Shahid Rajaee Teacher Training University

Journal of Electrical and Computer Engineering Innovations (JECEI)

Vol. 13 No. 1, Winter-Spring 2025

Journal of Electrical and Computer Engineering Innovations (JECEI) Vol. 13 No. 1, Winter-Spring 2025

JECEI

Journal of Electrical and Computer Engineering Innovations (JECEI)

Semiannual Publication

pISSN 2322-3952
eISSN 2345-3044

Volume 13, Issue 1, Winter-Spring 2025

Journal of Electrical and Computer Engineering Innovations
http://jecei.sru.ac.ir
EISSN: 2345-3044
ISSN: 2322-3952
JECEI
SRTTU
JECEI

License Holder: Shahid Rajaee Teacher Training University (SRTTU)

Address: Lavizan, 16788-15811, Tehran, Iran.

Journal of Electrical and Computer Engineering Innovations

Vol. 13; Issue 1: 2025

Contents

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Design and Stability Analysis of a Novel Path Planner for Autonomous Underwater Vehicles

Z. K. Pourtaheri *

Department of Mechatronics, Higher Education Complex of Bam, Bam, Iran.

Article Info

Article History:
Received 24 April 2024
Reviewed 20 June 2024
Revised 29 July 2024
Accepted 06 August 2024

Keywords:
Autonomous underwater vehicles
Path planning
Heuristic algorithms
Optimization
Stability analysis

*Corresponding Author's Email Address:
z.pourtaheri@bam.ac.ir

Abstract

Background and Objectives: According to this fact that a typical autonomous underwater vehicle consumes energy for rotating, smoothing the path in the process of path planning will be especially important. Moreover, given the inherent randomness of heuristic algorithms, stability analysis of heuristic path planners assumes paramount importance.
Methods: The novelty of this paper is to provide an optimal and smooth path for autonomous underwater vehicles in two steps by using two heuristic optimization algorithms called Inclined Planes system Optimization algorithm and genetic algorithm; after finding the optimal path by Inclined Planes system Optimization algorithm in the first step, the genetic algorithm is employed to smooth the path in the second step. Another novelty of this paper is the stability analysis of the proposed heuristic path planner according to the stochastic nature of these algorithms. In this way, a two-level factorial design is employed to attain the stability goals of this research.
Results: Utilizing a Genetic algorithm in the second step of path planning offers two advantages; it smooths the initially discovered path, which not only reduces the energy consumption of the autonomous underwater vehicle but also shortens the path length compared to the one obtained by the Inclined Planes system optimization algorithm. Moreover, stability analysis helps identify important factors and their interactions within the defined objective function.
Conclusion: This proposed hybrid method has implemented for three different maps; 36.77%, 48.77%, and 50.17% improvements in the length of the path are observed in the three supposed maps while smoothing the path helps robots to save energy. These results confirm the advantage of the proposed process for finding optimal and smooth paths for autonomous underwater vehicles. Due to the stability results, one can discover the magnitude and direction of important factors and the regression model.

Introduction

Nowadays, autonomous mobile robots have become an inseparable part of the growing world. These robots have the capability to perform difficult and sensitive tasks in high-risk environments and, therefore, have attracted a lot of attention. One example of an important type of these robots is the Autonomous Underwater Vehicle (AUV).

Path planning is one of the most important research topics in the field of Autonomous Underwater Vehicles, i.e., determining the movement path of the AUVs from the starting point to the target point to carry out a specific mission. The AUV's path from origin to destination can be pre-programmed, and extensive research has been

conducted on optimizing these paths. However, during its movement, it's crucial for the AUV to avoid collisions with both fixed and moving obstacles, such as other AUVs and underwater creatures. So, it can be said that determining the appropriate path for AUV movement from origin to destination remains an important challenge.

Path planning for AUVs is a challenging problem that heavily relies on optimization techniques and so far, various methods have been presented to solve this problem. The most important weaknesses of these methods are the high probability of getting stuck into local optima, the large volume of calculations and also, the existence of deficiencies in facing the dynamic environment. These cases can make them inappropriate for long distances. But due to the ability of heuristic optimization algorithms to solve complex optimization problems with high dimensions, these algorithms are suitable candidates for solving the path planning problem. The most important advantages of heuristic algorithms are their flexibility, high compatibility, high speed and efficiency, and their global search characteristic. These algorithms can avoid local optima and converge to global optima by using special schemes. A significant body of research has been conducted in the field of heuristic optimization for AUV path planning.

For instance, genetic algorithm has been successfully applied in this area, as demonstrated in [1]. The proposed method in this study introduces a new operator to the algorithm to guarantee convergence to the global minimum value in case of multiple minima Reference [2] introduces a hierarchical approach based on genetic algorithm for path planning of AUVs. The method first divides the workspace of the AUV into obstacle-free and obstacle-filled regions. Then, it utilizes genetic algorithm to search for a path among the obstacle-free regions. Reference [3] proposes a genetic algorithm-based path planning method for AUVs. The method first discretizes the three-dimensional space between the start point and the end point into a grid. Then, it employs genetic algorithm to search for the optimal path among these grid points. Each chromosome represents a sequence of these grid points, and the path between the start and the target points is constructed by connecting these points. The objective function is designed to minimize the energy consumption of the AUV along the path. In this study, some grid points are randomly selected as obstacles to simulate static obstacles in the problem space. Therefore, the path represented by each chromosome should not include these obstacle points. References [4]-[7] provide further examples of AUV path planning in known environments with static obstacles using genetic algorithm.

Particle swarm optimization algorithm and its variant, quantum particle swarm optimization algorithm, have been widely used for path planning of autonomous underwater vehicles. A combination of particle swarm optimization algorithm with differential evolution algorithm is proposed for offline path planning of AUVs in [8]. This proposed approach reduces the computational cost while increasing the ability of the PSO algorithm to find the optimal path. An improved quantum particle swarm optimization algorithm is proposed in [9] for the path planning of AUVs. Safety, path length, and path angle are considered to define the fitness function, and a cubic spline interpolation algorithm is used to smooth the path. Reference [10] presents a method for path planning of AUVs, which consists of two parts. First, a general path between the origin and destination is determined using genetic algorithm. The goal of this part is to find the shortest possible path between the origin and destination. To find this path, the space between the origin, and destination is discretized and points that can be considered as positions are extracted. The genetic algorithm then finds the best possible sequence of points from the available points to find the path. The next part is responsible for processing the sequence found by genetic algorithm. In this module, a particle swarm optimization algorithm is used to search the space between each pair of consecutive points in the sequence to find a suitable path between these two points. Therefore, a general path is first determined, and then other suitable paths are found between each pair of points on the general path. In the second module, the particle swarm algorithm searches the space with the goal of finding the shortest path. A new path planning method for underwater environments using an enhanced quantum particle swarm optimization algorithm is introduced in [11]. This method leverages a technique called Deep Q-Network to learn and adapt its behavior. The algorithm analyzes data about the particles' positions and utilizes neural networks to choose the most suitable action from a set of five options. This approach empowers the particles to make informed decisions in various situations, leading to a significant improvement in the algorithm's ability to explore the entire search space effectively. Furthermore, the accuracy is enhanced by fine-tuning operations. Additionally, a custom fitness function is designed specifically for underwater environments. This function considers factors like path length, deflection angles, and currents, allowing the algorithm to navigate underwater environments more efficiently and locate the path with the least energy consumption.

A combination of ant colony optimization (ACO) algorithm with A* algorithm is used for path planning in [12]. The results of this research show that using this method, the AUV can successfully navigate through an area with dense obstacles. The challenge of two-dimensional autonomous path planning for AUVs

operating in environments with ocean currents and obstacles is addressed in [13], and an improved Fireworks-Ant Colony Hybrid Algorithm is proposed to tackle this problem. First, a two-dimensional Lamb vortex ocean current environment scheme that incorporates randomly distributed obstacles is created. Subsequently, a mathematical model is formulated for path planning, considering factors such as navigation time, energy consumption, and total distance traveled. A multi-objective ant colony algorithm for path planning of AUV is presented in [14]. This approach goes beyond traditional, single-objective path planning by considering path length, energy consumption, and safe navigation. To address the challenge of finding optimal paths for autonomous underwater vehicles in complicated environments, an improved ant colony optimization algorithm merged with particle swarm optimization algorithm is presented in [15]. Considering the limitations faced by AUVs, such as constrained energy and visual interval, the proposed algorithm employs a modified pheromone update rule and heuristic function informed by PSO. This allows the AUV to navigate efficiently by connecting designated points while abstaining from collisions with the static obstacles. By incorporating PSO, this improved ACO algorithm overpowers the limitations of the traditional approach. A new algorithm called dynamic multi-role adaptive collaborative ant colony optimization is presented in [16] to address the limitations of slow convergence and poor diversity in the traditional ant colony algorithm. The results of applying this algorithm in robot path planning, illustrate its successfulness in solving this problem. A new approach for path planning of AUVs during dam inspections is proposed in [17]. The goal is to create safe and reliable paths that avoid obstacles while minimizing sharp turns. This method improves upon the traditional ACO algorithm by incorporating a "corner-turning heuristic function." This function helps the AUV select straighter paths, reducing turning times and improving overall efficiency. The reference [18] focuses on enhancing the underwater path-planning capabilities of AUVs by addressing limitations inherent to traditional algorithms like the ant colony algorithm and the artificial potential field algorithm. To overcome these limitations, an optimized scheme for the artificial potential field ant colony algorithm is proposed. Compared to conventional ant colony and other benchmark algorithms, the proposed algorithm achieved significant improvements: path length reductions of 1.57% and 0.63% (simple environment) and 8.92% and 3.46% (complex environment). Additionally, this algorithm demonstrated faster convergence, with iteration time reductions of approximately 28.48% and 18.05% (simple environment) and 18.53% and 9.24% (complex environment).

A novel method for planning safe paths for AUVs navigating environments filled with obstacles is presented in [19]. To address this, a hybrid approach is introduced that combines the strengths of two nature-inspired algorithms: Grey Wolf Optimization (GWO) and GA. This combined method, called Hybrid Grey Wolf Optimization, allows AUVs to find safe paths while minimizing travel distance. The proposed algorithm tackles GWO's weakness of random initialization by using GA to generate a good starting point for the search. In this research, the ideal path considers both the distance traveled and the penalties incurred from avoiding obstacles.

Many conventional heuristic algorithms struggle with two limitations: slow progress towards optimal solutions and getting stuck on suboptimal ones too early. These issues are addressed in [20] by introducing a novel hybrid heuristic algorithm. It combines the strengths of genetic algorithms, ant colony optimization, and simulated annealing. The proposed heuristic fusion incorporates a novel mutation operator inspired by ant colony optimization. This operator allows individuals from different generations to exchange information, leading to better solutions and faster convergence. Additionally, a mechanism is introduced that dynamically adjusts the probability of genetic operations, similar to simulated annealing.

To effectively navigate AUVs in intricate environments, an improved differential evolution algorithm is proposed in [21]. This approach incorporates a novel adaptive elite neighborhood learning strategy to achieve a balance between the exploitation and exploration capabilities of improved differential evolution when tackling complex problems. Additionally, a rank-guided crossover probability selection strategy is introduced to ensure effective preservation of information from elite individuals. Finally, the study explores a novel distance-greedy selection strategy, which improves population diversity while maintaining convergence accuracy. Moreover, this research introduces a new double-layer coding model for eliminating invalid path points.

Energy consumption is one of the challenges of AUVs due to the limited battery power. An AUV requires more energy to probe the coastal waters over a large path against the rough environmental situations dominant in the sea. Therefore, AUVs must supply the best detection performance with decreased search distance. An optimal and efficient path planning algorithm should be applied in AUVs [22].

Due to the unmanned nature of AUVs and the importance of saving battery power in them, the issue of the optimal path is more critical in this case. Therefore, in this paper, AUVs are specifically discussed.

According to the importance of energy consumption, in this paper, a two-step method is presented to reduce

not only the path but also the energy consumption; first, the Inclined Planes system Optimization algorithm as a powerful heuristic algorithm is employed to obtain the path with optimal length and then Genetic Algorithm is utilized to smooth the obtained path in order to decrease the energy consumption of the AUV. Finally, the stability of the presented heuristic path planner is analyzed. In this part, the effect of two structural parameters of applied heuristic algorithms on the designed path planner is investigated. It's worth mentioning that stability analysis of the heuristic path planner of AUVs is addressed in this paper for the first time.

Recently, many heuristic algorithms have been introduced. Some of these methods include the Orchard algorithm [23], the Meerkat optimization algorithm [24], the Artificial Rabbits optimization algorithm [25], and the Arithmetic Optimization algorithm [26]. Evaluating the performance and capabilities of each algorithm in different applications is one of the research areas of interest. Therefore, in this paper, the IPO algorithm is used for the first time in the field of AUV path planning to evaluate its performance. This research has shown that this algorithm is suitable for the path planning problem to meet expectations.

Genetic Algorithm is employed in this paper for several reasons. Firstly, GA has a theoretical foundation for convergence and guarantees global optimality. Secondly, it has been successfully applied to robot path planning in numerous prior studies. Thirdly, it is utilized as a hybrid approach in this research. However, it is crucial to emphasize that there are numerous alternative methods that could be investigated to examine their applicability in the path planning problem. Anti-coronavirus optimization algorithm [27], Backtracking search optimization algorithm [28] and Seasons optimization algorithm [29] are Some of these methods. In addition, a large number of algorithms can be extracted from [30] because more than three hundred researches related to bio-inspired and nature-inspired algorithms are reviewed in this paper.

The rest of this paper is organized as follows: first, the employed method for stability analysis is presented. After that, a review of Inclined Planes system Optimization algorithm is described. Then the proposed combinational method for achieving the optimal and smooth path of AUVs and the stability analysis of this heuristic path planner are presented. The experimental results are reported in the next section. Finally, conclusion of the paper is explained.

Two-Level Factorial Designs

When it's necessary to investigate the joint effects of several parameters on output in the experiments, factorial designs are extensively exerted. Joint effects implicate interactions and original effects. A significant case in this field is when two levels for each of the parameters exist; this type is named 2^k factorial designs as regards every replicate owning precisely 2^k experimental runs. 2^k factorial designs are helpful when screening tests should be performed to discover significant factors.

Adjusting a first order Response Surface Model (RSM) and acquiring the estimate of factor effect are the other applications of them.

One can employ factorial designs to determine the influence of several independent factors upon one dependent variable. There are two factors in 2^2 factorial designs (*A* and *B*), and two levels are defined for each parameter. The expressions high and low are employed for these levels. *A* and *B* indicate the impact of parameters *A* and *B*, respectively. Moreover, *AB* refers to the *AB* interaction. In this scheme, + and - are applied to show high and low levels related to each factor. Table 1 shows the design matrix, which specifies four treatment combinations of 2^2 design.

Table 1: The design matrix

Run	A	B
1	-	-
2	+	-
3	-	+
4	+	+

Small letters also illustrate the four runs; small letter related to each factor indicates the high level of it and the miss of one letter specifies the low level of that factor. So, *a* betokens the situation in which the level of *A* is high and the level of *B* is low and *ab* means the levels of two parameters are high. When the levels of all parameters are low, *1* is applied.

To calculate the original effect of *A* the difference of two averages is employed; the average of two combinations where the level of *A* is high ($\bar{y}_{A^+}$) and the average of two combinations where the level of *A* is low ($\bar{y}_{A^-}$). Thus, the main effect of *A* is specified as (1).

$$A = \bar{y}_{A^+} - \bar{y}_{A^-} = \frac{ab+a}{2n} - \frac{b+1}{2n} = \frac{ab+a-b-1}{2n} \quad (1)$$

In the same way, the main effect *B* is measured in (2).

$$B = \bar{y}_{B^+} - \bar{y}_{B^-} = \frac{ab+b}{2n} - \frac{a+1}{2n} = \frac{ab+b-a-1}{2n} \quad (2)$$

The interaction effect *AB* is the average of the difference of the effect *A* at low and high levels of *B*. So, the interaction *AB* is specified as (3).

$$AB = \frac{1}{2}\left\{\frac{[ab-b]}{n} - \frac{[a-1]}{n}\right\} = \frac{ab+1-a-b}{2n} \quad (3)$$

In many experiments of 2^k designs, both the direction and magnitude of the parameter effects are studied to discover significant parameters. Comparing the magnitudes of the effects in terms of their related standard errors is an advantageous method for advising the importance of the effects. To measure the standard error of *A, B,* and *AB*, one can compute the sums of squares for effects that are specified by SS_A, SS_B, and SS_{AB} in (4)-(6), respectively.

$$SS_A = \frac{(ab+a-b-1)^2}{4n} \quad (4)$$

$$SS_B = \frac{(ab+b-a-1)^2}{4n} \quad (5)$$

$$SS_{AB} = \frac{(ab+1-a-b)^2}{4n} \quad (6)$$

The total sums of squares, i.e. SS_T is measured by using (7) where y_{ijk} is the outcome of each run and $y_{...}$ is the sum of all runs.

$$SS_T = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{n} y_{ijk}^2 - \frac{y_{...}^2}{4n} \quad (7)$$

Finally, the error sum of squares (SS_E) is measured by using (8).

$$SS_E = SS_T - SS_A - SS_B - SS_{AB} \quad (8)$$

Considering the degrees of freedom of SS_E i.e., $4\times(n-1)$, the mean square error, MS_E, is specified as (9).

$$MS_E = \frac{SS_E}{4\times(n-1)} \quad (9)$$

Therefore, the standard error of an effect is calculated by using (10).

$$se(effect) = \sqrt{\frac{1}{n}MS_E} \quad (10)$$

Finally, for each effect estimate, two standard error limits exist as (11).

$$\begin{aligned} &A \pm 2\times se(effect) \\ &B \pm 2\times se(effect) \\ &AB \pm 2\times se(effect) \end{aligned} \quad (11)$$

Due to the considered analysis, if the interval of an effect estimate does not include zero, it is introduced a significant effect. At the end, it should be mentioned that the coefficients for the regression model are half of the corresponding factor effect estimates [31].

Inclined Planes System Optimization Algorithm (IPO)

The movement of several globular things on a frictionless ramp is the main basis of the IPO algorithm; these objects want to arrive at the lowest place on the ramp. In this algorithm, some tiny balls, as algorithm agents, probe the search space to discover the optimal point. The principal scheme of this algorithm is to attribute the height to every object according to a reference point. This height value is received from the objective function; the obtained values are an approximation of potential energy of the agents at different points, and as the balls descend, this energy is converted into kinetic energy and thus caused the balls to accelerate downwards. Therefore, the agents repeatedly move in the exploring space to discover a better answer and hence acquire an acceleration [32].

In a supposed search space with *N* agents, the position of the *i*-th agent is calculated as (12):

$$x_i = \left(x_i^1, \mathrm{K}, x_i^d, \mathrm{K}, x_i^n\right), \quad for\ \ i = 1,2,\mathrm{K},N \quad (12)$$

in which, x_i^d is the position of *i*-th agent in the *d*-th dimension in an *n*-dimensional system. The angle between the *i*-th and *j*-th agents in dimension *d*, i.e., φ_{ij}^d is measured by (13):

$$\varphi_{ij}^d(t) = \left(\tan^{-1}\left(\frac{f_j(t)-f_i(t)}{x_i^d(t)-x_j^d(t)}\right)\right), \quad for\ d = 1,\mathrm{K},n\ and\ i,j = 1,2,\mathrm{K},N, i \neq j \quad (13)$$

where, $f_i(t)$ is the value of the objective function, i.e., height for the *i*-th agent at the time *t*. A certain agent wants to move toward the lowest heights on the ramp, therefore the agents with lower height values are the only agents used in calculating the acceleration. The direction and amplitude of acceleration for the *i*-th agent in dimension *d* and at the time *t*, is demonstrated in (14) where, *U*(.) means the unit step function:

$$a_i^d(t) = \sum_{j=1}^{N} U\left(f_j(t)-f_i(t)\right).\sin\left(\varphi_{ij}^d(t)\right) \quad (14)$$

Finally, (15) is employed to update the position of the balls:

$$\begin{aligned} x_i^d(t+1) = {} & k_1.rand_1.a_i^d(t).\Delta t^2 \\ & + k_2.rand_2.v_i^d(t).\Delta t + x_i^d(t) \end{aligned} \quad (15)$$

$rand_1$ and $rand_2$ are two random parameters distributed uniformly on the [0,1] interval. $v_i^d(t)$ is the velocity related to the *i*-th agent at time *t* and in dimension *d*. k_1

and k_2 are applied to control the exploring process of the algorithm. These factors are defined by using (16) and (17):

$$k_1(t) = \frac{c_1}{1+\exp\left((t-shift_1)\times scale_1\right)} \quad (16)$$

$$k_2(t) = \frac{c_2}{1+\exp\left((t-shift_2)\times scale_2\right)} \quad (17)$$

$v_i^d(t)$ is defined in (18), where x_{best} is placed in numerator to demonstrate the agent's desire to achieve the best position in each run:

$$v_i^d(t) = \frac{x_{best}^d(t) - x_i^d(t)}{\Delta t} \quad (18)$$

Design and Stability Investigation of Heuristic Path Planner for AUVs

This research proposes a hybrid approach that combines the strengths of IPO and Genetic GA to achieve an optimal and smooth path for Autonomous Underwater Vehicles. This two-step method leverages IPO for global optimization and GA for path refinement, ultimately leading to efficient and safe AUV navigation. In the first step, the IPO algorithm is used to find the optimal path. The objective function is one of the important issues that should be defined properly when employing heuristic algorithms for optimization. Path length is considered as objective function in this research. It is expected that this objective function will be minimized by using IPO algorithms.

The path length is considered as a criterion for measuring the quality of the path in path planning problem. Therefore, the shorter the path it is, the better the fitness function it is.

In this research, the Euclidean distance is used to calculate the path length. When the system is implemented in the real world, the unit of the obtained path length will be in terms of the actual distance between the start and the end points. An important point to note here is the path length depends on various factors, including the speed of the AUV, the run speed of the algorithm, and the sonar accuracy. For example, suppose that based on sonar accuracy, the sonar detects an obstacle ten meters away, and the speed of the AUV is one meter per second. Additionally, the algorithm requires 10 runs to reach the next point, which takes one second. Therefore, during this time, the AUV has moved one meter, and until the sonar doesn't warn, the algorithm continues on its path and begins to explore the next point.

In the path planning problem, the goal is to find a possible path from the starting point to the end point of the movement, and after optimizing the length of this route, the path should have the shortest length. While safety is a paramount concern for any path, a safe path should be both collision-free and minimize its overall length. So, one can conclude that there is a constrained optimization problem to solve. Here, the penalty function method is used to solve this constrained optimization problem. For this purpose, the objective function is defined as the summation of the path length with a penalty function in the form of (19).

$$Objective\ Function = L\times(1+\beta\times V) \quad (19)$$

where L is the path length, V is the penalty function, and β is the coefficient of the penalty function. For all points of the path, the penalty is calculated, and finally, the penalty function will be the average of all calculated values. Obviously, if a point of the path does not collide with an obstacle, the penalty for that point is zero.

After detecting the optimal path using IPO, genetic algorithm is employed to smooth the obtained path. The importance of smoothing path is due to the energy consumption of AUV when turning.

So, if the path improves in such a way that the AUV turns with less angles, as a result, less energy is consumed.

It's noteworthy that path smoothing with a genetic algorithm can offer a twofold benefit for AUVs. First, it reduces energy consumption by minimizing unnecessary rotations. Second, it shortens the overall path compared to IPO's obtained path by eliminating redundant waypoints.

The second goal of this research is using two-level factorial designs to investigate the stability of designed heuristic path planner. So, two structural parameters which are common in both algorithms, are selected to check the stability of the proposed path planner. These parameters are number of iterations and population size.

Results and Discussion

The proposed method used for designing heuristic path planner is implemented for three different maps and the results are reported as follows. The dimensions of the search agents are equal to the number of points in the cubic spline Interpolation.

It's worth mentioning the values of parameters in the IPO algorithm are considered as below:

$c_1 = 0.225$.

$c_2 = 2.283$.

$shift_1 = 121.044$.

$shift_2 = 149.675$.

$scale_1 = 0.056$.

$scale_2 = 0.525$.

A. Map 1

In this case, in the first step, IPO algorithm has reached

the path with a length of 20.1796. In the next step, after applying genetic algorithm to this problem, two important consequences have obtained; firstly, the path has become smooth and secondly, the length of the path has reduced to 12.7596; That is, the path length has improved 36.77% by using the proposed method. The path found by IPO and the smooth path discovered by the combined algorithm are shown in Fig. 1 and Fig. 2.

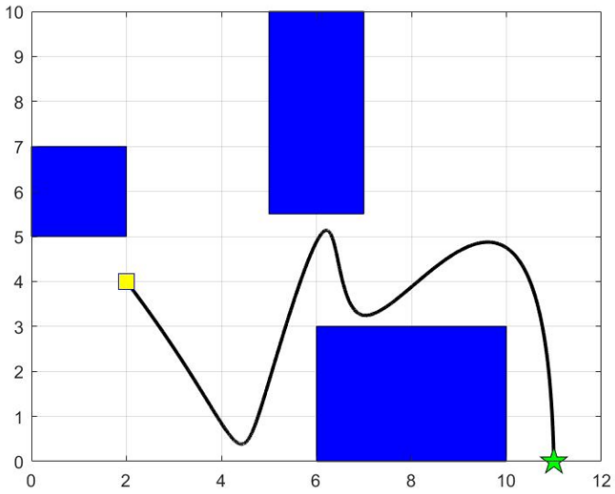


Fig. 1: Optimal path for map 1 using IPO algorithm.

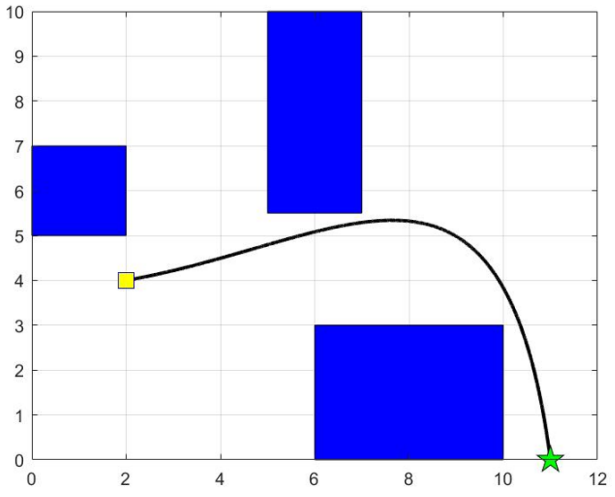


Fig. 2: Optimal and smooth path for map 1 using the hybrid algorithm of IPO-GA.

B. *Map 2*

In this case, at the first step, the path length obtained by IPO algorithm is 27.2011. The application of the genetic algorithm in the second step yielded two key results. First, the path was significantly smoothed. Second, the path length was reduced to 13.5540, representing a remarkable improvement of 50.17% compared to the original path. The path found by IPO and the smooth path discovered by the combined algorithm for map 2 are shown in Fig. 3 and Fig. 4.

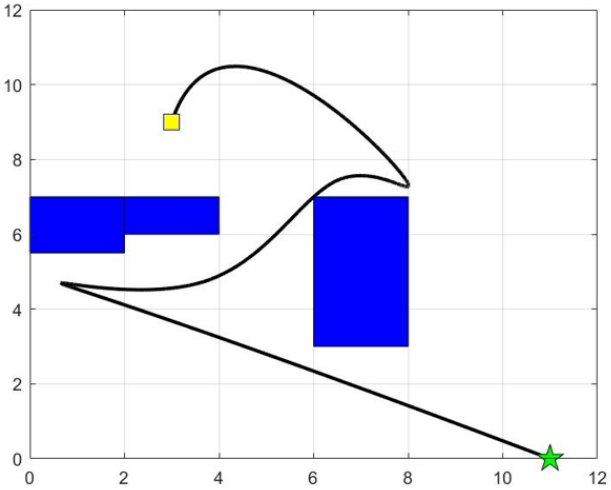


Fig. 3: Optimal path for map 2 using IPO algorithm.

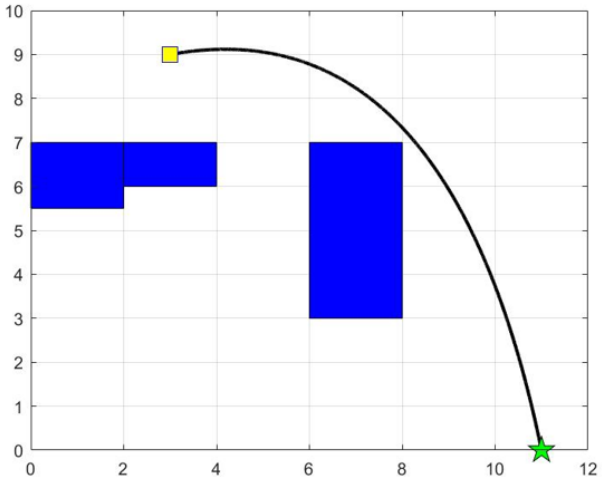


Fig. 4: Optimal and smooth path for map 2 using the hybrid algorithm of IPO-GA.

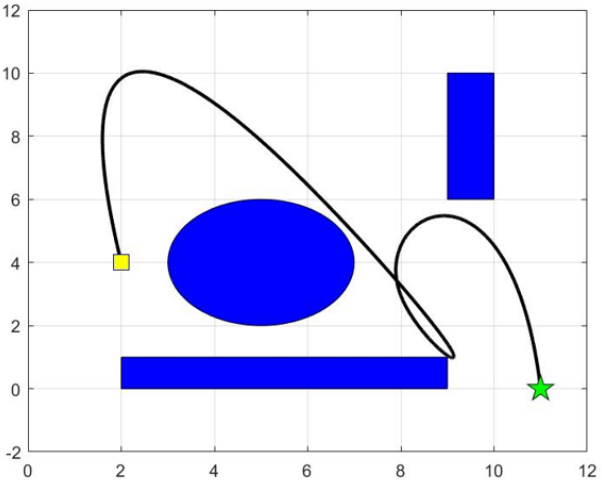


Fig. 5: Optimal path for map 3 using IPO algorithm.

C. *Map 3*

During the first step, the IPO algorithm was utilized to

optimize path length, resulting in a path with a length of 29.4538. The final step involving the genetic algorithm yielded two significant improvements. First, the path became noticeably smoother. Second, the path length was reduced to 15.0879, representing a noteworthy improvement of 48.77% achieved through the proposed hybrid method.

The path found by IPO and the smooth path discovered by the combined algorithm for map 3 are shown in Fig. 5 and Fig. 6.

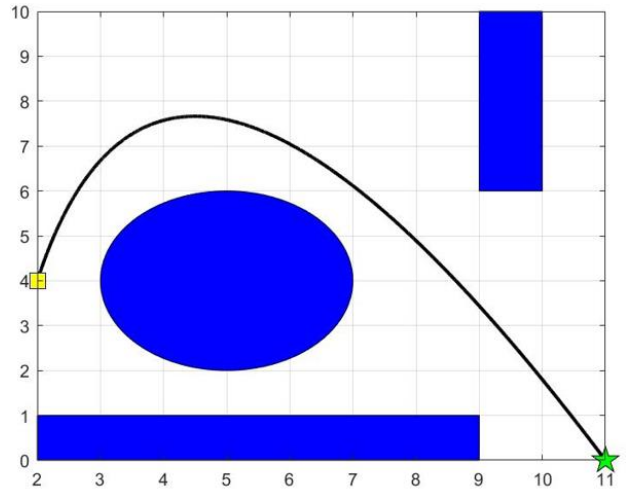


Fig. 6: Optimal and smooth path for map 3 using the hybrid algorithm of IPO-GA.

The obtained results in each step for three maps are shown in Table 2.

Table 2: The path length in each step

Map	First Step	Second Step	Improvement (%)
1	20.1796	12.7596	**36.77**
2	27.2011	13.5540	**50.17**
3	29.4538	15.0879	**48.77**

The maps used in this research are synthetic maps designed for the present study to evaluate the performance of the proposed method for AUV path planning. These maps use obstacles with different geometric shapes and variant sizes to evaluate the effectiveness of the proposed method. Therefore, a one-to-one comparison with other researches is not possible. However, computational cost can be considered as a metric for comparison with previous studies, in which case the number of iterations can be an appropriate choice. Table 3 shows the comparison results between the proposed method and several other methods in terms of number of iterations. According to this comparison, the number of iterations required by the proposed method is significantly lower than those of other methods.

Table 3: A comparative analysis of Number of iterations between the proposed method and other existing methods

Method	Number of iterations
[16]	2000
[33]	1000
[34] _PSO	2400
[34]_ACO	480
[35]	1000
Proposed Method	500

It’s worth to mention that we can never accurately determine the actual operational delay of the above methods because the coding style and instructions are crucial for delays.

