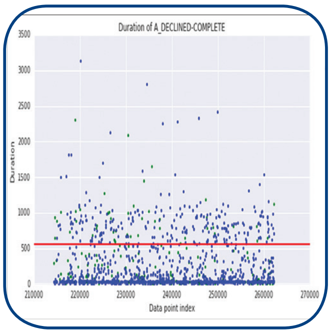
Volume 15, Number 6, 2024

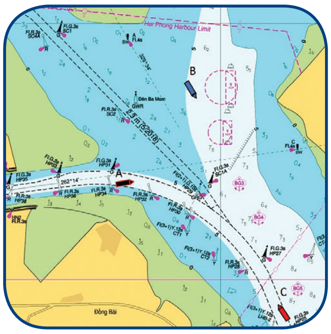
ISSN 1847-6996

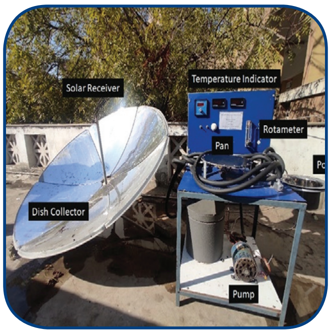
FERIT
FACULTY OF ELECTRICAL ENGINEERING, COMPUTER SCIENCE AND INFORMATION TECHNOLOGY **OSIJEK**

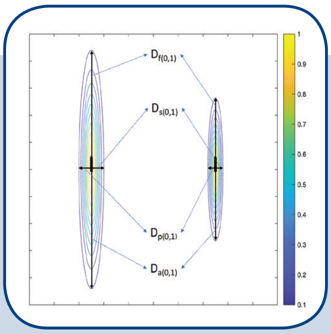
IJECES International Journal of Electrical and Computer Engineering Systems

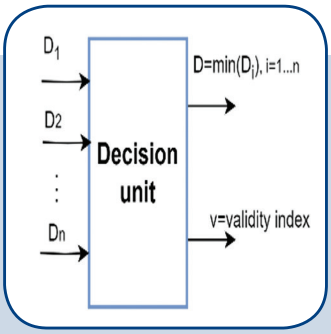
International Journal of Electrical and Computer Engineering Systems

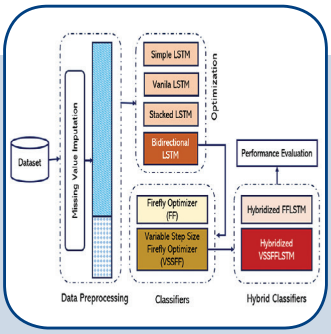












ISSN: 1847-6996

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 15, Number 6, 2024 | Pages 469 - 552

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

Proofreader

Ivanka Ferčec
J.J. Strossmayer University of Osijek, Croatia

Editing and technical assistence

Davor Vrandečić
J.J. Strossmayer University of Osijek, Croatia

Stephen Ward
J.J. Strossmayer University of Osijek, Croatia

Dražen Bajer
J.J. Strossmayer University of Osijek, Croatia

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

Cluster-based Improvised Time Synchronization Algorithm for Multihop IoT Networks

Original Scientific Paper

Neha Dalwadi*
Shri K. J. Polytechnic, Bharuch,
Department of Computer Engineering
Bholav, Bharuch, India
neha.dalwadi@gmail.com

Mamta Padole
The Maharaja Sayajirao University of Baroda,
Faculty of Technology and Engineering, Department of Computer Science and Engineering
Kala bhavan, Vadodara, India
mpadole29@rediffmail.com

*Corresponding author

Abstract *– Achieving precise time synchronization among wireless sensor devices within Internet-of-Things (IoT) networks poses a significant challenge. Various approaches have been proposed to efficiently synchronize time in wireless sensor networks (WSNs) used in the IoT. However, these solutions typically involve extensive message exchanges to achieve synchronization, leading to notable communication and energy overheads. In this context, we introduce a clustering approach aimed at enhancing the Reference Broadcast Synchronization (RBS) protocol to suit large multihop IoT networks. This paper discusses existing cluster-based time synchronization methods and compares their effectiveness. Moreover, our proposed clustering approach seamlessly integrates existing time synchronization protocols, thereby enhancing both power efficiency and synchronization accuracy, which are specifically tailored for multihop IoT networks. To validate the effectiveness of our approach, we conducted emulations, which demonstrated a significant improvement in minimizing synchronization error by 78% compared to existing RBS methods, along with a 40% reduction in the power consumption of reference nodes. Overall, our proposed method yields satisfactory results with less overhead in scalable IoT networks.*

Keywords*:* *Clustering Algorithms, Internet of Things, Time Synchronization, Wireless Sensor Network, Network Topology*

Received: February 28, 2024; Received in revised form: April 18, 2024; Accepted: May 10, 2024

1. INTRODUCTION

The IoT describes a network of interconnected physical devices and other objects that are integrated with sensors, actuators, software and communicate data throughout the network. Time synchronization among IoT devices plays an important role in IoT applications such as smart grids, smart parking systems, health monitoring systems, mobile communications, environment monitoring systems, and distributed resource allocation [1-5]. For the optimal operation of IoT networks, precise time coordination is essential. Proper synchronization minimizes communication collisions, reduces retransmissions, and contributes to energy conservation. While the study of clock synchronization in wireless sensor networks has been performed for many years, IoT devices have some challenges. IoT devices use small components, which usually have limited battery power, constrained memory resources, low cost and less precise crystal oscillators, and low-power sensors and actuators compared to traditional WSN nodes.

To overcome issues of limited battery power, we proposed clustering approach. Due to limited battery power in IoT devices, the prime concern is to conserve power. With the help of clustering communication needs to be done within subset of nodes of the network, which in turn also reduces the distance of communication. Short distance communication reduces transmission delays as well as power consumption. There are less chances of node failure due to power conservation which in turn avoids synchronization error and improves synchronization accuracy.

These small components cause disparities in clock time between devices in a network. As prevalent clock sources in electronic devices, crystal oscillators furnish a reference frequency for timekeeping. Despite their widespread use, these oscillators are not flawless and introduce various constraints that compromise the precision and stability of time synchronization in IoT devices [6]. Crystal Oscillator Constraints are as follows:

- Crystal oscillators possess a defined frequency accuracy that may shift over time due to factors such as temperature changes. These inaccuracies can result in drift, leading to time discrepancies over prolonged periods [7, 8].
- Fluctuations in temperature can affect the oscillator's frequency, introducing variations in timekeeping. This sensitivity is particularly relevant in IoT devices deployed under diverse environmental conditions.
- The aging effect (gradual changes in frequency over time) can lead to a loss of accuracy in timekeeping, and this is especially important in applications where precise time synchronization is critical.
- Despite their overall power efficiency, crystal oscillators can still impact battery life in energy-constrained IoT devices, particularly for devices operating in remote or battery-powered scenarios.

In applications with strict synchronization requirements, the precision of crystal oscillators may not be sufficient.

To mitigate the impact of these limitations on time synchronization in IoT devices, various methods, such as the use of an external time server to recalibrate the device's internal clock, the implementation of algorithms that can compensate for the aging effect and temperature effects, and the implementation of mechanisms to stabilize temperature effects around crystal oscillators, need to be implemented [7].

Numerous clock synchronization approaches have been proposed for WSNs. The network time protocol (NTP) is widely utilized to synchronize computers over the internet, making it advantageous for application in this new context as well [9-11], its accuracy rarely meets the typical requirements of IoT applications and IoT networks. Moreover, NTP is not suitable for large mesh networks, as several nodes in mesh need to communicate using a gateway (or border router) with an NTP time server. Usually, a gateway is a high-traffic area that may cause delays [11, 12]. A drawback of NTP is its lack of secure time synchronization. To address this issue, the work presented in [13] introduces an NTP-based time synchronization method incorporating trust management and blockchain techniques. A simplified version of the NTP protocol (SNTP) [9] is commonly used in embedded systems. While the SNTP is used in computers with low processing power and in microcontrollers where accuracy is not an issue, it is helpful where scalability and low overheads are needed. It is not suitable for large mesh networks because it works only on high-speed networks such as Ethernet. Numerous clock synchronization protocols have been specifically devised for both wired and wireless networks, including the precision time protocol (PTP) designed for IEEE 802.11 networks. In a modified iteration of PTP, beacon frames are utilized to synchronize the mobility of access points (APs) with one another, as well as to broadcast timestamps acquired by the APs [14]. The study referenced in [15] investigates the use of PTP technology for time synchronization in Industrial IoT. However, their findings are derived from a methodological review of existing data, thus not offering conclusive evidence regarding the reliability of PTP in Industrial IoT. The efficacy of time synchronization hinges on various factors, such as hardware clock precision, sensor accuracy, and environmental influences. Tailored approaches for time synchronization in specific applications have also been introduced. For instance, a three-step method was developed in [16] to estimate clock skew and offsets, specifically for receiver-only based time synchronization in underwater applications. Another example is found in [17, 18], where a synchronized health monitoring system was described. This synchronization was accomplished through the utilization of high-precision external oscillators and GPS systems.

In recent years, various time synchronization protocols have been developed for wireless sensor networks that can work on IEEE 802.11. Many IoT applications have been developed over the IEEE 802.11 network, and they can deploy a time synchronization protocol for time accuracy. Time synchronization algorithms are primarily categorized into centralized and distributed synchronization algorithms. Centralized approaches utilize a reference node to synchronize all nodes within a network. Conversely, distributed algorithms are receiver–receiver-based synchronization methods in which no single reference node is employed. Section 2 provides an overview of existing algorithms in this context. However, not all these protocols provide accurate time synchronization in multihop networks, so providing a time synchronization protocol that is compatible with both single-hop and multihop networks is challenging. Our proposed cluster-based approach for multihop networks incorporates clustering methods to synchronize the whole network. This paper describes the existing RBS approach for time synchronization and proposes the use of clustering methods with RBS for time synchronization in IoT networks. The main goal of this work is to apply a clustering approach with RBS to provide time synchronization among all nodes (network-wide synchronization), particularly in multihop networks.

Achieving precision in time synchronization poses a significant challenge, particularly regarding comparing time information across network-wide nodes. Each node must assess its clock drift and skew based on the time received from a reference node. The accuracy of a clock is heavily influenced by the delays incurred in transmitting time information between locations, which consequently leads to synchronization errors.

Consequently, nodes further away from the reference node experience increased synchronization errors. Efficient synchronization routines necessitate minimizing the number of messages sent by each node and reducing energy consumption. However, in a multihop environment, achieving efficiency in terms of reducing power consumption and latency among nodes, correcting time values, and minimizing synchronization errors are particularly challenging. Additionally, considerations must be made for node failures and node mobility within the network.

Successfully achieving time synchronization in the IoT necessitates a meticulous examination of the tradeoffs between precision and diverse resource limitations. These constraints span energy consumption, communication overhead, scalability, latency, and robustness. Striking an appropriate balance customized to the precise needs of IoT deployment is pivotal for maximizing performance and efficiency. As seen from existing time synchronization algorithms [19-28], as discussed in section 2, while flooding the network with synchronization messages may offer high accuracy, it results in significant communication overhead. Minimizing this overhead while maintaining acceptable synchronization levels is vital, especially in resource-constrained IoT setups. Some synchronization methods excel in small-scale deployments but struggle to maintain accuracy as the network expands. Designing protocols that scale effectively while preserving accuracy is key for large IoT deployments. In latency-sensitive applications, minimizing synchronization latency may be prioritized over achieving perfect accuracy. Synchronization methods must consider the system's robustness against network disruptions and node failures. Trade-offs may arise between perfect synchronization and ensuring that the system can recover quickly from disruptions.

Section 2 elaborates on the related work conducted in the field of time synchronization. Section 3 describes the clustering techniques employed in both wireless sensor networks (WSNs) and Internet of Things (IoT) networks. The cluster-based time synchronization approach outlined in section 4 addresses the challenges associated with message overhead by enabling direct communication between nodes for clock corrections. This approach effectively manages node failures by periodically forming clusters at predefined resynchronization intervals. Additionally, it ensures network scalability during synchronization by incorporating new nodes into the cluster. Furthermore, the power consumption is minimized by reducing the latency at various stages of the synchronization process, including the send time, receive time, and propagation time.

2. RELATED WORK IN TIME SYNCHRONIZATION ALGORITHMS

Time synchronization stands as a focal point in the realm of wireless sensor networks and IoT networks, attracting widespread attention in research. Considerable research has been dedicated to the time synchronization of sensor nodes. Nonetheless, there remains an opportunity to refine existing time synchronization algorithms to effectively operate with IoT end devices, prioritizing low power consumption and heightened synchronization accuracy.

Traditional time synchronization algorithms typically follow a Sender–Receiver approach [19], where a root node serves as the time server, and other nodes synchronize with it. This method, often termed centralized time synchronization, has a drawback: if the root node is compromised, the entire network may suffer, resulting in incorrect clock values. Examples of such algorithms include the flooding time synchronization protocol (FTSP) [20], lightweight tree-based synchronization (LTS) [21], the timing synchronization protocol for sensor networks (TPSN) [22], and the flooding with clock speed agreement (FCSA) protocol proposed in [23], aimed at achieving skew synchronization among neighboring nodes. This protocol is tailored to mitigate synchronization errors that escalate with the number of hops in the FTSP.

In contrast, receiver–receiver-based algorithms [19] synchronize based on the arrival time of synchronization messages from other nodes in the network and use estimated offset values to correct their own clock time. However, this approach has limitations, such as high message complexity due to additional message exchanges between nodes and potential message collisions. The algorithms in this category include reference broadcast synchronization (RBS) [24] and time diffusion synchronization protocol (TDSP) [25].

In multihop scenarios, various cluster-based time synchronization approaches have been proposed. The methodologies outlined in references [26-28] employ a cluster-based approach to reduce synchronization errors by minimizing the average hop count from the root node. However, these methods are ill suited for networks with dynamic topologies. Moreover, effective mechanisms for handling node failures during synchronization and ensuring network scalability are lacking. Alternatively, other time synchronization approaches, as described in [29, 30], facilitate the construction of distributed networks and exhibit robustness to dynamic topologies and node failures. However, these algorithms suffer from a significant drawback in the form of packet collisions. This issue increases the overall message complexity of the network, impeding efficient data transmission and potentially leading to network congestion and reduced performance. The proposed C-sync [31], a clustering-based energy efficient decentralized time synchronization protocol, aims to achieve scalability by incorporating multiple reference nodes. However, the protocol's involvement of multiple reference nodes introduces message overhead, thereby increasing the time required to synchronize the entire network. To improve upon the traditional RBS algorithm, [32] introduced adaptive clock synchronization in sensor networks. While this solution seeks to alleviate the high overhead linked with flooding the network with reference packets, it also introduces tradeoffs such as latency, reliance on sensor nodes, synchronization reliability, and implementation complexity. In their work, [33] introduced a clustering-based hierarchical time synchronization method to facilitate multihop synchronization. This approach utilized long radio ranges and clustering to reduce average hop counts. However, the increased number of referenced messages overhead resulted in delays in the synchronization process. Additionally, the applied overhearing method, while effective in reducing hops, consumed more power, making it unsuitable for power-constrained IoT devices. Furthermore, the method did not support topology changes, posing limitations in dynamic network environments. The approach outlined in [34] introduces a multihop clustering mechanism for scalable IoT networks, with the objective of minimizing the number of Internet connections while maximizing the number of hops to its coordinator. However, it relies on Dijkstra's shortest path first algorithm to calculate the distances among all possible pairs of nodes, with a time complexity of $O(|N|(|N|\log|N|+|E|))$. Additionally, the complexity of obtaining clusters is $O(|N|2\log|N|)$, thereby contributing to an overall increase in the complexity of the clustering approach. In [35], a clock synchronization strategy based on precision time protocol (PTP), aimed at synchronizing clocks between IoT devices and the Cloud, which is interconnected within a distributed network framework, was proposed. Here, the Software as a Service (SaaS) cloud service is used to gather data for analysis and initiate corresponding actions on IoT devices. The approach described in [36] introduces the energy efficient clustering algorithm (EECA) for wireless sensor networks (WSNs). In this algorithm, clusters are formed based on the center of the sensing field, after which the synchronization process commences. Each node synchronizes with its respective cluster head (CH) within the network. In [37], a novel time synchronization method called cluster-based maximum consensus time synchronization (CMTS) was introduced. This method incorporates a rotational cluster head scheme. Synchronization is achieved by exchanging timestamp messages between cluster heads and cluster members and then computing the clock offset. An enhanced time synchronization approach for home automation systems has been proposed in [38], based on the Elastic Timer Protocol (ETP). This approach introduces synchronization overhead resulting from the dynamic adjustment of timer values and synchronization parameters. Work presented in [39] achieves time synchronization with heterogeneous technologies in IoT network based on Cross-Technology Communication technique (CTC). However, CTC introduce additional complexity and computational demands on devices impacts power consumption.

We delve into a detailed examination of existing strategies such as the RBS, FTSP, and TPSN in the context of multihop scenarios for time synchronization. This analysis serves to facilitate a comprehensive comparison with our proposed approach for time synchronization in multihop networks.

1. RBS in Multihop Network

The reference broadcast synchronization (RBS) algorithm operates on a receiver–receiver basis for time synchronization. In this method, a reference node broadcasts reference messages across the network. Neighborhood nodes record the timestamp of received broadcast messages and exchange their local time with other nodes in the network. Subsequently, all nodes calculate the average offset value and estimate their own clock value.

Fig. 1. [24] illustrates a scenario for a multihop network. In the depicted scenario, both Node A and Node B send synchronization pulses at times PA and PB, respectively. Receiver node 4 (R4) captures both sync pulses and forward the clock information from one neighborhood to another. Receivers R1 and R7 detect events at times E1R1 and E7R7, respectively. R4 leverages both A's and B's reference broadcasts to establish the best-fit line for adjusting clock values from R1 to R4 and from R4 to R7, respectively. However, this scheme necessitates a lengthy process for R4 to compute the correct time, leading to delays.

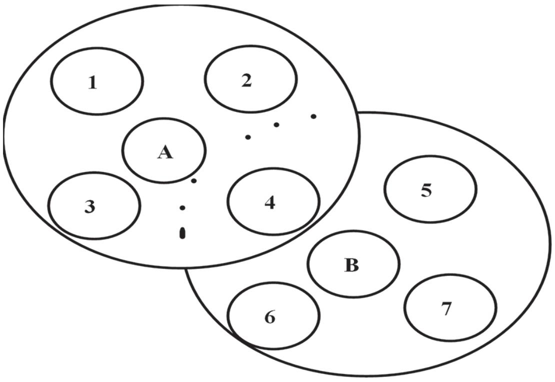


Fig. 1. RBS in the multihop network

Another drawback arises from implicit skew correction for all three nodes—R1, R4, and R7—during each time base conversion. Here, R4 listens to two sync messages and requires a series of timestamp conversions for clock skew calculation. If a node listens to more than two sync messages, this task becomes more challenging. Consequently, the RBS strategy does not adequately support large multihop networks for time synchronization, resulting in scalability issues.

2. FTSP in Multihop Network

The FTSP, on the other hand, utilizes a sender-receiver-based approach [20], supporting both single-hop and multihop networks for time synchronization. In this method, the root node transmits a sync message, and non-root nodes synchronize their clocks with their neighbors based on the root message. Each node, excluding the root node, utilizes timestamps from multiple neighbors to determine its local clock time and achieve synchronization. The synchronization process in a multihop FTSP relies on reference points established by broadcast messages periodically transmitted by the synchronization root node. However, this approach exhibits longer propagation times for leaf nodes in the network and may be susceptible to compromised nodes assuming the role of the root node and disseminating incorrect synchronization messages.

Another approach for multihop FTSP, as described in reference [30], accomplishes network synchronization without depending on an external time source. In this method, each node is allocated a unique identifier, and synchronization is attained through MAC layer timestamping. Clock skew estimation is performed by computing the average of multiple timestamp values, followed by the application of linear regression to estimate the clock offset. However, this approach entails increased timestamp overhead and is limited in handling small network traffic. Moreover, it demonstrates greater message complexity than does the RBS method.

3. The TPSN in the MultiHop Network

The TPSN employs a sender-receiver approach for time synchronization, comprising two phases: the establishment of a hierarchical topology followed by the synchronization phase. In the hierarchical topology phase, nodes at the ith level are connected with at least one node at the (i-1)th level. During the synchronization phase, child nodes synchronize with the root node at each level. Each pair of nodes is considered a root-child node, with the child node becoming the root for the subsequent node in the tree.

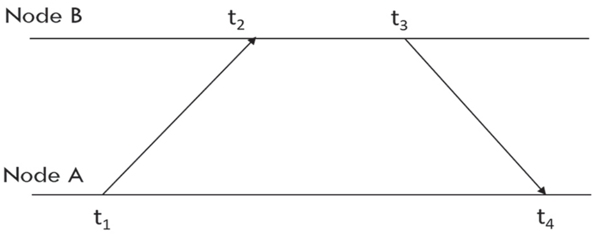


Fig. 2. TPSN sync message transmission

In Fig. 2, at time t_1, sender node A transmits a synchronization pulse packet to node B. Node B receives the packet at time t_2 and responds with an acknowledgment packet containing timestamp values t_1, t_2, and t_3. Node A acknowledges the receipt of the acknowledgment at time t_4. The clock offset is then calculated as $\Delta t = [(t_2 - t_1) - (t_4 - t_3)]/_2$, while the propagation delay is calculated as $d = [(t_2 - t_1) - (t_4 - t_3)]/2$.

In a multihop scenario, the TPSN adopts post facto synchronization, where nodes synchronize only as needed. Consequently, the receiver utilizes the TPSN to synchronize its clock after receiving a packet before forwarding it to the next hop. However, a drawback of this approach is that if any root node computes an incorrect offset during any point of the synchronization phase, this error will propagate down the tree. Furthermore, the TPSN transmits a large number of messages to synchronize a network, resulting in high data traffic.

Several aspects overlooked in the above approaches may hinder their implementation for high-level synchronization in multihop wireless networks. These aspects include ensuring quick network synchronization, which is particularly crucial in dense network environments where frequent synchronization is needed. Additionally, streamlining synchronization to minimize message overhead and enable multihop synchronization in a scalable manner, even when synchronization regions do not intersect, is essential. To address these challenges, we have implemented cluster-based and receiver–receiver-based approaches for time synchronization to support scalability and flexibility in multihop networks. The following section outlines clustering approaches applicable to multihop networks for time synchronization.

2.1. RELATED WORK FOR CLUSTERING IN WSNS AND THE IOT

Clustering offers a promising solution to address numerous challenges encountered in the IoT, including energy efficiency, scalability, and mobility. Its resemblance to wireless sensor networks (WSNs) makes it particularly advantageous for tackling these issues [40]. In cluster, a cluster head (CH) is a pivotal node serving as the central coordinator within a group of nodes. Cluster head plays a crucial role in organizing, managing, and optimizing the performance of a cluster of nodes.

In the implementation of time synchronization, the rapid exchange of synchronization messages among all network nodes without congestion is crucial. Clustering methods offer a solution by dividing the network into smaller regions, enabling simultaneous synchronization of message exchange among cluster nodes via cluster heads (CH). This approach enhances efficiency in terms of parallel processing, fault tolerance, and organizing dense networks effectively. Several clustering algorithms have been proposed to partition networks into smaller clusters based on criteria such as distance, link quality, and path.

- Clustering Methods in WSNs

In wireless sensor networks (WSNs), clustering algorithms are categorized as centralized or distributed approaches. Examples of clustering algorithms include the following:

In low-energy adaptive clustering hierarchy (LEACH) [41], nodes calculate their likelihood of becoming cluster heads (CHs) and broadcast them, with each node selecting the cluster requiring the least amount of communication energy to reach the CH. However, the CH distribution may not be uniform, leading to uneven energy dissipation. LEACH-C [42] selects a CH based on energy information and load balancing, with a cluster setup performed by the base station. While ensuring load balancing, this approach is less scalable and consumes more energy. The hybrid energy efficient distributed (HEED) method [43] selects a CH based on the residual energy and the energy required for intra-cluster communication. This ensures a uniform CH distribution and applies load balancing. The energy efficient clustering scheme (EECS) [44] selects a CH based on the maximum residual energy and minimum distance, extending the cluster formation approach of LEACH. However, CH deaths may occur due to congestion near the base station. The linked cluster algorithm (LCA) [45] selects the CH based on the highest ID among the node neighbors. Despite identity-based selection, nodes exhibit low energy efficiency, and these algorithms are designed for homogeneous networks and lack support for node mobility and data aggregation at the CH.

However, for IoT networks, which require support for data aggregation and mobility, these algorithms are not suitable, necessitating the development of new clustering strategies.

- Clustering Methods in IoT Networks

Efficient operation of IoT applications requires energy efficiency, low communication overhead, mobility support, data aggregation, and compatibility with heterogeneous environments. Several clustering algorithms address these requirements:

The heuristic clustering algorithm [46] forms clusters based on the number of neighboring nodes and residual energy. The node with the maximum residual energy is selected as the CH, facilitating one-hop communication but requiring re-clustering if the topology changes, leading to increased energy consumption. The graph-based clustering algorithm [47] uses graph theory to form clusters, selecting the vertex with the maximum degree and maximal residual energy as the CH. It supports node mobility with energy efficiency. The hybrid energy-aware clustered protocol for heterogeneous IoT [48] enhances node energy utilization and prolongs network lifetime. CH selection is based on various weighted election probabilities, such as residual energy. Cluster formation adjusts the number of CHs and the cluster length to optimize connectivity and energy utilization in multihop networks.

In summary, clustering approaches for IoT networks minimize communication overhead and maximize node connectivity, improving the efficiency of time synchronization algorithms while prolonging network lifetime and enhancing energy utilization.

3. IMPROVISED RBS ALGORITHM USING CLUSTER

Our research endeavors center on the development of time synchronization algorithms capable of withstanding significant differences in clock drift and offset, which is particularly relevant for extensive IoT network deployments. It has been noted that minimizing communication distance aids in reducing clock drift and offset. To achieve this, before initiating the synchronization process, an RSSI-based clustering approach is employed to partition the large network into smaller clusters. This allows reference nodes to communicate directly with their nearest cluster-head nodes, which then handle synchronization among the nodes within their respective clusters. This localized approach minimizes the energy expenditure required for time synchronization across the entire network. By reducing communication distance and overhead, this approach reduces power consumption and effectively reduces delays that can occur at various stages of the synchronization process, thereby minimizing synchronization errors. In this scheme, each cluster-head node serves as the reference node for its cluster nodes.

In the implementation of time synchronization, it is imperative to ensure rapid dissemination of synchronization messages across all network nodes without causing congestion. Various clustering algorithms have been suggested for partitioning networks into smaller clusters based on diverse criteria. Drawing from extensive surveying and identifying research gaps, we advocate for a clustering approach tailored for IoT network applications based on the received signal strength indicator (RSSI). This approach aims to minimize power consumption, enhance the packet reception ratio, and enable data aggregation at cluster heads within heterogeneous IoT networks.

To address the challenges posed by dynamic topology alterations, node failures, and scalability issues, we devised a strategy wherein the network undergoes re-clustering after each synchronization interval. This adaptive approach ensures resilience to changing network conditions while bolstering the system's robustness and scalability. Our time synchronization methodology for multihop IoT networks is structured into two phases: Phase-1 involves cluster formation through cluster-head selection, while Phase-2 encompasses the synchronization method between nodes. In the subsequent sections, we elaborate on our novel cluster approach followed by the time synchronization methodology tailored for multihop IoT networks.

- Cluster Algorithm for IoT network

  * Phase-1: Cluster Formation

Cluster Head Selection: The selection of cluster heads (CHs) is achieved by analyzing the radio signal strength indicator (RSSI) of each node within the network. This entails measuring the power present in signals transmitted by each node. By utilizing the received signals from neighboring nodes, each node maintains a count of the number of neighboring nodes (denoted as 'm') covered by the RSSI. The node with the highest value of 'm' is designated the CH. This selection process continues until all nodes are encompassed within the network.

Clustering: During this phase, each CH identifies its cluster nodes based on a predefined RSSI threshold value. Nodes are compared with the threshold value, and if their RSSI value is lower than the threshold, they are included as cluster nodes for that particular cluster head. This process iterates until all nodes are assigned to one of the clusters within the network. The algorithm outlining the cluster formation process is depicted in Algorithm 1.

Algorithm 1. Cluster Formation

```
Input: ($N^i$ is the $i^{th}$ node in the network), $L_i$ = list of neighbor nodes, $M$= Maximum no. of nodes covered in one cluster, $m$= No. of clusters.
Output: Create $m$ no. of clusters.
Initialize all the nodes $N_i$, where $i$=1,2, 3…$n$. in a network.
For each node $N^i$
        $RN_i \leftarrow RSSI\ (N_i)$, $i=i+1$
        (Sort $RN_i$, choose nearest node in $L_i$)
For each node $RN_i$
        If $RN_i < RN_i+1$
        $L_i < RN_i$
End.
For each node $N^i$
//From the sorted list $L_i$, the nodes which covers maximum no. of Neighbor nodes are declared as cluster heads.
//Choose other nodes in the range of $CH_i$ as member nodes in cluster $C_i$, where $i \in N_i$. Such that $M<=m$, where $m<=N/2$
In case of conflict, where the node may fall in more than one clusters, then the node joins the cluster having minimum distance with the $CH_i$.
Repeat steps after each Synchronization Interval $S_i$.
```

- Proposed Cluster-based Time Synchronization Approach for Multihop IoT

Time synchronization commences after the completion of cluster formation as outlined in Algorithm 1 during Phase-1. Phase-2 encompasses the synchronization process among the nodes and is logically subdivided into two stages.

Phase 2(a) entails Algorithm 2(a), synchronization among cluster head nodes with a designated reference node, referred to as inter-cluster synchronization. This phase is crucial for rectifying offset discrepancies by initially addressing the time differential between the primary synchronized reference node and the cluster heads of other clusters within the network. To achieve this, the reference node transmits beacon messages to the cluster heads, recording the timestamp of these beacon messages from the synchronized reference node. Subsequently, all nodes exchange their local time information with one another and estimate their respective clock offsets.

Phase 2(b) involves synchronization among cluster nodes as in Algorithm 2(b) with their respective cluster heads, termed as intra-cluster synchronization.

Algorithm 2(a). Inter-Cluster Synchronization

```
Assumption: Cluster formed using algorithm-1
Initialization:  S- Reference node
$CH_i$- $i^{th}$ Cluster head node, $C_i$ - $i^{th}$ Cluster node
$LT_i$ - $i^{th}$ local timestamp of $i^{th}$ cluster head node.
$OCH_{ij}$ - Clock Offset between nodes $i$ and $j$
$CHT_i$ – Adjusted new clock value
Reference Node $S$: broadcast () //Broadcast Beacon
For each $CH_i$
        Call broadcast_receive()  // $CH$ Receive broadcast message
        $LT_i$ = Clock_time() //Record local timestamp of received message
        Call unicast ($LT_i$, $CH_j$) // Send LTi to other $CH_j$ nodes where $i≠j$.
End
For each $CH_i$
        $OCH_{ij} = \frac{\sum LT_i - LT_j}{n}$        (1)
//Compute clock offset at each cluster head node. Where, $n$ = no. of received timestamp from neighbor cluster heads $CH_j$. $m<=N/2$, and $M$ depends on number of nodes in the network.
        $CHT_i = OCH_{ij} + LT_i$        (2)
// Compute new clock value $CHT_i$, and synchronize with reference node $S$.
End.
```

Algorithm 2(b). Intra-Cluster Synchronization

```
//In continuation of algorithm 2(a)
For each node $CH_i$
        Call multicast ($CHT_i$, $C_i$)
// $CH_i$ nodes multicast their updated time value $CHT_i$ to their cluster nodes $C_i$
        Call multicast_receive()
        $CT_i$ = Clock_time()
// Each $C_i$ records timestamp of received message as $CT_i$ and unicast to other nodes $C_j$, ($i≠j$)
End
For each $C_i$ node
        Call unicast ($CT_i$, $C_j$) // Send $CT_i$ to other $C_j$ nodes where $i≠j$.
// Compute clock offset
        $OC_{ij} = \frac{\sum CT_i - CT_j}{CN}$        (3)
```

// where CN = No. of nodes within cluster. And OC_{ij} is an offset between two cluster nodes in one cluster.

// Cluster node computes new clock value

$$TC_i = OC_{ij} + CT_i \tag{4}$$

// Where, TC_i is the new adjusted clock time of ith cluster node in one cluster

End.

- Key Features of the Proposed Cluster-Based Algorithm

Minimizes communication distance for synchronization message transmission, leading to potential reductions in power consumption and synchronization errors. Decreases collision occurrences owing to smaller cluster sizes and reduced message overhead. Enhances network reliability and scalability through the implementation of cluster-based time synchronization. Reduces communication distance for sync-message transmission, thereby lowering the power consumption of the reference node and the network overall. Improves the probability of packet reception ratio (PRR) for individual nodes.

4. IMPLEMENTATION AND RESULTS

4.1. IMPLEMENTATION

The time synchronization algorithm for a multihop IoT network is implemented using Contiki OS with the Cooja emulator [49]. Initially, all nodes are initialized with random startup times. The proposed cluster-based algorithm is compared with the existing RBS algorithm in terms of synchronization accuracy and power consumption on the same platform. Emulation parameters and configurations are summarized in Table 1.

Table 1. Emulation Parameters on Cooja

Parameters	Values
Number of Nodes	6 / 12 / 20
Type of Network	Multihop
Emulation Time	15min
Synchronization Interval	30 s
OS	Contiki
Topology	Random
Radio	CC2420 (2.4 GHz)
MAC / RDC layer Protocol	CSMA with ContikiMAC
Network Stack	Rime
Radio Medium	UDGM
Channel Check Rate	8Hz
Mote Type	Tmote Sky

To establish a multihop environment, the first experiment involves generating results with a random topology consisting of 10 nodes. Node id-1 is designated as the reference node for other nodes in the network. Following Algorithm 1, all nodes except node id-1 calculate their RSSI values and identify the number of nodes within their communication range. After selecting cluster heads in the network and forming clusters, the reference node begins broadcasting beacons throughout the network. Only cluster head (CH) nodes receive these beacons and exchange their local time of reception with other CH nodes in the network, as outlined in Algorithm 2, for synchronization. After inter-cluster synchronization of CHs with the reference node, cluster heads proceed with intra-cluster synchronization within their respective clusters.

4.2. RESULTS

To assess the performance of the proposed cluster-based algorithm in terms of time synchronization error and power consumption, it is implemented on different platforms. Specifically, the proposed cluster-based multihop RBS algorithm is executed on Sky motes [50], with configurations detailed in Table 1. Power consumption is evaluated at each node in the network.

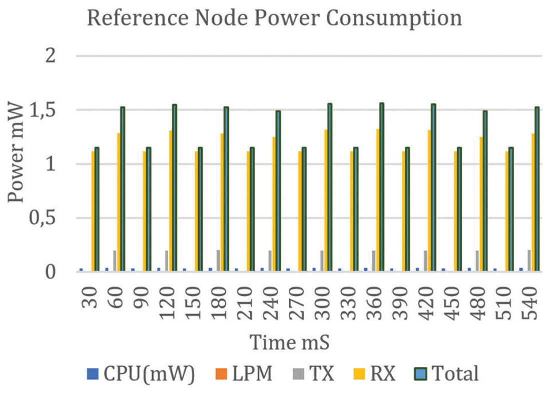


Fig. 3(a). Power consumption of reference node in proposed cluster-based RBS algorithm for multihop network

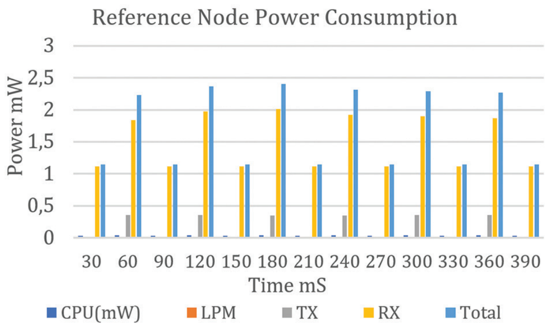


Fig. 3(b). Power consumption of reference node in RBS algorithm for multihop network

The sink node (Reference node) broadcasts beacon messages at specified regular intervals to synchronize cluster heads. The algorithm is tested under various scenarios, including random node startup times and cluster head node failures, to ensure continued network connectivity, wherein remaining nodes reform the connected network and reassign cluster heads as per Phase-1. The results, depicted in Fig. 3(a), illustrate the power consumption of the reference node in the proposed cluster-based RBS algorithm, while Fig. 3(b) displays the power consumption of the reference node in the RBS algorithm. Power consumption is assessed using the power tracer tool within the Cooja emulator. The formula utilized for calculating power consumption at each node is as follows [50].

$$\mathrm{CPU}_{\mathrm{power}} = \frac{(\mathrm{Energest\_Value} \times \mathrm{Current} \times \mathrm{Voltage})}{(\mathrm{Number\ of\ ticks\ per\ second} \times \mathrm{Runtime})} \qquad (5)$$

Power consumption, data gathering, and analysis are conducted at various stages of the node lifetime, including transmission power, receiving power, CPU power, and during low power mode. To facilitate a fair comparison between both algorithms, identical configurations (as per Table 1) are applied to assess performance. Graphical analysis reveals that the reference node's power consumption in the RBS algorithm is approximately 2.5mW, which is higher compared to the 1.5mW power consumption observed for the proposed cluster-based algorithm. Estimated percentage of reduction in power consumption using proposed cluster-based approach as follows:

$$\%\mathrm{Reduction} = \frac{\mathrm{Power\ Consumption}_{\mathrm{RBS}} - \mathrm{Power\ Consumption}_{\mathrm{Cluster\ based\_RBS}}}{\mathrm{Power\ Consumption}_{\mathrm{RBS}}} \times (6)$$

$$100\ \%$$

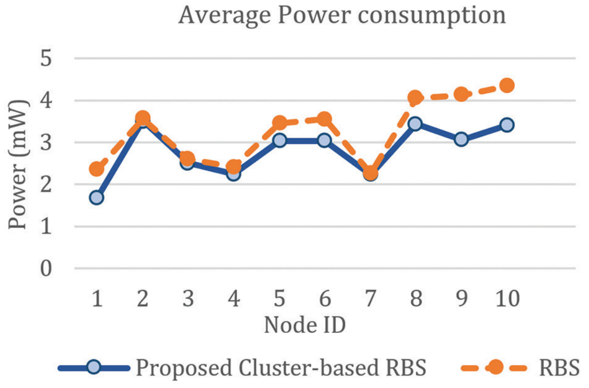


Fig. 4. Average power consumption at each node of RBS and proposed Cluster-based Improvised RBS multihop IoT network

Fig. 4 illustrates the average power consumption at each node within the multihop network. It is evident that the overall power consumption of the multihop network with the RBS algorithm is approximately greater than or equal to 4.5 mW, whereas with the proposed cluster-based approach, it is approximately less than 3.5 mW. This highlights a notable enhancement in power efficiency with the proposed cluster-based approach for time synchronization. Additionally, it is observed that nodes 2 and 3 exhibit nearly identical power consumption for both algorithms. This similarity arises because both nodes fall within the interference range of each other and of node 1, necessitating transmission to more than one cluster, resulting in increased power consumption.

The subsequent step involves calculating clock offset and clock skew to determine the synchronization error between two nodes. Clock offset is estimated using equation (1) for inter-cluster nodes and equation (3) for intra-cluster nodes. Synchronization error is assessed using the linear regression method, where the least squares method is employed to predict the nearest correct time value, minimizing the error sum of squares as per the principle of least squares method [51]. The following formula is utilized to calculate predicated time say P_t and Synchronization Error say S_e for this purpose:

$$P_t = \frac{((x-\bar{x})\times(y-\bar{y})}{(x-\bar{x})^2} \times x + \delta \qquad (7)$$

Where x denotes the timestamp value of the sent beacon message, y represents the timestamp value of the received beacon message, and δ denotes the clock skew, representing the difference between two clock frequencies.

$$S_e = A_t - P_t \qquad (8)$$

Where A_t denoted Actual Received Time. After determining the absolute synchronization error between two nodes, the subsequent step involves calculating the delta error. The proposed cluster-based approach effectively minimizes the overall power consumption of each node. Furthermore, using equation (6) the average power consumption of the reference node has been notably reduced by 40%, consequently improving the overall performance and lifetime of the network. Emulation results also indicate approximately 30% improvement in minimizing the average power consumption of each node in the network, apart from the reference node.

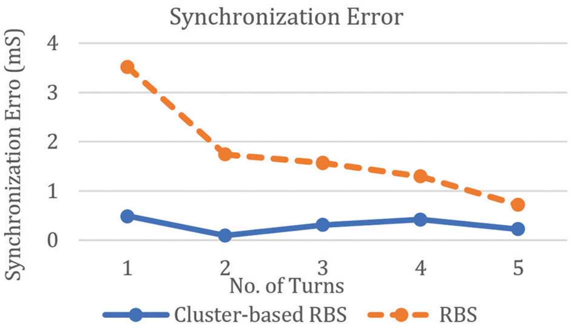


Fig. 5. Synchronization Error in RBS and Cluster-based RBS

Fig. 5 depict the synchronization error graphs for RBS and RBS with clustering, calculated using the linear regression method. In the case of RBS, the synchronization error ranges from 0.5 mS to 4 mS. However, with the incorporation of clustering, there is a reduction in the synchronization error rate, ranging from 0.5 mS to 0 mS. To mitigate the overall synchronization error, the re-synchronization interval is determined by assessing the absolute error between the estimated offset and the corrected offset at each offset synchronization point.

Consequently, the overall synchronization error gradually diminishes, tending toward zero.

Compared to the existing RBS approach, the proposed cluster-based approach demonstrates an average 78% improvement in minimizing node synchronization error.

4.3. IMPACT OF DIFFERENT TOPOLOGIES ON TIME SYNCHRONIZATION

Both synchronization algorithms, namely RBS and the proposed cluster-based approach, underwent testing using an experimental setup outlined in Table 1. Three distinct network topologies - Random, Ellipse, and Linear - were tested for each synchronization algorithm, specifically designed for multihop networks. Tables 2, 3, and 4 present the performance of both approaches for time synchronization across random, ellipse, and linear topologies, respectively. The evaluation includes synchronization error and power consumption for various numbers of nodes within the system. An Average Synchronization Error, measured in milliseconds, indicates the level of synchronization accuracy for each configuration, with smaller values indicating preferable synchronization. Standard Deviation measures dispersion in synchronization errors; lower values indicate consistent errors, ensuring system stability [51]. Here, Standard Error helps to understand the likely range within which the true mean synchronization error falls.

Table 2. Synchronization Error and Power Consumption for Random Topology

Synchronization Approach	RBS	Proposed Approach (with Cluster)	RBS	Proposed Approach (with Cluster)	RBS	Proposed Approach (with Cluster)
No. of Nodes	6		12		20	
Average (mS) (Synchronization Error)	1.53	0.36	1.46	0.3	3.16	0.64
Standard Deviation	1.30	0.28	0.98	0.15	1.72	0.30
Standard Error of (Synchronization Error) (mS)	1.12	0.39	0.44	0.07	0.77	0.13
Average Power Consumption (mW)	2.33	1.27	3.3	1.00	1.15	2.8

Table 3. Synchronization Error and Power Consumption for Ellipse Topology

Synchronization Approach	RBS	Proposed Approach (with Cluster)	RBS	Proposed Approach (with Cluster)	RBS	Proposed Approach (with Cluster)
No. of Nodes	6		12		20	
Average (mS) (Synchronization Error)	2.52	1.96	9.12	5.32	9.6	7.91
Standard Deviation	1.29	0.88	6.18	2.96	6.88	5.6
Mean Synchronization Error (Standard Error) (mS)	1.02	0.39	2.77	1.32	3.07	2.5
Average Power Consumption (mW)	2.76	1.98	3.24	2.46	4.6	2.1

Table 4. Synchronization Error and Power Consumption for Linear Topology

Synchronization Approach	RBS	Proposed Approach (with Cluster)	RBS	Proposed Approach (with Cluster)	RBS	Proposed Approach (with Cluster)
Number of Nodes	6		12		20	
Average (mS) (Synchronization Error)	3.12	2.48	9.12	7.20	11.04	8.08
Standard Deviation	2.87	2.59	6.81	3.03	9.71	7.56
Mean Synchronization Error (Standard Error) (mS)	1.28	1.16	3.05	1.36	4.34	3.38
Average Power Consumption (mW)	1.73	1.4	3.28	2.93	3.36	2.11

The results displayed in Tables 2, 3, and 4 show that the proposed clustering-based approach generally outperforms the baseline RBS approach in terms of synchronization error. Across all node configurations (6, 12, and 20 nodes), the proposed approach consistently exhibits a lower average synchronization error. Moreover, the standard deviation in the proposed cluster-based approach is generally lower, indicating more consistent and stable synchronization errors across different trials. In summary, these results indicate that the proposed cluster-based synchronization approach achieves lower synchronization errors and greater power efficiency across diverse numbers of nodes in various topologies than does the baseline RBS approach. Additionally, the results suggest that the random topology yields optimized results compared to the ellipse and linear topologies, which represent the worst-case scenarios.

4.4. IMPACT OF HOP DISTANCE ON TIME SYNCHRONIZATION

Table 5 illustrates the performance of the proposed cluster-based RBS multihop algorithm implemented using random and linear topologies.

Table 5. Results of synchronization error for the implementation of the proposed cluster-based RBS for multiple hops using random and linear topologies (absolute values)

Hop Distance	Random Topology	Linear Topology
	Average Error (mS)	Average Error (mS)
1-hop	0.33	0.24
2-hop	0.47	9.48
3-hop	0.93	10.36
4-hop	1.29	8.72
5-hop	1.45	8.75

In the linear topology implementation, each cluster head has only one cluster node within its range, representing a worst-case scenario for synchronization accuracy at each node. Conversely, the random topology is considered the best-case scenario for evaluating performance based on the number of hops in a multihop network.

As evidenced by the smaller variations in average error values depicted in Fig. 6, in the linear topology, there are greater variations in average errors, particularly for larger hop distances. Conversely, the random topology demonstrates relatively stable and consistent synchronization performance across different hop distances. Consequently, the proposed cluster-based RBS algorithm for multihop networks utilizing a random topology outperforms other implementations.

