined.

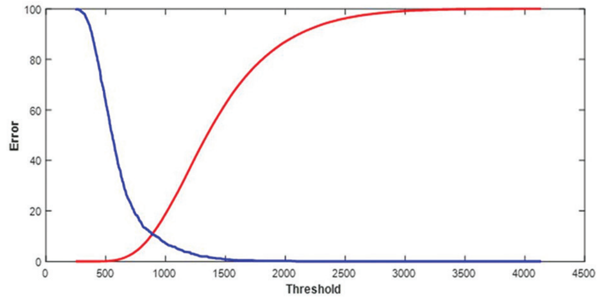


Fig. 6. FAR versus FRR for the proposed identification

The suggested FOK pattern detection method's False Accept Rate (FAR) against False Reject Rate (FRR) is plotted in Fig. 6. The error rate is shown on the x-axis, while the threshold is on the y-axis. As the plot shows, a decreasing FAR corresponds to an increasing FRR, and vice versa, which illustrates a trade-off between the FAR and FRR. The plot can be used to choose the best threshold for a particular application and aids in visualizing how well the suggested strategy performs at various threshold values.

A Receiver Operating Characteristic (ROC) curve plot of the suggested FOK pattern identification method's performance is shown in Fig. 7. Genuine Attempts Accepted (1-FRR) is shown on the x-axis, and Impostor Attempts Accepted (1-FAR) is. The True Positive Rate (TPR) against False Positive Rate (FPR) at various threshold settings is plotted on the ROC curve. As the curve shows, a greater TPR leads to a higher FPR, which illustrates a trade-off between the two metrics. The overall performance of the approach is measured by the Area Under the Curve (AUC) of ROC, with a greater AUC indicating better performance. The ROC curve and AUC can be used to compare the effectiveness of various biometric identification techniques and to establish the ideal operating point for a particular application.

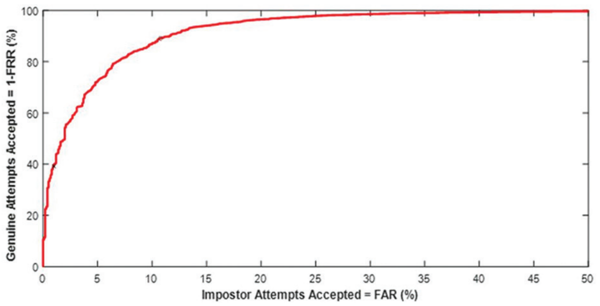


Fig. 7. ROC curve for the proposed identification

An illustration of the correlation between the false reject rate (FRR) and false acceptance rate (FAR) is called a detection error trade-off (DET) curve. The x-axis shows the FRR and the FAR is represented by the y-axis. Fig. 8 illustrates how the DET curve is often shown on a log scale to better easily visualize subtle variations in error rates.

When the cost of a false rejection is greater than that of a false acceptance, the DET curve helps assess the performance of biometric systems. A decision threshold must be selected in these systems to balance the FRR and FAR trade-off. This trade-off is visualized by the DET curve, which also enables the appropriate decision threshold to be chosen based on the intended operating point. A successful system should have a curve that is as close as feasible to the bottom left corner of the graph to indicate low error rates for both erroneous acceptances and false rejections. The outcomes of a FOK and LDN with carefully selected parameters are shown in Fig. 9.

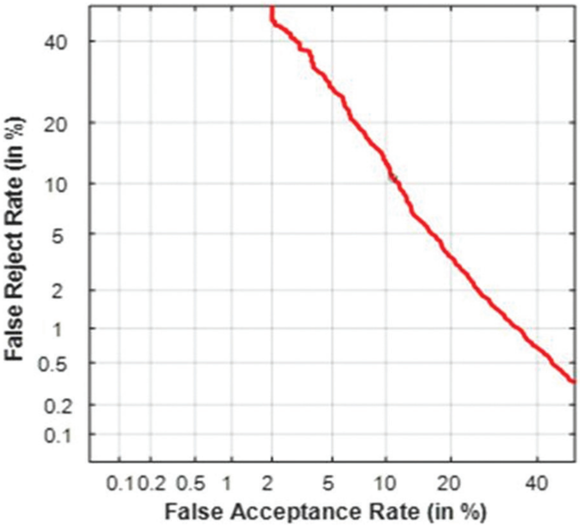


Fig. 8. DET curve for the proposed identification

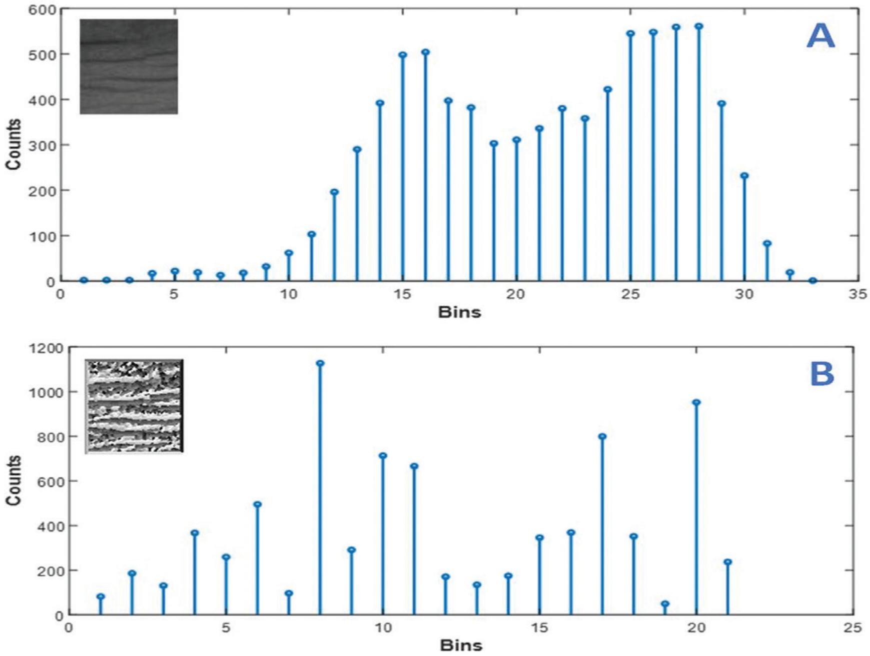


Fig. 9. Results of an FOK and LDN (with best-chosen parameters), A: Histogram counts for an FOK image before the LDN, B: Histogram counts for an FOK image after the LDN

4.4. COMPARISONS

Comparisons are considered with many feature extractions including state-of-the-art methods. Table 4 provides comparisons with various feature extraction methods for the identification based on the FOK. These are the Surrounded Patterns Code (SPC) [1], Enhanced Local Line Binary Pattern (ELLBP) [2, 3], Local Binary Patterns (LBP) [4], Centralized Binary Patterns (CBP) [7], Center-Symmetric Local Binary Pattern (CSLBP) [8] and Local Binary Patterns for FOK (LBP-FOK) [5], respectively.

The table provides the Equal Error Rates (EER) for different feature extraction techniques for personal identification based on the finger's outer knuckle. The EER measures the accuracy of a biometric system, with a lower EER indicating better performance.

Table 4. Comparisons with various feature extraction methods for the identification based on the FOK

Reference	Feature extraction	Parameters	EER (%)
[27]	SPC	---	45.9
[28]	ELLBP	N=17, w1=0.7 and w2=0.3	29.53
[29]	LBP	P=8 and R=1	28.37
[30]	CBP	P=8 and R=3	23.33
[31]	CSLBP	P=8 and R=2	23.26
[32]	LBP-FOK	N=5	14.03
Proposed method	LDN	Mask of type Gaussian and Sigma=0.85	10.78

Among the traditional feature extraction methods, LBP and CLBP have lower EER than the other traditional methods (SPC and ELLBP). Specifically, LBP has an EER of 28.37%, which is slightly better than the EER of CBP (23.33%) and CSLBP (23.26%), but worse than the EER of LBP-FOK (14.03%) and LDN (10.78%).

LBP-FOK and LDN, which are more advanced feature extraction techniques, have significantly lower EER than traditional methods. LDN has the lowest EER (10.78%), followed by LBP-FOK (14.03%).

Overall, LDN performs best among the feature extraction techniques in the table, followed by LBP-FOK, CBP, CSLBP, LBP, ELLBP, and SPC in descending order of performance.

5. CONCLUSION

The LDN method offered a thorough and efficient methodology to evaluate FOK patterns for identification. LDN improves accuracy and robustness in identifying persons based on their FOKs by utilizing the input image, assessing edge responses in several directions and creating a robust encoding technique.

The result showed the proposed method's effectiveness and robustness with EER equal to 10.78%. It can be concluded that using the LDN pattern identification method is possible to identify people based on their FOKs in a trustworthy and accurate manner. This is because it creates a six-bit binary code that distinguishes between variations in texture intensity.

Various future studies can be suggested as exploiting the LDN for the patterns of fingerprint, finger inner knuckle and finger nail. In addition, more work can be carried out for using the LDN with FOKs in terms of verification.

6. REFERENCES

[1] A. Attia, M. Chaa, Z. Akhtar, Y. Chahir, “Finger kunckcle patterns based person recognition via bank of multiscale binarized statistical texture features”, Evolving Systems, Vol. 11, 2020, pp. 625-635.

[2] E. Perumal, S. Ramachandran, “A multimodal biometric system based on palmprint and finger knuckle print recognition methods.”, International Arab Journal of Information Technology, Vol. 12, No. 2, 2015.

[3] E. Rani, R. Shanmugalakshmi, “Finger knuckle print recognition techniques a survey”, International Journal of Engineering Science, Vol. 2, No. 11, 2013, pp. 62-69.

[4] H. M. Ahmed, M. Y. Kashmola, “A proposed architecture for convolutional neural networks to detect skin cancers”, International Journal of Artificial Intelligence, Vol. 11, No. 2, 2022, pp. 1-9.

[5] R. R. O. Al-Nima, M. A. M. Abdullah, M. T. S. Al-Kaltakchi, S. S. Dlay, W. L. Woo, J. A. Chambers, “Finger texture biometric verification exploiting multi-scale sobel angles local binary pattern features and score-based fusion”, Digital Signal Processing, Vol. 70, 2017, pp. 178-189.

[6] H. M. Ahmed, M. Y. Kashmola, “Generating digital images of skin diseases based on deep learning”, Proceedings of the 7th International Conference on Contemporary Information Technology and Mathematics, Mosul, Iraq, 25-26 August 2022, pp. 179-184.

[7] V. Yadav, V. Bharadi, S. K. Yadav, “Texture feature extraction using hybrid wavelet type I & II for finger knuckle prints for multi-algorithmic feature fusion”, Procedia Computer Science, Vol. 79, 2016, pp. 359-366.

[8] V. Yadav, V. Bharadi, S. K. Yadav, “Feature vector extraction based texture feature using hybrid wavelet type I & II for finger knuckle prints for multi-instance feature fusion”, Procedia Computer Science, Vol. 79, 2016, pp. 351-358.

[9] A. M. Aljuboori, M. H. Abed, “Finger knuckle pattern person identification system based on LDP-NPE and machine learning methods”, Bulletin of Electrical Engineering and Informatics, Vol. 11, No. 6, 2022, pp. 3521-3529.

[10] J. Kim, K. Oh, B.-S. Oh, Z. Lin, K.-A. Toh, “A line feature extraction method for finger-knuckle-print verification”, Cognitive Computation, Vol. 11, 2018.

[11] R. Hammouche, A. Attia, S. Akrouf, “A novel system based on phase congruency and gabor-filter bank

for finger knuckle pattern authentication", Journal of Image and Video Processing, Vol. 10, No. 3, 2020, pp. 2125-2131.

[12] R. Ranjbarzadeh, S. Dorosti, S. J. Ghoushchi, S. Safavi, N. Razmjooy, N. T. Sarshar, S. Anari, M. Bendechache, "Nerve optic segmentation in CT images using a deep learning model and a texture descriptor", Complex & Intelligent Systems, Vol. 8, No. 4, 2022, pp. 3543-3557.