In fact, even if we measure the order of computations, we can never accurately determine the computational cost, as there are many important factors that affect it during implementation, such as the type of hardware, the program itself, and the programmer. If all of these factors are identical, there are still important considerations during operational implementation; even if we have a highly optimized code running on powerful hardware, there are still operational factors that can affect execution time such as sensor delay, sensor accuracy, data acquisition rate, data processing time, and data decoding; these operational considerations introduce a layer of variability that makes it difficult to precisely determine the computational cost of a method in a real-world setting.

In practice, we often rely on approximations, and metrics to get a general sense of the computational efficiency of different methods. However, it's important to recognize that these estimates may not reflect the exact execution time in a specific application or environment.

It is also important to consider that the maps, the programming style, and the methods employed in this research are all unique.

It is precisely after taking these factors into account, it appears that the number of iterations is the most logical measure and criterion that can be considered as an estimate, rather than saying that because the number of iterations is a certain value, the computational cost must also be the same.

In fact, the claim has been that since it was not possible to implement these methods, the number of iterations

was simply used as a handy and documented metric for comparison. So, without a doubt the number of iterations of an algorithm may not accurately indicate its speed or efficiency, but this can still be used as a metric. Now, we can take a step further and complete the before statement: even the number of iterations reported is not entirely reliable as all heuristic methods are problem based and random methods.

This means that if a heuristic method finds a solution to a problem in an experiment with a specific number of iterations, it may take more or fewer iterations to find the solution to the same problem in a different experiment. However, in the absence of a documented metric, it seems that one of the logical metrics to use is the number of iterations. This is because the search loop is the core component of the body of all heuristic methods. In fact, all heuristic methods have a search loop that constitutes the largest part of the body of a heuristic method, including all the operators that need to be performed and the termination conditions. In fact, the search loop is typically the most computationally expensive part of a heuristic method. Therefore, the number of iterations of this part, as a common and substantial body of all heuristic algorithms, is an appropriate metric to understand how many times calculations were needed to reach the main solution.

As mentioned before, two selected factors for stability analysis are population size (*A*) and number of iterations (*B*) and two levels are assumed for each factor; 100 and 200 for population size, 500 and 700 for number of iterations in IPO. 50 and 100 for population size, 700 and 1000 for number of iterations in GA. Moreover, map 3 is used to accomplish stability analysis.

It's worth mentioning that for each treatment combination, the experiment is repeated two times in order to obtain needful data for stability investigation. Obtained data from two replicates for stability analysis of IPO are indicated in Table 4.

Table 4: Observed Data for IPO

Treatment Combination	Replicate	Path Length
1	I	27.4652
	II	33.7509
a	I	23.3083
	II	29.3106
b	I	23.8117
	II	19.7694
ab	I	28.193
	II	33.1777

Two standard error limits on the effects related to supposed fitness function (path length) is calculated by employing observed data. Obtained interval for each effect is shown in (20).

$$\begin{aligned} A &: 2.2981 \pm 5.4020 \\ B &: -2.2208 \pm 5.4020 \\ AB &: 6.5967 \pm 5.4020 \end{aligned} \tag{20}$$

Due to above intervals, it's obvious that effect *AB* in path length, optimized with IPO, is important because its interval does not include zero.

Now, the regression model for path length measure for IPO can be specified as (21). Where x_1 and $x2$ are the design factors *A* and *B*, respectively, on the coded (-1, +1) scale and β_0 is the mean of all observations of path length measure.

$$\begin{aligned} y &= \beta_0 + \beta_{12} x_1 x_2 \\ &= 27.3484 + \left(\frac{6.5967}{2} \right) x_1 x_2 \end{aligned} \tag{21}$$

The direction of each factor, which is also extracted from this analysis, is another important point; the effect *AB* in path length is positive in this path planning problem; i.e., if *AB* increases from the low to the high level, the path length measure will increase.

Obtained data from 2 replicates for stability analysis of GA are shown in Table 5.

Table 5: Observed Data for GA

Treatment Combination	Replicate	Path Length
1	I	17.2608
	II	19.9797
a	I	10.7910
	II	13.8029
b	I	13.2827
	II	13.4881
ab	I	12.9068
	II	15.8683

Two standard error limits on the effects related to the path length obtained by GA, is defined by employing observed data. The related interval for each effect is shown in (22).

$$\begin{aligned} A &: -2.6606 \pm 2.5138 \\ B &: -1.5721 \pm 2.5138 \\ AB &: 3.6627 \pm 2.5138 \end{aligned} \tag{22}$$

Due to obtained intervals, one can conclude that effect *A* and *AB* in path length, optimized by applying GA, are significant because these intervals do not include zero. Hence, the regression model for path length measure for GA can be described as (23).

$$y = \beta_0 + \beta_1 x_1 + \beta_{12} x_1 x_2 = 14.6725 + \left(\frac{-2.6606}{2}\right) x_1 + \left(\frac{3.6627}{2}\right) x_1 x_2 \quad (23)$$

The effect *A* in path length measure is negative; i.e., increasing *A* from the low to the high level will decrease the path length measure. Also, *AB* is positive in this path planning problem; i.e., increasing the level of *AB* from low to high leads the measure of path length to increase.

Conclusion

In this study, a hybrid heuristic method is developed to detect the optimal and smooth path from the starting point to the target point for AUVs. The proposed method is implemented in two steps by using IPO and GA. In the first step, the only goal is to find an optimal path for AUV by applying IPO. In the next step, GA is employed to smooth the detected path obtained in the previous step. The results of the last step are an optimal and smooth path, i.e., the GA not only smooths the path but also decreases the length of the path. This method reduces AUV energy consumption by eliminating unnecessary turns.

So, it is an efficient method for the path planning of AUVs.

The IPO-GA algorithm has applied on three different maps. In all three cases, the results confirm the efficiency of the proposed method; after using GA, the path becomes smooth and also shorter. In the best case, an improvement of 50.17% is seen in the length of the path.

After developing the proposed heuristic path planner, it's time to study the stability of this heuristic path planner.

Due to the random nature of heuristic algorithms, this part seems to be necessary. By applying the two-level factorial design, one can efficiently evaluate how each parameter (population size and number of iterations) and their potential interactions affect the objective function (path length).

This will help you identify the optimal configuration for your genetic algorithm in optimizing AUV path planning. This approach can find significant effects in each situation and also, can specify regression model related to the defined objective function.

Finally, it is necessary to emphasize that some suggestions can be made for future work in this field. One of these suggestions is using other new algorithms with a new objective function. It is also possible to use multi-objective heuristic algorithms with several objective functions.

Another important suggestion is to investigate the stability of the proposed heuristic method with other approaches and also by considering other parameters.

Author Contributions

Z. K. Pourtaheri designed the framework of the research. She designed the experiments and interpreted the results and wrote the manuscript.

Acknowledgment

This work is completely self-supporting, thereby no any financial agency's role is available.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

IPO	Inclined Planes system Optimization
AUV	Autonomous Underwater Vehicle
GA	Genetic Algorithm
RSM	Response Surface Model
PSO	Particle Swarm Optimization
ACO	Ant Colony Optimization

References

[1] A. Alvarez, A. Caiti, R. Onken, "Evolutionary path planning for autonomous underwater vehicles in a variable ocean," IEEE J. Oceanic Eng., 29(2): 418-429, 2004.

[2] Q. R. Zhang, "A hierarchical global path planning approach for AUV based on genetic algorithm," in Proc. 2006 International Conference on Mechatronics and Automation, 2006.

[3] W. Hong-Jian, B. Xin-Qian, Z. Jie, D. Fu-Guang, X. Guo-Qing, "A GA path planner based on domain knowledge for AUV," in Proc. Oceans '04 MTS/IEEE Techno-Ocean '04, 3: 1570-1573, 2004.

[4] J. Cao, Y. Li, S. Q. Zhao, X. S. Bi, "Genetic-algorithm-based global path planning," in Proc. 9th International Symposium on Computational Intelligence and Design, 2016.

[5] Y. Sun, R. B. Zhang, "Research on global path planning for AUV based on GA," in Proc. Mechanical Engineering and Technology: 311-318, 2012.

[6] Q. Li, X. H. Shi, Z. Q. Kang, "The application of an improved genetic algorithm in the AUV global path planning," Appl. Mech. Mate., 246: 1165-1169, 2013.

[7] S. K. Yan, F. Pan, "Research on route planning of AUV based on genetic algorithms," in Proc. IEEE International Conference on Unmanned Systems and Artificial, 2019.

[8] H. S. Lim, S. S. Fan, C. K. H. Chin, S. H. Chai, N. Bose, "Particle swarm optimization algorithms with selective differential evolution for AUV path planning," Int. J. Rob. Autom., 9(2): 94-112, 2020.

[9] L. Wang, L. L. Liu, J. Y. Qi, W. P. Peng, "Improved quantum particle swarm optimization algorithm for offline path planning in AUVs," IEEE Access, 8: 143397-143411, 2020.

[10] Z. Zeng, K. Sammut, L. Lian, F. He, A. Lammas and Y. Tang, "A comparison of optimization techniques for AUV path planning in environments with ocean currents," Rob. Auton. Syst., 82: 61-72, 2016.

[11] H. Zhang, X. Shi, "An improved quantum-behaved particle swarm optimization algorithm combined with reinforcement learning for AUV path planning," J. Rob., 2023.

[12] X. Yu, W. N. Chen, T. Gu, H. Q. Yuan, H. Zhang, J. Zhang, "ACO-A*: Ant colony optimization plus A* for 3-D traveling in environments with dense obstacles," IEEE Trans. Evol. Comput., 23(4): 617-631, 2019.

[13] Y. Ma, Z. Y. Mao, T. Wang, J. Qin, W. J. Ding, X. Meng, "Obstacle avoidance path planning of unmanned submarine vehicle in ocean current environment based on improved firework-ant colony algorithm," Comput. Electr. Eng., 87, 106773, 2020.

[14] C. L. Hu, F. Zhang, "Research on AUV global path planning based on multi-objective ant colony strategy," in Proc. Chinese Automation Congress, 2019.

[15] G. F. Che, L. J. Liu, Z. Yu, "An improved ant colony optimization algorithm based on particle swarm optimization algorithm for path planning of autonomous underwater vehicle," J. Ambient Intell. Hum. Comput., 11(8): 3349-3354, 2020.

[16] D. Zhang, X. You, S. Liu, H. Pan, "Dynamic multi-role adaptive collaborative ant colony optimization for robot path planning," IEEE Access, 8: 129958-129974, 2020.

[17] M. Ronghua, C. Xinhao, W. Zhengjia, D. Xuan, "Improved ant colony optimization for safe path planning of AUV," Heliyon, 10(7), 2024.

[18] G. Chen, D. Cheng , W. Chen, X. Yang, T. Guo , "Path planning for AUVs based on improved apf-ac algorithm," Comput. Mater. Continua, 78(3), 2024.

[19] S. P. Sahoo, B. Das, B.B. Pati, F. P. Garcia Marquez, I. Segovia Ramirez, "Hybrid path planning using a bionic-inspired optimization algorithm for autonomous underwater vehicles," J. Mar. Sci. Eng., 11(4): 761, 2023.

[20] J. Wen, J. Yang, T. Wang, "Path planning for autonomous underwater vehicles under the influence of ocean currents based on a fusion heuristic algorithm," IEEE Trans. Veh. Technol., 70(9): 8529-8544, 2021.

[21] J. Fan, L. Qu , "Innovative differential evolution algorithm with double-layer coding for autonomous underwater vehicles path planning in complex environments," Ocean Eng., 303, 2024.

[22] O. Marceau, J. M. Vanpeperstraete, "AUV optimal path for leak detection," in Proc. OCEANS–Anchorage, 2017.

[23] M. Kaveh, M. S. Mesgari, B. Saeidian, "Orchard Algorithm (OA): A new meta-heuristic algorithm for solving discrete and continuous optimization problems," Math. Comput. Simul., 208: 95-135, 2023.

[24] S. Xian, X. Feng, "Meerkat optimization algorithm: A new meta-heuristic optimization algorithm for solving constrained engineering problems," Expert Syst. Appl., 231, 2023.

[25] L. Wang, Q. Cao, Z. Zhang, S. A. Mirjalili, "Artificial rabbits optimization: A new bio-inspired meta-heuristic algorithm for solving engineering optimization problems," Eng. Appl. Artif. Intell., 114, 2022.

[26] L. Abualigah, A. Diabat, S. A. Mirjalili, M. Abd Elaziz, "The arithmetic optimization algorithm," Comput. Meth. Appl. Mech. Eng., 376: 2021.

[27] H. Emami, "Anti-coronavirus optimization algorithm," Soft Comput., 26: 4991-5023, 2022.

[28] P. Civicioglu, "Backtracking search optimization algorithm for numerical optimization problems," Appl. Math. Comput., 219(15): 8121-8144, 2013.

[29] H. Emami, "Seasons optimization algorithm," Eng. Comput., 38: 1845-1865, 2022.

[30] M. Daniel, J. Poyatos, J. Del Ser, S. García, A. Hussain, F. Herrera, "Comprehensive taxonomies of nature-and bio-inspired optimization: Inspiration versus algorithmic behavior," Cognit. Comput., 12: 897-939, 2020.

[31] H. Myers Raymond, C. Montgomery Douglas, M. Anderson-Cook Cristine, Response surface methodology: process and product optimization using designed experiments, John Wiley & Sons, 2016.

[32] H. Mozaffari, M. H. Abdy, S. H. Zahiri, "IPO: an inclined planes system optimization algorithm," Comput. Inf., 35(1): 222-240, 2016.

[33] L. Dogan, U. Yuzgec, "Robot path planning using gray wolf optimizer," in Proc. International Conference on Advanced Technologies, Computer Engineering and Science, 2018.

[34] M. Yousif, A. Salim, W. Jummar, "A robotic path planning by using crow swarm optimization algorithm," Int. J. Math. Sci. Comput., 7(1): 20-25, 2021.

[35] X. Li, D. Wu, J. He, M. Bashir, M. Liping, "An improved method of particle swarm optimization for path planning of mobile robot," J. Control Sci. Eng., 2020: 1-12, 2020.

Biographies

Zeinab Khatoun Pourtaheri received her B.Sc. and M.Sc. degrees in Electrical Engineering from Shahid Bahonar University of Kerman in 2010 and 2012, respectively, and Ph.D. degree in Electrical Engineering from University of Birjand in 2017. She is an Assistant Professor with the Department of Mechatronic Engineering at Higher Education complex of Bam. Her major research interests include Path planning, heuristic algorithms, optimization, ensemble classification, and stability analysis of heuristic methods.

- Email: z.pourtaheri@bam.ac.ir
- ORCID: 0000-0003-0465-5488
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
Z. K. Pourtaheri, "Design and stability analysis of a novel path planner for autonomous underwater vehicles," J. Electr. Comput. Eng. Innovations, 13(1): 1-12, 2025.

DOI: 10.22061/jecei.2024.10624.725

URL: https://jecei.sru.ac.ir/article_2170.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Robust Fuzzy Control of Uncertain Two-axis Inertially Stabilized Platforms Using a Disturbance Observer: A Backstepping-based Adaptive Control Approach

M. Ghalehnoie [*]***, A. Azhdari, J. Keighobadi***

Faculty of Electrical Engineering, Shahrood University of Technology, Shahrood, Iran.

Article Info

Article History:
Received 08 May 2024
Reviewed 27 June 2024
Revised 30 July 2024
Accepted 09 August 2024

Keywords:
Inertially stabilized platform
Backstepping control
Disturbance observer
Fuzzy approximation
Model-free control

[*]Corresponding Author's Email Address:
ghalehnoie@shahroodut.ac.ir

Abstract

Background and Objectives: The two-axis inertially stabilized platforms (ISPs) face various challenges such as system nonlinearity, parameter fluctuations, and disturbances which makes the design process more complex. To address these challenges effectively, the main objective of this paper is to realize the stabilization of ISPs by presenting a new robust model-free control scheme.

Methods: In this study, a robust adaptive fuzzy control approach is proposed for two-axis ISPs. The proposed approach leverages the backstepping method as its foundational design mechanism, employing fuzzy systems to approximate unknown terms within the control framework. Furthermore, the control architecture incorporates a model-free disturbance observer, enhancing the system's robustness and performance. Additionally, novel adaptive rules are devised, and the uniform ultimate boundedness stability of the closed-loop system is rigorously validated using the Lyapunov theorem.

Results: Using MATLAB/Simulink software, simulation results are obtained for the proposed control system and its performance is assessed in comparison with related research works across two scenarios. In the first scenario, where both the desired and initial attitude angles are set to zero, the proposed method demonstrates a substantial mean squared error (MSE) reduction: 96.2% for pitch and 86.7% for yaw compared to the backstepping method, and reductions of 75% for pitch and 33.3% for yaw compared to the backstepping sliding mode control. In the second scenario, which involves a 10-degree step input, similar improvements are observed alongside superior performance in terms of reduced overshoot and settling time. Specifically, the proposed method achieves a settling time for the pitch gimbal 56.6% faster than the backstepping method and 58% faster for the yaw gimbal. Moreover, the overshoot for the pitch angle is reduced by 53.5% compared to backstepping and 35.5% compared to backstepping sliding mode control, while for the yaw angle, reductions of 43.6% and 37.6% are achieved, respectively.

Conclusion: Through comprehensive simulation studies, the efficacy of the proposed algorithm is demonstrated, showcasing its superior performance compared to conventional control methods. Specifically, the proposed method exhibits notable improvements in reducing maximum deviation from desired angles, mean squared errors, settling time, and overshoot, outperforming both backstepping and backstepping sliding mode control methods.

Introduction

Unmanned aerial inspection systems face challenges in maintaining optical imaging sensors' direction due to angular disturbances from vehicle motion, wind, and measurement errors. To tackle this, inertially stabilized platforms (ISPs) with gimbal assemblies are commonly employed [1]-[4]. However, the complex dynamics of

multi-axis ISPs, including strong nonlinearity and various uncertainties, make stabilization challenging [5]-[10].

The PID controller stands out as a popular choice in practical applications for its simple implementation. However, its derivative term can exacerbate high-frequency disturbances, leading to rapid saturation of the controller output. To address this, modern modifications are used to automatically adjust control coefficients in different operating conditions, such as using the fuzzy approach in ISPs [11], [12]. Yet, increasing the number of fuzzy partitions complicates controller design and implementation. Thus, the fuzzy controller structure is often kept simple, impacting the PID controller's performance in ISP environments with uncertainties and disturbances.

Researchers employ robust control strategies to tackle uncertainties in nonlinear systems. For example, in [13], authors investigate magnetically suspended gimbals and develop an H_∞ method. Also, [14] proposes a mixed sensitivity H_∞ controller for ISPs, aiming to strike a balance between robustness and performance. These static strategies assume bounded variables, and their complexity increases with uncertainty levels, impacting system efficiency.

Besides, sliding mode control (SMC) methods are favored for severely nonlinear systems like ISPs due to their robustness against uncertainty [15]. As an example, [5] introduces a standard SMC to counter disturbances, while [4] proposes the super-twisting method to address chattering issues. Integral sliding mode control (ISMC), suggested in [16] and [17], further mitigates nonlinear disturbances and uncertainties in ISPs. However, the growing complexity of dynamic models and uncertainties diminishes SMC effectiveness. So, some combine SMC with the backstepping approach, as seen in [18]-[22]. Nevertheless, designing a robust backstepping-based controller for ISPs remains a challenge. For instance, [23] introduces an innovative adaptive neural network model integrated with backstepping control to address the difficulties posed by unknown disturbances and dynamics in nonlinear three-degree-of-freedom (3-DOF) ISPs. ISPs can feature either two or three DOF, each configuration offering distinct benefits. A 2-DOF ISP generally consists of a two-axis gimbal assembly, providing stabilization over two axes such as azimuth and elevation [24]. Conversely, a 3-DOF ISP comprises a three-axis gimbal, enabling stabilization over three axes, thus facilitating more complex motion compensation and greater flexibility in target tracking. The additional degree of freedom in a 3-DOF ISP allows for more comprehensive control over the line of sight and offers enhanced disturbance rejection capabilities, making it suitable for high-accuracy applications [25], [26]. In contrast, a 2-DOF ISP is often more cost-effective and simpler to implement due to its reduced complexity and fewer moving parts [27]. However, it may face limitations in compensating for certain disturbances and body motions, especially in scenarios involving large payloads or significant external disturbances.

To the best of author's knowledge, the most effective strategy for mitigating lumped uncertainties involves using disturbance observers and rejection methods alongside well-designed controllers [23], [28]-[30]. For instance, [5] introduces continuous terminal sliding mode control with high-order sliding mode observers for estimating state variables and uncertainties, while [17] combines terminal sliding mode control with extended state observers in ISPs. Additionally, [31] presents a model predictive control method using a discrete-time disturbance observer, and [4] proposes continuous SMC with finite time disturbance observers. Besides, [32] presents an adaptive SMC algorithm for ISPs using disturbance observers. Also, [33] employs an Uncertainty and Disturbance Estimator (UDE) to estimate the composite disturbance and enhance the robustness of a Feedback Linearization-based controller designed for a 3-DOF known nominal ISP system. These model-based methods enhance control performance by estimating disturbances, although model's accuracy significantly impact their efficacy.

In contrast to model-based disturbance observers, whose performance depends on the system model's accuracy, model-free disturbance observers estimate lumped uncertainties using techniques such as neural networks. For example, [8] proposes an RBFNN-based adaptive disturbance control method for effective uncertainty estimation. However, using a linear system model significantly increases uncertainty levels and impacts observer performance.

Another drawback is the neglect of rate of change in disturbances or estimation error. As well, [22] and [34] combines a backstepping sliding mode control with an adaptive radial basis function neural network estimator to address parametric uncertainties, friction, mass imbalance, and uncertain disturbances. They believe the sampling estimation period can be small in comparison to these variations.

The above studies show that the dynamic model of ISPs is so complex and highly nonlinear that accurate mathematical modeling of all physical aspects is impossible. On the other hand, increasing the complexity of the model makes the control design and its implementation challenging. Therefore, researchers use models that include a variety of uncertainty resources, including parametric and structural uncertainties as well as internal and external disturbances. Controllers that use inaccurate dynamics or do not consider existing

disturbances or those that approximate the uncertainties based on the inaccurate model are doomed to failure in practice. However, the dynamical structure of ISPs is in a particular class of nonlinear systems in which the benefits of backstepping can be used. Despite the advantages of backstepping control, not using precise dynamics and not paying attention to the lumped uncertainties leads to irreparable consequences. Inspired by [35], [36], to tackle these issues, fuzzy approximators and disturbance observers can be coupled to estimate unknown dynamics and disturbances, respectively.

Inspired by the related literature, the main difficulties can be stated as follows:

- The existence of time-varying disturbances and highly nonlinearities in the system model makes it challenging to design a learning-based robust adaptive controller to deal with both of them.
- Considering the system dynamics as unknown terms brings us closer to the real-world applications. Thus, how to design a disturbance observer on the basis of unknown system dynamics is an important issue.
- Achieving robust stability of the overall system in the presence of various uncertainties is a serious issue.

So, this paper introduces a fuzzy disturbance observer-based backstepping control to track desired trajectories amidst uncertainties. It employs a nonlinear model-free disturbance observer to approximate time-varying disturbances, uncertainties, and fuzzy errors. This method enhances system performance by sharing information between the fuzzy approximator and disturbance observer. Besides, stability is verified via recursive Lyapunov-based analysis.

pitch frame (p) with the coordinate system $\{x_p, y_p, z_p\}$,

As clearly observed in Table 1, compared with the current relevant studies, the primary contributions of the suggested controller are as follows:

i) For the ISP systems, this study is the first attempt to handle the various uncertainties, including time-varying disturbances and highly nonlinearities, by employing an adaptive fuzzy-based disturbance observer.

ii) Integrating the fuzzy learning algorithm with the model-free disturbance observer improves the system performance while the system information is not required during the process.

iii) The design mechanism is based on the backstepping method, where the overall system's stability is proven by compensating for the error of each subsystem for the ISP systems. In other words, if an adverse result occurs, it is easier to find its origination.

The rest of this paper is arranged as follows. Section 2 discusses the modeling of a two-axis ISP system, concluding that the resulting state space has a low triangular structure and contains structural and parametric uncertainty. Section 3 uses this structure to develop a controller based on the backstepping approach. However, to improve performance, the nonlinear dynamics of the model are assumed to be unknown and are identified by a fuzzy approximation. Also, a model-free disturbance observer estimates the disturbances and the fuzzy approximation errors. The control signal exploits the approximated dynamics and disturbances, and the stability conditions are expressed. Section 4 presents the simulation results, followed by Section 5's conclusions.

Table 1: Comparative analysis of implementation strategies for ISP systems

Reference	Model Linearity	Model-based/free	DOB-based	Disturbance Type
[3]	nonlinear	model-free	No	time-invariant and bounded
[4]	nonlinear	model-based	No	time-varying
[5], [19]	linear	model-based	Yes	not determined
[16], [29], [31]	linear	model-based	yes	time-varying
[8]	linear	model-free	yes	time-invariant and bounded
[17], [33], [37], [38]	nonlinear	model-based	yes	time-invariant and bounded
[6], [22], [34]	nonlinear	model-free	yes	time-invariant and bounded
proposed approach	nonlinear	model-free	yes	time-varying

Modeling of Two-Axis ISPs

Fig. 1 illustrates a typical two-axis ISP architecture, comprising three frames crucial for coupling analysis: the yaw frame (a) with $\{x_a, y_a, z_a\}$, and the base frame (b) with $\{x_b, y_b, z_b\}$. Notably, a two-axis gyro is incorporated for line-of-sight control, influenced by the

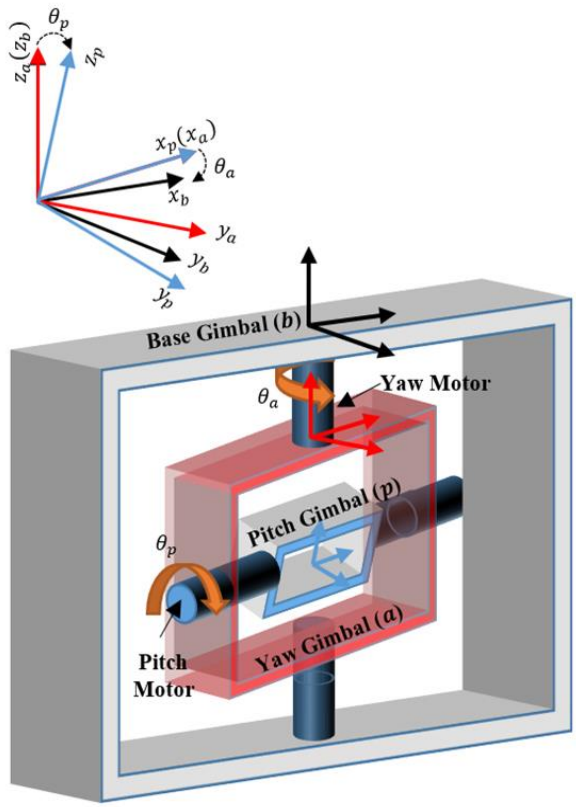


Fig. 1: A typical configuration diagram for two-gimbal ISPs and the related coordinate frames.

movements of the pitch and yaw gimbals.

Encoders installed on the pitch and yaw gimbals measure their angular positions. The relative angular displacement between the base plate and the yaw gimbal is represented by θ_a, while θ_p denotes the relative angular displacement between the yaw and pitch gimbals. Moreover, the angular rates between the gimbal coordinates are denoted as $\dot{\theta}_a$ and $\dot{\theta}_p$. By employing Euler transformation matrices, namely C_b^a and C_a^p, which pertain to rotations about the x-axis and the z-axis respectively, the angular rates of gimbals are specified as,

$$\begin{aligned}\omega_{na}^{a} &= \left[\omega_{nax}^{a}, \omega_{nay}^{a}, \omega_{naz}^{a}\right]^T = C_b^a \omega_{nb}^{b} + \left[0, 0, \dot{\theta}_a\right]^T \\ &= \begin{bmatrix} \omega_{nbx}^{b}\cos\theta_a + \omega_{nby}^{b}\sin\theta_a \\ -\omega_{nbx}^{b}\sin\theta_a + \omega_{nby}^{b}\cos\theta_a \\ \omega_{nbz}^{b} + \dot{\theta}_a \end{bmatrix},\end{aligned} \tag{1}$$

$$\begin{aligned}\omega_{np}^{p} &= \left[\omega_{npx}^{p}, \omega_{npy}^{p}, \omega_{npz}^{p}\right]^T = C_a^p \omega_{na}^{a} + \left[\dot{\theta}_p, 0, 0\right]^T \\ &= \begin{bmatrix} \omega_{nax}^{a} + \dot{\theta}_p \\ \omega_{nay}^{a}\cos\theta_p + \omega_{naz}^{a}\sin\theta_p \\ -\omega_{nay}^{a}\sin\theta_p + \omega_{naz}^{a}\cos\theta_p \end{bmatrix}.\end{aligned} \tag{2}$$

Since the base frame is tied to the helicopter, $\omega_{nb}^{b} = \left[\omega_{nbx}^{b}, \omega_{nby}^{b}, \omega_{nbz}^{b}\right]^T$ is the helicopter's angular velocity.

A. Dynamics of Gimbals

If the gimbals are treated as rigid bodies, their motion equations can be derived using the Newton–Euler theory.

The total external torques are expressed as follows,

$$T_a = \dot{H}_a + \omega_{na}^{a} \times H_a, \qquad T_P = \dot{H}_p + \omega_{np}^{p} \times H_p, \tag{3}$$

where $T_a = \left[T_{ax}, T_{ay}, T_{az}\right]^T$ denotes the total external torque about the yaw gimbal, $T_p = \left[T_{px}, T_{py}, T_{pz}\right]^T$ represents the total external torque applied to the pitch gimbal, and H_a and H_p signify the total angular momentum of the gimbals. Assuming symmetry for each gimbal with respect to its coordinate and neglecting inertia products, the moment of inertia for the two gimbals is defined as,

$$\begin{aligned} J_a &= diag\left(J_{ax}, J_{ay}, J_{az}\right), \\ J_p &= diag\left(J_{px}, J_{py}, J_{pz}\right) \end{aligned} \tag{4}$$

Here, it is assumed that the gimbals have balanced masses. By substituting the angular momentums, angular rates, and total external torques about the yaw gimbal's x-axis and the pitch gimbal's z-axis into (3), the dynamic model of the gimbals is obtained. The pitch gimbal's angular momentum is,

$$H_p = J_p \omega_{np}^{p}. \tag{5}$$

Substituting (5) in (3) leads to the pitch gimbal's angular momentum about the x-axis,

$$T_{px} = J_{px}\, \dot{\omega}_{npx}^{p} + \left(J_{pz} - J_{py}\right) \omega_{npy}^{p}\, \omega_{npz}^{p}. \tag{6}$$

Due to the connection between the pitch and yaw gimbals, the yaw gimbal's inertial angular momentum is,

$$H_a = C_p^a\, H_p + J_a\, \omega_{na}^{a} \tag{7}$$

Thus, by utilizing (3), the projection of the resultant angular momentum along the yaw gimbal's z-axis can be derived as,

$$\begin{aligned} T_{az} = &\left(J_{az}\, \omega_{naz}^{a}\right. \\ &+ \left(J_{py}\cos^2\theta_p + J_{pz}\sin^2\theta_p\right) \omega_{naz}^{a}\Big)' \\ &+ \Big(\left(J_{py} - J_{pz}\right)\cos\theta_p \sin\theta_p\, \omega_{nay}^{a}\Big)' \\ &+ \left(J_{ay} + J_{py}\, \cos^2\theta_p\right. \\ &\left. + J_{pz}\, \sin^2\theta_p\right)\omega_{nax}^{a}\, \omega_{nay}^{a} \\ &- J_{ax}\omega_{nax}^{a}\, \omega_{nay}^{a} - J_{px}\, \omega_{npx}^{p}\, \omega_{nay}^{a} \\ &+ \left(J_{py} - J_{pz}\right)\, \cos\theta_p \sin\theta_p\, \omega_{naz}^{a}\, \omega_{nax}^{a} \end{aligned} \tag{8}$$

where the prime operator signifies the time derivative. To finalize the development of a dynamic model for a two-axis ISP, the subsequent section elaborates on formulating a dynamical model for a DC motor linked to

each gimbal.