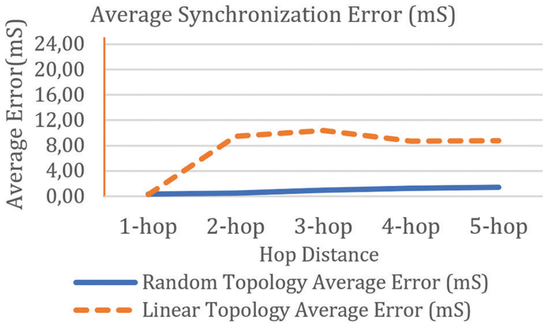


Fig. 6. Synchronization error in the proposed cluster-based RBS with increasing number of hop distances

4.5. APPLICATIONS OF PROPOSED WORK

Our research findings offer valuable applications for real-time water quality monitoring and controlling system [52] requires accurate time synchronization to enable solenoid valves to react promptly to sensor data by opening and closing as needed. In this system, nodes exchange beacon messages with adjacent nodes, periodically adjusting their clocks to maintain synchronization. Without effective time synchronization, there is a potential risk of distributing contaminated water which cannot be use for drinking purpose. Our approach minimizes communication distance, thereby conserving power which prevents node failure and ensuring reliable real-time operations. Similarly, our research findings are relevant to Cyber-Physical Systems (CPS) [53] requiring time synchronization, ensuring precise data timestamping among system components and maintains event ordering. Our synchronization method guarantees accurate timestamping and facilitates sequence of event by reducing communication overhead between server node and client nodes. In Patient's Real-time Health Monitoring System, precise time synchronization facilitates real-time recording and transmission of patient data, including vital signs and medication schedules. Synchronized data transmission minimizes network congestion and conserves energy, extending the battery life of wearable medical devices employed in patient monitoring. Furthermore, timely and accurate data transmission supported by time synchronization enhances patient safety by enabling healthcare providers to promptly identify and address critical situations, thereby reducing the likelihood of medical errors and adverse outcomes.

Having implemented our proposed approach using the Cooja emulator with sky motes, we observed consistent results comparable to real-world sky mote scenarios, with minimal discrepancies.

5. CONCLUSION

While numerous time synchronization algorithms exist for wireless sensor networks (WSNs), there is no elaborate work done on IoT networks, where explicit challenge is power conservation. They often prove less effective in IoT networks due to constraints such as low power availability, limited memory, and unreliable crystal clocks inherent in IoT devices. In response to these challenges, a novel approach to time synchronization in IoT networks has been proposed. By implementing RSSI-based clustering method to segment the network into smaller regions. By leveraging RSSI values, the proposed algorithm aims to minimize communication distance, reduce power consumption, enhance the packet reception ratio, and enhance synchronization accuracy. The incorporation of an adaptive re-clustering strategy after each synchronization interval is another novel aspect of this work. This adaptive approach ensures power conservation and thus, there is less chances of node failure, resilience to changing network conditions, bolstering the system's robustness and scalability. The proposed algorithm offers flexibility and adaptability crucial for IoT deployments. The cluster-based approach aims to minimize the power consumption of the reference node and the overall network while also reducing synchronization errors to ensure accurate event ordering.

A comprehensive overview of time synchronization approaches for multihop networks has been presented. Furthermore, we analyzed three key performance metrics in the multihop IoT network for comparison: power consumption, synchronization error, and scalability. The study demonstrates the effectiveness and robustness of the cluster-based approach in different deployment scenarios such as diverse network topologies (random, ellipse, linear) and hop distances. The emulation results demonstrate that the proposed cluster-based approach has minimized the power consumption by 40% of the reference node and 30% of the overall network. A significant 78% reduction in synchronization error is achieved. Building upon reference-broadcast synchronization principles, this study explores an alternative method for synchronizing multihop networks, offering enhanced precision, flexibility, and resource efficiency compared to traditional algorithms. Through the cluster-based RBS approach, the communication distance between nodes involved in time synchronization is minimized, resulting in reduced propagation delay and synchronization errors within the cluster nodes.

6. REFERENCES:

[1] N. Dalwadi, M. Padole, "An Insight into Time Synchronization Algorithms in IoT", Data, Engineering and Applications, Springer, 2019, pp. 285-296.

[2] A. Zanella, N. Bui, A. Castellani, L. Vangelista, M. Zorzi, "Internet of Things for Smart Cities", IEEE Internet of Things Journal, Vol. 1, No. 1, 2014, pp. 22-32.

[3] X. Liu, Z. Qin, Y. Gao, J. A. McCann, "Resource Allocation in Wireless Powered IoT Networks", IEEE Internet of Things Journal, Vol. 6, No. 3, 2019, pp. 4935-4945.

[4] G. Xu, W. Shen, X. Wang, "Applications of Wireless Sensor Networks in Marine Environment Monitoring: A survey", Sensors, Vol. 14, No. 9, 2014, pp. 16932-16954.

[5] E. Xu, Z. Ding, S. Dasgupta, "Target Tracking and Mobile Sensor Navigation in Wireless Sensor Networks", IEEE Transactions on Mobile Computing, Vol. 12, No. 1, 2013, pp. 177-186.

[6] T. Schmid, Z. Charbiwala, J. Friedman, Y. Cho, M. Srivastava, "Exploiting Manufacturing Variations for Compensating Environment-induced Clock Drift in Time Synchronization", Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS, Annapolis, MD, USA, 2-6 June 2008, pp. 97-108.

[7] F. Tirado-Andrés, A. Araujo, "Performance of Clock Sources and their Influence on Time Synchronization in Wireless Sensor Networks", International Journal of Distributed Sensor Networks, Vol. 15, No. 9, 2019.

[8] E. Coca, V. Popa, "A Practical Solution for Time Synchronization in Wireless Sensor Networks", Advances in Electrical and Computer Engineering, Vol. 12, No. 4, 2012, pp. 57-62.

[9] D. Mills, "Internet Time Synchronization: The Network Time Protocol", IEEE Transactions on Communications, Vol. 39, No. 10, 1991, pp. 1482-1493.

[10] D. Mills, U. Delaware, J. Martin, J. Burbank, W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", IETF, RFC 5905, June 2010, http://www.rfceditor.org/rfc/rfc5905.txt (accessed: 2024)

[11] A. N. Novick, M. A. Lombardi, "Practical Limitations of NTP Time Transfer", Proceedings of the Joint Conference of the IEEE International Frequency Control Symposium & the European Frequency and Time Forum, Denver, USA, 2015, pp. 570-574.

[12] J. Elson, K. Romer, "Wireless Sensor Networks: A New Regime for Time Synchronization", Proceedings of the First Workshop on Hot Topics In Networks (HotNets-I), Princeton, NJ, USA, 28-29 October 2002.

[13] K. Fan, Z. Shi, R. Su, Y. Bai, P. Huang, K. Zhang, H. Li, Y. Yang, "Blockchain-Based Trust Management for Verifiable Time Synchronization Service in IoT", Peer-to-Peer Networking and Applications, Vol. 15, No. 2, 2022, pp. 1152-1162.

[14] A. Mahmood, G. Gaderer, H. Trsek, S. Schwalowsky, N. Kero, "Toward high accuracy in IEEE 802.11 based clock synchronization using PTP", Proceedings of the International IEEE Symposium on Precision Clock Synchronization for Measurement Control and Communication, Munich, Germany 2011, pp. 13-18.

[15] K. Balakrishnan, R. Dhanalakshmi, B. B. Sinha, R. Gopalakrishnan, "Clock Synchronization in Industrial Internet of Things and Potential Works in Precision Time Protocol: Review, Challenges and Future Directions", International Journal of Cognitive Computing in Engineering, Vol. 4, 2023, pp. 205-219.

[16] G. Liu, S. Yan, L. Mao, "Receiver-Only Based Time Synchronization Under Exponential Delays in Un-

derwater Wireless Sensor Networks", IEEE Internet of Things Journal, Vol. 7, No. 10, 2020, pp. 9995-10009.

[17] M. AbdelRaheem, M. Hassan, U.S. Mohammed, A. A. Nassr, "Design and Implementation of a Synchronized IoT-based Structural Health Monitoring System", Internet of Things, Vol. 20, 2022, p. 100639.

[18] G. Cena, I. Bertolotti, S. Scanzio, A. Valenzano, C. Zunino, "Synchronize your Watches: Part I: General-Purpose Solutions for Distributed Real-Time Control", IEEE Industrial Electronics Magazine, Vol. 7, No. 1, 2013, pp. 18-29.

[19] Y. Wu, Q. Chaudhari, E. Serpedin, "Clock Synchronization of Wireless Sensor Networks", IEEE Signal Processing Magazine, Vol. 28, No. 1, 2011, pp. 124-138.

[20] M. Maroti, B. Kusy, G. Simon, A. Ledeczi, "The Flooding Time Synchronization Protocol", Proceedings of the 2nd International Conference on Embedded Networked Sensor Systems, New York, NY, USA, 2004, pp. 39-49.

[21] J. V Greunen, J. Rabaey, "Lightweight Time Synchronization for Sensor Networks", Proceedings of the 2nd ACM International Workshop on Wireless Sensor Networks and Applications, New York, NY, USA, September 2003.

[22] S. Ganeriwal, R. Kumar, M. B. Srivastava, "Timing-sync Protocol for Sensor Networks", Proceedings of the 1st International Conference on Embedded Networked Sensor Systems, Los Angeles, CA, USA, 5-7 November 2003, pp. 138-149.

[23] K. Yildirim, A. Kantarci, "Time Synchronization based on Slow-Flooding in Wireless Sensor Networks", IEEE Transactions on Parallel and Distributed Systems, Vol. 25, No. 1, 2014, pp. 244-253.

[24] J. Elson, L. Girod, D. Estrin, "Fine-Grained Network Time Synchronization using Reference Broadcasts", ACM SIGOPS Operating Systems Review, Vol. 36, No. SI, 2002, pp. 147-163.

[25] W. Su, I. F. Akyildiz, "Time Diffusion Synchronization Protocol for Wireless Sensor Networks", IEEE/ACM Transactions on Networking, Vol. 13, No. 2, 2005, pp. 384-397.

[26] H. Kim, D. Kim, and S.-e. Yoo, "Cluster-Based Hierarchical Time Synchronization for Multihop Wireless Sensor Networks", Proceedings of 20th International Conference on Advanced Information Networking and Applications, Vienna, Austria, April 2006, pp. 318-322.

[27] B. J. Choi, H. Liang, X. Shen, W. Zhuang, "DCS: Distributed Asynchronous Clock Synchronization in Delay Tolerant Networks", IEEE Transactions on Parallel and Distributed Systems, Vol. 23, No. 3, 2012, pp. 491-504.

[28] M. Leng, Y. C. Wu, "Distributed Clock Synchronization for Wireless Sensor Networks using Belief Propagation", IEEE Transactions on Signal Processing, Vol. 59, No. 11, 2011, pp. 5404-5414.

[29] M. Akar, R. Shorten, "Distributed Probabilistic Synchronization Algorithms for Communication Networks", IEEE Transactions on Automatic Control, Vol. 53, No. 1, 2008, pp. 389-393.

[30] M. Maroti, B. Kusy, G. Simon, A. Ledeczi, "Robust Multi-Hop Time Synchronization in Sensor Networks", Proceedings of the International Conference on Wireless Networks, Las Vegas, NV, USA, 21-24 June 2004, pp. 454-460.

[31] N. Shivaraman, P. Schuster, S. Ramanathan, A. Easwaran, S. Steinhorst, "Cluster-Based Network Time Synchronization for Resilience with Energy Efficiency", Proceedings of the IEEE Real-Time Systems Symposium, Dortmund, Germany, 7-10 December 2021, pp. 149-161.

[32] S. Palchaudhuri, A. K. Saha, D. B. Johnsin, "Adaptive Clock Synchronization in Sensor Networks", Proceedings of the Third International Symposium on Information Processing in Sensor Networks, Berkeley, CA, USA, 26-27 April 2004, pp. 340-348.

[33] H. Kim, D. Kim, S. Yoo, "Cluster-Based Hierarchical Time Synchronization for Multihop Wireless Sensor Networks", Proceedings of the 20th International Conference on Advanced Information Networking and Applications, Vienna, Austria, 18-20 April 2006.

[34] Y. Sung, S. Lee, M. Lee, "A Multi-Hop Clustering Mechanism for Scalable IoT Networks", Sensors, Vol. 18, No. 4, 2018, p. 961.

[35] R. Jha, P. Gupta, “Clock Synchronization in IoT Network Using Cloud Computing”, Wireless Personal Communications, Vol. 97, 2017, pp. 6469-6481.

[36] G. Gautam, N. Chand, “A Novel Cluster Based Time Synchronization Technique for Wireless Sensor Networks”, Wireless Sensor Network, Vol. 9, No. 5, 2017, pp. 145-165.

[37] Z. Wang, P. Zeng, M. Zhou, D. Li, J. Wang, “Cluster-Based Maximum Consensus Time Synchronization for Industrial Wireless Sensor Networks”, Sensors, Vol. 17, No. 1, 2017, p. 141.

[38] K. Mehta, Y. Kumar, A. Aayushi, "Enhancing Time Synchronization for Home Automation Systems", ECS Transactions, Vol. 107, No. 1, 2022, pp. 6197-6208.

[39] D. Gao, Y. Liu, B. Hu, L. Wang, W. Chen, Y. Chen, T. He, “Time Synchronization based on Cross-Technology Communication for IoT Networks”, IEEE Internet of Things Journal, Vol. 10, No. 22, 2023, pp. 19753-19764.

[40] S. Sholla, S. Kaur, G. Rasool, R. Naaz Mir and M. A. Chishti, “Clustering Internet of Things: A Review”, Journal of Science and Technology: Issue on Information and Communications Technology, Vol. 3, No. 2, 2017, pp. 21-27.

[41] W. R. Heinzelman, A. Chandrakasan, H. Balakrishnan, “Energy- Efficient Communication Protocol for Wireless Microsensor Networks”, Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, Maui, HI, USA, 7 January 2000.

[42] W. Xinhua, S. Wang, “Performance Comparison of LEACH and LEACH-C Protocols by NS2”, Proceedings of the 9th International Symposium on Distributed Computing and Applications to Business Engineering and Science, Hong Kong, China, 10-12 August 2010, pp. 254-258.

[43] O. Younis, S. Fahmy, “HEED: A Hybrid, Energy-Efficient, Distributed Clustering Approach for Ad Hoc Sensor Networks”, IEEE Transactions on Mobile Computing, Vol. 3, No. 4, 2004, pp. 366-379.

[44] M. Ye, C. Li, G. Chen, J. Wu, “EECS: An Energy Efficient Clustering Scheme in Wireless Sensor Networks”, Proceedings of the 24th IEEE International Performance, Computing, and Communications Conference, Phoenix, AZ, USA, 7-9 April 2005, pp. 535-540.

[45] P. Kumarawadu, D. J. Dechene, M. Luccini, A. Sauer, “Algorithms for Node Clustering in Wireless Sensor Networks: A Survey”, Proceedings of the 4th International Conference on Information and Automation for Sustainability, Colombo, Sri Lanka, 12-14 December 2008, pp. 295-300.

[46] J. S. Kumar, M. A. Zaveri,” Clustering Approaches for Pragmatic Two-Layer IoT Architecture”, Wireless Communications and Mobile Computing, Vol. 21, 2018, pp. 1-16.

[47] J. S. Kumar, M. A. Zaveri, “Hierarchical Clustering for Dynamic and Heterogeneous Internet of Things”, Proceedings of the 6th International Conference on Advances in Computing & Communications, Cochin, India, 6-8 September 2016, pp. 276-282.

[48] A. R. Sadek, “Hybrid Energy Aware Clustered Protocol for IoT Heterogeneous Network”, Future Computing and Informatics Journal, Vol. 3, No. 2, 2018, pp. 166-177.

[49] Contiki Operating System home page, http://www.contiki-os.org (accessed: 2022)

[50] N. Dalwadi, M. Padole, “Performance Analysis of Wireless Motes in IoT”, ICT Analysis and Applications, Lecture Notes in Networks and Systems, Springer, Vol. 517, 2023, pp. 383-397.

[51] K. Molugaram, G. Rao, S. “Analysis of Time Series, Statistical Techniques for Transportation Engineering”, Elsevier, 2017, pp. 463-489.

[52] N. Dalwadi, M. Padole, "The Internet of Things Based Water Quality Monitoring and Control", Proceedings of Smart Systems and IoT: Innovations in Computing, Smart Innovation, Systems and Technologies, Springer, Vol. 141, 2020, pp. 409-417.

[53] E. Lisova, E. Uhlemann, J. Åkerberg, M. Björkman, "Monitoring of Clock Synchronization in Cyber-Physical Systems: A Sensitivity Analysis", Proceedings of the International Conference on Internet of Things, Embedded Systems and Communications, Gafsa, Tunisia, 20-22 October 2017, pp. 134-139.

Dahlin Deadbeat Internal Model Control for Discrete MIMO Systems

Original Scientific Paper

Nahla Touati
Esprit School of Engineering, André Ampère, Ariana, 2083, Tunisia,
University of Tunis El Manar, National Engineering School of Tunis,
Automatic Research Laboratory, Tunis, Le Belvédère, 1002, Tunisia
nahla.karmani@gmail.com

Imen Saidi*
University of Tunis El Manar, National Engineering School of Tunis,
Automatic Research Laboratory, Tunis, Le Belvédère, 1002, Tunisia.
Université de Tunis, Ecole Nationale Supérieure d'Ingénieurs de Tunis, Taha Hussein Montfleury, Tunis, 1008, Tunisia
imen.saidi@gmail.com

*Corresponding author

Abstract *– Controlling Multiple Input Multiple Output (MIMO) systems present a considerable challenge, particularly when dealing with time delays, nonlinearities, and disturbances. While the Dahlin algorithm and deadbeat control can offer good performance for such systems especially for systems requiring aperiodic responses or those where overshoot and setteling time need to be minimized, their effectiveness can diminish if the model parameters are inaccurate or in the presence of disturbances which lead to steady-state errors. To address these limitations, we propose combining these approaches with Internal Model Control, known for its robustness in handling variations in process dynamics, ensuring accurate setpoint tracking and disturbance rejection. In this paper, we introduce the Dahlin Deadbeat Internal Model Control (DDIMC) for discrete MIMO systems. Initially designed for linear processes with multiple time delays, this control strategy addresses complex control challenges arising from coupling effects and time delays. For nonlinear processes, we extend this controller using a multimodal control strategy which involves describing the nonlinear system with multiple linear discrete models, each paired with a Dahlin Deadbeat controller. A fusion technique is then employed to select the most suitable controller for application. Simulation case studies performed using the MATLAB software validate the effectiveness of these strategies, demonstrating their ability to consistently ensure satisfactory dynamic and robust performance.*

Keywords*:* *Dahlin Deadbeat control, Discrete Systems with Time Delays, Internal Model Control, Multimodal control, linear Multiple Input Multiple Output systems, nonlinear systems*

Received: February 8, 2024; Received in revised form: March 28, 2024; Accepted: April 24, 2024

1. INTRODUCTION

Controlling Multiple Input Multiple Output (MIMO) systems presents a significant challenge in control theory due to their inherent complexity arising from intricate variables interactions, time delays, and nonlinear characteristics [1, 2].

Various control laws have been developed to handle these difficulties and to achieve effective nominal performance. Conventional controllers like PID are commonly used due to their simplicity. However, they frequently yield inadequate performance leading to issues like instability, large overshoots, and slow responses [3]. With the advancement of intelligent control techniques, algorithms such as fuzzy control [4, 5], neural network control [6] and predictive control [7, 8] have been introduced for the control of MIMO systems with time delays. However, due to their complexity, these algorithms present challenges in practical applications [9]. In recent decades, the Deadbeat control stands out as an approach that aims to achieve the desired output behavior while minimizing settling time and eliminating steady-state error [10]. It is based on the use of a model to calculate the inputs that eliminate the current errors in finite time intervals. MIMO deadbeat control was proposed in [11] for linear continuous time systems with several constraints in time or frequency domain. In [12] the Deadbeat Algorithm was proposed to regulate the conical tank system. The nonlinear dynamics of this system were identified through mathematical modeling and approximated to a first-order system. The robustness of this control strategy becomes critical in the presence of non-linearities, parameter variation, or other mismatches [13]. To address these issues, the Deadbeat controller integrated with other strategies like PID, as presented in [14], was proposed to control a nonlinear higher-order system. The Dahlin Controller is an extension of the Deadbeat controller and well known especially for controlling deadbeat processes offering stability and nominal performance [15]. In [16], modulator based current control strategies (Deadbeat, PI and Dahlin controller) for permanent magnet synchronous motors were compared. Although all investigated control strategies exhibit stability, the Dahlin Controller stands out as offering better robustness properties for the closed-loop control system. For nonlinear systems. the operating-range scheduled robust Dahlin Algorithm was proposed in [9], for a class of SISO nonlinear systems represented by a nominal first-order inertia plus pure delay model. To eliminate steady state error, the integration control action is added when the output is close to the setting value. In [17], a modified Dahlin algorithm was proposed for level control in a nonlinear tank system, which was linearized around its equilibrium point. The proposed approach achieves better performance compared to conventional PID controllers.

While the Dahlin controller is known for its effectiveness, it faces challenges such as steady state errors and diminished robustness due to inaccuracies in model parameters or constraints on the control as discussed in [18]. To address these issues, Dahlin algorithm was combined to robust control methods or adaptive control algorithms [18].

The Dahlin Deadbeat algorithm can be combined to discrete internal model known for its nominal performance and robustness, while considering the model structure of the process [19, 20]. It was proposed to control the manipulator's positioning system in [21]. An IMC–Dahlin temperature control method based on relay feedback self-tuning identification was proposed and validated through real application on a thermostat in [22]. In this paper, the Dahlin Deadbeat based IMC, DDIMC, was initially proposed for MIMO linear discrete systems [23]. The promising outcomes achieved in controlling such systems prompted its broader application to multivariable nonlinear discrete-time systems by considering multimodeling strategy [24]. Multimodel methodologies have gained significant traction in both modeling and controlling nonlinear systems [25, 26]. This novel approach involves initially developing a model base to describe the MIMO nonlinear system. Each linear model is paired with its correspondent Dahlin deadbeat controller. The main key of the multi-model approach lies in the selection, at each sampling time, of the most fitting model that accurately approximates the current state of the process around an operational point. Subsequently, its corresponding controller is applied to the entire system.

This paper studies control challenges of MIMO systems. The DDIMC is initially proposed for linear systems with time delays and then extended to nonlinear systems using the DDIMMC control. The main objectives consist of ensuring good dynamic performance while maintaining robustness.

The remainder of this paper is organized as follows: the Dahlin Deadbeat Internal Model Control (DDIMC) is provided in Section 2. Dahlin Deadbeat Internal Multimodal Model Control (DDIMMC) is proposed in Section 3. Section 4 explores the results obtained from numerical simulations, while Section 5 presents some conclusions.

2. DAHLIN DEADBEAT INTERNAL MODEL CONTROL FOR LINEAR MIMO SYSTEMS

The DDIMC control is proposed for linear MIMO processes with time delays and particularly when there are requirements for fast response and robustness [2]. The proposed approach combines the advantages of the Dahlin Deadbeat control and the Internal model control within a unified structure. In a dead-beat controller, the system tracks a step input that is delayed by a few sampling times [10]. The Dahlin controller [13], which is built upon the dead-beat controller, generates a smoother exponential response in comparison to the standard dead-beat controller. As for the Internal Model Control (IMC), it is known for its robustness in handling both disturbances and uncertainties by incorporating a detailed model of the process [19].

2.1. THE DISCRETE IMC CONTROL FOR MIMO SYSTEMS

The discrete IMC structure, depicted in Fig. 1, incorporates a stable MIMO process $G(z)$, the internal model $M(z)$ and a controller C_{CMI} (z) arranged to act as the model inverse. These components are described by transfer matrices of dimension $(n \times n)$. $u(z)$ and $y(z)$ represent respectively the input actions and the output vectors of dimension $(n \times 1)$. $r(z)$ and $d(z)$ are respectively the reference vector of dimension $(n \times 1)$ and the disturbance vector that may affect the system. The input actions are simultaneously applied to the process and its model. The outputs mismatch is considered to adjust the controller's input $e(z)$.

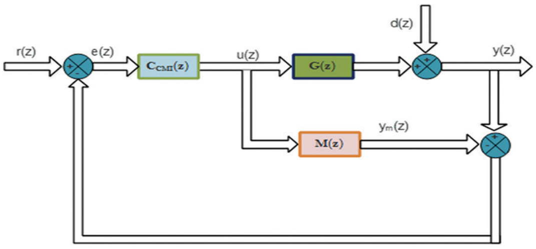


Fig. 1. The MIMO IMC structure [20]

From Fig. 1, we can deduce the following equation for the input action vector $u(z)$ [23]:

$$u(z) = (I_n + C_{CMI}(z)(G(z) - M(z)))^{-1} C_{CMI}(z)(r(z) - d(z)) \text{ (1)} \quad (1)$$

$$\begin{gathered} y(z) = G(z)(I_n + C_{CMI}(z)(G(z) - M(z)))^{-1} C_{CMI}(z) r(z) \\ +(I_m - G(z)(I_n + C_{CMI}(z)(G(z) - M(z)))^{-1} C_{CMI}(z)) d(z) \end{gathered} \quad (2)$$

In conventional IMC theory, when the controller is chosen as the model inverse, perfect control is achieved. However, for many physical systems, the inversion task isn't feasible. An approximate inverse is then required [26, 27].

The IMC controller for non-minimum and delayed systems, depicted in Fig. 2, can be designed as proposed in [20, 23].

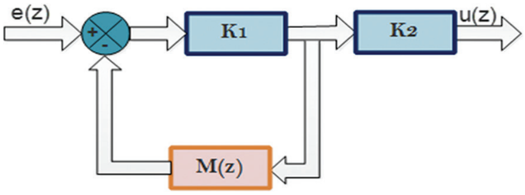


Fig. 2. Structure of the internal model controller [23]

The internal model controller $C_{CMI}(z)$ is then described as the following:

$$C_{CMI}(z) = K_1 K_2 (I_n + K_1 M(z))^{-1} \tag{3}$$

The gain matrix K_1 is crucial for ensuring the stability of the controller, while K_2 is considered to compensate for system's static errors.

K_2 is described below:

$$K_2 = (I_n + K_1 M(1))(K_1 M(1))^{-1} \tag{4}$$

where $M(1)$ represents the model's static matrix gain.

The proposed controller steady-state gain is equal to the inverse of the model steady-state gain. Offset-free control is then obtained for constant setpoints and output disturbances [28].

The IMC control structure illustrated in Fig. 1, can be modified to a classical feedforward control as presented in Fig. 3 below.

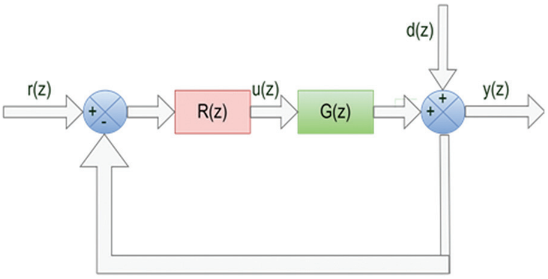


Fig. 3. Classical closed-loop control structure

where:

$$R(z) = \left(I_n - C_{CMI}(z)M(z)\right)^{-1} C_{CMI}(z) \tag{5}$$

$$H(z) = R(z)G(z) \times \left(I_n + R(z)G(z)\right)^{-1} \tag{6}$$

2.2. THE DAHLIN-DEADBEAT CONTROLLER

Deadbeat control is a control strategy aiming to drive the system outputs to the desired value within a few sampling times. Fast and accurate tracking of references signals are then ensured.

For MIMO systems that occur frequently in the processing industry, it's desirable to eliminate the coupling effects between the loops for MIMO systems. The proposed controller, in this paper, is chosen to handle both interactions and time delays, that may exist, within a single design [29]. For that reason, the desired closed-loop transfer matrix $H(z)$ is chosen to have a diagonal form and is defined as follows.

$$H(z) = \begin{bmatrix} z^{-k} & 0 & \ldots & 0 \\ 0 & z^{-k} & \ldots & 0 \\ M & M & O & M \\ 0 & 0 & \ldots & z^{-k} \end{bmatrix}, \ k \geq 1 \tag{7}$$

The Dahlin algorithm is an extension of the deadbeat control that was proposed specifically for the system with pure time delay. The key idea of the Dahlin algorithm is to design an anticipant closed-loop transfer function. The system behaves similarly to a continuous first order process with time delay [13]. The transfer matrix H(s) is chosen as follows:

$$H(s) = \begin{bmatrix} h_{11}(s) & 0 & L & 0 \\ 0 & h_{22}(s) & L & 0 \\ M & M & O & M \\ 0 & 0 & L & h_{nn}(s) \end{bmatrix} = \left(h_{ij}(s)\right)_{1 \leq i,j \leq n} \tag{8}$$

where: $h_{ii}(s)=exp(-T_i s)/(\tau_i s+1), 1 \leq i \leq n; T_i$ is the time delay selected as: $T_i = N \times T_s$, T_s is the sampling time and τ_i is the time constant.

The discrete form of the transfer functions $h_{ii}(s)$, $1 \leq i \leq n$, obtained with a zero-order hold is then described below:

$$h_{ii}(z) = \left(\frac{1-exp\left(\frac{-T_s}{\tau_i}\right)}{\left(z-exp\left(\frac{-T_s}{\tau_i}\right)\right)} \right) z^{-N} \tag{9}$$

The Dahlin deadbeat controller R(z) is then described below:

$$R(z) = \left(I_n - H(z)\right)^{-1} H(z) G(z)^{-1} \tag{10}$$

2.3. THE DAHLIN DEADBEAT IMC CONTROL

The proposed DDIMC control strategy uses the Internal Model Control (IMC), as depicted in Fig. 4. Initially, the desired closed-loop dynamics are selected according to Eq. (8) and Eq. (9), followed by the design of the Dahlin controller described by Eq. (10).

The IMC controller considered in the DDIMC structure is then described as follows:

$$C_{CMI}(z) = R(z) \times \left(I_n + R(z)M(z)\right)^{-1} \tag{11}$$

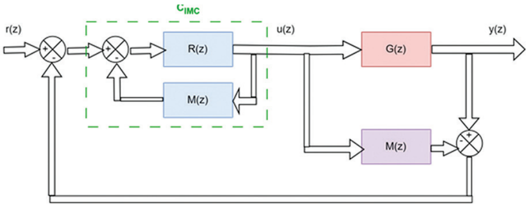


Fig. 4. The DDIMC structure

3. DAHLIN DEADBEAT INTERNAL MULTIMODEL CONTROL FOR NONLINEAR MIMO SYSTEMS

Modern industrial processes often exhibit nonlinearity. Linear models can't capture the dynamics of complex systems due to the presence of strong nonlinearities. The effects of these nonlinearities are mostly undesirable and can greatly affect the performance of controllers [26]. To tackle these challenges, multimodal approaches are emerging as promising alternatives to conventional linearization methods. These methods involve segmenting the system's operational range into distinct zones and considering localized linear models for each zone [17]. The Multimodal principle is depicted in Fig. 5.

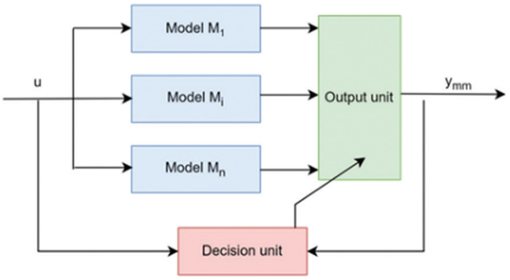


Fig. 5. Multimodel Control

The algorithm of the proposed method is given by:

Step 1: A base of several discrete MIMO linear models is defined to describe the nonlinear system across its entire operating ranges.

Step 2: The desired closed loop transfer matrix H(z) is specified based on Eq (9).

Step 3: For each linear MIMO model, a specific Dahlin Deadbeat controller is designed based on Eq10.

Step 4: At each sampling time, the model that closely matched the process dynamics is selected based on the switching technique illustrated in Fig. 6.

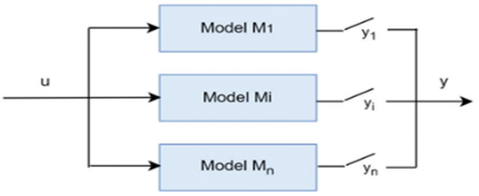


Fig. 6. Switching technique [26]

The errors between model outputs and the actual system responses should then be evaluated.

For each model M_i, a distance vector D_i, describing model outputs y_{M_i} and the system outputs y mismatch, is represented by the Eq. (12).

$$D_i = \sum_{j=1}^{m} \left\| y_j(k) - y_{M_{ij}}(k) \right\| \; i = 1 \ldots n \qquad (12)$$

For each model M_i, i=1...n, a validity index v_i needs to be assessed. A validity index v_i of 1 is assigned to the model with the smallest distance vector, indicating its superior relevance in describing the nonlinear system. Conversely, for the other models in the set, v_i is set to 0. The multimodal vector of outputs aligns then with the vector of the chosen model's outputs (cf. Fig. 7).

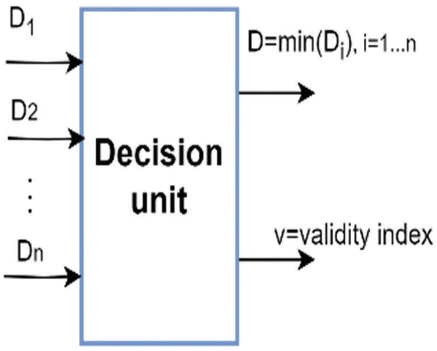


Fig. 7. Basic diagram of the model validation method [24]

Step 5. Once the model is validated, its corresponding DDIMC controller, is applied to control the entire nonlinear system.

The new DDIMMC, proposed for nonlinear discrete systems is depicted in Fig. 7.

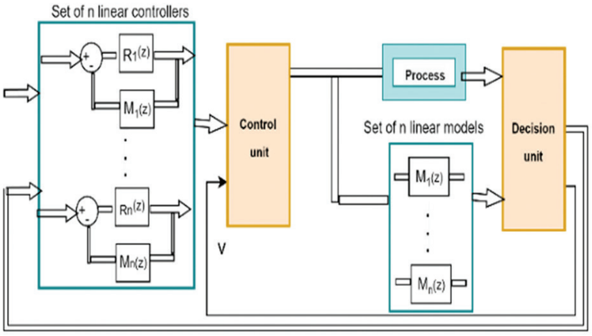


Fig. 8. The DDIMMC structure

4. SIMULATION CASE STUDIES

To demonstrate the effectiveness of the proposed control structures, two case studies were introduced. For the first case, a linear MIMO discrete system, specifically a neonatal incubator is proposed. As for the second case, it concerns a nonlinear discrete MIMO system: stirred tank reactor (CSTR) process.

4.1. DDIMC FOR A LINEAR MIMO DISCRETE SYSTEM: A NEONATAL INCUBATOR SYSTEM

- System description

Let's consider a linear MIMO neonatal incubator system described by the following transfer matrix [30]:

$$\begin{bmatrix} Y_H(s) \\ Y_T(s) \end{bmatrix} = \begin{bmatrix} \frac{0.3145}{1.753s+1} e^{-0.184s} & \frac{-0.01649}{0.3065s+1} e^{-0.496s} \\ \frac{-0.3483}{11.29s+1} e^{-1.31s} & \frac{0.2356}{26.07s+1} e^{-1.46s} \end{bmatrix} \begin{bmatrix} U_H(s) \\ U_T(s) \end{bmatrix} \qquad (13)$$

where $Y_H(s)$, $U_H(s)$, $Y_T(s)$, $U_T(s)$ are the outputs and control actions related respectively to the humidity and temperature inside the incubator.

The discrete transfer matrix is described as follows with a sampling time of $Ts = 1.2$ seconds.

$$\begin{bmatrix} Y_1(z) \\ Y_2(z) \end{bmatrix} = \begin{bmatrix} \frac{0.1383z+0.01755}{z-0.5043} z^{-1} & \frac{-0.01483z-0.00133}{z-0.01994} z^{-1} \\ \frac{-0.03205z-0.00366}{z-0.8992} z^{-2} & \frac{0.008344z+0.002255}{z-0.9924} z^{-2} \end{bmatrix} \begin{bmatrix} U_1(z) \\ U_2(z) \end{bmatrix} \quad (14)$$

Two scenarios are presented. The first one considers the nominal case without any disturbances, while the second one tests the robustness towards external disturbances.

- First scenario: Nominal case

Fig. 9 illustrates simulation results for this scenario. All the responses accurately settle the setpoints. The overall performance is better when applying the DDIMC compared to the discrete IMC [20]. The proposed approach has less overshoot and shorter settling time as presnted in Table 1 which illustrates a quantitative comparison of the obtained results, to validate the effectiveness of the proposed control approach compared to the IMC and its ability to ensure satisfactory performance.

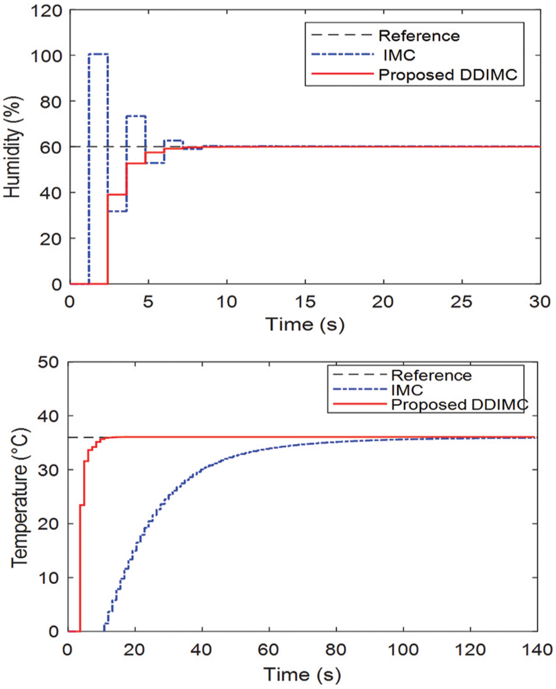


Fig. 9. Humidity and Temperature levels (Nominal case)

Table 1. Performance of the transient responses with the DDIMC and IMC [20]

	IMC		Proposed DDIMC	
	Humidity	Temperature	Humidity	Temperature
Rise Time (s)	0.57	36.52	2.53	2.72
Setting Time (s)	6.49	82.48	5.75	8.84
Overshoot (%)	67.46	0	8.2. 10^{-5}	0

- Second scenario: In the presence of disturbances

The robustness towards external disturbances of the proposed approach is presented in this scenario. Step type output disturbances of 10% occur at t=10s on the humidity level and 2°C occur at t=100s on the temperature level, respectively. Fig. 10 displays the responses for the DDIMC control. The system remains stable, and the disturbances are completely rejected after about 8 and 12 sampling times for the humidity and the temperature levels, respectively.

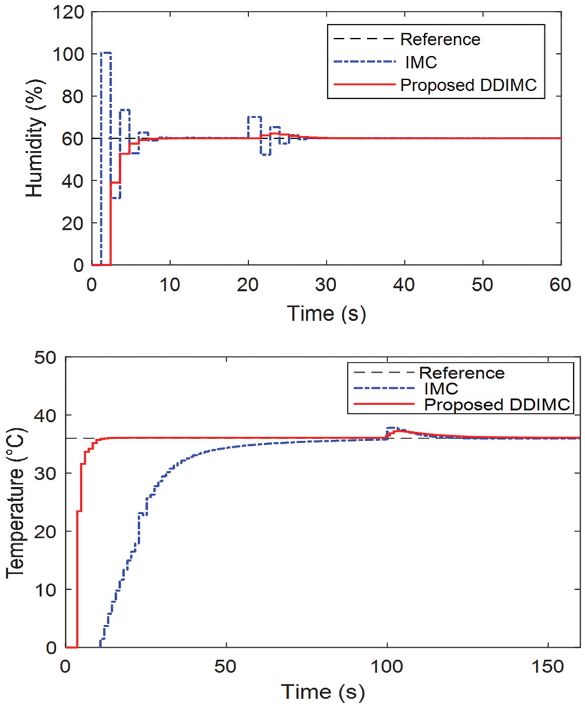


Fig. 10. Humidity and Temperature levels (robustness towards disturbances)

4.2. DDIMC FOR A NONLINEAR MIMO DISCRETE SYSTEM: STIRRED TANK REACTOR

- System description

Let's consider a MIMO stirred tank reactor (CSTR) process, which consists of an irreversible, exothermic reaction, A → B, in a constant volume reactor cooled by a single coolant stream. It can be modeled by the following nonlinear equations [31]:

$$\begin{cases} \dot{C}_A(t) = \frac{q}{V}[C_{A0} - C_A(t)] \\ \quad - k_0 C_A(t) e^{-(E/RT(t))} \\ \dot{T}(t) = \frac{q}{V}[T_0 - T(t)] + k_1 C_A(t) e^{-(E/RT(t))} \\ \quad + k_2 q_c(t)[1 - e^{-(k_3/q_c(t))}][T_{c0} - T(t)] \end{cases} \quad (15)$$

The system's inputs are the flow rate q and coolant flow rate q_c. The outputs are respectively the concentration C_A, and the temperature T. Table 2 displays the CSTR's parameter values.

Parmeter	Description	Value
C_{A0}	Feed concentration	1 mol/l
T_{C0}	Inlet coolant temperature	350 K
h_A	Heat transfer term	7×10^5 cal/minK
E/R	Activation energy term	10^4 K
ρ,ρ_c	Liquid densities	10^3 g/l
q	Process flow rate	100 l min^{-1}
T_0	Feed temperature	350K
V	CSTR volume	100 l
k_0	Reaction rate constant	7×10^{10} min^{-1}
ΔH	Heat of reaction	-2×10^5 cal/mol
C_p, C_{pc}	Specific heats	1 cal $g^{-1}K^{-1}$

$$k_1 = \frac{-\Delta H k_0}{\rho C_p} \qquad k_2 = \frac{-\rho_c C_{pc}}{\rho C_p V} \qquad k_3 = \frac{h_A}{\rho_c C_{pc}}$$

After linearization around three operating points, local linear models are obtained. In the discrete state-space representation with a sampling time of $T_s = 0.1$ seconds, these models are represented below [31]:

$$\begin{cases} x(k+1) = A_i x(k) + B_i u(k) \\ \quad y(k) = Cx(k) + Du(k) \end{cases} i = 1,2,3$$

where, $x(k)$, $u(k)$ and $y(k)$ represent respectively the states, inputs, and outputs vectors:

$$A_1 = \begin{bmatrix} 0.1552 & -0.004 \\ 143.4331 & 1.5794 \end{bmatrix}, B_1 = \begin{bmatrix} 0.0007 & 0.0002 \\ -0.0463 & -0.1122 \end{bmatrix}; 16)$$

$$A_2 = \begin{bmatrix} -0.0225 & -0.0039 \\ 175.1096 & 1.5464 \end{bmatrix}, B_2 = \begin{bmatrix} 0.0006 & 0.0002 \\ -0.0343 & -0.1203 \end{bmatrix}$$

$$A_3 = \begin{bmatrix} 0.1801 & -0.0044 \\ 137.5519 & 1.6376 \end{bmatrix}, B_3 = \begin{bmatrix} 0.0007 & 0.0002 \\ -0.0509 & -0.1117 \end{bmatrix};$$

$$C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } D = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Three discrete MIMO transfer matrices are obtained and described as follows:

$$M_1 = \begin{bmatrix} \dfrac{0.0007z - 0.0009}{z^2 - 1.735z + 0.8189} & \dfrac{0.0002z + 0.0001}{z^2 - 1.735z + 0.8189} \\ \dfrac{-0.0463z + 0.1076}{z^2 - 1.735z + 0.8189} & \dfrac{-0.1122z + 0.0461}{z^2 - 1.735z + 0.8189} \end{bmatrix};$$

$$M_2 = \begin{bmatrix} \dfrac{0.0006z - 0.0008}{z^2 - 1.524z + 0.6481} & \dfrac{0.0002z + 0.0002}{z^2 - 1.524z + 0.6481} \\ \dfrac{-0.0343z + 0.1043}{z^2 - 1.524z + 0.6481} & \dfrac{-0.1203z + 0.0323}{z^2 - 1.524z + 0.6481} \end{bmatrix}; \quad (17)$$

$$M_3 = \begin{bmatrix} \dfrac{0.0007z - 0.0009}{z^2 - 1.818z + 0.9002} & \dfrac{0.0002z + 0.0002}{z^2 - 1.818z + 0.9002} \\ \dfrac{-0.0509z + 0.1055}{z^2 - 1.818z + 0.9002} & \dfrac{-0.1117z + 0.0476}{z^2 - 1.818z + 0.9002} \end{bmatrix}$$

The desired closed loop transfer matrix is described below.

$$H(z) = \begin{bmatrix} \frac{0.09516}{z-0.9048} z^{-1} & 0 \\ 0 & \frac{0.09516}{z-0.9048} z^{-1} \end{bmatrix} \tag{18}$$

Two scenarios are presented. In the first one, the nominal case without any disturbances is considered, while the second one tests the robustness towards external disturbances.

- First scenario: Nominal case

Fig. 11 illustrates simulation results for this scenario. The setpoints are 0.09 mol/l and 450 K for the concentration and the temperature respectively. We can notice that the outputs of the CSTR system track the reference signals with zero steady state errors. Moreover, the system demonstrates better transient responses compared to the Discrete Internal Multimodel Control strategy [24]. The DDIMMC yields responses with minimized overshoot and undershoot, and shorter setting time as detailed in Table 3. In fact, the transient response performance is not explicitly considered in the IMMC controller design, whereas optimizing transient response is a primary concern for the proposed DDIMMC controller.