[13] Y. Gao, J. Wang, L. Zhang, "Robust ROI localization based on image segmentation and outlier detection in finger vein recognition", Multimedia Tools and Applications, Vol. 79, 2020, pp. 20039-20059.

[14] K. Kapoor, S. Rani, M. Kumar, V. Chopra, G. S. Brar, "Hybrid local phase quantization and grey wolf optimization based SVM for finger vein recognition", Multimedia Tools and Applications, Vol. 80, 2021, pp. 15233-15271.

[15] C. F. G. Dos Santos et al. "Gait Recognition Based on Deep Learning: A Survey", ACM Computing Surveys, Vol. 55, No. 2, 2023, pp. 1-34.

[16] G. Jaswal, R. C. Poonia, "Selection of optimized features for fusion of palm print and finger knuckle-based person authentication", Expert Systems, Vol. 38, No. 1, 2021, p. e12523.

[17] A. Arjiah, W. El-Tarhouni, A. Lawgali, "Finger Knukle Print Recognition based on the Combination of the Multi Shift Local Binary Pattern Descriptor with Discrete Fourier Transform", Proceedings of the IEEE 2nd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering, Sabratha, Libya, 23-25 May 2022.

[18] Vensila C., A. B. Wesley, "Authentication-based multimodal biometric system using exponential water wave optimization algorithm", Multimedia Tools and Applications, Vol. 82, 2023, pp. 30275-30307.

[19] A. K. Gautam, R. Kapoor, "A review on Finger vein based Recognition", Proceedings of the IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering, Dehradun, India, 11-13 November 2021, pp. 1-6.

[20] W. Jia et al. "A survey on dorsal hand vein biometrics", Pattern Recognition, Vol. 120, 2021, p. 108122.

[21] J. O. H. Engineering, "Retracted: Biometric Recognition of Finger Knuckle Print Based on the Fusion of Global Features and Local Features", Journal of Healthcare Engineering, Vol. 2022, 2022, p. 9820927.

[22] W. Liet et al. "Biometric recognition of finger knuckle print based on the fusion of global features and local features", Journal of Healthcare Engineering, Vol. 2022, 2022.

[23] R. Kapoor et al. "Completely Contactless Finger-Knuckle Recognition using Gabor Initialized Siamese Network", Proceedings of the International Conference on Electronics and Sustainable Communication Systems, Coimbatore, India, 2-4 July 2020, pp. 867-872.

[24] A. R. Rivera, J. R. Castillo, O. O. Chae, "Local directional number pattern for face analysis: Face and expression recognition", IEEE Transactions on Image Processing, Vol. 22, No. 5, 2013.

[25] C. Kant, S. Chaudhary, "A multimodal biometric system based on finger knuckle print, fingerprint, and palmprint traits", Proceedings of ICICV Innovations in Computational Intelligence and Computer Vision, 2021, pp. 182-192.

[26] IIT Delhi Finger Knuckle Database version 1.0, http://www.comp.polyu.edu.hk/~csajaykr/knuc kle/iitd_knuckle.htm (accessed: 2023)

[27] R. R. O. Al-Nima, M. Al-Kaltakchi, S. Al-Sumaidaee, S. Dlay, W. Woo, T. Han, J. Chambers, "Personal verification based on multi-spectral finger texture lighting images", IET Signal Processing, Vol. 12, Issue. 9, 2018.

[28] R. R. O. Al-Nima, "Signal Processing and Machine Learning Techniques for Human Verification Based on Finger Textures", School of Engineering, Newcastle University, UK, 2017, PhD thesis.

[29] T. Ojala, M. Pietikäinen, D. Harwood, "A comparative study of texture measures with classification based on featured distributions", Pattern Recognition, Vol. 29, No. 1, 1996.

[30] X. Fu, W. Wei, "Centralized binary patterns embedded with image euclidean distance for facial expression recognition", Proceedings of the Fourth International Conference on Natural Computation, Jinan, China, 18-20 October 2008.

[31] M. Heikkilä, M. Pietikäinen, C. Schmid, "Description of interest regions with center-symmetric local binary patterns", Computer vision, graphics and image processing, Springer, Berlin, Heidelberg, 2006.

[32] R. R. O. Al-Nima, M. K. Jarjes, A. W. Kasim, S. S. M. Sheet, "Human Identification using Local Binary Patterns for Finger Outer Knuckle", Proceedings of the IEEE 8th Conference on Systems, Process, and Control, Melaka, Malaysia, 11-12 December 2020, pp. 7-12.

Transformer Faults Classification Based on Convolution Neural Network

Original Scientific Paper

Maha A. Elmohallawy
Department of electrical Engineering, Zagazig
Higher Institute of Engineering and Technology,
Zagazig, Egypt
ounelmaha35@gmail.com

Amir Yassin Hassan
Department of Power Electronics
and Energy Conversion,
Electronics Research Institute, Cairo, Egypt
amir@eri.sci.eg

Amal F. Abdel-Gawad
Faculty of computer and informatics,
Zagazig University, Zagazig, Egypt.
amgawad2001@yahoo.com

Sameh I. Selem
Electrical Power & Machines Department,
Faculty of Engineering, Zagazig University,
Zagazig, Egypt
sameh.eb@gmail.com

Abstract *– This paper studies the latest advances made in Deep Learning (DL) methods utilized for transformer inrush and fault currents classification. Inrush and fault currents at different operating conditions, initial flux and fault type are simulated. This paper presents a technique for the classification of power transformer faults which is based on a DL method called convolutional neural network (CNN) and compares it with traditional artificial neural network (ANN) and other techniques. The inrush and fault current signals of the transformer are simulated within MATLAB by using Fourier analyzers that provides the 2nd harmonic signal. The 2nd harmonic peak and variance statistic values of input signals of the three phases of transformer are used at different operating conditions. The resulted values are aggregated into a dataset to be used as an input for the CNN model, then training and testing the CNN model is performed. Consequently, it is obvious that the CNN algorithm achieves a better performance compared to other algorithms. This study helps with easy discrimination between normal signals and faulty signals and to determine the type of the fault to clear it easily.*

Keywords*: Machine learning, Transformer, inrush, fault classification, Artificial intelligence, Deep learning, CNN algorithm*

1. INTRODUCTION

The difference between normal signals and faulty signals must be distinguished even when disturbances occur and protective devices should deal with faulty signals to keep continuity of supply [1]. Numerous faults in power systems are unavoidable due to the complex circumstances and a variety of human or natural factors. For more effective power supply restoration and fault cause analysis, fault categorization is crucial [2]. Preventing a costly outage of electrical network system requires efficient fault diagnosis [3].

Artificial intelligent (AI) proved effectiveness in solving many vital challenges [4]. There are different types of faults such as asymmetrical faults (line to ground, line to line, and two- lines to ground) and symmetrical fault. The fault classification by utilizing AI algorithms have received much attention in recent years. However, most of work has been focused on the fault classification problem in power systems [5]. Power transformer is a vital element in power grid. Its failure may affect the continuity of supply of electrical energy to the consumers [6].

Transformers' inrush current can be significant, ranging from five to seven times the rated current [7]. Nowadays, with the development and spread of DL usage, smart grid faults diagnosis based on DL should be considered [8].

Machine learning (ML) techniques have been widely used for power systems faced challenges and achieve good results. ML has been used in solving nonlinear problems (detection, classification, recognition, etc.) [9].

Rao et al. uses ML algorithms in transformer dissolved gas analysis [10]. For the purpose of diagnosing faults in oil-immersed power transformers, a bi-level ML technique with a multi-classification model and a binary imbalanced classification model is suggested; study is made to explain that the inrush current is rich with 2nd harmonic content [11].

For power transformers, differential relays are blocked by using the 2nd harmonic component, and for many researchers, this subject meets a great concern. Therefore, detecting the 2nd harmonic component and fault current wave forms is significant [12-16].

Many researchers had an interest to recognize the current signals using the 2nd harmonic component as transformers' inrush current waveforms includes 62% of 2nd harmonics and 55% of the DC component [17]. The two-instantaneous-value-product algorithm has been used for recognizing fault and inrush currents by extracting the current amplitude variations [18]. Krstivojevic and Milenko presents an algorithm that prevents false tripping of the restricted earth fault relay during the transformer energization [19].

For power transformers, based on adaptive neuro-fuzzy inference systems and discrete wavelet transform, Salama, et al. presented a hybrid algorithm for simulating the faults [20].

Many researchers use MATLAB simulation in modeling and classification. Different types of ANN and their applications are used in solving power systems challenges. ANN used as a classifier by using back propagation method for discrimination between inrush current and the fault current [21].

Both Radial Basis Neural Networks and Back Propagation Neural Networks are frequently employed. The multilayer perceptron, which has at least three layers (input layer, output layer, and hidden layer), is the most common architecture of this computing paradigm [22].

A convolutional neural network (CNN) is a specific category within machine learning. It belongs to a range of ANNs that are utilized for diverse purposes and data formats. CNNs are a type of network structure designed for DL algorithms and are particularly employed for tasks involving pixel data processing, such as recognition tasks [23]. There are further categories of ANNs in DL, but for objects recognition and identification. CNN is the most widely used type of ANNs specialized in classification. The main characteristic of CNN makes it better than standard ANNs for recognition [23]. CNN is an effective tool for recognizing multi-spectrograms that are structured into numerical data for diagnosing faults, eliminating the requirement of selecting the vibration axis beforehand [24]. The proposed ML method uses a CNN framework that performs discrimination between inrush and faulty currents.

Fault detection refers to the requirement of having knowledge about the system's health limited to two possible conditions (normal or abnormal). The normal state indicates that the system is functioning correctly without any worrisome indications. On the other hand, the abnormal state signifies that certain system symptoms fall outside the range of what is considered normal. The system being developed must be capable of identifying and distinguishing between these two states [25].

To prevent undesired tripping due to magnetizing inrush current, a novel approach is introduced for distinguishing internal fault current from inrush current. Transformer inrush currents can reach significant magnitudes, often ranging from five to seven times the rated current of the transformer. The second harmonic component is employed to inhibit the activation of differential relays in power transformers. False triggering of protection systems during inrush situations remains a prominent issue associated with transformer inrush currents.

The main objective of this study is to identify the inrush current and the type of fault based on two methods; variance statistical inference on three phase transformer signal and Fourier analyzers used to analyze the input signal and provide us with second harmonic signal. The peak value of 2nd harmonic input signals of the three phases of transformer. The two methods are used at different operating conditions to train the network.

In this research work, an efficient ML algorithm - which is CNN - is learned to determine the faults conditions and their type. A study is made to explain that the inrush current is rich with 2nd harmonic content and Fourier analyzers are used to analyze the input signal of three phase transformer. The following sections of this research include the ML and Fourier analysis, preparing the dataset of normal and faulty current signals, training the CNN, results and testing of network, comparison with other algorithms and the conclusion.

2. MACHINE LEARNING AND FOURIER ANALYSIS

In this research work, an efficient ML algorithm is learned to determine the faults conditions and their type, this is CNN. CNN includes the pooling, dropout, and fully connected (FC) layers. The phases of the applied ML technique are preparing the dataset of normal and faulty current signals, building the CNN model, splitting the data into train and test, training and testing the model, evaluation, and changing the parameters to enhance the performance [26].

The input layer of CNN is numerical data of current includes the three phases current signals (Red, Yellow, and Blue) each represented by 1041 samples data for variance signal value with a matrix (1041*3) and 1118 samples data for second harmonic signal value with a matrix (1118*3). Some of these data is utilized in the CNN model training and the rest is utilized for testing the proposed model. These parameters taken under different operation condition to train CNN giving different type of current signals (inrush current or different type of faulty current) as the target of CNN that is shown in Table 1.

Table1. The target of CNN

1	Inrush
2	F(A-B)
3	F(A-B-C)
4	F(A-C)
5	F(A-G)
6	F(B-C)
7	F(B-G)
8	F(C-G)
9	F(A-B-G)
10	F(A-C-G)
11	F(B-C-G)

In this study, Fourier analyzers present the 2nd, 3rd and 5th harmonic contents of the transformer input current signals for the current signal model of (normal-inrush-faulty). Peak value of the 2nd harmonic inrush and fault current signals are recorded and some numerical samples of them are selected to be used as an input to the CNN algorithm in order to train it to be ready for the needed fault classification process. Variance statistic values of input signals of the three phases of transformer are used at different operating conditions.