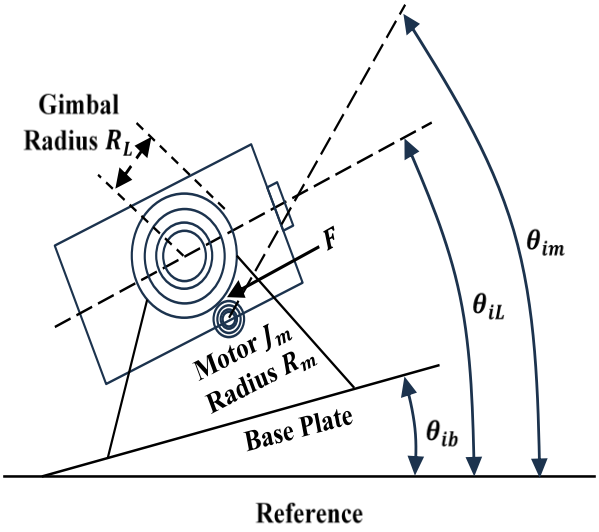


Fig. 2: A single gimbal gear-drive system [3].

B. Dynamical Model of Motor

The high imaging loads necessitate the use of a DC motor with a gearbox, as depicted in Fig. 2, instead of a direct-driven torque motor to ensure stable control [3]. In Fig. 2, F signifies the interacting force between the motor gear and the gimbal gear, with R_m and R_L denoting their respective radii, and where $L \in \{p, a\}$ represents the pitch and yaw gimbals. Furthermore, $N = R_L/r$ stands for the gear ratio.

In a gear-driven system, the torque applied to the gimbal is given by,

$$T_L = R_L F + T_{dL} \tag{9}$$

while the motor's angular acceleration is described by,

$$\ddot{\theta}_{nm} = (K_t\, i_m - rF + T_{dm})/J_m \tag{10}$$

where θ_{nm} is the attitude of the gimbal's motor, K_t denotes the torque constant, and i_m stands for the motor's armature current. Also, external torque perturbations affecting the gimbal and the motor are represented by T_{dL} and T_{dm}, respectively, predominantly reflecting the influences of mass imbalance, bearing friction, and gearing friction.

Considering the electrical characteristics of a DC motor's equivalent armature circuit, the armature voltage u is defined as,

$$u_L = K_e\, \dot{\theta}_{m/b} + R_m\, i_m + L_m\, di_m/dt \tag{11}$$

Here, L_m and R_m signify the motor's inductance and resistance, respectively, while K_e stands for the back electromotive force constant. Moreover, $\theta_{m/b}$ represents the motor's motion relative to the base plate. Additionally, the kinematic relations of the system are as follows,

$$\theta_{nL} = \theta_{L/b} + \theta_{nb} \tag{12}$$

$$\theta_{nm} = \theta_{m/b} + \theta_{nb} \tag{13}$$

$$\theta_{m/b} = N\, \theta_{L/b} \tag{14}$$

In these equations, θ_{nL} and θ_{nb} denote the attitudes of the gimbal and the base with regard to inertial space, respectively, while $\theta_{L/b}$ represents the gimbal's motion relative to the base plate.

By disregarding the negligible value of motor inductance and utilizing (11) to (14), we obtain,

$$T_m = K_t i_m = K_t/R_m \left(u_L - K_e\left(\dot{\theta}_{nm} - \dot{\theta}_{nb}\right)\right) \tag{15}$$

Also, by substituting (15) into (10) and subsequently utilizing (9), T_L is derived as follows,

$$\begin{aligned} T_L = {} & K_t/R_m \left(u_L - K_e\left(\dot{\theta}_{nm} - \dot{\theta}_{nb}\right)\right) \\ & - N^2\, J_m\, \ddot{\theta}_{nL} + (N\, T_{dm} + T_{dL}) \\ & + N(N-1)\, J_m\, \dot{\theta}_{nb} \end{aligned} \tag{16}$$

In (16), accounting for the yaw gimbal, the subscript L corresponds to a, thus $\theta_{nb} = \theta^a_{nbz}$, $\theta_{nL} = \theta^a_{na}$, and $T_L = T_{az}$. Similarly, for the pitch gimbal, we have $\theta_{nb} = \theta^p_{nbz}$, $\theta_{nL} = \theta^p_{np}$, and $T_L = T_{px}$.

Finally, considering (6), (8), and (16), one can conclude the dynamic model of a two-axis ISP as follows,

$$\dot{\omega}^p_{npx} = f_1(t) + b_1 u_1 + d_1 \tag{17}$$

$$\dot{\omega}^a_{naz} = f_2(t) + b_2 u_2 + d_2 \tag{18}$$

in which,

$$u_1 = u_p, \qquad u_2 = u_a \tag{19}$$

$$\begin{aligned} b_1 &= NK_t/\left(R_m\left(J_{px} + N^2\, J_m\right)\right) \\ b_2 &= \frac{NK_t \cos\theta_p}{R_m\left(J_{az} + N^2 J_m + J_{pz}\cos^2\theta_p\right)} \end{aligned} \tag{20}$$

$$\begin{aligned} f_1(t) = {} & \frac{K_t\, K_e\, N^2\, \left(\omega^p_{nax} - \omega^p_{npx}\right)}{\left(J_{px} + N^2\, J_m\right)R_m} \\ & + \frac{N\,(N-1)\, R_m\, J_m\, \dot{\omega}^p_{nax}}{\left(J_{px} + N^2\, J_m\right)R_m} \\ & - \frac{\left(J_{pz} - J_{py}\right)\omega^p_{npy}\omega^p_{npz}}{J_{px} + N^2\, J_m} \end{aligned} \tag{21}$$

$$
\begin{aligned}
& f_2(t) \\
&= \frac{\cos\theta_p\,\omega_{nay}^a\left(\omega_{npx}^p J_{px} - \omega_{nax}^a\left(J_{ay} - J_{ax}\right)\right)}{J_{az} + N^2 J_m + J_{pz}\cos^2\theta_p} \\
&+ \frac{N^2 K_t K_e \cos\theta_p\,(\omega_{nbz}^a - \omega_{naz}^a)}{R_m\left(J_{az} + N^2 J_m + J_{pz}\cos^2\theta_p\right)} \\
&+ \frac{N(N-1)\cos\theta_p\,R_m J_m\,\dot{\omega}_{nbz}^a}{R_m\left(J_{az} + N^2 J_m + J_{pz}\cos^2\theta_p\right)} \\
&+ \frac{J_{pz}\,\omega_{npz}^p \sin 2\theta_p\left(\dot{\theta}_p + \omega_{nax}^a\right)}{2\left(J_{az} + N^2 J_m + J_{pz}\cos^2\theta_p\right)} \\
&- \frac{J_{py}\,\omega_{npy}^p\,\omega_{nax}^a\,\cos^2\theta_p}{J_{az} + N^2 J_m + J_{pz}\cos^2\theta_p} \\
&- \frac{\Psi(t)}{J_{az} + N^2 J_m + J_{pz}\cos^2\theta_p}
\end{aligned}
\tag{22}
$$

$$
\begin{aligned}
\Psi(t) = {} & \cos\theta_p\,\left(J_{py}\,\omega_{npy}^p \sin\theta_p\right)' \\
& + \left(J_{az} + J_{py}\cos^2\theta_p + J_{pz}\sin^2\theta_p\right. \\
& \left. + N^2 J_m\right)\left(\omega_{naz}^a\,\sin\theta_p\,\dot{\theta}_p\right. \\
& \left. + \left(\omega_{nay}^a\,\sin\theta_p\right)'\right)
\end{aligned}
$$

$$
d_1 = \left(N\,T_{dm} + T_{dp} + \Lambda_1\right)/\left(J_{px} + N^2 J_m\right) \tag{23}
$$

$$
d_2 = \frac{N T_{dm} + T_{da} + \Lambda_2}{J_{az} + N^2 J_m + J_{pz}\cos^2\theta_p}\cos\theta_p \tag{24}
$$

where other angular velocities can be calculated using (1) and (2) in terms of the helicopter's angular velocity ω_{nb}^b.

In ISP systems, the moments of inertia (i.e., J_a and J_p) are subject to estimation due to imprecise knowledge. These estimation errors, along with other parametric and structural uncertainties, are accounted for in dynamic equations (17)-(24) by Λ_1 and Λ_2. However, considering all uncertainties in system dynamics, particularly in f_1 and f_2, which are highly complex and nonlinear, can lead to excessive uncertainty levels and consequently poor system performance. Therefore, in the following, only f_1 and f_2 are treated as unknown and approximated using a Takagi-Sugeno fuzzy approximator. In addition, the fuzzy estimate errors in each subsystem, along with terms d_1 and d_2, are regarded as a lumped disturbance. To approximate this lumped disturbance, disturbance observers utilize the approximated nonlinear terms f_1 and f_2.

Hence, by defining the state variables $x_1 = \theta_{nx}^p$, $x_2 = \dot{\theta}_{nx}^p = \omega_{npx}^p$, $x_3 = \theta_{nz}^a$, and $x_4 = \dot{\theta}_{nz}^a = \omega_{naz}^a$, the ISP system's state-space model is obtained as,

$$\dot{x}_1 = x_2 \tag{25}$$

$$\dot{x}_2 = f_1 + b_1 u_1 + d_1 \tag{26}$$

$$\dot{x}_3 = x_4 \tag{27}$$

$$\dot{x}_4 = f_2 + b_2 u_2 + d_2 \tag{28}$$

here, the nonlinear terms f_1 and f_2 are generally unknown, with all parametric and structural uncertainties encapsulated in the d_1 and d_2. This model comprises two subsystems with a low-triangular structure, enabling the utilization of backstepping control in its controller design.

Main Results

This section aims to design an adaptive fuzzy backstepping controller equipped with a model-free disturbance observer for the system (25)-(28). Following the backstepping method, the proposed approach involves four steps. As mentioned, to increase the efficiency of the designed controller, $f_i,\ i \in \{1,2\}$ is considered unknown, which is approximated by a Sugeno-type fuzzy logic system,

$$R^k: IF\ x_1 is\ F_{i1}^k\ and\ \dots and\ x_n is\ F_{in}^k\ Then\ \hat{f}_i = \Gamma_{ik}$$

where $x = (x_1, \dots, x_n)^T$ is the inputs of fuzzy system, $k = 1, 2, \dots, l$ is the rule number and l is the number of rules, F_{i1}^k is the fuzzy set with the membership function $\mu_{F_{ij}^k}(x_j)$, and Γ_{ik} is a constant value. Using singleton fuzzifier, product inference and weighted average defuzzification, we obtain [39],

$$\hat{f}_i = \Gamma_i^T \phi_i(x)$$

where $\phi_i(x) = [\phi_{i1}(x), \phi_{i2}(x), \dots, \phi_{il}(x)]^T$, and $\phi_{ik}(x) = \prod_{j=1}^n \mu_{F_{ij}^k}(x_j) / \sum_{m=1}^l \prod_{j=1}^n \mu_{F_{ij}^m}(x_j)$ is the membership function of the j^{th} rule's antecedent part. Besides, $\Gamma_i^T = [\Gamma_{i1}, \Gamma_{i2}, \dots, \Gamma_{il}]^T$ is the fuzzy weight vector. Considering the fuzzy approximation error ϵ_i,

$$f_i(x) = \Gamma_i^{*T} \phi_i(x) + \epsilon_i \tag{29}$$

where $\Gamma_i^* = [\Gamma_{i1}^*, \Gamma_{i2}^*, \dots, \Gamma_{il}^*]^T$is the optimal fuzzy weight vector. Since f_i and its corresponding optimal fuzzy weight vector Γ_i^* are unknown, their estimations is used in the control signals. In other words, $\hat{f}_i = \hat{\Gamma}_i^{*T} \phi_i$ is employed in the proposed control signals, in which $\hat{\Gamma}_i^*$ is the estimated fuzzy weight vector such that the error of the fuzzy weight vector $\tilde{\Gamma}_i = \Gamma_i^* - \hat{\Gamma}_i$ should converge to zero ultimately. However, the fuzzy approximation error ϵ_i and the uncertainties $g_i d_i$ are considered lumped

disturbance $D_i = d_i + \epsilon_i$ which is observed using a disturbance observer $\widehat{D}_i$. This leads to enhancing the robust behavior of the control system. It is assumed that positive real scalars bound the lumped disturbance D_i and its variation. In other words, we assume $\dot{D}_i^T \dot{D}_i \leq \zeta_i^2$, where $\zeta_i > 0$ is a known positive real scalar. Here, the conservative assumption of zero variation is not considered. Furthermore, since the input membership functions $\phi_{ij}(x_i)$, $i \in \{1,2\}$ and $j \in \{1,2,\dots,l\}$ are known, it is clear that $\phi_i^T \phi_i \leq \eta_i^2$ where $\eta_i > 0$ is known. More details are given in the following.

Step 1) For the first subsystem (25) the tracking error is defined as,

$$\omega_1 = x_1 - x_1^d \tag{29}$$

in which x_1^d is the desired reference signal. We choose the first virtual control,

$$\beta_1 = -k_1\omega_1 + \dot{x}_1^d \tag{30}$$

where k_1 is a positive real constant. For the 2nd subsystem (26), the error surface $\omega_2 = x_2 - \beta_1$ is defined. Thus, the derivative of the tracking error ω_1 is as follows,

$$\dot{\omega}_1 = \dot{x}_1 - \dot{x}_1^d = x_2 - \dot{x}_1^d = \omega_2 + \beta_1 - \dot{x}_1^d \tag{31}$$

Substituting (30) into (31) yields,

$$\dot{\omega}_1 = -k_1\omega_1 + \omega_2 \tag{32}$$

Now, choosing the Lyapunov function $V_1 = \frac{1}{2}\omega_1^2$ and considering (32) leads us to,

$$\dot{V}_1 = -k_1\omega_1^2 + \omega_1\omega_2 \tag{33}$$

Step 2) The derivative of the 2nd subsystem's error surface is,

$$\dot{\omega}_2 = \dot{x}_2 - \dot{\beta}_1 = \Gamma_1^{*T}\phi_1 + b_1u_1 - \dot{\beta}_1 + D_1 \tag{34}$$

where $D_1 = d_1 + \epsilon_1$ is the lumped disturbance. Considering the control signal,

$$u_1 = \left(-\hat{\Gamma}_1^{\mathrm{T}}\phi_1 - k_2\omega_2 - \omega_1 + \dot{\beta}_1 - \widehat{D}_1\right)/b_1 \tag{35}$$

in which $k_2 > 0$ is a real scalar and $\widehat{D}_1$ is the lumped disturbance's estimation, and then substituting it into (34) gives,

$$\dot{\omega}_2 = \tilde{\Gamma}_1^T\phi_1 + \tilde{D}_1 - \omega_1 - k_2\omega_2 \tag{36}$$

In this step, the Lyapunov function is chosen as,

$$V_2 = V_1 + \frac{1}{2}\omega_2^2 + \frac{1}{2}\tilde{\Gamma}_1^T\gamma_1^{-1}\tilde{\Gamma}_1 + \frac{1}{2}\tilde{D}_1^2 \tag{37}$$

where $\gamma_1 > 0$ is the learning rate of adaptation mechanism. Hence, one can obtain,

$$\begin{aligned}\dot{V}_2 = &-k_1\omega_1^2 - k_2\omega_2^2 + \tilde{\Gamma}^T{}_1\left(\omega_2\phi_1 - \gamma_1^{-1}\dot{\hat{\Gamma}}_1\right)\\ &+ \tilde{D}_1\left(\omega_2 + \dot{D}_1 - \dot{\hat{D}}_1\right)\end{aligned} \tag{38}$$

Using the fuzzy approximation for f_1, we define the following model-free disturbance observer,

$$\begin{aligned}\widehat{D}_1 &= L_1(x_2 - \chi_2)\\ \dot{\chi}_2 &= \hat{\Gamma}_1^{\mathrm{T}}\phi_1 + b_1u_1 + \widehat{D}_1 - L_1^{-1}\omega_2\end{aligned} \tag{39}$$

in which L_1 is a positive real constant; and the derivative of $\widehat{D}_1$ is,

$$\dot{\hat{D}}_1 = L_1(\dot{x}_2 - \dot{\chi}_2) = L_1\left(\tilde{\Gamma}_1^T\phi_1 + \tilde{D}_1\right) + \omega_2 \tag{40}$$

Employing (40) in (38),

$$\begin{aligned}\dot{V}_2 = &-k_1\omega_1^2 - k_2\omega_2^2 + \tilde{\Gamma}_1^T\left(\omega_2\phi_1 - \gamma_1^{-1}\dot{\hat{\Gamma}}_1\right)\\ &+ \tilde{D}_1\left(\dot{D}_1 - L_1\left(\tilde{\Gamma}_1^T\phi_1 + \tilde{D}_1\right)\right)\end{aligned} \tag{41}$$

Now, we choose the first adaptation law,

$$\dot{\hat{\Gamma}}_1 = \gamma_1\left(\omega_2\phi_1 - \delta_1\hat{\Gamma}_1\right) \tag{42}$$

where $\delta_1 > 0$ is a real scalar. Using Young inequality, one can find,

$$\begin{aligned}\tilde{D}_1\dot{D}_1 &\leq \frac{1}{2}\tilde{D}_1^2 + \frac{1}{2}\zeta_1^2\\ -\tilde{D}_1\tilde{\Gamma}_1^T\phi_1 &\leq \frac{1}{2}\omega_1\tilde{D}_1^2\eta_1^2 + \frac{1}{2\omega_1}\tilde{\Gamma}_1^T\tilde{\Gamma}_1\\ \tilde{\Gamma}_1^T\hat{\Gamma}_1 &\leq -\frac{1}{2}\tilde{\Gamma}_1^T\tilde{\Gamma}_1 + \frac{1}{2}\|\Gamma_1^*\|^2\end{aligned} \tag{43}$$

in which $\omega_1 > 0$. Then, $\dot{V}_2$ is obtained as,

$$\begin{aligned}\dot{V}_2 \leq &-k_1\omega_1^2 - k_2\omega_2^2 - \left(\frac{\delta_1}{2} - \frac{L_1}{2\omega_1}\right)\tilde{\Gamma}_1^T\tilde{\Gamma}_1\\ &- \left(L_1 - \frac{L_1\omega_1}{2}\eta_1^2 - \frac{1}{2}\right)\tilde{D}_1^2\\ &+ \left(\frac{\delta_1}{2}\|\Gamma_1^*\|^2 + \frac{1}{2}\zeta_1^2\right)\end{aligned} \tag{44}$$

Step 3) Considering the desired trajectory x_3^d for the state variable x_3 and define the error surface $\omega_3 = x_3 - x_3^d$, we have,

$$\dot{\omega}_3 = \dot{x}_3 - \dot{x}_3^d = x_4 - \dot{x}_3^d = \omega_4 + \beta_2 - \dot{x}_3^d \tag{45}$$

Another virtual control law is constructed as $\beta_2 = -k_3\omega_3 + \dot{x}_3^d$, where k_3 is a positive constant. This is then substituted in (17), resulting in,

$$\dot{\omega}_3 = -k_3\omega_3 + \omega_4 \tag{46}$$

Considering the Lyapunov function $V_3 = V_2 + \frac{1}{2}\omega_3^2$ as well as (44) and (46), one can obtain,

$$\begin{aligned}\dot{V}_3 \leq &-k_1\omega_1^2 - k_2\omega_2^2 - k_3\omega_3^2 + \omega_3\omega_4\\ &- \left(\frac{\delta_1}{2} - \frac{L_1}{2\omega_1}\right)\tilde{\Gamma}_1{}^T\tilde{\Gamma}_1\\ &- \left(L_1 - \frac{L_1\omega_1}{2}\eta_1^2 - \frac{1}{2}\right)\tilde{D}_1^2\\ &+ \left(\frac{\delta_1}{2}\|\Gamma_1^*\|^2 + \frac{1}{2}\zeta_1^2\right)\end{aligned} \tag{47}$$

Step 4) The error surface for the last subsystem (28) is defined as $\omega_4 = x_4 - \beta_2$. Hence, its derivative is,

$$\dot{\omega}_4 = \Gamma_2^{*T}\phi_2 + b_2 u_2 + D_2 - \dot{\beta}_2 \tag{48}$$

in which $D_2 = \epsilon_2 + d_2$ is the total disturbance include the fuzzy approximation error ϵ_2 for the term f_2 and the other modeling uncertainties d_2. The second control law is proposed as,

$$u_2 = \left(-\hat{\Gamma}_2^{\mathrm{T}}\phi_2 - k_4\omega_4 - \omega_3 + \dot{\beta}_2 - \hat{D}_2\right)/b_2 \tag{49}$$

where $k_4 > 0$ and $\hat{D}_2$ is the estimation of D_2. Substituting (49) into (48) results in,

$$\dot{\omega}_4 = \hat{\Gamma}_2^{\mathrm{T}}\phi_2 + \tilde{D}_2 - k_4\omega_4 - \omega_3 \tag{50}$$

The final Lyapunov function is chosen as,

$$V_4 = V_3 + \frac{1}{2}\omega_4^2 + \frac{1}{2}\tilde{\Gamma}_2^T\gamma_2^{-1}\tilde{\Gamma}_2 + \frac{1}{2}\tilde{D}_2^2 \tag{51}$$

where $\gamma_2 > 0$. The derivative of V_4 can be obtained as,

$$\begin{aligned}\dot{V}_4 \le\ & -k_1\omega_1^2 - k_2\omega_2^2 - k_3\omega_3^2 - k_4\omega_4^2 \\ & + \tilde{\Gamma}_2^T\left(\omega_4\phi_2 - \gamma_2^{-1}\dot{\hat{\Gamma}}_2\right) \\ & + \tilde{D}_2\left(\omega_4 + \dot{D}_2 - \dot{\hat{D}}_2\right) - \left(\frac{\delta_1}{2} - \frac{L_1}{2\omega_1}\right)\tilde{\Gamma}_1^{\ T}\tilde{\Gamma}_1 \\ & - \left(L_1 - \frac{L_1\omega_1}{2}\eta_1^2 - \frac{1}{2}\right)\tilde{D}_1^2 \\ & + \left(\frac{\delta_1}{2}\|\Gamma_1^*\|^2 + \frac{1}{2}\zeta_1^2\right)\end{aligned} \tag{52}$$

Similar to the step 2, the following model-free disturbance observer is designed,

$$\begin{aligned}\hat{D}_2 &= L_2(x_4 - \chi_4) \\ \dot{\chi}_4 &= \hat{\Gamma}_2^T\phi_2 + b_2u_2 + \hat{D}_2 - L_2^{-1}\omega_4\end{aligned} \tag{53}$$

in which L_2 is a positive real constant. So,

$$\dot{\hat{D}}_2 = L_2(\dot{x}_4 - \dot{\chi}_4) = L_2\left(\tilde{\Gamma}_2^T\phi_2 + \tilde{D}_2\right) + \omega_4 \tag{54}$$

Considering the adaptation law,

$$\dot{\hat{\Gamma}}_2 = \gamma_2\left(\omega_4\phi_2 - \delta_2\hat{\Gamma}_2\right) \tag{55}$$

where $\delta_2 > 0$ and the following inequalities that obtained by Young inequality lemma,

$$\begin{aligned}\tilde{D}_2\dot{D}_2 &\le \frac{1}{2}\tilde{D}_2^2 + \frac{1}{2}\zeta_2^2 \\ -\tilde{D}_2\tilde{\Gamma}_2^T\phi_2 &\le \frac{1}{2}\omega_2\tilde{D}_2^2\eta_2^2 + \frac{1}{2\omega_2}\tilde{\Gamma}_2^T\tilde{\Gamma}_2 \\ \tilde{\Gamma}_2^T\hat{\Gamma}_2 &\le -\frac{1}{2}\tilde{\Gamma}_2^T\tilde{\Gamma}_2 + \frac{1}{2}\|\Gamma_2^*\|^2\end{aligned} \tag{56}$$

in which $\omega_2 > 0$, Then, we have from (52),

$$\begin{aligned}\dot{V}_4 \le\ & -k_1\omega_1^2 - k_2\omega_2^2 - k_3\omega_3^2 - k_4\omega_4^2 \\ & - \left(\frac{\delta_1}{2} - \frac{L_1}{2\omega_1}\right)\tilde{\Gamma}_1^T\tilde{\Gamma}_1 \\ & - \left(L_1 - \frac{L_1\omega_1}{2}\eta_1^2 - \frac{1}{2}\right)\tilde{D}_1^2 \\ & + \left(\frac{\delta_1}{2}\|\Gamma_1^*\|^2 + \frac{1}{2}\zeta_1^2\right) - \left(\frac{\delta_2}{2} - \frac{L_1}{2\omega_2}\right)\tilde{\Gamma}_2^T\tilde{\Gamma}_2 \\ & - \left(L_2 - \frac{L_2\omega_2}{2}\eta_2^2 - \frac{1}{2}\right)\tilde{D}_2^2 \\ & + \left(\frac{\delta_2}{2}\|\Gamma_2^*\|^2 + \frac{1}{2}\zeta_2^2\right)\end{aligned} \tag{57}$$

Expressing equation (57) as $\dot{V}_4 \le -AV_4 + B$, where $A = \min\{k_i, \bar{\gamma}_j, \bar{D}_j\}$, $i \in \{1, \dots, 4\}$, $j \in \{1, 2\}$, and $B = \left(\frac{\delta_1}{2}\|\Gamma_1^*\|^2 + \frac{1}{2}\zeta_1^2 + \frac{\delta_2}{2}\|\Gamma_2^*\|^2 + \frac{1}{2}\zeta_2^2\right)$, it is evident that the overall system (25)-(28) exhibits uniformly ultimate boundedness (UUB) stability, ensuring that the signals involved in V_4 remain bounded.

For a clearer understanding of the proposed approach, a block diagram illustrating the structure of the model-free control scheme is provided in Fig. 3.

Simulation Results

The efficacy of the suggested technique is evaluated by applying it to a model with parameters outlined in [3] and [6] (see Table 2) and compares with sliding mode control and backstepping sliding mode. While various types of uncertainties, including motor specs, mass imbalance, friction torque, and wind disturbance, are considered, simulation conditions for all three control methods are identical.

This ensures that the noise and disturbances applied to the system in the simulation of the proposed method exactly match those applied during the simulation of the other two methods.

In typical scenarios, the load is often not centered at the gimbals' rotation center, leading to mass imbalance torque [40]. Additionally, bearing friction introduces a nonlinear torque acting as a disturbance. To simulate the influence of these disturbances, we consider the following torque expression,

$$T_{dL} = 1.7(rnd - 0.5)(\sin(\omega t) + \sin(2\omega t)), \tag{58}$$

where $L \in \{p, a\}$ represents pitch or yaw, respectively. Furthermore, to explore the nonlinear effects of gearing friction and other disturbances on the electric motors' torque, we propose,

$$T_{dm} = 0.04(\sin(\omega t) + \sin(2\omega t)) \tag{59}$$

Moreover, fluctuations in the gimbals' moment of inertia are set to be 20% of the nominal moment value. Thus, J_p and J_a can be represented as,

$$J_L = J_L^{nominal}\big(1 + 0.4\,(rnd - 0.5)\big), \qquad (60)$$

where $L \in \{p, a\}$. Simultaneously, random attitude perturbations of the helicopter's stationary base plate caused by wind disturbances are represented as,

$$\omega_{nb\tau}^{b}(t) = 0.6\,(rnd - 0.5), \qquad \tau \in \{x, y, z\} \qquad (61)$$

In the first scenario, with both the desired and initial attitude angles set to zero, Fig. 4 illustrates the attitude angles of pitch and yaw gimbals under the influence of the three control techniques.

Notably, the proposed control method exhibits significantly lesser deviation from the desired angles compared to both the backstepping control method and the backstepping sliding mode control method for both gimbals. This superiority is further evidenced by the mean squared errors presented in Table 3.

Table 3: Nominal values for the parameters of model

Parameter	Nominal Value	Unit
N	50	
J_m	2.7×10^{-4}	$Kg.m^2$
K_t	0.143	Nm/Amp
K_e	0.143	$V\,sec/rad$
R_m	7.56	Ω
$[J_{ax}, J_{ay}, J_{az}]^T$	$[0.540, 0.475, 0.162]^T$	$Kg.m^2$
$[J_{px}, J_{py}, J_{pz}]^T$	$[0.460, 0.267, 0.200]^T$	$Kg.m^2$
$[\omega_{nbx}^b, \omega_{nby}^b, \omega_{nbz}^b]^T$	$[0, 0, 0]^T$	rad/sec

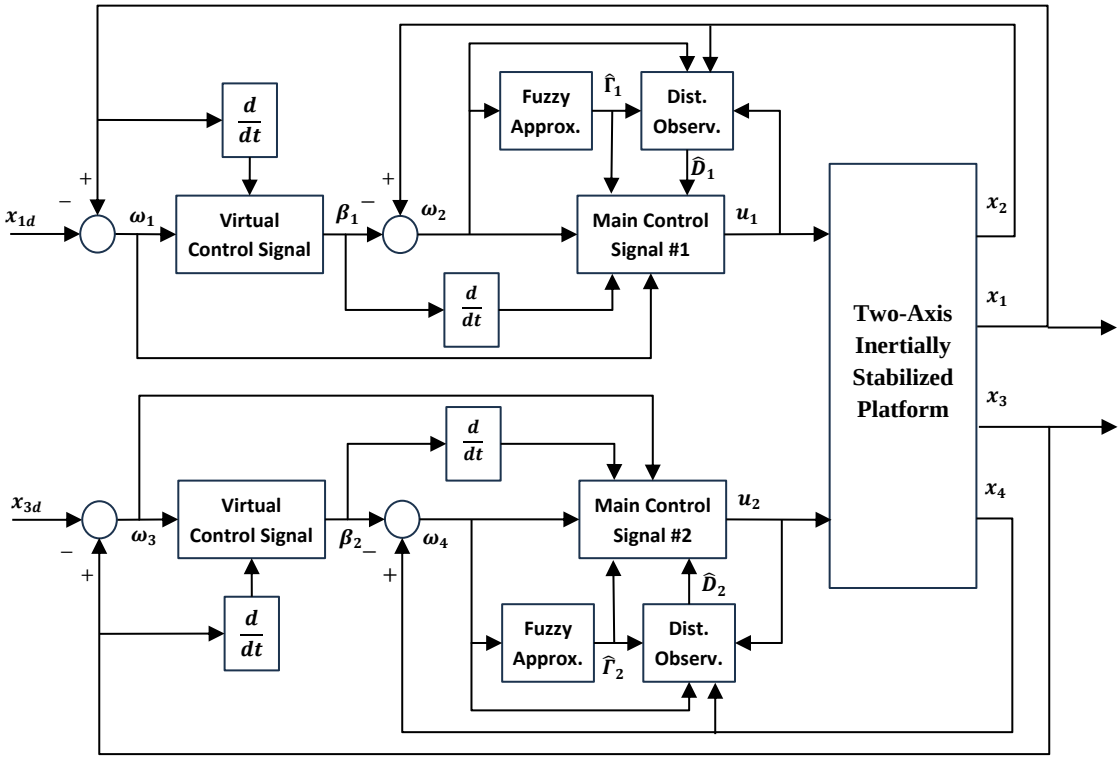


Fig. 3: Block Diagram of the proposed model-free control scheme for two-axis ISPs.

Table 2: Quantitative comparison of steady-state response in different scenarios for the evaluated control approaches

		Attitude angle of Pitch gimbal (deg.)			Attitude angle of Yaw gimbal (deg.)		
Scenario	**Approach**	$\max(\lvert e\rvert)$	$mse(e)$	$std(e)$	$\max(\lvert e\rvert)$	$mse(e)$	$std(e)$
	Backstepping	0.1535	0.0026	0.0507	0.1340	0.0015	0.0390
First	Backstepping SMC	0.0807	0.0004	0.0195	0.0759	0.0003	0.0175
	Proposed Method	0.0404	0.0001	0.0118	0.0474	0.0002	0.0136
	Backstepping	14.2217	5.3999	2.2386	11.2344	3.0565	1.6932
Second	Backstepping SMC	10.2482	1.5565	1.2276	10.1431	1.2727	1.1060
	Proposed Method	9.9973	0.4145	0.6341	10.0034	0.4096	0.6303

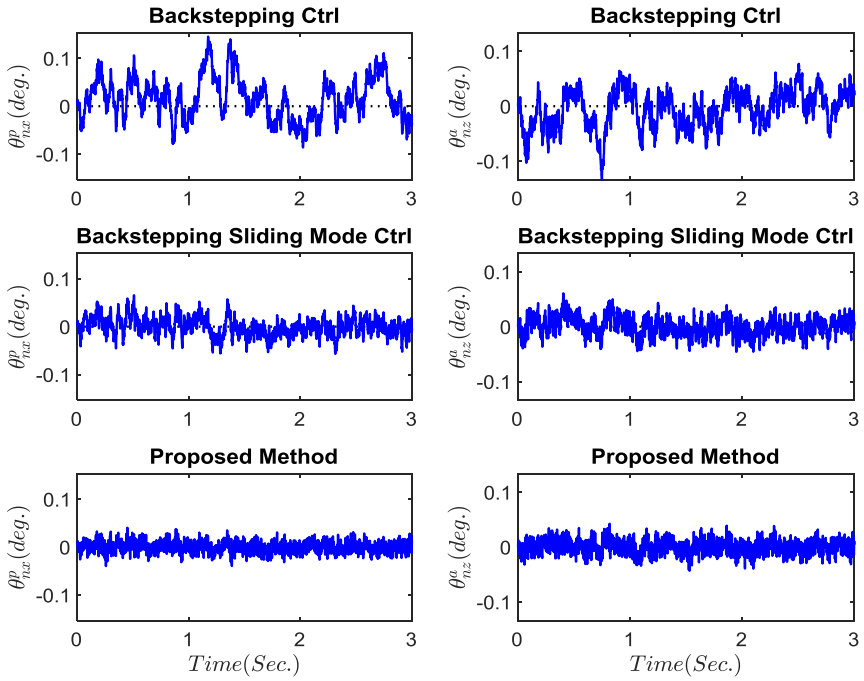


Fig. 4: Comparison of the steady-state response for both pitch and yaw gimbals in different control methods.