Table 3. Performance of the transient responses with the DDIMMC and IMMC [24]

	IMMC		Proposed DDIMMC	
	Concentration	Temperature	Concentration	Temperature
Setting Time (s)	0.96	1.03	0.22	0.21
Overshoot (%)	6.4	0.62	2.99	0.78
Peak	9.64×10-2	453	9.23×10^{-2}	452.3

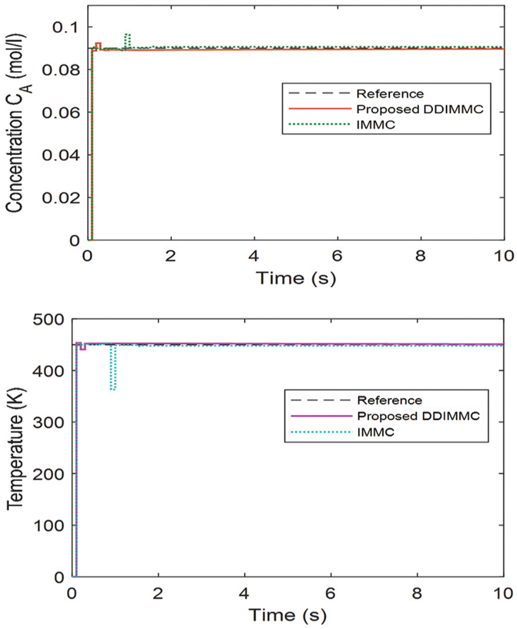


Fig. 11. Concentration and Temperature levels (Nominal case)

- Second scenario: In the presence of disturbances

The robustness towards external disturbances of the proposed approach is presented in this scenario. Persistent disturbances of 0.02 mol/l and 10 K occur at t=0.5 s on the concentration and the temperature levels respectively. Fig. 12 displays simulation results for this scenario. We can notice that despite the presence of persistent disturbances, the outputs remain able to follow the reference inputs. The DDIMMC controller has proven its ability to ensure good performance despite the disturbances.

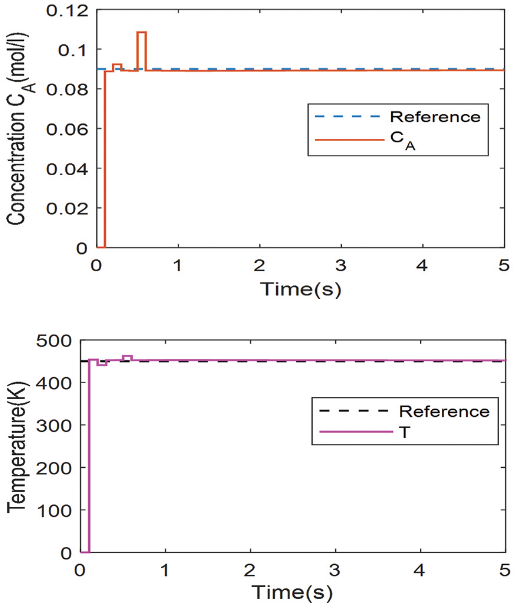


Fig. 12. Concentration and Temperature levels (Robustness towards disturbances)

5. CONCLUSION

The Dahlin deadbeat Internal Model Control was proposed in this paper for MIMO systems. It was designed on the principles of the Dahlin deadbeat control and the internal model control. The controller, proposed initially for linear systems, is easy to implement, robust and has good dynamic control performance. A simulation study on a linear MIMO neonatal incubator illustrates the effectiveness of this approach in ensuring good transient performance, accurate tracking, and robustness towards disturbances. Beyond linear systems, the DDIMC was extended to control MIMO nonlinear systems by incorporating a multi-modeling strategy based on describing the nonlinear system by a set of multiple linear models. At each sampling time, the most appropriate model is selected and its corresponding Dahlin deadbeat controller is applied to the entire system. This novel Dahlin deadbeat internal multimodal control method (DDIMMC) demonstrates its effectiveness through simulations on a stirred tank reactor (CSTR) process involving two inputs and outputs. The proposed control approach has proven its ability to ensure satisfactory nominal and robustness performances.

Future work may involve conducting experimental tests on real systems using DDIMC and DDIMMC methods aiming to prove the effectiveness of these approaches in real-world applications. Furthermore, extending these control strategies to address the complexities of non-square systems, which pose additional challenges in control, could be explored.

6. REFERENCES

[1] S. Gambhire, D. Kishore, M. Kumar, S. Pawar, “Design of Super Twisting Integral Sliding Mode Control for Industrial Robot Manipulator”, International Journal of Electrical and Computer Engineering Systems, Vol. 13, No. 9, 2022, pp. 815-821.

[2] B. Xu, “On delay independent stability of large scale systems with time delays”, IEEE Transactions on Automatic Control, Vol. 40, No. 5, 1995, pp. 930-933.

[3] A. Ajiboye, F. Opadiji, O. Popoola, O. Adebayo, “Selection of PID controller design plane for time delay systems using genetic algorithm”, International Journal of Electrical and Computer Engineering Systems, Vol. 13, No. 10, 2022, pp. 917-926.

[4] B. Chen, X. Liu, K. Liu, C. Lin, “Adaptive fuzzy tracking control of nonlinear MIMO systems with time-varying delays”, Fuzzy Sets and Systems, Vol. 217, 2013, pp. 1-21.

[5] T. T. Pham, C. N. Nguyen, “Adaptive Sliding Mode Control Based on Fuzzy Logic and Low Pass Filter for Two-Tank Interacting System”, International Journal of Electrical and Computer Engineering Systems, Vol. 13, No. 9, 2022. pp. 477-483.

[6] W. Ruliang, M. Kunbo, C. Chao-Yang, L. Yanbo, M. Hebo, Y. Zhifang, "Adaptive neural control for MIMO nonlinear systems with state time-varying delay”, Journal of Control Theory and Applications, Vol. 10. No. 3, 2012, pp. 309-318.

[7] W. Tong, S. Wang, “Fault Tolerant Control of Lipschitz Nonlinear Switched Delay Systems: A Model Predictive Control Scheme”, Journal of Applied Science and Engineering, Vol. 27, No. 7, 2023, pp. 2735-2745.

[8] R. Hološ, V. Bobál, "Model predictive control of time-delay systems with measurable disturbance compensation", Proceedings of the 20th International Conference on Process Control, Strbske Pleso, Slovakia, 9-12 June 2015, pp. 209-214.

[9] X. Tian, H. Peng, X. Luo, S. Nie, F. Zhou, X. Peng, “Operating Range Scheduled Robust Dahlin Algorithm to Typical Industrial Process with Input Constraint”, International Journal of Control, Automation and Systems, Vol. 18, 2020, pp. 897-910.

[10] I. Mahmoud, I. Saidi, “A dead-beat internal model control for trajectory tracking in discrete- time linear system”, Przeglad Elektrotechniczny, Vol. 99, No. 1, 2022, pp. 116-119.

[11] K. Tsumura, H. Nakanishi, E. Nobuyama, "Multiobjective control for continuous time MIMO systems via deadbeat regulation", Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, USA, 12 December 1997, pp. 466-471.

[12] D. Marshiana, P. Thirusakthimurugan, "Design of Deadbeat Algorithm for Nonlinear Conical Tank System", Procedia Computer, Vol. 57, 2015, pp. 1351-1358.

[13] E. B. Dahlin, "Designing and tuning digital controllers", Instruments and control systems, Vol. 41, No. 6, 1968, pp. 77-83.

[14] A. J. Bedekar, S. Shinde, "Robust Deadbeat Control of Twin Rotor Multi Input Multi Output System", International Journal of Engineering Research & Technology, Vol. 4, No. 5, 2015, pp. 703-709.

[15] F. C. Teng, G. F Ledwich, A. C. Tsoi, "Extension of the Dahlin-Higham controller to multivariable systems with time delays", International Journal of Systems Science, Vol. 25, No. 2, 1994, pp. 337-350.

[16] S. Walz, R. Lazar, G. Buticchi, M. Liserre, "Dahlin-Based Fast and Robust Current Control of a PMSM in Case of Low Carrier Ratio", IEEE Access, Vol. 7, 2019, pp. 102199-102208.

[17] T. Dlabač, S. Antić, M. Ćalasan, A. Milovanović, N. Marvučić, "Nonlinear Tank-Level Control Using Dahlin Algorithm Design and PID Control", Applied Sciences, Vol. 13, No. 9, 2023, pp. 1-25.

[18] K. Kawaguchi, J. Endo, H. Shibasaki, R. Tanaka, Y. Hikichi, Y. Ishida, "The Control of the Pneumatic Actuator Using Dahlin Algorithm", International Journal of Modeling and Optimization, Vol. 3, No. 244, 2013, pp. 98-100.

[19] C. G. Garcia, M. Morari, "Internal model control. A unifying review and some results", Industrial Engineering Chemistry Process Design and Development, Vol. 21, No. 2, 1982, pp. 308-323.

[20] I. Bejaoui, I. Saidi, D. Soudani, "Internal model control of discrete non-minimum phase over-Actuated systems with multiple time delays and uncertain parameters", Journal of Engineering Science and Technology Review, Vol. 12, No. 2, 2019.

[21] I. Clitan, V. Muresan, A. Clitan, A. M. Abrudean, "Discrete Control for an Industrial Manipulator Using Both Dahlin Algorithm and Internal Model Control Design Approaches", Applied Mechanics and Materials, Vol. 656, 2014, pp. 360-368,

[22] P. Li, Z. Zhao, "IMC - Dahlin Temperature Regulator for Thermostat", Proceedings of the 29th Chinese Control And Decision Conference, Chongqing, China, 28-30 May 2017, pp. 4238-4242.

[23] I. Saidi, N. Touati, "Dahlin Deadbeat Internal Model Controller Design for Discrete Systems with Time Delay", Proceedings of the 9th International Conference on Control", Decision and Information Technologies, Rome, Italy, 3-6 July 2023, pp. 2633-2638.

[24] C. Othman, I. Ben Cheikh, D. Soudani, "On the Internal Multi-Model Control of Uncertain Discrete-Time Systems", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 9, 2016, pp. 88-98.

[25] A. Dhahri, I. Saidi, D. Soudani, "Internal model control for multivariable over-actuated systems with multiple time delay", Proceedings of the 4th International Conference on Control Engineering & Information Technology, Hammamet, Tunisia, 16-18 December 2016, pp. 623-626.

[26] N. Touati, I. Saidi, A. Dhahri, D. Soudani "Internal multimodel control for nonlinear overactuated systems", Arabian Journal for Science and Engineering, Vol. 44, No. 3, 2019, pp. 2369-2377.

[27] N. Touati, I. Saidi, "Internal model control for underactuated systems based on novel virtual inputs method", Przegląd Elektrotechniczny, Vol. 97, No. 9, 2021, pp. 95-99.

[28] I. Saidi, N. Touati, A. Dhahri, D. Soudani, "A comparative study on existing and new methods to design internal model controllers for non-square systems", Transactions of the Institute of Measurement and Control, Vol. 41, No. 13, 2019, pp. 3637-3650.

[29] F. C. Teng, G. F. Ledwich, A. C. Tsoi, "Extension of the Dahlin-Higham controller to multivariable systems with time delays", International Journal of Systems Science, Vol. 25, No. 2, 1994, pp. 337-350.

[30] W. Bezerra-Correia, B. Claure-Torrico, R. D. Olímpio-Pereira, "Optimal control of MIMO dead-time linear systems with dead-time compensation structure1", DYNA, Vol. 84, No. 200, 2017, pp. 62-71.

[31] J. Du, T. A. Johansen, "A gap metric based weighting method for multimodel predictive control of MIMO nonlinear systems", Journal of Process Control, Vol. 24, No. 9, 2014, pp. 1346-1357.

Is the development of objective image quality assessment methods keeping pace with technological developments?

Preliminary Communication

Md. Abdur Rahman*
Hongik University,
Department of Electronic & Electrical Engineering
Mapo-gu, Seoul, Republic of Korea
mdabdur.rahman.1995@ieee.org

Hanif Bhuiyan
Monash University,
Monash Data Futures Institute
Clayton, Victoria, Australia
hanif.bhuiyan@monash.edu

*Corresponding author

Abstract *– Visual data proliferation with technological innovations is progressively moving forward, bringing automation, digitization, and smartness to various aspects of daily lives. These visual data-dependent technologies need to be assessed during their development to ensure the quality of services. Consequently, when appropriate quality assessment metrics are unavailable, the advancement of image processing technologies is hindered. This preliminary work addresses the disparity between the rapid advancement of image processing-based algorithms and the lag in developing image quality assessment (IQA) methods while briefly discussing future prospects. The shortcomings in IQA development are discussed by broadly categorizing application areas into human visual system (HVS)-based and computer vision-based, while highlighting the trends and necessity of developing application-specific IQA methods. Despite the existence of several emerging IQAs, this preliminary communication indicates significant opportunities for objective IQA developments in both application areas to support technological advancements. Hence, this work can serve academic and industrial communities by providing guidance for expediting IQA developments.*

Keywords: *image quality assessment (IQA), human vision, computer vision*

Received: January 13, 2024; Received in revised form: April 26, 2024; Accepted: April 26, 2024

1. INTRODUCTION

The widespread adoption of visual media is facilitated by an increasing number of smart devices and Internet of Things (IoT) based systems. The availability of visual media, such as images and videos, has paved the way for new applications and widened the aspects of the existing application areas in various sectors while offering more facilities for consumers. Quality evaluation metrics play a crucial role in ensuring the quality of services for these applications. Since humans are the primary customers or end-users of visual media in most cases, subjective quality evaluation is a well-known approach for evaluating the quality of images and videos during any process or application. Despite receiving ample acceptance, subjective evaluation is less desirable for real-time application due to its expense, time consumption, impracticality for real-time process, and implementational complexity [1, 2]. The difficulties associated with employing subjective metrics for image quality assessment (IQA) have led to the development of objective metrics [2, 3]. Also, even for a specific application, the subjective evaluation process is very complicated, making it challenging to ensure generalized quality assessments. The difficulty of employing subjective IQA arises from the fact that humans' responses to any visual media can vary depending on the environment of observations, the subject's age, display systems, and other factors. Consequently, although subjective metrics are used to validate objective metrics in most cases, objective metrics are usually applied when developing algorithms for optimizing them [3]. Hence, this work focuses on objective metrics to address the gap in IQA development compared to technological advancements.

The objective IQA metrics were initially developed based on mathematical algorithms, such as mean squared error (MSE) and peak signal-to-noise ratio (PSNR), to evaluate the difference between original/source/reference images and processed/modified/transformed images in terms of mathematical errors and signal fidelity, respectively. Though they gained popularity initially, their shortcomings in evaluating image quality became apparent due to their sole focus on the mathematical perspective. As the human visual system (HVS) perceives images in a more complex manner, this motivated the incorporation of HVS-related theories in developing objective IQA metrics. Till now, several objectives have been and are being developed considering spatial and frequency domain behaviors of HVS, which are summarized in existing review studies for a specific area as well as from a broad perspective [3-7].

Despite the availability of several theories to explain the behaviors of HVS, most IQA metrics consider some of those theories to resemble HVS when evaluating image quality. The IQAs are application-specific in most cases [8], which indicates the need for generalized methods to be implementable in any case, irrespective of application area. Additionally, it is essential to note that the source/original image is not always available as a reference for evaluating image quality. Based on the availability of reference images, objective IQAs can be categorized into full-reference (FR) IQA, reduced-reference (RR) IQA, and no-reference (NR) or blind IQA metrics.

As the developments of IQAs are always driven by the necessities or problems faced in different scenarios, they are hardly keeping up with the progressive developments of image processing algorithms. The image processing applications still need sophisticated IQAs, where the application areas of image processing can be broadly categorized into HVS-based and computer vision-based, considering the end-user. Irrespective of the IQA category, objective IQAs for HVS-based applications aim to ensure perceptual quality. In contrast, objective IQAs for computer vision-based applications prioritize meeting the machine-operational requirement to achieve predefined goals. The existing literature reviewing these objective metrics focuses on a single IQA category or multiple IQA categories in most studies. However, no study has adequately addressed the lag in IQA developments relative to technological advancements.

This preliminary study aims to highlight this issue by discussing the lag in IQA developments for different applications. As the selection of the category of objective metrics for an application typically depends on the application requirements and corresponding available resources, this study addresses the gaps in IQA developments without emphasizing any specific category. Although a detailed categorization of IQAs based on their application areas is feasible, this work discusses the gaps in IQA development by broadly categorizing the areas into two groups based on end-users, including HVS-based and computer vision-based areas. The contribution of this work is outlined as follows.

- A preliminary study is presented to investigate the trend of objective IQA development in response to technological advancements.
- Limitations in IQA developments in HVS-based and computer vision-based application areas are identified.
- Guidelines for future research are provided to harmonize IQA developments with technological advancements.

The rest of the manuscript is organized as follows. Sections 2 and 3 address the trends of the existing IQA development to keep pace with the technological innovations in each application area while pointing out some notable limitations. This work is concluded in Section 4 with a brief discussion of future prospects.

2. HVS-BASED APPLICATIONS

HVS-based applications are one of the primary motivators for developing various IQAs to ensure the quality of services for HVS. Considering the requirements of existing and emerging applications, several factors need to be accounted for when developing IQAs, as presented below.

2.1. DISTORTIONS OR ARTIFACTS

In HVS-based applications, IQAs are frequently employed to assess image quality across multiple stages. These stages encompass various processes, including image capture, compression, transmission, decompression, and enhancement, among others. Fig. 1 illustrates the image processing pipeline, from capture to display for end-users.

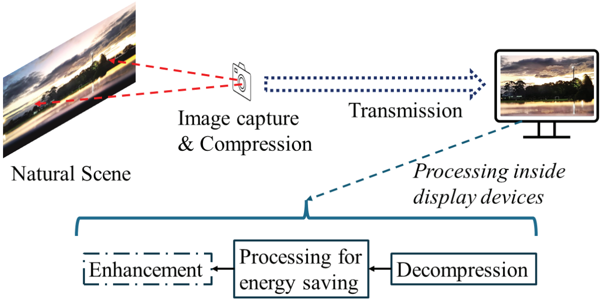

Natural Scene
Image capture & Compression
Transmission
Processing inside display devices
Enhancement
Processing for energy saving
Decompression

Fig. 1. A block diagram of image processing from source to end users.

Due to the possibility of different distortions during these steps, several conventional IQA algorithms have been developed, focusing on them. However, conventional IQAs may fall short in accurately assessing quality when multiple distortions are present simultaneously in an image. As a solution, multi-distortion IQA is developed [9]. Additionally, distortions or artifacts can be caused by various energy-saving technologies employed within the display devices or systems, presenting a challenge for existing IQAs. Although distortions stemming from external processes at various stages differ from internal distortions, IQAs capable of accounting for all these cumulative distortions are yet to be developed. Hence, the development of new IQAs is essential to effectively assess image quality while considering internal and external distortions.

On the other hand, the requirements of new applications and the problems in assessing the existing applications have created the urge to develop new methodologies. Several emerging IQAs, alongside conventional ones, have been discussed in [3]. While these applications-driven emerging IQAs are promising, they highlight the ongoing race in IQA developments to keep pace with new technologies. Moreover, there is a pressing need for new IQA methodologies, especially in areas involving distorted or degraded image processing. For example, dehazing IQA [10] is developed to evaluate image quality during the dehazing process, as a hazy source image has less information as a reference. A new NR IQA is discussed in [10], as conventional IQAs perform inadequately, and synthetic images are insufficient reference for this image reconstruction process. Similar to dehazing IQAs, proper IQAs are required to assess images enhanced from various sources, currently evaluated by NR IQAs or cumbersome subjective scores. For instance, images captured in night vision mode by the closed-circuit television (CCTV) often require enhancement for better visualization and monitoring. Despite its critical role in security and monitoring purposes, optimal IQAs to ensure service quality post-enhancement remain elusive.

2.2. CONTENT SPECIFICATION

Differences in image content compared to natural images necessitate the development of new IQAs for their evaluation. For instance, images with textual and pictorial information differ from conventional images, leading to the development of screen content IQA [11]. Another instance involves cartoon images, which prioritize structural and color features. Since conventional IQAs designed to evaluate the perceptual quality of natural images are insufficient for cartoon images, a new IQA tailored specifically for cartoon images is discussed in [12]. Additionally, remote sensing, an important area for image processing applications, requires good IQA to ensure the quality of service while monitoring remote regions. As the existing IQAs are found to be insufficient, a low-level and deep-level feature combination-based IQA has been discussed recently in [13] as a solution. Furthermore, retargeted image quality assessment (RIQA), an important branch of IQA for assessing the quality of content-aware image retargeting in multimedia applications for HVS, is still in the early stage, as addressed in [14]. As technological advancement is still in progress for this application, developing an appropriate IQA is essential. These content-oriented IQA developments address the limitations of existing IQAs and highlight the need for new IQAs tailored to application-specific content requirements for existing applications.

The trend of content-oriented new IQA developments is particularly pronounced in emerging technologies, notably in the realm of 3D visual content. One such example is the emergence of 360° images and videos. The progress in virtual reality (VR) has spurred the creation of 360° images and videos, necessitating new technologies to handle the spherical nature of these visual contents. As addressed in [15], new perceptual and deep learning-based methods (i.e., subjective assessment of multimedia panoramic video quality (SAMPVIQ), Craster parabolic projection PSNR (CPP-PSNR), viewport-based CNN (V-CNN) and others) are developed to ensure the quality of services for this emerging technology. Similarly, the rise of 3D visual media has spurred the development of stereoscopic/3D IQA [16]. Although 3D/stereoscopic IQAs are notable, they may not be suitable enough with new technologies used for 3D view generation. As the concept of stereoscopy has limitations in providing depth information and has vergence-accommodation conflict, multifocal displays [17] are emerging as a promising solution in generating 3D views for which new IQAs will be required.

2.3. SURROUNDING ENVIRONMENTS' EFFECTS

In addition to distortions and visual contents, surroundings that affect the sensitivity of HVS need to be considered for perceptual quality evaluations in different environmental situations. Despite being a critical factor affecting technological advancements, no IQA has incorporated this consideration to support existing and emerging applications. Backlight dimming is an example of an established application area where challenges arise from the absence of appropriate IQAs while developing new image processing algorithms. In [18], an energy-efficient dimming technology is discussed, which takes advantage of changes in HVS sensitivity with variations in viewing distance and ambient light. However, due to the lack of IQAs with surroundings consideration, image quality was assessed by capturing reference and processed images in the same environment for that research work. This approach proves to be inefficient and impractical, especially for large datasets.

Similar problems can be evident in emerging technologies, illustrated by the case of augmented reality (AR) or mixed reality (MR). While VR IQA [19] exists to evaluate the quality of VR content, it can be ineffective for AR or MR. As AR or MR technology merges the real world with the virtual world, virtual visual content needs to be clearly visible amidst real-world objects. An example of such a scenario is an automotive head-up display (HUD), where vital information is provided using transparent displays. Hence, new IQAs are required to evaluate the quality of such virtual content with consideration of the real world in the background. With the emergence of the metaverse concept, the importance of such IQAs can not be ignored.

Nonetheless, beyond the examples outlined in this section, there are several promising fields where the lack of proper IQAs hinders the growth of technologies to ensure the quality of their services while offering various advantages to human users.

3. COMPUTER VISION-BASED APPLICATIONS

Due to computer vision applications prioritizing accuracy over visual quality, various accuracy-oriented metrics are used due to the absence of appropriate IQAs. As the perception of visual media by HVS differs from that of computer vision, the performance of computer vision applications can not be ensured by perceptual quality resembling the perception of HVS. At the same time, accuracy metrics or mathematical algorithms are more suitable for such situations, as computer vision deals with factors like signal quality, data, and features in visual media. Pixel-wise accuracy, mean absolute error (MAE), MSE, and PSNR are examples of popular mathematical algorithms used in computer vision. However, these metrics may prove inadequate to evaluate application-wise requirements. In [20], an objective/goal-oriented IQA (GO-IQA) to support computer vision applications such as image segmentation has been discussed, pointing out the necessity of new IQAs. The research work on GO-IQA has notably addressed the ineffectiveness of focusing on perceptual quality for computer vision applications when the objectives significantly differ from those of HVS-based applications.

The necessity of task-oriented IQA development becomes evident when considering the broad application areas of computer vision. As illustrated in Fig. 2, computer vision tasks include image classification, image segmentation, object detection and recognition, as well as action and activity recognition, all of which find applications across various domains. The application areas of computer vision include education, healthcare, construction projects, commercial/industrial productions, agriculture, livestock operations, unmanned aerial vehicles, intelligent transportation systems, underwater activities, and so on. These tasks can suffer due to having images with noise, bad quality, inadequate contrast, or degraded structural/feature information. For these applications, whether the collected images are suitable for the desired task can be easily determined if appropriate IQAs are available. Furthermore, leveraging the scores of the appropriate IQAs, the pre-processing step can be invoked to process images according to application requirements. However, till now, most cases focus on distortion-free image collections with sufficient resolution, which indicates focusing on visual quality but may not always be useful, as addressed earlier.

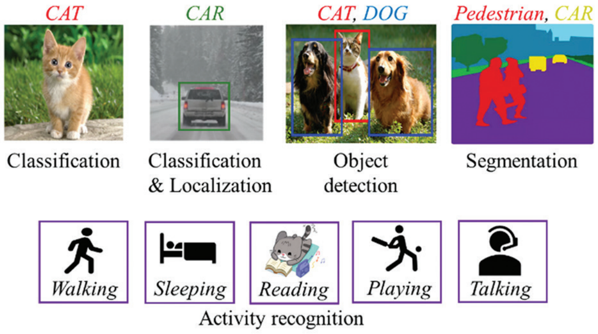


Fig. 2. Examples of different computer vision tasks

Another example highlighting the limitations of relying solely on perceptual IQAs is found in adversarial examples. As adversarial attacks introduce barely visible perturbations to images, HVS-based IQAs usually suggest good quality. However, these perturbations can significantly impact machine learning or deep learning models. If application-oriented IQAs are available for assessing quality from a computer vision perspective instead of HVS, they can identify the poor quality of corrupted images for that specific task. Similarly, steganography, which conceals important information within visual data while maintaining its invisibility, requires IQA to assess its imperceptibility. However, as addressed in [21], existing IQA methods appear inefficient for such an objective, indicating the necessity of appropriate IQA development. Furthermore, similar to HVS-based applications, computer vision-based applications require surrounding considerations. For example, unique metrics such as underwater image contrast measure and underwater image quality measure are suitable for evaluating images collected in a blueish aquatic environment [22].

New technological advancements in computer vision address the necessity for new IQAs. For instance, text-to-image generation, initiated from the urge to improve user experience and facilitate different sectors while presenting an economical solution for the content creators, requires appropriate IQAs. While objective NR IQAs may be used for evaluating artificial intelligence (AI) generated visual content, they often fall short as they primarily focus on real or natural scenarios during their development. Consequently, new IQAs tailored to artificial scenarios have emerged, including Fréchet inception distance (FID) [23] as an algorithmic approach and learned perceptual image patch similarity (LPIPS) [24] as a deep learning-based approach. Synthetic images are another example of AI-generated visual content widely used in computer vision, especially for training AI-based models. As synthetic images differ from natural images, conventional IQAs concerned with real-world images are inadequate. Moreover, as synthetic images are generated for computer vision-oriented tasks, modifications are required even for utilizing perceptual metrics [25].

All these discussions mentioned above highlight the challenges associated with using perceptual metrics for computer vision applications and signify the necessity of developing IQAs considering the specific requirements of computer vision. The discussions also indicate the attempts of IQA development to align with technological developments, even for computer vision-based applications.

4. DISCUSSIONS

Regardless of the application area, image processing applications often outpace the developments of IQAs. This issue can be a significant obstacle to technological innovations, leading to the use of subjective methods in most cases at the initial stage of any technological advancements. However, relying on subjective assessments is a time-consuming and expensive process, and it introduces the possibility of bias due to the personal preferences of the observers. The situation worsens when dealing with a large volume of images. In recent years, deep learning-based approaches have started to reduce the gap between IQA development and technological advancements. As deciding whether to opt for FR or NR IQA can be a concern for any application, a flexible deep learning-based IQA method is discussed in [26], offering the ability to switch between FR and NR IQA as needed. Deep learning-based approaches have also received notable attention for their success as NR IQAs [27], offering an alternative to the conventional complex modeling of IQAs required by theoretical perception-based schemes.

However, it is premature to envision predicting technological innovations and preemptively developing IQAs. Considering the requirement for sufficient datasets for both technological advancements and IQA developments, it is prudent to explore them concurrently whenever feasible. The simultaneous development of new technologies and corresponding IQAs can help to avoid the problems associated with the lack of proper IQAs and to ensure the effective and reliable performance of new technologies. Moreover, during the development of IQAs, it is essential to consider the characteristics of end-users, such as HVS or computer vision, to enhance their effectiveness. Developing a generalized IQA for both HVS and computer vision can be quite challenging. Therefore, one feasible solution is to focus on developing or modifying application-specific IQA development or modification can be one feasible solution. In addition to adhering to existing norms, the following issues can be considered in IQA development efforts.

- Is it tailored to a specific application?
- Does it account for all relevant distortions/artifacts?
- Are the distortions considered to be confined to their mathematical representations?
- Does it consider the sensitivity/responses of end-users (e.g., computer vision, HVS) from a broad perspective?
- Does it take relevant environmental factors into account?
- Is the available dataset comprehensive enough to cover all cases for verification?

As the aforementioned challenges are minimum requirements for IQA development, the research community always needs to find new factors affecting IQAs' effectiveness to keep up with technological advancements. AI-based approaches are promising in this regard due to their ability to learn. In [28], an interesting approach is presented for IQA development, where textual information is considered part of the quality assessment reference. This concept can be a potential pathway for future IQA development. In the future, a generalized AI-based IQA can be developed as a prospective solution, capable of taking text-based guidance or some critical factors into account to find appropriate principles and tune the model according to application requirements. This type of model can be considered as an NR IQA, where textual information is the reference for evaluating an image. The adaptivity of such IQA can be increased by incorporating the ability to extract requirements or guidance from pictorial references in the pre-processing stage to make the model applicable as an FR IQA. Nonetheless, considering the technological innovations, several sectors lack appropriate IQA methods. However, there are different potential ways for developing IQAs to accommodate these sectors, which are broader than the examples addressed in this work.

5. REFERENCES:

[1] H. Sheikh, A. Bovik, "Image Information and Visual Quality", IEEE Transactions on Image Processing, Vol. 15, No. 2, 2006, pp. 430-444.

[2] Z. Wang, "Applications of Objective Image Quality Assessment Methods [applications corner]", IEEE Signal Processing Magazine, Vol. 28, No. 6, 2011, pp. 137-142.

[3] G. Zhai, X. Min, "Perceptual Image Quality Assessment: A Survey", Science China Information Sciences, Vol. 63, 2020, pp. 1-52.

[4] W. Lin, M. Narwaria, "Perceptual image quality assessment: recent progress and trends", Visual Communications and Image Processing 2010, Vol. 7744, 2010, pp. 33-41.

[5] G. Zhai, W. Sun, X. Min, J. Zhou, "Perceptual Quality Assessment of Low-light Image Enhancement", ACM Transactions on Multimedia Computing, Communications, and Applications, Vol. 17, No. 4, 2021, pp. 1-24.

[6] W. Lin, W. Yuxuan, X. Lishi, C. Weiling, Z. Tiesong, W. Hongan, "No-reference quality assessment for

low-light image enhancement: Subjective and objective methods", Displays, Vol. 78, 2023, p. 102432.

[7] S. Cheng, H. Zeng, J. Chen, J. Hou, J. Zhu, K.-K. Ma, "Screen Content Video Quality Assessment: Subjective and Objective Study", IEEE Transactions on Image Processing, Vol. 29, 2020, pp. 8636-8651.

[8] A. George, S. J. Livingston, "A Survey on Full Reference Image Quality Assessment Algorithms", International Journal of Research in Engineering and Technology, Vol. 2, No. 12, 2013, pp. 303-307.

[9] K. Okarma, P. Lech, V. V. Lukin, "Combined Full-Reference Image Quality Metrics for Objective Assessment of Multiply Distorted Images", Electronics, Vol. 10, No. 18, 2021, p. 2256.

[10] X. Lv, T. Xiang, Y. Yang, H. Liu, "Blind Dehazed Image Quality Assessment: A Deep CNN-Based Approach", IEEE Transactions on Multimedia, Vol. 25, 2023, pp. 9410-9424.

[11] H. Yang, Y. Fang, W. Lin, "Perceptual Quality Assessment of Screen Content Images", IEEE Transactions on Image Processing, Vol. 24, No. 11, 2015, pp. 4408-4421.

[12] H. Chen, X. Chai, F. Shao, X. Wang, Q. Jiang, X. Meng, Y.-S. Ho, "Perceptual Quality Assessment of Cartoon Images", IEEE Transactions on Multimedia, Vol. 25, 2023, pp. 140-153.

[13] Y. Wang, G. Liu, L. Wei, L. Yang, L. Xu, "A Method to Improve Full-resolution Remote Sensing Pansharpening Image Quality Assessment via Feature Combination", Signal Processing, Vol. 208, 2023, p. 108975.

[14] B. Asheghi, P. Salehpour, A. M. Khiavi, M. Hashemzadeh, "A Comprehensive Review on Content-aware Image Retargeting: From Classical to State-of-the-art Methods", Signal Processing, Vol. 195, 2022, p. 108496.

[15] M. Xu, C. Li, S. Zhang, P. Le Callet, "State-of-the-Art in 360° Video/Image Processing: Perception, Assessment and Compression", IEEE Journal of Selected Topics in Signal Processing, Vol. 14, No. 1, 2020, pp. 5-26.

[16] K. Sim, J. Yang, W. Lu, X. Gao, "Blind Stereoscopic Image Quality Evaluator Based on Binocular Semantic and Quality Channels", IEEE Transactions on Multimedia, Vol. 24, 2022, pp. 1389-1398.

[17] T. Zhan, J. Xiong, J. Zou, S.-T. Wu, "Multifocal Displays: Review and Prospect", PhotoniX, Vol. 1, 2020, pp. 1-31.

[18] M. A. Rahman, J. You, "Human visual sensitivity based optimal local backlight dimming methodologies under different viewing conditions", Displays, Vol. 76, 2023, p. 102338.

[19] A. K. R. Poreddy, R. B. C. Ganeswaram, B. Appina, P. Kokil, R. B. Pachori, "No-Reference Virtual Reality Image Quality Evaluator Using Global and Local Natural Scene Statistics", IEEE Transactions on Instrumentation and Measurement, Vol. 72, 2023, pp. 1-16.

[20] S. Kiruthika, V. Masilamani, "Goal Oriented Image Quality Assessment", IET Image Processing, Vol. 16, 2022, pp. 1054-1066.

[21] De R. I. M. Setiadi, S. Rustad, P. N. Andono, G. F. Shidik, "Digital Image Steganography Survey and Investigation (Goal, Assessment, Method, Development, and Dataset)", Signal Processing, Vol. 206, 2023, p. 108908.

[22] N. V. Dharwadkar, A. M. Yadav, M. A. Kadampur, "Improving the Quality of Underwater Imaging using Deep Convolution Neural Networks", Iran Journal of Computer Science, Vol. 5, No. 2, 2022, pp. 127-141.

[23] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter. "GANs trained by a two time-scale update rule converge to a local nash equilibrium", Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4-9 December 2017, pp. 6629-6640.

[24] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18-23 June 2018, pp. 586-595.

[25] J. Chaudhary, D. R. Pant, S. Pokharel, J.-P. Skön, J. Heikkonen, R. Kanth, "Image Quality Assessment by Integration of Low-level & High-level Features:

Threshold Similarity Index", Proceedings of the IEEE 31st International Symposium on Industrial Electronics, Anchorage, AL, USA, 1-3 June 2022, pp. 135-141.

[26] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, W. Samek, “Deep Neural Networks for No-reference and Full-reference Image Quality Assessment”, IEEE Transactions on Image Processing, Vol. 27, No. 1, 2017, pp. 206-219.

[27] X. Yang, F. Li, H. Liu, "A Survey of DNN Methods for Blind Image Quality Assessment", IEEE Access, Vol. 7, 2019, pp. 123788-123806.

[28] Y. Watanabe, R. Togo, K. Maeda, T. Ogawa, M. Haseyama, "Assessment of Image Manipulation using Natural Language Description: Quantification of Manipulation Direction", Proceedings of the IEEE International Conference on Image Processing, Bordeaux, France, 16-19 October 2022, pp. 1046-1050.

A Study on A Novel Collision Risk Prediction Map for Maritime Traffic Surveillance Based on Ship Domain

Original Scientific Paper

Van Quang Nguyen
Vietnam Maritime University,
Department of Personel & Administration
Lach Tray Street, Haiphong, Vietnam
nguyenvanquang@vimaru.edu.vn

Tu Nam Luong*
Vietnam Maritime University,
Faculty of Navigation
Lach Tray Street, Haiphong, Vietnam
luongtunam@vimaru.edu.vn

Van Luong Tran
Vietnam Maritime University,
International School of Education
Lach Tray Street, Haiphong, Vietnam
tranvanluong@vimaru.edu.vn

*Corresponding author

Abstract *– Recently, a regional model for assessing the risk of multi-ship collision has been developed to reduce the risk of ship collision in territorial sea areas such as trade ports and entry waterways and improve the safety and efficiency of ship traffic. The focus is on marine traffic in the visualized waters with the risk of ship collision. However, due to the lack of information from experts with sufficient knowledge and experience in a given area, they also have some limitations in adequately and comprehensively representing the risk of collision, especially in busy waterways where encounters of more than two ships often appear. In addition, they could not visualize the location of the proximity collision and the exact risk value in real time. Therefore, to overcome the limitations of previous studies, this paper proposes a new regional collision risk visualization system, which combines density-based spatial clustering of applications with noise (DBSCAN) and analysis and knowledge-based ship domains and uses AIS data to intuitively and accurately map the dynamic collision risk of water areas at successive moments, predict areas where collisions can happen by dynamic risk index and warn the ships. Identifying high-risk collision areas between multi-ships can be enhanced using the developed system, which allows for reliable and accurate analysis to help implement safety measures.*

Keywords*:* *ship collision, collision risk, maritime traffic, heat map, ship domain*

Received: January 22, 2024; Received in revised form: April 26, 2024; Accepted: April 30, 2024

1. INTRODUCTION

Frequent encounters between multiple ships in congested waters are one of the main factors causing ship collisions. However, there are significant challenges to understanding complex multi-ship encounter situations. Most accidents are caused by human error, which cannot be prevented as long as people operate the ship. Collisions with ships at sea can cause severe loss of life and property, and some may even seriously impact the marine ecological environment. Therefore, detecting and judging collision risk is the primary task of intelligent navigation of ships. The increased ship traffic at a given moment can complicate ship traffic, make congested waters more congested, and increase the likelihood of ship collisions.

Multi-ship encounters at sea are highly complex and uncertain, which can be a big challenge for ships. A quick and correct grasp of the current situation is needed to make appropriate maneuvering decisions and perform well in multi-ship encounters. Most recent frameworks need to provide a general picture of a complex problem. Influencing factors such as speed, heading, environment, and maneuverability should be considered. The size-related factor that affects the encounter, such as the ship's length, should also be considered. The larger the size, the higher the risk of collisions between ships at similar distances.

It is necessary to evaluate the potential danger between ships in real time [1] and express the collision risk in the chart, which can also provide a better reference for collision avoidance operations [2]. Many researchers have recognized that it is helpful for navigators to be vigilant about real-time collision risks, facilitating decision-making. A correct and complete understanding of complex multi-ship encounters is an essential means and premise to prevent ship collisions and ensure the safety of ship navigation.

To solve the above problems, this study proposes a new framework for early detection of collision risk between ships in congested waters to notify the Officer on Watch (OOW) or Vessel Traffic Service Officer (VTSO) of potential collisions so that the person in charge can make decisions by observation. Considering the area's complexity, a new ship domain (SD) concept will determine the near collision. Near collision is the situation when SDs overlap. Overlapping areas can be displayed as indexes. In addition, a method of using an index to display the areas where the ship may collide with other ships is proposed, which is presented in the form of a collision heat map so that OOWs and VTSOs can quickly identify the danger areas where collisions may occur in congested waters. As a result, risk awareness has also increased. The main contributions of our framework are:

- Propose Heat Ship Domain can calculate continuous and dynamic potential collision risk levels around a ship in a maritime waterway. It considers the ship's static and dynamic characteristics and experts' knowledge of particular water.
- Propose a Dynamic Collision Risk Index that calculates the impact of multiple ships on an area and detects high-value index areas by using the overlap function of the Heat Ship Domain.
- Propose a Collision Risk Prediction Map that assesses a waterway's regional risk and identifies and displays high-risk hotspots based on the Dynamic Collision Risk Index. This map would allow OOWs and VTSOs to make evasive decisions for collision avoidance by observing it.

The remainder of the paper is organized as follows: Section 2 provides a literature review of work related to collision risk alert systems, focusing on their merits and demerits. Section 3 discusses the methodology, including techniques and algorithms for constructing a new SD based on the Kernel Density Function and using this SD to establish a collision heat map that displays collision risk areas in congested waters. Section 4 presents a case study of multi-ship encounter scenarios to validate the proposed methodology. Finally, Section 5 presents conclusions about the results.

2. RELATED WORK

Ship collision risk identification is attractive, particularly in specific water areas. To improve the safety of navigation, many scholars have been studying solutions to visualize the geographical distribution of collision risk. To achieve this goal, the risk of collision between ships should first be accurately quantified. In various models and navigation practices, the distance to the closest approach point (DCPA) and the time to the closest approach point (TCPA) are the essential criteria for "collision risk" and the most critical parameters [3]. However, DCPA and TCPA are only used to show the risk of collision based on subjective judgment and cannot generate applicable quantitative values for collision risk.

Furthermore, DCPA and TCPA are hard to apply in busy waterways with high ship density. Therefore, another index, named the Collision Risk Index (CRI), was introduced, which evaluates the probability of a collision [4]. Fuzzy logic has been used in various collision risk assessment and collision avoidance decision-making methods [5, 6]. However, geometric information needs to be considered, and the risk of collision with multiple ships is not discussed. The velocity obstacles-based framework assesses the risk of colliding velocities [7, 8]. Although this research has effectively calculated collision risk, they can only apply in open sea areas. In congested waterways with smaller room for ship maneuvering and complex conditions, the collision risk could be more comprehensively achieved, especially when there are more than two ships.

Still, it is difficult to assess the risk of a collision given the geographical conditions. Although these methods can determine collision risk based on encounter conditions and ship maneuvering, they need evaluation results that combine geometric information. Therefore, the concept of "ship domain" is used to find available maneuvering space and geometric information can be used to assess collision risk. The ship domain was first defined as "the domain around a ship underway which most navigators of following ships would avoid entering" [9] and "the effective area around a ship which a navigator would like to keep free to other ships and stationary objects" [10]. The ship domain model is a quantitative tool for assessing the risk of collision when another ship intrudes into the range of another ship and has been widely used for different purposes, such as ship collision avoidance [11, 12], near misses, and hotspot identification [13, 14]. In the above models, ship domains are assumed deterministic, and the domain parameters have not been extended. If a target is outside the boundary, it is safe; no action needs to be taken; if the target is inside the boundary, it is dangerous, and action must be taken to keep it out of the boundary. The collision risk is only 0 and 1; the level of risk can not be exactly presented. There are some advantages and disadvantages to ship domain model applications. Subjectivity is a problem in knowledge-based and analysis-based ship domain models, as the models rely heavily on the judgment of navigators, experts, and researchers—the human factor is not considered in the empirical ship domain model [15]. The human factor is crucial because when a ship has an accident, the leading causes are human error, which is caused by insufficient knowledge of the operation, receiving wrong or inadequate information to make judgments, or unfamiliarity with the environmental characteristics of the water area [16]. A framework is needed to formulate a mathematical model that considers ship size, speed, and a human factor component.

To improve safety, several studies have developed a framework for assessing regional collision risk, which combines density complexity and multi-shi collision risk. The risk of a collision off the coast of Portugal is evaluated by predicting future distances between ships based on AIS data. This approach can be only used to identify collision candidates in complex traffic patterns in the long term but not in real-time [17]. Maritime traffic around the Shetland Islands is visualized in the form of AIS pings maps, ship density maps, ship trajectory maps, ship length maps, etc., to ensure the safety of navigation in marine space and development planning [18]. The molecular collision theory establishes an encounter probability map of the Istanbul Strait [19]. After summarizing the risks through the radial distribution function, the spatial interpolation technique was used to identify the geographical distribution of the collision risks in the Bohai Strait [20]. While these methods effectively visualize maritime traffic in waters, some things could be improved. The spatial distribution of the encounter probability within these models is calculated for large areas. This means that the whole area will have the same index value.

A kinematics feature-based vessel conflict ranking operator is introduced to evaluate ship collision risk by integrating the relative position vector and the relative velocity, accounting for static and dynamic information of AIS to quantify ship collision risk and identify high collision risk areas. However, this paper needs to consider the impact of multi-ship, which is only available in open-sea regions [21]. Another method for identifying ship navigation risks is combining the ship domain with AIS data to increase collision risk identification prediction accuracy for ship navigation in complex waterways. This method constructs a ship domain model based on the ship density map drawn using AIS data. Then, the collision time with the target ship is calculated based on the collision hazard detection line and safety distance boundary, forming a method for dividing the danger level of the ship navigation situation. The risk level is only evaluated when the target ship is inside the outside ship domain and the intersection of the boundary [22]. Fuzzy logic calculates the collision avoidance maneuver for the selected ship, considering the closest point of approach, relative bearing, and the ship's speed. Evaluate the collision risk and navigation situation based on COLREG rules, sort the target vessels, and determine the most dangerous vessel. Multi-ship encounters are considered but only in the vast open sea [23]. An anchorage collision risk model was established in microscopic, macroscopic, and complexity aspects, which considered ship relative motion, anchorage characteristics, and ship traffic complexity. In modeling complexity, it would be better to incorporate the factors of ship motion to make the consideration of traffic complexity more sufficient [24]. A dynamic elliptical ship domain based on AIS data combines the relative motion between ships in different encounter situations to assess the level of ship intrusion in the domain. However, during the movement, the size of the ship domain is static, not changing with speed [25].

Furthermore, these studies have been used for maritime traffic analysis, and the risk index used for the location of future collisions needs to be taken into account. It is difficult to distinguish the collision risk between collision candidates and obtain an accurate collision risk value. In addition, identifying appropriate potential collision areas in congested waters is challenging. Still, no standard scale for measuring the degree of collision criticality exists. Therefore, adopting specific criteria to determine the collision risk and warn the OOWs and VTSOs is critical.