The 2nd, 3rd, and 5th harmonic contents of faulty model are shown in figures 1, 2 and 3 respectively. Table 1 illustrates various harmonic contents.

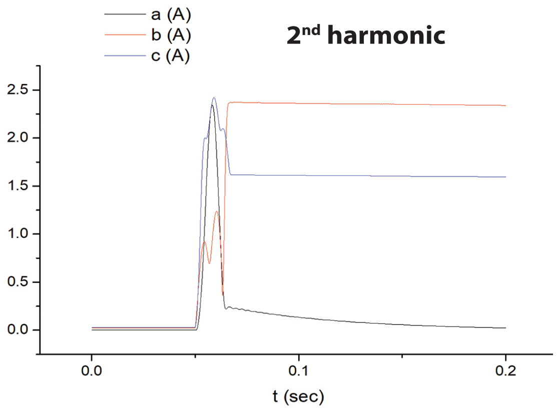


Fig. 1. 2nd harmonic of current signal

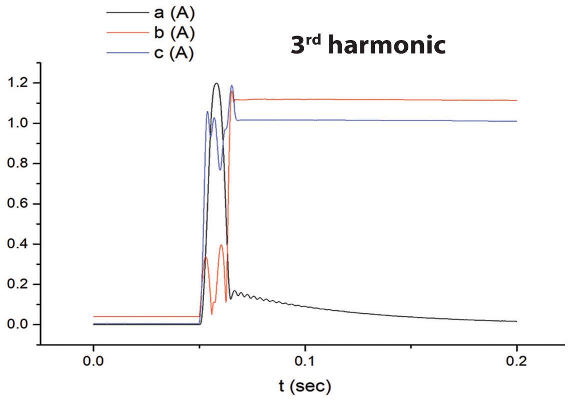


Fig. 2. 3rd harmonic of current signal

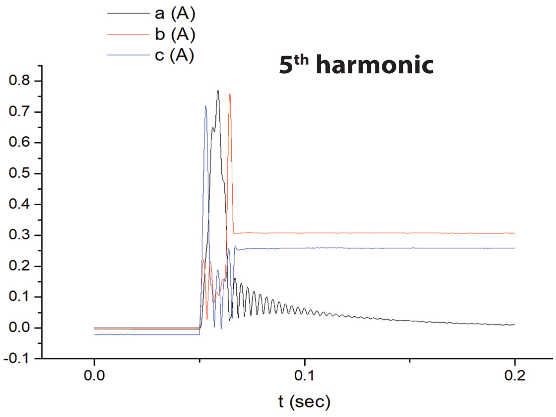


Fig. 3. 5th harmonic of current signal

The data in table 2 confirms that the 2nd harmonic is the dominating harmonic during transformer energization as the greatest value of three different kinds of current is the 2nd harmonic content.

Table 2. Harmonic spectrum

Current type	2nd(A)	3rd(A)	5th(A)
Normal	0.6	0.15	0.1
Inrush	2.4	1.2	0.75
Faulty	(1.5-2.6)	(1-1.5)	(0.25-0.3)

3. DATA SET PREPARATION AND CNN TRAINING

Among the different types of neural networks (others include recurrent neural networks (RNN), long short-term memory (LSTM), artificial neural networks (ANN), etc.), CNNs are easily the most popular. These convolutional neural network models are ubiquitous in the image data space. They work phenomenally well on computer vision tasks like image classification, object detection, image recognition [27].

Input data is processed using sklearn pre-processing function called MinMaxScaler with feature_range = (0, 1). CNN algorithm is used to improve the accuracy of classification. Parameters of this method are adjusted to improve performance where 50 epochs is used with a batch size = 1. a dams optimizer is used and achieves higher performance in most of DL methods. The function "Get dummies" from "pandas" library is used to convert categorical variable of output data into dummy/indicator variables. A sequential CNN model with two dense layers is used with 'relu' activation function for the first input layer and 'softmax' activation function for the second output layer.

A variety of samples of various operation conditions have been chosen and the 2nd harmonic is recorded by Fourier analysis for the three phase current signals. Tables 3 and 4 shows the maximum 2nd harmonic current and variance statistic values respectively for some of these samples under different operating conditions and various current conditions to be input for CNN.

Fig. 4 shows the CNN architecture where it contains an input layer with activation function 'relu', four blocks of hidden layers (convolution / pooling) and FC layers (flatten/dense).

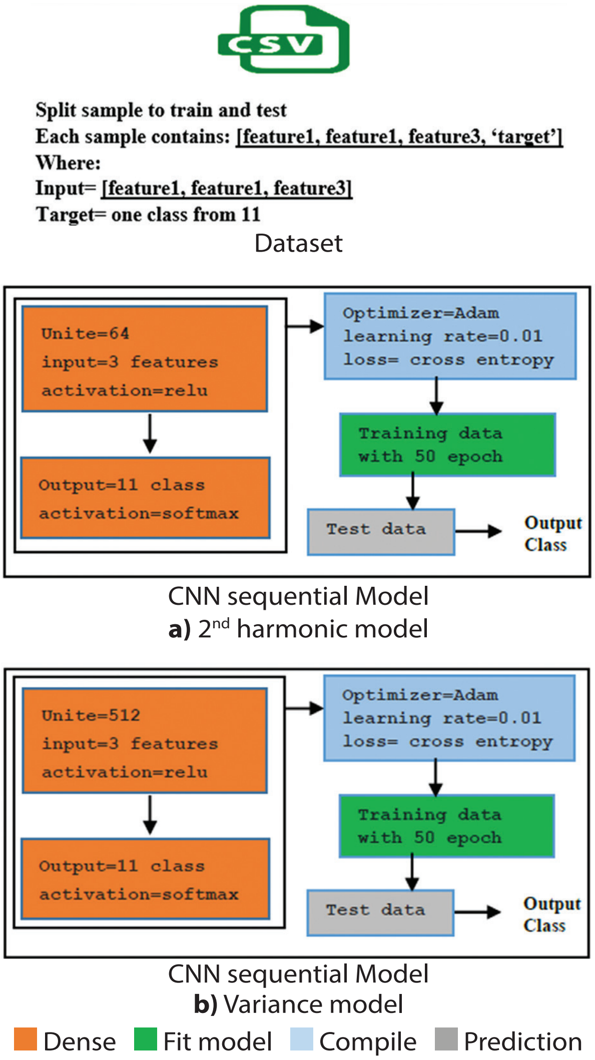
Table 3. Maximum value of 2nd harmonic current

Conditions of operating			Maximum value of 2nd harmonic		
Flux value	Connection of winding	Signal type	I/Ph/ Red	I/Ph/ Yellow	I/Ph/Blue
0.4, -0.2, 0.2	Yg-Yg	Faulty (A-B-C)	2.2	0.0	2.0
0.4, -0.2, 0.2	Yg-Yg	Inrush	0.730	0.010	0.001
0.4, -0.2, 0.2	Y-Y	Faulty (A-B-C)	2.5	2.0	2.0
0.4, -0.2, 0.2	Y-D	Inrush	0.4	0.2	0.3
0.4, -0.2, 0.2	Y-Y	Inrush	0.13	0.13	0.12
0.4, -0.2, 0.2	Y-Y	Faulty (B-G)	0.13	0.12	0.13

0.4, -0.2, 0.2	Y-Y	Faulty (A-B)	2.00	2.00	0.01
0.4, -0.2, 0.2	D-D	Inrush	0.360	0.345	0.054
0.4, -0.2, 0.2	D-D	Faulty (C-G)	0.40	0.35	0.06

Table 4. Variance value signals

Conditions of operating			variance value		
Flux value	Connection of winding	Signal type	I/Ph/ Red	I/Ph/ Yellow	I/Ph/ Blue
0.4, -0.2, 0.2	Yg-Yg	Inrush	0.00	0.00	0.08
0.2, 0, 0	Y-Y	Inrush	0.003	0.003	0.007
0.4, -0.2, 0.2	Yg-Yg	Faulty (A-B)	0.001	8.03	8.72
0.4, -0.2, 0.2	D-D	Faulty (A-B-C)	11.8	13.2	14.4
0.4, -0.2, 0.2	Y-Y	Faulty (A-C)	8.51	0.01	8.49
0.3, 0, 0	Y-Y	Inrush	0.0017	0.0015	0.0035
0.4, -0.2, 0.2	Yg-Yg	Faulty (B-G)	0.0003	11.005	1.100
0.4, -0.2, 0.2	D-D	Faulty (C-G)	0.01	0.10	0.10
0.6, -0.3, 0.3	D-D	Inrush	0.0002	0.0026	0.0038
0.4, -0.2, 0.2	Y-Y	Faulty (B-C-G)	6.63	6.48	0.21



Split sample to train and test
Each sample contains: [feature1, feature1, feature3, 'target']
Where:
Input= [feature1, feature1, feature3]
Target= one class from 11

Dataset

CNN sequential Model
a) 2nd harmonic model

CNN sequential Model
b) Variance model

🟧 Dense 🟩 Fit model 🟦 Compile ⬜ Prediction

Fig. 4. Convolutional neural network architecture

The primary metric for comparing classifiers was the F1-score. F1-score, recall and Precision are computed as shown in the following equations 1, 2 and 3 [24]. Tables 5 and 6 show the parameters of the variance and the 2nd harmonic sequential models respectively.

$$\text{Precision} = \text{Truepositive}/(\text{TruePositive}+\text{FalsePositive}) \quad (1)$$

$$\text{Recall} = \text{Truepositive}/(\text{TruePositive}+\text{FalseNegative}) \quad (2)$$

$$\text{F1\_ score} = (2*\text{Recall}*\text{Precision})/(\text{Recall}+\text{Precision}) \quad (3)$$

Table 5. Variance Sequential Model

Layer (type)	Output Shape	Parameters
dense (Dense)	(None, 512)	2048
dense_1 (Dense)	(None, 11)	5643

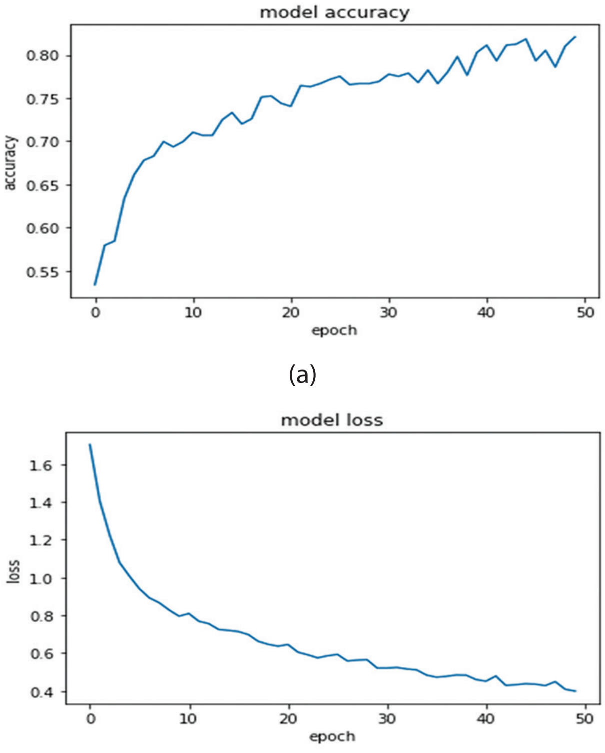
Table 6. Second harmonic Sequential Model

Layer (type)	Output Shape	Parameters
dense (Dense)	(None, 64)	256
dense_1 (Dense)	(None, 11)	715

4. RESULTS AND TESTING OF CNN

Different percentages for training and testing are applied and the best results are with training by 60% and testing by 40% of the data for 2nd - harmonic model and with training by 80% and testing by 20% of the data for variance model.