Although the proposed method performs best, however, the better performance of the backstepping sliding mode control compared to the pure backstepping technique is not far from expected because the sliding mode method is insensitive to parameter variations and external disturbances [6].

In the second scenario, a step of 10 degrees is introduced to the attitude angles of the yaw and pitch gimbals to assess the dynamic response characteristics of the control methods. Simulation results pertaining to this scenario are presented in Fig. 5 and Table 3, where once again, the proposed method outperforms the others. In terms of dynamic response, the proposed method settles the pitch gimbal in 0.1953 seconds, which is significantly faster than the pure backstepping control (0.4497 seconds) and the backstepping sliding mode method (0.4595 seconds). Similarly, the overshoot in the proposed method is 76% and 36% less than in pure backstepping and backstepping sliding mode control, respectively. Further details regarding the dynamic response characteristics of the investigated methods for both yaw and pitch gimbals are summarized in Table 4.

Besides, to evaluate the control effort exerted by the proposed method and assess its effectiveness, we compare the control effort u_2 for the Yaw gimbal in Fig. 6. It is evident that, generally, there is not a significant difference in control effort required. However, at the moment of angle change (1.5 seconds), the control effort in SMC-based methods is notably lower compared to the pure backstepping approach. Furthermore, the performance of the proposed method in terms of lamped disturbance tracking is depicted in Fig. 7, demonstrating the effective capability of the proposed approach to simultaneously estimate disturbances and uncertainties present in the system.

Table 4: Quantitative comparison of the transient response caused by using the control methods in pitch and yaw gimbals

Approach	Settling Time (Sec.)	Peak (deg.)	Overshoot (%)
Values for pitch gimbal evaluation			
Backstepping	0.4497	24.2217	142.2168
Backstepping SMC	0.4595	20.2482	102.4815
Proposed Method	0.1953	16.6126	66.1255
Values for yaw gimbal evaluation			
Backstepping	0.4438	21.2344	112.3436
Backstepping SMC	0.4533	20.1431	101.4314
Proposed Method	0.1875	16.3361	63.3610

Since fuzzy systems and disturbance observers are sharing information with each other, one cannot precisely determine whether fuzzy estimator can approximate d_2. Hence, the aim of the proposed controller is achieved in view of the estimation task, if the estimation can track the lumped uncertainty with high precision. This job is confirmed by Fig. 7.

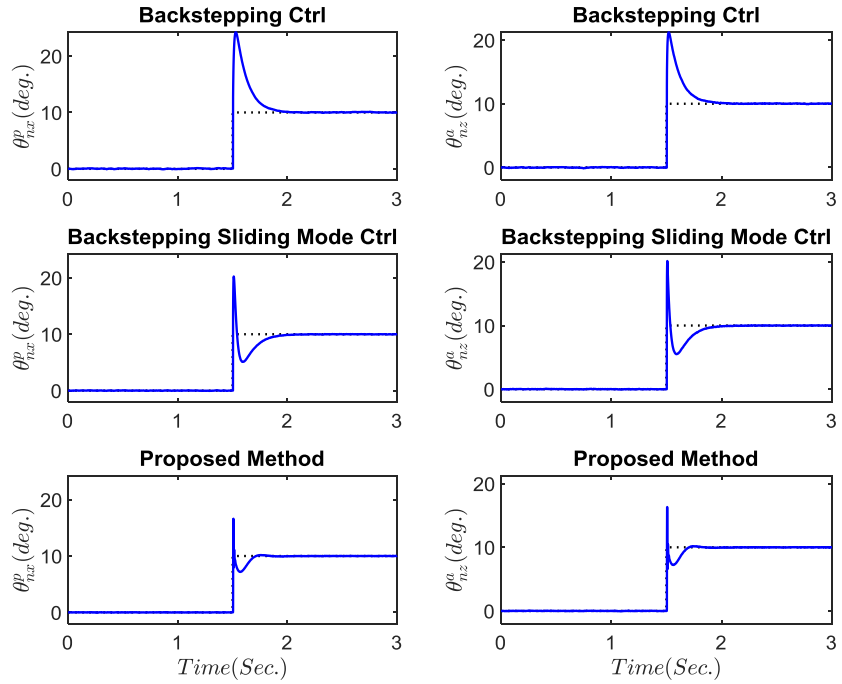


Fig. 5: Comparing transient response to a 10-degree step change in pitch and yaw gimbals attitudes.

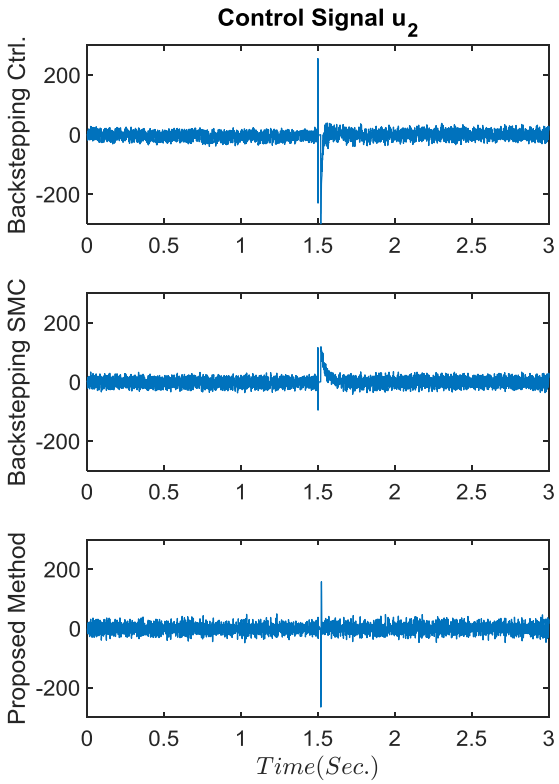


Fig. 6: Comparison of control signals for yaw gimbal in the second scenario.

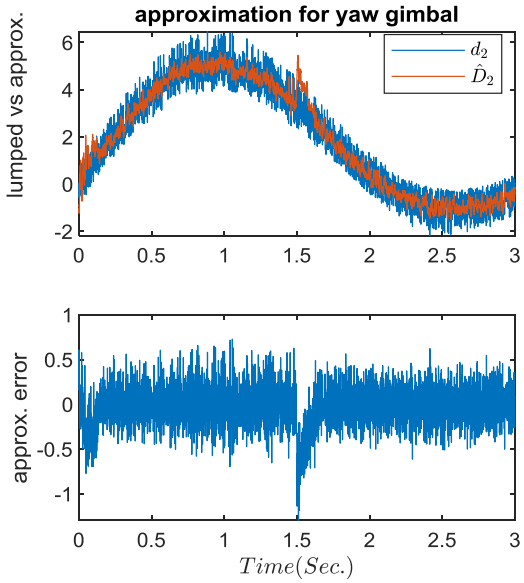


Fig. 7: Performance of the disturbance estimator for lumped disturbances of yaw gimbal in Scenario 2.

Conclusion

In order to enhance the control performance of the ISP system, a disturbance observer-based adaptive fuzzy backstepping controller is developed in this paper. Integrated with a model-free disturbance observer, it ensures high-performance control in uncertain environments. In addition, the stabilization control with high accuracy is also provided in the presence of various uncertainties. The recursive Lyapunov-based analysis confirms the uniformly ultimate boundedness stability of the overall system.Different simulations and comparisons with two relevant control techniques, namely the backstepping control and the backstepping sliding mode control, demonstrate the proposed controller's superiority in the perspective of the transient

response and the steady-state response. Inspired by the current study, we will present an adaptive constrained model-free fault-tolerant control scheme for the ISP system in the future.

Author Contributions

In the current study, the roles of each individual were as follows: M. Ghalehnoie, J. Keighobadi, and A. Azhdari conducted the experimental design. Software development and simulations were carried out by M. Ghalehnoie and A. Azhdari. Following this phase, M. Ghalehnoie collected and analyzed the data. The initial drafting of the manuscript was undertaken collaboratively by M. Ghalehnoie, A. Azhdari, and J. Keighobadi. Subsequently, M. Ghalehnoie and J. Keighobadi provided critical revision of the manuscript. Additionally, supervision of the research execution was provided by M. Ghalehnoie.

Acknowledgment

The authors would like to express their sincere gratitude to the reviewers and editors of JECEI for their constructive comments and valuable suggestions, which have significantly enhanced the quality of this article. Additionally, the authors are thankful to the editorial board for their support and professional handling of the manuscript throughout the review process.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

DOB	Disturbance Observer
ISMC	Integral Sliding Mode Control
ISP	Inertially Stabilized Platform
PID	Proportional Integral Derivative
SMC	Sliding Mode Control

References

[1] J. M. Hilkert, "Inertially stabilized platform technology Concepts and principles," IEEE Control Syst, 28(1): 26-46, 2008.

[2] M. K. Masten, "Inertially stabilized platforms for optical imaging systems," IEEE Control Syst, 28(1): 47-64, 2008.

[3] X. Lei, Y. Zou, F. Dong, "A composite control method based on the adaptive RBFNN feedback control and the ESO for two-axis inertially stabilized platforms," ISA Trans, 59: 424-433, 2015.

[4] J. Mao, S. Li, Q. Li, J. Yang, "Design and implementation of continuous finite-time sliding mode control for 2-DOF inertially stabilized platform subject to multiple disturbances," ISA Trans, 84: 214-224, 2019.

[5] J. Mao, J. Yang, X. Liu, S. Li, Q. Li, "Modeling and robust continuous TSM control for an inertially stabilized platform with couplings," IEEE Trans. Control Syst. Technol., 28(6): 2548-2555, 2020.

[6] F. Dong, X. Lei, W. Chou, "A dynamic model and control method for a Two-Axis inertially stabilized platform," IEEE Trans. Ind. Electron., 64(1): 432-439, 2017.

[7] X. Zhou, H. Zhang, R. Yu, "Decoupling control for two-axis inertially stabilized platform based on an inverse system and internal model control," Mechatronics, 24(8): 1203-1213, 2014.

[8] Q. Mu, G. Liu, X. Lei, "A RBFNN-Based adaptive disturbance compensation approach applied to magnetic suspension inertially stabilized platform," Math. Probl. Eng., 2014: 1-9, 2014.

[9] X. Zhou, G. Gong, J. Li, H. Zhang, R. Yu, "Decoupling control for a three-axis inertially stabilized platform used for aerial remote sensing," Trans. Inst. Meas. Control, 37(9): 1135-1145, 2015.

[10] S. Liu, H. Che, L. Sun, "Research on stabilizing and tracking control system of tracking and sighting pod," J. Control Theory Appl., 10(1): 107-112, 2012.

[11] F. Liu, H. Wang, "Fuzzy PID tracking controller for two-axis airborne optoelectronic stabilized platform," Int. J. Innovative Computi. Inf. Control, 13(4): 1307-1322, 2017.

[12] N. Ghaeminezhad, W. Daobo, F. Farooq, "Stabilizing a gimbal platform using self-tuning fuzzy PID controller," Int. J. Comput. Appl., 93(16): 13-19, 2014.

[13] Q. Guo, G. Liu, B. Xiang, T. Wen, H. Liu, "Robust control of magnetically suspended gimbals in inertial stabilized platform with wide load range," Mechatronics, 39: 127-135, 2016.

[14] S. Hong, K. D. Cho, C. H. Park, W. S. Kang, "Trajectory generation and H∞ robust control for inertially stabilized system," in Proc. 2011 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM): 695-700, 2011.

[15] A. Toloei, H. Asgari, "Quaternion-based finite-time sliding mode controller design for attitude tracking of a rigid spacecraft during high-thrust orbital maneuver in the presence of disturbance torques," Int. J. Eng., 32(3): 430-437, 2019.

[16] T. Wen, B. Xiang, W. Wong, "Coupling analysis and cross-feedback control of three-axis inertially stabilized platform with an active magnetic bearing system," Shock Vibr., 2020: 1-17, 2020.

[17] X. Zhou, Y. Shi, L. Li, R. Yu, L. Zhao, "A high precision compound control scheme based on non-singular terminal sliding mode and extended state observer for an aerial inertially stabilized platform," Int. J. Control Autom. Syst., 18(6): 1498–1509, 2020.

[18] H.-C. Park, S. Chakir, Y. B. Kim, T. Huynh, "a nonlinear backstepping controller design for high-precision tracking applications with input-delay gimbal systems," J. Mar. Sci. Eng., 9(5): 530, 2021.

[19] J. Deng, W. Xue, X. Zhou, Y. Mao, "On disturbance rejection control for inertial stabilization of long-distance laser positioning with movable platform," Meas. Control, 53(7–8): 1203-1217, 2020.

[20] Y. Wang, H. Lei, J. Ye, X. Bu, "Backstepping sliding mode control for radar seeker servo system considering guidance and control system," Sensors, 18(9): 2927, 2018.

[21] R. Yazdanpanah, J. Soltani, "Robust backstepping control of induction motor drives using artificial neural networks and sliding mode flux observers," Int. J. Eng., 20(3): 221-232, 2007.

[22] M. M. Zohrei, A. Roosta, "Constrained adaptive backstepping sliding mode control for inertial stable platform," Iran. J. Sci. Technol. Trans. Electr. Eng., 46(3): 753-764, 2022.

[23] M. M. Zohrei, A. Roosta, B. Safarinejadian, "Robust backstepping control based on neural network stochastic constrained for three axes inertial stable platform," J. Aerosp. Eng., 35(1): 2022.

[24] I. S. Azzam, A. G. Wassal, S. A. Maged, "Line of sight control strategies for inertial stabilization platforms: comparative study," in Proc. 2021 16th International Conference on Computer Engineering and Systems (ICCES): 1-7, 2021.

[25] M. F. Reis, J. C. Monteiro, R. R. Costa, A. C. Leite, "Super-twisting control with quaternion feedback for a 3-DoF inertial stabilization platform," in Proc. 2018 IEEE Conference on Decision and Control (CDC): 2193-2198, 2018.

[26] S. Dey, T. K. Sunil Kumar, S. Ashok, S. K. Shome, "Robust cascade control strategy for trajectory tracking to decouple disturbances using 3-degree-of-freedom inertial stabilized platform and its experimental validation," Trans. Inst. Meas. Control, 2023.

[27] H. Khodadadi, M. R. J. Motlagh, M. Gorji, "Robust control and modeling a 2-DOF Inertial Stabilized Platform," in Proc. International Conference on Electrical, Control and Computer Engineering 2011 (InECCE): 223-228, 2011.

[28] A. Assoud, A. V. Polynkov, "Improving the stabilization accuracy of a platform using active disturbance rejection control and field-oriented control," in AIP Conference Proceeding, 2549(1), 2023.

[29] F. Wang, R. Wang, E. Liu, W. Zhang, "Stabilization control mothed for two-axis inertially stabilized platform based on active disturbance rejection control with noise reduction disturbance observer," IEEE Access, 7: 99521-99529, 2019.

[30] S. Asgari, M. B. Menhaj, A. A. Suratgar, M. G. Kazemi, "A disturbance observer based fuzzy feedforward proportional integral load frequency control of microgrids," Int. J. Eng., 34(7): 1694-1702, 2021.

[31] X. Liu, J. Mao, J. Yang, S. Li, K. Yang, "Robust predictive visual servoing control for an inertially stabilized platform with uncertain kinematics," ISA Trans, 114: 347-358, 2021.

[32] D. Tian, M. Wang, F. Wang, R. Xu, "Adaptive sliding-mode-assisted disturbance observer-based decoupling control for inertially stabilized platforms with a spherical mechanism," IET Control Theory Appl., 16(12): 1194-1207, 2022.

[33] A. Kodhanda, J. P. Kolhe, M. M. Kuber, V. V. Parlikar, "Uncertainty and disturbance estimation based control of three-axis stabilized platform," Int. J. Latest Trends Eng. Technol., 3(3): 289-297, 2014.

[34] Z. Ding, F. Zhao, Y. Lang, Z. Jiang, J. Zhu, "Anti-disturbance neural-sliding mode control for inertially stabilized platform with actuator saturation," IEEE Access, 7: 92220-92231, 2019.

[35] X. Yan, M. Chen, G. Feng, Q. Wu, S. Shao, "Fuzzy robust constrained control for nonlinear systems with input saturation and external disturbances," IEEE Trans. Fuzzy Syst., 29(2): 345-356, 2021.

[36] B. Xu, "Composite learning control of flexible-link manipulator using NN and DOB," IEEE Trans Syst. Man Cybern. Syst., 48(11): 1979-1985, 2018.

[37] M. M. Zohrei, H. R. Javanmardi, "Nonlinear observer-based control design for a three-axis inertial stabilized platform," J. Appl. Res. Electr. Eng., 2(2): 158-172, 2024.

[38] L. Wang, X. Li, Y. Liu, D. Mao, B. Zhang, "High-precision control of aviation photoelectric-stabilized platform using extended state observer-based kalman filter," Sensors, 23(22): 9204, 2023.

[39] H. Gorjizadeh, M. Ghalehnoie, S. Negahban, A. Nikoofard, "Fuzzy controller design for constant bottomhole pressure drilling under operational/physical constraints," J. Pet. Sci. Eng., 212: 110335, 2022.

[40] N. H. Giap, J. H. Shin, W. H. Kim, "Robust adaptive neural network control for XY table," Intell. Control Autom., 04(03): 293-300, 2013.

Biographies

Mohsen Ghalehnoie was born in Shahrood, Iran, in 1982, and holds Bachelor's, Master's, and Doctoral degrees in Control Engineering. He earned his B.Sc. and M.Sc. degrees from Iran University of Science and Technology and the University of Tehran in 2005 and 2008, respectively. Additionally, he received his Ph.D. from Ferdowsi University of Mashhad in 2018. Currently, he serves as an assistant Professor of Control Engineering at Shahrood University of Technology, Shahrood, Iran. His work focuses on control systems theory, optimization, fuzzy control, data fusion, and expert systems, especially for hybrid switched systems and industrial processes.

- Email: ghalehnoie@shahroodut.ac.ir
- ORCID: 0000-0001-8012-262X
- Web of Science Researcher ID: AFL-9790-2022
- Scopus Author ID: 27367994200
- Homepage: https://shahroodut.ac.ir/fa/as/?id=S917

Ali Azhdari was born in Shiraz, Iran, in 1995. He received his Bachelor's degree in Control Electrical Engineering from Fasa University in 2018 and his Master's degree in the same field from Shahrood University of Technology in 2021. His current research focuses on Buck and Multi-level converters. He is particularly interested in Classic control methodologies such as back-stepping control, Super twisting algorithms, Sliding Mode Control, and adaptive control. Additionally, he has expertise in fuzzy controllers and optimization methods.

- Email: aliazhdari.pro@gmail.com
- ORCID: 0000-0002-9779-1683
- Web of Science Researcher ID: JHT-1749-2023
- Scopus Author ID: NA
- Homepage: https://www.linkedin.com/in/ali-azhdari-2ab243180

Javad Keighobadi his B.Sc., M.Sc. and Ph.D. degrees from Shahrood University of Technology, in 2012, 2014 and 2020, respectively. Currently, He is an assistant professor with the Faculty of Electrical Engineering at Shahrood University of Technology. His research interests include Nonlinear Control, Fault-tolerant Systems, Robotics and Intelligent Control.

- Email: javad_keighobadi@shahroodut.ac.ir
- ORCID: 0000-0001-6474-5499
- Web of Science Researcher ID: KYQ-0669-2024
- Scopus Author ID: 57211945901
- Homepage: https://shahroodut.ac.ir/fa/as/index.php?id=S1148

How to cite this paper:
M. Ghalehnoie, A. Azhdari, J. Keighobadi, ” Robust fuzzy control of uncertain two-axis inertially stabilized platforms using a disturbance observer: A backstepping-based adaptive control approach,” J. Electr. Comput. Eng. Innovations, 13(1): 13-26, 2025.

DOI: 10.22061/jecei.2024.10886.746

URL: https://jecei.sru.ac.ir/article_2171.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Persian Slang Text Conversion to Formal and Deep Learning of Persian Short Texts on Social Media for Sentiment Classification

M. Khazeni, M. Heydari *, A. Albadvi

Department of IT Engineering, Faculty of Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran.

Article Info

Article History:
Received 19 April 2024
Reviewed 17 June 2024
Revised 08 August 2024
Accepted 14 August 2024

Keywords:
Natural language processing
Persian conversational text
Sentiment analysis
Deep learning

*Corresponding Author's Email Address:
m_heydari@modares.ac.ir

Abstract

Background and Objectives: The lack of a suitable tool for the analysis of conversational texts in Persian language has made various analyzes of these texts, including Sentiment Analysis, difficult. In this research, it has we tried to make the understanding of these texts easier for the machine by providing PSC, Persian Slang Convertor, a tool for converting conversational texts into formal ones, and by using the most up-to-date and best deep learning methods along with the PSC, the sentiment learning of short Persian language texts for the machine in a better way.

Methods: Be made More than 10 million unlabeled texts from various social networks and movie subtitles (as dialogue texts) and about 10 million news texts (as official texts) have been used for training unsupervised models and formal implementation of the tool. 60,000 texts from the comments of Instagram social network users with positive, negative, and neutral labels are considered as supervised data for training the emotion classification model of short texts. The latest methods such as LSTM, CNN, BERT, ELMo, and deep processing techniques such as learning rate decay, regularization, and dropout have been used. LSTM has been utilized in research, and the best accuracy has been achieved using this method.

Results: Using the official tool, 57% of the words of the corpus of conversation were converted. Finally, by using the formalizer, FastText model and deep LSTM network, the accuracy of 81.91 was obtained on the test data.

Conclusion: In this research, an attempt was made to pre-train models using unlabeled data, and in some cases, existing pre-trained models such as ParsBERT were used. Then, a model was implemented to classify the Sentiment of Persian short texts using labeled data.

Introduction

With the increasing accessibility of digital platforms, such as websites and social networks, there is a continual surge in the volume of data generated from these sources. This data serves as a valuable resource for managers and researchers, facilitating diverse analyses. An aspect worthy of examination involves the emotional content within texts, a task traditionally performed manually by human resources in the past. Contemporary advancements enable the delegation of this analysis to machines, handling larger datasets with heightened efficiency, precision, and comprehensiveness. Nevertheless, data sourced from social networks is inherently unstructured and characterized by its complexities. This research endeavors to introduce a system designed for the analysis of conversational and concise texts originating from Persian language social networks. The ensuing section refers to the delineation, objectives, significance, and historical context of the

subject matter. Emphasizing the analysis of text emotions offers multifaceted applications across commercial, economic, educational, political, and cultural domains. In recent years, a spectrum of methodologies leveraging text mining algorithms, natural language processing, and emotion dictionaries has emerged for emotion analysis and opinion mining. Notably, the proliferation of deep neural networks has gained prominence in this domain, reflecting their computational prowess, akin to their widespread adoption in various analytical fields.

Related Works

Within the sentiment analysis domain of short texts in Persian, several research initiatives have been executed. Nevertheless, none of these investigations have relied on extensive and dependable conversational datasets, thus lacking requisite comprehensiveness and confidence. Additionally, the historical trajectory of studies concerning Persian conversational language extends considerably, predominantly centering on the elucidation of features, characteristics, and rules inherent to spoken Persian. However, a limited number have explored the conversion of slang texts into formal language. This study stands as the inaugural substantive exploration of transforming colloquial texts into formalized expressions, addressing a notable gap in existing research endeavors [1].

The progression of natural language processing and text mining has spurred a surge in research devoted to refining sentiment analysis methods. The primary objective is to adeptly handle expansive datasets with enhanced precision and efficiency. Recognizing the importance of sentiment analysis and acknowledging the distinct challenges posed by both formal and conversational nuances within the Persian language, there arises a pressing demand for innovative models and methodologies to effectively scrutinize sentiments within extensive textual datasets originating in Persian [2].

The diverse applications and lucrative implications of sentiment analysis, coupled with the linguistic intricacies presented by the Persian language, underscore a significant and pressing issue. Addressing this matter necessitates foundational solutions derived from comprehensive and extensive studies within this domain [3].

A prevalent concern within text mining pertains to sentiment analysis, also known as opinion mining. This process entails the computational scrutiny of opinions, evaluations, attitudes, and emotions conveyed by individuals regarding various entities, such as individuals, issues, events, topics, and their respective attributes [4].

In nearly all sources, sentiment analysis is commonly examined at three primary levels: document, sentence, and aspect. At the document level, the goal is to determine whether the overall sentiment of the document is positive or negative. At the sentence level, the sentiment is assessed for each individual sentence. Finally, at the aspect level, the sentiment is investigated for each feature or entity mentioned within the sentence [5].

Challenges encountered in machine learning for short texts include a) Brevity: Short texts, comprising only a few words, may lead to insufficient representation of the document for comprehension or learning. b) Feature Limitations: Short texts have limited length, and this restricted capacity must be utilized to express diverse topics for users, each using their own vocabulary and writing style. Therefore, a specific topic may have diverse content, making it challenging to extract precise features from short texts. c)Swift Processing: Due to its practical application, a short text needs to be processed very quickly, and results need to be conveyed promptly. d) Spelling Errors and Informal Writing: In many cases, especially in opinions expressed on microblogs and social networks, a short text is summarized, resulting in numerous spelling errors, informal writing, or colloquial expressions [6].

Naemi et al., addressed conversion of Persian informal words to formal words by using the spell-checking approach. They extracted two datasets included formal and informal words from the four most visited news websites in Persian. Results show that their proposed system can detect approximately 94% of the Persian informal words, with the ability to detect 85% of the best equivalent formal words [7].

Tajalli et al., building a parallel corpus of 50,000 sentence pairs with alignments in the word/phrase level. The sentences were attempted to cover almost all kinds of lexical and syntactic changes between informal and formal Persian, therefore both methods of exploring and collecting from the different resources of informal scripts and following the phonological and morphological patterns of changes were applied to find as much instances as possible. The corpus has about 530,000 alignments and a dictionary containing 49,397 word and phrase pairs [8].

Rasooli et al., proposed an effective standardization approach based on sequence-to-sequence translation. They designed an algorithm for generating artificial parallel colloquial-to-standard data for learning a sequence-to-sequence model and annotated publicly available evaluation data consisting of 1912 sentences. Their model improves English-to-Persian machine translation in scenarios for which the training data is from colloquial Persian with 1.4 absolute BLEU score difference in the development data, and 0.8 in the test data [9].

Mazoochi et al., constructed a user opinion dataset called ITRC-Opinion in a collaborative environment and insource way contained 60,000 informal and colloquial

Persian texts from social microblogs such as Twitter and Instagram. They proposed a new architecture based on the convolutional neural network (CNN) model for more effective sentiment analysis of colloquial text in social microblog posts. The constructed datasets are used to evaluate the presented architecture. Some models, such as LSTM, CNN-RNN, BiLSTM, and BiGRU with different word embeddings, including FastText, Glove, and Word2vec, investigated our dataset and evaluated their results. Their model reached 72% accuracy [10].

Momtazi et al., explored the issue of informal texts, and suggested a framework for transforming informal texts into formal texts in the Persian language. Two cutting-edge sequence-to-sequence models, specifically the encoder-decoder and transformer-based models, are employed for this purpose. Alongside neural network models, a series of guidelines for converting informal text to formal text are introduced, and the effects of integrating these guidelines with each of the two models are analyzed. The evaluation of their proposed frameworks reveals that the optimal performance, with an accuracy of 70.7% in the SacreBLEU metric, is achieved through the utilization of the transformer-based model in conjunction with the set of guidelines [11].

Nezhad et al., investigations on sarcasm detection technique in Persian tweets was examined through the integration of various machine learning and deep learning approaches. A series of feature sets encompassing diverse forms of sarcasm were introduced, specifically deep polarity, sentiment, part of speech, and punctuation features. These features were employed for the categorization of tweets into sarcastic and nonsarcastic categories. The deep polarity feature was formulated by executing a sentiment analysis utilizing a deep neural network architecture. Moreover, a Persian sentiment lexicon comprising four sentiment classifications was constructed to extract the sentiment feature. Additionally, a novel Persian proverb lexicon was incorporated during the preparatory phase to enhance the precision of the proposed model. The model's performance was assessed through a range of standard machine learning techniques. The experimental outcomes demonstrated that the method surpassed the baseline approach, achieving an accuracy rate of 80.82%. The research also delved into the significance of each feature set proposed and appraised their contribution to the classification process [12].

Golazizian et al., in their research, which is the first attempt at irony detection in Persian language, emoji prediction is used to build a pretrained model. The model is finetuned utilizing a set of hand labeled tweets with irony tags. A bidirectional LSTM (BiLSTM) network is employed as the basis of our model which is improved by attention mechanism. Additionally, a Persian corpus for irony detection containing 4339 manually labeled tweets is introduced. Experiments show the proposed approach outperforms the adapted state-of-the-art method tested on Persian dataset with an accuracy of 83.1% and offers a strong baseline for further research in Persian language [13].

Hajiabdollah et al., explored on Improving polarity identification in sentiment analysis using sarcasm detection and machine learning algorithms in Persian tweets. To accomplish their study, 8000 Persian tweets that have emotional labels and examined for the presence or absence of sarcasm have been used. The innovation of their research is in extracting keywords from sarcastic sentences. In their research, a separate classifier has been trained to identify irony of the text. The output of this classifier is provided as an added feature to the text recognition classifier. In addition to other keywords extracted from the text, emoticons and hashtags have also been used as features. Naive Bayes, support vector machines, and neural networks were used as baseline classifiers, and finally the combination of classifiers was used to identify the feeling of the text. The results of this study show that identifying the irony in the text and using it to identify emotions increases the accuracy of the results [14].

Najafi-Lapavandani et al., investigated on Humor Detection in Persian. As one of the early efforts for detecting humor in Persian, their research proposes a model by fine-tuning a transformer-based language model on a Persian humor detection dataset. The proposed model has an accuracy of 84.7% on the test set. Moreover, their research introduced a dataset of 14,946 automatically-labeled tweets for humor detection in Persian [15].

Sharma et al. proposed a model for classifying short sentimental sentences using a CNN-enhanced with fine-tuned Word2Vec embeddings. Their approach demonstrated improved classification performance, highlighting the efficacy of CNNs in handling short text sentiment analysis [16].

Muhammad et al. conducted sentiment analysis on Indonesian hotel reviews utilizing a combination of Word2Vec embeddings and LSTM networks. Their study achieved significant improvements in sentiment classification accuracy, underscoring the effectiveness of integrating Word2Vec with LSTM for capturing contextual information in text [17].

Ouchene et al. explored sentiment analysis on Algerian tweets using FastText embeddings combined with LSTM networks. Their empirical study demonstrated that this hybrid approach effectively captured sentiment nuances in Algerian Arabic tweets, offering a robust methodology for sentiment analysis in low-resource languages [18].

Patel et al. introduced a hybrid deep learning approach for rumor detection, combining ELMo embeddings with CNN. This model achieved enhanced accuracy in detecting rumors, leveraging the contextual embeddings from ELMo and the feature extraction capabilities of CNNs to address the complexities of rumor detection in text [19].

Farahani et al. developed ParsBERT, a transformer-based model tailored for Persian language understanding. ParsBERT significantly advanced the state of natural language processing in Persian, providing a powerful tool for various downstream tasks such as sentiment analysis, text classification, and named entity recognition [20].

Pires et al. investigated the multilingual capabilities of BERT, analyzing its performance across multiple languages. Their findings revealed that multilingual BERT can effectively handle a variety of languages, including those with limited resources, making it a versatile tool for multilingual natural language processing tasks [21].