Based on these circumstances, this study developed a clustering-based regional collision risk prediction model for a collision area using a new ship domain and considering geographical patterns. The method aims to detect collision risk quickly and dynamically using AIS data from short intervals. The establishment of our model is presented in Section 3.

3. PROPOSED METHOD

Collisions in highly complex maritime traffic pose significant risks. In high-density waters, it is expected to encounter a group of ships. If the complexity exceeds the threshold, the likelihood of a near collision rises significantly. In crowded waters, the ship risks colliding with multiple target ships. The degree of collision between multiple ships needs to be integrated. The degree of danger increases with the number of target ships at risk of collision. In this section, the methodology of real-time collision risk assessment indicators using the ship domain is developed as the basis for the safe navigation of ships. The diagram of the proposed framework is shown in Fig. 1.

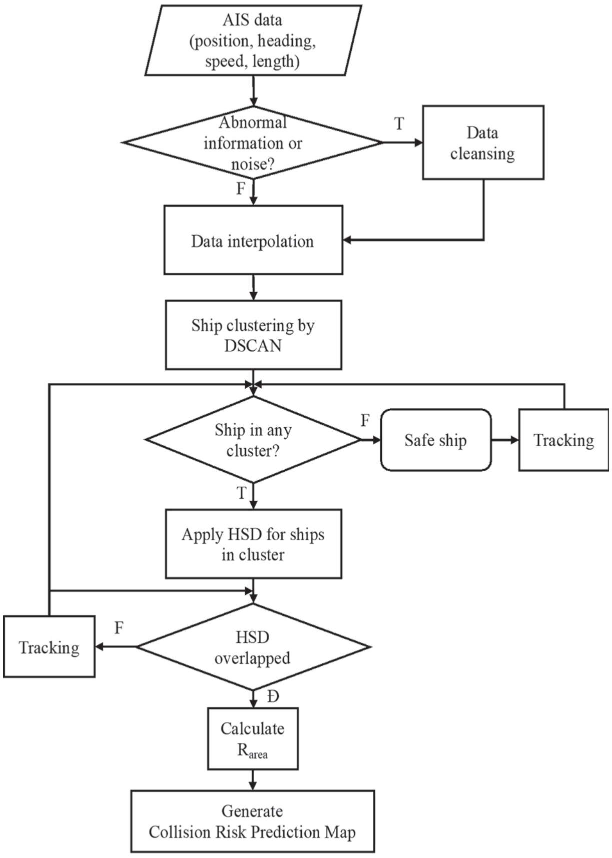


Fig. 1. Diagram of proposed framework

3.1. DATA PROCESS

AIS data is increasingly being used as a valuable source of information on ship traffic in maritime traffic engineering and maritime transport safety studies. The AIS identifies each ship equipped with an AIS transmitter and transmits static information about the ship (call sign, IMO number, destination, cargo, etc.) and frequent updates about the ship's position, speed, and heading [26]. The AIS data is decoded by extracting position, speed, heading, size, time, and MMSI information.

The main task of this step is to make the AIS data reliably used for calculation. More specifically, AIS data should be updated and interpolated correctly over time intervals.

Abnormal information or noise can significantly affect the regional ship collision risk assessment. Therefore, given the integrity of the real-time data, it is not appropriate to delete those noisy records, which should be cleaned and updated. A 4-step process was used for the data cleansing step in this study [12]. According to Newton's laws of motion, the average speed can be calculated as the ratio of the distance traveled to the travel time. Therefore, the ship's position recording and acceleration and deceleration capabilities can be used to check whether the speed record is within a reasonable range. Correspondingly, the updated speed data can be used to clean the location data based on the same principle.

Step 1: Check the reasonableness of the speed data.

Step 2: Update the irrational speed data.

Step 3: Check the reasonableness of position data.

Step 4: Update the position data.

In addition, AIS data is sent randomly at different times. To prevent the temporal dispersion of AIS data, the updated AIS data is interpolated every 30 seconds to obtain information simultaneously [27]. Calculate the movement of each ship and estimate the position in 30 seconds. After extensive cleansing and pre-processing, the original AIS database becomes a suitable dataset for analysis.

3.2. SHIP DOMAIN CONSTRUCTION

Given the complexity, this subsection intends to construct a new ship domain to identify a potential collision that is defined as one that occurs when the ship domains of the local ship (OS) and the target ship (TS) overlap, as shown in Fig. 2.

The concept of ship domain is reflected in COLREG, where ships must pass through each other and obstacles at a safe distance. This safety distance represents the domain of the ship. Several ship domains have been proposed to express the hazard level within the domain [28-30]. However, choosing the size and shape of the ship domain best suited for ship navigation takes a lot of work, especially in congested waters. Most of the existing collision risk identification methods are based on a geometric perspective and use indicators such as distance to measure collision risk, which requires more expert knowledge.

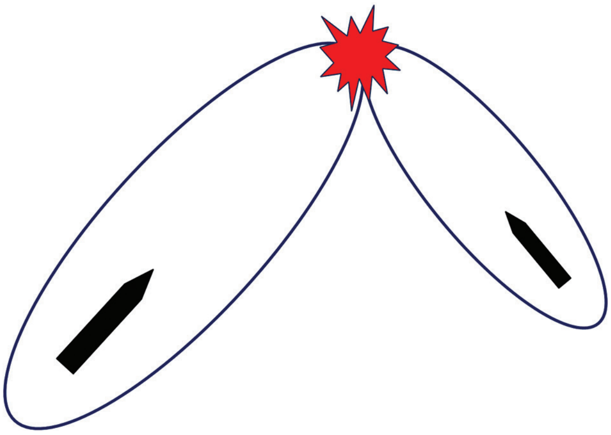


Fig. 2. Ship domains overlap

The ship's domain is essential for classifying according to the severity of the encounter since the encroachment on the domain implies a certain level of proximity that the navigator usually wishes to avoid. In addition, other contextual characteristics should be considered. In meetings between ships, larger ships have larger domains. Each ship has its turning circle. The turning circle of a ship determines the ease and rapidness with which a ship can change its course or direction. The greater the size, the larger the turning circle. Larger ships need more space for maneuvering than smaller ones. Some authors use geographic information technology and AIS data to calculate how the space around ships is used (or kept free) during their movements. They have found that the ship's size affects its domain, and smaller ships tend to meet slightly closer than larger ones. This means that safety areas increase along with the size of the ship. One can notice that the bigger the ship, the bigger the domain becomes. This means that for larger ships in a given encounter, the situation may be classified as dangerous for larger ships but not for smaller ships, where the situation may still be considered safe. This problem can be solved using a new ship domain using field theory.

The concept of "field" is abstracted as a mathematical concept used to describe the distribution of a particular physical quantity or mathematical function in space. There are both connections and differences between the various fields. Fields can be expressed abstractly with mathematical models. Any object can form a field, and different objects produce different fields. The field theory-based method has been widely used in vehicle safety research, but there are relatively few research results in the safe navigation of ships [31]. Due to the differences in traffic characteristics between navigation areas, using one type of ship domain for each area is difficult. Therefore, inspired by field theory, a new ship domain is introduced to measure the degree of collision risk around ships. The new ship domain applies to regions and the probability levels of advanced decision-making systems better suited for navigational risk detection.

The coordinates of the ship encounter are shown in Fig. 3. The origin represents the own ship (OS). It is located in the OS's center, and the OS's velocity vector relative to the target ship constitutes the Y-axis through the origin. The X axis is perpendicular to the Y axis and passes through the origin. The line of motion of the target ship (TS) relative to the OS is parallel to the Y axis. Assuming that the TS is located at point P with coordinates (x_p, y_p), then point A is the projection of P on the X-axis. The PA line is the relative line of motion of the target ship at P relative to the OS, and the projection point A is the CPA of the target ship to the OS, i.e., the distance (d_{OA}) from the origin O to point A is the DCPA from the operating system to the target ship. The distance from point P to point A (d_{PA}) is from the target ship to the CPA, which is the product of the time required to reach TCPA and the speed V [32]. Moreover, assume that except for real target ship P, there are a lot of imaginary target ships P_n with no speed. The non-dimensional of d_{OA} and d_{PA} are calculated as follows:

$$d'_{OA} = \frac{DCPA_p}{L} = \frac{|x_p|}{L} \qquad (1)$$

$$d'_{PA} = \frac{TCPA_p \times V}{L} = \frac{|y_p|}{L} \qquad (2)$$

where

d'_{OA} is non-dimensional of dOA;

d'_{PA} is non-dimensional of dPA;

V is the speed of the ship;

L is the length of the ship;

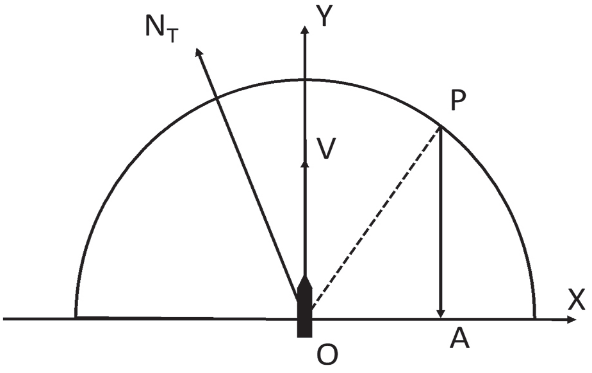


Fig. 3. The coordinates of the ship encounter

Kernel density estimation (KDE) generates a smoothed empirical probability density function based on individual locations across all sample data. This estimate better represents the "true" probability density function of a continuous variable [33].

The radial kernel estimator is based on the Euclidean distance between an arbitrary point {x,y} and sample point {x_i, y_i}, i = 1,2, ..., n:

$$\hat{f}(x,y) = \frac{1}{nh_xh_y}\sum_{i=1}^{n} K(\sqrt{\left(\frac{x_i - x}{h_x} + \frac{y_i - y}{h_y}\right)^2}) \qquad (3)$$

where

n is the number of sample points;

K is the kernel function;

h_x, h_y are the smoothing parameters in the X-axis and Y-axis.

KDE is applied to establish a new dynamic collision risk (DCR) around OS, employed d'_{OA} and d'_{PA}. DCR value is calculated for every point around a ship:

$$DCR(P) = \frac{1}{nh_xh_y}\sum_{i=1}^{n} K(\sqrt{\left(\frac{d'_{OA}}{h_x} + \frac{d'_{PA}}{h_y}\right)^2}) \qquad (4)$$

Assume that the smoothing parameter in X-axis and Y-axis have the same value ($h_x = h_y = h$). After applying the asymmetric Gaussian function as a kernel function, the DCR of a point P can be expressed as follows:

$$DCR(P) = \frac{1}{nh^2}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_p+y_p)^2}{2L^2h^2}} \qquad (5)$$

where

$DCR(P)$ is the Dynamic Collision Risk Index of point P;

h is the influencing parameter, which represents water areas and is named the area parameter.

One ship will have different sizes of ship domains in different navigation areas.

DCR is proposed to analyze and measure the collision-risk degree of every point around OS, including real and imaginary target ships, considering DCPA, TCPA, ship length, and speed.

The risk index at a point is related to the coordinate value at this point. Every point in the vicinity of the ships has a value of DCR, and points with the same value of DCR will be shown as contour lines or the same color to indicate the same degree of risk. Fig. 4 shows the visualization of DCR in two forms: contour lines or colors of DCR. DCR can be depicted as an elliptical area surrounding the ship, consisting of multiple levels of risk. This area is called the Heat Ship Domain (HSD). Each level within the HSD is represented by a color or isotherm that connects the points with the same value in the field to make the iso risk index line. The fundamental concept is that when the ships move closer, their respective ship domains will overlap. The overlapped area, which can be understood as the potential collision area, is the collision position. This area will vary at different moments, and the changes in the overlapped area also demonstrate the degree of collision probability at the moment of encounter. In case of a multi-ship encounter, at each moment when ships are approaching each other, values of DCR of points between these ships increase according to their positions to ships. The influence of these ships in this area is higher than in other areas. Due to the change of DCR, the color describes the area with a probability of collision accident changing from cool to hot. The DCR is a cost-like value. It tends to be higher for the higher of the collision risk. The points with high values of DCR indicate that there will probably be a collision there if the ships involved do not perform the evasive action and the magnitude of the action required to clear the situation.

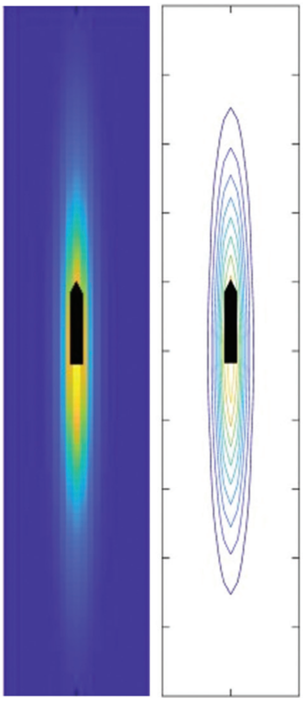


Fig. 4. Appearances of Heat Ship Domain

The speed and heading of the ship determine the extension and direction of the longitudinal axis of HSD. The area parameter also specifies the coverage of HSD. The next step is to determine the edge of HSD, corresponding to iso risk line 0,1 (DCR = 0,1).

Many factors affect the size of the ship domain, but only a few can be considered in the domain size determination process for practical reasons. The first is the human factor, which includes navigators' skills, knowledge, and mental and physical abilities [34]. Also, according to experts, another critical factor is the type of water used [35].

Four points are analyzed for the boundary of HSD to reveal the ship’s passing distance. The domain proposed in this paper considers ship speed, ship length, and area parameter h. For a ship of a specific size and speed, different h will lead to varying edges of HSD, corresponding to iso risk line 0,1 (DCR = 0,1) (as shown in Fig. 5). The value of h depends on navigators and the characteristics of the water area and can be determined by expert knowledge methods.

The navigator's knowledge of the assessment of the navigation situation provides the basis for determining the safe distance between ships in one particular area with lengths overall, respectively: under 115m, 116 – 145m, 146 – 175m, and over 175m m. Suppose ships are traveling at a speed of 8 knots. An expert study was conducted to assess ships' encounters in Haiphong Port waters in conditions of good visibility. The participants were navigators, including captains, OOWs, sea pilots, and VTSOs with different sea experiences. The study took the form of a questionnaire.

Respondents were required to give answers about safety distances in four directions: fore (a), aft (b), port (c), and starboard (d) from OS with four ship lengths, as mentioned in Fig. 6.

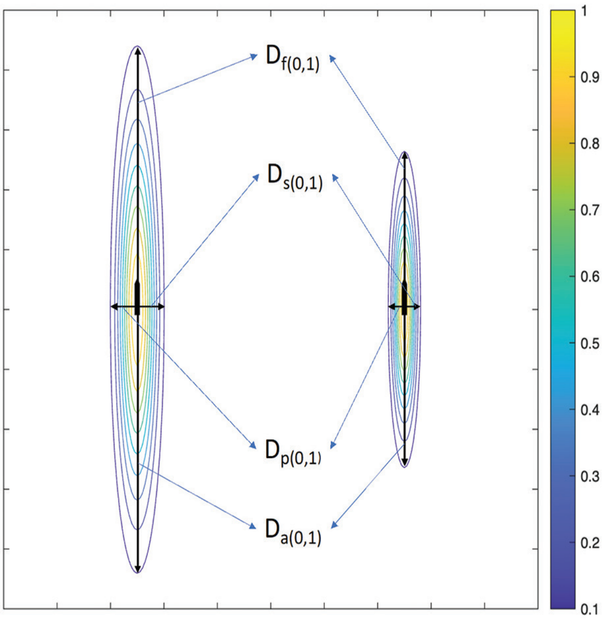


Fig. 5. Size of HSD changes with area parameter

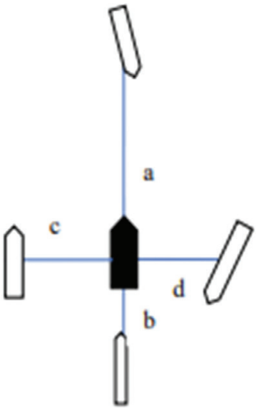


Fig. 6. Safe distances around a ship

The survey was conducted over three months and had more than 300 participants. The results of the survey are shown in Table 1-4. With each type of ship, the safe distances from OS to four directions are collected as minimum, maximum, and mean values. The pilots have a lot of experience navigating in the study area, so they believe the safe distance between ships in front of the ship can be manageable. On the contrary, a more considerable distance is required for the officers, especially the 3rd officer, to ensure no collision risk. According to experts, the dangerous distance on both sides has relatively uniform results. Due to the characteristics of Haiphong Ports water – narrow width (average 80m) and depth, it is unavailable for the long passing distance from starboard and port sides. It can be seen that, in this area, experts all believe that the collision avoidance distance on both sides is similar. Some empirical studies that used the AIS data in narrow channels or restricted areas to measure the space around a ship during navigation have shown different results in the size of ship domains. However, the dangerous distances on both sides of the ship are similar.

Table 1. Safe distances of ships under 115m

	Minimum (m)	Maximum (m)	Mean (m)
Fore (a)	350	900	580
Aft (b)	350	900	520
Port (c)	15	40	20
Starboard (d)	15	40	25

Table 2. Safe distances of ship 116 – 145m

	Minimum (m)	Maximum (m)	Mean (m)
Fore (a)	400	900	670
Aft (b)	400	900	560
Port (c)	15	40	22
Starboard (d)	15	40	25

Table 3. Safe distances of ship 146 – 175m

	Minimum (m)	Maximum (m)	Mean (m)
Fore (a)	450	1300	750
Aft (b)	450	1300	650
Port (c)	20	40	28
Starboard (d)	20	40	30

Table 4. Safe distances of vessels over 175m

	Minimum (m)	Maximum (m)	Mean (m)
Fore (a)	600	1600	850
Aft (b)	600	1600	780
Port (c)	20	40	32
Starboard (d)	40	40	35

These survey results will be used as subject data for the approximation process to determine area parameter h. As can be seen, a particular area will have a value of h. For this aim, the Least Squares method is used to find very close or the same solution as the optimal values [36]. The boundary of the iso risk index line 0,1 of HSD will be differently generated by each value of h and compared with the collected data. The best-fit boundary takes the minimum value of the fitness function and returns the optimal value of h. The fitness function is as follows:

$$\begin{cases} \Delta d_f(h) = (D_{f(0,1)}(h) - a)^2 \\ \Delta d_a(h) = (D_{a(0,1)}(h) - b)^2 \\ \Delta d_s(h) = (D_{s(0,1)}(h) - c)^2 \\ \Delta d_p(h) = (D_{p(0,1)}(h) - d)^2 \end{cases} \tag{6}$$

$$\Delta d(h) = \sum_{i=1}^{4} \Delta d_i(h) \tag{7}$$

where

$D_{f(0,1)}$, $D_{a(0,1)}(h)$, $D_{s(0,1)}(h)$, $D_{p(0,1)}(h)$ are the radii of iso risk index line 0,1 of HSD in fore, aft, starboard, and port side, respectively;

$\Delta d_f(h)$, $\Delta d_a(h)$, $\Delta d_s(h)$, $\Delta d_p(h)$ are squared distance differences at fore, aft, starboard, and port side, respectively;

$\Delta d(h)$ is the sum of squared distance differences.

The fitness function of the approximation for the size of iso risk index line 0,1 of HSD in the study area involving area parameter h is calculated by Equation 7. The parameter value of h can be estimated using the sum of squared distance differences between radii in the proposed model and the surveyed data. It means that after finding the minimum value of the fitness function ($min\Delta d$), the corresponding parameter can be selected as in Table 5.

Table 5. Area parameter h for different ship lengths

Ship length	Under 115 m	116-145 m	146-175 m	Over 175 m
h	0.22	0.21	0.2	0.2

The motivation of this study is to utilize the overlap between ship domains to identify potential collision risks and visualize this area. A questionnaire about the visualization of HSD was carried out with VTSOs in Haiphong Ports water. They realized that if the elliptical-shaped domains are used, it will lead to a misunderstanding of the collision situation for VTSOs and confuse them in cases where a ship was following or crossing aft of another, the forward part of one ship domain overlaps the aft part of another ship's domain (as shown in Fig. 7).

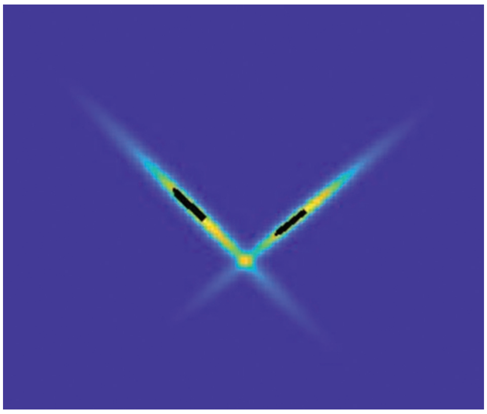


Fig. 7. False potential collision area

Although an overlapped area happens, there is no collision risk in this case. Therefore, the aft part of HSD will be shortened to the value of the starboard and port sides, according to the questionnaire with VTSOs, as shown in Fig. 8, and samples of HSD with speed 8kn in Haiphong Port water are presented in Fig. 9. Different ship lengths will have different sizes of HSDs.

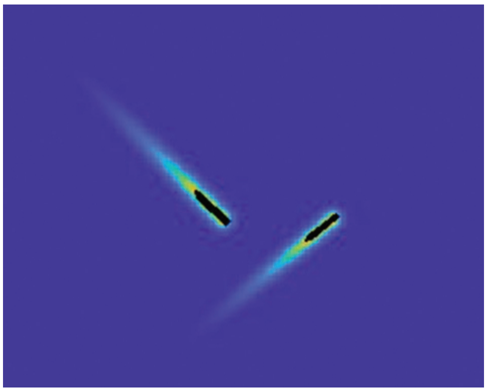


Fig. 8. Shorten HSDs

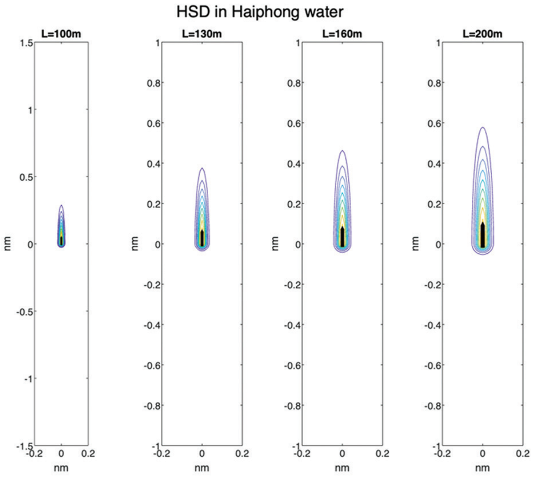


Fig. 9. HSDs of ships with different lengths in Haiphong water

The resulting past ship domain with speed 10kn in restricted areas is reproduced in Fig. 10 along with the edge of HSD, corresponding to iso risk line 0,1. The comparison shows that the ship domain of Coldwell [37] is reasonably compatible with the proposed ship domain on the forward but overestimates the space requirement on the starboard and port sides. Other ship domains have shorter fore parts, but three other sides appear overestimated. The reason is that HSD is constructed based on the characteristics of the Haiphong waterway based on experts' experiences. It should be noted that the Fujii et al. [9] and Hansen et al. [11] models still cannot take into account the effect of change in speed, while Wang and Chin's model does not enlarge significantly with increasing speed [38]. With the more extended fore parts and changing during navigation, HSD is suitable for early detection risk, especially when crossing narrow channels.

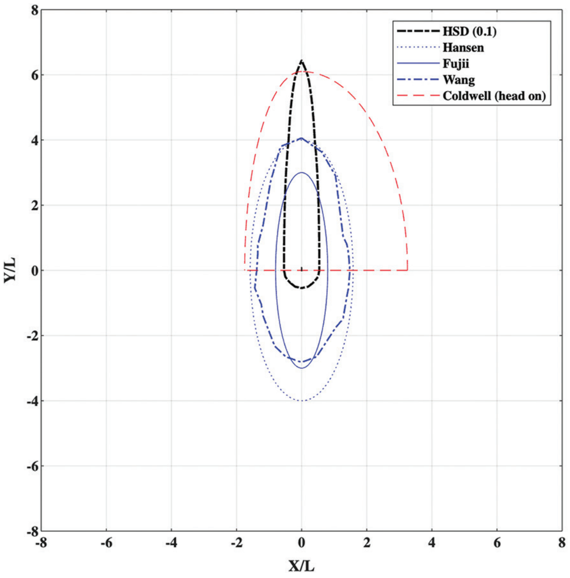


Fig. 10. Comparisons of HSD and restricted area ship domains

One problem is that if many HSDs appear simultaneously, it is easier for OOWs or VTSOs to grasp the dangerous areas quickly. Thus, the method for grouping encounter ships to apply HSD will be presented in the next section.

3.3. IDENTIFYING ENCOUNTER SHIPS BY DBSCAN CLUSTERING

DBSCAN is a clustering algorithm that considers data with a constant density in the same cluster or group [39]. Each data item represented by a coordinate point is divided into three types: core (red dot), boundary (orange dot), and noise (black dot), as shown in Fig. 11.

This algorithm is suitable for ship spatial clustering compared to others because it generates clusters in any shape and can eliminate the noise, which can be considered the singular object in space [40]. There are two main parameters in DBSCAN: Eps, which refers to the radius of the neighborhood area, and MinPts, which refers to the minimum number of objects in a cluster. With clustering, as shown in Fig. 12, the calculation of collision risk can be simplified because it is no longer necessary to consider every ship. In addition, it can also be found that some ships are not in any clusters, are considered noise ships, and are eliminated in the process.

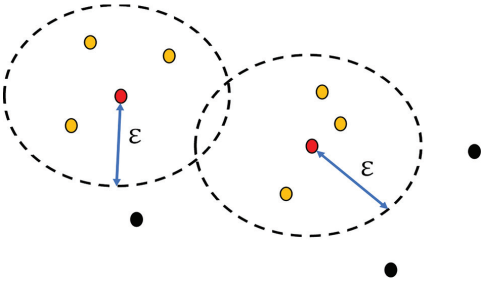


Fig. 11. Illustration of DBSCAN

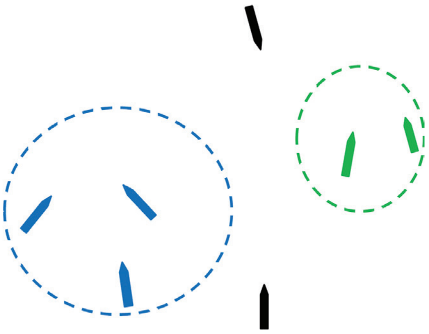


Fig.12. Ship clustering by DBSCAN

Assume x is a data point in the radius Eps of data set X. x can be defined as follows:

$$x = x_{core} \; if \; N(X) \geq MinPts \forall x \epsilon X;$$
$$x = x_{border} \; if \; N(X) < MinPts \forall x \epsilon X; \qquad (8)$$
$$x = x_{noise} \; if \; N(X) \geq MinPts \forall x \epsilon X;$$

To establish a collision risk prediction map, HSD will be applied to ships identified as core or borders. When the ship leaves its cluster, HSD will not be employed. Although the collision risk around the noise ships is not calculated, they are still tracked until the distance from them to others is smaller than the defined radius. This function will make the high collision risk areas or hot spots easier for users to focus on.

3.1. ESTABLISHING COLLISION RISK PREDICTION MAP

This section proposes a process for presenting the geographical distribution of collision risk or a regional collision risk prediction map. In busy waters, the ship risks colliding with multiple target ships. Therefore, there is a need for a function to estimate the area with a high possibility of collision. The degree of danger increases with the number of target ships at risk of collision. These potential collision areas, or hot spots, are where HSDs overlap, and their indexes (R_{area}) are calculated by Equation 9.

$$R_{area} = \sum_{i=1}^{m} DCR(P_i) \qquad (9)$$

where

R_{area} is a collision risk index of overlapped areas;

$DCR(P_i)$ is the collision risk index of point P_i;

m is a number of ships with HSDs that impact point P_i.

This collision risk index of domain-overlapped areas can measure the detail level or degree of collision risk. The greater the index, the more likely the collision will happen if there is no change in the speed and course of at least one ship. If there is no overlap between HSDs, the status of the encounter situation can be suggested as no collision risk. If an overlapped area happens (color becomes hotter), the situation can be considered a pre-collision state if the distance between ships is smaller. The effectiveness of the map with risk prediction function will be verified in Section 4.

4. CASE STUDY AND RESULTS

At present, there are three types of collision risk assessment methods: risk at the micro, macro, and regional level. From a micro point of view, the risk of collision between the pair of ships is usually calculated so that the ships can take action to avoid the collision. However, there are always many ships in the congested sea area, and the possibility of collision between more than two ships will occur frequently. Therefore, this study uses actual AIS data of ships sailing in the waters of Haiphong Ports for experiments to illustrate and validate the proposed model's effectiveness in macroscopic and regional views. Haiphong Ports are located in Northern Vietnam, and there is a vast and rapidly increasing amount of goods due to the development of Vietnam's economy, especially the development of the sea economy. It leads to a rise in the daily flow and density of ships and limits the maneuvering space between ships. Therefore, an urgent need is to analyze and assess the risk of ships colliding in this water.

Geographically, the studied water area is positioned between latitudes 20°46'28.31" N to 20°52'12.57" N, 106°43'36.11" E to 106°55'37.91" E (in Fig. 13).

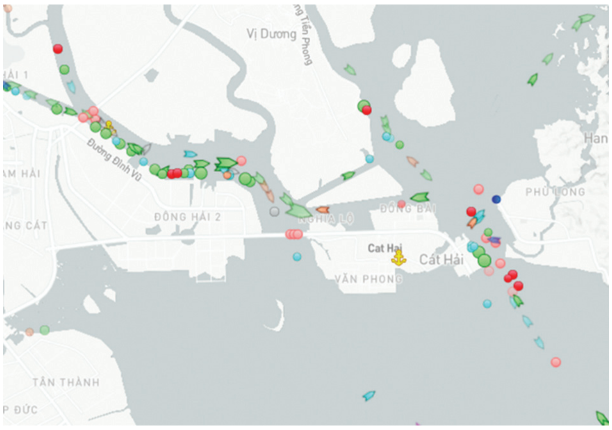


Fig. 13. Study area

In recent years, the growth in the amount of goods and the number of ships passing through Haiphong Ports has caused tremendous pressure on traffic here. Haiphong Ports are along the rivers, with many winding sections, narrow channel width, limited depth, and complicated flow sections intersecting many dangerous areas. There is no TSS here, and the average width of the channel bottom is about 80m. In some areas, avoiding or overtaking each other is difficult, while the number of ships passing to enter the ports increases. There are some passages where ships can only go one way with the control of VTSOs.

The source of the AIS data was on 28th August 2023. To prevent the temporal dispersion of the AIS data, the extracted AIS data is processed through the cleansing and interpolation process described in Section 3.1. Ships' position data in longitude and latitude are converted into Descartes coordinates with axes in nautical miles (NM).

The next step is to group the ships via DBSCAN. The encounter must be formed inside at least two ships, so $MinPts$ should be defined as 2. Eps is the radius at which the ships are connected to form an encounter cluster and varies depending on the situation between the ships, the ship's maneuverability, the sea conditions, meteorological conditions, etc. In general, the designation of Eps should follow the following rules: the Eps value for confined waters should be smaller, and the Eps value for open water should be more significant. For practical purposes in each area, Eps should be matched with the recommendations of experienced captains, OOWs, pilots, and VTSOs of the study area. In this Haiphong Port water case study, the value of Eps was selected as 1.5 NM. After specifying the parameters, DBSCAN is applied to cluster at least two ships into groups and filter out single ships as noises. In previous studies, noise ships were not included, and it is advisable to ignore them, considering the simplification that contributes to the calculations. However, in this study, ships designated as safe are still tracked if they approach other ships and become the border or core of a cluster. The HSD will then be applied to assess the risk. The result is a 2D heat map showing the spatial distribution of the potential risk of collision based on different moments.

First, a numerical simulation that acquires AIS data for three approaching ships was carried out to evaluate the algorithm for determining the macroscopic risk of ship collision based on HSD. The positions and information of the ships at the beginning are shown in Fig. 14 and Table 6. Ship trajectories are illustrated in Fig. 15, with ship A in black, ship B in blue, and ship C in red.

Table 6. Information of ships at the beginning

	Length (m)	Position	Speed (kn)	Course (°)
Ship A	200	20°49.132N 106°53.083E	7,2	081
Ship B	182,9	20°49.658N 106°53.295E	6,0	148
Ship C	222,2	20°48.520N 106°54.341E	6,0	325

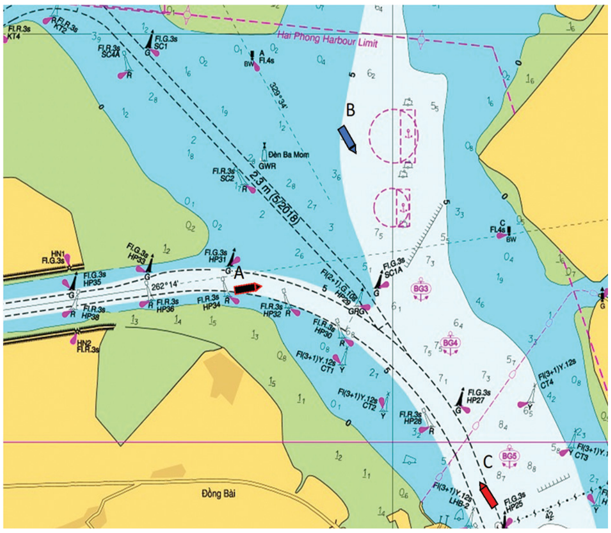


Fig. 14. Positions of ships at the beginning

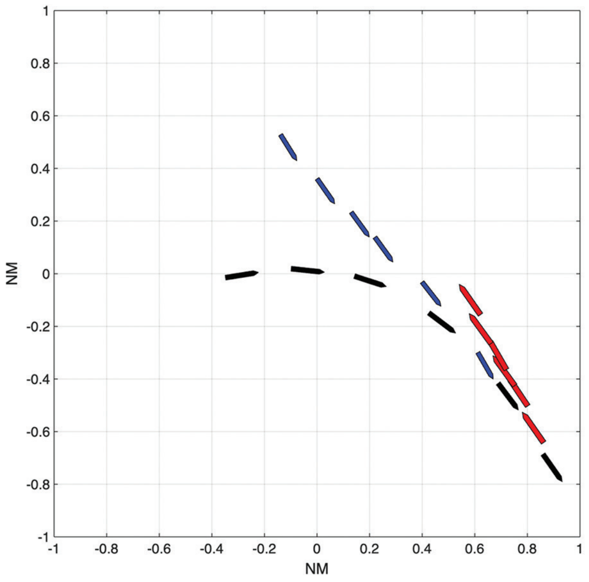


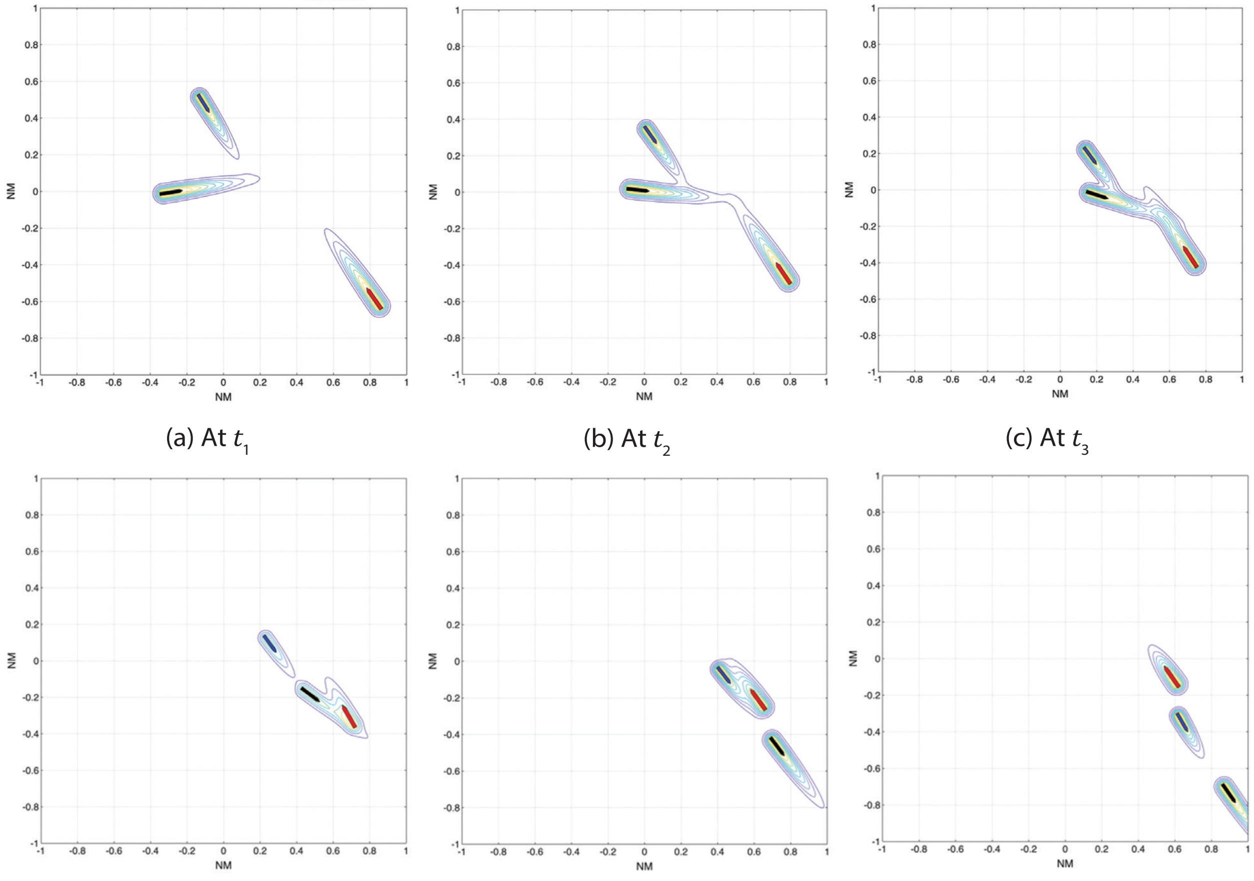
Fig. 15. Trajectories of ships

Due to the distances between ships in the scenario being smaller than 1.5 NM, these ships are clustered in one group, and HSD with $h = 0.2$ is applied to them. At the macroscopic level, the HSDs are presented with iso risk index lines from 0.1 (the most inside line) and 0.9 (the most outside line). The results are shown in Fig. 16 with ship A in black, ship B in blue, and ship C in red. At t_1, when there is no interference between HSDs, the fore parts of the HSDs are about 0.5 NM in length. At t_2, it begins to be observed that there is an overlapped area 1 between the port side of HSD_A and the fore of HSD_B with a value of 0.2 and an overlapped area 2 between the fore of HSD_A and HSD_C with a value of 0.2. Thus, it can be seen that when the overlapped area between HSDs has a value of 0.2, the risk of collision begins to form. At t_3, when the three ships approach each other, the overlapped areas 1 and 2 values increase to 0.5 and 0.4, respectively. The ships began to take action to avoid collision: ship A continued to turn to starboard while ship B and C slowed down. At time t_4, the HSDB became smaller and separated while the HSDA and HSDC still had interference. The most significant value at the overlapped area is 0.6. The two ships, A and C, changed their course to starboard until the HSDs of the two ships separated, and the risk between them did not exist anymore. However, the HSDB and HSDC interfered with each other at t_5, and these two ships continued maneuvering, so there was no interference at t_6.

Although ECDIS can observe areas where ships are assembled, obtaining the exact collision risk values for these locations is impossible. In other words, any location with the same number of ships looks similar, and the collision risk values or indexes of these locations are indistinguishable. Therefore, constructing a Collision Risk Prediction Map is necessary to quickly identify accurate regional collision risk values of high-risk areas by combining multi-ship clusters. Specifically, when the proposed map is adopted, it is possible to obtain an overlapped area of HSDs with value R_{area}. An area with a high rate of R_{area} value during a period can be defined as a hot spot. Each color corresponds to each collision risk level. The areas with more yellow color represent the greater risk.

AIS data from 1005 to 1020 on August 28, 2023, with 12 ships, was processed and then applied DBSCAN and HSD. The results are Collision Risk Prediction Maps of these moments with 5-minute intervals, as shown in Fig. 17). At 1005, 2 clusters of encounter ships were detected. In cluster 1 at the top left corner of the map, two ships were approaching each other. HSDs of two ships overlapped, and this area appeared to be yellow. The index of this area was from 0.88 – 0.93. In cluster 2 on the right side of the map, a ship was passing through the channel, and two other ships were preparing to go in. At this moment, although there was an overlapped area of two HSDs with the value of index 0.33, we can see that this index would increase when the other ship of the cluster came, which would be more dangerous.

At 1010, there was no more overlapped area in cluster 1 due to two ships having taken action to prevent a collision. However, because the distance between these ships is smaller than 1.5 NM (Eps value), HSD was still applied for them. In cluster 2, the index of the overlapped area raised to 0.91 due to the contribution of all three ships. The ships could be advised to start paying attention to it and take substantive actions as soon as possible by judging the index.



(a) At t_1

(b) At t_2

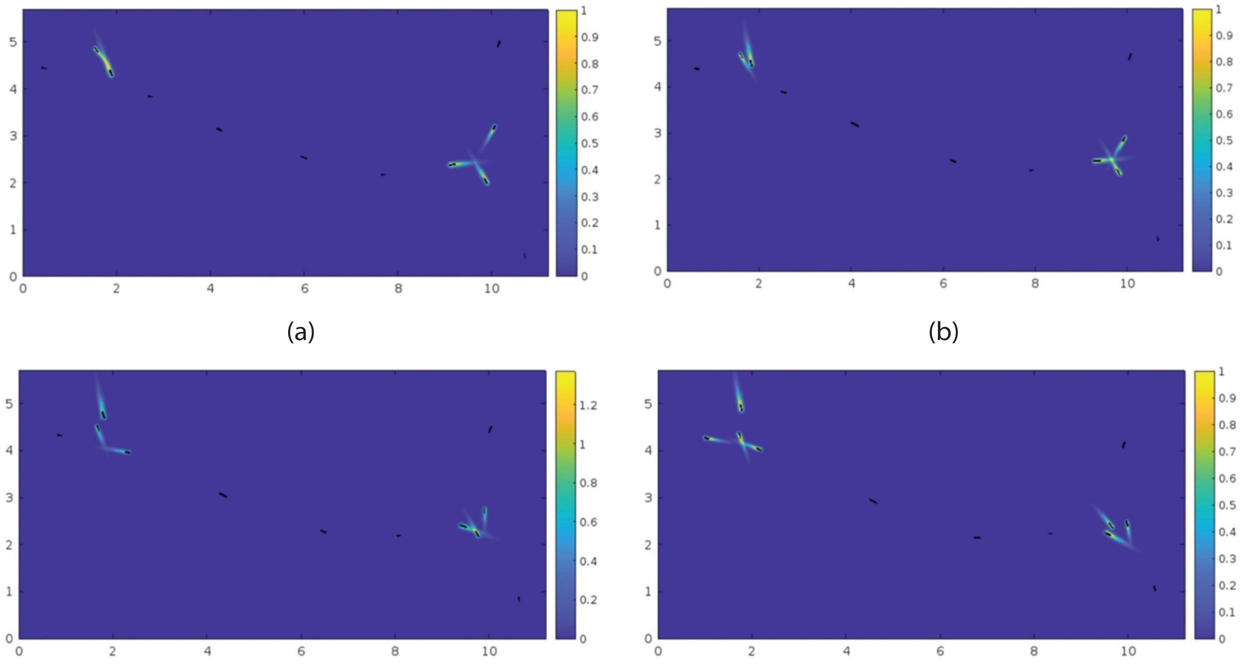
(c) At t_3

(d) At t_1

(e) At t_2

(f) At t_3

Fig. 16. Encounter of ships with HSDs: (**a**) At t_1 there is no interference between HSDs; (**b**) At t_2 there is an overlapped area 1 between the port side of HSDA and the fore of HSDB and an overlapped area 2 between the fore of HSD_A and HSD_C; (**c**) At t_3 three ships continue to approach each other; (**d**) At t_4 there is only interference of HSD_A and HSD_C; (**e**) At t_5 there is only interference of HSD_B and HSD_C; (**f**) At t_6 there is no longer any overlap



(a)

(b)

(c)

(d)

Fig. 17. Collision Risk Prediction Map: (**a**) At 1005, clusters of encounter ships were detected; (**b**) At 1010, there was no overlapped area in cluster 1, and the index of the overlapped area in a cluster; (**c**) At 1015 cluster 1 had one more ship and collision risk became obvious in cluster 2; (**d**) At 1020 four ships formed cluster 1 and HSDs separated in cluster 2

At 1015, cluster 1 had one more ship because a noise ship that had not applied HSD in previous moments started to become a border ship, and its HSD was displayed. Although their HSDs did not violate each other, these ships should be monitored due to their proximity. In cluster 2, before collision risk became apparent, one ship had taken effective collision avoidance measures by changing course and reducing speed. However, the index of the overlapped area still increased to 1.2 due to the close encounter of the other two ships. At this moment, the collision risk was very urgent. At 1020, a ship from the left joined the cluster 1. This cluster now included four ships with the overlapped area of 2 HSDs with an index of 0.85. These two ships were required to take the most helpful action to avoid a collision; otherwise, the collision would occur. In cluster 2, three ships have maneuvered to avoid a collision, and their HSDs are separated, showing safe status. The hot spots area is usually concentrated in the traffic intersection area in the study area.

In narrow areas or narrow channels, especially in the study area (Haiphong waters), it is difficult for ships to cross too close aft side of another. However, this kind of encounter still sometimes happens at the intersections. When these situations happen, the VTSOs should keep continuous communication with all parties, such as captains and maritime pilots, and require all ships to use sound and light signals if necessary. Due to the characteristics of ships and conditions of the channel (depth, width, obstructions, other ships coming…), VTSOs can refer to the proposed framework by observing overlapped areas of HSDs with DCR values (as represented in Table 7) and determine the actions of each ship such as: slowing down or remaining speed, changing or keeping heading… to prevent the collision, and ensure continuous traffic in the monitoring area.