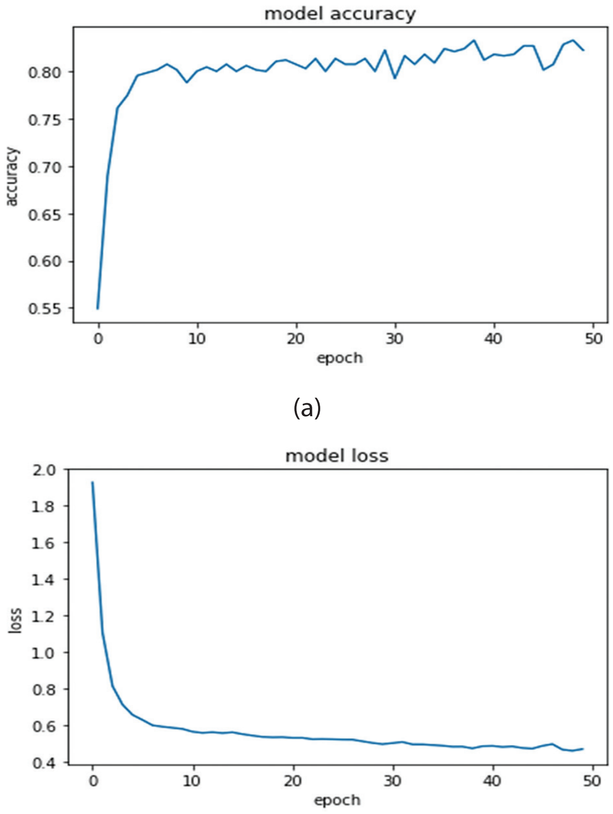
Figs. 5 and 6 shows the accuracy and the loss of CNN model for 50 epochs when training with both variance and 2nd harmonic numerical values.



(a)

(b)

Fig. 5. Performance of training the CNN model with variance values. (**a**) accuracy and (**b**) Loss



(a)

(b)

Fig. 6. Performance of training the CNN model with second harmonic values. (**a**) accuracy and (**b**) Loss

Two dense layers are applied with specific total parameters equals 7,691 for variance values and equals 971 for 2nd harmonic values. Different optimizers are applied; the best optimizer is a dams. Tables 7 and 8 present the test results of the CNN model with both variance and 2nd harmonic values.

Table 7. Test results of the Sequential Model with variance values

Target	Precision	Recall	f1-score
1	0.96	1.00	0.98
2	0.53	1.00	0.70
3	0.78	1.00	0.88
4	1.00	0.33	0.50
5	1.00	0.73	0.84
6	0.00	0.00	0.00
7	0.86	1.00	0.92
8	0.80	0.44	0.57
9	1.00	0.42	0.59
10	0.67	0.80	0.73
11	0.56	1.00	0.71

The results in Table 7 shows that the recall classifier is better in target classes 1, 2, 3, 7, 10 and 11, while the precision classifier gives better results in target classes 4, 5, 8 and 9.

Table 8. Test results of the Sequential Model with 2nd harmonic values

Target	Precision	Recall	f1-score
1	0.87	1.00	0.93
2	0.82	1.00	0.90
3	1.00	1.00	1.00
4	0.95	1.00	0.74
5	1.00	0.25	0.40
6	0.95	1.00	0.74
7	1.00	0.48	0.65
8	1.00	0.18	0.31
9	0.00	0.00	0.00
10	0.00	0.00	0.00
11	0.00	0.00	0.00

The results in Table 8 shows that the recall classifier is better in target classes 1, 2, 3, 4 and 6, while the precision classifier gives better results in target classes 4, 5, 6,7 and 8. The test results show an accuracy of 83% when using 2nd harmonic signals and 86% with variance signals. The result shows that the sequential CNN model achieves good performance.

5. COMPARISON WITH OTHER ALGORITHMS

In order to rate the proposed CNN model, it is compared with other algorithms that are used for transformer inrush and fault currents classification. Table 9 presents a comparison between the proposed CNN algorithm performance and the performance of other models. The comparison shows that the proposed CNN model achieves a higher accuracy in fault classification than other compared models.

Table 9. Comparison of the proposed model performance with other research works

Model	Proposed CNN model	ANN [29]	Based language ML models [28]	DLNN with auto-encoders (SAE) [30]	Stacked sparse auto encoder DL [31]
Acc.	86%	80.4%	72.3%	71.3%	79.94%

6. CONCLUSION

In this paper, CNN is used to classify transformer faults. Matlab-Simulink is used to simulate the faults at different operating conditions. The current harmonic contents are extracted by using Fourier analysis and become clear that the 2nd harmonic content is the predominant. The accuracy of the CNN model is improved by training with numerical data of variance and 2nd harmonics values. The proposed CNN model achieves an accuracy of 83% when learned with 2nd harmonic values and 86% with variance values. It is obvious that the variance data set yields better performance. A comparison with other techniques is performed and the CNN model presents a higher improved accuracy by about 5.6% more than using ANN.

7. 7. REFERENCES:

[1] P. Arboleya, G. Diaz, J. G. Aleixandre, “A solution to the dilemma inrush/fault in transformer relaying using MRA and wavelets”, Electric Power Components and Systems, Vol. 34, No. 3, 2006, pp. 285 – 301.

[2] M.-F. Guo, N.-C. Yang, W.-F. Chen, "Deep-learning-based fault classification using Hilbert–Huang transform and convolutional neural network in power distribution systems", IEEE Sensors Journal, Vol. 19, No. 16, 2019, pp. 6905-6913.

[3] Y. Ran et al. "A survey of predictive maintenance: Systems, purposes and approaches", arXiv:1912.07383, 2019.

[4] A. Yassin, M. Badr, S. Wahsh, “dSP ACE DS 1202 Based Real Time Implementation of Cuckoo Search Optimized FDTC of PMSM", Proceedings of the 20th MEPCON Conference, Egypt, December 2018, pp. 1126-1133.

[5] T. Rajić, Z. Stojanović. "An algorithm for longitudinal differential protection of transmission lines", International Journal of Electrical Power & Energy Systems, Vol. 94, 2018, pp. 276-286.

[6] A. Zitouni, "Power Transformer Differential Relay Reliability Assessment Using False Trip Root Cause Analysis”, Proceedings of the International Conference on Electrical Engineering, Istanbul, Turkey, 25-27 September 2020.

[7] S. Hodder et al. "Low second-harmonic content in transformer inrush currents-Analysis and practical solutions for protection security”, Proceedings of the 67th Annual Conference for Protective Relay Engineers, College Station, TX, USA, 2014.

[8] S. Barrios et al. "Partial discharge classification using deep learning methods—Survey of recent progress”, Energies, Vol. 12, No. 13, 2019, p. 2485.

[9] A. K. Ozcanli, F. Yaprakdal, M. Baysal, "Deep learning methods and applications for electrical power systems: A comprehensive review”, International Journal of Energy Research, Vol. 44, No. 9, 2020, pp. 7136-7157.

[10] U. M. Rao et al. "Identification and application of machine learning algorithms for transformer dissolved gas analysis”, IEEE Transactions on Dielectrics and Electrical Insulation, Vol. 28, No. 5, 2021, pp. 1828-1835.

[11] D. Zhang et al. "A bi-level machine learning method for fault diagnosis of oil-immersed transformers with feature explainability”, International Journal of Electrical Power & Energy Systems, Vol. 134, 2022, p. 107356.

[12] R. P., F. B. C. Medeiros, K. M. Silva, "Power transformer differential protection using the boundary discrete wavelet transform”, IEEE Transactions on Power Delivery, Vol. 31, No. 5, 2015, pp. 2083-2095.

[13] K. Behrendt, N. Fischer, C. Labuschagne, "Considerations for using harmonic blocking and harmonic restraint techniques on transformer differential relays”, SEL Journal of Reliable Power, Vo. 2, No. 3, 2011.

[14] S. Krishnamurthy, B. E. Baningobera, “IEC61850 standard-based harmonic blocking scheme for power transformers”, Protection and Control of Modern Power Systems, Vol. 4, No. 1, 2019, p. 10.

[15] Z. Moravej, D. N. Vishwakarma “ANN-based harmonic restraint differential protection of power transformer”, Journal Institution of Engineers India: Part Electrical Engineering Division, 2003, pp. 1-6.

[16] M. Thompson, J. R. Closson “sing IOP characteristics to troubleshoot transformer differential relay misoperation”, Proceedings of the International Electrical Testing Association Technical Conference, Kansas City, Missouri, USA, 2001.

[17] P. C. Ling, A. Basak, “Investigation of magnetizing inrush current in a single-phase transformer”, IEEE Transactions on Magnetics, Vol. 24, No. 6, 1988, pp. 3217-22.

[18] J. Ma et al. "A new algorithm to discriminate internal fault current and inrush current utilizing feature of fundamental current”, Canadian Journal of Electrical and Computer Engineering, Vol. 36, No. 1, 2013, pp. 26-31.

[19] J. Krstivojevic, M. Djuric, "A new algorithm for avoiding maloperation of transformer restricted earth fault protection caused by the transformer magnetizing inrush current and current trans-

former saturation", Turkish Journal of Electrical Engineering and Computer Sciences, Vol. 24, No. 6, 2016, pp. 5025-5042.

[20] A. M. Salama et al. "A new hybrid protection algorithm for protection of power transformer based on discrete wavelet transform and ANFIS inference systems", International Journal of Emerging Electric Power Systems, Vol. 19, No. 3, 2018.

[21] W. Zhang et al. "Application of deep learning algorithms in geotechnical engineering: a short critical review", Artificial Intelligence Review, Vol. 54, No. 8, 2021, pp. 5633-5673.

[22] M. Stanbury, Z. Djekic, "The Impact of Current Transformer Saturation on Transformer Differential Protection", IEEE Transactions on Power Delivery, Vol. 30, No. 3, 2015, pp. 1278-1287.

[23] L. G. Nachtigall, R. M. Araujo, G. R. Nachtigall, "Classification of apple tree disorders using convolutional neural networks", Proceedings of the 28th International Conference on Tools with Artificial Intelligence, San Jose, CA, USA, 6-8 November 2016.

[24] D. Łuczak, S. Brock, K. Siembab, "Cloud Based Fault Diagnosis by Convolutional Neural Network as Time–Frequency RGB Image Recognition of Industrial Machine Vibration with Internet of Things Connectivity", Sensors, Vol. 23, 2023, p. 3755.

[25] D. Łuczak, S. Brock, K. Siembab, "Fault Detection and Localisation of a Three-Phase Inverter with Permanent Magnet Synchronous Motor Load Using a Convolutional Neural Network", Actuators, Vol. 12, 2023, p. 125.

[26] B. R. G. Elshoky et al. "Comparing automated and non-automated machine learning for autism spectrum disorders classification using facial images", ETRI Journal, Vol. 44, No. 4, 2022.

[27] P. Dhruv, S. Naskar, "Image classification using convolutional neural network (CNN) and recurrent neural network (RNN): a review", Proceedings of Machine Learning and Information Processing: 2019, pp. 367-381.

[28] M. A. Al-Garadi et al. "Text classification models for the automatic detection of nonmedical prescription medication use from social media", BMC medical informatics and decision making, Vol. 21, No. 1, 2021, pp. 1-13.

[29] M. A. Goda et al. "Discrimination between inrush and fault currents of transformers using Artificial Neural Network Tools", The Egyptian International Journal of Engineering Sciences and Technology, Vol. 37, No. 2, 2022, pp. 34-38.

[30] W. Yixing M. Liu, Z. Bao, "Deep learning neural network for power system fault diagnosis", Proceedings of the 35th Chinese Control Conference, Chengdu, China, 27-29 July 2016, pp. 6678–6683.

[31] Z. Zhang, S. Li, Y. Xiao, Y. Yang, "Intelligent simultaneous fault diagnosis for solid oxide fuel cell system based on deep learning", Applied Energy, Vol. 233-234, 2019, pp. 930-942.