Alkhalifa et al., addresses the challenge of humor detection in natural language processing, particularly for Arabic, a language with limited resources. The authors collected and annotated humorous tweets in both Arabic dialects and Modern Standard Arabic (MSA). They evaluated seven Arabic pre-trained language models (PLMs)—AraBERTv02, Arabertv02-twitter, QARIB, MarBERT, MARBERTv2, CAMeLBERT-DA, and CAMeLBERT-MIX—by fine-tuning them on this dataset. The results indicated that CAMeLBERT-DA performed best, achieving an F1-score and accuracy of 72.11% [22].

Eke et al., tackles the challenge of sarcasm detection in natural language processing by proposing a context-based feature technique using both deep learning and conventional machine learning models. Traditional models often focus solely on content, neglecting contextual information and sentiment polarity, leading to ineffective sarcasm detection. The study introduces three models: (1) a deep learning model with Bi-LSTM and GloVe embeddings for context learning, (2) a Transformer-based model using the BERT architecture, and (3) a feature fusion model combining BERT, sentiment-related features, syntactic features, and GloVe embeddings with conventional machine learning. Evaluations on Twitter and Internet Argument Corpus (IAC-v2) datasets show high precision rates of 98.5% and 98.0%, demonstrating the effectiveness of the proposed approach [23].

Shatnawi et al., presented BFHumor, a BERT-Flair-based humor detection model designed to identify humor in news headlines. The model combines several state-of-the-art pre-trained NLP techniques in an ensemble approach. Evaluated using SemEval-2020 public humor datasets, BFHumor achieved notable results with a Root Mean Squared Error (RMSE) of 0.51966 and an accuracy of 0.62291. The study also explores the reasons for the model's effectiveness through experiments on the BERT model, revealing that BERT captures surface knowledge in lower layers, syntactic features in middle layers, and semantic understanding in higher layers [24].

Annamoradnejad et al., introduces a novel approach for detecting and rating humor in short texts, leveraging a well-known linguistic theory of humor. The method involves separating sentences within a text, generating embeddings using the BERT model, and feeding these embeddings into a neural network to analyze congruity and latent relationships between sentences for humor prediction. The approach is validated using a newly created dataset of 200,000 labeled short texts for binary humor detection. Additionally, the model was tested in a live machine-learning competition on Spanish tweets, achieving F1 scores of 0.982 and 0.869. These results outperform both general and state-of-the-art models. The study highlights that the effectiveness of the model is significantly attributed to the use of sentence embeddings and the incorporation of humor's linguistic structure in the model design [25].

Sadjadi et al., addresses the challenge of measuring semantic similarity in Persian informal texts, which has been poorly served by previous methods. Traditional approaches have struggled with both accuracy and handling colloquial language. To overcome these limitations, the study introduces a new transformer-based model, FarSSiBERT, specifically designed for Persian informal short texts from social networks. This model is built using the BERT architecture, trained from scratch on approximately 104 million Persian informal texts, and supported by a novel tokenizer that effectively handles informal language. Additionally, a new dataset, FarSSiM, has been created with real social network data and annotated by linguistic experts. The FarSSiBERT model outperforms existing models like ParsBERT, laBSE, and multilingual BERT in measuring semantic similarity, and shows promise for broader NLP tasks involving colloquial Persian text and informal tokenization [26].

Falakaflaki et al., addresses the challenge of formality style transfer in Persian, a task complicated by the growing use of informal language on digital platforms. The goal is to convert informal text into formal text while preserving its original meaning, considering both lexical and syntactic differences. The authors propose a new model, Fa-BERT2BERT, which builds on the Fa-BERT architecture and integrates consistency learning with gradient-based dynamic weighting. This model enhances understanding of syntactic variations and improves balance in loss components during training. Evaluation against existing methods using new metrics tailored to syntactic and stylistic changes shows Fa-BERT2BERT's superior performance across BLEU, BERT score, Rouge-l,

and other metrics. This advancement enhances Persian language processing by improving the accuracy and functionality of NLP tools, which can streamline content moderation, enhance data mining, and support effective cross-cultural communication [27].

Dashti et al., presented an advanced Persian spelling correction system that integrates deep learning with phonetic analysis to improve accuracy and efficiency in NLP. The system employs a fine-tuned language representation model to combine deep contextual understanding with phonetic insights, effectively addressing both non-word and real-word spelling errors. It is particularly adept at handling the complexities of Persian spelling, such as its intricate morphology and homophony. Evaluations on a comprehensive dataset reveal the system's exceptional performance, with F1-Scores of 0.890 for detecting real-word errors, 0.905 for correcting them, and 0.891 for non-word error correction. These results demonstrate the effectiveness of incorporating phonetic analysis into deep learning models for spelling correction, advancing Persian language processing and highlighting a valuable approach for future research in the field [28].

Kebriaei et al., addresses the issue of hate and offensive language on social networks, focusing on Twitter and Persian language content. Due to the scarcity of resources for Persian, the researchers compiled a dataset of 38,000 Persian tweets containing hate and offensive language, using keyword-based selection and crowdsourced lexicons. The dataset includes a Persian offensive lexicon and nine target-group lexicons. Manual annotation was performed by multiple annotators to ensure accuracy. The study also evaluated potential biases in the dataset using two assessment criteria (FPED and pAUCED) and adjusted the dataset to reduce bias. The results show that while bias was significantly reduced, the F1 score was minimally affected, demonstrating the effectiveness of the bias mitigation strategy [29].

Vakili et al., explores advanced sentiment analysis techniques for Persian Twitter content by combining multiple approaches: the Naive Bayes classifier, a custom rule-based model, and the BERT transformer model. While traditional models like SVM, Naive Bayes, and MLP show limitations in isolation, the hybrid model developed in this study integrates these methods and achieves notable improvements. The hybrid approach, which combines Naive Bayes and a bespoke rule-based model with BERT, outperforms BERT alone, reaching an accuracy of 89% compared to BERT's 86%. Despite being slightly more complex, this hybrid model maintains comparable computational efficiency to BERT fine-tuning and enhances sentiment classification effectiveness for social media applications [30].

Table 1: Related works comparison

Study	Objective	Dataset	Method	Results
Naemi	Conversion of Persian informal words to formal words	Formal and informal word datasets from Persian news websites	Spell-checking approach	94% detection of informal words; 85% detection of formal equivalents
Tajalli	Building a parallel corpus for informal-to-formal Persian text transformation	50,000 sentence pairs with 530,000 alignments and a dictionary of 49,397 pairs	Resource collection and phonological/morphological pattern analysis	Comprehensive corpus and dictionary for informal-to-formal conversion
Rasooli	Standardization of colloquial to formal Persian using sequence-to-sequence translation	1,912 annotated sentences	Sequence-to-sequence model for artificial parallel data generation and standardization	1.4 BLEU score improvement for English-to-Persian translation
Mazoochi	Sentiment analysis of colloquial Persian texts	60,000 informal and colloquial Persian texts from microblogs	CNN-based architecture with various embeddings	Achieved 72% accuracy in sentiment analysis
Naemi	Conversion of Persian informal words to formal words	Formal and informal word datasets from Persian news websites	Spell-checking approach	94% detection of informal words; 85% detection of formal equivalents
Tajalli	Building a parallel corpus for informal-to-formal Persian text transformation	50,000 sentence pairs with 530,000 alignments and a dictionary of 49,397 pairs	Resource collection and phonological/morphological pattern analysis	Comprehensive corpus and dictionary for informal-to-formal conversion
Rasooli	Standardization of colloquial to formal Persian using sequence-to-sequence translation	1,912 annotated sentences	Sequence-to-sequence model for artificial parallel data generation and standardization	1.4 BLEU score improvement for English-to-Persian translation
Mazoochi	Sentiment analysis of colloquial Persian texts	60,000 informal and colloquial Persian texts from microblogs	CNN-based architecture with various embeddings	Achieved 72% accuracy in sentiment analysis

Study	Objective	Dataset	Method	Results
Momtazi	Formalization of informal Persian texts using sequence-to-sequence and transformer models	No specific dataset mentioned, evaluation on existing data	Encoder-decoder and transformer models with integration guidelines	70.7% accuracy using transformer-based model with guidelines
Nezhad	Sarcasm detection in Persian tweets	8,000 Persian tweets with emotional labels	Machine learning and deep learning techniques with various feature sets	80.82% accuracy in sarcasm detection
Golazizian	Irony detection in Persian tweets	4,339 manually labeled Persian tweets	BiLSTM network with attention mechanism	83.1% accuracy in irony detection
Hajiabdollah	Improving sentiment analysis by incorporating sarcasm detection	8,000 Persian tweets with emotional labels	Classifiers with sarcasm detection feature integration	Increased accuracy in sentiment analysis by incorporating irony detection
Najafi	Humor detection in Persian tweets	14,946 automatically labeled Persian tweets	Fine-tuned transformer-based language model	84.7% accuracy in humor detection
Sharma	Sentiment classification of short sentences using CNN with Word2Vec embeddings	Not specified, general short sentences	CNN with fine-tuned Word2Vec embeddings	Improved classification performance for short sentences
Muhammad	Sentiment analysis of hotel reviews in Indonesian using Word2Vec and LSTM	Indonesian hotel reviews	Word2Vec embeddings combined with LSTM networks	Significant improvement in sentiment classification accuracy
Ouchene	Sentiment analysis of Algerian tweets using FastText and LSTM	Algerian Arabic tweets	FastText embeddings with LSTM networks	Effective sentiment analysis in low-resource language
Patel	Rumor detection using hybrid deep learning approach	Not specified, rumor detection dataset	ELMo embeddings combined with CNN	Enhanced accuracy in rumor detection
Farahani	Persian language processing advancements with ParsBERT	Various Persian language tasks	Transformer-based ParsBERT model	Significant advancements in Persian NLP tasks
Pires	Multilingual capabilities of BERT	Various multilingual datasets	Analysis of multilingual BERT performance	Effective handling of multiple languages, including limited-resource languages
Alkhalifa	Humor detection in Arabic using pre-trained language models	Dataset of humorous tweets in Arabic	Fine-tuned pre-trained Arabic language models	CAMeLBERT-DA model achieved an F1-score of 72.11%
Eke	Sarcasm detection using context-based features	Twitter and Internet Argument Corpus datasets	Bi-LSTM with GloVe, BERT-based model, and feature fusion	High precision rates of 98.5% and 98.0%
Shatnawi	Humor detection in news headlines using BERT and Flair	SemEval-2020 public humor datasets	BERT-Flair-based ensemble model	RMSE of 0.51966 and accuracy of 0.62291
Annamoradnejad	Humor detection and rating in short texts based on linguistic theory	200,000 labeled short texts	Sentence embeddings with BERT and neural network	F1 scores of 0.982 and 0.869 in humor detection
Sadjadi	Measuring semantic similarity in Persian informal texts	FarSSiM dataset with 104 million Persian informal texts	Transformer-based FarSSiBERT model with novel tokenizer	Outperformed ParsBERT, laBSE, and multilingual BERT in similarity measurement
Falakaflaki	Formality style transfer in Persian, converting informal to formal text	Not specified	Fa-BERT2BERT model with consistency learning and dynamic weighting	Superior performance in BLEU, BERT score, Rouge-l, and other metrics
Dashti	Persian spelling correction integrating deep learning and phonetic analysis	Comprehensive spelling correction dataset	Fine-tuned language model with phonetic analysis	F1-Scores: 0.890 (real-word errors), 0.905 (corrections), 0.891 (non-word errors)
Kebriaei	Identification of hate and offensive language in Persian tweets	38,000 Persian tweets with hate and offensive language annotations	Keyword-based data selection and crowdsourced lexicons	Effective bias mitigation with minimal impact on F1 score
Vakili	Advanced sentiment analysis for Persian Twitter content integrating Naive Bayes, rule-based model, and BERT	Persian Twitter content	Hybrid model combining Naive Bayes, rule-based, and BERT	Achieved 89% accuracy, outperforming BERT's 86%

Objectives

This research will employ quantitative research methodology to derive results. The approach involves the observation and analysis of authentic data, aimed at extracting and scrutinizing pertinent characteristics and variables, as well as evaluating the impact of each. Subsequently, the findings, along with the identification of optimal variables, will be reported based on the empirical data. The procedural steps for conducting this research include:

A. Preliminary Study in the Required Fields

A thorough and adequate examination of the definitions, terms, and existing literature relevant to the subject within the field is imperative.

B. Conversion of Slang Texts into Formal

The investigation focuses on delineating the definition of informal language and discerning its distinctions from formal language, in addition to elucidating the rules governing informal language. The goal is to devise a tool capable of transforming slang texts into formal expressions within the Persian language context. Furthermore, a comprehensive review of existing efforts in the domain of converting slang texts into formal language will be undertaken to inform and enrich the study. Given the Persian colloquial and formal grammar, a significant portion of colloquial words have been converted to formal ones using rule-based methods. Each of these methods not only affects the target words but also introduces errors in other words. Therefore, the effectiveness of these methods had to be evaluated through trial and error to ensure that if the generated error rate was negligible, the method would be chosen for converting colloquial words to formal ones.

C. Sentiment Analysis of Persian Short Texts

A comprehensive investigation within the domain of sentiment analysis is warranted, encompassing an exploration of definitions, levels, and practical applications within the field. Additionally, attention will be devoted to examining the characteristics of short texts and the associated challenges posed to machine learning algorithms. Furthermore, an inquiry into existing endeavors concerning sentiment analysis of both Persian and English short texts will be conducted to inform the research comprehensively.

D. Data Gathering

This research involves the collection of data from two distinct categories.

1. **Unlabeled Text:** To facilitate the conversion of conversational texts into formal expressions and to facilitate deep learning of short texts, the research will rely on two primary sources of data. The first source encompasses raw and untagged texts extracted from social networking platforms such as Instagram, Twitter, and Telegram, as well as subtitles from movies, which will serve as conversational data. The second source comprises texts sourced from Persian-language news agency websites, which will serve as formal texts for the study.
2. **Labeled Text:** For the implementation of the sentiment analysis model, supervised learning will be employed using tagged data derived from Instagram users' comments.

E. Data Standardization

To prepare the data for subsequent steps, a crucial pre-processing and standardization phase is imperative. Texts from social networks exhibit considerable complexity, attributable to diverse characters, hashtags, links, emoticons, and generally non-standard writing practices. Prior to utilization and processing, it is essential to standardize these data. While numerous basic natural language processing tools are available for the Persian language, their performance on conversational texts from social networks is often suboptimal. Consequently, there arises a necessity to implement specific equalization techniques tailored for these texts.

F. Method

Through a meticulous examination of the data and methodologies applied in addressing both the challenges of converting slang texts into formal expressions and analyzing the sentiments of short texts, the research will make informed decisions regarding the selection of methods. These methods may draw inspiration from prior research efforts or be entirely novel, tailored to the specific requirements of the current study. The decision-making process will be guided by a comprehensive understanding of the intricacies presented by the data and the specific objectives of the research.

G. Implementation

Based on the outcomes derived from the preceding stage, the chosen methods will be implemented utilizing the research data. Unlabeled data will be leveraged to train unsupervised models, while labeled data will be instrumental in training supervised models. Considering the substantial volume of data involved in this research, the utilization of appropriate hardware and up-to-date software packages is deemed necessary.

H. Evaluation

A pivotal stage in any system involves the evaluation and validation of the stated claims. In the context of this research, a comprehensive examination will be conducted to assess both the efficacy of converting slang texts into formal expressions and the accuracy of the sentiment analysis applied to short texts. This meticulous investigation aims to verify the reliability and effectiveness of the proposed methods within the defined scope of the research.

1. **Formal evaluation of the Proposed Tool:** Considering the scarcity of labeled data and the encompassing diversity within these datasets, the evaluation of this tool will pivot on the total number of converted words. This approach accounts for the comprehensive nature of the data, providing a broader assessment of the tool's performance in handling various linguistic nuances and expressions encountered in conversational texts.
2. **Short text Sentiment Classification Evaluation:** The model, constructed using the training data, is subsequently applied to the test data. The predictions generated by the model are then compared with the actual labels, and various evaluation criteria, including accuracy, precision, recall, and F-score, are computed. This meticulous assessment serves to gauge the performance of the model and validate its effectiveness in accurately predicting outcomes across the test dataset.
3. **PSC Evaluation for Short Text Sentiment Classification:** Furthermore, each classification involving slang texts and their corresponding converted expressions undergoes evaluation through the formalizer tool. This analysis aims to scrutinize the impact of the formalizer tool on the transformation process and assess its effectiveness in achieving the desired conversion of slang texts into formalized language.

Data Section

The totality of the data used in this research is described in this section. The utilized data comprises two parts with labels and without labels. The labeled data are used to train the sentiment classification model.

A. Data Gathering

Based on the conducted investigations, it appears that authentic conversational texts in the Persian language are scarcely available. However, there are corpora that includes complete sets of formal texts, such as the Hamshahri corpus containing several years' worth of news from this agency. While the news agency's content may seem comprehensive, a notable challenge lies in the fact that not all sentences within this corpus are strictly formal. Instances may arise where the Hamshahri news agency quotes colloquial expressions from individuals, introducing a layer of complexity to the categorization of formal and informal language within the dataset.

Given the inadequacy of Persian texts meeting the requirements of this treatise, it became imperative to procure colloquial and formal textual data from diverse sources. Based on the conducted surveys, the most suitable source for data collection emerged as the formal texts from news agencies. Numerous Farsi-language news agency websites are accessible, facilitating data extraction through web crawling. Additionally, recognizing that colloquial language is prevalent in everyday communication, social networks stand out as a rich source for collecting colloquial data. Many social networks offer programming interfaces that can be leveraged for this purpose.

1. **Data Crawling***:* Data crawling from the web involves extracting information and the structural elements of a web page, subsequently storing the acquired data in a database for individual retrieval of texts, images, videos, and other components. The initial phase of this information extraction process entails creating a robot (crawler) capable of recognizing the structure of a web page, executing parsing operations, and then storing the parsed data in a database. In this research, a custom-designed crawler was employed to fulfill these tasks.
2. **Social Networks APIs**: The process of data collection is significantly streamlined using interfaces, obviating the need to develop and program a crawler from scratch. Social networks offer interfaces to enhance profitability and foster business activities within their networks. Notably, tools such as the Telegram bot for Telegram and the Instagram API for Instagram have been employed for data collection in this research. The Telegram bot facilitates the collection and storage of diverse data from groups and channels, while the Instagram interface enables the gathering and storage of posts and comments from public Instagram networks.

B. Dataset

The data collected for this research is presented in the Table 2. The dataset in the study is categorized into two main types: raw data and labeled data. This chapter will focus on the discussion of unlabeled data, while Table 3 will provide insights into the labeled data.

Table 2: Dataset statistic

Data Subject	Counts
Persian news agencies	10,000,000
Instagram Comments	6,000,000
Movies Subtitles	4,000,000
Telegram Groups	4,000,000
X (Twitter) Tweets	4,000,000
Instagram Comments	80,000
Sum	28,080,000

For unsupervised learning methods like Word2Vec, FastText, and BERT, which involve pre-training models, the efficacy and generalization of these models are highly dependent on the use of large datasets. The expansive datasets serve to impart a broad understanding of

language patterns and nuances, enabling the models to capture a rich representation of the underlying linguistic structures. The use of substantial and diverse datasets is crucial in enhancing the robustness and performance of pre-trained models in various natural language processing tasks.

C. Data Preprocessing

Before conducting any operations on textual data, preprocessing is essential to make the data usable. Text processing tools offer various functionalities such as equalization, stemming, tokenization, and sentence segmentation. Among these tools, Hazm is widely used for Persian text processing, particularly for formal texts. However, a separate tool was implemented for unifying conversational texts. The prevalence of non-standard words in everyday communication, especially in social networks, poses a challenge for machines to comprehend text content. These non-standard words can impact the performance of natural language processing tools, including machine translators, text summarizers, and text component taggers. To achieve the highest accuracy, text unification is crucial before any processing. Some of the equalization operations include Removal of links, IDs, and phrases related to social networks, such as retweets on Twitter.

- Separation of emoticons, punctuation, numbers, hashtags, English letters, and other characters,
- Unification of Persian and Arabic letters, as well as Persian and English numbers,
- Removal and correction of spaces and semi-spaces,
- Removal of Arabization, and
- Unification of polysyllabic words, such as "America" and its Persian equivalent.

These equalization operations collectively contribute to preparing the textual data for subsequent natural language processing tasks.

D. Constructing Term-Frequency (TF) Corpus

Some text platforms predominantly contain colloquial text, while others are primarily formal; these can be utilized to establish word frequencies. For instance, raw comments from social networks can serve as colloquial texts, while raw texts from news agencies can be considered formal. However, a crucial consideration here is that certain texts, like comments on the Instagram social network, may lack entirely colloquial sentences, featuring formal expressions or even sentences in other languages or dialects. It is evident that such a corpus is more diverse and liberal, making these sentences and expressions more authentic and valid. After standardizing the corpora, the news corpus was designated as the formal corpus, and the aggregate of tweet corpora, Telegram messages, Instagram comments, and video subtitles were merged as the colloquial corpus. Subsequently, two corpora were created to quantify the frequency of formal and colloquial words. A corpus was specifically crafted under the title of "Pure Conversational." Pure Conversational is a corpus of words that adheres to specific criteria.

1. They must be present in the Conversational Corpus.
2. If they exist in the formal corpus, the number of repetitions in the colloquial corpus must be more than five times the number of repetitions in the formal corpus.

This coefficient was obtained by trial and error. The number of repetitions of words in each figure is available in the Table 3.

Table 3: Term Frequency (TF) corpus

Frequency Corpus	Unique Words	Sum of Frequency
Formal	570,092	726,905,260
Slang	1,465,468	70,236,528
Pure Slang	1,279,607	16,136,606

Also, the first 10 words of each figure and the number of repetitions in that figure are listed in the Table 4.

Table 4: The first 10 words of each figure and the number of repetitions in that figure

Pure Slang		Slang		Formal	
476,752	یه	1,626,397	و	33,567,634	و
214,569	باشه	1,230,548	به	27,528,074	در
204,356	دیگه	1,211,883	که	22,548,591	به
184,633	داره	964,338	از	17,483,954	از
177,246	میشه	878,994	رو	14,038,740	این
156,974	اگه	752,524	این	13,996,189	که
81,062	میکنه	714,285	تو	11,053,024	با
80,765	بشه	712,672	من	10,914,328	است

PSC: Persian Slang Text Conversion to Formal

Many slang words can be formalized by examining the colloquial grammar rules explained in the previous chapter and studying the pure colloquial corpus through the application of a limited set of rules. These rules have been derived using the intelligent search method and visual observation to identify patterns and structures within colloquial language.

1. **Words Direct Conversion**: The number of repetitions of words in the Pure Conversational Corpus follows the Fig. 1. This chart follows a large head long tail distribution. By converting a few key words, many

words in any colloquial sentence can be easily transformed. The 1000 most frequent words from the pure conversational corpus were directly and manually converted into their formal forms. By obtaining a pure colloquial corpus through the comparison of colloquial and formal data, and subsequently reviewing this corpus, interesting rules naturally come to mind.

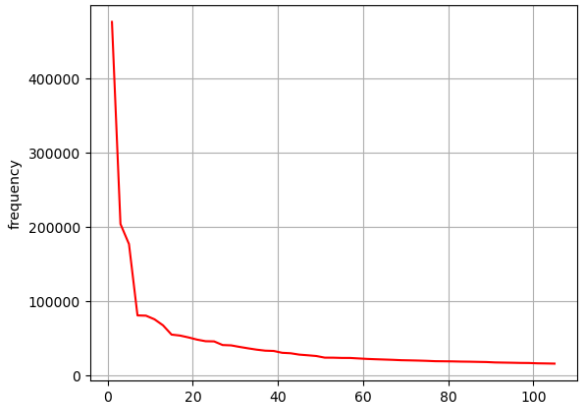


Fig. 1: The number of repetitions of words in the body of pure conversation.

- For example, after observing the words "*اون*" (oon), "*همون*" (hamoon), and "*خونه*" (khooneh), we conclude that by converting "*ون*" (oon) to "*ان*" (aan), several colloquial words can be transformed into their formal equivalents. However, it is inevitable to consider exceptions for each rule, such as the word "*خون*" (khoon) for this specific rule.
- **"*و*" to "*را*" Conversion** Strings that end with the letter "*و*" and are in the object sentence. First, it is checked that this word does not have more than 10 repetitions in the formal corpus. This is so that it does not become a mistake if the formal word existed with this form. After removing the "*و*" at the end of the word, the number of strings is checked in the formal body, and if they were more than 20, they become that substring plus "*را*". for example:
  - گلابیو : گلابی را
  - میزو : میز را
- **"ون" to "ان" Conversion** Strings ending in "*ون*". First, it is checked that this word does not have more than 10 repetitions in the official corpus. This is so that it does not become a mistake if the formal word existed with this form. Strings are checked for their number after removing "*ون*" at the end of the word in the formal body, and if they were more than 20, they become that substring plus "*ان*". For example:
  - خیابون : خیابان
  - صابون : صابون
- **Plural Words Conversion**: Strings ending in "ا" or "ای". First, it is checked that this word does not have more than 10 repetitions in the formal text. This is so that it does not become a mistake if the formal word existed with this form. After removing the "ا" at the end of the word, the number of strings is checked in the formal text, and if they were more than 20, they become that substring plus a half space plus "ها". For example:
  - خودکارا : خودکار ها
  - دارا : دارا
- **Repetition of Letters Conversion**: Sometimes users repeat one or more letters more when typing a word to emphasize or express a feeling. First, it is checked that this word does not have more than 10 repetitions in the formal text. This is so that it does not become a mistake if the formal word existed with this form. After correcting the repeated letters in the formal corpus, the number of strings is checked and if the number of repetitions is more than 20, they are converted. For example:
  - خخخخخخخخخخ : خ
  - پاییز : پاییز
  - مممممحححسسسسننننن : محسن
- **Colloquial Verbs Conversion**: In the last chapter, it was observed that all colloquial verbs, like formal verbs, have rules that can be converted by knowing this rule: prefix + past participle or participle + suffix. that by converting each of these parts into their formal forms, the entire verb can be converted. As:
  - بخون : بخوان
  - ترسوندم : ترساندم
  - رفتن : رفتن
  - ندونه : نداند

  Before the conversion, it is checked that if the number of repetitions of this word in the formal text is more than 10, this conversion will not be done. Also, a list was created for this section under the title of exception list. This list includes all colloquial and formal infinitives that cannot be converted directly. This is because all infinitives such as "*رفتن*" can be used in the third person plural form as well as in the form of a noun.
- **Possessive Pronouns Conversion**: All formal possessive pronouns have a rule that can be converted by knowing this rule. This is so that it does not become a mistake if the formal word existed with this form. Then it is checked if it ends with one of the possessive connected pronouns, they are divided into two substrings, the number of the first substring in the formal corpus is checked, and if they were more than

20, it becomes the first substring and the converted form of the second substring.

- خودکاراتون : خودکارهایتان
- هوام : هوایم
- قیافش : قیافه‌اش
- همسایمون : همسایه‌مان

- **The results of conversion of conversation to formal on pure slang corpus:** After implementing all the rules, each of these rules was executed on the pure slang corpus. In the table below, the ten most frequently converted words from each rule can be seen in the corpus of pure conversation. You can see the result. The number of unique converted words along with the total number of repetitions of these words for each rule is given in the table below.

Table 5 "ون" to "ان" conversion

"ون" to "ان" Conversion		
Word	**Converted**	**No. of Occurrence**
آقایون	آقایان	5144
اقایون	اقایان	2379
یکیشون	یکیشان	2049
شیطون	شیطان	1898
برگردون	برگردان	1388
حالشون	حالشان	1360
اولشون	اولشان	1334
دمشون	دمشان	1299
زندگیشون	زندگیشان	1232
جفتشون	جفتشان	985

Table 6: "و" to "را" conversion

"و" to "را" Conversion		
Word	**Converted**	**No. of Occurrence**
خودمو	خودم را	5017
داستانو	داستان را	3132
همینو	همین را	3087
دهنتو	دهنت را	2713
چیزو	چیز را	2606
اسمشو	اسمش را	2520
مملکتو	مملکت را	2291
اینجارو	اینجا را	2020
صدامو	صدام را	1944
حالمو	حالم را	1892

Table 7: Words direct conversion

Words Direct Conversion		
Word	**Converted**	**No. of Occurrence**
یه	یک	476752
باشه	باشد	214569
دیگه	دیگر	204356
داره	دارد	184633
میشه	می‌شود	177246
اگه	اگر	156974
میکنه	می‌کند	81062
بشه	بشود	80765
واسه	برای	79297
آره	بله	75793

Table 8: Possessive pronouns conversion

Possessive Pronouns Conversion		
Word	**Converted**	**No. of Occurrence**
پیداش	پیدایش	5694
شبتون	شبتان	5586
لطفتون	لطفتان	5212
اینجام	اینجایم	3438
بابام	بابایم	3432
دمتون	دمتان	3380
بابات	بابایت	3219
چشمات	چشم‌هایت	2926
شیش	شیه‌اش	2892
صبحتون	صبحتان	2755

Table 9: Repetition of letters conversion

Repetition Of Letters Conversion		
Word	**Converted**	**No. of Occurrence**
خخخخ	خ	16972
خخخ	خ	14413
خخخخخ	خ	11548
خخخخخخ	خ	6561
خخخخخخخ	خ	3586
جووون	جون	2915
عاااالی	عالی	2833
واای	وای	2636
عاالی	عالی	2633
هههه	ه	2628

Table 10: Colloquial verbs conversion

Colloquial Verbs Conversion		
Word	**Converted**	**No. of Occurrence**
میکنی	می‌کنی	60881
نکنه	نکند	16695
می‌کنه	می‌کند	14846
میکردم	می‌کردم	14347
میخوان	می‌خواهند	11014
می‌خوام	می‌خواهم	10983
میزنم	می‌زنم	10759
میخواستم	می‌خواستم	10007
بکنه	بکند	8124
نمیده	نمی‌دهد	8065

Table 11: Plural words conversion

Plural Words Conversion		
Word	**Converted**	**No. of Occurrence**
دخترا	دخترها	5607
بعضیا	بعضی‌ها	5373
حرفای	حرف‌های	5273
دوستای	دوست‌های	3371
بهترینها	بهترین‌ها	2728
چشمای	چشم‌های	2466
مردا	مردها	2208
دخترای	دخترهای	2204
خانوما	خانوم‌ها	1900
پستای	پست‌های	1771

Table 12: Comparison of rules statistics

Rule	No. of UCW	% of UCW	No. of CW	% of CW
Words Direct Conversion	1000	0.078	6514099	40.36
"را" to "و" Conversion	22556	1.76	329726	0.043
"ان" to "ون" Conversion	3012	0.23	119466	0.74
Plural Words Conversion	31874	2.49	305020	1.89
Repetition Of Letters Conversion	65298	5.10	377351	2.33
Colloquial Verbs Conversion	4522	0.35	1021381	6.32
Possessive Pronouns Conversion	44288	3.46	574996	3.56
All Rules	172550	13.48	9242039	57.27

At the data section of our study, various types of rule-based conversion is shown in Table 5, Table 6, Table 7, Table 8, Table 9, Table 10, and Table 11. Also, statistics of rules comparison is shown in Table 12.

2. **Sentiment Labeled Data**: The dataset used in this research comprises 60,000 comments from the Instagram social network. Each comment has been labeled as positive, negative, or neutral by three users. The final labeling decision assigns the tag of the same class to comments that have at least two identical tags, achieving a consensus-based approach for classification. In the Table 13.

Table 13: Labeled data classification

Label	Count	Sample
Positive	25779	درود بر بزرگ مرد اخلاق سینمای ایران وجهان
Neutral	15366	کمیاب ترین کتاب‌ها در کانال ما
Negative	19657	تأسف برای آن دسته از ادمهایی که انقدر بی ادب هستند
Sum	60802	درود بر بزرگ مرد اخلاق سینمای ایران وجهان

To address data imbalance and ensure an equal number of instances for all classes, 15,000 sentences were randomly subsampled from both the positive and negative classes, resulting in an equal number for all three classes. Consequently, a total of 45,000 labeled data points were generated. In many research scenarios, a certain proportion of labeled data is randomly set aside for training and testing purposes. While some studies employ a three-way split (training, validation, and testing), others utilize the K-Fold method, although this is less common for larger datasets due to its computational complexity. In this research, as it is shown in Fig. 2 after performing pre-processing on all labeled and matched data, 70% of the data was allocated for training, with 15% each for validation and testing.

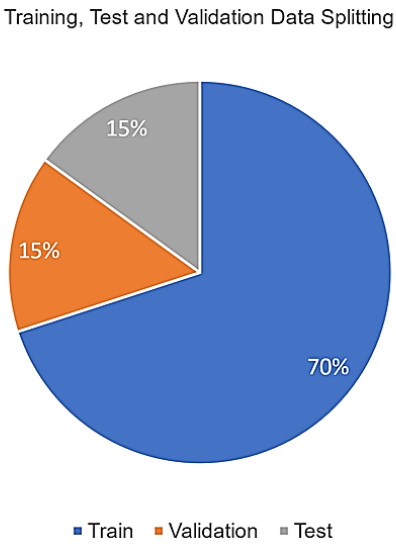


Fig. 2: Splitting rate of train, test and validation data.