Table 7. DCR range

Status	DCR range	Action
Outlier		Consider as safe but is still tracking
Ship clusterd in group		Pay more attention when HSD is applied
Ship domain overlapped	0 – 0.2	Collision risk begin to exist
Ship domain overlapped	0.2 – 0.5	Collision risk becomes apparent
Ship domain overlapped	0.5 – 1	Collision risk becomes critical
Ship domain overlapped	Greater than 1	Collsion risk becomes extremely critical

According to the COLREGS, the collision risk must be evaluated before determining whether to take a collision-avoidance action or change the current sailing condition. Before the risk of collision exists, ships are free to take any action, when a ship is located in a safe area and is not close to any ship. When a ship is clustered with others, they are advised to pay attention. In this situation, the HSD is applied to help visualize VTSOs and OOWs. When HSDs are overlapped, the overlapped area has $0 < DCR \leq 0.2$, the risk of collision first begins, and the ships must take early and substantial action to achieve a safe passing distance.

When 0.2 < DCR ≤ 0.5, ships should take more appropriate actions in compliance with the COLREGs to avoid collision. Similarly, when DCR > 0.5, ships must take the best aid to avoid a collision; otherwise, there will be a collision.

In particular, the proposed framework focuses on the index of overlapped areas of the ship domains to predict the location of collision, which is different from a previous method, which considered the position of the target ship inside the domain of its ship to evaluate collision risk. In fact, in congested areas, it is necessary to construct a framework that can detect risk as early as possible. The situation when a ship violates another ship's domain is urgent, and ships may not have enough time to act. This Collision Risk Prediction Map is accurate and stable by transforming domain overlap problems into the distribution of high-risk areas. The final results of the case study demonstrate that the parameters are appropriate; however, other sea waters need to adopt another value of area parameter due to their specific characteristics and conditions as well as the requirements and policies of the maritime traffic surveillance systems. The proposed framework shows its collision risk detection function at some past moments. If real-time AIS data is applied, the whole procedure of data processing and clustering encounter ships by DBSCAN and HSD should be implemented continuously due to the change of ships' positions, speeds, headings, and traffic density. As a result, the collision risk prediction map can change constantly. Finally, the detection of high-risk areas is a dynamic process over time.

5. CONCLUSIONS

This study proposes the Collision Risk Prediction Map based on AIS data. To simplify the computation and make it quicker to detect the hot spots, DBSCAN was used to cluster the ships in the water area, and the contribution of each ship to potential collision within the cluster is calculated as a function of the Heat Ship Domain. This domain is dynamic and expresses the risk around the ship by index. Collision risk was identified if an overlapped area between HSD happens. Finally, the geographical distribution of collision risk was visualized to establish a collision risk prediction map. To validate the effectiveness of the new map, a case study was conducted in the waters of Haiphong Ports in Vietnam. The results show that the visualization obtained by the framework can effectively reflect the collision risk of specific waters through the index in three perspectives: micro, macro, and region. Unlike other collision risk identification models, the parameter can be easily adjusted based on experts' knowledge of the study area. The proposed framework can help maritime traffic operators or administrations better understand the overall collision risk and its distribution in the water during collision risk monitoring.

6. ACKNOWLEDGEMENT

This research is supported by Vietnam Maritime University.

7. REFERENCES:

[1] L. P. Perera, J. P. Carvalho, C. G. Soares, "Fuzzy logic-based decision-making system for collision avoidance of ocean navigation under critical collision conditions", Journal of marine science and technology, Vol. 16, 2010, pp. 84-99.

[2] B. Wu, T. L. Yip, X. Yan, C. G. Soares, "Fuzzy logic-based approach for ship bridge collision alert system", Ocean Engineering, Vol. 187, 2019.

[3] Y. Huang, L. Chen, P. Chen, R. R. Negenborn, P.H.A.J.M. van Gelder, "Ship collision avoidance methods: State-of-the-art", Safety Science, Vol. 121, 2020, pp. 451-473.

[4] X. Yuan, D. Zhang, J. Zhang, M. Zhang, C. G. Soares, "A novel real-time collision risk awareness method based on velocity obstacle considering uncertainties in ship dynamics", Ocean Engineering, Vol. 220, 2021.

[5] T. Brcko, A. Androjna, J. Srse, R. Boc, "Ship Multi-Parametric Collision Avoidance Decision Model: Fuzzy Approach", Journal of Maritime Science and Engineering, Vol. 9, No. 1, 2021.

[6] R. Fiskin, O. Atik, H. Kisi, E. Nasibov, T. A. Johansen, "Fuzzy domain and meta-heuristic algorithm-based collision avoidance control for ships: experimental validation in virtual and real environment", Ocean Engineering, Vol. 220, 2021.

[7] P. Chen, Y. Huang, J. Mou, P.H.A.J.M. van Gelder, "Ship collision candidate detection method: a velocity obstacle approach", Ocean Engineering, Vol. 170, 2018, pp. 186-198.

[8] P. Chen, Y. Huang, E. Papadimitriou, J. Mou, P.H.A.J.M. van Gelder, "An improved time discretized non-linear velocity obstacle method for multi-ship encounter detection", Ocean Engineering, Vol. 196, 2020.

[9] Y. Fujii, K. Tanaka, "Traffic capacity", Journal of Navigation, Vol. 24, 1971, pp. 543-552.

[10] E. M. Goodwin, "A statistical study of ship domains", Journal of Navigation, Vol. 28, 1975, pp. 328-344.

[11] M. G. Hansen, T. K. Jensen, T. Lehn-Schiøler, K. Melchild, F. M. Rasmussen, F. Ennemark, "Empirical

ship domain based on AIS data", Journal of Navigation, Vol. 66, 2013, pp. 931-940.

[12] X. Qu, Q. Meng, L. Suyi, "Ship collision risk assessment for the Singapore Strait", Accident Analysis & Prevention, Vol. 43, 2011, pp. 2030-2036.

[13] C. Zhou, L. Ding, M. J. Skibniewski, H. Luo, S. Jiang, "Characterizing time series of near-miss accidents in metro construction via complex network theory", Safety Science, Vol. 98, 2017, pp. 145-158.

[14] S. L. Yoo, "Near-miss density map for safe navigation of ships", Ocean Engineering, Vol. 163, 2018, pp. 15-21.

[15] B. Kundakçı, S. Nas, L. Gucma, "Prediction of ship domain on coastal waters by using AIS data", Ocean Engineering, Vol. 273, 2023.

[16] J. Weng, D. Yang, T. Chai, S. Fu, "Investigation of occurrence likelihood of human errors in shipping operations", Ocean Engineering, Vol. 182, 2019, pp. 28-37.

[17] P. A. M. Silveira, A. P. Teixeira, C. G. Soares, "Use of AIS data to characterise marine traffic patterns and ship collision risk off the coast of Portugal", Journal of Navigation, Vol. 66, 2013, pp. 879-898.

[18] R. L. Shelmerdine, "Teasing out the detail: how our understanding of marine AIS data can better inform industries, developments, and planning", Marine Policy, Vol. 54, 2015, pp. 17-25.

[19] Y. C. Altan, E. N. Otay, "Spatial mapping of encounter probability in congested waterways using AIS", Ocean Engineering, Vol. 164, 2018, pp. 263-271.

[20] Z. Liu, Z. Wu, Z. Zheng, "A molecular dynamics approach for modeling the geographical distribution of ship collision risk", Ocean Engineering, Vol. 217, 2020.

[21] Z. Liu, B. Zhang, M. Zhang, H. Wang, X. Fu, "A quantitative method for the analysis of ship collision risk using AIS data", Ocean Engineering, Vol. 272, 2023.

[22] Z. Wang, Y. Wu, X. Chu, C. Liu, M. Zheng, "Risk Identification Method for Ship Navigation in the Complex Waterways via Consideration of Ship Domain", Journal of Maritime Science and Engineering, Vol. 11, 2023.

[23] T. Brcko, B. Luin, "A Decision Support System Using Fuzzy Logic for Collision Avoidance in Multi-Vessel Situations at Sea", Journal of Maritime Science and Engineering, Vol. 11, 2023.

[24] Z. Liu, D. Zhou, Z. Zheng, Z. Wu, L. Gang, "An Analytic Model for Identifying Real-Time Anchorage Collision Risk Based on AIS Data", Journal of Maritime Science and Engineering, Vol. 11, 2023.

[25] W. Li, L. Zhong, Y. Liu, G. Shi, "Ship Intrusion Collision Risk Model Based on a Dynamic Elliptical Domain", Journal of Maritime Science and Engineering, Vol. 11, 2023.

[26] W. Zhang, F. Goerlandt, P. Kujala, Y. Wang, "An advanced method for detecting possible near miss ship collisions from AIS data", Ocean Engineering, Vol. 124, 2016, pp. 141-156.

[27] L. Zhang, H. Wang, Q. Meng, "Big data-based estimation for ship safety distance distribution in port waters", Transportation Research Record: Journal of the Transportation Research Board, Vol. 2479, 2014, pp. 16-24.

[28] N. Im, T. N. Luong, "Potential risk ship domain as a danger criterion for real-time ship collision risk evaluation", Ocean Engineering, Vol. 194, 2019.

[29] L. Zhang, Q. Meng, "Probabilistic ship domain with applications to ship collision risk assessment", Ocean Engineering, Vol. 186, 2019.

[30] P. Silveira, A. P. Teixeira, C. G. Soares, "A method to extract the quaternion ship domain parameters from AIS Data", Ocean Engineering, Vol. 257, 2022.

[31] Z. Qiao, Y. Zhang, S. Wang, "A Collision Risk Identification Method for Autonomous Ships Based on Field Theory", IEEE Access, Vol. 9, 2021, pp. 30539-30550.

[32] W. Ma, H. Wang, S. Wang, "Critical Collision Risk Index Based on the Field Theory", Journal of Maritime Science and Engineering, Vol. 10, 2022.

[33] S. Węglarczyk, "Kernel density estimation and its application", ITM Web of Conferences, Vol. 23, 2018.

[34] J. Ø. Strand, "Ship Domain in Restricted Waters - A study assessing Norwegian navigators´ perception of safe passing distance to a targeted ship

in restricted waters", University of South-Eastern Norway, Faculty of Technology, Natural Sciences and Maritime Sciences, Norway, Master Thesis, 2018.

[35] M. Wielgosz, "Declarative ship domains in restricted areas", Scientific Journals of the Maritime University of Szczecin, Vol. 46, No. 118, 2016, pp. 217-222.

[36] M. Wielgosz, "Ship domain in open sea areas and restricted waters: an analysis of influence of the available maneuvering area", International Journal on Marine Navigation and Safety of Sea Transportation, Vol. 11, 2017, pp. 99-104.

[37] T. G. Coldwell, "Marine traffic behaviour in restricted waters", Journal of Navigation, Vol. 36, 1983, pp. 430-444.

[38] Y. Wang, H. Chin, "An empirically-calibrated ship domain as a safety criterion for navigation in confined waters", Journal of Navigation, Vol. 69, 2016, pp. 257-276.

[39] M. Ester, H. Kriegel, J. Sander, X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2-4 August 1996, pp. 226-231.

[40] Z. Liu, Z. Wu, Z. Zheng, "A novel framework for regional collision risk identification based on AIS data", Applied Ocean Research, Vol. 89, 2019, pp. 261-272.

Deep Learning-Based Method for Detecting Parkinson using 1D Convolutional Neural Networks and Improved Jellyfish Algorithms

Original Scientific Paper

Arogia Victor Paul M*
Department of Computer Science and Engineering,
B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India
victorpaul_cse@crescent.education

Sharmila Sankar
Department of Computer Science and Engineering,
B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India
sharmilasankar@crescent.education

*Corresponding author

Abstract *– Parkinson's disease (PD) is a common disease that predominantly impacts the motor scheme of the neural central scheme. While the primary symptoms of Parkinson's disease overlap with those of other conditions, an accurate diagnosis typically relies on extensive neurological, psychiatric, and physical examinations. Consequently, numerous autonomous diagnostic assistance systems, based on machine learning (ML) methodologies, have emerged to assist in evaluating patients with PD. This work proposes a novel deep learning-based classification of Parkinson's disease (PD) using voice recordings of people into normal, idiopathic Parkinson, and familial Parkinson. The improved jellyfish algorithm (IJFA) is utilized for hyper-parameter selection (HPS) of a 1D convolutional neural network (1D-CNN). The proposed technique makes use of the significant elements of 1D-CNN and filter-based feature selection models. Because of their strong performance in dealing with noisy data, the filter-based algorithms Relief, mRMR, and Fisher Score were chosen as the top choices. Using just 62 characteristics, the combination of deep relief features and deep learning was able to discriminate between people. The competence of the proposed 1D-CNN with IJFA method was determined through specific network metrics. The proposed 1D-CNN with IJFA method attains a total accuracy of 98.6%, which is comparatively better than the existing techniques. The proposed model produced around 9.5% improvements in accuracy, respectively, when compared to the data obtained without dimensionality reduction.*

Keywords: *Parkinson Disease Classification, Convolutional Neural Network, Improved jellyfish algorithm, Filter-based feature selection model*

Received: January 10, 2024; Received in revised form: March 1, 2024; Accepted: April 8, 2024

1. INTRODUCTION

Parkinson Disease (PD) is a condition that touches the motor scheme of the dominant nervous organization of the human body. Motor symptoms and non-motor symptoms are the two categories that may be used to describe the symptoms of PD [1]. The motor symptoms involve trembling and a lack of energy in the hands and legs, constipation, difficulty doing daily tasks, and a shuffled gait when walking [2]. Parkinson's disease non-motor symptoms include a variety of problems, including weariness, constipation, difficulty speaking, memory loss, and exhaustion [3]. Studies suggest that voice problems arise in about 90 percent of PD cases [4]. The Procedure of Diagnosing (PD) using just certain qualitative criteria may make the process more difficult since other illnesses might also present with similar symptoms [5].

In recent years, a surge in Parkinson's disease research has leveraged machine learning (ML) for diagnosis [6]. Studies have utilized various data, including walking tracks, speech recordings, and brain electrical event recordings (EEG) [7]. Notably, speech-based diagnostic methods have shown promise, as speech difficulties often manifest early in PD [8]. These algorithms effectively distinguish between individuals with PD and healthy subjects using distinctive features extracted from raw speech data [9]. PD diagnostic systems employ diverse speech signal processing algorithms to extract clinically relevant vocal characteristics from voice recordings [10]. Essential features extracted from real-world datasets are input into machine learning models for PD diagnosis [11]. The performance of these models depends on the relevance of features utilized during training [12]. To address high dimensionality and sparse data issues, reducing dimensionality is crucial in PD research. This process emphasizes relevant characteristics, enhancing the success of machine learning models for diagnosis [13, 14]. The contribution of this paper is,

- In this work, proposed a novel deep learning-based classification of Parkinson's disease using voice recordings of people into normal, idiopathic Parkinson, and familial Parkinson.
- Initially, the CNN network is used as a feature extractor to extract the voice recordings.
- Then, the Improved Jellyfish Search Algorithm (IJFA) to select the features before the final layers is utilized for hyper-parameter selection (HPS) of 1D-convolutional neural network (1D-CNN).
- Finally, a 1D-CNN classifier was trained using the generated deep feature representations.
- The performance of the Proposed 1D-CNN with IJFA method was measured by parameters such as Specificity, Accuracy, F1 score, Precision, and Sensitivity.

In Section 2, we provide the work that is relevant to this study, and in Section 3, we provide an explanation of the suggested model. Sections 4 and 5 each provide a conclusion that summarises the findings of the validation study.

2. RELATED WORKS

Recent research has presented various deep learning-based methods for detecting Parkinson's disease.

In 2022, Sahu et al. [15] suggested an early Parkinson disease diagnosis method based on hybrid deep learning. The combination of two deep learning techniques, such as regression analysis (RA) and artificial neural networks (ANN), for efficient probability-based illness diagnosis. The accuracy of the suggested method is 93.46%.

In 2022, Vyas et al. [16] developed a method for deep learning (DL) that uses convolutional neural networks (CNNs) in two and three dimensions. The 2D model attained an accuracy of 72.22% with 0.50 area under the curve (AUC), whereas the 3D model features from the data were able to categorize the test data with an accuracy of 88.9% with 0.86 AUC.

In 2022, Hosny et al. [17] developed a brand-new deep learning model to identify subthalamic nuclei (STN) in signals from local field potentials (LFPs). The k-Nearest Neighbor (KNN) classifier receives the characteristics as input. According to the findings, KNN achieved an accuracy rate of 87.27% on average.

In 2022, Rajanbabu et al. [18] developed transfer learning-based deep learning architectures for effective PD diagnosis using MRI data. Based on the maximum chance of all the models chosen for PD classification, an ensemble model is suggested. The method primarily concentrates on providing an accurate PD diagnosis.

In 2022, Moradi et al. [19] offered a microarray dataset (GSE22491) that was given by GEO. The Limma package, which is part of the R program, was used to find DEGs and analyze and assess gene expression. Support vector machines (SVM) results show that using three genes together can lead to an 88% prediction accuracy.

In 2022, AlMahadin et al. [20] proposed a series of resampling methods to enhance the classification of tremor severity in Parkinson's disease. The suggested method combines three types of resampling and signal processing techniques: hybrid, under, and over-sampling. ANN-MLP, as suggested, has an overall accuracy of 93.81%.

In 2024, Canturk et al. [21] suggested utilizing voice signals and artificial intelligence to diagnose Parkinson's disease. AlexNet, GoogleNet, ResNet50, and the majority of voting-based hybrid systems are among the first classifiers used. The deep feature fusion method produced an accuracy of 0.95%.

In 2024, Aldhyani et al. [22] suggested that the public dataset PD Spiral Drawings be utilized for PD research and diagnosis. The suggested technique made use of a common dataset made up of 204 spiral and wave drawings made by people with Parkinson's disease. With 94% accuracy, pictures were used to train the DenseNet201 classifier.

The analysis highlighted earlier emphasizes the constraints of current research procedures and models. To overcome these limitations, this paper Improved jellyfish algorithm (IJFA) is utilised for hyper-parameter selection (HPS) of 1D-convolutional neural network (1D-CNN).

3. PROPOSED METHODOLOGY

3.1. DATASET DESCRIPTION

In this proposed method, the Parkinson's disease dataset can be used. In this dataset, there are 62 voice recordings of people. Also, this dataset consists of three class: 50% samples belonging to healthy and 50% samples belonging to patients. The data in PD dataset take from 62 patients with Parkinson Disease (30 men and 32 women) with ages ranging from 33 to 87. The PD dataset with the of 50-50% training and testing partition.

3.2. FEATURES SELECTION

In medical applications, feature selection has been the subject of several research, all of which have shown that it is both adequate and successful. Because it is a pre-processing method, it is able to single out the most important aspect of the issue. Maximum relevance redundant features as feasible (minimum redundancy). Maximum relevance seeks to identify the characteristics that it also seeks to give the feature subset comprising fewer and minimal redundant features as feasible.

According to the mRMR, the optimization condition ought to be expressed as follows:

$$x_j \in X - S_{k-1} Max \; [I\,(x_j, c) \; - \; \tfrac{1}{k-1} \textstyle\sum_{x_i \in S_{k-1}}^{n} I(x_j, x_i)] \quad (1)$$

When c is the target class, x_i is the ith feature, and X is the entire set of features. The mutual information between class c and feature x_j is represented by x_i.

The mRMR method improves classification accuracy while simultaneously reducing the number of variables used. This is accomplished by minimising the selection of duplicate features. Table 1 shows the different aspects of the human voice.

Table 1. Different aspects of human voice

Description	Voice measure
11-point Amplitude Perturbation Quotient	MDVP: APQ
Absolute jitter in microseconds	MDVP: Jitter (Abs)
Average vocal fundamental incidence	MDVP: F0 (Hz)
Maximum vocal fundamental incidence	MDVP: Fhi (Hz)
Relative Amplitude Perturbation	MDVP: RAP
Five-point Period Perturbation Quotient	MDVP: PPQ
Average absolute difference of differences among cycles, divided by the average period	Jitter: DDP
Shimmer Local amplitude perturbation	MDVP: Shimmer
Local amplitude perturbation	MDVP: Shimmer (db)
Average absolute difference among the amplitudes of consecutive aeras	Shimmer: DDA
Noise-to-Harmonics Relation	NHR
Harmonics-to-Noise Relation	HNR
Recurrence Period Density Entropy	RPDE
Correlation Dimension	D2
Fundamental frequency difference	Spread1
Fundamental frequency difference	Spread2
Pitch retro entropy	PPE

$$f(K) = \frac{\sum_{j=1}^{C} n_{j\left(\mu_j^l - \mu^l\right)^2}}{\sum_{j=1}^{C} n_j \left(\sigma_j^l\right) 2} \quad (2)$$

The number of occurrences of a feature is denoted by its mean, which is denoted by the symbol l, whereas the number of occurrences of a class is denoted by the symbol n j. During the process of feature selection using Fisher Score, all of the features are sorted in decreasing order beginning with the high scores are selected.

3.2.2. Relief

The significance of the features is computed via relief selection, which does this by illuminating the relationships that exist between the features and the class labels. The iterative process that was used in order to indicate the feature relevances may be seen in the equation (3).

$$W_i = W_{i-1} - (x_i - NearHit_i)^2 + (x_i - NearMiss_i)^2 \quad (3)$$

In the 'n' dimensions and records of n different characteristics. While the closest samples of the same class and those of different classes are denoted by the terms "NearHit" and "NearMiss," respectively.

3.3. 1D-CNN FOR CLASSIFICATION

In convolution neural network (CNN) models are used rather often for the purpose of image identification in two dimensions. On the other hand, the use of CNN models should not be limited to either two-dimensional or tasks in order to be utilised. It should come as no surprise that the 1D-CNN model has the same qualities as previous CNN models. A one-dimensional input signal, which will be indicated by S, and a kernel variable, which will be denoted by W will be used in the following description of the convolution procedure.

$$(S * W)_n = \textstyle\sum_{i=1}^{|W|} W(i) S(i + n - 1) \quad (4)$$

A feature map is the term used to refer to the final product of the convolution process. Let the limited matrix of the input matrix to the weight matrix be denoted by the notation $S_{|W(i,j)}n$. $S_{|W(i,j)}n$ is a representation of the elements of S from n up to the dimension of $W(i,j)$. As a result, the output matrix is capable of being characterised by a generic formula, which may be found in Equation (5):

$$O_n^l = \left(S_{|W(i,j)} * W(i,j)\right)_n \quad (5)$$

The final part of the CNN model, which usually consists of a neural network layer, handles the classification task. This layer is in charge of the final level and is referred to as a completely connected layer. The input consists of the pre-processed signal segments, each of which contains 50% samples. In the first layer of the model, the signals are convoluted using 64 x 5 filters and three stride ratios in order to build feature maps with sizes ranging from 64 x 999 to 64 x 999. The second layer of the model is also a convolution layer has 128 filters by 5 rows. This layer produces brand new feature maps by using the results of the previous layer.

The Max Pool layer combines the two output vectors' maximum values in two-unit areas into a single value. This value is the result of the condensing process. These steps are repeated in a similar way in each of the model's subsequent layers, but each time, a range of distinct filter sizes are used. Dropout layers are built into the model to reduce the issue of overfitting. The dimensions determined in the flattened layer must be changed to fit the thick layers. The final layer, the softmax layer, is where the input signals are mapped onto the output signals. Therefore, the sum of classes (nb class) and the sum of units (nb unit) in this layer are equal to one another. In CNN network at the end of the last convolution, they apply the Jellyfish Search

Algorithm to select the features before the final layers. Table 2 contains comprehensive parameter representations of all of the model's layers. The Proposed 1D-CNN model is shown in Fig. 1.

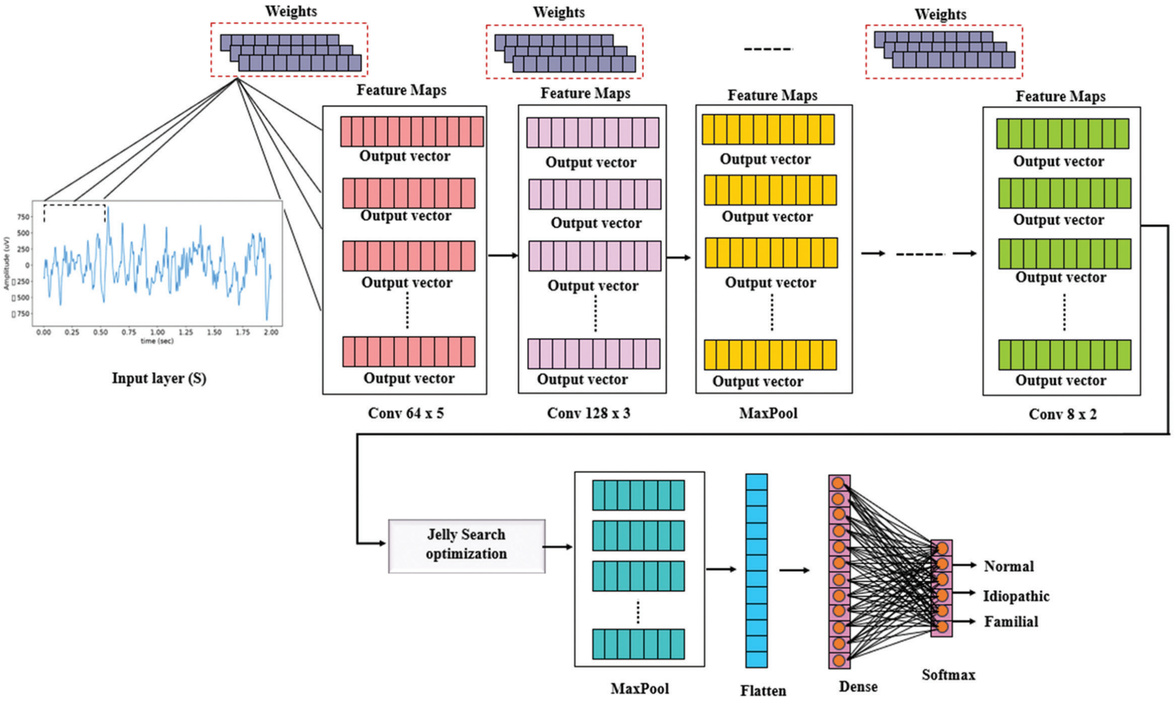


Fig. 1. The Architecture of the proposed 1D-CNN model

Table 2. Particulars of layers and strictures in the projected 1D-CNN perfect

Layer Name	Sum of Filterx Kernel Size	Sum of Trainable Parameters	Layer Parameters	Region/Unit Size	Output Scope
1D Conv	64 × 5	384	ReLU, S = 3	-	64 × 999
1D Conv	128 × 5	24,704	ReLU, Stride = 1	-	128 × 997
MaxPool	-	0	S = 2	2	128 × 489
Dropout	-	0	Rate = 0.2	-	128 × 498
1D Conv	128 × 13	213,120	ReLU, S = 1	-	128 × 346
1D Conv	256 × 7	229,632	ReLU, S = 1	-	256 × 480
MaxPool	-	0	Stride = 2	2	256 × 240
1D Conv	256 × 7	262,272	ReLU, S = 1	-	128 × 322
1D Conv	64 × 4	32,832	ReLU, S = 1	-	64 × 230
MaxPool	-	0	Stride = 2	2	64 × 54
1D Conv	8 × 5	2568	ReLU, S = 1	-	8 × 50
1D Conv	8 × 2	136	ReLU, S = 1	-	8 × 49
MaxPool	-	0	Stride = 2	2	8 × 42
Flatten	-	0	-	-	1 × 192
Dense	-	12,352	ReLU, Drop = 0.2	64	1 × 64
Dense	-	195	Softmax	nb_class	1 × nb_class

3.3.1. Jellyfish Search Algorithm

The search-feeding behaviour and drive designs of jellyfish in the water served as an inspiration for the development of the algorithm. The following is a rundown of the three rules that are a part of the JS algorithm:

Rule 1: Jellyfish are able to move in two different ways: one is to follow the currents of the ocean, and the other is to move about within their own population. The transition between these two different forms of drive is accomplished via a time- process.

Rule 2: Jellyfish habit, and when they are looking for food in the water, they are drawn to areas that have a greater concentration of food for them to consume.

Rule 3: stipulates that the quantity of food that jellyfish look for is reliant on the geographical location of the food as well as the goal purpose of the reaction. This model includes mechanisms for group movement, movement in response to time, and movement of jellyfish that follow the movement of ocean currents.

(1) Following ocean current movement

The term "trend" is used to describe the direction in which the current is moving.

$$trend \rightarrow = \frac{1}{n_{pop}} \sum_{i=1}^{n_{pop}} trend_i \rightarrow \tag{6}$$

where, n_{pop} is the present optimum position, and the equation that specifies its relationship is as follows:

$$trend_i \rightarrow = X^* - e_c X_i \tag{7}$$

where *X** represents the equation is derived by iterating over the previous equation until the desired result is achieved:

$$trend \rightarrow\ = \frac{1}{n_{pop}} \sum_{i=1}^{n_{pop}} (X^* - e_c X_i) = X^* - e_c \frac{1}{n_{pop}} \sum_{i=1}^{n_{pop}} X_i = X^* - e_c \mu \tag{8}$$

where *m* is the average location of all of the jellyfish. DF is represented as shadows:

$$DF = e_c \mu \tag{9}$$

The following is what you get when you plug Equation (9) into the equation that describes the ideal position:

$$trend\ \rightarrow = X^* - DF \tag{10}$$

The distribution of Jellyfish in the ocean is shown is Fig. 2. Where σ is the standard deviation of the normal distribution, and β signifies the distribution coefficient, which set to 3 in the algorithm. The range of $\pm\beta\sigma$ around the mean position μ contains the probability of all jellyfish positions.

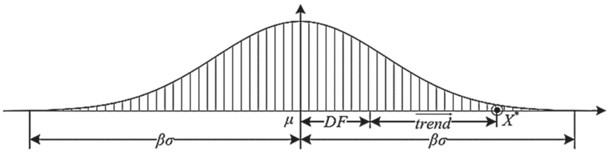


Fig. 2. Normal distribution of jellyfish in the ocean

A relative investigation demonstration that e_c can be uttered as shadows:

$$e_c = \beta * rand(0,1) \tag{11}$$

Following is the equation that was acquired in order to derive the equation

$$X_i(t+1) = X_i(t) + rand(0,1) *\ trend \rightarrow = X_i(t) + rand(0,1) * (X^* - \beta * rand(0,1) * \mu \tag{12}$$

where X i (t) represents the location of the Jellyfish at the instant in time when its position is being updated, and X i (t+1) represents the position of the jellyfish

(2) Movements made in groups

The jellyfish moves in a circle around its current position to indicate class A movement, and the following equation is used to update the jellyfish's position

$$X_i(t+1) = X_i(t) + \gamma * rand(0,1) * (U_b - L_b) \tag{13}$$

where *g* is the jellyfish's mobility coefficient, *Lb* is the lower bound, and *Ub* is the upper limit of the search space.

The following formula can be used to update the location of jellyfish participating in Class B drive: jellyfish that gather with the intention of consuming food when there is more food around:

$$X_i(t+1) = X_i(t) + step \rightarrow \tag{14}$$

where, *step*→ signifies the step length of jellyfish

defines *step*→ and is uttered as follows:

$$step \rightarrow = rand(0,1) *\ D \rightarrow \tag{15}$$

where, *D*→ shows the direction in which the jellyfish are swimming i. The following is an expression of the formula that may be used to determine the direction of motion:

$$D\ \rightarrow = \left\{ X_j(t) - X_i(t), iff\left(X_i(t)\right) \geq f\left(X_j(t)\right) X_i(t) - X_j(t), iff(X_j(t)) \geq f(X_i(t)) \right\} \tag{16}$$

where X_i (*t*) signifies the current location of jellyfish I, X_j (*t*) represents tof jellyfish *j*, function *f* represents the objective function with regard to *X*, and *X* signifies the collection of all jellyfish.

(3) A strategy for managing time

A temporal control system had to be conceived of, developed, and put into operation in order to successfully reproduce and materialise the switching that jellyfish are capable of doing between their three different modes of motion. The mechanism in question was described as a time control function denoted by the letter *c*. (*t*).

This is an expression of the formula that defines the variable *c* as follows: (*t*):

$$c(t) = \left| \left(1 - \frac{t}{T}\right) * (2 * rand(0,1) - 1) \right| \quad (17$$

where *t* is the current number of iterations, *T* is the number of iterations, and *c*(*t*) is a value that is randomly generated between 0 and 1 for each iteration. It was found that the range of values that controlled the jellyfish's movement in reaction to ocean currents was *c*(*t*) 0.5.

3.3.2. Improved Jellyfish Search Procedure

The (IJS) algorithm is presented in this section, and a thorough explanation of it can be found as follows:

(1) Development of a better technique for the initialization of population placement

Both the Sobol arrangement and the chaotic mapping starting strategy were used in order to create fifty percent of the total population. Under the illness that the search range limitation is satisfied, the Sobol sequence has the potential to create the beginning location of the jellyfish population in a more consistent manner. The functional representation of tent mapping may be described as follows:

$$x_{t+1} = \left\{ \frac{x_t}{\alpha} x_t \in [0, \alpha] (1 - x_t)/(1 - \alpha)\ x_t \in (\alpha, 1) \right\} \tag{18}$$

where x t is the created chaotic sequence, t2 is a series of numbers from 1 to n, n is the sum that has to be initialised, and an is an adjustment parameter with a value of 0.5.

(2) A sinusoidal component in the dynamic adaptation

A sinusoidal was included into the artificial jellyfish search algorithm in order to enhance its capacity for doing local searches. The expression of this factor may be expressed as follows:

$$S = 1 + sin\ sin \frac{\pi(2T+t)}{2T} \tag{19}$$

where S stands for the sinusoidal lively adaptation factor, T for the maximum iterations, and t for the iterations that are currently being performed.

(3) The functioning of the population difference

As the following explains, the addition of the random variation operation to the artificial jellyfish search method aimed to enhance the capacity to do worldwide searches and broaden the population's diversity:

Operation 1: After the jellyfish had finished migrating in accordance with the position update formula and had computed their respective fitness values, a particular jellyfish from the existing population was selected and given the name X_k. This particular jellyfish was picked at random. Next, three different jellyfish individuals were selected at random, and their relative fitness values were ranked from best to worst in order to determine X_a, X_b, and X_c. These individuals' fitness values are represented by the letters $f\,a$, $f\,b$, and $f\,c$, respectively. Finally, the following is an expression of the formula that may be used to calculate the new location of X_k:

$$X_k = X_a + \delta(X_b - X_c) \tag{20}$$

where d represents the variational operator, and the formula for it may be written as follows:

$$\delta = \delta_l + (\delta_u - \delta l) * \frac{f_b - f_a}{f_c - f_a} \tag{21}$$

where δ_u and dl represent the top and lower boundaries of variability, with 0.9 and 0.1 being the values used, respectively.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The overall performance of the Proposed 1D-CNN with IJFA method was evaluated built on the specific parameters viz., Accuracy, Precision, F1 score, Sensitivity, and Specificity.

$$\text{Sensitivity} = \frac{TP}{TP+FP} \tag{22}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{23}$$

$$Accuracy = \frac{\#Number\ of\ correct\ predictions}{\#Total\ number\ of\ predictions} \tag{24}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{25}$$

$$F1Score = \frac{2TP}{2TP+FP+FN} \tag{26}$$

False positives and true negatives of the MRI images are designated as TP and FP, respectively, whereas false positives and true negatives are designated as TN and FN, respectively. Table 2 shows the parameters for PD classes that are used to determine the proposed model's performance analysis.

Table 3. The overall Performance analysis of the Proposed model

Class	Accuracy	Precision	Sensitivity	F1-score	Specificity
Normal	97.54	96.32	94.08	93.69	92.58
Idiopathic	95.09	94.28	93.75	92.07	91.19
Familial	96.46	95.32	94.66	93.97	93.33

Table 1 illustrations the classification of various classes of Parkinson disease with specific metrics. The average Specificity, F1 score, Accuracy, Sensitivity, and Precision of the proposed 1D-CNN with IJFA method with the specific metrics. The proposed 1D-CNN with IJFA method has an average precision, sensitivity, F1score, and specificity of 95.3%, 94.16%, 93.24%, and 92.36%, respectively. Fig. 3 shown the performance parameters for three classes.

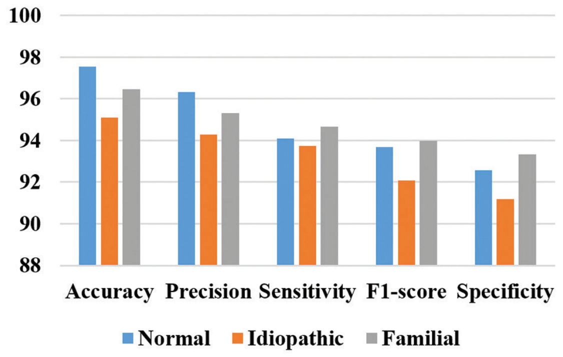


Fig. 3. The overall performance analysis of the proposed model

4.1. FEATURES SELECTION RESULTS

Maximum relevance minimal redundancy: selected the traits that were the most significant to us by using the mRMR, which stands for minimum redundancy (22 features). The PPE is the single most important variable in predicting the result. There is a noticeable difference between the first feature and the rest of the features combined because the first feature has a significantly lower score. The algorithm is positive; it has selected the most significant predictor because of the notable drop in the value of the relevant variable. Other research has demonstrated that, in contrast to earlier evaluations, PPE is resistant to the effects of noisy auditory environments and is also sensitive to changes in PD speech. Subsequently, the entropy is computed using the probability distribution of the semitone variations to define the PPE measure. Converting a speech pitch pattern into a logarithmic semitone measure is the first step in creating the probability distribution. Fig. 4 illustrates a sampling of the findings from this criterion selection. Fig. 5 displays the 22 characteristics in the feature ranking discovered by Relief.

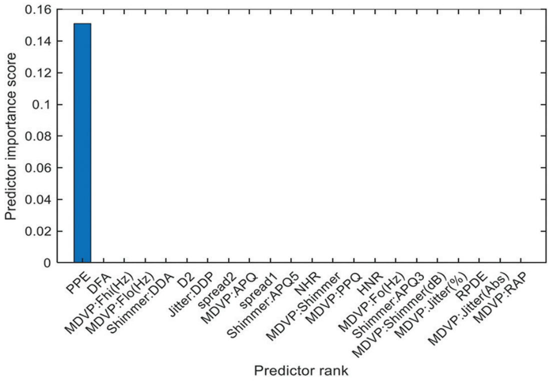


Fig. 4. Samples of the Feature findings

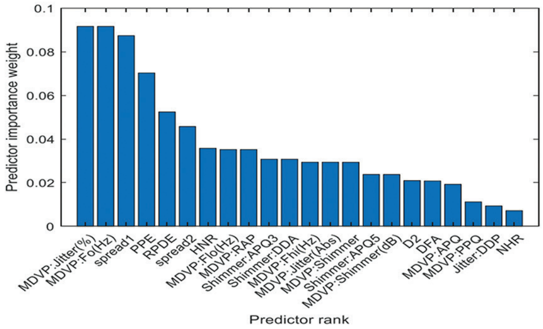


Fig. 5. The 22 characteristics of feature Ranking for Relief

The degree of relevance weights of all characteristics is figured out with the help of the bar plot. The Performance Analysis for Classifier is shown in Table 3. Fig. 6 provides the graphical analysis of proposed model with existing techniques.

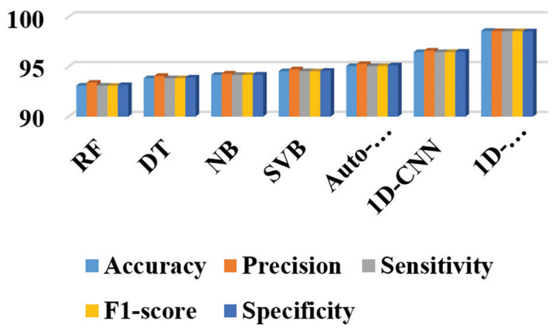


Fig. 6. Comparative Analysis of Proposed Model

Table. 4. Comparison of the proposed and the existing models

Authors	Methods	Accuracy
Sahu [15]	ANN-RA	93.46%
Vyas [16]	2D-CNN	88.9%
Hosny [17]	CNN-KNN	87.27%
AlMahadin [20]	ANN-MLP	93.81%
Proposed	1D-CNN with IJFA	98.6%

From Table 4, the comparison of several deep learning techniques based on their accuracy in the PD signals.

The Proposed 1D-CNN with IJFA method advances the overall accuracy by 5.21%, 9.83%, 11.4%, and 4.85% better than ANN-RA [13], 2D-CNN [15], CNN-KNN [16], and ANN-MLP [19] respectively. It is obvious from Table 4 that our innovative network outperforms the current methods. As a result, the proposed 1D-CNN with IJFA method fallouts may be used to accurately classify the PD disease.

Table 5. Performance Analysis for Classifier

Model	Accuracy	Precision	Sensitivity	F1-score	Specificity
RF	93.11	93.4	93.11	93.09	93.16
DT	93.85	94.08	93.85	93.84	93.91
NB	94.18	94.31	94.18	94.17	94.21
SVB	94.55	94.74	94.55	94.54	94.59
Auto-encoder	95.07	95.27	95.07	95.07	95.14
1D-CNN	96.47	96.62	96.47	96.47	96.51
1D-CNN with IJFA	98.6	98.57	98.56	98.56	98.54

5. CONCLUSION

This paper presents a novel deep learning-based classification of Parkinson's disease using voice recordings of people into normal, idiopathic Parkinson, and familial Parkinson. The improved jellyfish algorithm (IJFA) is utilized for hyper-parameter selection (HPS) of a 1D convolutional neural network (1D-CNN). To differentiate Parkinson's patients from healthy individuals at an early stage, the 1D-CNN method, coupled with IJFA approaches, was employed. The proposed technique takes use of the significant elements of 1D-CNN and filter-based feature selection models. In comparison between the other techniques the proposed 1D-CNN with IJFA methods with the accuracy of 98.6%. In the future, deep feature representations will be extracted from various kinds of data sources collected from wearable sensors, and then these data sources will be combined with various multi-modal techniques.

6. REFERENCES:

[1] N. Bhore, E. C. Bogacki, B. O'Callaghan, P.-H. Favreau, P. A. Lewis, S. Herbst, "Common genetic risk for Parkinson's disease and dysfunction of the endo-lysosomal system", Philosophical Transactions of the Royal Society B, Vol. 379, No. 1899, 2024, p. 20220517.

[2] S. Parveen, J. Moore, "Current Perceptions and Unmet Needs of People with Parkinson Disease and their Families", Clinical Archives of Communication Disorders, Vol. 8, No. 2, 2023, pp. 57-69.

[3] P. Ananthavarathan, B. Patel, S. Peeros, R. Obrocki, N. Malek, "Neurological update: non-motor symptoms in atypical parkinsonian syndromes", Journal of Neurology, Vol. 270, 2023, pp. 4558-4578.

[4] T. Romero Arias, I. R. Cortés, A. P. del Olmo, "Biomechanical parameters of voice in Parkinson's disease patients", Folia Phoniatrica et Logopaedica, Vol. 76, No. 1, 2024, pp. 91-101.

[5] M. T. García-Ordás, J. A. Benítez-Andrades, J. Aveleira-Mata, J. M. Alija-Pérez, C. Benavides, Determining the severity of Parkinson's disease in patients using a multi task neural network. Multimedia Tools and Applications, Vol. 83, No. 2, 2024, pp. 6077-6092.

[6] A. H. Al-nefaie, T. H. Aldhyani, D. Koundal, "Developing System-based Voice Features for Detecting Parkinson's Disease Using Machine Learning Algorithms", Journal of Disability Research, Vol. 3, No. 1, 2024, p. 20240001.

[7] S. Zhao, G. Dai, J. Li, X. Zhu, X. Huang, Y. Li, M. Tan, L. Wang, P. Fang, X. Chen, N. Yan, "An interpretable model based on graph learning for diagnosis of Parkinson's disease with voice-related EEG", NPJ Digital Medicine, Vol. 7, No. 1, 2024, p. 3.

[8] L. Cardenas, K. Parajes, M. Zhu, S. Zhai, "Auto-Health: Advanced LLM-Empowered Wearable Personalized Medical Butler for Parkinson's Disease Management", Proceedings of the IEEE 14th Annual Computing and Communication Workshop and Conference, Las Vegas, NV, USA, 8-10 January 2024, pp. 375-379.

[9] L. Ali, A. Javeed, A. Noor, H. T. Rauf, S. Kadry, A. H. Gandomi, "Parkinson's disease detection based on features refinement through L1 regularized SVM and deep neural network", Scientific Reports, Vol. 14, No. 1, 2024, p. 1333.

[10] T. D. Pham, S. B. Holmes, L. Zou, M. Patel, P. Coulthard, "Diagnosis of pathological speech with streamlined features for long short-term memory learning", Computers in Biology and Medicine, Vol. 170, 2024, p. 107976.

[11] J. Varghese, A. Brenner, M. Fujarski, C. M. van Alen, L. Plagwitz, T. Warnecke, "Machine Learning in the Parkinson's disease smartwatch (PADS) dataset", NPJ Parkinson's Disease, Vol. 10, No. 1, 2024, p. 9.

[12] L. Ali, C. Chakraborty, Z. He, W. Cao, Y. Imrana, J. J. Rodrigues, "A novel sample and feature dependent ensemble approach for Parkinson's disease detection", Neural Computing and Applications, Vol. 35, No. 22, 2023, pp. 15997-16010.

[13] R. Kapila, S. Saleti, "Optimizing Predictive Models for Parkinson's Disease Diagnosis", Intelligent Technologies and Parkinson's Disease: Prediction and Diagnosis, IGI Global, 2024, pp. 255-275.

[14] M. Shivakoti, S. C. Medaramatla, D. Godavarthi, N. Shivakoti, "Prognoza: Parkinson's Disease Prediction Using Classification Algorithms", EAI Endorsed Transactions on Pervasive Health and Technology, Vol. 9, 2023.

[15] L. Sahu, R. Sharma, I. Sahu, M. Das, B. Sahu, R. Kumar, "Efficient detection of Parkinson's disease using deep learning techniques over medical data", Expert Systems, Vol. 39, No. 3, 2022, p. e12787.