Experimental Procedure for Determining the Remanent Magnetic Flux Value Using the Nominal AC Energization

Original Scientific Paper

Dragan Vulin
J. J. Strossmayer University of Osijek,
Faculty of Electrical Engineering, Computer Science and Information Technology Osijek
Kneza Trpimira 2 B, Osijek, Croatia
dragan.vulin@ferit.hr

Denis Pelin
J. J. Strossmayer University of Osijek,
Faculty of Electrical Engineering, Computer Science and Information Technology Osijek
Kneza Trpimira 2 B, Osijek, Croatia
denis.pelin@ferit.hr

Mario Franjković
J. J. Strossmayer University of Osijek,
Faculty of Electrical Engineering, Computer Science and Information Technology Osijek
Kneza Trpimira 2 B, Osijek, Croatia
mario.franjkovic@student.ferit.hr

Abstract *– The laboratory setup and corresponding experimental procedure for determining the remanent magnetic flux in the magnetic core of a single-phase transformer are presented in this paper. Using the proposed method, the remanent flux can be determined without prior knowledge of any parameter or past states of the transformer which is a significant advantage compared to previously known methods. Furthermore, reliable information about the remanent flux could be obtained using less equipment than other methods. Only electrical measurements are needed, without any physical intervention in the core or some other parts of the transformer. However, the major drawback is that some new unknown value of the remanent flux is set after the measuring procedure. Various initial conditions of the remanent flux and the closing voltage angle are set before each energization of the transformer to prove the validity of the proposed method, which can be used to obtain some characteristics of the remanent flux, such as stability over time or its dependence on some external factors.*

Keywords*: inductance, magnetic cores, magnetic flux, transformers*

1. INTRODUCTION

A magnetic core will contain a certain amount of the remanent magnetic flux (Φ_R), also known as residual flux, remanent magnetization or remanence, after the de-energization. An example of a ferromagnetic material's major magnetic hysteresis loop is shown in φ-i characteristics (Fig. 1).

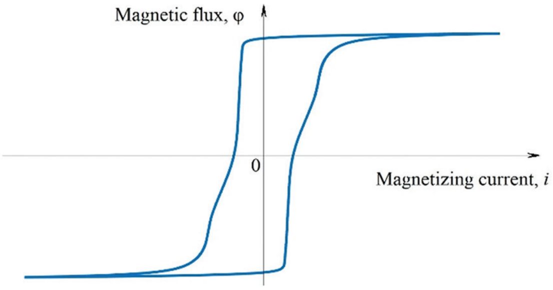


Fig. 1. An example of the magnetic hysteresis loop of a ferromagnetic material

The value of the remanent flux is essential in several areas in practice. One refers to reducing a coil or transformer inrush current by controlled switching [1-4]. Another application area where the remanent flux has an important impact is avoiding current transformer saturation [5-7]. Also, the remanent flux is important as one of the initial conditions in the ferroresonant circuit [8]. In almost all previously mentioned application areas, a magnetic core forms a closed loop, so the remanent flux is closed within the core itself and cannot be measured directly without physical intervention in the core.

However, some methods indirectly determine the remanent flux. The most widely used method is the determination of the remanent flux when de-energizing a coil or transformer by measuring the transformer terminal voltage during the de-energization [8, 9]. The basic idea is to determine the magnetic flux at the instant of de-energization by integrating the port-voltage.

This method is usually used to reduce the inrush current by controlled switching. It is quite simple, but it is unusable if the terminal voltage was not previously measured. Furthermore, the remanent flux can change its value while the coil is not energized if system transients occur [10], or even if there are no external impacts, due to the phenomenon called magnetic viscosity [11]. In that case, the method is unreliable.

Determining the remanent flux can also be done by measuring the leakage flux – the flux near the core is measured and then from the obtained results the remanent flux in the core is estimated [12-14]. Unlike the previous method, all possible changes of the remanent flux after the de-energization are taken into account. Disadvantages are its inaccuracy and high implementation costs – it is challenging to install a magnetic field sensor inside the power transformer tank in an aggressive environment and high temperature. On the other hand, installing the sensor outside the tank will not yield satisfactory results.

The remanent flux can also be determined using the low-voltage DC source for energization [15]. However, an additional DC voltage source is needed to determine the remanent flux utilising this method, while using the proposed method only the nominal voltage source available on-site is required. Furthermore, applying this method, the major hysteresis loop of the observed transformer needs to be obtained before determining the remanent flux, which is not the case in the proposed method. This method can be used in the same application areas as the proposed method.

The remanent flux could also be determined using the inductance value of the winding of a transformer [16]. As in [15], the transformer should be tested before using this method, and the correlation between the remanent flux and the inductance must be established. The conclusion is that the inductance will decrease if the remanent flux is high. However, this method is quite inaccurate – if the remanent flux value changes from zero to the maximum value, the inductance will change only 5% of its value.

Furthermore, the remanent flux could be determined using a minor hysteresis loop without any data regarding the last de-energization [17]. However, the transformer should be tested before using this method, and the relation between the remanent flux and the parameter WQ has to be established. But, it could be used for reducing the inrush current by controlled switching because the determined remanent flux will be preserved after the measurement process.

Finally, there is the method for determining the remanent flux which uses the nonlinear magnetizing characteristics of the core [18]. The basic idea is to energize a coil or transformer with the low voltage DC source and analyze the transient current. Prior to application of the method, the observed transformer should be tested and transient currents should be obtained for all possible remanent flux values. However, after applying this method, some unknown value of the remanent flux after the determination will be established.

There are also demagnetization and prefluxing techniques which actually do not determine the remanent flux, but set it to zero and the maximum value, respectively [19-21]. The basic idea of prefluxing is to set the remanent flux to the maximum positive or negative value prior to operation [22-25]. After demagnetization or prefluxing, the optimal switching angle for reducing the inrush current can easily be calculated. The devices used for demagnetization and prefluxing are simple in construction and operate using low DC voltage.

There is no adequate method for determining the remanent flux which can obtain some features, such as stability over time or how it is influenced by some external factors, except the method shown in [15]. Most of the previously mentioned methods cannot give satisfactory results regarding the precise and reliable value of the remanent flux in a closed magnetic core.

The laboratory setup and corresponding method for determining the remanent flux value will be presented in this paper. Although this paper will discuss only the remanent flux in the transformer core, the proposed method could also be appropriately applied to the iron core coil.

In most cases in practice, the possibility of setting the remanent flux to any value is not used for mitigating inrush current, due to the fact that setting magnetic flux value requests additional devices [20-24]. Thus, the methods most often used rely on integrating the measured port-voltage during the de-energization and assuming that the determined remanent flux will not change until the subsequent energization [2, 3, 9, 26]. The proposed method could be used to check this assumption, where one could consider the time-dependence of the remanent flux. In some cases, the time interval between de-energization and the next energization of the power transformer or coil (e.g. used for compensation of reactive power) could be a couple of months. Furthermore, it is proven that the remanent flux in the core can be changed even if the transformer is not energized [10]. Thus, using the proposed method, it can be investigated how system transients and external magnetic fields affect the remanent flux.

Finally, the remanent flux can be determined without any data about the past states. Furthermore, the parameters of the transformer, except the nominal voltage, should not be known. Only the voltage and current measurements should be conducted, meaning less equipment is required than in the other methods.

2. PROPOSED METHOD

The simple model of the winding of an unloaded transformer can be used as shown in Fig. 2 inside the dashed rectangle.

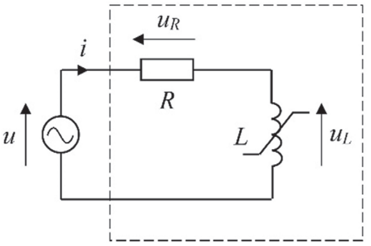


Fig. 2. The model of the winding of an unloaded transformer

The model consists of a linear resistance R in a series with a nonlinear inductance L with φ-i characteristics (Fig. 1) experimentally obtained. The resistance R represents the winding resistance. The model shown in Fig. 2 uses the φ-i characteristics of an unloaded transformer which including nonlinear characteristics with hysteresis. However, only the major magnetic hysteresis loop which is obtained for particular excitation (AC voltage, RMS 39 V, 50 Hz) is defined. The operating point or trajectory could be anywhere inside the loop for some arbitrary excitation. This makes the simulation model appropriate only for the steady state established for the previously mentioned excitation. However, there are some other models more appropriate for simulation, such as the lumped-circuit model by Chua and Stromsmoe [27], the Preisach model [28] and the Jiles-Atherton model [29]. The Chua mathematical model could not be used for the remanent flux simulation, because its nonlinear characteristics of the inductance and resistance are anhysteretic. On the other side, the Preisach and Jiles-Atherton models can be used to explain the remanent flux phenomenon as shown in [30] for the Preisach model. However, even if it is inappropriate for simulation in general, the role of model (Fig. 2) in this paper is to clarify the rationale behind our experimental procedure. Thereby, due to the straightforward physical explainability of the model, it was not necessary, given the focus of this paper, to use simulation as additional validation of the experiment. Our future research will address various already mentioned modeling and simulation methods, but also FEM and BEM modelling [31].

Assume that the AC source voltage is

$$u = \hat{U}\sin \omega t. \tag{1}$$

Kirchhoff's voltage law for the model shown in Fig. 2 equals

$$iR + N\frac{d\varphi}{dt} = \hat{U}\sin \omega t. \tag{2}$$

In the steady state, (2) for the DC components can be expressed as

$$I(0)R + N\frac{d\Phi(0)}{dt} = U(0), \tag{3}$$

where $I(0)$, $\Phi(0)$ and $U(0)$ are the DC components of the magnetizing current (i), the magnetic flux (φ) and the AC source voltage (u), respectively. Considering that the DC component of the AC source voltage, $U(0)$, is zero and concerning (3), the DC component of the magnetizing current, $I(0)$, must also be zero because $\Phi(0)$ has a constant value per definition and, thus, its derivative equals to zero.

Furthermore, considering that the magnetic flux (φ) is an odd function of the magnetizing current (i), as shown in Fig. 1, the DC component of the magnetic flux, $\Phi(0)$, in the steady state must also be zero.

The basic idea is to energize the unloaded transformer at the nominal AC voltage and measure the magnetizing current (i) and inductance voltage (u_L). The inductance voltage can be obtained as the difference between the measured transformer terminal voltage (u) and the product of the current (i) and the resistance (R):

$$u_L = u - iR. \tag{4}$$

Furthermore, the inductance voltage (uL) can also be obtained by measuring the voltage on the secondary unloaded winding and converting it to the primary side using the turns ratio. The magnetic flux (φ) in the core equals

$$\varphi(t) = \frac{1}{N}\int_0^t u_L(\tau)d\tau + \Phi_R, \tag{5}$$

where N is the number of the corresponding transformer winding turns, t and τ are the time variables, and Φ_R is the remanent flux value, that is, the magnetic flux at instant $t = 0$. Considering (5), the DC component of the magnetic flux (φ) in the steady state can be expressed as

$$\Phi(0) = \frac{1}{T}\int_{t_{ss}}^{t_{ss}+T}\left[\frac{1}{N}\int_0^t u_L(\tau)d\tau + \Phi_R\right]dt, \tag{6}$$

where T is the period of the AC source voltage. Considering that the DC component of the magnetic flux, $\Phi(0)$, in the steady state must be zero and concerning (6), the remanent flux (Φ_R) equals

$$\Phi_R = -\frac{1}{T}\int_{t_{ss}}^{t_{ss}+T}\left[\frac{1}{N}\int_0^t u_L(\tau)d\tau\right]dt, \tag{7}$$

Whereby it is crucial to choose the period for calculating the DC component in the steady state, not during the transient state. In other words, the instant t_{ss} must be in a steady state. Consequently, the remanent flux (Φ_R) can be obtained using the measured inductance voltage (u_L), that is, the calculated magnetic flux (φ_C):

$$\Phi_R = -\frac{1}{T}\int_{t_{ss}}^{t_{ss}+T}\varphi_C(t)dt, \tag{8}$$

$$\varphi_C(t) = \frac{1}{N}\int_0^t u_L(\tau)d\tau. \tag{9}$$

3. LABORATORY SETUP

The measurement circuit with the model shown in Fig. 3 is built to determine the remanent flux (Φ_R) by analyzing the unloaded transformer's inductance voltage (u_L) waveform. Various initial conditions of magnetic flux at the moment of de-energization (de-energization flux, Φ_D) and the closing voltage angle (α) will be set to prove the validity of the proposed method independently of the initial conditions.