Results and Discussion

The algorithms were applied using labeled data, and their respective results are presented in Table 14. Additionally, the PSC metric has been employed to assess its performance on Sentiment classification.

Table 14: Algorithms results comparison

Method	Precision	Accuracy	Recall	F1
Word2Vec+CNN	78.20	78.30	78.14	78.20
Word2Vec+CNN+PSC	78.90	78.91	78.83	78.86
word2Vec+LSTM	80.71	80.98	80.66	80.76
Word2Vec+LSTM+PSC	81.49	81.68	81.44	81.50
FastText+LSTM	81.21	81.12	81.10	81.11
FastText+LSTM+PSC	**81.91**	**81.89**	**81.84**	**81.85**
ELMo+CNN	77.61	77.87	77.41	77.11
ELMo+CNN+PSC	79.10	78.94	78.99	78.96
Parsbert	80.15	80.38	80.10	80.18
Parsbert+PSC	80.61	81.02	80.55	80.65
Bert_Multilingual	70.26	70.22	70.17	70.19

The Table 14 provide a comprehensive overview of the overall performance of the model across various deep learning methods. It is important to note that each cloud method involves numerous parameters, and the reported values represent the best-performing configurations in terms of accuracy. Upon analyzing the tables, the following conclusions can be drawn for deep learning methods:By using the PSC, the performance of all methods has improved, although this value is very low, and with the improvement of the PSC, the performance of the classifier also increases.

- The incorporation of PSC has demonstrated improvements in the performance of all methods, albeit the observed enhancement being relatively modest. Nevertheless, an increase in PSC corresponds to an improvement in classifier performance.
- The highest accuracy achieved is 81.91% using formal FastText vectors in conjunction with an LSTM network.
- Generally, deep processing methods exhibit markedly superior performance compared to machine learning methods. The FastText method outperforms the Word2Vec method, potentially because FastText considers word characters in addition to word embeddings.
- Although BERT-based methods were anticipated to yield superior results, the reduction in accuracy can be attributed to a mismatch in the text domains. The BERT models used in this research were pretrained on formal texts (Wikipedia and books), whereas the labeled data consists of conversational texts from social networks.

The PSC method has been evaluated using two parameters:

a. By counting the number of words converted in the pure colloquial corpus.
b. The effectiveness of this method is one of the best previous sentiment analysis methods.

The implementation of deep learning methods involved numerous hyperparameters. Initially, approximate values for most of these hyperparameters were set based on prior research. Subsequently, through a process of trial and error, the optimal value for each hyperparameter was sought in various iterations to achieve maximum accuracy. The early stopping method was employed to halt the training of the model.

If, after several training sessions, the error and accuracy of the model predictions for the validation data showed no improvement, the training process was terminated. Fig. 3 and Fig. 4 are illustrate the prediction error and accuracy of the model concerning repetitions in the training and validation data using the **FastText+LSTM+PSC** hybrid method.

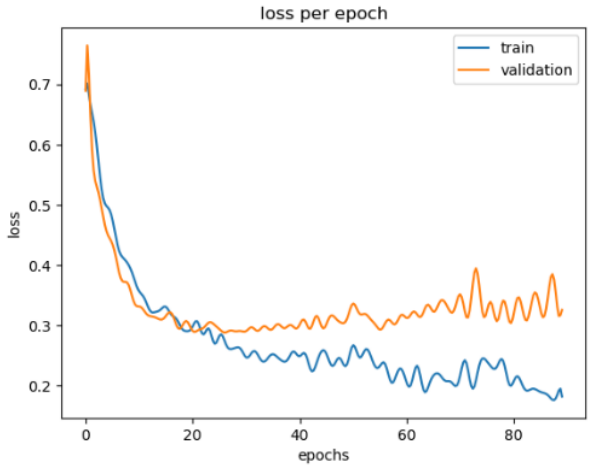


Fig. 3: Loss function.

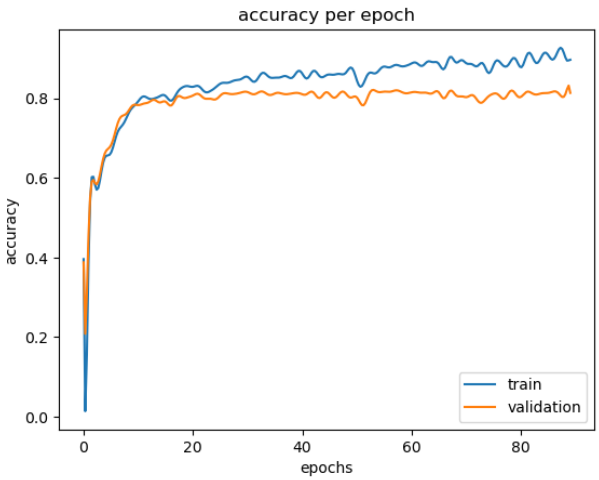


Fig. 4: Accuracy function.

After several iterations, the model's error increases for the validation data while decreasing for the training data, indicating a scenario of overfitting.

This phenomenon highlights the need for careful consideration and fine-tuning of hyperparameters to achieve optimal model performance.

Conclusion

Today, with the increasing utilization of social networks by users, a substantial amount of valuable data is generated for analysis within these networks. Due to the diverse user base encompassing various tastes and age groups, the prevalence of colloquial language, along with abbreviations and numerous spelling and writing mistakes, has significantly grown. The evident lack of basic and advanced conversational language processing tools in the Persian language became a focal point in this research.

The aim was to enhance sentiment analysis by addressing two primary challenges present in textual data from social networks in Persian: shorthand writing and the use of colloquial expressions. The richness and comprehensiveness of the data used in the research play a pivotal role in ensuring the accuracy and thoroughness of the study. The dataset incorporated both labeled and unlabeled data. Labeled data comprised 60,000 sentences with three sentiment classes (positive, negative, and neutral) from the Instagram social network. Unlabeled data consisted of over 10 million sentences from social networks including Instagram, Twitter, and Telegram, representing conversational data, along with over 10 million texts from various news agencies, serving as formal data.

In the scope of converting slang texts into formal language, limited prior research has been conducted. The perceived lack of attention to the importance of this area and its inherent complexity may contribute to this gap. This research addressed this challenge by presenting a hybrid approach involving statistical and rule-based methods for converting colloquial texts into formal ones.

Three forms of data: formal, conversational, and pure conversation were created using the collected data, forming the foundation for the slang-to-formal conversion method. An analysis of the pure colloquial corpus revealed that implementing several rules can successfully convert a significant portion of slang words into formal ones. Following the applied checks, 57.2% of pure colloquial words were successfully converted using the implemented rules.

This research also focused on pre-training models using unlabeled data, occasionally leveraging existing pre-trained models like ParsBERT. Subsequently, a model was implemented to classify the sentiment of Persian short texts using labeled data, achieving a maximum accuracy of 81.9%.

Future Works

As mentioned, there are several avenues for further exploration and enhancement in the field of converting slang texts into formal language, as well as in deep learning for short texts. The following points outline potential areas for continued research:

1. **Expand and Refine Rules:**
- Implement additional rules to enhance the formalization process, increasing the accuracy of current tools and rules.
- Consider contextual nuances, such as word position in a sentence or employing n-grams, to address errors in word conversion. For example, differentiating between "Selling their blood" and "Shed their blood."
- Explore sequence-to-sequence methods for sentence-level conversion, going beyond word-by-word transformation.

2. **Error Correction and Feature Preservation:**
- Investigate potential errors introduced by formalizing rules, especially in emotion-related texts. For instance, rules removing letter repetitions may inadvertently discard valuable features for emotion classification.
- Modify specific rules to improve the accuracy of emotion classification and ensure that important features are retained.

3. **Utilize NLP Tools for Improved Classification:**
- Leverage basic natural language processing tools, such as part-of-speech tagging and noun entity recognition, to enhance classifier performance.
- Assign higher weights to adjectives by incorporating part-of-speech tagging into feature vectors.
- Use noun entity recognition to remove nouns, allowing the model to grasp emotions from sentence style and context rather than learning specific nouns.

4. **Train Advanced Models in Persian:**
- Develop and train advanced models in Persian, akin to pre-trained models based on BERT available in English and other languages.
- Utilize large conversational datasets to train new models that could improve performance across various tasks, including emotion classification.
- Combine existing tagged data in Persian with additional datasets like Arman data to create a more robust model, especially in detecting allusions.

Continuing research in these areas could contribute to the refinement and advancement of tools and models, addressing challenges and optimizing performance in the analysis of emotions and the conversion of colloquial texts into formal language in the Persian language context.

Author Contributions

Each author role in the research participation must be mentioned clearly.

Example:

M. Khazeni proposed the problem in the Persian NLP domain and designed the research roadmap. M. Khazeni Crawled the data from scratch. M. Khazeni, M. Heydari carried out the data analysis. M. Khazeni, M. Heydari interpreted the results and wrote the manuscript. A. Albadvi Supervised the entire study.

Acknowledgment

The authors gratefully acknowledge the support and guidance of Professor Amir Albadvi for his work and supervision on proposing the research problem.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

In the final section of the article, abbreviations and their corresponding full forms are provided.

NLP	Natural Language Processing
LSTM	Long-Short Term Memory
BiLSTM	Bidirectional Long-Short Term Memory
CNN	Convolutional Neural Networks
BLEU	Bilingual Evaluation Understudy
BERT	Bidirectional Encoder Representations from Transformers
ELMO	Embeddings from Language Model
PSC	Persian Slang Text Convertor
MSA	Modern Standard Arabic
PLMs	Pre-trained Language Models
RMSE	Root Mean Squared Error
GloVe	Global Vectors for Word Representation
SVM	Support Vector Machine
MLP	Multilayer Perceptron
TF	Term-Frequency

References

[1] N. Armin, M. Shamsfard, “converting Persian colloquium text to formal by n-grams,” in Computer Society of Iran. for statistical machine translation, in Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP): 1724-1734, 2011.

[2] M. Heydari, “Sentiment analysis challenges in persian language,” arXiv Prepr. arXiv1907.04407, 2019.

[3] S. Zobeidi, M. Naderan, S. E. Alavi, “Opinion mining in Persian language using a hybrid feature extraction approach based on convolutional neural network,” Multimed. Tools Appl., 78(22): 32357-32378, 2019.

[4] B. Liu, L. Zhang, “A survey of opinion mining and sentiment analysis,” in Mining Text Data, C. C. Aggarwal and C. Zhai, Eds. Boston, MA: Springer US, pp. 415-463, 2012.

[5] B. Pang, L. Lee, “Opinion mining and sentiment analysis,” Found. Trends Inf. Retr., 2(1–2): 1-135, 2008.

[6] G. Song, Y. Ye, X. Du, X. Huang, S. Bie, “Short text classification: a survey,” J. Multimed., 9(5): 635-643, 2014.

[7] A. Naemi, M. Mansourvar, M. Naemi, B. Damirchilu, A. Ebrahimi, U. Kock Wiil, “Informal-to-formal word conversion for persian language using natural language processing techniques,” ACM Int. Conf. Proceeding Ser., 19: 1-7, 2021.

[8] V. Tajalli, F. Kalantari, M. Shamsfard, "Developing an informal-formal persian corpus," arXiv preprint arXiv:2308.05336, 2023.

[9] M. S. Rasooli, et al., "Automatic standardization of colloquial persian," arXiv preprint arXiv:2012.05879, 2020.

[10] M. Mazoochi, et al., "Constructing colloquial dataset for persian sentiment analysis of social microblogs," arXiv preprint arXiv:2306.12679, 2023.

[11] M. Adibian, S. Momtazi, "Using transformer-based neural models for converting informal to formal text in persian," Lang. Ling., 18(35): 47-69, 2022.

[12] Z. Bokaee Nezhad, M. A. Deihimi, "Sarcasm detection in persian," J. Inf. Commun. Technol., 20(1): 1-20, 2021.

[13] P. Golazizian et al., "Irony detection in Persian language: A transfer learning approach using emoji prediction," in Proc. Twelfth Language Resources and Evaluation Conference, 2020.

[14] M. Mirzarezaee, M. M. Pedram, "Improving polarity identification in sentiment analysis using sarcasm detection and machine learning algorithms in persian tweets," J. Inf. Commun. Technol. 53(14): 14-23, 2023.

[15] F. Najafi-Lapavandani, M. H. Shirali-Shahreza, “Humor detection in persian: a transformers-based approach,” Int. J. Inf. Commun. Technol. Res., 15(1): 56-62, 2023.

[16] A. K. Sharma, S. Chaurasia, D. K. Srivastava, “Sentimental short sentences classification by using CNN deep learning model with fine tuned Word2Vec,” Procedia Comput. Sci., 167: 1139-1147, 2020.

[17] P. F. Muhammad, R. Kusumaningrum, A. Wibowo, “Sentiment analysis using Word2vec and long short-term memory (LSTM) for Indonesian hotel reviews,” Procedia Comput. Sci., 179: 728-735, 2021.

[18] L. Ouchene, S. Bessou, “FastText embedding and LSTM for sentiment analysis: An empirical study on algerian tweets,” in Proc. 2023 International Conference on Information Technology (ICIT): 51-55, 2023.

[19] A. Patel, A. Kapoor, M. Mahato, S. Raut, B. B. Sinha, “Enhancing rumour detection: A hybrid deep learning approach with ELMo embeddings & CNN,” in Proc. 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), 2: 1-6, 2024.

[20] M. Farahani, M. Gharachorloo, M. Farahani, M. Manthouri, “ParsBERT: Transformer-based model for persian language understanding,” ArXiv, vol. abs/2005.1, 2020.

[21] T. Pires, E. Schlinger, D. Garrette, “How multilingual is multilingual BERT?,” arXiv Prepr. arXiv1906.01502, 2019.

[22] S. Mihi, B. Ait Benali, N. Laachfoubi, “Automatic sarcasm detection in Arabic tweets: resources and approaches,” J. Intell. & Fuzzy Syst., 45(6): 9483-9497, 2023.

[23] C. I. Eke, A. A. Norman, L. Shuib, “Context-based feature technique for sarcasm identification in benchmark datasets using deep

learning and BERT model," IEEE Access, 9: 48501-48518, 2021.

[24] F. Shatnawi, M. Abdullah, M. Hammad, M. Al-Ayyoub, "Comprehensive study of pre-trained language models: detecting humor in news headlines," Soft Comput., 27(5): 2575-2599, 2023.

[25] I. Annamoradnejad, G. Zoghi, "ColBERT: Using BERT sentence embedding in parallel neural networks for computational humor," Expert Syst. Appl., 249: 123685, 2024.

[26] S. M. Sadjadi, Z. Rajabi, L. Rabiei, M. S. Moin, "FarSSiBERT: A novel transformer-based model for semantic similarity measurement of persian social networks informal texts," arXiv Prepr. arXiv2407.19173, 2024.

[27] P. Falakaflaki, M. Shamsfard, "Formality style transfer in persian," arXiv Prepr. arXiv2406.00867, 2024.

[28] S. M. S. Dashti, A. Khatibi Bardsiri, M. Jafari Shahbazzadeh, "PERCORE: A deep learning-based framework for persian spelling correction with phonetic analysis," Int. J. Comput. Intell. Syst., 17(1): 1-23, 2024.

[29] E. Kebriaei et al., "Persian offensive language detection," Mach. Learn., 113(7): 4359-4379, 2024.

[30] Y. Z. Vakili, A. Fallah, S. Zakeri, "Enhancing sentiment analysis of persian tweets: A transformer-based approach," in Proc. 10th International Conference on Web Research (ICWR): 226-230, 2024.

Biographies

Mohsen Khazeni received his B.Sc degree in Computer Software Engineering, Iran University of Science and Technology and received his M.Sc. degree in Information Technology Engineering, Tarbiat Modares University, Tehran, Iran. His research interests are Natural Language Processing, Deep Learning, and Social Network Analysis.

- Email: m.khazeni@modares.ac.ir
- ORCID: NA
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://researchgate.net/profile/Mohsen-Khazeni-2

Mohammad Heydari received his B.Sc. degree in Computer Software Engineering, Technical and Vocational University, Tehan, Iran and received his M.Sc. degree in Information Technology Engineering, Tarbiat Modares University, Tehran, Iran. His research interests include Machine Learning, Big Data Engineering, and Graph Neural Networks.

- Email: m_heydari@modares.ac.ir
- ORCID: 0000-0002-7650-5924
- Web of Science Researcher ID: KZU-4848-2024
- Scopus Author ID: NA
- Homepage: https://scholar.google.com/citations?user=XBp6ipEAAAAJ&hl=en

Amir Albadvi is Full Professor of Information Systems. He received his Ph.D. from London School of Economics (LSE) in London and eventually earned his spot as technology thought leader for extra-large technology transformation projects with over 20 years of experiences in IT transformation and e-Strategy. After 10 years career as a change professional and winning the prestigious IT Deployment Award for his contribution in IT implementation, he decided it was time for a change of scenery (and weather) and moved to Beautiful British Columbia, Vancouver where he was offered visiting professor position at UBC and Victoria University. In addition, he focused on start-up ecosystem in the region, established Parallax Solutions Enterprise for technology and management consulting. Dr. Albadvi is now involved in another initiative in design thinking and technology-based innovation as team DNA named "Albadvi & Associates". A novel platform to promote social entrepreneurship among young generations.

- Email: albadvi@modares.ac.ir
- ORCID: 0000-0002-7758-9920
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://modares.ac.ir/~albadvi

How to cite this paper:
M. Khazeni, M. Heydari, A. Albadvi, "Persian slang text conversion to formal and deep learning of persian short texts on social media for sentiment classification," J. Electr. Comput. Eng. Innovations, 13(1): 27-42, 2025.

DOI: 10.22061/jecei.2024.10745.731

URL: https://jecei.sru.ac.ir/article_2172.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

A Fast and Accurate Yield Optimization Method for Designing Operational Amplifier Using Multi-Objective Evolutionary Algorithm Based on Decomposition

A. Yaseri [1], M. H. Maghami [2,*], M. Radmehr [1]

[1] Department of Electrical Engineering, Sari Branch, Islamic Azad University, Sari, Iran.

[2] Research Laboratory for Integrated Circuits, Faculty of Electrical Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.

Article Info

Article History:
Received 06 May 2024
Reviewed 20 July 2024
Revised 13 August 2024
Accepted 24 August 2024

Keywords:
Critical analysis
Multi-Objective evolutionary algorithm
Monte Carlo simulations
Improved MOEA/D
Yield optimization

[*]Corresponding Author's Email Address:
mhmaghami@sru.ac.ir

Abstract

Background and Objectives: In recent years, the electronics industry has experienced rapid expansion, leading to increased concerns surrounding the expenses associated with designing and sizing integrated circuits. The reliability of these circuits has emerged as a critical factor influencing the success of production. Consequently, the necessity for optimization algorithms to enhance circuit yield has become increasingly important. This article introduces an enhanced approach for optimizing analog circuits through the utilization of a Multi-Objective Evolutionary Algorithm based on Decomposition (MOEA/D) and includes a thorough evaluation. The main goal of this methodology is to improve both the speed and precision of yield calculations.
Methods: The proposed approach includes generating initial designs with desired characteristics in the critical analysis phase. Following this, designs that exceed a predefined yield threshold are replaced with the initial population that has lower yield values, generated using the classical MOEA/D algorithm. This replacement process results in notable improvements in yield efficiency and computational speed compared to alternative Monte Carlo-based methods.
Results: To validate the effectiveness of the presented approach, some circuit simulations were conducted on a two-stage class-AB Op-Amp in 180 nm CMOS technology. With a high yield value of 99.72%, the approach demonstrates its ability to provide a high-speed and high-accuracy computational solution using only one evolutionary algorithm. Additionally, the observation that modifying the initial population can improve the convergence speed and yield value further enhances the efficiency of the technique. These findings, backed by the simulation results, validate the efficiency and effectiveness of the proposed approach in optimizing the performance of the Op-Amp circuit.
Conclusion: This paper presents an enhanced approach for analog circuit optimization using MOEA/D. By incorporating critical analysis, it generates initial designs with desired characteristics, improving yield calculation efficiency. Designs exceeding a preset yield threshold are replaced with lower yield ones from the initial population, resulting in enhanced computational speed and accuracy compared to other Monte Carlo-based methods. Simulation results for a two-stage class-AB Op-Amp in 180 nm CMOS technology show a yield of 99.72%, highlighting the method's effectiveness in achieving high speed and accuracy with a single evolutionary algorithm.

Introduction

With the continuous development of the electronics industry in recent decades, there has been a growing concern regarding the cost associated with designing and sizing integrated circuits. Consequently, the yield value of these circuits has garnered significant attention. Yield is defined as the proportion of products that satisfy all

design constraints to the total number produced [1]-[3], and plays a crucial role in determining the overall success of circuit production. Therefore, the development of optimization algorithms that effectively enhance the yield value has become of utmost importance. The process of yield optimization typically encompasses several key steps, as shown in Fig. 1 [4]. These steps involve specifying design variables while considering desired constraints, generating initial designs using dedicated design tools based on the defined specifications, evaluating the yield value of the designs through simulation tools, iterating if the desired yield value is not achieved, and considering the optimization process complete once the desired yield value is attained.

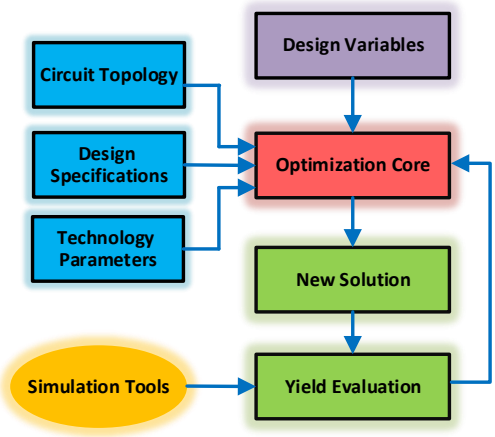


Fig. 1: General flowchart for yield optimization.

In general, yield optimization methods can be broadly categorized into two main categories [5]: statistical methods, such as Monte Carlo (MC) based methods and response-surface-based (RSB) methods, and non-statistical methods, including corner-based methods and performance-specific worst-case design (PSWCD) methods. Non-statistical methods offer advantages such as a lower number of simulations, simplicity, and faster sensitivity analysis. However, they have weaknesses such as design overhead at corner points and lower accuracy in estimation [5]. On the other hand, statistical methods provide good accuracy in calculating yield value and perform better at corner points compared to non-statistical methods. However, their main weakness lies in the large number of simulations required, which can be time-consuming [5].

Typically, yield optimization methods aim to maximize the yield value while minimizing computational time. Recent efforts have been made to optimize yield using evolutionary algorithms, which have shown promising results in increasing yield [6]-[9]. Additionally, some new methods have been proposed for yield estimation using machine learning instead of SPICE simulations including the work presented in [10]. However, it is important to note that this technique may not apply to all types of circuits, as in this method, numerous designs are generated from a circuit, and then a machine learning algorithm is trained using these designs. Therefore, when faced with other circuits that have not been trained by the algorithm, this method may not provide suitable responses.

The MC method is widely acknowledged as the most precise and commonly used approach for simulations. However, it does have a significant drawback, which is the extensive number of simulation iterations required, resulting in longer optimization process times. To address this challenge, various methods have been proposed to improve the efficiency of MC-based techniques. Examples of such methods include Latin Hypercube Sampling (LHS) [11] and Quasi MC simulation (QMC) [12]. The primary objective of speeding up MC simulations is to reduce the computational budget while maintaining yield accuracy. Therefore, yield optimization algorithms need to fulfill the following conditions: reducing the number of simulations, minimizing computational steps, and improving overall efficiency.

In this work, a combination of critical analysis and a Multi-Objective Evolutionary Algorithm [6], [13] based on Decomposition (MOEA/D) [14]-[15] is introduced. Indeed, to improve the efficiency of the optimization process, the conventional MOEA/D framework has been adapted by integrating critical analysis. This integration of critical analysis into the MOEA/D framework results in reduced simulation time and increased efficiency. The use of critical analysis helps to identify critical solutions that have a significant impact on yield improvement, while non-critical solutions are separated. To refine the simulation process further, this paper additionally utilizes the integration of optimal computing budget allocation (OCBA) [6] alongside critical analysis. This strategy controls the number of simulations needed for each design, thereby minimizing the simulation budget. As a result, the entire optimization process is accelerated. Pole-Zero analysis is also conducted to assess the stability of the circuit. Furthermore, this study considers a broader range of design characteristics compared to existing methods. In summary, the contributions of this work can be summarized as follows: enhancement of the classical MOEA/D method through the combination of critical analysis and MOEA/D, improvement of the yield value by replacing designs generated in both critical analysis and classical MOEA/D, acceleration of the yield calculation by reducing computational steps, and inclusion of circuit stability analysis and consideration of additional design characteristics.

The remainder of this paper is structured as follows. Initially, the background knowledge pertinent to this

study will be discussed in detail. Next, the specifics of the proposed method will be explored. This will be followed by an examination of the simulation results derived from the experiments. To wrap up, the key findings of the research presented will be summarized.

Background Knowledge of the Presented Technique

Since the proposed approach for yield optimization in this work is established based on OCBA, Multi-Objective Optimization (MOO), and MC simulation, these techniques are briefly reviewed in the following sub-sections. It should be noted that in the process of yield optimization, the goal is to identify a point that maximizes the yield value [5]. Therefore, the formula for calculating the yield can be expressed as follows:

$$d^* = \arg_{d \in D} \max\{Y(d)\} \quad (1)$$

where d is the design parameters such as transistor's dimensions, resistor and capacitor values, bias voltages, and current values. For each design parameter, an acceptable range of variations (upper and lower bands) is selected. The selection of this range of acceptable values is dependent on design knowledge, technological processes, or user preferences. In (1), the design space denoted by D. d^* represents the optimal point within the design space D that leads to maximizing the yield value. It is important to note that maximizing the yield is not always the goal; in some cases, the opposite is pursued, and the objective is to minimize the yield value (e.g., chip area in the case of yield-aware sizing). Accordingly:

$$Y(d) = E\{YS(d, s, \theta) \mid pdf(s)\} \quad (2)$$

where E is the expected value, θ indicates the environmental variables, and s represents the space of statistical parameters. In the case where all specifications are satisfied, $\mathrm{YS(d, s, \theta)}$ is set to 1, but if not, $\mathrm{YS(d, s, \theta)}$ is set to 0.

Optimal Computing Budget Allocation

The OCBA is one of the popular methods for ranking and selection in optimization methods [16]. In this method, the necessary number of iterations for each design is intelligently allocated based on the calculated mean value and variance for the designs. This enables a substantial reduction in the simulation budget by judiciously assigning simulation iterations to each design. With OCBA, the subsequent simulation step aims to identify the best solution by maximizing the likelihood of its discovery.

Multi-Objective Optimization

In the MOO technique, unlike single-objective optimization, multiple objective functions are employed to attain more accurate optimal solutions [17]. Consequently, in MOO, a set of solutions is obtained based on predefined objective functions. The optimization of multi-objective analog circuits is grounded in the use of multi-objective evolutionary algorithms (MOEAs). The MOO equation is given by:

$$\begin{aligned} &\text{Min/max } f_1(x), f_2(x), \ldots, f_n(x) \\ &\text{Subject to: } x \in U \end{aligned} \quad (3)$$

where the number of objective functions indicated by n , U is a set of feasible solutions, in this case, and $f_n(x)$ represents nth objective function. There are two types of object operations: min/max and x represents the solution. The proposed approach uses the MOEA/D multi-objective algorithm, which is discussed later.

Monte-Carlo Simulation

MC simulation delineates the process of translating uncertainties from a model's input to uncertainties in its output [6]-[9]. By employing statistical sampling, the MC method offers approximations for quantitative problems, facilitating an explicit and quantitative simulation of uncertainty. In MC simulation, inputs are designated as probability distributions, allowing for a clear representation of uncertainty. The predictability of system performance becomes uncertain when the inputs describing a system are uncertain. The outcome of any analysis involving inputs represented by probability distributions is likewise presented as a probability distribution. Typically, more than 1000 simulations are executed in the MC method, with each run referred to as a realization. Each realization involves sampling the distribution of each uncertain parameter, and selecting a random value for each parameter. Subsequently, specific input parameters are employed to simulate the system over time. This simulation yields performance metrics for the system. Ultimately, the system will traverse a potential path, and outputs are presented in the form of probability distributions.

Presented Method

After brief introduction of MC simulation, OCBA, and MOO technique, the presented method is described in this section. Fig. 2 shows the flowchart of the proposed method, which uses improved MOEA/D for yield optimization. Referred to Fig. 2, the combination of MOEA/D and critical analysis is employed to improve the classical MOEA/D and each is described separately here.

Critical Analysis

As previously mentioned, the MC method stands out as one of the most widely used techniques for calculating yield. Nevertheless, a significant challenge associated with this method is the extensive number of simulations, which can lead to system slowdown.

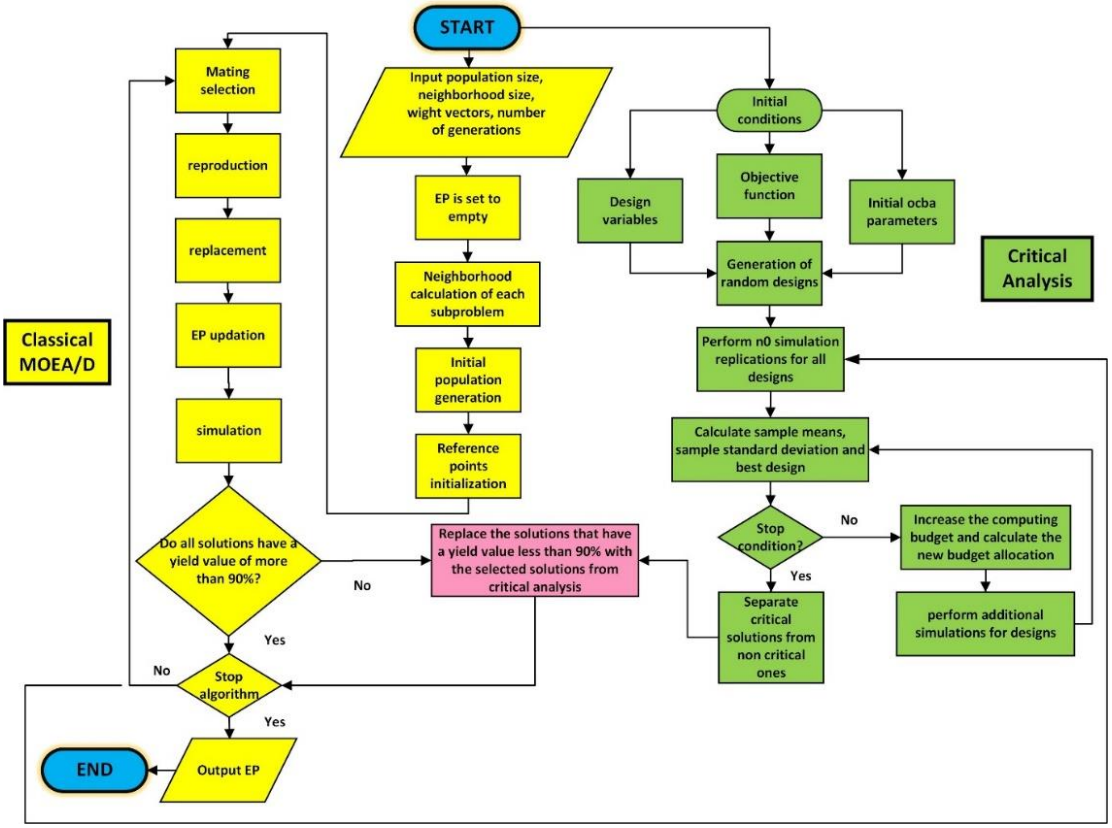


Fig. 2: Flowchart of the proposed approach for yield optimization.

A proposed solution to expedite the MC method involves reducing or eliminating simulation iterations assigned to non-critical solutions [6]. Critical solutions refer to designs that have a substantial impact on increasing the yield value, while non-critical solutions have a lesser effect on this calculation. Accordingly, the OCBA technique is employed to identify critical and non-critical solutions, intelligently allocating an appropriate number of simulation iterations to each [18], [19]. Consequently, this method of budget allocation between solutions leads to a reduction in computational time. The utilization of OCBA also contributes to a decrease in the yield estimator's variance. As a result, more simulation iterations are directed toward critical solutions, minimizing the time spent on non-critical ones. The details of this approach are outlined in Algorithm 1 [6].