[16] T. Vyas, R. Yadav, C. Solanki, R. Darji, S. Desai, S. Tanwar, "Deep learning-based scheme to diagnose Parkinson's disease", Expert Systems, Vol. 39, No. 3, 2022, p. e12739.

[17] M. Hosny, M. Zhu, W. Gao, Y. Fu, "A novel deep learning model for STN localization from LFPs in Parkinson's disease", Biomedical Signal Processing and Control, Vol. 77, 2022, p. 103830.

[18] K. Rajanbabu, I. K. Veetil, V. Sowmya, E. A. Gopalakrishnan, K. P. Soman, "Ensemble of Deep Transfer Learning Models for Parkinson's Disease Classification", Soft Computing and Signal Processing, Vol. 1340, Springer, 2022, pp. 135-143.

[19] S. Moradi, L. Tapak, S. Afshar, "Identification of Novel Noninvasive Diagnostics Biomarkers in the Parkinson's Diseases and Improving the Disease Classification Using Support Vector Machine", BioMed Research International, Vol. 2022, 2022.

[20] G. AlMahadin, A. Lotfi, M. M. Carthy, P. Breedon, "Enhanced Parkinson's disease tremor severity classification by combining signal processing with resampling techniques", SN Computer Science, Vol. 3, No 1, 2022, pp. 1-21.

[21] İ. Canturk, O. Günay, "Investigation of Scalograms with a Deep Feature Fusion Approach for Detection of Parkinson's Disease", Cognitive Computation, 2024, pp. 1-12.

[22] T. H. Aldhyani, A. H. Al-Nefaie, D. Koundal, "Modeling and diagnosis Parkinson disease by using hand drawing: deep learning model", AIMS Mathematics, Vol. 9, No. 3, 2024, pp. 6850-6877.

Enhancing Breast Cancer Diagnosis: A Hybrid Approach with Bidirectional LSTM and Variable Size Firefly Algorithm Optimization

Original Scientific Paper

Mandakini Priyadarshani Behera
Siksha 'O' Anusandhan (Deemed to be) University
Department of Computer Science and Engineering, Bhubaneswar, India
mandakini.beheracse@gmail.com

Archana Sarangi
Siksha 'O' Anusandhan (Deemed to be) University
Department of Electronics and Telecommunications, Bhubaneswar, India
archanasarangi@soa.ac.in

Debahuti Mishra*
Siksha 'O' Anusandhan (Deemed to be) University
Department of Computer Science and Engineering, Bhubaneswar, India
debahutimishra@soa.ac.in

*Corresponding author

Abstract *– Breast cancer stands as a significant global health challenge, ranking as the second leading cause of mortality among women. The increasing complexity of timely and accurate remote diagnosis has spurred the need for advanced technological solutions. Breast cancer prediction involves utilizing risk assessment models to identify individuals at higher risk, enabling early detection and personalized treatment strategies. This research meticulously assesses the effectiveness of various long short-term memory (LSTM) classifiers, including simple LSTM, Vanilla LSTM, Stacked LSTM, and Bidirectional LSTM, utilizing a comprehensive breast cancer dataset. Among these, the Bidirectional LSTM emerges as the preferred choice based on a thorough evaluation of accuracy, precision, recall, and F1-Score metrics. In a strategic move to further enhance precision, the Bidirectional LSTM integrates with the variable step-size firefly algorithm (VSSFF). Renowned for dynamically adjusting its step size, VSSFF offers adaptive exploration and exploitation capabilities in optimization tasks. The resulting hybrid model, HVSSFFLSTM, showcases superior performance in breast cancer prediction, suggesting potential applicability across diverse health conditions. Comparative analyses with other models highlight the exceptional accuracy rates of HVSSFFLSTM, achieving 99.78% (training) and 97.37% (testing), precision rates of 99.56% (training) and 97.22% (testing), recall rates of 100% (training) and 98.59% (testing), F1 scores of 99.82% (training) and 97.9% (testing) and specificity of 99.81% (training) and 99.15% (testing). This study not only underscores the adaptability of VSSFF as a valuable optimization tool but also emphasizes the promising prospects of the proposed hybrid model in advancing automated disease analysis. The results indicate its potential beyond breast cancer, suggesting broader applications in various medical domains.*

Keywords: *Simple LSTM, Vanilla LSTM, Stacked LSTM, Bidirectional LSTM, Firefly Optimization Algorithm, Variable Step Size Firefly Algorithm*

Received: February 16, 2024; Received in revised form: April 30, 2024; Accepted: May 2, 2024

1. INTRODUCTION

Breast cancer is a major health concern worldwide, underscoring the need for accurate risk assessment and early detection [1]. Machine learning (ML) and deep learning (DL) techniques are pivotal in improving detection methods [2]. ML algorithms like support vector machines (SVM), random forest, logistic regression (LR), decision trees (DT-C4.5), and k-nearest neighbours (k-NN) [3] aid in feature selection and classification. Recurrent neural networks (RNNs) [4], and long short-term memory (LSTM) [5] networks, excel in analysing mammographic images for abnormalities indicative of breast cancer [6]. Hybrid models integrating ML and

DL components offer a comprehensive approach, aiming to enhance detection accuracy. This study evaluates various LSTM classifiers—simple LSTM [5], Vanilla LSTM [7], Stacked LSTM [8], and Bidirectional LSTM [9] for breast cancer detection, with Bidirectional LSTM showing superior performance. The study introduces a hybrid model, VSSFFLSTM, combining Bidirectional LSTM with a variable step-size firefly algorithm (VSSFF) [10-12] to improve detection accuracy further. Utilizing well-established Breast Cancer Wisconsin (diagnostic) (WDBC) datasets [13] for training and validation ensures model consistency and performance.

The traditional firefly algorithm (FF) [14] static step-size hampers search effectiveness, necessitating dynamic adjustment. Initial larger step sizes are vital for identification and development, but dynamic alteration is needed with increased iterations for optimal performance. Whereas, the VSSFF improves upon FF by adapting the step size, resulting in better balance, faster convergence, and increased robustness. Integrating VSSFF with Bidirectional LSTM enhances breast cancer diagnosis by leveraging these improvements in optimization. The investigation introduces a hybrid model featuring several key contributions.

1. The hybrid breast cancer prediction model introduces a Bidirectional LSTM, improving predictive capabilities and overcoming hidden layer load challenges, marking a notable research innovation.
2. The VSSFF algorithm dynamically adjusts step size, enhancing the training efficiency of the Bidirectional LSTM by optimizing its weights, leading to an improved model with predictive accuracy and minimized mean square error (MSE) in the network output.
3. VSSFF further optimizes the Bidirectional LSTM by determining the optimal number of hidden neurons, utilizing initial random values and iterative refinement through the Adam optimizer. The integration of VSSFF with the Bidirectional LSTM forms a comprehensive strategy (HVSSFFLSTM).

This article follows a structured approach, beginning with a review of relevant literature in Section 2, followed by a comprehensive explanation of the methodologies in Section 3. Section 4 provides an analysis of the experiments conducted. Key findings are discussed in this section as well. Finally, Section 5 concludes the paper by discussing future avenues of research.

2. LITERATURE SURVEY

In Hazra et al. [15], an artificial neural network and a DT model are employed to scrutinize early-stage breast cancer characteristics, distinguishing between malignancy and benign nature. Another investigation by Naji et al. [16] analyses the BCWD dataset using five ML algorithms, providing valuable insights into their performance. A comprehensive evaluation of LSTM for breast cancer detection is conducted in a broader context by Behera et al. [17], providing insights into its capabilities. Mammographic image analysis and classification into normal, benign, and malignant classes are explored using CNN and Bidirectional LSTM architectures. According to Xia et al. [18] Innovative ensemble architectures, like the MTW CNN-BLSTM ensemble, aim to improve breast cancer prediction. In data mining methodologies, a statistical approach preprocesses data followed by a unique PSO framework for improved accuracy, sensitivity, and specificity. Multi-objective feature selection strategies incorporating ACO and PSO are developed for breast cancer diagnosis by Saturi et al. [19], enhancing detection probability by selecting relevant features. Additionally, a model named BPBRW with HKH-ABO mechanism is proposed for early-stage breast cancer diagnosis using breast magnetic imaging resonance data by Dewangan et al. [20].

The manuscript highlights a significant gap in breast cancer prediction research:

1. The absence of comprehensive comparative analyses among various ML and DL techniques.
2. Moreover, challenges in model interpretability, scalability, and generalizability remain unaddressed, indicating the need for further exploration.
3. However, there is a need for further exploration and development of ensemble techniques to enhance model accuracy and robustness.
4. Integrating innovative ensemble architectures, such as particle swarm optimization (PSO) and ant colony optimization (ACO), alongside emerging technologies like MRI data analysis, presents promising avenues for enhancing early-stage breast cancer diagnosis.
5. Therefore, future research efforts should prioritize rigorous comparative evaluations and innovative methodological advancements to bridge these critical gaps in breast cancer prediction and diagnosis.

3. METHODOLOGIES ADOPTED

In this work, simple LSTM and its variants such as; vanilla LSTM, stacked LSTM, and Bi-directional LSTM networks used. Along with this, the FF algorithm [14] is also used to optimize the positions based on fireflies' attractiveness, with intensity decreasing with the distance. Equation (1) guides the fireflies towards brighter positions, integrating attractiveness, distance, and randomness.

$$x_i = x_i + \beta_0 e^{-\gamma r_{ij}^2}\left(x_i - x_j\right) + \alpha(rand - \tfrac{1}{2}) \qquad (1)$$

3.1. VARIABLE STEP SIZE FIREFLY ALGORITHM (VSSFF)

The VSSFF [11] algorithm is an enhanced version of the FF, designed to overcome its limitations and improve convergence rates. It emphasizes a balance between global exploration and local exploitation to maximize benefits. In FF, a constant step size hampers effective searching, necessitating dynamic adjustment for optimal exploration-convergence equilibrium. Initial larger step sizes are needed for balanced identification and development in the early stages, gradually decreasing over iterations to maintain equilibrium. The choice between large or small step sizes depends on the optimization target's definition space. In [10] to sustain equilibrium between identification and development capabilities, the initial step size (α) should be relatively larger, gradually decreasing over iterations. In [12] the choice between a large or small search step size is contingent on the optimization target's definition space; a high-dimensional space requires a larger search step size, while a lower-dimensional space benefits from a smaller search step size, optimizing the algorithm's ability to address diverse optimization challenges as stated in Equation (2). Here, the number of existing iterations $max\ generation = maxmber\ of\ iterations$. The VSSFF operational steps are stated in Algorithm 1.

$$\alpha(t) = {}^{0.4}\Big/\left(1 + \exp\left[0.015 \times \frac{(t - \max generation)}{3}\right]\right) \quad (2)$$

Algorithm 1. VSSFF operational steps [11]

Step 1: Initialize each firefly randomly.

Step 2: Evaluate the fitness function value for the initialized population.

Step-3: Assess the light intensity.

Step 4: Determine the light absorption coefficient γ.

Step 5: Evaluate the non-constant step size α using Equation (2).

Step 6: Update the position of a specific firefly towards another attractive firefly based on Equation (1).

Step 7: Calculate the latest solution and update the light intensity.

Step 8: Modify the locations of fireflies based on their rank to obtain the current optimal solution.

Step-9: End if termination conditions are met and select the optimal solution; otherwise, return to Step 2.

3.2. PROPOSED BIDIRECTIONAL LSTM NETWORK WITH VSSFF (HVSSFFLSTM) FOR CLASSIFICATION

A novel hybrid approach, HVSSFFLSTM, integrates Bidirectional LSTM with VSSFF to enhance breast cancer classification accuracy. VSSFF collaborates with Bidirectional LSTM to optimize architecture and hyperparameters for this purpose. Bidirectional LSTM captures temporal dependencies, while VSSFF explores hyperparameter space for crucial configurations. The potential of this hybrid technique can be further realized through meticulous parameter tuning and rigorous model validation of Bidirectional LSTM. Prudent adjustment and robust validation promise enhanced performance and reliability of HVSSFFLSTM, contributing significantly to accurate breast cancer classification. In this framework, VSSFF optimizes Bidirectional LSTM parameters, particularly focusing on weight optimization. The VSSFF algorithm systematically assesses the ideal number of hidden neurons within each hidden layer. Initial random values are assigned to the primary weights of the network, and an Adam optimizer with maximum epoch=100, batch size=512, initial learning rate=0.001, grounded in gradient descent principles, is utilized to iteratively refine these network weights. Subsequently, the model undergoes comprehensive testing to gauge its performance following the adjustments made by the VSSFF algorithm. This approach not only underscores the pivotal role of weight optimization in fine-tuning the predictive capabilities of the Bidirectional LSTM but also highlights the significance of determining the optimal number of hidden neurons to elevate overall model efficacy. The integration of the VSSFF algorithm with the Bidirectional LSTM reflects a holistic strategy aimed at achieving optimal predictive accuracy in the context of breast cancer data classification. The workflow of the manuscript is presented in Fig.1. A concise mathematical representation of the hybrid model VSSFF is presented below as Equation (3), where let θ denote the set of parameters to be optimized, J(θ) represents the objective function related to breast cancer classification, and the VSSFF optimization process is denoted as by Equation (3).

$$\theta\ new = VSSFF\ (\theta\ old, J\ (\theta\ old)) \quad (3)$$

Let, w represents the weights of the hidden layers in Bidirectional LSTM. The model is denoted as Bidirectional LSTM (w). The optimized parameters θ* from the VSSFF algorithm are used to fine-tune the weights of the Bidirectional LSTM model. The VSSFFLSTM model is represented as (θ*,w*), where w* are the adjusted weights. The optimization process involves iteratively updating the parameters θ using the VSSFF algorithm as stated in Algorithm 2.

Algorithm 2. The proposed HVSSFFLSTM algorithm

Step 1: Let D represent the breast cancer dataset, with corresponding class labels (benign: $y = 0$, malignant: $y = 1$) and split D into training (D_{Train}) and testing (D_{Test}) sets.

Step 2: Initialize a population P of fireflies with random hyperparameters.

Step 3: Define the Bidirectional LSTM model architecture, specifying parameters such as the number of LSTM layers, units, and dropout rates.

Step 4: Train the Bidirectional LSTM model on D_{Train} to obtain initial weights $w_{initial}$ Evaluate the model's performance on the D_{Train} and D_{Test} using binary cross-entropy.

Step 5: Develop the VSSFF approach to optimize the hyperparameters of the Bidirectional LSTM model.

```
Step 6:  Define the hyperparameter space $\Theta$, including the number of LSTM layers, units, dropout rates, etc.
Step 7:  Create a firefly population $F$, where each firefly $f_i$ is represented by a set of hyperparameters $\theta_i \in \Theta$.
Step 8:  Define the fitness function $J(\theta_i, D_{Train})$ based on the training dataset.
Step 9:  Implement the variable step size method, adjusting the step size based on firefly brightness.
Step 10: Select the hyperparameters $\theta^*$ from the firefly population $F$ based on the highest brightness, optimizing the Bidirectional LSTM model.
Step 11: Train the Bidirectional LSTM model using the optimized hyperparameters $\theta^*$, resulting in final weights $w_{final}$. Evaluate the final model's performance on the $D_{Test}$.
```

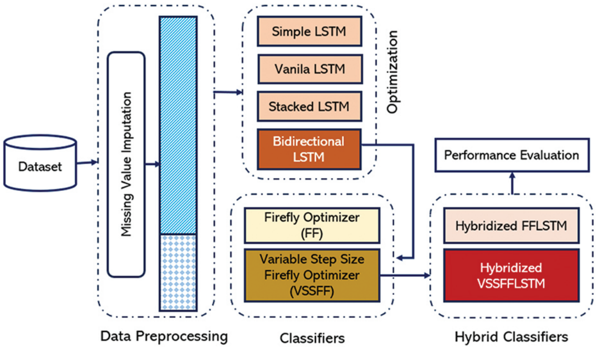


Fig. 1. Schematic layout of the proposed strategy for breast cancer prediction

3.3. DATASET AND MISSING VALUE IMPUTATION

The manuscript utilizes the WDBC dataset from the UCI ML repository [13], comprising 569 records with a distribution of 62.7% benign and 37.3% malignant breast cancer cases. Each record includes an ID number, diagnostic label (B for benign, M for malignant), and 30 real-valued input features representing significant cell nuclei characteristics such as radius, texture, perimeter, etc. Missing values in the dataset are imputed with 0 or 1 based on their respective values <50 or >=50. 80% of the dataset is used for training the model, while the remaining 20% is reserved for evaluating the model's performance.

4. EXPERIMENTAL ANALYSIS

In this section, we conduct comprehensive experimental analyses to assess the performance of our proposed model. We compare our results with various models using diverse performance metrics to gain insights into the effectiveness and superiority of our approach. The LSTM classifier is trained with 100 epochs, a batch size 512, and the Adam optimizer for optimization. The experiments were executed on a system equipped with a 1.80 GHz Intel(R) Core (TM) i5-8265U processor and 8.00 GB RAM, running on the Windows 10 operating system. All ML approaches discussed in this study were implemented using the Scikit-learn library and the Python programming language.

4.1. PARAMETERS USED

The training parameters used in this manuscript are outlined in Table 1, and Table 2 provides the parameter settings for the discussed hybrid models.

Table 1. Training Parameters

Optimizer	Maximum Epoch	Batch Size	Initial Learning Rate
Adam	100	512	0.001

Table 2. Parameter setting for hybrid models

Hybrid Models	Population Size	Iteration	Upper Bound	Lower Bound
HFFLSTM	50	200	5	-5
HVSSFFLSTM	50	200	5	-5

4.2. RESULTS ANALYSIS

This research follows a structured experimental approach consisting of two phases. Initially, four variants of LSTM networks are thoroughly explored, with Bi-directional LSTM showing superior performance across various evaluation metrics. The Bi-directional LSTM likely achieved the highest values for all metrics due to its ability to capture bidirectional contextual information, generate comprehensive feature representations, reduce information loss, effectively handle temporal dependencies, and maintain robustness to input variability, which collectively contribute to its superior performance compared to other LSTM variants. Encouraged by these results, the research progresses to the second phase, focusing on optimizing Bidirectional LSTM with FF and VSSFF algorithms. This transition marks a strategic progression, aiming to uncover and capitalize on the most effective configurations for robust performance in breast cancer prediction.

The initial experimentation phase, detailed in Table 3, meticulously analyses various LSTM model variants across training and testing datasets. Results highlight the Bidirectional LSTM's distinct superiority, demonstrating exceptional performance in both phases. In training, it achieves 96.70% accuracy, 97.90% precision, 96.89% recall, 97.40% F1-Score, and 97.79% specificity. In testing, the Bidirectional LSTM outperforms alternative variants with 96.49% accuracy, 97.18% precision, 97.18% recall, 97.18% F1-Score, and 97.18% specificity.

Table 3. Performance of different variants of LSTM

Execution Stages	Performance Metrics (in %)	Simple LSTM	Vanilla LSTM	Stacked LSTM	Bi-directional LSTM
Training	Accuracy	95.16	96.04	96.48	96.70
	Precision	94.59	98.22	97.56	97.90
	Recall	97.90	95.51	96.89	96.89
	F1-Score	96.21	96.58	97.23	97.40
	Specificity	94.78	94.97	95.86	97.79
Testing	Accuracy	95.61	94.73	92.10	96.49
	Precision	96.96	95.52	95.31	97.18
	Recall	95.52	95.52	91.04	97.18
	F1-Score	96.24	95.52	93.12	97.18
	Specificity	95.88	95.93	95.11	97.18

Table 4. Performance of HFFLSTM and HVSSFFLSTM models

Execution Stages	Performance Metrics (in %)	HFFLSTM	HVSSFFLSTM
Training	Accuracy	99.34	99.78
	Precision	98.96	99.56
	Recall	1.00	1.00
	F1-Score	99.47	99.82
	Specificity	98.85	99.81
Testing	Accuracy	96.49	97.37
	Precision	98.55	97.22
	Recall	95.77	98.59
	F1-Score	97.14	97.9
	Specificity	97.12	99.15

The empirical findings strongly support the Bidirectional LSTM as the most promising variant, leading to its strategic selection for optimization. The VSSFF algorithm is then employed to enhance its predictive capabilities further. The optimization aims to fine-tune and improve key performance metrics like accuracy, precision, recall, and F1-Score in breast cancer prediction, contributing to more reliable outcomes in medical diagnostics. The exploration extends to hybridized forms of Bidirectional LSTM, including HFFLSTM and HVSSFFLSTM. In HFFLSTM, the FF algorithm integrates seamlessly with Bi-directional LSTM, optimizing hyperparameters to enhance pattern recognition capabilities by fine-tuning parameters such as layer numbers, units, and dropout rates. Conversely, HVSSFFLSTM combines the VSSFF algorithm with Bi-directional LSTM, introducing dynamic step-size adjustment for efficient hyperparameter exploration. The FF algorithm optimizes hyperparameters, while VSSFF introduces dynamic step-size adjustment, facilitating more efficient exploration of the hyperparameter space. The experimental evaluations include accuracy plot analyses and comprehensive performance metrics. These assessments highlight the superior predictive capabilities of HVSSFFLSTM, showcasing its potential to advance breast cancer prediction models compared to HFFLSTM and other existing approaches. Subsequent sections provide a detailed exploration of this experimentation phase, offering insights into the optimization process intricacies and innovative strides toward enhancing breast cancer prediction models.

Within Table 4, we meticulously conduct a comparative analysis, delving into the nuanced distinctions between HFFLSTM and HVSSFFLSTM. The outcomes of this detailed examination underscore the consistent superiority of HVSSFFLSTM over HFFLSTM during the training phase, boasting remarkable metrics such as accuracy (99.78%), precision (99.56%), recall (100%), F1-Sscore (99.82%) and specificity (99.81%). In the testing phase, HVSSFFLSTM continues to excel, demonstrating impressive performance in accuracy (97.37%), recall (98.59%), F1-Sscore (97.9%), and specificity (99.15%). Albeit with a marginally lower precision (97.22%) when juxtaposed with the HFFLSTM model, which registers at (98.55%). This minor discrepancy is deemed manageable, affirming the overall robustness of HVSSFFLSTM.

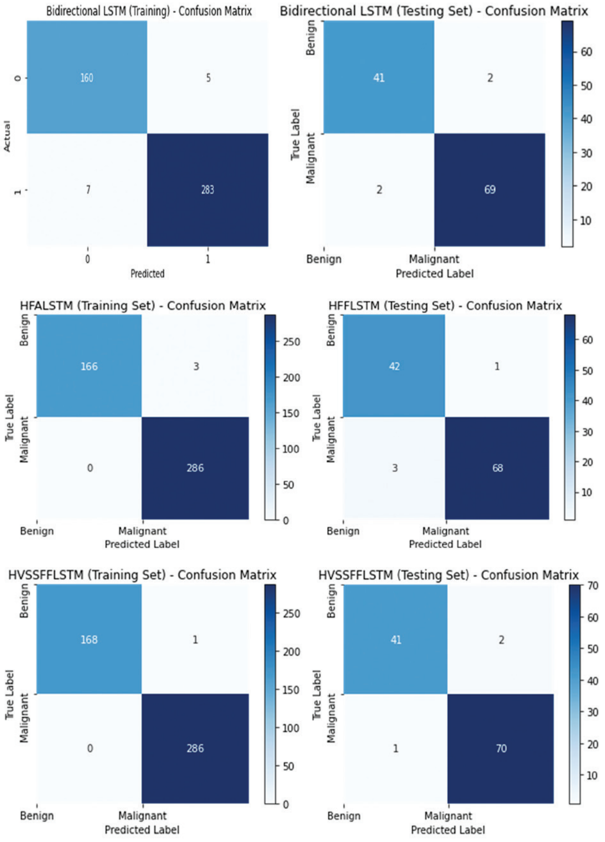


Fig. 2. Confusion matrix for training and testing with Bidirectional LSTM, HFFLSTM, and HVSSFFLSTM

Fig. 2. illustrates confusion matrices for Bidirectional LSTM, HFFLSTM, and VSSFFLSTM models in breast cancer prediction across training and testing phases. VSSFFLSTM stands out with significantly elevated accuracy compared to Bidirectional LSTM and HFFLSTM, indicating its superior performance. Subsequent meticulous assessment of classification performance, considering metrics like accuracy, precision, recall, and F1-Score, unveils the nuanced advantages of HVSSFFLSTM over its counterparts. This analysis showcases HVSSFFLSTM's ability to deliver accurate and reliable predictions in breast cancer prediction. Visualization of confusion matrices and detailed performance analysis not only quantitatively evaluates models but also highlights HVSSFFLSTM's strengths and capabilities. This empirical evidence substantiates the efficacy and potential superiority of the proposed model, emphasizing its significance in advancing breast cancer prediction methodologies.

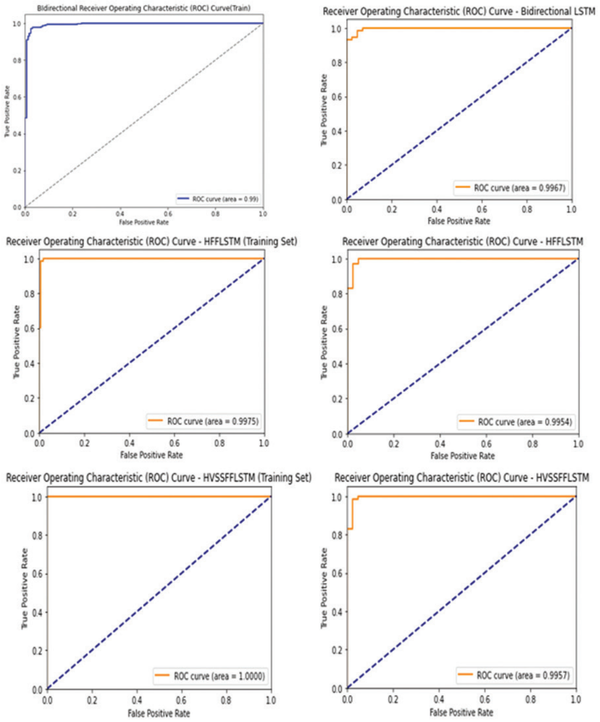


Fig. 3. ROC curve results for training and testing with Bidirectional LSTM, HFFLSTM, and HVSSFFLSTM

Fig. 3. illustrates the graphical ROC analysis for Bidirectional LSTM, HFFLSTM, and VSSFFLSTM in breast cancer prediction, showcasing the true positive rate (TPR) versus false positive rate (FPR) for both training and testing datasets. Specifically, the ROC values for Bidirectional LSTM are 99% for training and 99.67% for testing, while HFFLSTM achieves 99.75% for training and 99.54% for testing. Notably, the ROC curve values for the proposed HVSSFFLSTM algorithm stand at 100% for training and 99.57% for testing. These results highlight the exceptional discriminative performance of the HVSSFFLSTM model in effectively distinguishing between TP and FP during breast cancer prediction.

Fig. 4. presents graphical representations showing the dynamic fluctuations in accuracy across epochs during both training and testing phases for Bidirectional LSTM, HFFLSTM, and HVSSFFLSTM models in breast cancer prediction. These visuals offer insights into the evolution of accuracy for each model throughout the training and testing processes, aiding in understanding the learning trajectories and performance trends. The fluctuations in accuracy over epochs enable observation of how each model adapts and refines its predictive capabilities with iterative learning, which is crucial for evaluating stability, convergence, and overall learning efficiency. Fig. 4. serves as a visual narrative, providing a comprehensive overview of the learning dynamics exhibited by the models during breast cancer prediction, enhancing understanding of temporal aspects of model performance, and identifying key epochs influencing predictive power.

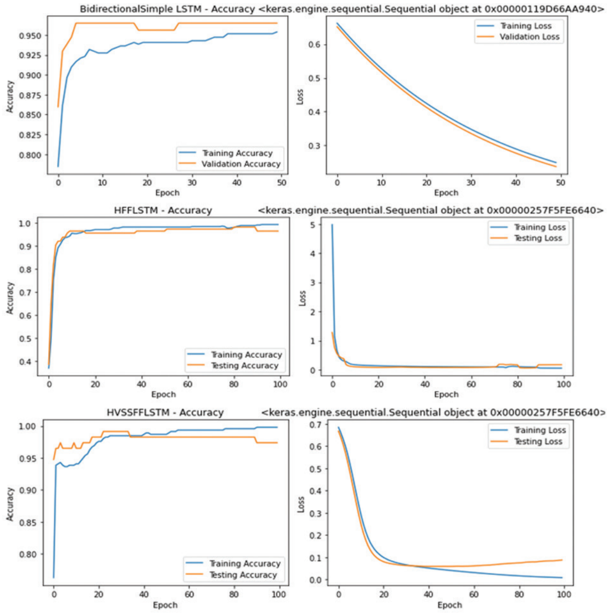


Fig. 4. Comparison of accuracy achieved for training and testing with Bidirectional LSTM, HFFLSTM, and HVSSFFLSTM

The efficacy of the VSSFF algorithm lies in its adaptive step size, dynamically balancing exploration and exploitation. This addresses the limitations of a fixed step size, preventing suboptimal results. The algorithm enables more effective navigation through intricate optimization landscapes and contributes to faster convergence toward optimal solutions by refining its exploration strategy through iterations. Its robustness across various optimization problems provides flexibility to adapt exploration strategies based on landscape characteristics. These features render the VSSFF algorithm more effective than the FF algorithm. Consequently, this research proposes a more robust model HVSSFFLSTM for breast cancer data classification, leveraging the enhanced capabilities of the VSSFF algorithm.

4.3 STATISTICAL VALIDATION AND EXECUTION TIME COMPARISON

Table 5 displays the McNemar test [21] results comparing HFFLSTM vs. Bidirectional LSTM and HVSSFFLSTM vs. HFFLSTM models in training and testing phases, revealing significant performance differences. HVSSFFLSTM demonstrates the shortest training and testing times, outperforming other models.

Table 5. McNemar's test results

Execution Stages	Tests and p-values	HFFLSTM vs. Bidirectional LSTM	HVSSFFLSTM vs. HFFLSTM
Training	McNemar Test Statistic	5.21	7.78
	p-value	0.022	0.005
Testing	McNemar Test Statistic	5.68	4.89
	p-value	0.0171	0.026

Fig. 5 compares execution times (in milliseconds) for Bidirectional LSTM, HFFLSTM, and HVSSFFLSTM models. HVSSFFLSTM demonstrates the shortest training time at 2.87 ms, followed by HFFLSTM at 3.46 ms and Bidirectional LSTM at 4 ms. In testing, HVSSFFLSTM also shows the fastest execution time at 2.11 ms, outperforming HFFLSTM (3.44 ms) and Bidirectional LSTM (3.55 ms). These results underscore HVSSFFLSTM's superior efficiency in both the training and testing phases.

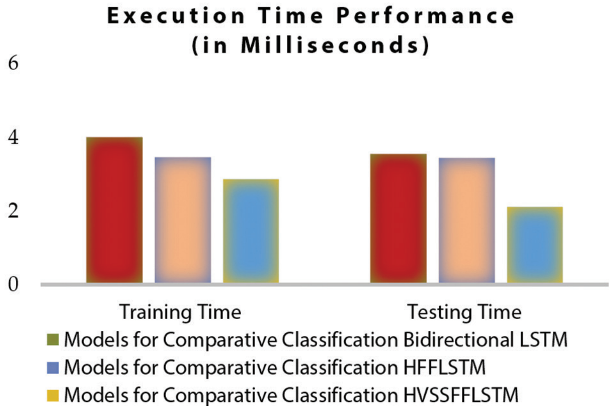


Fig. 5. The recorded performance in terms of execution time

4.4. PRINCIPAL INSIGHTS AND DISCUSSIONS

This section provides a detailed analysis of LSTM networks for breast cancer prediction in two phases. Four LSTM variants are initially explored, with Bidirectional LSTM identified as the most promising. Bidirectional LSTM is then optimized using FF and VSSFF algorithms to enhance predictive capabilities, leading to hybrid forms like HFFLSTM and HVSSFFLSTM. Experimental evaluations highlight HVSSFFLSTM's superior predictive capabilities, confirmed by accuracy plots, ROC analyses, and comprehensive metrics. Comparative analysis consistently favors HVSSFFLSTM, with statistical validation confirming its significance. HVSSFFLSTM also demonstrates computational efficiency, positioning it as a promising candidate for resource optimization. These findings contribute to a deeper understanding of model efficacy and computational efficiency in breast cancer prediction.

The proposed classification models consistently exhibit robust performance across training and testing phases, with HVSSFFLSTM showing superior performance. During training, HVSSFFLSTM achieves exceptional results with 99.78% accuracy, 99.56% precision, 100% recall, 99.82% F1 Score, and 99.81% specificity. Testing also demonstrates strong performance with 99.37% accuracy, 97.22% precision, 98.59% F1 Score, and 99.15% specificity. Statistical validation and execution time performance solidify HVSSFFLSTM as a noteworthy advancement in breast cancer detection research, showcasing its precision and reliability in classification.

5. CONCLUSION AND FUTURE SCOPE

This study thoroughly examines four LSTM algorithms and two hybrid models for breast cancer classification. Results consistently show the superiority of the proposed hybrid model. HVSSFFLSTM, HFFLSTM, and Bidirectional LSTM are ranked as the top three models. The study suggests avenues for future exploration, including additional hybrid models and diverse datasets. The proposed predictive methods demonstrate versatility, with potential applications in various medical conditions beyond breast cancer. This research sets the stage for continued innovation in medical predictive modelling.

6. REFERENCES

[1] A. B. Nassif, M. A. Talib, Q. Nasir, Y. Afadar, O. Elgendy, "Breast cancer detection using artificial intelligence techniques: A systematic literature review", Artificial Intelligence in Medicine, Vol. 127, 2022, p. 102276.

[2] S. Dadsetan, D. Arefan, W. A. Berg, M. L. Zuley, J. H. Sumkin, S. Wu, "Deep learning of longitudinal mammogram examinations for breast cancer risk prediction", Pattern Recognition, Vol. 132, 2022, p. 108919.

[3] M. A. Naji, S. E. Filali, K. Aarika, E. H. Benlahmar, R. A. Abdelouhahid, O. Debauche, "Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis", Procedia Computer Science, Vol. 191, 2021, pp. 487-492.

[4] G. R. Paul, J. Preethi, "A novel breast cancer detection system using SDM-WHO-RNN classifier with LS-CED segmentation", Expert Systems with Applications, Vol. 238, Part C, 2023, p. 1217.

[5] M. G. Lanjewar, K. G. Panchbhai, L. B. Patle, "Fusion of transfer learning models with LSTM for detection of breast cancer using ultrasound images", Computers in Biology and Medicine, Vol. 169, 2024, p. 107914.

[6] K. Atrey, B. K. Singh, N. K. Bodhey, R. B. Pachori, "Mammography and ultrasound-based dual modality classification of breast cancer using a hybrid deep learning approach, Biomedical Signal Processing and Control", Vol. 86, Part A, 2023, p. 104919.

[7] W. Fang, R. Zhu, J. C. Lin, "An air quality prediction model based on improved Vanilla LSTM with multichannel input and multiroute output", Expert Systems with Applications, Vo. 211, 2023, p. 118422.

[8] E. Alabdulkreem, N. Alruwais, H. Mahgoub, A. K. Dutta, M. Khalid, R. Marzouk, A. Motwakel, S. Drar, "Sustainable groundwater management using stacked LSTM with deep neural network", Urban Climate, Vol. 49, 2023, p. 101469.

[9] S. Tripathi, S. K. Singh, H. K. Lee, "An end-to-end breast tumor classification model using context-based patch modeling – A BiLSTM approach for image classification", Computerized Medical Imaging and Graphics, Vol. 87, 2021, p. 101838.

[10] M. P. Behera, A. Sarangi, D. Mishra, S. K. Mohapatra, "Variable Step Size Firefly Algorithm for Automatic Data Clustering", Proceedings of ICICC Intelligent and Cloud Computing, Bhubaneswar, Odisha, India, 22-23 October 2021, pp. 243-253.

[11] M. P. Behera, A. Sarangi, D. Mishra, S. K. Sarangi, "Optimizing Multi-Layer Perceptron using Variable Step Size Firefly Optimization Algorithm for Diabetes Data Classification", International Journal of Online and Biomedical Engineering, Vol. 19, No. 4, 2023, pp. 124-139.

[12] M. P. Behera, A. Sarangi, D. Mishra, P. K. Mallick, J. Shafi, P. N. Srinivasu, M. F. Ijaz, "Automatic Data Clustering by Hybrid Enhanced Firefly and Particle Swarm Optimization Algorithms", Mathematics, Vol. 10, 2022, p. 3532.

[13] UCI machine learning repository, http://archive.ics.uci.edu/ml/ (accessed: 2023)

[14] X. S. Yang, "Metaheuristic Optimization: Nature-Inspired Algorithms and Applications", Artificial Intelligence, Evolutionary Computing and Metaheuristics, Studies in Computational Intelligence, Vol. 427, Springer, 2013.

[15] R. Hazra, M. Banerjee, L. Badia, "Machine learning for breast cancer classification with ANN and Decision Tree", Proceedings of the 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference, Vancouver, BC, Canada, 4-7 November 2020, pp. 522-527.

[16] M. A. Naji, S. E. Filali, K. Aarika, E. H. Benlahmar, R. A. Abdelouhahid, O. Debauche, "Machine learning algorithms for breast cancer prediction and diagnosis", Procedia Computer Science, Vol. 191, 2021, pp. 487-492.

[17] M. P. Behera, A. Sarangi, D. Mishra, S. K. Sarangi, "Breast Cancer Prediction Using Long Short-Term Memory Algorithm", Proceedings of the 5th International Conference on Computational Intelligence and Networks, Bhubaneswar, India, 1-3 December 2022, pp. 1-6.

[18] T. Xia, Y. Song, Y. Zheng, E. Pan, L. Xi, "An ensemble framework based on convolutional bi-directional LSTM with multiple time windows for remaining useful life estimation", Computers in Industry, Vol. 115, 2020, p. 103182.

[19] R. Saturi, P. Premchand, "Multi-Objective Feature Selection Method by Using ACO with PSO Algorithm for Breast Cancer Detection", International Journal of Intelligent Engineering & Systems, Vol. 14, No. 5, 2021.

[20] K. K. Dewangan, D. K. Dewangan, S. P. Sahu, R. Janghel, "Breast cancer diagnosis in an early stage using novel deep learning with hybrid optimization technique", Multimedia Tools and Applications, Vol. 81, No. 10, 2022, pp. 13935-13960.

[21] D. A. Redelmeier, R. J. Tibshirani, "A simple method for analyzing matched designs with double controls: McNemar's test can be extended", Journal of Clinical Epidemiology, Vol. 81, 2017, pp. 51-55.

PMiner: Process Mining using Deep Autoencoder for Anomaly Detection and Reconstruction of Business Processes

Original Scientific Paper

Veluru Chinnaiah*
Department of CSE,
Vijaya Engineering College,Khammam,Telangana, India
vtchinna2k12@gmail.com

Vadlamani Veerabhadram
Department of C. S. E
CVR College of Engineering,
Hyderabad, Telangana, India
drbhadram@gmail.com

*Corresponding author

Ravi Aavula
Department of CSE - DS,
Anurag University, Hyderabad
aavularavi@gmail.com

Srinivas Aluvala
Department of Computer Science and Artificial Intelligence, SR University
srinu.aluvala@gmail.com

Abstract *– We proposed a deep learning-based process mining framework known as* ***PMiner*** *for automatic detection of anomalies in business processes. Since there are thousands of business processes in real-time applications such as e-commerce, in the presence of concurrency, they are prone to exhibit anomalies. Such anomalies if not detected and rectified, cause severe damage to businesses in the long run. Our Artificial Intelligence (AI) enabled framework* ***PMiner*** *takes business process event longs as input and detects anomalies using a deep autoencoder. The framework exploits a deep autoencoder technique which is well-known for Its ability to discriminate anomalies. We proposed an algorithm known as Intelligent Business Process Anomaly Detector (IBPAD) to realize the framework. This algorithm learns from historical data and performs encoding and decoding procedures to detect business process anomalies automatically. Our empirical results using the BPI Challenge dataset, released by the IEEE Task Force on Process Mining, revealed that* ***PMiner*** *outperforms state-of-the-art methods in detecting business process anomalies. This framework helps businesses to identify process anomalies and rectify them in time to leverage business continuity prospects.*

Keywords*:* *Process Mining, Artificial Intelligence, Deep Autoencoder, Long Short Term Memory, Deep Learning*

Received: October 21, 2023; Received in revised form: February 9, 2024; Accepted: February 9, 2024

1. INTRODUCTION

Enterprise business applications are very complex and involve several hundreds of business processes. Often millions of users across the globe use such applications. Each business process involved in the application can be accessed by thousands of users simultaneously. In other words, there is concurrent access to business processes. It may lead to anomalous behaviour of business processes in terms of sequence of events or temporal dimension. Detection of anomalies in business processes is a tedious and complex phenomenon [1]. To enable the discovery of business processes, business process event logs are generated. Process mining is the science of dealing with business processes and analysing them to discover potential faults in the execution of business processes [2]. Complex business processes should be understood from multiple perspectives towards discovering actionable knowledge [3]. Process mining research involves diversified activities aimed at monitoring, tracking and rectifying business processes.

Many researchers focused on process mining since it is crucial for enterprise-level businesses. Association rule learning is one of the techniques used in [1] and [4] for finding business anomalies. Machine learning approaches are widely used for process mining as discussed in [5] and [6]. Advanced neural network models or deep learning techniques are also used by researchers to leverage business processes. This kind of research includes repairing missing activities [4], process prediction [7], anomaly detection [8, 9] and outcome prediction [6]. Hybrid learning approaches are also found important for process mining as discussed in [4] and [10]. Business process anomaly classification is found significant as explored in [11] and [12]. From the literature, it is observed that process mining research focuses on different aspects. However, an integrated approach with process discovery, anomaly detection and enhancement still requires further research. Our contributions to this paper are as follows.

1. We proposed an Artificial Intelligence (AI) enabled framework known as as ***PMiner*** which takes business process events longs as input and detects anomalies using a learning-based approach. It also has provisions for rectifying anomalies to improve the quality of business processes.

2. We proposed an algorithm known as Intelligent Business Process Anomaly Detector (IBPAD) to realize the framework. This algorithm learns from historical data and performs encoding and decoding procedures to detect business process anomalies automatically.

3. We evaluated our framework using the BPI Challenge dataset, released by the IEEE Task Force on Process Mining, which revealed that ***PMiner*** outperforms state-of-the-art methods in detecting business process anomalies. This framework helps businesses to identify process anomalies and rectify them in time to leverage business continuity prospects.

The remainder of the paper is structured as follows. Section 2 reviews existing research on process anomaly detection. Section 3 presents the proposed framework for the automatic detection of process anomalies and rectifying them. Section 4 presents the results of our empirical study. Section 5 discusses important findings in our research along with limitations. Section 6 concludes our work besides providing scope for future research.

2. RELATED WORK

This section reviews existing methods of process mining involving anomaly detection and rectification. The literature review covers research from 2013 to 2023. The rationale behind choosing older references is that they do have credible process mining research. Sungkono et al. [1] observed that ERP systems manage business processes, generating extensive logs. This study integrates process mining, fuzzy decision-making, and association rule learning to detect anomalies, enhancing fraud detection accuracy at low confidence levels. Kovalchuk et al. [2] found that deep learning, specifically LSTM models, enhances process mining for business operations. This approach combines accuracy and explainability, generating informative graphs. Stefanini et al. [3] stated that process Mining is a valuable technique for business process analysis, though its managerial potential remains underexplored. This review identifies research gaps and proposes a research agenda for its application in various business contexts. Chen et al. [13] observed that process mining bridges process modelling and data mining. To propose an LSTM-based model to repair missing activity labels in event logs, outperforming existing methods. Future work includes expanding and optimizing the approach.

Koninck et al. [14] introduced representation-learning techniques for business processes, enabling low-dimensional vectors for activities, traces, logs, and models. Applications include trace clustering and process model comparison. Future research avenues include interpretability and incorporating additional data dimensions—Joaristi et al. [15] utilized event logs for business process analysis. Existing encoding methods focus on control flow, leaving out other aspects. Deep-TRace2Vec, a deep learning approach, produces superior trace representations considering multiple perspectives. In Future, the work includes anomaly detection and transformer neural networks. Dewandono et al. [4] proposed a hybrid method combining association rule learning and process mining to improve fraud detection accuracy with fewer false discoveries compared to process mining alone. Vasumathi and Vijayakamal [16] showed that enterprise applications with Service Oriented Architecture (SOA) became complex. A framework using auto encoders improves these aspects, especially with Probabilistic Auto Encoder based Anomaly Detection (PAE-AD). Empirical results support its efficiency. Future work includes deep learning integration.

Fettke et al. [7] used process mining to reconstruct business processes from digital traces. A systematic review examines 32 methods to identify strengths, weaknesses, and research gaps. Unified benchmarks could enhance future process prediction approaches. According to Dumas et al. [17] complex business systems generate event logs that can be analysed for predictive business constraint monitoring, allowing early intervention. Implemented in ProM, validated using cancer treatment data. Further enhancements could involve different similarity measures and classification techniques for more significant accuracy. Charles et al. [18] found that organizations face challenges in detecting process abnormalities. A novel approach using conformance analysis identifies abnormalities by comparing successful and failed process instances. Fitness scores predict anomalies. Alexander et al. [19] observed that detecting subtle changes and anomalies in business processes is crucial. A neural network-based system can filter noisy event logs and detect anomalies without prior knowledge. In Future, this work includes investigating frequent anomalies and different noise levels. Neural networks are applicable and can capture underlying process patterns in event logs.

Franczyk et al. [11] proposed a semi-supervised deep learning classification model that effectively identifies anomalies in business process event sequences. It considers time dependencies and outperforms existing approaches in accuracy. In Future, we need to improve time-related anomaly detection and integrate the model into real-time environments. Flammini et al. [20] improved process mining with IoT log analytics and machine learning to detect and fix IoT anomalies, enhancing resilience. Research should address protocols and error predictability. Consistent Event Logs are key to Self-Healing in IoT-based CPS. Hemmer et al. [21] used process mining to detect IoT system misbehaviours and attacks, even with heterogeneous platforms and protocols. It employs data pre-processing and clustering techniques for predictive security. Experiments demonstrate its effectiveness. Future work involves automated countermeasures and deep learning integration. Capurro et al. [22] said that process mining in healthcare analyses processes using data from information systems. A literature review examines 74 relevant papers, providing insights and guidance for future applications in healthcare.