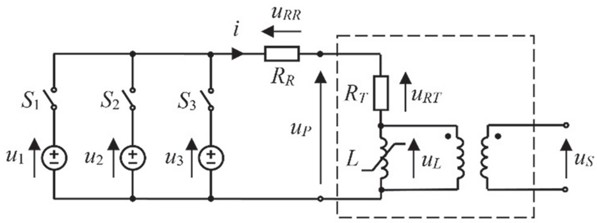


Fig. 3. The measurement circuit model

The measurement circuit model consists of the two winding transformer (inside the dashed rectangle), the resistance R_R, three variable AC voltage sources, u_1, u_2 and u_3, two electronically controlled switches S_1 and S_2, and the ordinary mechanical switch S_3. The transformer is modeled with the resistance R_T and the perfect transformer (ideal transformer with magnetizing inductance L included) connected in a series (Fig. 3). The magnetic characteristics of the nonlinear inductance L are shown in Fig. 1. Physical realization of the measurement circuit is shown in Fig. 4.

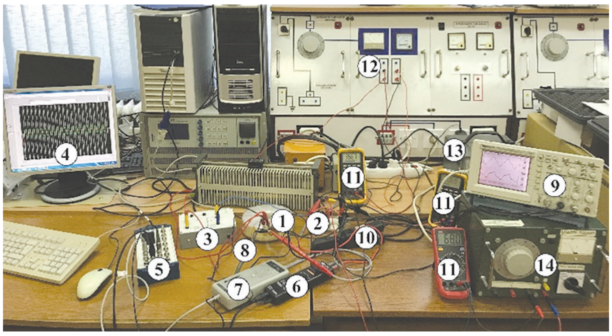


Fig. 4. Realization of the measurement circuit

The measurement circuit consists of: 1 – transformer under test, 2 – resistor R_R, 3 – electronic switching device, 4 – PC with installed data acquisition software, 5 – data acquisition card, 6 and 7 – active differential probes, 8 – passive voltage probe, 9 – oscilloscope, 10 – current probe, 11 – digital multimeters, 12 – variable AC voltage source u_1 used for the energization of the transformer, 13 – variable AC voltage source u_2 used for the setting of the de-energization flux, 14 – variable AC voltage source u_3 used for demagnetization.

The single-phase transformer (1, Fig. 4) is made of a toroidal core with two windings – 47 and 7 turns, both wired with triple wire (each has a round cross-section of 1.3 mm^2). The magnetic core (the cross-sectional area is 20 cm^2) consists of oriented transformer sheets (M5-type). The nominal power is 200 VA, nominal voltages are 30 V and 4.5 V, and nominal currents are 6.5 A and 44.5 A. The resistor R_T equals 0.19 Ω and inductance L equals 0.59 H in linear (non-saturated) area (Fig. 3). The purpose of the resistor R_R which equals 1.22 Ω (2, Fig. 4, also shown in Fig. 3) is to limit the inrush current. If not used, it could reach up to 120 A, devastating for the equipment used. The secondary winding terminals are used only for obtaining the voltage (u_S). The electronic switching device (3, Fig. 4) sets the initial remanent flux. All the measured values (magnetizing current, primary and secondary voltage) are obtained using the data acquisition (DAQ) card National Instruments NI-USB 6212 (5, Fig. 4) and the PC (4, Fig. 4). The waveform of the magnetizing current (i) is obtained indirectly by measuring the voltage on the additional resistor (R_R) using active differential probe GW Instek GDP-025 (6, Fig. 4). Furthermore, the primary voltage (u_P) is also obtained using active differential probe (7, Fig. 4) and the secondary voltage (u_S) using passive differential probe (8, Fig. 4). The sampling frequency was set to 50 kHz. The frequency of all the AC sources is 50 Hz. Furthermore, oscilloscope (9, Fig. 4) and digital multimeters (11, Fig. 4) were used only for monitoring the situation. All the measuring data used in this research is collected using the DAQ card and PC.

4. EXPERIMENTAL PROCEDURE

Every single measurement is carried out in three steps, but only to test the proposed method's validity. In possible application, only the third step of the experimental procedure should be carried out. The first step is AC demagnetization, carried out by slowly decreasing the voltage of the variable AC source u_3 from the RMS value of 36 V to zero in approximately 10 s, as shown in [32]. During the core demagnetization, the switches S_1 and S_2 are open. The demagnetization is important for setting the de-energization flux (Φ_D) value in the second step.

The second step follows up approximately 5 s after the first step, and it is done using the AC source u_2 and the switch S_2, while the switches S_1 and S_3 are open. The second step is described in detail in [15]. The value of the de-energization flux (Φ_D) is set by changing the RMS voltage of the variable AC source u_2 (U_2) and the instant of opening the switch S_2 (when the current is crossing zero value). In total, 25 different de-energization flux values are obtained in this experiment which corresponds to 13 different RMS voltages of the variable AC source u_2 (U_2), including zero value, and two different zero crossings of the magnetizing current in the φ-i characteristics. De-energization flux (Φ_D) values and corresponding RMS voltages U_2 are shown in Table 1. (only positive values due to the symmetry of the φ-i characteristics).

Table 1. Corresponding de-energization flux values and RMS voltages of the variable AC source u_2.

U_2 (V)	Φ_D (mVs)	U_2 (V)	Φ_D (mVs)
36	2.997	18	1.152
33	2.752	15	0.901
30	2.364	12	0.650
27	2.040	9	0.433
24	1.728	6	0.140
21	1.441	3	0.061

The third step follows up in less than 1 s after the second step. Thus, the value of the set de-energization flux (Φ_D) will be considered as the value of the remanent flux (Φ_R) in the core in the moment of energization of the transformer in the third step. This is the usual procedure when determining the remanent flux during the de-energization of the transformer by measuring the port-voltage [33, 34]. The third step of the experimental procedure is the only step which will be carried out in the possible application of the proposed method, and it is energizing the transformer using the AC source u_1 and the electronically controlled switch S_1, while the switches S_2 and S_3 are open. The RMS voltage of the AC source u_1 is set to 36 V which is 20% higher than the nominal voltage of the primary winding. It is done to obtain the steady state faster. Namely, the time constant that affects the length of the transient state is $L/(RT+RR)$ whereby the inductance L is not a constant value, but equals

$$L = \frac{d\varphi}{di}. \tag{10}$$

Thus, when the core reaches saturation, the inductance L is significantly lower than in the non-saturated region. Finally, the lower inductance L, the lower time constant means faster entry in the steady state. This is important because the remanent flux (Φ_R) value is obtained as the negative value of the DC component of the calculated magnetic flux (φ_C) in the steady state. The closing voltage angle (α), is set by the PC over the electronically controlled switch S_1. To prove that the different initial conditions of the de-energization flux (Φ_D) and closing voltage angle (α) do not affect the validity and accuracy of the proposed method, the measurements were carried out by varying the following parameters:

- U_2 = 36 V (-), 33 V (-), …, 0 V, …, 33 V (+), 36 V (+);
- α = 0°, 30°, 60°, 90°, 120°, 150°, 180°;

which in total gives 175 measurements. At the end, the obtained values of the remanent flux using the proposed method will be compared to the set values of the de-energization flux.

5. EXPERIMENTAL RESULTS

The obtained results for each measurement include the magnetizing current (i), the primary winding voltage (u_P), the secondary winding voltage (u_S), and the calculated magnetic flux (φ_C). Examples of the obtained waveforms are shown in Figs. 5, 6, 7, and 8, respectively.

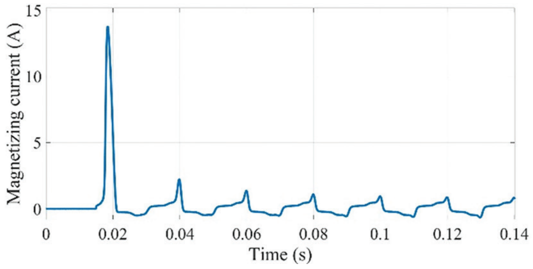


Fig. 5. Magnetizing current (i) for U_2 = 18 V (+) and α = 90°

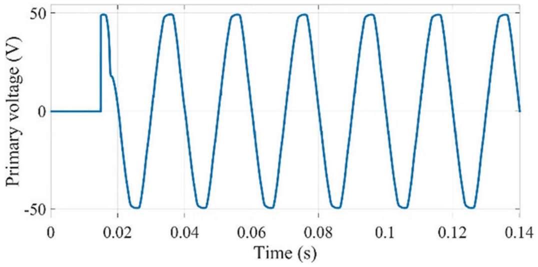


Fig. 6. Primary winding voltage (u_P) for U_2 = 18 V (+) and α = 90°

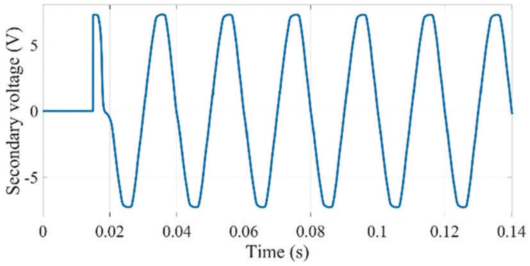


Fig. 7. Secondary winding voltage (u_S) for U_2 = 18 V (+) and α = 90°

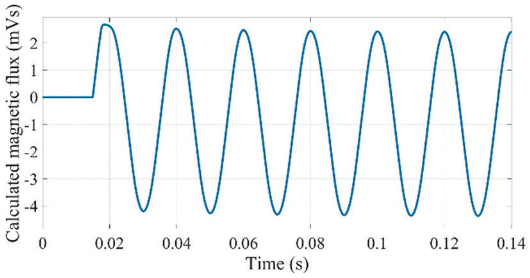


Fig. 8. Calculated magnetic flux (φ_C) for U_2 = 18 V (+) and α = 90°

For each measurement, the remanent flux (Φ_R) value is obtained as the negative value of the DC component of the calculated magnetic flux (φ_C) in the steady state. The obtained remanent flux (Φ_R) values are shown in Table 2, where parameter U_2 is marked with its value and sign (+) or (–) attached, depending on the sign of the set de-energization flux (Φ_D).

Table 2. Obtained remanent flux values

RMS voltage, U_2 (V)	Closing voltage angle, α						
	0°	30°	60°	90°	120°	150°	180°
36 (+)	3.082	3.083	3.085	3.080	3.078	3.079	3.077
33 (+)	2.752	2.788	2.757	2.747	2.748	2.751	2.766
30 (+)	2.123	2.179	2.164	2.369	2.330	2.178	2.337
27 (+)	1.805	1.881	2.104	2.078	1.872	1.864	2.055
24 (+)	1.605	1.721	1.557	1.705	1.664	1.750	1.609
21 (+)	1.323	1.455	1.378	1.344	1.536	1.309	1.242
18 (+)	1.160	1.326	1.041	1.018	1.252	1.189	1.280
15 (+)	1.004	1.072	0.868	1.026	0.913	0.864	0.892
12 (+)	0.584	0.508	0.606	0.687	0.544	0.645	0.650
9 (+)	0.377	0.378	0.445	0.370	0.454	0.353	0.429
6 (+)	0.273	0.331	0.339	0.310	0.340	0.312	0.267
3 (+)	0.167	0.163	0.186	0.151	0.181	0.143	0.148
0	–0.049	0.002	0.002	0.013	0.051	–0.037	–0.023

RMS voltage, U_2 (V)	Closing voltage angle, α						
	0°	30°	60°	90°	120°	150°	180°
3 (–)	–0.155	–0.128	–0.130	–0.139	–0.140	–0.156	–0.151
6 (–)	–0.308	–0.263	–0.278	–0.270	–0.307	–0.342	–0.335
9 (–)	–0.443	–0.386	–0.339	–0.457	–0.381	–0.438	–0.430
12 (–)	–0.714	–0.738	–0.713	–0.692	–0.644	–0.801	–0.614
15 (–)	–0.992	–0.756	–0.802	–0.974	–0.879	–0.887	–0.988
18 (–)	–1.065	–1.126	–1.087	–0.917	–1.006	–1.099	–1.239
21 (–)	–1.458	–1.261	–1.195	–1.447	–1.361	–1.360	–1.208
24 (–)	–1.543	–1.660	–1.771	–1.731	–1.502	–1.591	–1.847
27 (–)	–1.848	–1.825	–1.837	–2.067	–1.850	–2.054	–1.916
30 (–)	–2.159	–2.243	–2.295	–2.180	–2.238	–2.115	–2.291
33 (–)	–2.725	–2.735	–2.727	–2.702	–2.690	–2.749	–2.731
36 (–)	–3.051	–3.049	–3.028	–3.043	–3.048	–3.055	–3.057