Based on algorithm 1, T , K and b are the total budget of simulations, the number of competing designs and the best design, respectively. The variance and mean of the k solutions indicated by σ_i^2 , J_i . Every candidate solution runs n_0 simulations initially. N_i and N_j are the number of simulation replications allocated to design i , j respectively and $\delta_{b,i} = J_b - J_i$.

Multi-Objective Evolutionary Algorithm based on Decomposition

The MOEA/D [20], [21], developed by Jang in 2007, decomposes a multi-objective problem into multiple subproblems and optimizes them concurrently. This method combines the objective functions using a weight vector defined for each subproblem. Each individual in the population represents a solution obtained through an aggregation vector, with the population size matching the number of problems, so each solution corresponds to a member of the population. By the end of the search, each problem yields a Pareto-front answer, which is the best solution discovered for the respective subproblem constituting the population in each generation. The proximity between aggregation vectors establishes neighbor relations among sub-problems. During the search, a neighborhood member plays a role in contributing to the solution of the problem, marking this stage as the collaboration stage. For each subproblem following the current one in the neighborhood, a weighted aggregation vector is provided. If the solutions to the neighbors' problems surpass the original answers, these should replace the initial answers, constituting the competition stage. The processes of cooperation and competition are applied to all sub-problems, ensuring a continuous exchange of information between neighbors.

The subproblems within its neighborhood are leveraged to optimize each subproblem in the algorithm. The general framework of the MOEA/D can be considered as follows:

Algorithm 1. Critical Analysis

Initializing design variables and specifying a reasonable range for each of the design specifications.

Initializing $K, \mathrm{T}, \mathrm{n}_0, \Delta$ and let $l \leftarrow 0$. Then performing n_0 simulation replications for all designs.

$$N_1^l = N_2^l = \ldots = N_k^l = n_0$$

Calculate sample means and standard deviation, then finding the best design according to $b = \arg\min_i (J_i)$.

Construct solution set by critical solutions.

If the termination condition is satisfied, then, end the algorithm. Otherwise, increase the computing budget by Δ and calculate the new budget allocation.

$N_1^{l+1}, N_2^{1+1}, \ldots, N_k^{l+1}$

$$\frac{N_i}{N_j} = \left(\frac{\sigma_i / \delta_{b,i}}{\sigma_j / \delta_{b,j}} \right)^2$$

$$N_b = \sigma_b \sqrt{\sum_{i=1, i \neq b}^{k} \frac{N_i^2}{\sigma_i^2}}$$

$i, j \in \{1, 2, \ldots, k\}$ *and* $i \neq j \neq b$

Perform additional $\max(N_i^{l+1} - N_i^l, 0)$ simulations for the design i, $i = 1, 2, \ldots, k; l \leftarrow l + 1$

Go to step 3

$$\begin{aligned} &\text{Minimize}\,(f_1(x), f_2(x), \ldots, f_m(x)) \\ &\text{Subject to}\, g(x) \geq 0, X_L < x < X_H \end{aligned} \tag{4}$$

A given objective function is called fi(x), i = 1...m, m is the number of objectives, and x is the design variable. There are X_L and X_H for the lower and upper bounds, respectively. The vector g(x) ≥ 0 represents the design constraints.

Each non-dominated solution to the multi-objective optimization problem aligns with an optimal single-objective solution when utilizing a specific weight vector. Within MOEA/D, distinct weight vectors guide diverse searches across various regions of the objective space, forming a comprehensive set of weight vectors. In the context of a multi-objective optimization problem, the Tchebycheff method allows for the definition of N subproblems. In this method, the objective function of the jth (j=1, 2, ..., N) sub-problem is as follows:

$$g^{te}(x \mid \lambda^j, z^*) = \max_{1 \leq i \leq m} \left\{ \lambda_i^j \mid f_i(x) - z_i^* \mid \right\} \tag{5}$$

where $\lambda^j = \left(\lambda_1^j, \ldots, \lambda_m^j \right)^T$ demonstrates a weight vector, $z^* = (z_1^*, \ldots, z_m^*)^T$ represents the vector of reference points. For each Pareto optimal point x* there exists a weight vector so that x* is the optimal solution of (5). There are related references such as the work presented in [15] that describe methods for determining the weight vector. There are Pareto optimal solutions to the problem of (4) for each solution of (5). Some methods [14], [20] have pointed to the weakness of the MOEA/D when facing more complex circuits and a higher number of objectives in comparison with the NSGA-II. However, others [12], [21] have emphasized the capability of the MOEA/D in multi-objective optimization problems. Given this, the main goal of the proposed method in this article is to enhance the convergence speed of the algorithm while upholding a high level of accuracy. This goal is achieved through an effective integration of the critical analysis method and the MOEA/D algorithm. The comparison of the yield histogram between the proposed method and the classic MOEA/D, serves as evidence of the method's efficacy. Furthermore, in contrast to the CSNM [6] employing the NSGA-III, the proposed method, by integrating solution responses, demonstrates both remarkable speed and accuracy, as will be expounded upon in the upcoming simulation results section. It is essential to note that alternative methods, such as epsilon constraint methods or lexicographical methods, can also be considered valuable techniques for objective weighting.

The critical analysis section begins by generating a set of designs aiming to optimize the desired specifications, which serve as the objectives. These objectives include DC voltage gain, unity-gain bandwidth (UGBW), phase margin, common-mode rejection ratio (CMRR), total harmonic distortion (THD), output voltage swing, slew rate (SR), and power dissipation. Additionally, the stability of each solution is assessed through pole-zero analysis. The design variables governing the solutions encompass capacitor capacitance, transistor dimensions, the number of parallel transistors, and bias voltages. The designer selects the initial population size, the number of subproblems, the maximum iteration limit to conclude the algorithm, and values associated with critical analysis

parameters. In the multi-objective optimization algorithm, the goals involve the simultaneous minimization of THD and power consumption, while maximizing other specified objectives. It should be noted that as illustrated in Fig. 3, design parameters often trade off against each other, turning the design process into a multi-dimensional optimization. Successfully navigating these challenges requires a combination of intuition and experience to reach an acceptable compromise [22]. It seems that having too many objectives in an optimization problem hinders a multi-objective algorithm's ability to improve them simultaneously, especially if they all are correlated. Therefore, it may be beneficial to remove objectives that are of lesser importance in amplifier design. However, to achieve a highly effective design that excels in all aspects, it is crucial to consider all important objectives. In this particular design, parameters such as input resistance, output resistance, and noise were not included in the optimization problem to address the aforementioned issue.

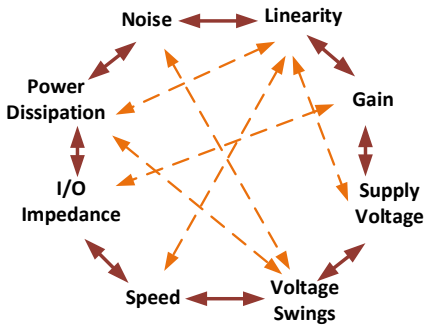


Fig. 3: Analog design octagon [22].

In this work, the optimization procedure begins by using critical analysis and OCBA to identify designs that meet the desired specifications. Concurrently, a decomposition-based optimization algorithm generates the initial population. A comparative analysis is then conducted to evaluate the production yield of the population generated by MOEA/D against that found through critical analysis. Solutions identified through critical analysis that exhibit higher yield values than those in the MOEA/D production population are subsequently substituted. At this stage, the population formed is anticipated to contain more optimal solutions than the initial production population. This iterative substitution process continues until the algorithm's stopping criteria are met. Specifically, the stopping criterion for the critical analysis phase is achieved when the cumulative number of simulations assigned to the designs meets or exceeds the total computational budget. Meanwhile, the termination criterion for the MOEA/D algorithm is reached when the predefined maximum number of iterations, set at the algorithm's inception, is attained.

It is important to note that if designs with yield values below the designer's desired threshold are replaced and the stop conditions are not met, these solutions will be redirected to the critical analysis section for reassessment of simulations. Through these iterations, the process helps generate a more optimal set of solutions in this phase. Conversely, if the designs produced in the MOEA/D section have yield values exceeding the desired threshold, there is no need for replacement, and the critical analysis step can be skipped. This approach not only accelerates computational speed but also improves production efficiency by preserving solutions generated in the MOEA/D stage during previous iterations.

Simulation Results

The proposed algorithm is tested on a two-stage class-AB Operational Amplifier (Op-Amp) shown in Fig. 4 [23] in a standard 0.18µm CMOS technology with a supply voltage of 1.8V. For performing MC simulations and evaluation of circuit performance parameters, MATLAB R2020 and Synopsys HSPICE are used, respectively. It should be noted that Op-Amp proposed in [23] is a fully differential two-stage amplifier employing a positive feedback technique and split-length transistors to increase the DC voltage gain without affecting the UGBW, stability, power dissipation, and output voltage swing compared to the conventional two-stage Op-Amp. A comprehensive analysis of the Op-Amp shown in Fig. 4 is provided in [23]. In Fig. 4, the first stage is a folded-cascade and the second stage is a common-source amplifier. Transistor pairs of $M_{16\text{-}19}$ and $M_{20\text{-}23}$ are used to build split-length transistors and by applying the output signal V_{out+} to the drain terminal of M_{22} and V_{out-} to the drain terminal of M_{18}, a positive feedback loop is created.

Fig. 5 illustrates the MATLAB-HSPICE link, which is employed for implementing the algorithm presented in this work. In this process, the user defines the design variables and circuit specifications in MATLAB. MATLAB then prepares the circuit netlist parameters and initiates the simulation of the circuit using HSPICE. Following the HSPICE circuit simulation, MATLAB scans the output file generated by HSPICE and extracts various metrics of the simulated circuit, such as SR, CMRR, and output voltage swing. This step involves MATLAB analyzing the simulation results and making adjustments to the circuit parameters in the subsequent iteration if the desired values for the output parameters are not achieved.

Table 1 shows the desired specifications for the Op-Amp shown in Fig. 4. The compensation capacitance, bias voltages, transistor dimensions, as well as parallel transistor count are the design variables in this work. According to the utilized technology file, the transistor's width ranges from 0.54µm to 100µm and their length can range from 0.18µm to 20µm. Moreover, the compensation capacitance which is utilized for stability concerns is set to be from 0.1pF to 10pF.

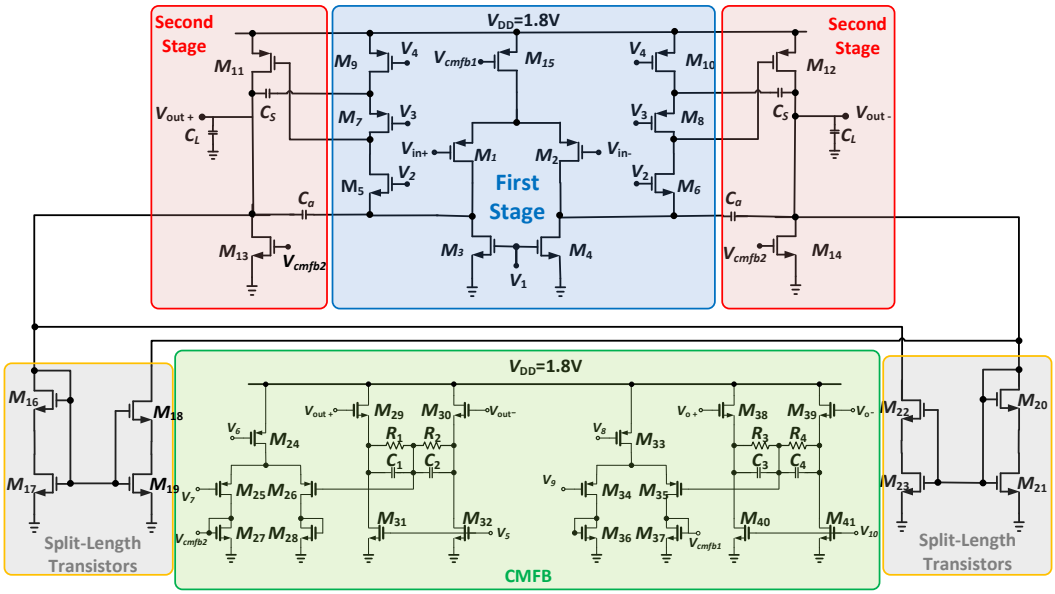


Fig. 4: Circuit schematic of the utilized two-stage class AB-OP-Amp [23].

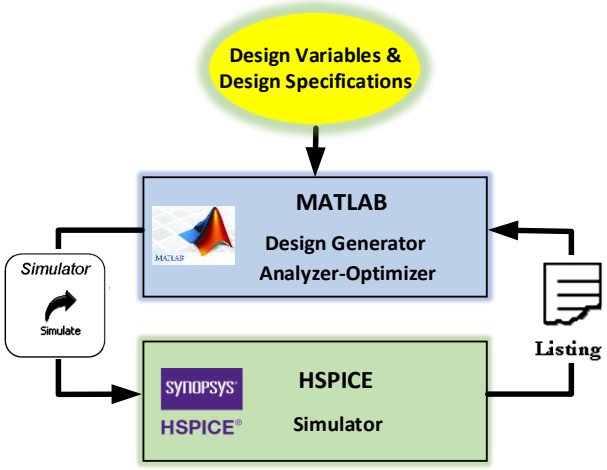


Fig. 5: Utilized MATLAB-HSPICE link.

Table 1: Desired specifications of the two-stage class AB-OP-Amp

Specifications	Desired value
DC Voltage Gain (dB)	80≥
Phase Margin (deg)	60°≤ PM≤ 80°
Power Dissipation (mW)	≤ 10
Slew Rate (V/µs)	600≥
Unity-Gain Bandwidth (MHz)	300≥
Common Mode Rejection Ratio (dB)	90≥
Power Supply Rejection Ratio (dB)	60≥
Output Voltage Swing (V)	2.5≥
Bandwidth (KHz)	1≥
Total Harmonic Distortion (dB)	≤ -30
Pole-Zero Analyze	Pole: Z=a+bj, Re(Z) < 0

In the presented work, critical analysis should be initiated by creating some random designs, based on the design specifications and design variables mentioned above. Next, critical analysis is used to select designs that comply with the design constraints. All the settings related to the critical analysis method are done exactly as the [24], [25], where n_0 is set to be 5 and $\Delta = 5$. Moreover, T is determined by:

$$T = M_1 \times sim_{ave} \tag{6}$$

where sim_{ave} represents the average budget for each candidate and M_1 denotes the number of critical solutions, which are set to be 50 and 100, respectively.

In the next step, random solutions are generated by the MOEA/D algorithm. Then, the necessary simulations are performed to evaluate the desired goals. The number of MOEA/D population is equal to 100 and also the maximum number of iterations to reach appropriate goals is equal to 100. Alternative critical analysis solutions with a yield value higher than 90% are substituted for solutions with yield values less than 90% in the set of the current MOEA/D population. If the set of solutions generated in the critical analysis section does not have a sufficient number of solutions with a yield value greater than the designer's desired value, solutions with a yield value greater than the solutions generated by the MOEA/D algorithm are replaced. At the end, the stop condition is checked and if the stop condition is not satisfied, the cycle of production and replacement of solutions will continue.

As mentioned above, after the replacement of the design produced in the MOEA/D section with a yield value of less than 90%, if the stop conditions are not met, this set of solutions will be sent to the critical analysis section

to reassign simulations to these designs. It is obvious that if the solutions created in the MOEA/D section have a yield value greater than the desired yield value by the designer, there is no need to replace the solutions. At this condition, the critical analysis step is removed at the next iteration. By replacing the selected solutions between two steps, the calculation speed and the yield value calculation accuracy will be increased.

Obtained from a presented algorithm, the values of the passive components, bias voltages, and transistors dimensions used in the two-stage class-AB Op-Amp can be found in Table 2.

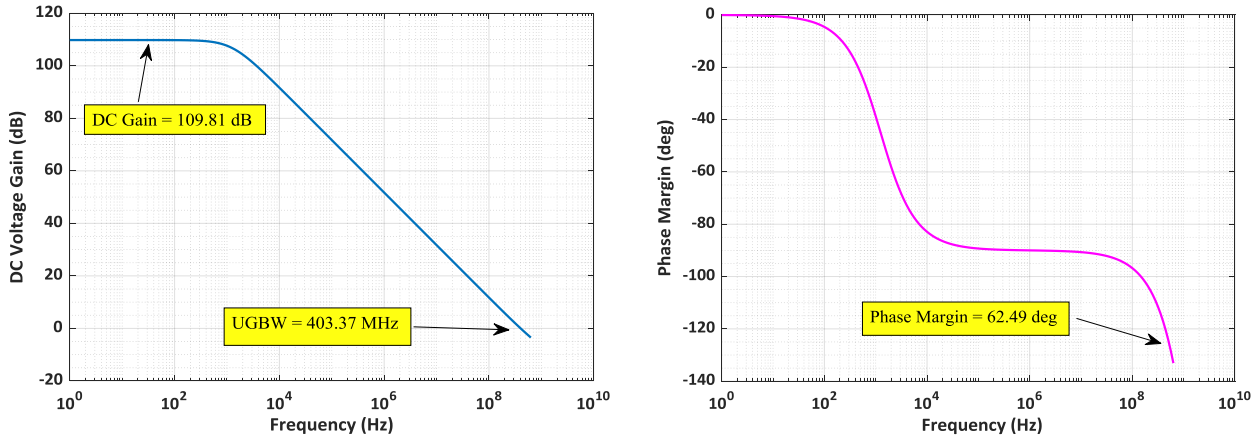
Moreover, Table 3 provides the simulation results for the DC gain, UGBW, phase margin, power dissipation, output voltage swing, CMRR, Power Supply Rejection Ratio (PSRR), amplifier Bandwidth (BW), THD, and SR of the designed two-stage class-AB Op-Amp in different process and temperature corners. As stated in Table 3, the simulation results for the two-stage class-AB Op-Amp indicate the following values: DC gain of 109.81dB, UGBW of 403.37MHz, phase margin of 62.49°, power dissipation of 7.94mW, output voltage swing of 3.4V, CMRR of 148.56dB, PSRR of 61.41dB, BW of 1.26 kHz, THD of -42.14dB, and SR of 667.93 V/µs. Fig. 6 illustrates the open-loop frequency responses of the two-stage class AB Op-Amp that has been designed using the values specified in Table 2. It's crucial to emphasize that the simulation results illustrated in Table 3 align with the expected outcomes. Additionally, Fig. 7, Fig. 8, and Fig. 9 show the plots for CMRR, PSRR, and output voltage swing of the designed Op-Amp, respectively. According to Fig. 10(a), the designed two-stage class-AB Op-Amp is utilized as a unity gain capacitor buffer to measure its slew rate [23], [26]. In this configuration, a square wave with 1Vpp amplitude and a frequency of 5 MHz was applied to the circuit, and the result is given in Fig. 10(b). The measured slew rate value of the two-stage class-AB Op-Amp is 667.93 V/µs.

Table 2: One solution of a two-stage class-AB Op-Amp (Fig. 4) based on transistor dimensions and passive components

Parameter	Value	Parameter	Value
$(W/L)_{1,2}$	$2\times 22.92\,\mu m/0.18\mu m$	$(W/L)_{38,39}$	$1\times 20.25\mu m/0.18\mu m$
$(W/L)_{3,4}$	$3\times 25.17\,\mu m/0.18\mu m$	$(W/L)_{40,41}$	$1\times 20.88\mu m/0.18\mu m$
$(W/L)_{5,6}$	$1\times 35.68\,\mu m/0.18\mu m$	C_s	$1.05pF$
$(W/L)_{7,8}$	$4\times 44.44\,\mu m/0.18\mu m$	C_L	$1pF$
$(W/L)_{9,10}$	$2\times 90.5\,\mu m/0.18\mu m$	C_a	$1pF$
$(W/L)_{11,12}$	$4\times 43.51\,\mu m/0.36\mu m$	$C_{1,2,3,4}$	$1.5\,pF$
$(W/L)_{13,14}$	$1\times 59.01\,\mu m/0.36\mu m$	$R_{1,2,3,4}$	$20K\Omega$
$(W/L)_{15}$	$5\times 29.4\,\mu m/0.18\mu m$	V_1	$0.6V$
$(W/L)_{16,17,20,21}$	$1\times 2\mu m/0.18\mu m$	V_2	$1V$
$(W/L)_{18,19,22,23}$	$1\times 2.58\mu m/0.18\mu m$	V_3	$0.77V$
$(W/L)_{24}$	$1\times 40.5\mu m/0.18\mu m$	V_4	$1.2V$
$(W/L)_{25,26}$	$1\times 22.03\mu m/0.18\mu m$	V_5	$0.685V$
$(W/L)_{27,28}$	$1\times 38.11\mu m/0.18\mu m$	V_6	$1.14V$
$(W/L)_{29,30}$	$2\times 10.43\mu m/0.18\mu m$	V_7	$0.276V$
$(W/L)_{31,32}$	$1\times 19.6\mu m/0.18\mu m$	V_8	$1.2V$
$(W/L)_{33}$	$1\times 44.2\mu m/0.18\mu m$	V_9	$0.5V$
$(W/L)_{34,35}$	$1\times 19.7\mu m/0.18\mu m$	V_{10}	$0.59V$
$(W/L)_{36,37}$	$1\times 4.37\mu m/0.18\mu m$		

Table 3: Specifications of the two-stage class-AB Op-Amp

Specifications	temperature corners		
	TT (27°C)	FF (-40°C)	SS (90°C)
DC-Gain (dB)	109.81	87	96.4
Phase Margin (°)	62.49	61.14	62.11
Power Dissipation (mW)	7.94	9.1	7.34
Slew Rate (V/µs)	667.93	843.87	575.1
UGBW (MHz)	403.37	511.6	321.65
CMRR (dB)	148.56	127.1	136.89
Output Swing (V)	3.4	3.4	3.4
THD (dB)	-42.14	-40.5	-41.9
PSRR (dB)	61.41	59.7	60
BW (KHz)	1.26	1.07	1.19



(a) (b)

Fig. 6: Frequency response of the two-stage class-AB Op-Amp: a) magnitude, b) phase.

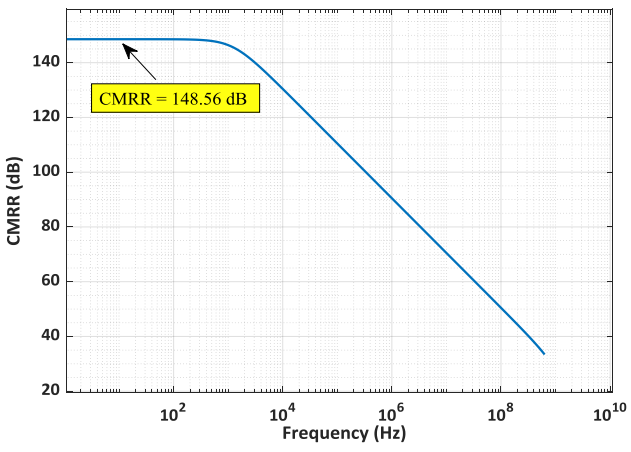


Fig. 7: CMRR behavior of the simulated two-stage class-AB Op-Amp.

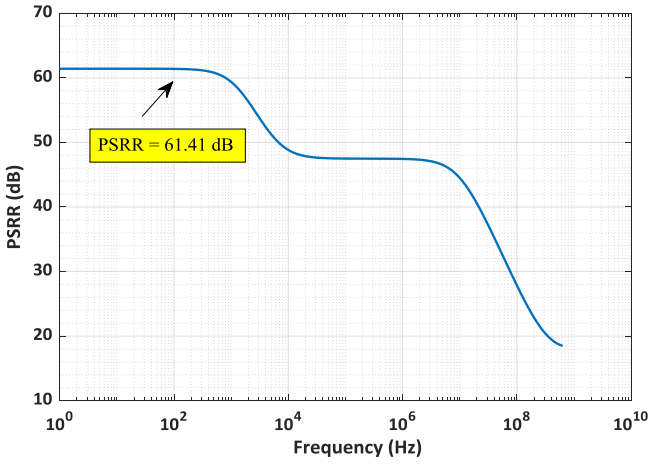


Fig. 8: PSRR behavior of the simulated two-stage class-AB Op-Amp.

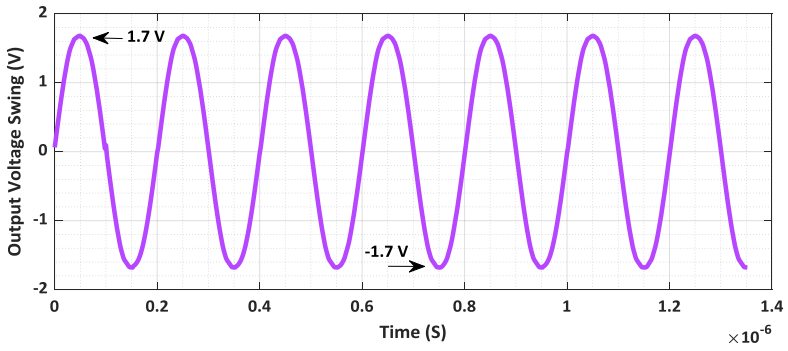
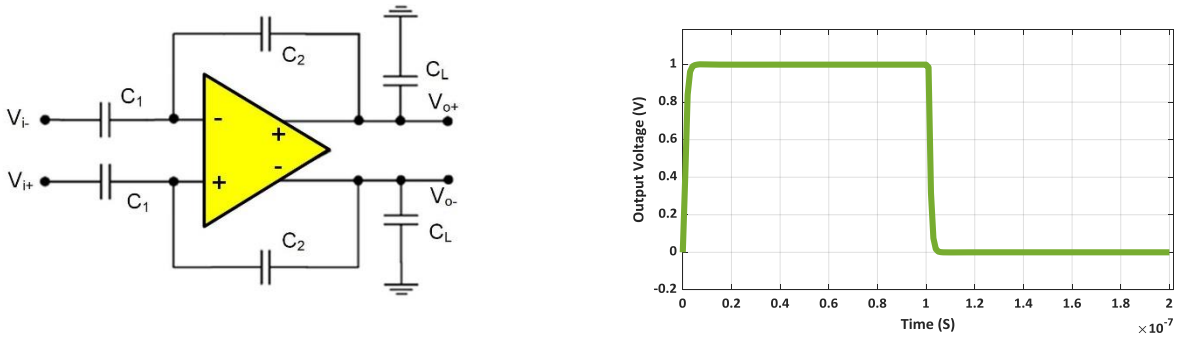
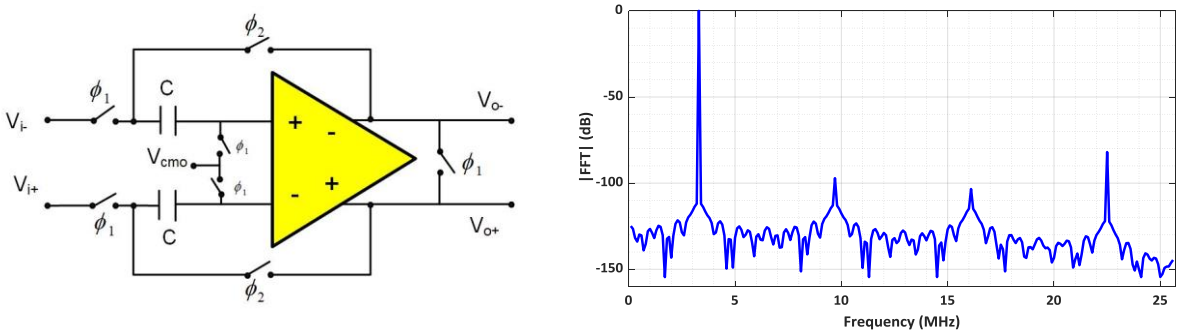


Fig. 9: Output Swing of the two-stage class-AB Op-Amp.



(a) (b)

Fig. 10: (a) Circuit schematic of a unity gain capacitive buffer [23], [26], (b) Op-amps large signal step responses.



(a) (b)

Fig. 11: (a) Circuit schematic of designed flip-around sample-and-hold [27], [28], (b) FFT plot of the output of the designed sample-and-hold.

Moreover, the designed two-stage class-AB Op-Amp in this work is used in a switched-capacitor flip-around sample-and-hold (SH) circuit shown in Fig. 11(a) [27], [28] to study its linearity in a closed-loop configuration. Fig. 11(b) depicts the circuit-level simulation outcome of the output voltage spectrum, showcasing the large signal transient response of the circuit to a 1Vpp input step voltage with two non-overlapping clocks at a frequency of 3.125MHz. According to the performed simulations, the output spectrum of the SHA shows a THD of -42.14dB.

The stability of the designed two-stage class-AB Op-Amp is thoroughly examined and verified through the implementation of Pole-Zero analysis, a powerful technique used to assess the system's stability characteristics. By analyzing the Pole-Zero plot, which is visually depicted in Fig. 12, valuable insight into the location and behavior of the poles and zeros of the Op-Amp's transfer function has been gained. This comprehensive analysis ensures that the Op-Amp operates within stable and desirable parameters, guaranteeing reliable and accurate performance in various applications.

To attain a more precise yield simulation, this study conducted the MC simulation with 2000 replications, as indicated by the values presented in Table 2. Fig. 13(a) shows a histogram of yields. According to Fig. 13(a), the

mean value for yield is 99.72%, with a standard deviation of 0.03%. In Fig. 13(b), a histogram of yields for the classic MOEA/D algorithm is depicted. According to Fig. 13(b), the mean value for yield is 89.95%, with a standard deviation of 0.03%. By comparing the two Figs, it is evident that in the proposed method, which combines critical analysis and MOEA/D, not only higher accuracy is achieved but also the issue of objective aggregation present in classic MOEA/D [14], [20] has been addressed.

To assess the effectiveness of the proposed method, the CSNM [6], Freeze-Thaw Bayesian optimization [30], and Mirzaei [7] algorithms were evaluated on the circuit shown in Fig. 4. To have a more accurate comparison between the methods examined, all of these methods have been implemented by the authors and the simulations performed on a workstation equipped with a CPU: Intel Core i7-4790K @4GHz, 16GB RAM, and a 64-bit operating system with an x64-based processor. The improved MOEA/D algorithm used in the proposed approach demonstrated faster performance compared to the other three algorithms, as shown in Table 4. Additionally, the proposed approach required fewer computational steps compared to the other three methods.

Based on these findings, it can be concluded that when combined with critical analysis, the proposed method can decrease the number of simulations needed for solutions with minimal effects on yield. Furthermore, replacing critical analysis with MOEA/D solutions can significantly enhance efficiency and reduce simulation time. However, it is worth noting that the CSNM, which utilizes OCBA, critical analysis, and two evolutionary algorithms, achieved a higher yield value compared to the proposed approach. So, the proposed approach, compared to three other existing methods, demonstrates lower complexity, fewer steps, and reduced computational time.

In Fig. 14, two diagrams related to the Pareto-front evaluation of the generated solutions (phase margin versus voltage gain, and UGBW versus voltage gain) are reported. As shown in Fig. 14(a) and according to Table 1, a voltage gain value greater than 80 dB and a phase margin range between 60 to 80 degrees are obtained. Also, the simulation results for the plot of UGBW versus voltage gain also comply with the conditions stated in Table 1. It should be noted that based on what is observable in Fig. 14, the simulation results show more scattering in regions with a voltage gain of around 110 dB, a phase margin of 63 degrees, and an UGBW of 400 MHz.