Cristina Nicoleta [23] discussed Industry 4.0 reliability and safety, suggesting a method for real-time robotic process verification with IIoT and Celonis. It enhances quality control and cuts errors, costs, and downtime. While focusing on a synthetic robotic arm, it offers a blueprint for boosting real-time industrial automation. Vanhoof et al. [24] focused on corporate fraud, particularly internal transaction fraud, which is costly. Process mining helps detect fraud by analysing event logs. A case study confirms its benefits in mitigating internal transaction fraud, especially in auditing and compliance checking. Pauwels and Calders et al. [25] automated modelling of behaviour captured in complex log files, enabling anomaly detection and concept drift identification using extended Dynamic Bayesian Networks. Luettgen et al. [5] proposed auto encoder-based approach for detecting and interpreting anomalies in business processes, achieving an F1 score of 0.87. Gyunam et al. [26] opined that process mining extracts insights but lacks actionable improvements. This framework connects monitoring with automated actions for process enhancement, successfully tested on real systems.

Okubo and Kaiya [27] Introduced a method for enhancing security in the DevOps lifecycle, focusing on threat analysis, attack detection, vulnerability extraction, and countermeasure assessment. Tested in a development case, it proves effective. Clemente et al. [8] proposed a 5G-oriented cyber defence architecture that employs deep learning for efficient cyber threat detection and self-adaptation to network traffic fluctuations. Experiments demonstrate its effectiveness. In Future, the work includes optimizing deep learning models and real-data training. Benedi et al. [28] presented emotive process mining algorithms for analysing human behaviour patterns in ambient assisted living environments. Fathalla et al. [29] introduced a deep reinforcement learning approach for business process anomaly detection, using limited labelled data and exploring unlabelled data. The model outperforms existing methods. Lagraa [29] discussed an approach using process mining to investigate and track malicious activities in authentication events, improving defence systems against such events. Guha and Samanta [10] presented a hybrid model for anomaly detection (AD) in title insurance using autoencoders (AE) and one-class support vector machines (OSVM). This approach shows promise but requires improvements in training and data-generative techniques.

Ashok Kumar et al. [30] focused on Conformance Checking (CC) which assesses the alignment between process models and real execution. Process Mining aids analysis, validation, and improvement. Challenges include data volume, control flow focus, and tool efficiency, suggesting room for future enhancements. Kratsch et al. [6] Predictive process monitoring anticipates business process behaviour. Deep learning outperforms classical machine learning, especially with high variant-to-instance ratios and imbalanced variables. Future research should consider broader log types and develop decision models. Luettgen et al. [12] explored BINet, a neural network for real-time multi-perspective anomaly detection in business process event logs. It outperforms other methods on synthetic and real-life datasets. BINet is adaptable for autonomous operation and can handle concept drift. In future, the work may discuss issues of forgetting in repeated event sequences. Folino and Pontieri [31] stated that process mining research is extending to less structured logs from non-process-aware systems. However, interpreting deep neural networks remains challenging. Research in Explainable DL aims to address this, and informed PM methods are being developed to utilize expert guidance. From the literature, it is observed that process mining research focuses on different aspects. However, an integrated approach with process discovery, anomaly detection and enhancement still requires further research.

3. PROPOSED FRAMEWORK

This section presents a proposed framework and the underlying methodology for the automatic detection of business process anomalies and solving the problem.

3.1. PROBLEM DEFINITION

Provided a set of business processes in the form of event logs, developing a process mining framework using deep autoencoder for automatic detection and rectification of anomalies is the challenging problem considered.

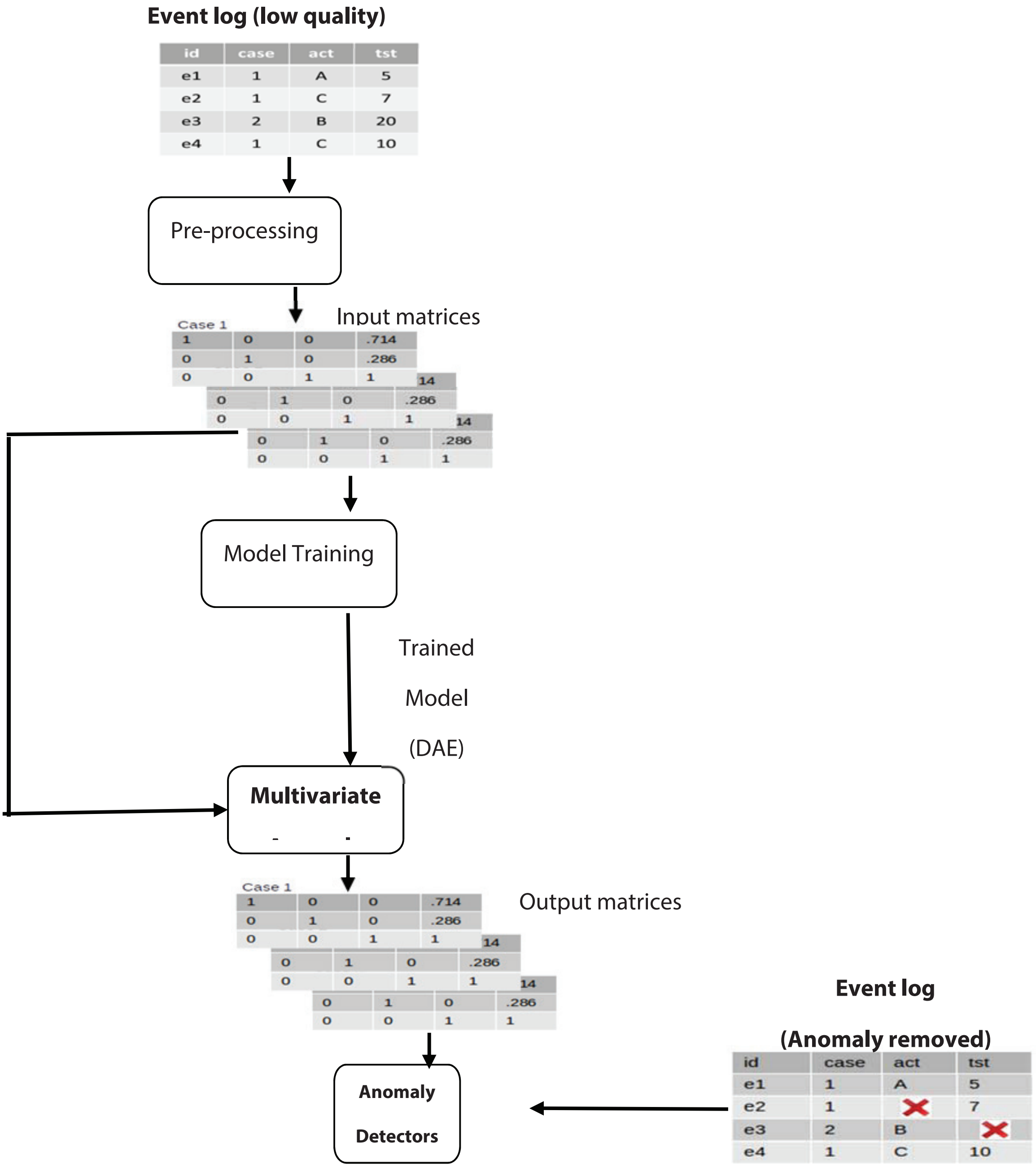
3.2. OUR FRAMEWORK

We proposed a deep learning-based process mining framework known as PMiner for the automatic detection of anomalies in business processes. Since there are thousands of business processes in real-time applications such as e-commerce, in the presence of concurrency, they are prone to exhibit anomalies. PMiner with its underlying mechanisms helps in the detection of anomalies and solves them automatically. PMiner is illustrated in terms of its anomaly detection in Fig. 1 and the reconstruction process in Fig. 2. We used the BPI

Challenge 2020 dataset collected from [32]. This dataset provides real-life event logs for research. However, the data was anonymized to preserve privacy. This section, later, illustrates an excerpt from the dataset while discussing the proposed methodology. Notations used in the proposed system are provided in Table 1.

Notation	Meaning
g_φ	Denotes encoder
f_θ	Denotes decoder
x^i	Original input
$f_\theta(g_\varphi(x^i))$	Reconstructed input
$P(x)$	Probability of input x
$P(x\|y)$	Denotes conditional probability
$P(y\|x)$	Denotes posterior probability
$P(y)$	Denotes prior probability
$P(x\|y)/P(y)$	Denotes likelihood ratio

Table 1. Notations used in the proposed system



Event log (low quality)

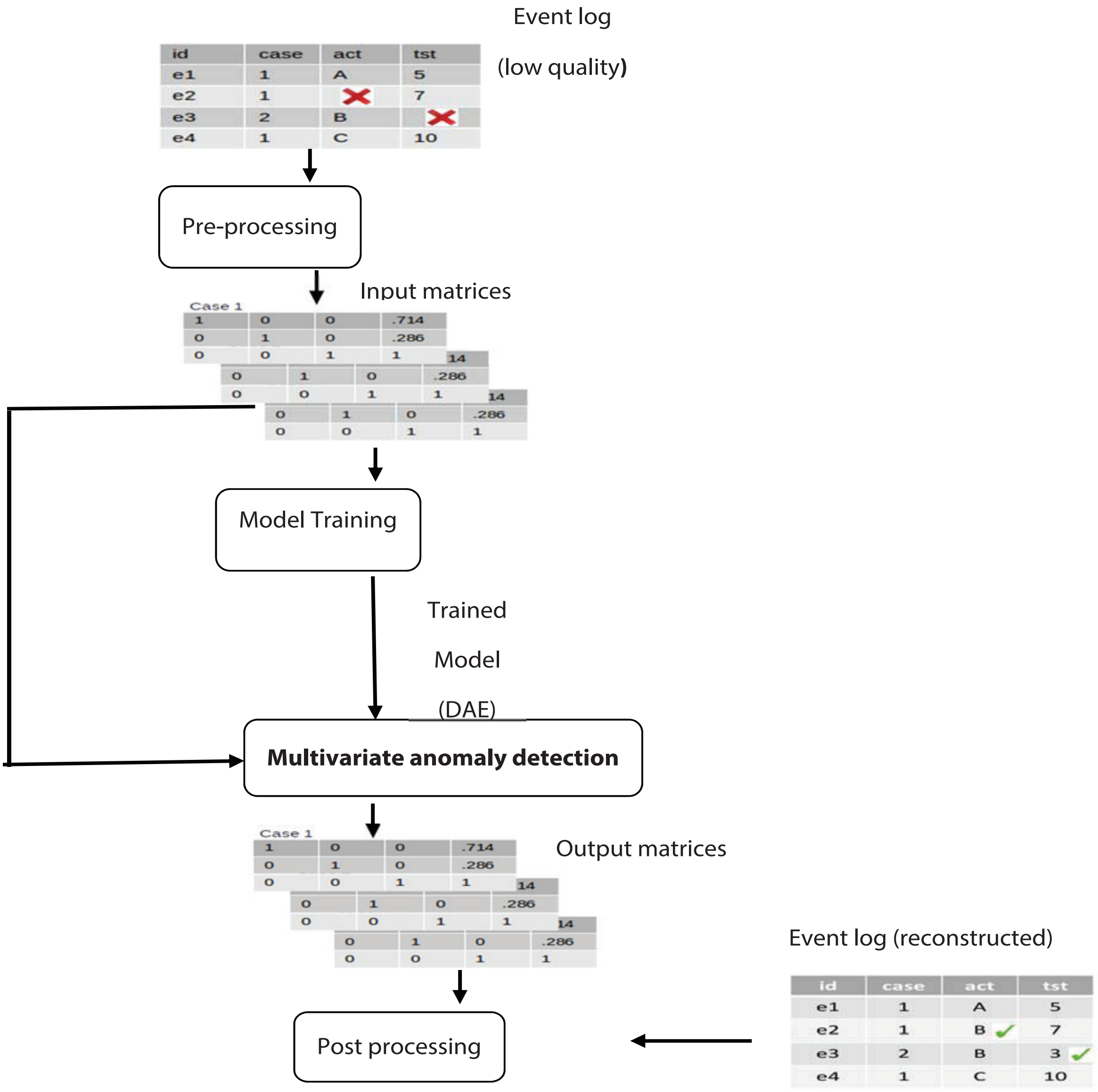
id	case	act	tst
e1	1	A	5
e2	1	C	7
e3	2	B	20
e4	1	C	10

Event log

(Anomaly removed)

id	case	act	tst
e1	1	A	5
e2	1	✗	7
e3	2	B	✗
e4	1	C	10

Fig. 1. PMiner framework reflecting process anomaly detection process



id	case	act	tst
e1	1	A	5
e2	1	✗	7
e3	2	B	✗
e4	1	C	10

id	case	act	tst
e1	1	A	5
e2	1	B ✓	7
e3	2	B	3 ✓
e4	1	C	10

Fig. 2. PMiner framework reflecting process anomaly rectification process

PMiner takes event log data as input. The event log is a text file containing log entries reflecting a set of cases represented as L ∈. Each case contains several events and attributes. The presence of a value and absence of value for a given attribute are denoted as $\#_a(c)$ and $\#_a(c) = \perp$ respectively. A sequence of events in the given trace or case is denoted as $\#_{trace}(c) \in \varepsilon^*$.

An event in the log entry is an activity involved in a process. In a given case there are several events denoted as e ∈ε. The activity attribute associated with data is $ddetedas\#_{act.}(e) \in A$. Similarly, $\#_{time}(e) \in T$ denotes the $ttimestampattributeOther$ attributes such as cost, resource and transaction are denoted as $\#cost(e)$, $\#_{resourse}(e)$ and $\#transe(e)$ respectively. As shown in Fig. 1, the given dataset is subjected to pre-processing. Table 2 shows an excerpt from the event log.

The data presented in Table 1 is subjected to attribute standardization where event ID and case attributes contain the identity of the event and case respectively.

The rest of the two columns do have discrete and continuous values. The normalization process has resulted in Table 3.

Id	Case	Act	Test
e1	1	A	5
e2	1	B	7
e3	2	B	3
e4	1	C	10

Table 2. An excerpt from the event log dataset

Id	Case	C_A	C_B	C_C	C_{tst}
e1	1	1	0	0	-0.42
e2	1	0	1	0	0.25
e3	2	0	1	0	-1.09
e4	1	0	0	1	1.26

Table 2. An excerpt from the event log dataset

After completion of processing, input matrices are generated. These matrices are used to train deep autoencoders as part of the encoding process. In the decoding process, the deep autoencoder generates output matrices. These outputs enable the framework to derive two kinds of anomaly detectors. They are generated based on activity and time. The selection criterion for these two is that the anomaly is generally based on inconsistency in activity or time in which events occur. This is the rationale for generating those two types of anomaly detectors. Detection of these two kinds of anomalies is very important for owners of businesses that make use of an enterprise application that relies on several business processes. These anomaly detectors are used by the framework to detect anomalies and remove them as illustrated in Fig. 1.

The anomaly rectification process of PMiner takes the output of the process illustrated in Figure 1. This output containing log entries with events where anomalies are removed is subjected to pre-processing. As in the anomaly detection phase, pre-processing generates input matrices and a deep autoencoder model is trained with those matrices. Then the trained model is used to generate output matrices that help in the reconstruction of log entries in the form of post-processing. Fig. 3 shows the learning process resulting in labelling through reconstruction error and finally detecting anomalies.

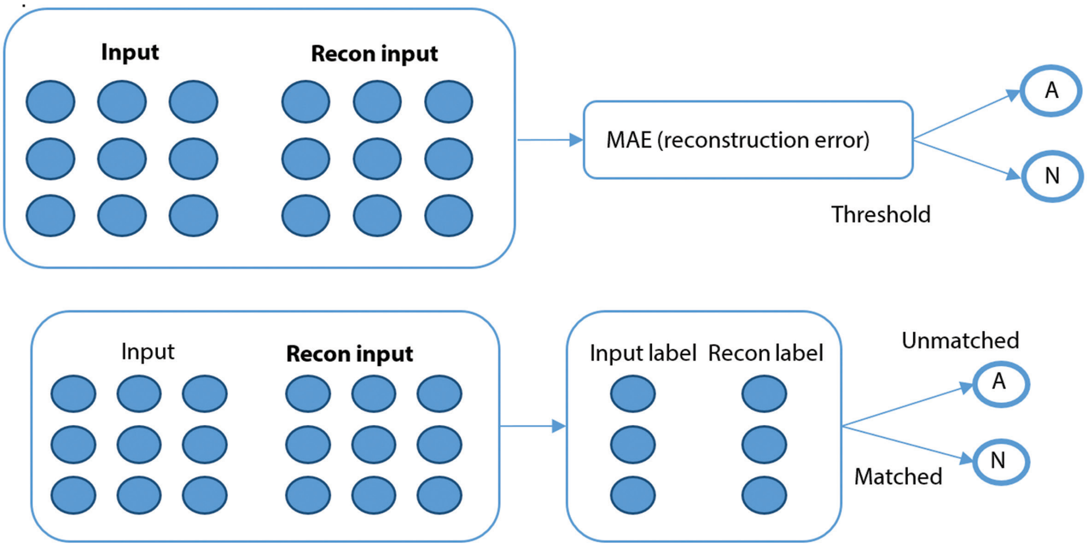


Fig. 3. Outlines the learning process involved in PMiner

In each phase of PMiner, there are deep encoding and decoding procedures involved as illustrated in Figure 4.

The deep autoencoder maps inputs to a distribution, in terms of two vectors such as mean and standard deviation, instead of fixed vector.

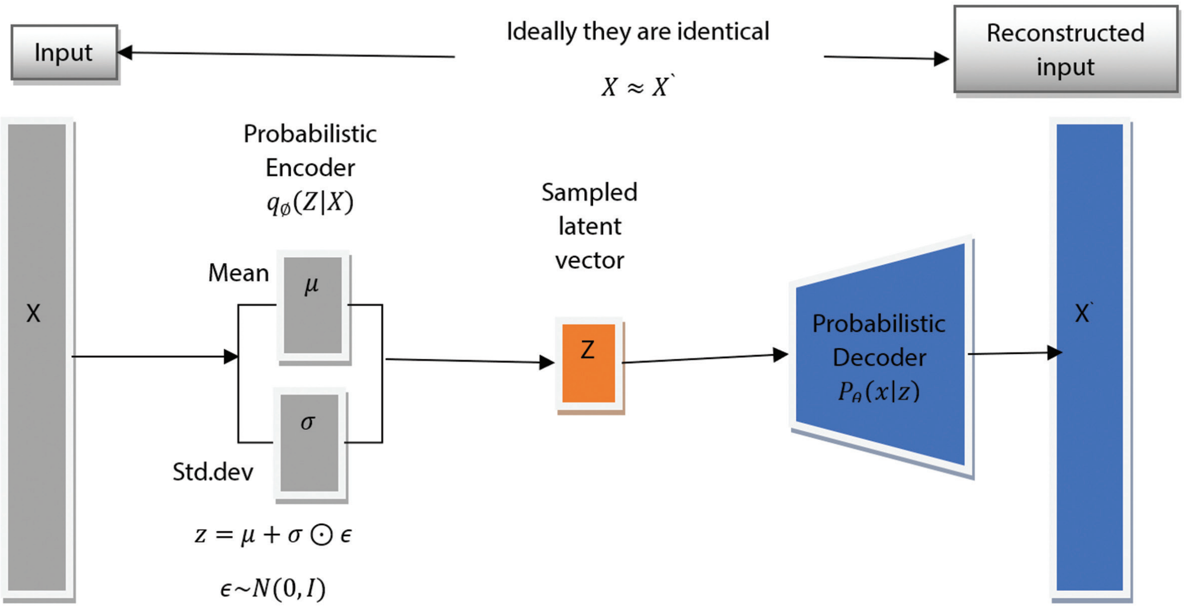


Fig. 4. Deep autoencoder used in the PMiner framework

The encoder and decoder functionalities in the autoencoder help in realizing anomalies in the business processes. The input X is mapped to mean vector μ and standard deviation vector σ. The encoding process results in the compressed nature of sampled latent vector z. The loss function associated with the autoencoder has two terms such as reconstruction loss and regularizer as expressed in Eq. 1.

$$L(\theta,\varphi)=-E_{Z\sim q_\theta}\,[P_\theta\,(x|z)]+D_{KL}\,(q_\varphi\,(z|x)//p_\theta\,(z)) \qquad (1)$$

The autoencoder functions based on probability theory. Given a random variable x, its probability is defined as $P(x)$ and its conditional probability is denoted as $P(x|y)$. Therefore, the probability theory can be expressed as in Eq. 2.

$$P(y|x)=\frac{P(x|y)P(y)}{P(x)} \qquad (2)$$

This theory is based on the well-known Baye's theorem where the likelihood ratio is denoted as $p(x|y)$ / $p(x)$, prior probability is denoted as p(y) while posterior probability is denoted as $P(y|x)$. Then theorem of total probability is expressed as in Eq. 3.

$$P(x)=\sum_{i-1}^{n}P(x|y)P(y) \qquad (3)$$

Given input variable x, the expected value associated with the random variable is weighted as per the probability of the event. Therefore, $E(x)$ of a random variable is computed as in Eq. 4.

$$E(x)=\sum x_i p(x=x_i) \qquad (4)$$

3.3. ALGORITHM DESIGN

We proposed an algorithm known as Intelligent Business Process Anomaly Detector (IBPAD) to realize the framework. This algorithm learns from historical data and performs encoding and decoding procedures to detect business process anomalies automatically.

```
Algorithm 3: Intelligent Business Process Anomaly
Detector (IBPAD)
Input: Event logs $L=\{e1, e2, \dots en\}$ for training
Output: $L_{(x,\hat{x})}$//reconstruction error
        $\varphi, \theta \leftarrow$ network parameter initialization
repeat
        $X^M \leftarrow$ obtain random points containing data
        points
$\epsilon \leftarrow$ nnoisebasedrandom samples $p(\epsilon)$
; g $\leftarrow \nabla_{\theta,\varphi}\, \tilde{L}^M\, )(\theta,\varphi; X^M,\epsilon)$//gradients
        $\varphi, \theta \leftarrow$ parameter update
until parameter convergence $(\varphi, \theta)$
$\varphi, \theta \leftarrow$ trained parameters
$\alpha \leftarrow$ threshold as per training data
repeat
        for i=1 to N do
Compute $L(x,\hat{x})$
            $L_{(\varphi,\theta;xi)}=\sum_i \|x_i\text{-}g_\theta\,(f_\varphi(x_i))\|^2$
if $L(x,\hat{x})> \alpha$ then
$x_i$ is considered anomaly
else
$x_i$ has no anomaly
end if
end for
```

Algorithm 1. Intelligent Business Process Anomaly Detector (IBPAD)

Algorithm 1 takes event log entries as input and detects anomalies through deep autoencoder based approach. It has training process where the algorithm gains knowledge which is then used in the anomaly detection process. Provided $L=\{e_1, e_2, \dots e_n\}$ as input, the algorithm eventually results in $L(x,\hat{x})$. Since event logs reflect activities of a business process that occur in temporal order, the proposed methodology and underlying algorithm learn from the huge data associated with business processes and finds anomalies. Once anomalies are detected, it is possible to rectify them from the knowledge gained in the process of detecting abnormality. The proposed system considers two kinds of anomalies such as time related and also activity related anomalies.

4. EXPERIMENTAL RESULTS

We implemented the proposed framework PMiner using Python language and process mining library. Anaconda distribution is used for building prototype. Environment used for the implementation is a PC with i3-1215U processor, 8GB RAM and Windows 11 operating system. BPI challenge 2020 dataset [32] is used in our empirical study. The dataset is freely available for usage by researchers. This section presents experimental results along with performance evaluation.

4.1. EXPLORATORY DATA ANALYSIS

This section presents data distribution dynamics in the data collected from [32]. The data is analysed in terms of anomalous data and normal data.

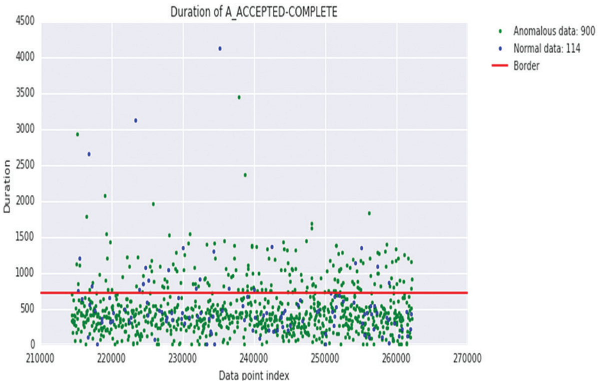


Fig. 5. Data distribution dynamics of A_ACCEPTED-COMPLETE attribute

As presented in Fig. 5, data point index against duration are visualized reflecting number of normal data points (114) and number of anomalous data points (900) distributed in the dataset.

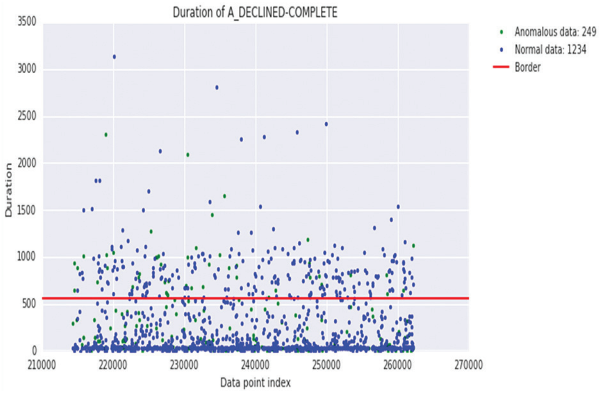


Fig. 6. Data distribution dynamics of A_DECLINED-COMPLETE attribute

As presented in Fig. 6, the data point indexes against duration are visualized reflecting number of normal data points (1234) and number of anomalous data points (249) distributed in the dataset.

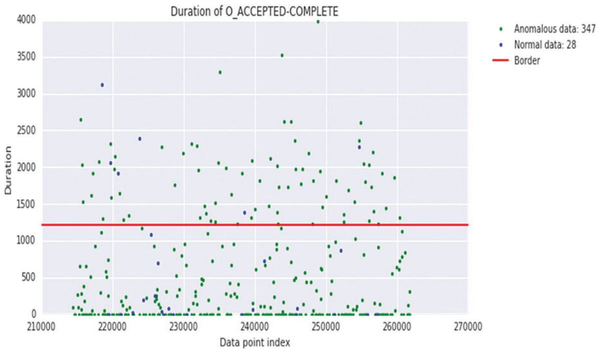


Fig. 7. Data distribution dynamics of O_DECLINED-COMPLETE attribute

As presented in Fig. 7, data point index against duration are visualized reflecting number of normal data points (28) and number of anomalous data points (347) distributed in the dataset.

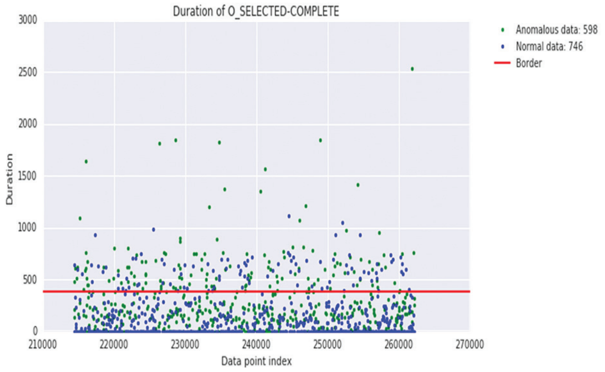


Fig. 8. Data distribution dynamics of O_SELECTED-COMPLETE attribute

As presented in Fig. 8, data point index against duration are visualized reflecting number of normal data points (746) and number of anomalous data points (598) distributed in the dataset.

4.2. TIME BASED ANOMALY DETECTION

This section presents time based anomaly detection results using the proposed PMiner framework. It covers reconstruction error, confusion matrix and AUC.

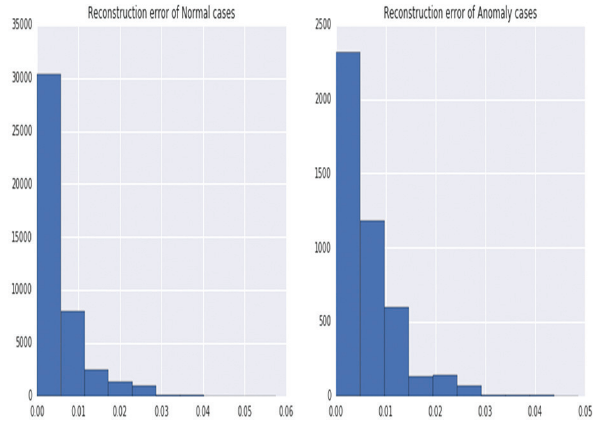


Fig. 9. Reconstruction error for normal and anomaly classes pertaining to time based anomalies

As presented in Fig. 9, it shows reconstruction error for normal class and also anomaly class. The proposed methodology has tested the entire dataset and the confusion matrix reflecting its detection process is presented in Fig. 10.

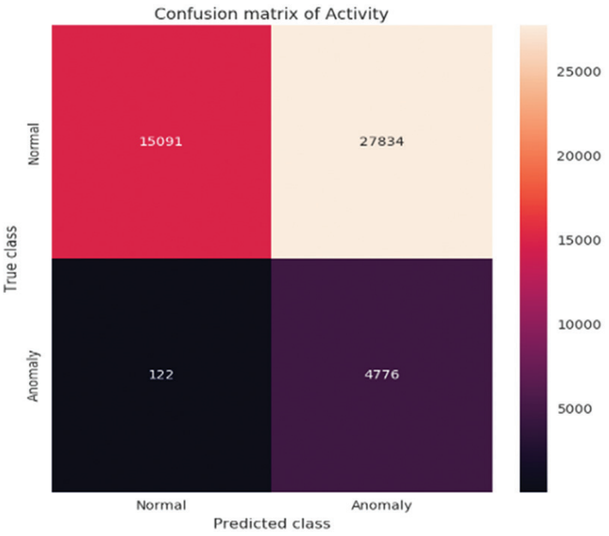


Fig. 10. Confusion matrix for time based anomaly detection

The confusion matrix visualizes the summary of results containing ground truth and also prediction results. It shows 4776 true positives, 15091 true negatives, 122 false positives and 27834 false negatives. Fig. 11 shows the AUC curve reflecting the performance of the proposed system.

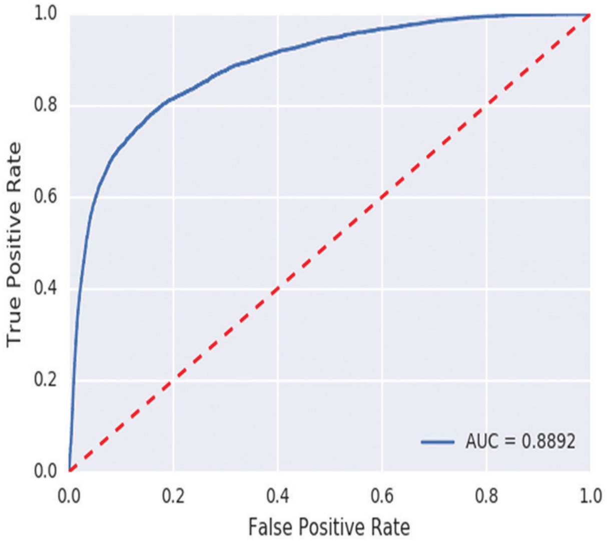


Fig. 11. AUC performance of the proposed system for time based anomaly detection

Area Under Curve (AUC) measure is used to assess the performance of the proposed system. AUC curve is computed as in Eq. 5.

$$\mathrm{AUC} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{6}$$

AUC of the proposed system for time based anomaly detection is 0.8892. Higher in AUC indicates better performance.

4.3. ACTIVITY BASED ANOMALY DETECTION

This section presents activity-based anomaly detection results using the proposed PMiner framework. It covers reconstruction error, confusion matrix and AUC.

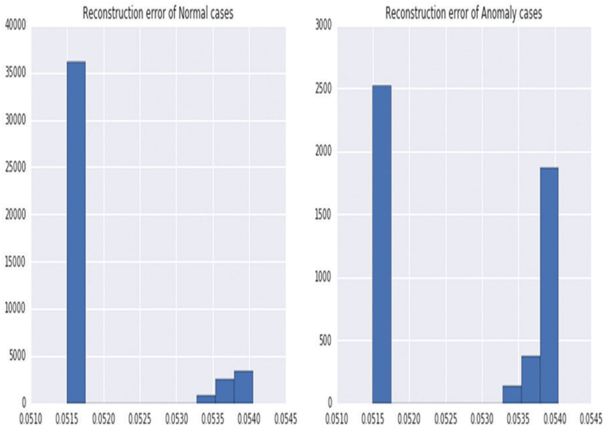


Fig. 12. Reconstruction error for normal and anomaly classes pertaining to activity based anomalies

As presented in Fig. 12, it shows reconstruction error for normal class and also anomaly class. The proposed methodology has tested the entire dataset and the confusion matrix reflecting its detection process is presented in Fig. 13.

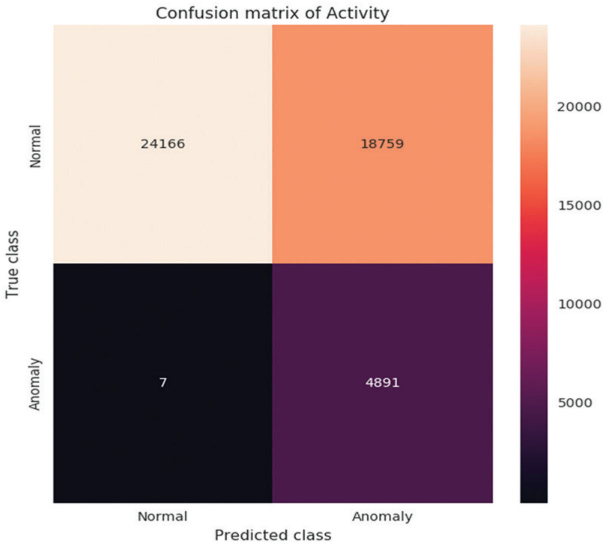


Fig. 13. Confusion matrix for activity based anomaly detection

The confusion matrix visualizes the summary of results containing ground truth and also prediction results. It shows 4891 true positives, 24166 true negatives, 7 false positives and 18759 false negatives. Fig. 14 shows AUC curve reflecting the performance of the proposed system.

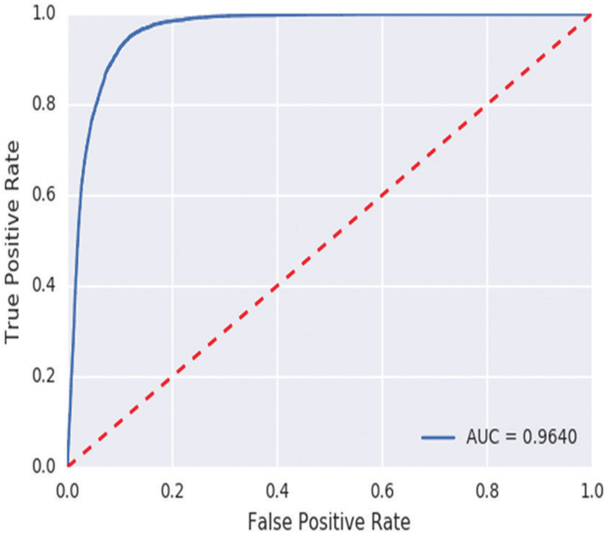


Fig. 14. AUC performance of the proposed system for activity-based anomaly detection

Area Under Curve (AUC) measure is used to assess the performance of the proposed system. With activity-based anomaly detection, the AUC of the proposed model is 0.9640. The activity-based anomaly detection performance is found to be better than that of time based anomaly detection.

4.4. PERFORMANCE COMPARISON

The proposed model is compared against simple autoencoder that does not make use of probability theory.

Table 4. Shows performance comparison among models

Anomaly Detection Model	Precision	Recall	F-Measure
Simple Autoencoder	0.9	0.76	0.824096
Proposed (DAE)	0.95	0.89	0.919022

As presented in Table 4, the performance of the proposed model is compared against simple autoencoder with the proposed framework.

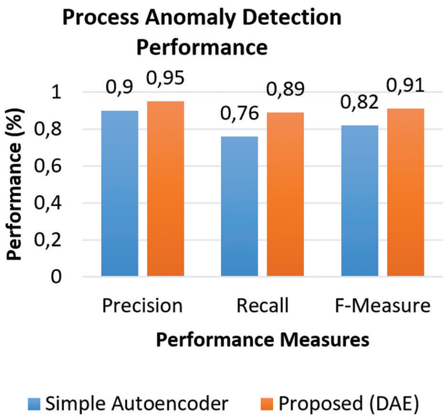


Fig 15. Performance comparison of process anomaly detection

As presented in Fig. 15, the performance of the proposed framework PMiner is compared against deep autoencoder (proposed) and simple autoencoder. It is observed that PMiner is capable of detecting anomalies and rectifying them. However, it could work better with the proposed deep autoencoder which is based on probabilistic theory. The precision achieved by a simple autoencoder with PMiner framework is 90%, recall 76% and F1-Score 82%. The PMiner framework with deep autoencoder could achieve 95% precision, 89% recall and 91% F1-Score. Therefore, the proposed PMiner framework along with the proposed algorithm based on deep autoencoder achieved the highest performance in process anomaly detection and rectification.

5. DISCUSSION

This section discusses important questions like why process mining? how does the proposed method achieve process anomaly detection and rectification? and what is the implication of this research for future endeavours? Enterprise applications in the real world, in the contemporary era, are running businesses through distributed applications. Such applications have several thousands of business processes. Due to the high complexity of the business processes and the concurrency nature of the processes in multi-user environments, there is ever possibility of anomalies in the execution of processes. Such execution dynamics are generally saved into log files known as process logs. If there is an anomaly which is not detected can lead to potential errors in the application. This, in turn, leads to a deterioration of customer satisfaction besides attracting legal issues. Therefore, it is indispensable to monitor process log entries to detect any sort of anomalies and rectify them. Therefore, process mining plays an important role in improving business process consistency. The proposed framework named PMiner in this paper is very useful for this purpose as it can automatically detect business process anomalies and rectify them. The research in this paper has implications that lead to further research endeavours in future.

5.1. LIMITATIONS

Though the proposed framework is capable of detecting and rectifying business process anomalies, it has several limitations. First, it is evaluated with the BPI Challenge 2020 dataset. Though this dataset is close to real-time processes in businesses, the proposed framework has not yet been evaluated by deploying in real enterprise premises with live data. Second, the dataset used for evaluation is relatively smaller in size (7.20 MB) and belongs to a specific domain. Therefore, it is important to evaluate our framework further with data from multiple domains and also with large data. Third, business process log entries grow dynamically. Therefore, it is desired to consider big data environment and computing frameworks to deal with streaming data.These limitations can be overcome by using live streaming of process event logs of enterprises, increasing the data for implicit training of autoencoder and usage of MapReduce kind of parallel processing framework.

6. CONCLUSION AND FUTURE WORK

A process mining framework known as ***PMiner*** is proposed for automatic detection of anomalies in business processes. The framework is designed to take real life business process event logs as input and detect anomalies using a deep autoencoder as it has potential to discriminate anomalies. An algorithm named IBPAD is proposed to realize the framework. This algorithm is able to process business process event logs with the proposed deep autoencoder, detect anomalies and rectify the same. BPI Challenge dataset released by IEEE Task Force on Process Mining is used for the empirical study. The proposed algorithm could achieve highest F1-Score 91% outperforming its existing autoencoder counterpart. In future, we intend to improve our framework to evaluate it with real enterprise application's live streaming business process event logs.

7. REFERENCES

[1] S. Riyanarto, S. Fernandes, S. K. Rossa, "Anomaly detection in business processes using process mining and fuzzy association rule learning", Journal of Big Data, Vol. 7, No. 1, 2020, pp. 1-19.

[2] H. K. Muzzammil, K. Yevgeniya, G. M. Medhat, "A Graph-Based Approach to Interpreting Recurrent

Neural Networks in Process Mining", IEEE Access, Vol. 8, 2020, pp. 172923-172938.

[3] P. Zerbino, A. Stefanini, D. Aloini, "Process Science in Action: A Literature Review on Process Mining in Business Management", Technological Forecasting and Social Change, Vol. 172, 2021, pp. 1-20.

[4] R. Sarno, R. D. Dewandono, T. Ahmad, M. FaridNaufa, "Hybrid Association Rule Learning and Process Mining for Fraud Detection", IAENG International Journal of Computer Science, Vol. 42, No. 2, 2015, pp. 1-14.

[5] N. Timo, L. Stefan, S. Alexander, M. Max, "Analyzing business process anomalies using autoencoders", Machine Learning, Vol. 107, 2018, pp.1-19.

[6] K. Wolfgang, M. Jonas, R. Maximilian, S. Johannes, "Machine Learning in Business Process Monitoring: A Comparison of Deep Learning and Classical Approaches Used for Outcome Prediction", Business & Information Systems Engineering, Vol. 63, 2020, pp. 261-276.

[7] D. A. Neu, J. Lahann, P. Fettke, "A systematic literature review on state-of-the-art deep learning methods for process prediction", Artificial Intelligence Review, Vol. 55, 2021, pp. 801-827.

[8] M. L. Fernandez, G. A. L. Perales, C. F. J. Garcia, P. M. Gil, P. G. Martinez, "A Self-Adaptive Deep Learning-Based System for Anomaly Detection in 5G Networks", IEEE Access, Vol. 6, 2018, pp. 7700-7712.

[9] E. A. Elaziz, R. Fathalla, M. Shaheen, "Deep reinforcement learning for data-efcient weakly supervised business process anomaly detection", Journal of Big Data, Vol. 10, 2023, p. 33.

[10] G. Abhijit, S. Debabrata, "Hybrid Approach to Document Anomaly Detection: An Application to Facilitate RPA in Title Insurance", International Journal of Automation and Computing, Vol. 18, No. 1, 2020, pp. 1-18.

[11] B. Franczyk, "Semi-Supervised Anomaly Detection in Business Process Event Data using Self-Attention based Classification", International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, Vol. 192, 2021, pp. 39-48.

[12] N. Timo, L. Stefan, S. Alexander, M. Max, "BINet: Multi-perspective business process anomaly classification", Information Systems, Vol. 103, 2022, p. 101458.

[13] Y. Lu, Q. Chen, S. K. Poon, "A Deep Learning Approach for Repairing Missing Activity Labels in Event Logs for Process Mining", Information, Vol. 13, No. 5, 2022, p. 234.

[14] W. Mathias, M. Marco, W. Ingo, V. Jan, "act2vec, trace2vec, log2vec, and model2vec: Representation Learning for Business Processes", Proceedings of the International Conference on Business Process Management, Sydney, NSW, Australia, 9-14 September 2018, pp. 305-321.

[15] A. Guzzo, M. Joaristi, A. Rullo, E. Serra, "A multi-perspective approach for the analysis of complex business processes behaviour", Expert Systems with Applications, Vol. 177, 2021, pp. 1-13.

[16] M. Vijayakamal, D. Vasumathi, "Unsupervised Learning Methods for Anomaly Detection and Log Quality Improvement Using Process Event Log", International Journal of Advanced Science and Technology, Vol. 29, No. 1, 2020, pp. 1109-1125.

[17] F. M. Maggi, C. D. Francescomarino, M. Dumas, C. Ghidini, "Predictive Monitoring of Business Processes", Lecture Notes in Computer Science, Springer, 2014, pp. 457-472.

[18] Z. Tariq, D. Charles, S. McClean, I. McChesney, Pau, "Anomaly Detection for Service-Oriented Business Processes Using Conformance Analysis", Algorithms, Vol. 15, No. 8, 2022, pp. 1-25.

[19] C. Toon, C. Michelangelo, M. Donato, "Unsupervised Anomaly Detection in Noisy Business Process Event Logs Using Denoising Autoencoders", Proceedings of the International Conference on Discovery Science, Bari, Italy, 19-21 October 2016, pp. 442-456.

[20] P. Singh, M. S. Azari, F. Vitale, F. Flamm "Using log analytics and process mining to enable self healing in the Internet of Things", Environment Systems and Decisions, Vol. 42, 2022, pp. 234-250.

[21] H. Adrien, B. Remi, C. Isabelle, "A Process Mining Approach for Supporting IoT Predictive Security", Proceedings of the IEEE/IFIP Network Operations and Management Symposium, Budapest, Hungary, 20-24 April 2020, pp. 1-9.

[22] R. Eric, M. G. Jorge, S. Marcos, C. Daniel, "Process mining in healthcare: A literature review", Journal of Biomedical Informatics, Vol. 61, 2016, pp. 224-236.

[23] T. C. Nicoleta, "Process Mining on a Robotic Mechanism", Proceedings of the IEEE International Conference on Software Testing, Verification and Validation Workshops, Porto de Galinhas, Brazil, 12-16 April 2021, pp. 1-8.

[24] M. Jans, J. Martijn, V. Werf, N. Lybaert, K. Vanhoof, "A business process mining application for internal transaction fraud mitigation", Expert Systems with Applications, Vol. 38, No. 10, 2011, pp. 13351-13359.

[25] S. Pauwels, T. Calders, "An anomaly detection technique for business processes based on extended dynamic bayesian networks", Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, April 2019, pp. 494-501.

[26] G. Park, M. P. Wil, V. Aalst, "Action-oriented process mining: bridging the gap between insights and actions", Progress in Artificial Intelligence, 2022, pp. 1-22.

[27] T. Okuboa, H. Kaiya, "Efficient secure DevOps using process mining and Attack Defense Trees", Procedia Computer Science, Vol. 207, 2022, pp. 446-455.

[28] C. Fernández-Llatas, J.-M. Benedi, J. García-Gómez, V. Traver, "Process Mining for Individualized Behavior Modeling Using Wireless Tracking in Nursing Homes", Sensors, Vol. 13, No. 11, 2013, pp. 15434-15451.

[29] L. Sofiane, S. Radu, "Process mining-based approach for investigating malicious login events", Proceedings of the IEEE/IFIP Network Operations and Management Symposium, Budapest, Hungary, 20-24 April 2020, pp.1-5.