For each parameter of U_2, the obtained remanent flux (Φ_R) value should be the same, regardless of the initial condition of the closing voltage angle (α). To evaluate this, for each parameter of RMS voltage U_2, the average remanent flux value ($\Phi_{R\_average}$) is calculated as

$$\Phi_{R\_average} = \frac{\sum_{i=1}^{7} \Phi_{Ri}}{7}. \tag{11}$$

Also, the standard deviation (σ) of the obtained remanent flux (Φ_R) values for each parameter of RMS voltage U_2 is calculated as

$$\sigma = \sqrt{\frac{\sum_{i=1}^{7} \left(\Phi_{Ri} - \Phi_{R\_average}\right)^2}{6}}. \tag{12}$$

Furthermore, the relative standard deviation ($\sigma_{\%}$) is calculated for each parameter of U_2 as

$$\sigma_{\%} = \frac{\sqrt{\frac{\sum_{i=1}^{7} \left(\Phi_{Ri} - \Phi_{R\_average}\right)^2}{6}}}{\Phi_{R\_average}} \cdot 100\ \%. \tag{13}$$

Table 3. Standard deviation and relative standard deviation of the obtained remanent flux values

RMS voltage U_2 (V)	Average remanent flux value, Φ_{R}–average (mVs)	Standard deviation, σ (mVs)	Relative standard deviation, $\sigma_{\%}$
36 (+)	3.080	0.003	0.09%
33 (+)	2.759	0.014	0.49%
30 (+)	2.240	0.093	4.17%
27 (+)	1.951	0.114	5.82%
24 (+)	1.659	0.066	3.95%
21 (+)	1.370	0.091	6.63%
18 (+)	1.181	0.109	9.19%
15 (+)	0.949	0.078	8.21%
12 (+)	0.603	0.058	9.69%
9 (+)	0.401	0.037	9.35%
6 (+)	0.310	0.028	9.01%
3 (+)	0.163	0.015	9.42%
0	–0.006	0.031	535.62%
3 (–)	–0.143	0.011	7.43%
6 (–)	–0.300	0.029	9.63%
9 (–)	–0.410	0.039	9.60%
12 (–)	–0.702	0.057	8.06%
15 (–)	–0.897	0.087	9.69%
18 (–)	–1.077	0.092	8.59%
21 (–)	–1.327	0.100	7.51%
24 (–)	–1.664	0.117	7.03%
27 (–)	–1.914	0.097	5.06%
30 (–)	–2.217	0.063	2.84%
33 (–)	–2.723	0.019	0.69%
36 (–)	–3.047	0.009	0.30%

The measurement results show that the relative standard deviation ($\sigma_{\%}$) does not exceed 10% in any case, except for $U_2 = 0$ V. That exception is because the denominator value (average remanent flux) is near zero when calculating the relative standard deviation ($\sigma_{\%}$). Also, the relative standard deviation ($\sigma_{\%}$) lowers when the RMS voltage U_2 rises. The reason for these deviations could be in the second step of the experimental procedure when magnetizing the transformer, that is, setting the de-energization flux (Φ_D). In the first step of the experimental procedure, the transformer is demagnetized and this is done accurately. But the second step is critical in terms of accuracy, especially at lower magnetizing voltage values. At higher magnetizing voltage values, the relative standard deviation ($\sigma_{\%}$) is less than 1%. In these cases, the transformer goes into saturation even at the steady state, while at lower magnetizing voltage values it does not go into saturation at all. So, at lower magnetizing voltage values, the transformer will slowly enter a steady state because of the higher time constant in these cases. As a result, the magnetizing process in the second step does not always set the aimed de-energization flux (Φ_D) value accurately.

Finally, the obtained average remanent flux values for each parameter U_2 are compared to the corresponding de-energization flux (Φ_D) values shown in Table 1. The results are shown in Fig. 9.

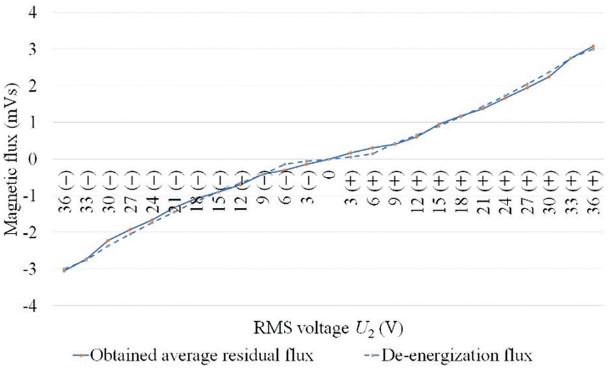


Fig. 9. Obtained average remanent flux values and corresponding de-energization flux values for each parameter of RMS voltage U_2

The obtained average remanent flux value is almost the same as the corresponding de-energization flux value for each parameter U_2, as shown in Fig. 9. Also, symmetry of the results can be seen in Fig. 9, which is the expected result because of symmetrical *φ-i* characteristics. The ratios of the obtained average values of remanent flux (Φ_R) and corresponding values of de-energization flux (Φ_D) are shown in Table 4 for each parameter U_2.

Table 4. Ratios of the obtained average values of remanent flux and corresponding values of de-energization flux

U_2 (V)	Remanent flux / De-energization flux	U_2 (V)	Remanent flux / De-energization flux
36 (+)	102.79%	36 (–)	101.69%
33 (+)	100.23%	33 (–)	98.93%
30 (+)	94.74%	30 (–)	93.78%
27 (+)	95.65%	27 (–)	93.80%
24 (+)	95.97%	24 (–)	96.25%
21 (+)	95.01%	21 (–)	92.07%
18 (+)	102.47%	18 (–)	93.46%
15 (+)	105.28%	15 (–)	99.55%
12 (+)	92.88%	12 (–)	108.08%
9 (+)	92.56%	9 (–)	94.81%
6 (+)	222.05%	6 (–)	214.98%
3 (+)	266.70%	3 (–)	233.93%

The results in Table 4 show that the average remanent flux values are between 93% and 108% of the corresponding values of de-energization flux in most cases. The exceptions appear only for parameters with RMS voltage $U_2 = 3$ V (+), $U_2 = 3$ V (–), $U_2 = 6$ V (+), and $U_2 = 6$ V (–). The reason for these exceptions could be the relatively small absolute values of the obtained remanent flux and corresponding de-energization flux which can cause higher measurement uncertainty. Another reason could be the imprecise setting of the de-energization flux in the second step of the experimental procedure.

6. CONCLUSION

The remanent flux can be determined using the presented experimental procedure without any data about parameters or past states of a transformer, using only electrical measurements. Some of the previously mentioned methods determine the remanent flux during the de-energization, which is very useful for reducing the inrush current by controlled switching. However, most of these methods cannot be used for investigating stability over time and the impact of the external excitations (system transients and magnetic fields) on the remanent flux in the core because they determine it in the moment of de-energization. Namely, those external excitations could change the remanent flux value while the transformer is not even connected to the grid. This means that the remanent flux can have a different value at the end of such an idle state, compared to the de-energization instant. On the other hand, the proposed method determines the remanent flux in the moment of energization of the transformer, that is, after an idle state. This means that the proposed method is unusable for reducing the inrush current by controlled switching because some new value of the remanent flux is established in the core after conducting the experimental procedure. It also means that the determined value will depend on the moment of the previous de-energization. Still, every change of the remanent flux between the previous de-energization and new energization will be taken into account, contrary to the other methods of determination during the de-energization. Thus, some new value of the remanent flux is established in the core after conducting the proposed method, but in investigating the stability over time and impacts of external excitations on the remanent flux, this drawback is not crucial because the goal is to determine how remanent flux was changed during the idle state, that is, its new value established at the end of the experimental procedure is not significant. Although there is a method similar to the proposed one which uses the low voltage DC source to determine the remanent flux, the proposed method could be more applicable because it uses the nominal voltage for energization, which could be easier to obtain on-site. Furthermore, the proposed method does not demand any data about the observed transformer, except the nominal voltage, which is not the case when using the method with the DC energization.

Finally, the other methods which determine the remanent flux during de-energization and the proposed method go along in investigating the stability over time and impacts of external excitations on the remanent flux. Namely, using the previously known methods, the remanent flux will be determined in the moment of de-energization and using the proposed method, the remanent flux will be determined in the moment of new energization, enabling comparison of these two values, that is, detecting the remanent flux changes due to the impact of external excitations.

7. REFERENCES:

[1] U. Parikh, B. R. Bhalja, "Mitigation of magnetic inrush current during controlled energization of coupled un-loaded power transformers in presence of residual flux without load side voltage measurements", International Journal of Electrical Power & Energy Systems, Vol. 76, 2016, pp. 156-164.

[2] J. H. Brunke, K. J. Fröhlich, "Elimination of transformer inrush currents by controlled switching - Part I: Theoretical considerations", IEEE Transactions on Power Delivery, Vol. 16, No. 2, 2001, pp. 276-280.

[3] S. Fang, H. Ni, H. Lin, S. L. Ho, "A Novel Strategy for Reducing Inrush Current of Three-Phase Transformer Considering Residual Flux", IEEE Transactions on Industrial Electronics, Vol. 63, No. 7, 2016, pp. 4442-4451.

[4] J. Song, T. Zheng, Y. Guo, L. Pan, "Suppression Strategy of the Inrush Current in the Transformer Phase Selection Closing Considering the Residual Flux", Proceedings of the 6th Asia Conference on Energy and Electrical Engineering, Chengdu, China, 2023, pp. 140-145

[5] E. Hajipour, M. Salehizadeh, M. Vakilian, M. Sanaye-Pasand, "Residual Flux Mitigation of Protective Current Transformers Used in an Autoreclosing Scheme", IEEE Transactions on Power Delivery, Vol. 31, No. 4, 2016, pp. 1636-1644.

[6] J. Duan, Z. Jin, Y. Lei, "Residual flux suppression of protective current transformers for autoreclosure process", Proceedings of the IEEE International Transportation Electrification Conference and Expo, Asia-Pacific, Harbin, China, 7-10 August 2017.

[7] S. Sanati, Y. Alinejad-Beromi, "Fast and Complete Mitigation of Residual Flux in Current Transformers Suitable for Auto-Reclosing Schemes Using Jiles-Atherton Modeling", IEEE Transactions on Power Delivery, Vol. 37, No. 2, 2022, pp. 765-774.

[8] K. Milicevic, D. Vinko, D. Vulin, "Experimental investigation of impact of remnant flux on the ferroresonance initiation", International Journal of Electrical Power & Energy Systems, Vol. 61, 2014, pp. 346-354.

[9] Y. Husianycia, M. Rioual, "Determination of the residual fluxes when de-energizing a power transformer/comparison with on-site tests", Proceedings of the IEEE Power & Energy Society General Meeting, San Francisco, CA, USA, 16 June 2005, pp. 449-454.

[10] D. Tishuai, Z. Bi-De, F. Chun-En, L. Wei, R. Xiao, C. Chuanjiang, "Influence of system transients on the residual flux of three-phase transformers", Proceedings of the 4th International Conference on Electric Power Equipment - Switching Technology, Xi'an, China, 22-25 October 2017, pp. 970-973.

[11] M. F. Lachman, V. Fomichev, V. Rashkovski, A. M. Shaikh, "Frequency response analysis of transformers and influence of magnetic viscosity", Proceedings of the 77th Annual International Doble Client Conference, 2010, pp. 1-18.

[12] D. Cavallera, V. Oiring, J. L. Coulomb, O. Chadebec, B. Caillault, F. Zgainski, "A new method to evaluate residual flux thanks to leakage flux, application to a transformer", IEEE Transactions on Magnetics, Vol. 50, No. 2, 2014.

[13] J. Horiszny, "Method of determining the residual fluxes in transformer core", Proceedings of the 18th International Symposium on Electromagnetic Fields in Mechatronics, Electrical and Electronic Engineering, Lodz, Poland, 14-16 September 2017, pp. 2-3.

[14] H. Zhang et al. "A New Method to Measure the Residual Flux by Magnetic Sensors and a Finite-Element Model", IEEE Transactions on Instrumentation and Measurement, Vol. 72, pp. 1-10, 2023.

[15] D. Vulin, I. Biondic, K. Milicevic, D. Vinko, "Laboratory setup for determining residual magnetic flux value using low voltage DC source", Electrical Engineering, Vol. 102, 2020, pp. 1707-1714.

[16] C. Wei, X. Li, M. Yang, Z. Ma, H. Hou, "Novel remanence determination for power transformers based on magnetizing inductance measurements", Energies, Vol. 12, No. 24, 2019.

[17] D. Vulin, K. Milicevic, I. Biondic, G. Petrovic, "Determining the Residual Magnetic Flux Value of a Single-Phase Transformer Using a Minor Hysteresis Loop", IEEE Transactions on Power Delivery, Vol. 8977, No. C, 2020, pp. 1-10.

[18] W. Ge, Y. Wang, Z. Zhao, X. Yang, Y. Li, "Residual flux in the closed magnetic core of a power transformer", IEEE Transactions on Applied Superconductivity, Vol. 24, No. 3, 2014, pp. 3-6.

[19] T. Zheng et al. "Fast, in situ demagnetization method for protection current transformers", IEEE Transactions on Magnetics, Vol. 52, No. 7, 2016, pp. 1-4.

[20] B. Kovan, F. De Leon, D. Czarkowski, Z. Zabar, L. Birenbaum, "Mitigation of inrush currents in network transformers by reducing the residual flux with an ultra-low-frequency power source", IEEE Transactions on Power Delivery, Vol. 26, No. 3, 2011, pp. 1563-1570.

[21] F. De Leon, A. Farazmand, S. Jazebi, D. Deswal, R. Levi, "Elimination of residual flux in transformers by the application of an alternating polarity dc voltage source", IEEE Transactions on Power Delivery, Vol. 30, No. 4, 2015, pp. 1727-1734.

[22] V. O. De Castro Cezar, L. L. Rouve, J. L. Coulomb, F. X. Zgainski, O. Chadebec, B. Caillault, "Elimination of inrush current using a new prefluxing method. Application to a single-phase transformer", Proceedings of the International Conference on Electrical Machines, Berlin, Germany, 2-5 September 2014, pp. 1717-1723.

[23] P. J. Kotak, P. Jaikaran, "Prefluxing technique to mitigate inrush current of three-phase power transformer", International Journal of Scientific and Engineering Research, Vol. 4, No. 6, 2013, pp. 135-141.

[24] D. I. Taylor, J. D. Law, B. K. Johnson, N. Fischer, "Single-phase transformer inrush current reduction using prefluxing", IEEE Transactions on Power Delivery, Vol. 27, No. 1, 2012, pp. 245-252.

[25] Y. Pan, X. Yin, Z. Zhang, B. Liu, M. Wang, X. Yin, "Three-Phase Transformer Inrush Current Reduction Strategy Based on Prefluxing and Controlled Switching", IEEE Access, Vol. 9, 2021, pp. 38961-38978.

[26] T. Liu, H. Siguerdidjane, M. Petit, T. Jung, J. P. Dupraz, "Reconstitution of power transformer's residual flux with CVT's measurement during its de-energization", Proceedings of the IEEE International Conference on Control Technology and Applications, Yokohama, Japan, 8-10 September 2010 pp. 206-209.

[27] L. O. Chua, K. A. Stromsmoe, "Lumped-circuit models for nonlinear inductors exhibiting hysteresis loops", IEEE Transactions on Circuit Theory, Vol. 17, No. 4, 1970, pp. 564-574.

[28] F. Preisach, "On the magnetic aftereffect", IEEE Transactions on Magnetics, Vol. 53, No. 3, 2017, pp. 1-11.

[29] D. C. Jiles, J. B. Thoelke, M. K. Devine, "Numerical determination of hysteresis parameters for the modeling of magnetic properties using the theory of ferromagnetic hysteresis", IEEE Transactions on Magnetics, Vol. 28, No. 1, 1992, pp. 27-35.

[30] A. Wilk, M. Michna, A. Cichowski, "Simulation of the remanence influence on the transient states of the single-phase transformer including feedback preisach model", Proceedings of the 40th Annual Conference of the IEEE Industrial Electronics Society, Dallas, TX, USA, 2014, pp. 875-880.

[31] M. Marković, Ž. Štih, B. Ćućić, "Power transformer main insulation design improvement using BEM and FEM", Proceedings of Eurocon 2013, Zagreb, Croatia, 2013, pp. 1553-1560

[32] D. Lovejoy, "Demagnetization", Magnetic particle inspection, Springer, Dordrecht, 1993, pp. 149-169.

[33] M. Rioual et al. "Field application of a synchronous controller based on the measurement of residual fluxes for the energization of a step-up transformer", Proceedings of the IEEE Power & Energy Society General Meeting, 2011, pp. 1-8.

[34] G. Petrović, T. Kilić, S. Milun, "Remanent flux measurement and optimal energization instant determination of power transformer", Proceedings of the XVII IMEKO World Congress 22-27 June 2003, Dubrovnik, Croatia, pp. 952-955.

INTERNATIONAL JOURNAL OF ELECTRICAL AND COMPUTER ENGINEERING SYSTEMS

Published by Faculty of Electrical Engineering, Computer Science and Information Technology Osijek,
Josip Juraj Strossmayer University of Osijek, Croatia.

About this Journal

The International Journal of Electrical and Computer Engineering Systems publishes original research in the form of full papers, case studies, reviews and surveys. It covers theory and application of electrical and computer engineering, synergy of computer systems and computational methods with electrical and electronic systems, as well as interdisciplinary research.

Topics of interest include, but are not limited to:

- Power systems
- Renewable electricity production
- Power electronics
- Electrical drives
- Industrial electronics
- Communication systems
- Advanced modulation techniques
- RFID devices and systems
- Signal and data processing
- Image processing
- Multimedia systems
- Microelectronics
- Instrumentation and measurement
- Control systems
- Robotics
- Modeling and simulation
- Modern computer architectures
- Computer networks
- Embedded systems
- High-performance computing
- Parallel and distributed computer systems
- Human-computer systems
- Intelligent systems
- Multi-agent and holonic systems
- Real-time systems
- Software engineering
- Internet and web applications and systems
- Applications of computer systems in engineering and related disciplines
- Mathematical models of engineering systems
- Engineering management
- Engineering education

Paper Submission

Authors are invited to submit original, unpublished research papers that are not being considered by another journal or any other publisher. Manuscripts must be submitted in doc, docx, rtf or pdf format, and limited to 30 one-column double-spaced pages. All figures and tables must be cited and placed in the body of the paper. Provide contact information of all authors and designate the corresponding author who should submit the manuscript to https://ijeces.ferit.hr. The corresponding author is responsible for ensuring that the article's publication has been approved by all coauthors and by the institutions of the authors if required. All enquiries concerning the publication of accepted papers should be sent to ijeces@ferit.hr.

The following information should be included in the submission:

- paper title;
- full name of each author;
- full institutional mailing addresses;
- e-mail addresses of each author;
- abstract (should be self-contained and not exceed 150 words). Introduction should have no subheadings;
- manuscript should contain one to five alphabetically ordered keywords;
- all abbreviations used in the manuscript should be explained by first appearance;
- all acknowledgments should be included at the end of the paper:
- authors are responsible for ensuring that the information in each reference is complete and accurate. All references must be numbered consecutively and citations of references in text should be identified using numbers in square brackets. All references should be cited within the text;
- each figure should be integrated in the text and cited in a consecutive order. Upon acceptance of the paper, each figure should be of high quality in one of the following formats: EPS, WMF, BMP and TIFF;
- corrected proofs must be returned to the publisher within 7 days of receipt.

Peer Review

All manuscripts are subject to peer review and must meet academic standards. Submissions will be first considered by an editor-in-chief and if not rejected right away, then they will be reviewed by anonymous reviewers. The submitting author will be asked to provide the names of 5 proposed reviewers including their e-mail addresses. The proposed reviewers should be in the research field of the manuscript. They should not be affiliated to the same institution of the manuscript author(s) and should not have had any collaboration with any of the authors during the last 3 years.

Author Benefits

The corresponding author will be provided with a .pdf file of the article or alternatively one hardcopy of the journal free of charge.

Units of Measurement

Units of measurement should be presented simply and concisely using System International (SI) units.

Bibliographic Information

Commenced in 2010.
ISSN: 1847-6996
e-ISSN: 1847-7003

Published: semiannually

Copyright

Authors of the International Journal of Electrical and Computer Engineering Systems must transfer copyright to the publisher in written form.

Subscription Information

The annual subscription rate is 50€ for individuals, 25€ for students and 150€ for libraries.

Postal Address

Faculty of Electrical Engineering,
Computer Science and Information Technology Osijek,
Josip Juraj Strossmayer University of Osijek, Croatia
Kneza Trpimira 2b
31000 Osijek, Croatia

IJECES Copyright Transfer Form

(Please, read this carefully)

This form is intended for all accepted material submitted to the IJECES journal and must accompany any such material before publication.

TITLE OF ARTICLE (hereinafter referred to as "the Work"):

COMPLETE LIST OF AUTHORS:

The undersigned hereby assigns to the IJECES all rights under copyright that may exist in and to the above Work, and any revised or expanded works submitted to the IJECES by the undersigned based on the Work. The undersigned hereby warrants that the Work is original and that he/she is the author of the complete Work and all incorporated parts of the Work. Otherwise he/she warrants that necessary permissions have been obtained for those parts of works originating from other authors or publishers.

Authors retain all proprietary rights in any process or procedure described in the Work. Authors may reproduce or authorize others to reproduce the Work or derivative works for the author's personal use or for company use, provided that the source and the IJECES copyright notice are indicated, the copies are not used in any way that implies IJECES endorsement of a product or service of any author, and the copies themselves are not offered for sale. In the case of a Work performed under a special government contract or grant, the IJECES recognizes that the government has royalty-free permission to reproduce all or portions of the Work, and to authorize others to do so, for official government purposes only, if the contract/grant so requires. For all uses not covered previously, authors must ask for permission from the IJECES to reproduce or authorize the reproduction of the Work or material extracted from the Work. Although authors are permitted to re-use all or portions of the Work in other works, this excludes granting third-party requests for reprinting, republishing, or other types of re-use. The IJECES must handle all such third-party requests. The IJECES distributes its publication by various means and media. It also abstracts and may translate its publications, and articles contained therein, for inclusion in various collections, databases and other publications. The IJECES publisher requires that the consent of the first-named author be sought as a condition to granting reprint or republication rights to others or for permitting use of a Work for promotion or marketing purposes. If you are employed and prepared the Work on a subject within the scope of your employment, the copyright in the Work belongs to your employer as a work-for-hire. In that case, the IJECES publisher assumes that when you sign this Form, you are authorized to do so by your employer and that your employer has consented to the transfer of copyright, to the representation and warranty of publication rights, and to all other terms and conditions of this Form. If such authorization and consent has not been given to you, an authorized representative of your employer should sign this Form as the Author.

Authors of IJECES journal articles and other material must ensure that their Work meets originality, authorship, author responsibilities and author misconduct requirements. It is the responsibility of the authors, not the IJECES publisher, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it.

- The undersigned represents that he/she has the authority to make and execute this assignment.
- For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.
- The undersigned agrees to indemnify and hold harmless the IJECES publisher from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.

______________________________ ______________________________

Author/Authorized Agent **Date**

CONTACT
International Journal of Electrical and Computer Engineering Systems (IJECES)
Faculty of Electrical Engineering, Computer Science and Information Technology Osijek
Josip Juraj Strossmayer University of Osijek
Kneza Trpimira 2b
31000 Osijek, Croatia
Phone: +38531224600,
Fax: +38531224605,
e-mail: ijeces@ferit.hr