[30] S. A. Kumar, R. Kamra, U. Shrivastava, "Conformance Checking Techniques of Process Mining: A Survey", Recent Trends in Intensive Computing, 2021, pp. 335-341.

[31] F. Folino, L. Pontieri, "AI-Empowered Process Mining for Complex Application Scenarios: Survey and Discussion", Journal on Data Semantics, Vol. 10, 2021, pp. 77-106.

[32] BPI Challenge 2020 dataset, https://www.tf-pm.org/competitions-awards/bpi-challenge/2020 (accessed: 2023)

Comparative Predictive Analysis through Machine Learning in Solar Cooking Technology

Original Scientific Paper

Karankumar Chaudhari*
Department of Mechanical Engineering, G H Raisoni College of Engineering,
Nagpur, India
karan.chaudhari01@gmail.com

Pramod Walke
Department of Mechanical Engineering, G H Raisoni College of Engineering,
Nagpur, India
pramod.walke@raisoni.net

Sagar Shelare
Department of Mechanical Engineering, Priyadarshini College of Engineering, Nagpur, India
Centre for Research Impact and Outcomes, Chitkara University, Rajpura, Punjab, India
sagmech24@gmail.com

*Corresponding author

Abstract – *Renewable energy technology has helped solve global environmental issues in recent years. Solar cooking technology is a sustainable alternative to conventional cooking, particularly in regions with ample sunlight. Although there is a growing interest into solar cooking, however, there is a lack of comprehensive comparison research upon the machine learning models predictive accuracy. Prior studies frequently concentrate upon individual models or fail to conduct comprehensive comparative analyses, resulting in a knowledge deficit regarding the most effective predictive methodologies for solar cooking technology. This research article compares solar cooking with special types of cooking utensils used for indoor cooking by predictive analysis of different kinds of machine learning models. To achieve proper cooking, the temperature of both pan and pot is to be monitored constantly. For this, a machine learning (ML) system model was constructed for predicting pan and pot temperature as a response parameter. By leveraging datasets encompassing time duration of the cooking, mass flow rate of heat transfer fluid, type of heat transfer fluid, and global solar radiations, a range of machine learning algorithms, including decision tree regressor, linear regression, extreme gradient boosting, and random forest regressor algorithms, are employed for predicting pan and pot temperature of solar cookers. Extreme gradient boosting is the best machine learning model for solar utensil temperature, with maximum R2 and minimum mean squared error, mean absolute error, and root mean squared error values that perfectly predict all answers. Also, extreme gradient boosting predicts well on training and testing datasets, whereas Random forest predicts well on training datasets but poorly on test data, causing overfitting. This research shows that machine learning could revolutionize solar cooking technology, promising a future for renewable energy and sustainable living.*

Keywords: *Solar Cooking, Machine Learning, Regression Analysis, XGBoost, Statistical Analysis*

Received: February 12, 2024; Received in revised form: April 22, 2024; Accepted: April 23, 2024

1. INTRODUCTION

In recent years, the global quest for sustainable and eco-friendly practices has gained unprecedented momentum, prompting a critical reevaluation of conventional processes across various sectors. Clean, renewable energy is necessary to combat climate change, environmental degradation, and the depletion of fossil fuels. One of the domains that needs a paradigm shift is cooking. Traditional methods use environmentally harmful non-renewable energy. Solar energy in culinary applications overcomes environmental concerns connected with conventional fuel sources, reduces climate change, and supports global sustainable development. Modern cooking consumes considerable energy, adding to greenhouse gas emissions and resource depletion. This article examines the constraints of conventional cooking and the potential benefits of solar energy to demonstrate how sustainable energy solutions improve the culinary sector. Through an examination of existing research, technological advancements, and successful case studies, we will explore the multifaceted advantages of integrating solar energy into the cooking field. From reducing carbon footprints and minimizing reliance on finite energy sources to fostering community empowerment and technological innovation, the use of solar energy in cooking holds the promise of a more sustainable and socially responsible culinary future. So, the research has been conducted in the field of solar cooking.

To demonstrate recent technological developments and the present state of solar-based cooking technology, Aramesh et al. [1] provide a comprehensive assessment of current experimental and analytical economics research on solar cookers. With exemplary examples from India, a methodology for estimating the level of many incentives necessary to ensure the financial appeal of institutional solar cooking is described. In terms of cost to the government, an accelerated depreciation is demonstrated to be least expensive method for an incentivizing institutional solar cooking, followed by viability gap financing, interest subsidy, and investment tax credit in that order. Solar Dish Stirling Systems (SDSS) design requirements, thermal performance analysis, opto-geometrical parameters, techno-economic factors, and thermodynamic optimization are discussed. SDSS applications include hybridization and storage, solar power plants, solar cookery, water desalination, and micro co-generation. Solar cooking is a viable option since it is both economical and expandable. Arunachala et al. [2] give a survey of such cookers to unveil the cost-effective solar cooker concepts. Materials used in solar thermal storage include fatty acids paraffin and non-paraffin, hydrated salts as well as material that use the thermo-chemical processes, sensible heat energy. Ndukwu et al. [3] discuss the various exergy methodologies used for various solar systems such as solar still, hybrid solar water heating, solar dryers-heaters, solar cookery systems and solar space heating. Because greater temperatures are attained in a shorter period of time, parabolic solar cookers outperform conventional box solar cookers. Lentswe et al. [4] provide an in-depth evaluation of a thermal energy storage (TES) based parabolic solar cookers, which are sustainable cooking option for some underdeveloped nations. This study predicts pan and pot temperatures. For optimal solar cooking system operation, prior temperature information is helpful. ML algorithms are the most advanced prediction systems today. Many researchers utilise ML.

Qahwaji et al. [5] investigate the use of sunspot relationships and ML for autonomous short term of a prediction of solar flare. It uses ML to anticipate automated short-term solar flare retrieval and convert McIntosh categorization of each sunspot into a numerical representation for ML algorithms. Colak et al. [6] provide short term predictions of a big solar flares using automated hybrid computer system. A ML based system will analyze years of a sunspot and flare data to generate associations Ahmed et al.'s [7] work uses feature selection, ML, and advanced feature extraction to forecast solar flares. Flare prediction is more accurate than SMART MFs and ML. Bobra et al. [8] utilize a machine learning algorithm called Support Vector Machine (SVM) and data of four years from the Solar Dynamics Observatory's Magnetic and Helioseismic Imager. Researchers want to forecast X- and M-class solar outbursts. Voyant et al.'s [9] provided an overview of ML-based solar irradiation forecasting techniques. ML has recently advanced to the point that a wide range of solar prediction works have been produced. In the continental United States, seven sites, five climatic zones, and three sky conditions [10] are employed to evaluate hourly predicting performances of total 68 ML algorithm. In their evaluation of several ML regression algorithms, Cornejo-Bueno et al. [11] tackle the topic of estimating worldwide solar radiation using data from geostationary satellites. Viscondi et al. [12] present a literature review utilizing big data model to forecast generation of solar photovoltaic electricity. The review considers the data used to solve the problem and each project proposal. Artificial Neural Network (ANN), support vector machine (SVM), deep learning (DL), and K-nearest neighbor (KNN) are the four ML methods used in the study. The analysis found that the ANN algorithm fits best. However, all examined algorithms can reliably anticipate daily global solar radiation statistics. Mahmood et al. [13] describe the fundamentals of ML and standard operating procedures. Additionally, the author has made several recommendations that may improve ML's value for companies researching organic solar cells.

The use of ML in solar engineering is from last decades. Most of the researchers has used ML in prediction of solar radiation, solar power, and application of solar energy. The table 1 shows the authors with their applied ML algorithm and evaluation metrics. Umit et al. [14] has forecasted daily global sun radiation using the KNN, SVM, DL, and ANN ML algorithms. The author finds R^2 values between 0.855 and 0.936 for all four techniques. Cetina et al. [15] has applied ANN, SVM, and linear regression (LR) used to predict daily solar global radiations. Author assessment measures include R^2, root mean square error (RMSE), mean average error (MAE), and mean square error (MSE). Linear regression (LR) ML's maximum R^2 is 0.9917. Tagnamas et al [16] has predicted the two parameters such as atmosphere temperature and thickness of beetroot using catboost ML algorithm. Author has used R^2, RMSE, MSE, and MAE for the evaluation of ML algorithm purpose. The author gets R^2 value of 0.9999 for this algorithms. Ledmaoui et al. [17] has applied total six algorithms i.e. ANN, Support Vector Regression (SVR), Decision Tree (DT), Generalized Additive Model (GAM) Random Forest (RF), and Extreme Gradient Boosting (XGBOOST) to predict the electricity production of solar energy. The R2, RMSE, MSE and MAE are the evaluation metrics considered by the author. The maximum R^2 value for the XGB ML algorithm is 0.99. Elgendi et al. [18] has predicted the yield of solar still using ANN and LR ML algorithm.

The experiment has conducted on the pyramid solar still. Author has used R^2, RMSE, and MAE for the evaluation of ML algorithm purpose. The author gets R^2 value of 0.956 for ANN algorithms.

Kameni et al. [19] has used six algorithms i.e. LR, DT, SVM, DL, RF and Gradient Boosted Trees (GBT) to predict global solar radiation. The maximum R^2 value for the GBT ML algorithm is 0.985. Oh et al. [20] has predicted the diffuse and direct solar radiation using XGB, Light Gradient Boosting Machine (LGBM), Kier and ANN ML algorithm. Author has used R^2, RMSE, and MAE for the evaluation of ML algorithm purpose. The author gets R^2 value of 0.955 for Kier algorithms.

Table 1. Summary of machine learning applications in solar energy

Authors	Parameters	Response	ML Algorithm	R2	Evaluation Parameter
Ağbulut et al. [14]	daily maximum and minimum ambient temperature, daily extraterrestrial solar radiation, cloud cover, solar radiation and day length	daily global solar radiation	SVM, ANN, KNN and DL	0.855 to 0.936	R^2, rRMSE, RMSE, MABE, MAPE, and MBE
Cetina et al. [15]	solar irradiance, Solar dryer type, ambient temperature, relative humidity, and wind velocity	daily global solar radiation	ANN, SVM, LR	0.9917	R^2, MSE, MAE, RMSE
Tagnamas et al. [16]	absorber plate, drying chamber outlet air temperatures and solar collector outlet air	temperature and thickness of the beetroot slices	Catboost model	0.9999	R^2, MSE, MAE, RMSE
Ledmaoui et al. [17]	the irradiation, total energy, daily energy, and the temperature	solar energy production	ANN, SVR, RF, DT, XGB and GAM	0.99	R^2, MSE, MAE, RMSE
Elgendi et al. [18]	the atmosphere temperature, relative humidity, air velocity	the yield of solar still	ANN and LR	0.956	MAE, R^2, and RMSE
Kameni et al. [19]	wind speed (va), daily air temperature(ta), solar radiation, and relative humidity(rh)	global solar radiation	LM, DT, SVM, DL, RF, AND GBT	0.985	ARE,AAE,RMSE, and R^2
Oh et al. [20]	Relative humidity, Dry-bulb temperature, Extraterrestrial irradiance, Solar azimuth angle, Solar zenith angle, Turbidity, Clearness index	direct and diffuse solar irradiance	XGB, LGBM, ANN, KIER	0.955	RMSE, MAE, and R^2
Khosravi et al. [21]	local time, pressure, temperature, relative humidity, and wind speed	hourly solar irradiance	MLFFNN, RBFNN, SVR, FIS, and ANFIS	0.9999	RMSE, MAE, and R^2
Muhammed et al. [22]	sunshine, temperature, meteorological parameters and day number	global horizontal solar irradiation	MLP, ANFIS and SVM	0.85	RMSE, MSE, and R^2
Alhamrouni et al. [23]	--	Day temperature and solar radiation	SVM LR, KNN, and RF	0.9948	--
Feng et al. [24]	air temperature	global solar radiation	ANN, MNEA, RF, AND WNN	0.885	RMSE, MSE, and R^2, RRMSE, MAE
Quej et al. [25]	daily minimum and maximum air temperature, rainfall, and extraterrestrial solar radiation	daily global solar radiation	ANN, ANFIS and SVM	0.737	MSE, RMSE, MAE and R^2
Citakoglu [26]	calendar month number (M), average air temperature (Tmean), extraterrestrial radiation (Ra), and average relative humidity (RHmean)	Solar radiation	RF, KNN, XGB	0.9436	MAE, RMSE, R^2,

Khosravi et al. [21] has used radial basis function neural network (RBFNN), multilayer feed forward neural network (MLFFNN), SVR, adaptive neuro-fuzzy inference system (ANFIS), and fuzzy inference system (FIS) for the prediction of hourly based solar radiation. The maximum R^2 value for the ANFIS ML algorithm is 0.9999. Muhammed et al. [22] has predicted the global horizontal solar irradiation using Multi-layer perceptron (MLP), ANFIS and SVM ML algorithm. The sunshine, temperature, meteorological parameters and day number were the parameters considered for the prediction purpose. The author gets R^2 value of 0.85 for SVM algorithms. Alhamrouni et al. [23] has used LR, RF, KNN and SVM for the prediction of solar radiation and temperature. R^2, RMSE, MSE and MAE are the evaluation metrics considered. The maximum R2 value for the SVM ML algorithm is 0.9948. Feng Yu et al. [24] to predict the global solar radiation using ANN, mind evolutionary algorithm (MNEA), RF, Wavelet neural network (WNN) ML algorithm. The author gets R^2 value of 0.885 for ANN algorithms. Victor et al. [25] has used ANFIS, ANN and SVM for the prediction of daily based global solar radiations. The maximum R^2 value for the SVM ML algorithm is 0.737. Citakoghu [26] has predicted the solar radiation using RF, KNN and XGB ML algorithm. The extraterrestrial radiation (Ra), calendar month number (M), average relative humidity (RHmean), and average air temperature (Tmean) were the parameters considered for the prediction purpose. Author has used RMSE, R^2 and MAE for the evaluation of ML algorithm purpose. The author gets R^2 value of 0.9436 for XGB algorithms.

Despite the tremendous improvement in renewable energy technologies, machine learning in solar cooking still needs to be explored. Although there is a growing interest into solar cooking, however there are lack of comprehensive comparison research upon the machine learning models predictive accuracy. Prior studies frequently concentrate upon individual models or fail to conduct comprehensive comparative analyses, resulting in a knowledge deficit regarding most effective predictive methodologies for solar cooking technology. Most of the researchers has used ML algorithm for prediction of solar radiation and solar dryers. It gives scope of the use of such algorithms for solar cooking also. Also, data-driven methods to analyze and enhance it are still being determined. By utilizing machine learning predictive analysis, anyone can bridge the gap and gain valuable insights through data-driven methods.

This study led to the development of split-type solar cooking. The experiment examined pan and pot cooking performance. The study found that sun intensity varies with time of day. Heat transfer fluid type and oil mass flow rate are most significant. The four ML methods investigated were linear regression (LR), decision tree (RF), random forest (RF), and extreme gradient boosting. These programmes predicted the future using available data. This research found that the extreme gradient boosting algorithm has the lowest mean square error, root mean square error, mean absolute error, and greatest R^2 value. It was suggested to use XGBoost for the project.

2. RESEARCH METHODOLOGY

The solar cooking system is a need of the future. The day to day development has been takes place in this system. In this research, the indirect solar-powered cooking system has been created. As per discussion in the introduction, different types of cooking system are available. But it has been observed that, the research on the cooking utensils was not conducted. In this research, special types of cooking utensils were developed which can cook the Indian food. The basic cooking utensils for the Indian food are pan and pot. So, the research was carried out to develop the cooking pan and pot for the indirect solar cooking system, which gives the comfort of cooking food inside house like cooking on LPG gas.

The testing of these utensils were conducted on the indirect solar cooking system. The indirect solar cooking system was developed as shown in Fig 1. The solar cooking system consist of parts like parabolic dish collector, solar receiver, pump, pipelines, pan and pot. In order to check the performance of system, the temperature indicators were installed. In this system, the solar energy was collected by the parabolic dish collector and transferred to the solar receiver. The receiver gets heated due to solar energy. The heat transferred fluid i.e. Therminol 55 and Soyabean oil was used to transfer heat from solar receiver to cooking utensils. Therminol 55 and soybean oil have optimal heat transfer characteristics. Therminol 55 is a high-quality heat transfer fluid due to its excellent thermal stability, lower viscosity, and higher thermal conductivity. Soybean oil is an ideal heat transfer fluid and having widespread availability, lower cost, and environmentally friendly with the good thermal properties. These heat transfer oils have working temperature range from 200°C to 2500°C, also has high specific heat. In order to transfer the fluid, 0.5 hp centrifugal pump was used. The heat absorbed from the solar receiver was transfer to the utensils and utensils were gets heated. The heated utensils (heat from the utensils) were used to cook the food. The pan was used to cook the Indian food like roti, chapatti, paratha, dosa etc., while pot was used to cook the Indian food like, dal, rice, curry etc.

The testing was conducted in the month of April 2023 at Nagpur, India. The testing was conducted from 9 am to 5 pm. The solar intensity were recorded The observations were recorded after equal interval of one hour. As discussed earlier two different heat transfer fluid were used i.e. Therminol 55 and Soyabean oil. The specific heat is the important parameter for the selection of these fluid. The examinations were conducted by varying the mass flow rate of fluid. The table 2 shows parameters used for the prediction system. The original dataset of 54 size used for the study and to predict the temperature of pan and pot. The maximum obtained temperature of pan is 1920°C for 5 hours of heating till 1 PM with a solar intensity of 635 w/m2 and mass flow rate of 12 lpm.

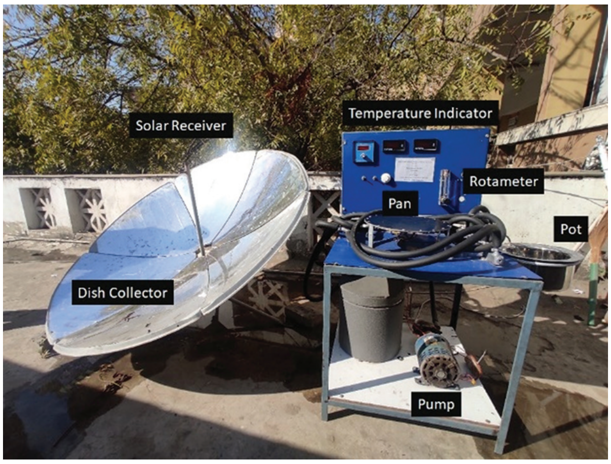


Fig.1. The Experimental Setup of Indirect Solar Cooking System with utensils

Table 2. Parameters used for the experimentation

	P1	P2	P3	P4	P5	P6	P7	P8	P9
Time (hr)	0	1	2	3	4	5	6	7	8
Solar Intensity (W/m2)	300	350	400	450	500	550	600	650	700
Mass Flow Rate (LPM)	6	9	12	-	-	-	-	-	-
HTF (kJ/KgK)	1.98	2.54	-	-	-	-	-	-	-

3. MACHINE LEARNING APPROACH

In this study, ML approach including LR, DT, RF and XGB were used to predict the temperature of cooking utensils i.e. pan and pot and select the suitable algorithm for accurate prediction. The suggested regression learning methodology utilized in the ML.

3.1. LINEAR REGRESSION

Two hypotheses make use of the linear regression approach [27]. Researchers first apply linear regression analyses in forecasting and prediction, where they closely resemble the use of ML. Linear regression analysis is a useful tool in some scenarios to ascertain the relationship between the dependent and independent variable. It is important to consider that regressions demonstrate relationships between dependent variables and a defined dataset comprising various factors.

Linear regression models [28] predict the dependent variables based on the independent variables. Linear regression analysis is used to estimate value of dependent variable, y, since independent variables, x, has a range of values [29, 30]. There are two categories of regression [31]: simple linear regression and polynomial linear regression. In this study, simple linear regression is used.

3.1.1. Simple Linear Regression

A simple linear regression is a model with only one independent variable [32]. Simple linear regression defines the variable's dependence as $y=\beta 0+\beta 1x+\varepsilon$. The effect of independent factors is distinguished from interaction of dependent variables by simple regression. Fourier multivariate regression (MLR) is statistical method that uses many explanatory factors to predict answer variable's outcome. Modeling the linear connection between independent variables x and dependent variable y that will be examined is the goal of MLR.

3.2. DECISION TREE REGRESSOR

Among many ML approaches is the decision tree. Although this approach is usually applied to classification data, it may also be applied to regression data. An approach that is transparent and simple to comprehend is the DT technique, as opposed to employing an artificial neural network as a black box.

The aim in this study is a continuous value since a decision tree technique is utilized for a regression problem. In order to minimize the impurity function and choose the best sites for future data splits, regression criteria such as mean squared error (MSE) and mean absolute error (MAE) may be used. The mean values for MSE can be used to minimize an error [33].

However, this approach has problems with stability, scalability, and robustness when it comes to large-scale data processing [34]. The utilization of extensive data samples leads to increased complexity, which must be addressed. To reduce the complexity of a decision tree, metrics such as total number of leaves, the total number of nodes, number of attributes, and tree depth can be adjusted [35]. Ensemble DT are utilized instead as they are more reliable and can handle these problems in some situations.

3.3. RANDOM FOREST REGRESSION

A supervised learning technique called Random Forest may be used to solve decision tree and classification issues. A "Random Forest" is a collection of numerous trees, where each tree depends on the value of a random vector, which is equally and independently sampled from each tree in the forest. [36]. By combining many decision trees, the random forest method may greatly improve the decision tree's predictive performance [37].

Two reasons contribute to the randomness of this algorithm:(1) each split node in the DT formation process selects sample chunk of m variables from the original data set, and the best one is used in that node; (2) every tree develops at random on a distinct bootstrap sample derived from the training set. A useful ML technique for prediction is Random Forest. The RFR Model is suggested by Harrison et al. [38] for nutrient concentration estimation utilising high-frequency sensor data. Since the method is suitable for multivariate datasets with multicollinearity among predictors, nonlinear correlations between predictor and response variables, and highly skewed data, it is well-suited for this application. The benefit of Random Forest Regression over least squares regression, according to the study in [39], is higher R squared (R^2) value.

3.4. EXTREME GRADIENT BOOSTING

Chen and Guestrin [40] developed the XGBoost method. Given its efficacy as a tree-based ensemble learning method, data scientists view it as a potent instrument. Based on gradient boosting architecture [41], XGBoost estimates the outcomes makes use of a variety of complement functions.

3.5. PERFORMANCE EVALUATION

Three performance statistical error functions, including the coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE), were taken into consideration in order to assess the performance of the ML models under examination. Generally speaking, five-fold cross-validation (CV) involves randomly dividing all of the data into k folds ($k = 5$ in this example), training the model on the k - 1 folds, and leaving one fold for testing. There are k repetitions of this process. But before any data is utilised in this study, it is divided into training and testing datasets, with the purpose of using the training dataset for cross-validation. The 95% confidence intervals (CI) and model accuracy are estimated using the repeated cross-validation procedure [55]. For every model, the 5-fold cross-validation is carried out 50 times. CV accuracy is the average of all repetitions, and 95% confidence intervals are computed from the repeated cross-validation data. The last step in assessing the model's performance is to determine if the testing accuracy falls within the 95% confidence interval. The model is deemed acceptable if the testing accuracy is within the 95% confidence interval. If the testing accuracy falls outside of this range and the difference is statistically significant, underfitting or overfitting is thought to be present. The cross-validation process does not use the testing data, which is a separate dataset.

4. RESULTS AND DISCUSSION

As previously stated, the goal of this research is to predict the temperature of the pan (Tpn) and the pot (Tpt) in a solar cooking system using four ML models: linear regression (LR), decision tree (DT), random forest (RF), and XGBoost (XGB). First, using the Scikit-Learn Python module, the following two PR-based meta-models are created based on the training dataset.

$$T_{pn} = 11.493 \times T + 0.301 \times E - 1.415 * mf - 8.396 \times htf - 25.162 \quad (1)$$

$$T_{pt} = 10.09 \times T + 0.265 \times E - 1.24 * mf - 7.192 \times htf - 19.527 \quad (2)$$

The first ML model applied for the data given in table 2 is Linear regression model. In this model, the sklearn library was used. In order to consider intercept for this model, fit intercept is considered as a True in nature. The normalize is kept deprecated to fit the model intercepted. For the faster computation, number of jobs is considered as a 2. The 80% of data i.e. 43 samples are used for the training purpose while 20% of data i.e. 11 samples are used for the testing purpose. A machine learning model was developed and tested for predicting the temperature of a pan and pot. The regression equation obtain from the model can be seen in equation 1 and 2 for the T_{pn} and T_{pt} respectively.

Decision Tree Machine Learning algorithm also applied on the given data. The decision tree is one of the advanced technique of the regression model. It is node based algorithm. In order to apply the algorithm on the data, the criteria for evaluation was considered as a "squared error", which helps to reduce the variance as a feature selection and minimize the $L2$ loss. The "best" strategy is considered for the split at node. The maximum depth of the tree is restricted to 10, in order to avoid overfitting of model. The minimum sample split is set default as 1. This all parameters are considered while developing decision tree algorithm. Here also, the data is divided as 80% for training and 20% for testing.

A machine learning algorithm has also been applied to the provided data. The decision tree is one of the advanced technique of the regression model. It is node based algorithm. In order to apply the algorithm on the data, the criteria for evaluation was considered as a "squared error", which helps to reduce the variance as a feature selection and minimize the $L2$ loss. The "best" strategy is considered for the split at node. The maximum depth of the tree is restricted to 10, in order to avoid overfitting of model. The minimum sample split is set default as 1. This all parameters are considered while developing decision tree algorithm. Here also, the data is divided as 80% for training and 20% for testing.

The random forest is a bagging techniques. The bagging techniques helps to improve the accuracy and overcome the problem of overfitting in the decision tree. The random forest regression model was applied on the data with spilt of 80:20 for training and testing. "n_estimator" is set to 100 with criterion "squared error" to run multiple decision trees in parallel and determine the final outcome. The depth of the tree was restricted to 10 with minimum sample split 2 and minimum sample leaf as 1. The maximum features considered as an "auto" means all available features are considered for the model. In order to avoid the overfitting, pruning takes place. All these parameters were considered for the development of random forest regression ML model.

The other method to improve the performance of decision tree algorithm is boosting techniques. In boosting techniques series approach was used. The output of one tree is used for the nest decision tree. One of such algorithm was used for testing of data. The XGBoost algorithm is one of the most advance boosting algorithm. For this algorithm, "n_estimator" considered as 100 while criterion as "squared error". The depth of the tree was restricted to 10 with minimum sample split 2 and minimum sample leaf as 1. The maximum features considered as an "auto" means all available features are considered for the model. The learning rate as 0.1 and "n_job" as a 10 in order to speed up the performance of the algorithm.

An effort is now made to estimate the values of T_{pn} and T_{pt} for the solar kitchenware once all created ML models have been properly trained using the dataset under consideration. For each of these ML models, Table 2 shows a predicted and target responses values. Plotting a predicted and target values for T_{pn} and T_{pt}, respectively, allows for a more clear understanding of the prediction performance of the ML models. These numbers show that within a ± 15% error band, all of the created ML models can accurately anticipate both of these answers.

As can be seen from Fig. 2, all of the ML models had almost excellent estimates for Tpn prediction on training data, with the majority of the data points either hugging or resting upon a diagonal identity line. But, in LR ML model, some training points are not on the line while three points are beyond the ±15% error band. But in case of RF only 3 points are found such that they are not on line but are in the ±15% error band. The other two ML model DT and XGB are tuning perfectly with the line and almost all the data points are on the line. This shows that the DT and XGB model has good generalisation and no overtraining. Both ML model shows identical performance.

When ML models predict Tpt, a trend comparable to that of Tpn is observed. The LR ML model, have quite similar prediction as seen as for Tpn. There are most of the data points are on the line and some are beyond the line. Out of some distracted data points only three data points are outside the error band which can be seen poor prediction towards the loser data points. The better prediction of Tpt can be seen for the RF model than the LR model. In RF no data points are beyond the error band. The DT and XGB shows the best performance model than LR and RF. Most of the data points of Tpt can be seen on the line.

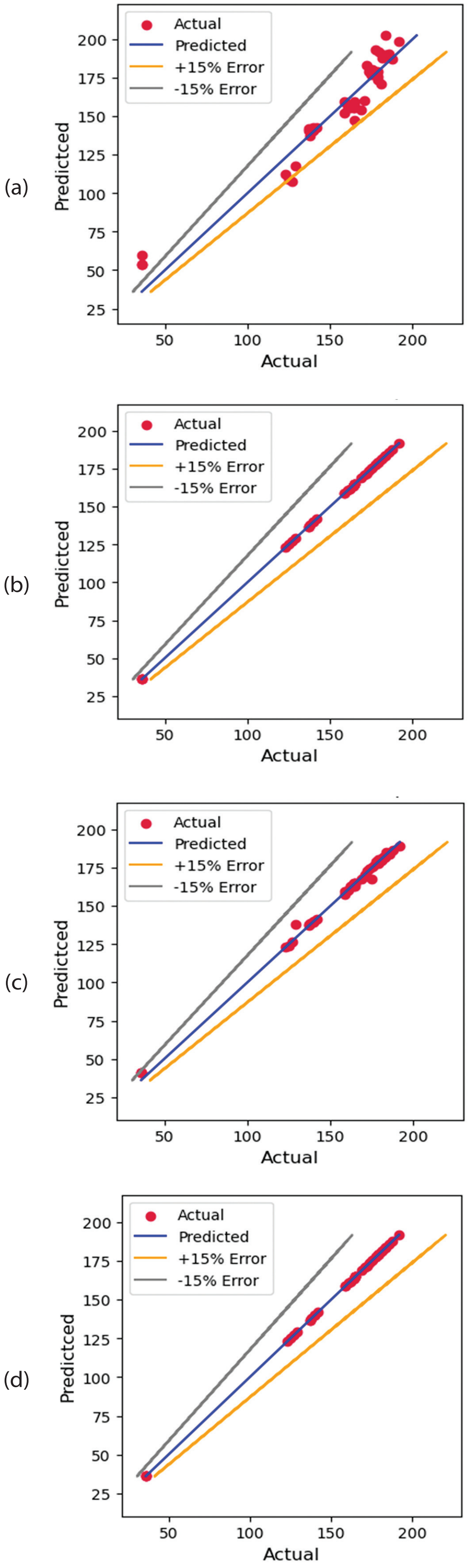


Fig. 2. Target vs. predicted T_{pn} values comparison for (a) LR, (b) DT, (c) RF and (d) XGB

Fig. 3 shows the residuals, or differences between the goal and anticipated response values, for each of the created ML models. The zero line in Fig. 3 denotes zero prediction error, while the points above and below it shows underprediction (i.e., predicted value less than target value) and overprediction (i.e., projected value more than target value), respectively. This is important to notice. All of the ML models' test data performances—aside from LR's—are generally comparable to their related data performances. This suggests that the instruction is sufficient. Additionally, there is no discernible pattern in the residuals' dispersion, which suggests that there is no bias.

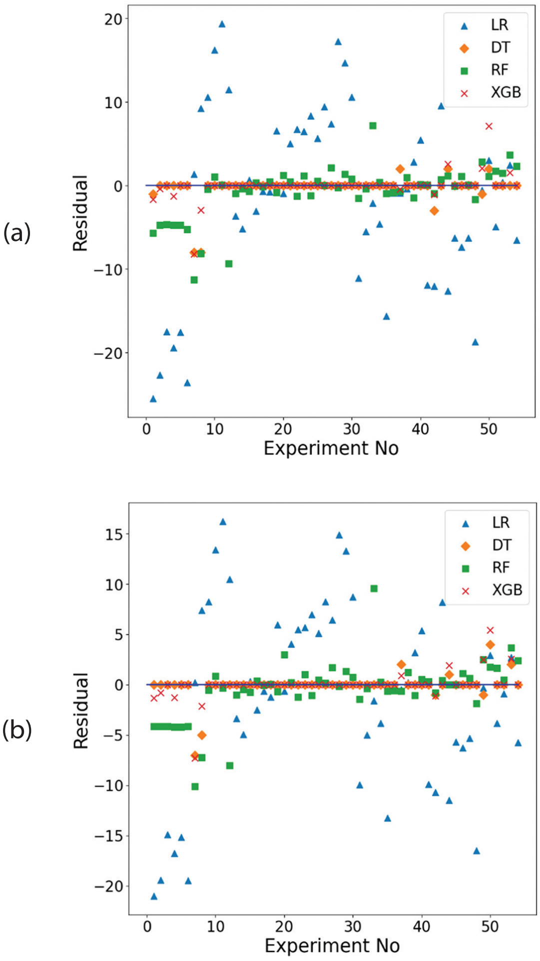


Fig 3. (a) Residuals of predicted T_{pn} and (b) residuals of predicted T_{pt}

Now that Figs. 2, and 3 have been closely examined, it is clear that the RF, DT, and XGB ML models all perform rather well in terms of T_{pn} prediction. The predicted T_{pn} values for each of these three ML models exhibit incredibly little deviations from the matching goal values. With minor variations of the projected Tpn values from the goal, the implementation of the LR ML model yields average prediction results. T_{pt} response is also observed in a similar manner. It is challenging to identify which of the produced ML models has the greatest prediction performance for the case under consideration simply by looking at the aforementioned numbers. To do this, the values of four model accuracy metrics—MSE, RMSE, MAE, and R^2—are calculated, as shown in Table 3. It is important to note that lower values of MSE, RMSE, and MAE and higher values of R^2 are always preferred for any of the predictive models [42].

Table 3. Metrics representing the accuracy of the models for both responses

Response	ML Model	Dataset	MSE	RMSE	MAE	R2
Tpn	LR	Testing	176.89	13.30	9.98	0.96
		Overall	117.13	10.82	8.51	0.94
	DT	Testing	13.00	3.61	2.82	1.00
		Overall	2.65	1.63	0.57	1.00
	RF	Testing	26.89	5.19	4.09	0.99
		Overall	10.10	3.18	1.99	1.00
	XGB	Testing	13.23	3.64	2.66	1.00
		Overall	2.69	1.64	0.54	1.00
Tpt	LR	Testing	128.82	11.35	8.49	0.96
		Overall	85.78	9.26	7.31	0.95
	DT	Testing	13.18	3.63	2.00	1.00
		Overall	2.69	1.63	0.41	1.00
	RF	Testing	20.90	4.57	3.56	0.99
		Overall	8.89	2.98	1.86	0.99
	XGB	Testing	10.04	3.17	2.48	1.00
		Overall	2.05	1.43	0.51	1.00

Table 3 shows that, when it comes to Tpn, the XGB ML model has the best R2 values, coming in at 0.9997 on the training dataset. Based on the training dataset, the MSE, RMSE, and MAE corresponding values of 1.9423, 1.3936, and 0.3718, respectively, further support the outstanding performance of the XGB ML model. However, DT has the highest R2 values at 0.9977 when taking into account the performance in relation to the test dataset. The minimal MSE, RMSE, and MAE values for the identical test data are 13, 3.605, and 2.4212, respectively. Based on the entire dataset (training and testing), Tpn prediction shows that XGB is the best ML model, with the highest R2 (0.9987) and lowest MSE (2.6945), RMSE (1.6415), and MAE (0.5424) values. Additionally, the LR ML model has the lowest R2 (0.9424) and the poorest MAE (8.5133), RMSE (10.8226), and MSE (117.13) values based on the entire dataset. In summary, the DT and RF ML models come in second and third place, respectively, when it comes to predicting Tpn values throughout the whole dataset.

Similar findings are also observed when estimating Tpt values for the solar cooking tool. On the training dataset with the maximum R2, minimum MSE, RMSE, and MAE values, and maximum R2 accuracy, the XGB performs best. However, XGB proves to be the most accurate ML model when it comes to predicting Tpt values using the test data, with the highest R2 (0.9967) and lowest MSE (10.04), RMSE (3.1688), and MAE (2.4817) values. In terms of all model accuracy measures, XGB performs the best across the board for the dataset, with DT and RF ML models following closely behind. Out of all four measures, the LR displays the poorest results for the whole dataset. As a result, it is seen that LR's performance is very variable for both Tpn and Tpt replies. It's interesting to note that while it performs well on training datasets, it performs poorly on testing and general datasets. On the other hand, for the two replies that are being examined, XGB consistently possesses an accuracy level over the whole dataset.

5. CONCLUSION

This work develops 4 ML models—linear regression, decision tree, random forest, and extreme gradient boosting to accurately predict solar cooking utensil temperatures. For the utensils temperature that is temperature of Pan, and temperature of pot; duration of time, solar intensity, type of heat transfer fluid, and mass flow rate of heat transfer fluid are treated as the input parameters. The prediction performance of the four developed ML models is compared in terms of four model accuracy metrics—R2, mean squared error, root mean squared error, and mean absolute error using these experimental datasets as a basis. Based on the comprehensive comparative analysis of the Machine Learning models' overall performance, the following conclusions can be drawn:

- In case of both the temperature of solar utensils, extreme gradient boosting emerges out as the best machine learning model with maximum R2 and a minimum value of mean squared error, mean absolute error and root mean squared error values perfectly predict all of the answers that are being considered.
- For prediction, extreme gradient boosting consistently yields good results on training and testing datasets.
- Moreover, while random forest performs exceptionally well in predictions on training datasets, its accuracy on test data is low, which causes the model to become over fit.
- Extreme gradient boosting has a wide range of tuning parameters, yet it may be argued that, for forecasting response values of the temperature of utensils under consideration, it is an effective prediction tool. Finding the ideal mix of those tuning parameters to get the highest prediction performance out of the extreme gradient boosting machine learning model is still a difficult challenge.

In the future, it may be possible to investigate the application potential of more machine learning models, such as the Adaboost regressor, support vector regressor, and Naïve Bias regressor, and compare how well they predict response values for utensil temperature. The demonstrative examples provided use tiny datasets, each with a mere 54 experimental observations, to validate the prediction performance of all the ML models. A data repository with a sizable amount of experimental data may be created in order to improve the image and be used for the training and testing of those ML models.

6. REFERENCES

[1] M. Aramesh et al. "A review of recent advances in solar cooking technology", Renewable Energy, Vol. 140, 2019, pp. 419-435.

[2] U. C. Arunachala, A. Kundapur, "Cost-effective solar cookers: A global review", Solar Energy, Vol. 207, pp. 903-916, 2020.

[3] M. C. Ndukwu, L. Bennamoun, M. Simo-Tagne, "Reviewing the exergy analysis of solar thermal systems integrated with phase change materials", Energies, Vol. 14, No. 3, 2021, p. 724.

[4] K. Lentswe, A. Mawire, P. Owusu, A. Shobo, "A review of parabolic solar cookers with thermal energy storage", Heliyon, Vol. 7, No. 10, 2021.

[5] R. Qahwaji, T. Colak, "Automatic short-term solar flare prediction using machine learning and sunspot associations", Solar Physics, Vol. 241, 2007, pp. 195-211.

[6] T. Colak, R. Qahwaji, "Automated Solar Activity Prediction: A hybrid computer platform using machine learning and solar imaging for automated prediction of solar flares", Space Weather, Vol. 7, No. 6, 2009.

[7] O. W. Ahmed, R. Qahwaji, T. Colak, P. A. Higgins, P. T. Gallagher, D. S. Bloomfield, "Solar flare prediction using advanced feature extraction, machine learning, and feature selection", Solar Physics, Vol. 283, 2013, pp. 157-175.

[8] M. G. Bobra, S. Couvidat, "Solar flare prediction using SDO/HMI vector magnetic field data with a machine-learning algorithm", Astrophysical Journal, Vol. 798, No. 2, 2015, p. 135.

[9] C. Voyant et al. "Machine learning methods for solar radiation forecasting: A review", Renewable Energy, Vol. 105, 2017, pp. 569-582.

[10] G. M. Yagli, D. Yang, D. Srinivasan, "Automatic hourly solar forecasting using machine learning models", Renewable and Sustainable Energy Reviews, Vol. 105, 2019, pp. 487-498.

[11] L. Cornejo-Bueno, C. Casanova-Mateo, J. Sanz-Justo, S. Salcedo-Sanz, "Machine learning regressors for solar radiation estimation from satellite data", Solar Energy, Vol. 183, 2019, pp. 768-775.

[12] G. de Freitas Viscondi, S. N. Alves-Souza, "A Systematic Literature Review on big data for solar photovoltaic electricity generation forecasting", Sustainable Energy Technologies and Assessments, Vol. 31, 2019, pp. 54-63.

[13] A. Mahmood, J.-L. Wang, "Machine learning for high performance organic solar cells: current scenario and future prospects", Energy & Environmental Science, Vol. 14, No. 1, 2021, pp. 90-105.

[14] Ü. Ağbulut, A. E. Gürel, Y. Biçen, "Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison", Renewable and Sustainable Energy Reviews, Vol. 135, 2021, p. 110114.

[15] A. J. Cetina-Quiñones, G. Santamaria-Bonfil, R. A. Medina-Esquivel, A. Bassam, "Techno-economic analysis of an indirect solar dryer with thermal energy storage: An approach with machine learning algorithms for decision making", Thermal Science and Engineering Progress, Vol. 45, 2023, p. 102131.

[16] Z. Tagnamas, A. Idlimam, A. Lamharrar, "Predictive models of beetroot solar drying process through machine learning algorithms", Renewable Energy, Vol. 219, 2023, p. 119522.

[17] Y. Ledmaoui, A. El Maghraoui, M. El Aroussi, R. Saadane, A. Chebak, A. Chehri, "Forecasting solar energy production: A comparative study of machine learning algorithms", Energy Reports, Vol. 10, 2023, pp. 1004-1012.

[18] M. Elgendi, M. Atef, "Calculating the impact of meteorological parameters on pyramid solar still yield using machine learning algorithms", International Journal of Thermofluids, Vol. 18, 2023, p. 100341.

[19] M. K. Nematchoua, J. A. Orosa, M. Afaifia, "Prediction of daily global solar radiation and air temperature using six machine learning algorithms; a case of 27 European countries", Ecological Informatics, Vol. 69, 2022, p. 101643.

[20] M. Oh et al. "Analysis of minute-scale variability for enhanced separation of direct and diffuse solar irradiance components using machine learning algorithms", Energy, Vol. 241, 2022, p. 122921.

[21] A. Khosravi, R. N. N. Koury, L. Machado, J. J. G. Pabon, "Prediction of hourly solar radiation in Abu Musa Island using machine learning algorithms", Journal of Cleaner Production, Vol. 176, 2018, pp. 63-75.

[22] M. A. Hassan, A. Khalil, S. Kaseb, M. A. Kassem, "Potential of four different machine-learning algorithms in modeling daily global solar radiation", Renewable Energy, Vol. 111, 2017, pp. 52-62.

[23] Y. Zahraoui, I. Alhamrouni, S. Mekhilef, M. R. B. Khan, "Machine learning algorithms used for short-term PV solar irradiation and temperature forecasting at microgrid", Applications of AI and IoT in Renewable Energy, Elsevier, 2022, pp. 1-17.

[24] Y. Feng, D. Gong, Q. Zhang, S. Jiang, L. Zhao, N. Cui, "Evaluation of temperature-based machine learning and empirical models for predicting daily global solar radiation", Energy Conversion and Management, Vol. 198, 2019, p. 111780.

[25] V. H. Quej, J. Almorox, J. A. Arnaldo, L. Saito, "ANFIS, SVM and ANN soft-computing techniques to estimate daily global solar radiation in a warm subhumid environment", Journal of Atmospheric and Solar-Terrestrial Physics, Vol. 155, 2017, pp. 62-70.

[26] H. Citakoglu, "Comparison of artificial intelligence techniques via empirical equations for prediction of solar radiation", Computers and Electronics in Agriculture, Vol. 118, 2015, pp. 28-37.

[27] D. Wu, J. Liu, Y. Yang, Y. Zheng, "Nitrogen/Oxygen Co-Doped Porous Carbon Derived from Biomass for Low-Pressure CO2Capture", Industrial & Engineering Chemistry Research, Vol. 59, No. 31, 2020, pp. 14055-14063.

[28] H. Roopa, T. Asha, "A linear model based on principal component analysis for disease prediction", IEEE Access, Vol. 7, 2019, pp. 105314-105318.

[29] G. A. F. Seber, A. J. Lee, "Linear regression analysis", John Wiley & Sons, 2003.

[30] D. C. Montgomery, E. A. Peck, G. G. Vining, "Introduction to linear regression analysis", John Wiley & Sons, 2021.

[31] S. Kavitha, S. Varuna, R. Ramya, "A comparative analysis on linear regression and support vector regression", Proceedings of the Online International Conference on Green Engineering and Technologies, 2016, pp. 1-5.

[32] A. Abdulazeez, B. Salim, D. Zeebaree, D. Doghramachi, "Comparison of VPN Protocols at Network Layer Focusing on Wire Guard Protocol", International Journal of Interactive Mobile Technologies, Vol. 14, No. 18, 2020.

[33] A. J. Smola, B. Schölkopf, "A tutorial on support vector regression", Statistics and Computing, Vol. 14, 2004, pp. 199-222.

[34] T. Aluja-Banet, E. Nafria, "Stability and scalability in decision trees", Computational Statistics, Vol. 18, No. 3-4, 2003, pp. 505-520.

[35] S. B. Kotsiantis, "Decision trees: a recent overview", Artificial Intelligence Review, Vol. 39, 2013, pp. 261-283.

[36] L. Breiman, "Random forests", Machine Learning, Vol. 45, 2001, pp. 5-32.

[37] G. James, D. Witten, T. Hastie, R. Tibshirani, "An introduction to statistical learning", Vol. 112, Springer, 2013.

[38] J. W. Harrison, M. A. Lucius, J. L. Farrell, L. W. Eichler, R. A. Relyea, "Prediction of stream nitrogen and phosphorus concentrations from high-frequency sensors using Random Forests Regression", Science of The Total Environment, Vol. 763, 2021, p. 143005.

[39] M. Schonlau, R. Y. Zou, "The random forest algorithm for statistical learning", Stata Journal, Vol. 20, No. 1, 2020, pp. 3-29.

[40] T. Chen, C. Guestrin, "Xgboost: A scalable tree boosting system", Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13-17 August 2016, pp. 785-794.

[41] J. H. Friedman, "Greedy function approximation: a gradient boosting machine", The Annals of Statistics, Vol. 29, No. 5, 2001, pp. 1189-1232.

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