Table 4: Yield simulation results and the run time for different techniques applying to the two-stage class-AB Op-Amp of Fig. 4

Technique	Best	Worst	Mean	Run time (h)	Computer
Freeze–Thaw Bayesian [30]	99.24	99.13	99.23	19	CPU Intel Core i7-4790K @4GHz with 16GB RAM
CSNM [6]	99.97	98.77	99.87	20	
Mirzaei method [7]	99.11	98.91	99.01	22	
Proposed Method	99.82	99.62	99.72	16	

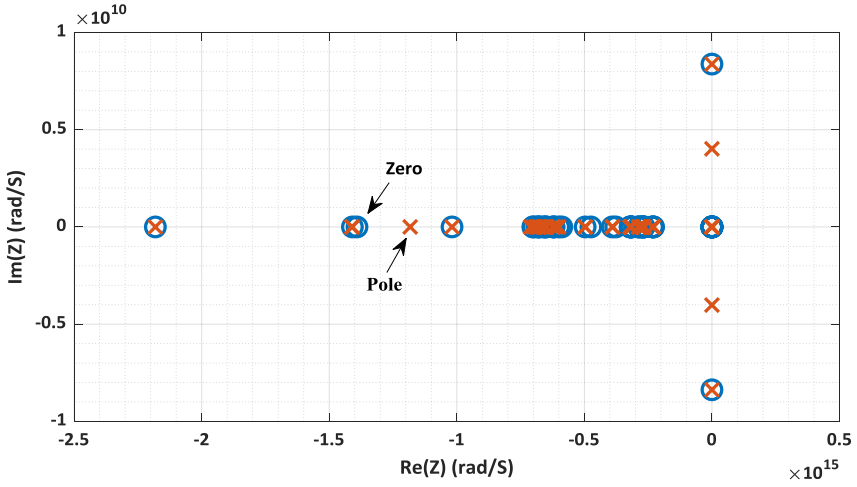
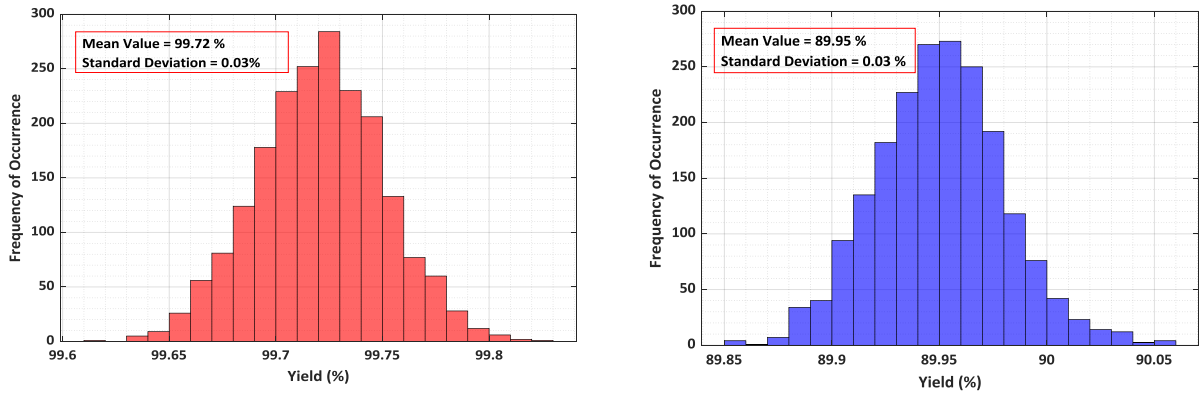
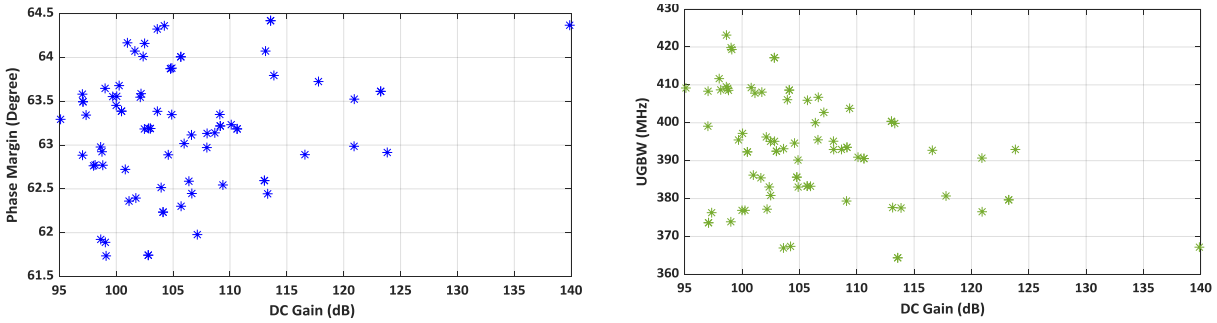


Fig. 12: Pole-Zero plot of the designed Op-Amp.



(a) (b)

Fig. 13: The yield histogram for the MC simulation with 2000 iterations of the utilized Op-Amp, (a) Propose approach, (b) Classic MOEA/D.



(a) (b)

Fig. 14: Pareto front of the generated solutions, (a) DC gain versus Phase Margin, (b) DC gain versus UGBS.

Conclusion

In this paper, an enhanced approach for MOEA/D based on Decomposition, utilizing critical analysis is presented to enhance the computational speed and accuracy of yield calculation in analog circuit optimization. The critical analysis generates initial designs with desired characteristics. Subsequently, designs surpassing a predefined yield threshold are replaced with the initial population having lower yield values, which is generated using the classical MOEA/D. This approach significantly improves yield efficiency and computational speed compared to other MC-based methods. The simulation results for a two-stage class-AB Op-Amp in 180 nm CMOS technology demonstrate a yield value of 99.72%. This computational approach stands out as a high-speed and high-accuracy technique, employing only one evolutionary algorithm. Furthermore, by modifying the initial population, improvements in both the convergence speed and yield value of the evolutionary algorithm have been observed. The efficiency of the proposed technique is validated through extensive simulation results.

Author Contributions

Conceptualization and design, A. Yaseri; formal analysis, A. Yaseri; software, A. Yaseri; investigation, M. H. Maghami.; writing—original draft preparation, A. Yaseri; writing—review and editing, M. H. Maghami. supervision, M. H. Maghami, and M. Radmehr.

Acknowledgment

This work was supported by Shahid Rajaee Teacher Training University under grant number 5973/89.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

MC	Monte Carlo
RSB	response-surface-based
PSWCD	performance-specific worst-case design
LHS	Latin hypercube sampling
QMC	Gaussian Monte Carlo simulation

MOEA/D	Multi-Objective Evolutionary Algorithm based on Decomposition
OCBA	optimal computing budget allocation
MOO	Multi-Objective Optimization
MOEA	multi-objective evolutionary algorithm
CA	Critical Analysis
UGBW	unity-gain bandwidth
CMRR	common-mode rejection ratio
THD	total harmonic distortion
SR	slew rate
PSRR	Power Supply Rejection Ratio
BW	Bandwidth

References

[1] E. D. Sandru, E. David, I. Kovacs, A. Buzo, C. Burileanu, G. Pelz, "Modeling the dependency of analog circuit performance parameters on manufacturing process variations with applications in sensitivity analysis and yield prediction," IEEE Trans. Comput.-aided Des. Integr. Circuits Syst., 41(1): 129-142, 2022.

[2] N. Mirzaie, H. Shamsi, G. S. Byun, "Automatic design and yield enhancement of data converters," J. Circuits Syst. Comput., 26(01): 1750018, 2017.

[3] S. Kondamadugula, S. R. Naidu, "Parameter-importance based Monte-Carlo technique for variation-aware analog yield optimization," in Proc. the 26th edition on Great Lakes Symposium on VLSI, 2016.

[4] M. Fakhfakh, E. Tlelo-Cuautle, Computational intelligence in analog and mixed-signal (AMS) and radio-frequency (RF) circuit design. Springer International Publishing, 2015.

[5] N. Mirzaie, H. Shamsi, G. S. Byun, "Yield-aware sizing of pipeline ADC using a multiple-objective evolutionary algorithm: Yield-aware sizing of pipeline ADC," Int. J. Circuit Theory Appl., 45(6): 744-763, 2017.

[6] A. Yaseri, M. H. Maghami, M. Radmehr, "A four-stage yield optimization technique for analog integrated circuits using optimal computing budget allocation and evolutionary algorithms," IET Comput. Digit. Tech., 2022.

[7] N. Mirzaie, G. S. Byun, "An optimal design methodology for yield-improved and low-power pipelined ADC," IEEE Trans. Semicond. Manuf., 31(1): 130-135, 2018.

[8] A. Canelas et al., "Hierarchical yield-aware synthesis methodology covering device-, circuit-, and system-level for radiofrequency ICs," IEEE Access, 9: 124152-124164, 2021.

[9] A. Canelas, R. Povoa, R. Martins, "FUZYE: A fuzzy c -means analog IC yield optimization using evolutionary-based algorithms," IEEE Trans. Comput. Aided Des. Integr. Circuits Syst., 39(1): 1-13, 2020.

[10] G. İslamoğlu, T. O. Çakıcı, Ş. N. Güzelhan, E. Afacan, G. Dündar, "Deep learning aided efficient yield analysis for multi-objective analog integrated circuit synthesis," Integration, 81: 322-330, 2021.

[11] M. Stein, "Large sample properties of simulations using Latin hypercube sampling," Technometrics, 29(2): 143, 1987.

[12] M. Pak, F. V. Fernandez, G. Dundar, "Comparison of QMC-based yield-aware pareto front techniques for multi-objective robust analog synthesis," Integration, 55: 357-365, 2016.

[13] K. Deb, S. Agrawal, A. Pratap, T. Meyarivan, "A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II," in Parallel Problem Solving from Nature PPSN VI, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 849-858, 2000.

[14] E. Sağlican, E. Afacan, "MOEA/D vs. NSGA-II: A comprehensive comparison for multi/many objective analog/RF circuit optimizations through a generic benchmark," ACM Trans. Des. Automat. Electron. Syst., 29(1): 1-23, 2024.

[15] Q. Xu, Z. Xu, T. Ma, "A survey of multiobjective evolutionary algorithms based on decomposition: Variants, challenges and future directions," IEEE Access, 8: 41588-41614, 2020.

[16] D. Q. Mayne, "Yu-chi ho, Qian-Chuan Zhao and Qing-Shan Jia: Ordinal optimization: Soft optimization for hard problems: Springer, 2007," J. Optim. Theory Appl., 145(3): 613-615, 2010.

[17] K. Deb, Multi-Objective Optimization using Evolutionary Algorithms. Chichester, England: John Wiley & Sons, 2014.

[18] C. H. Chen, J. Lin, E. Yücesan, S. E. Chick, "Simulation budget allocation for further enhancing the efficiency of ordinal optimization," Discrete Event Dyn. Syst., 10(3): 251-270, 2000.

[19] I. Guerra-Gomez, E. Tlelo-Cuautle, L. G. de la Fraga, "OCBA in the yield optimization of analog integrated circuits by evolutionary algorithms," in Proc. 2015 IEEE International Symposium on Circuits and Systems (ISCAS), 2015.

[20] N. Srinath, I. O. Yilmazlar, M. E. Kurz, K. Taaffe, "Hybrid multi-objective evolutionary meta-heuristics for a parallel machine scheduling problem with setup times and preferences," Comput. Ind. Eng., 185: 109675, 2023.

[21] M. Nohtanipour, M. H. Maghami, M. Radmehr "A placement and routing method for layout generation of CMOS operational amplifiers using multi-objective evolutionary algorithm based on decomposition," Inf. MIDEM, 51(3), 2021.

[22] B. Razavi, Design of Analog CMOS Integrated Circuits, 2nd ed. New York, NY: McGraw-Hill Professional, 2016.

[23] S. M. Anisheh, H. Shamsi, M. Mirhassani, "Positive feedback technique and split-length transistors for DC-gain enhancement of two-stage op-amps," IET Circuits Devices Syst., 11(6): 605-612, 2017.

[24] C. H. Chen, L. H. Lee, Stochastic Simulation Optimization: Stochastic Simulation Optimization: An Optimal Computing Budget allocation. World Scientific Publishing, 2010.

[25] B. Liu, F. V. Fernandez, G. G. E. Gielen, "Efficient and accurate statistical analog yield optimization and variation-aware circuit sizing based on computational intelligence techniques," IEEE Trans. Comput.-aided Des. Integr. Circuits Syst., 30(6): 793-805, 2011.

[26] R. Assaad, J. Silva-Martinez, "Enhancing general performance of folded cascode amplifier by recycling current," Electron. Lett., 43(23): 1243, 2007.

[27] S. M. Anisheh, H. Shamsi, "Placement and routing method for analogue layout generation using modified cuckoo optimization algorithm," IET Circuits Devices Syst., 12(5): 532-541, 2018.

[28] P. Gray, R. G. Meyer, P. J. Hurst, S. Lewis, Analysis and design of analog integrated circuits, 6th ed. Brisbane, QLD, Australia: John Wiley and Sons (WIE), 2024.

[29] E. R. Ziegel, W. Winston, "Simulation modeling using @risk," Technometrics, 39(3): 345, 1997.

[30] X. Wang et al., "Analog circuit yield optimization via freeze–thaw Bayesian optimization technique," IEEE Trans. Comput.-aided Des. Integr. Circuits Syst., 41(11): 4887-4900, 2022.

Biographies

Abbas Yaseri was born in 1979 and received the B.Sc. and M.Sc. degrees in Electrical Engineering from University of Mazandaran. He is a specializing in electronics, artificial intelligence and computer vision. He has joined the Hadaf Higher Education since 2009.

- Email: abbas.yaseri@gmail.com
- ORCID: 0000-0001-9136-4122
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Mohammad Hossein Maghami was born in Mashhad, Iran, in 1984. He received the B.Sc. degree from Ferdowsi University of Mashhad, Mashhad, Iran, in 2006, the M.Sc. degree from Amirkabir University of Technology, Tehran, Iran, in 2009, and the Ph.D. degree from K. N. Toosi University of Technology, Tehran, Iran, in 2015, all in Electrical Engineering. He carried out part of his Ph.D. research work at Polytechnique Montreal as a visiting research scholar. Since September 2016 he is with Shahid Rajaee Teacher Training University, Tehran, Iran, as an Assistant Professor. His main areas of interests are implantable biomedical microsystems, high-speed low-power A/D converters, and mixed-mode integrated circuits.

- Email: mhmaghami@sru.ac.ir
- ORCID: 0000-0002-7932-9161
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://www.sru.ac.ir/en/faculty/school-of-electrical-engineering/mohammad-hossein-maghami/

Mehdi Radmehr was born in 1974 and received the B.Sc., M.Sc., and PhD degrees in Electrical Engineering from University of Tehran, Tarbiat Modares, and Islamic Azad University, Science and Research campus, Tehran, Iran, in 1996, 1998, and 2006 respectively. He is a specializing in power electronics, motor drives and power quality. He has worked for Mazandaran Wood and Paper Industries as an advisor since 1997 before starting his Ph.D. study. He has joined the scientific staff of Islamic Azad University, Sari branch since 1998.

- Email: maradmehr@gmail.com
- ORCID: 0000-0003-1678-9758
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
A. Yaseri, M. H. Maghami, M. Radmehr, "A fast and accurate yield optimization method for designing operational amplifier using multi-objective evolutionary algorithm based on decomposition," J. Electr. Comput. Eng. Innovations, 13(1): 43-56, 2025.

DOI: 10.22061/jecei.2024.10814.737

URL: https://jecei.sru.ac.ir/article_2195.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Noise Folding Compensation in Compressed Sensing based Matched-Filter Receiver

M. Kalantari *

Faculty of Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.

Article Info

Article History:
Received 18 May 2024
Reviewed 15 July 2024
Revised 14 August 2024
Accepted 31 August 2024

Keywords:
Compressed sensing
Matched-filter
Noise folding

*Corresponding Author's Email Address: *mkalantari@sru.ac.ir*

Abstract

Background and Objectives: Compressed sensing (CS) of analog signals in shift-invariant spaces can be used to reduce the complexity of the matched-filter (MF) receiver, in which we can be approached the standard MF performance with fewer filters. But, with a small number of filters the performance degrades quite rapidly as a function of SNR. In fact, the CS matrix aliases all the noise components, therefore the noise increases in the compressed measurements. This effect is referred to as noise folding. In this paper, an approach for compensating the noise folding effect is proposed.

Methods: An approach for compensating of this effect is to use a sufficient number of filters. In this paper the aim is to reach the better performance with the same number of filter as in the previous work. This, can be approached using a weighting function embedded in the analog signal compressed sensing structure. In fact, using this weighting function we can remedy the effect of CS matrix on the noise variance.

Results: Comparing with the approach based on using the sufficient number of filters to counterbalance the noise increase, experimental results show that with the same numbers of filters, in terms of probability of correct detection, the proposed approach remarkably outperforms the rival's.

Conclusion: Noise folding formation is the main factor in CS-based matched-filter receiver. The method previously presented to reduce this effect demanded using the sufficient number of filters which comes at a cost. In this paper we propose a new method based on using the weighting function embedded in the analog signal compressed sensing structure to achieve better performance.

Introduction

In many emerging and important applications, the Nyquist sampling rate, a rate equals to twice the highest frequency, is so high that we encounter with far too many samples that must be processed and stored in high-capacity memory. Meanwhile, in applications in which inputs are wideband signals, it is so costly and mostly physically impossible to build analog to digital convertor capable of acquiring samples at Nyquist rate [1], [2]. So, in the past few years the vast interest in the area of compressed sensing (CS) caused the sampling theory has again been revived [3]-[5]. Compressed sensing is a framework for sensing and compression of finite-dimensional vectors simultaneously [3], [4], [6]. The main idea in CS is that, instead of sampling at a high rate and then compressing the samples, we want to have a way to measure the data in a compressed form directly [7]. For this, the finite-dimensional signal must have a sparse or compressible representation in a known basis [3]. In this way, we capture only the essential information embedded in a signal. Formerly, the CS was a mathematical theory for measuring finite-dimensional

vector. In fact, CS was a framework for sampling of discrete-time signals and reconstruction from a finite number of samples. Despite the widespread literature in the area of extending the ideas of CS to analog domain, it remains a difficult challenge [8]-[11]. Eldar extended the CS to consider sub-Nyquist sampling of continuous-time signals in shift-invariant spaces via combining ideas from compressed sensing with analog sampling results [7]. A shift-invariant (SI) subspace is a space in which signals can be represented as a linear combination of shifts of a set of generators [12]-[16]. The subspace of bandlimited signals, multiband signals, the spline functions, the communication transmission such as PAM (pulse amplitude modulation) and QAM (quadratic amplitude modulation) are some important examples of SI subspace [17]-[24]. So, the compressed sensing of analog signals in SI spaces leads to low-rate (sub-Nyquist) sampling of a broad set of analog signals. This sub-Nyquist samples can be processed directly without having to upsample them back to the Nyquist rate, leading to low-rate processing as well.

The idea of analog CS can be used to standard detection problem, concerned in communication systems, for reducing the receiver complexity [7], [25], [26]. In fact, analog CS enables us to convey more information over the channel with the same receiver. It is a well-known result that the MF receiver which consists the same number of filters (correlators) as the number of transmitted signals, say $\boldsymbol{N}$, maximize the probability of correct detection. Nevertheless, it can be shown that, using the idea of analog CS and in a noise-free environment, regardless of the number of signals for transmission, only two filters is required to detect the transmitted signal exactly [7]. In the presence of noise, in order to achieve good performance, the number of correlators must be increased. In fact, when noise is present, for strictly maximizing the probability of correct detection we require $\boldsymbol{N}$ filters. However we can get very good performance with fewer correlators. But with a smaller number of filters the performance degrades quite rapidly as a function of SNR. In fact, the CS matrix aliases all the noise components, therefore the noise increases in the compressed measurement. This effect is referred to as noise folding [27]. An approach for compensating this effect is to use sufficient number of filters. It is shown that approximately $\boldsymbol{\log N}$ filters are needed to countervail this increase in noise [28].

In this paper the aim is to reach the better performance with the same number of filter as in the previous work in [28], [7]. This, can be approached using a weighting function embedded in the analog signal compressed sensing structure. In fact, using this weighting function we can remedy the effect of CS matrix on the noise variance.

This paper is organized as follows. The second section reviews the fundamentals of analog signal compressed sensing. The third section presents the proposed method for amending the noise folding effect. Comparisons with the rival method are presented in the fourth section, and the end section concludes the paper.

Compressed Sensing of Analog Signals and CS based Matched-filter Receiver

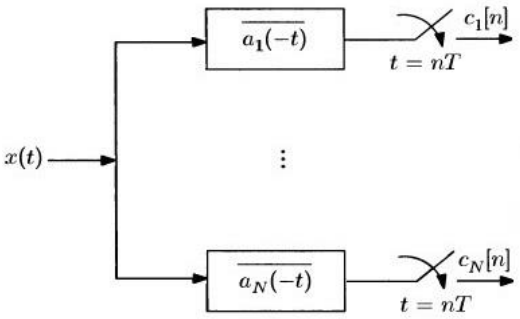
This section shortly explains the theory of analog signal compressed sensing [7], [29]. The formulation provided in this section will be utilized in the third section to develop the proposed method.

A. Sampling and Reconstruction in SI Spaces

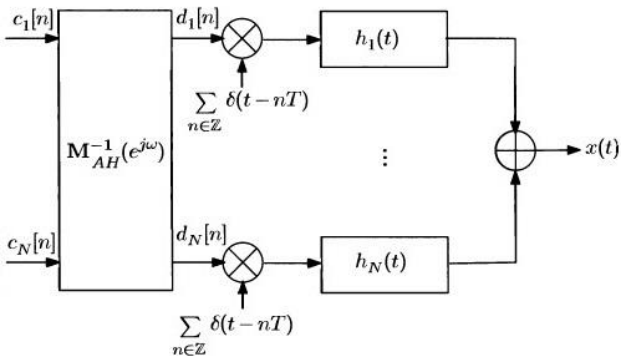
A mentioned in the previous section, SI signals are specified by a set of generators $\{h_\ell(t), 1 \le \ell \le N\}$, where N may be finite or infinite. So, any signal in SI space can be written as

$$x(t) = \sum_{\ell=1}^{N} \sum_{n\in\mathbb{Z}} d_\ell[n] h_\ell(t - nT) \qquad (1)$$

where $d_\ell[n] \in \ell_2, 1 \le \ell \le N$, ℓ_2 is the space of discrete-time finite-energy signals and T is the period. It is well known that such signals can be recovered from sample at a rate of N/T. Sampling and reconstruction of signals in SI space have been depicted in Fig. 1.



(a)



(b)

Fig. 1: (a) Sampling, (b) Reconstruction in shift- invariant spaces [7].

In this sampling method, $x(t)$ is passed through a bank of N filters, each with almost arbitrary impulse response

of $\overline{a_\ell(-t)}$. Then, the outputs are sampled with period T unformly, resulting the sample sequnces $c_\ell[n]$. The vector containing the DTFTs of $c_\ell[n], 1 \le \ell \le N$, denoted by $\mathbf{c}(e^{j\omega})$, and the vector collecting the DTFTs of $d_\ell[n], 1 \le \ell \le N$, denoted by $\mathbf{d}(e^{j\omega})$.
It can be shown that

$$\mathbf{d}(e^{j\omega}) = \mathbf{M}_{AH}^{-1}(e^{j\omega})\mathbf{c}(e^{j\omega}) \tag{2}$$

where $\mathbf{M}_{AH}(e^{j\omega})$ is an $N \times N$ matrix, with entries

$$[\mathbf{M}_{AH}(e^{j\omega})]_{i\ell} = \frac{1}{T}\sum_{k\in\mathbb{Z}} \overline{A_\iota\left(\frac{\omega}{T} - \frac{2\pi k}{T}\right)} H_\ell\left(\frac{\omega}{T} - \frac{2\pi k}{T}\right) \tag{3}$$

$A_i(\omega)$ and $H_\ell(\omega)$ are the CTFTs of $a_i(t)$ and $h_\ell(t)$ respectively, and $\mathbf{M}_{AH}^{-1}(e^{j\omega})$ is the inverse of $\mathbf{M}_{AH}(e^{j\omega})$. The reconstruction of $x(t)$ is accomplished via modulating each output sequence $d_\ell[n]$ by a periodic impulse train $\sum_{n\in\mathbb{Z}}\delta(t-nT)$ with period T, followed by filtering with analog filter $h_\ell(t)$. Similar to finite interpolation in the Shannon-Nyquist theorem, if $h_\ell(t)$ decay fast enough, interpolation with finitely many samples leads to sufficiently accurate reconstruction.

For signals that can be represented by k generator, $k < N$, chosen from a finite set of N functions, we have

$$x(t) = \sum_{|\ell|=k}\sum_{n\in\mathbb{Z}} d_\ell[n]h_\ell(t-nT) \tag{4}$$

If we know the k active generators then we can uniformly sample the output of k appropriate filters with sampling period of T as in Fig. 1, resulting a sampling rate of k/T. On the other hand, if we know that only k out of N generators are active but don't know in advance which one, then the minimal sampling rate is at least $2k/T$. Thus the lack of knowledge about the exact subspace to which $x(t)$ belongs, leads to an increase of at least a factor 2 in the minimal sampling rate [7]. By combining ideas from analog sampling and CS, this minimal rate has been achieved [29].

B. Compressed Sensing in Sparse Unions

Suppose that $\mathbf{d}[n]$ is a vector whose ℓth component is given by $d_\ell[n]$, in which only k out of the N sequences $d_\ell[n]$ are nonzero. Compressively measuring the vector sequence $\mathbf{d}[n]$ can be accomplished by a $p \times N$ sensing matrix $\mathbf{A}$, $p < N$, that allows recovery of k-sparse vectors. The choice $p < N$, reduces the sampling rate below the Nyquist rate. In fact, a compressive sampling system produces a vector of low-rate samples $\mathbf{y}[n] = \left[y_1[n], \dots, y_p[n]\right]^T$ satisfying the relation

$$\mathbf{y}[n] = \mathbf{A}\mathbf{d}[n], \quad \|\mathbf{d}[n]\|_0 \le k \tag{5}$$

where $\|\mathbf{d}[n]\|_0$ is the number of nonzero elements of $\mathbf{d}[n]$. For each n,

$$\mathbf{y}[n] = \mathbf{A}\mathbf{d}[n], \quad n \in \mathbb{Z} \tag{6}$$

Equation (6) is an infinite measurement vector (IMV) problem and for each n, $\mathbf{d}[n]$ is k-spares [7], [29]. Meanwhile, the infinite set of vectors $\{\mathbf{d}[n], n \in \mathbb{Z}\}$ shares a joint sparsity pattern: at most k of the sequences $d_\ell[n]$ are nonzero. These set of equations can be transformed into an equivalent multiple measurement vector (MMV) problem using the continuous to finite (CTF) block technique [7]. The perfect recovery of $\mathbf{d}[n]$ (or recovery with high probability) is guaranteed because $\mathbf{A}$ was designed to enable CS techniques.

The frequency-domain counterpart of (6) is as follows,

$$\mathbf{y}(e^{j\omega}) = \mathbf{A}\mathbf{d}(e^{j\omega}), \quad 0 \le \omega \le 2\pi \tag{7}$$

where $\mathbf{y}(e^{j\omega})$, $\mathbf{d}(e^{j\omega})$ are the vectors containing the DTFTs $Y_\ell(e^{j\omega}), D_\ell(e^{j\omega})$ respectively. $\mathbf{d}[n]$ may also be recovered from

$$\mathbf{y}(e^{j\omega}) = \mathbf{W}(e^{j\omega})\mathbf{A}\mathbf{d}(e^{j\omega}), \quad 0 \le \omega \le 2\pi \tag{8}$$

where $\mathbf{W}(e^{j\omega})$ is any invertible $p \times p$ matrix with elements $W_{i\ell}(e^{j\omega})$. In this way we can generalize the class of sensing operators. This extra freedom can be used in the proposed method for noise folding compensation as we will see in the following sections. The analog compressed sampling can be seen in Fig. 2.

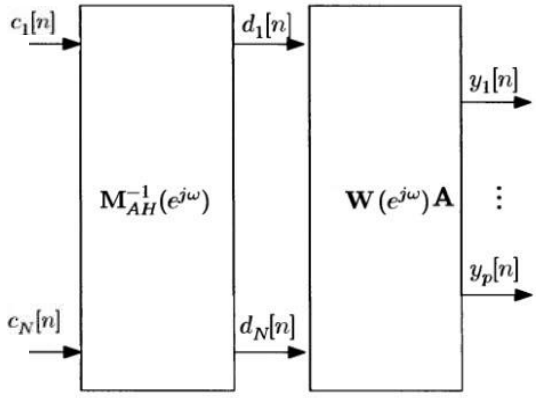


Fig. 2: Analog compressed sampling [7].

The inputs in Fig. 2 come from Fig. 1(a). Though the sampling method of Fig. 2 leads to compressed measurements $\{y_\ell[n]\}$, the sampling rate is still N/T. Eldar reduced this rate to p/T where $2k \le p < N$ via proving the following theorem [29],

Theorem 1. $\{y_\ell[n]\}, 1 \le \ell \le p$, in Fig. 2 can be obtained by filtering $x(t)$ in (4) with p filters $\overline{\{s_\ell(-t)\}}$ and sampling the outputs at rate $1/T$, where

$$\mathbf{s}(\omega) = \overline{\mathbf{W}(e^{j\omega T})}\mathbf{A}\mathbf{v}(\omega) = \overline{\mathbf{W}(e^{j\omega T})\mathbf{A}\mathbf{M}_{AH}^{-1}(e^{j\omega T})}\mathbf{a}(\omega) \tag{9}$$

where $\mathbf{s}(\omega), \mathbf{a}(\omega)$ are the vectors with ℓth elements $S_\ell(\omega), A_\ell(\omega)$ respectively, and $V_\ell(\omega)$, the components of $\mathbf{v}(\omega)$, are Fourier transform of generators $v_\ell(t)$ such that

$\{v_\ell(t-nT)\}$ are biorthogonal to $\{h_\ell(t-nT)\}$. In the time domain we have,

$$s_i(t) = \sum_{\ell=1}^{N}\sum_{r=1}^{p}\sum_{n\in\mathbb{Z}} \overline{w_{ir}[-n]\mathbf{A}_{r\ell}}\, v_\ell(t-nT) \tag{10}$$

in which $w_{ir}[n]$ is the inverse DTFT of $W_{ir}(e^{j\omega})$, the elements of matrix $\mathbf{W}(e^{j\omega})$, and

$$v_i(t) = \sum_{\ell=1}^{N}\sum_{n\in\mathbb{Z}} \overline{\varphi_{i\ell}[-n]}\, a_\ell(t-nT) \tag{11}$$

here $\varphi_{i\ell}[n]$ is the inverse DTFT of $[\mathbf{M}_{AH}^{-1}(e^{j\omega T})]_{i\ell}$. When $\mathbf{W}(e^{j\omega}) = \mathbf{I}$,

$$s_i(t) = \sum_{\ell=1}^{N} \overline{\mathbf{A}_{i\ell}}\, v_\ell(t). \tag{12}$$

Fig. 3 shows the compressed sensing of analog signals with sampling rate of p/T.

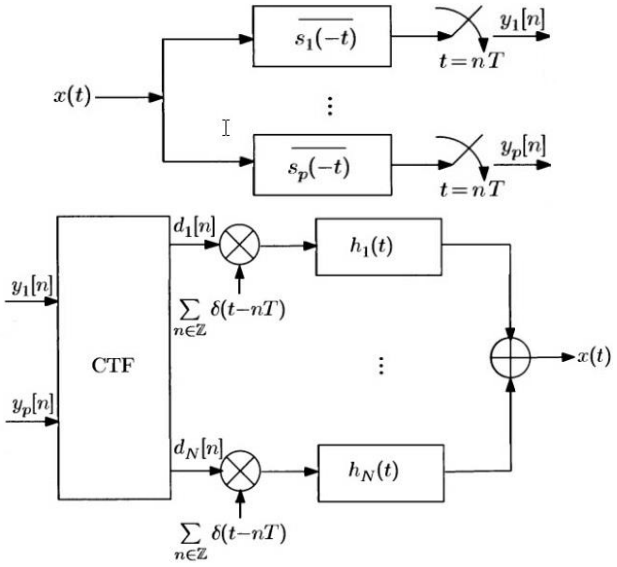


Fig. 3: Compressed sensing of analog signal [7].

C. Compressed-Sensing Based Matched- Filter Receiver

Suppose that a basic communication system transmits digital data to a receiver by sending one of a set of N linearly independent known signals $\{h_i(t), 1 \le i \le N\}$ over a symbol duration of T. The channel add a zero-mean white Gaussian noise $n(t)$ with variance σ^2 to the signal, so the received signal is as

$$y(t) = h_\ell(t) + n(t) \tag{13}$$

for some index ℓ. The goal is to determine the index ℓ in order to decode the transmitted symbol. Demodulator for a typical matched-filter receiver is shown in Fig. 4 in which $\overline{s_\ell(-t)} = \overline{h_\ell(-t)}$.

It is well known that a MF receiver is a sufficient statistic for detection; that is, the optimal detector can be computed based on the MF output providing the noise is Gaussian. The maximum-likelihood detector for MF receiver is as

$$\ell = \arg\max_i \mathcal{R}\{y_i\} \tag{14}$$

where $\mathcal{R}\{y_i\}$ is the real part of y_i.

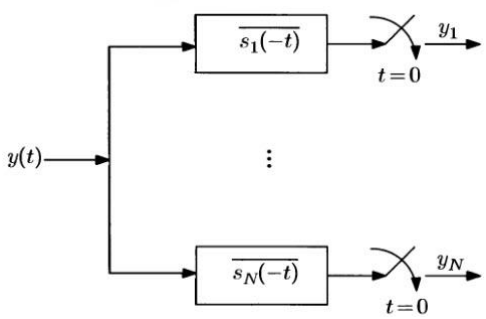


Fig. 4: Demodulator for a matched-filter receiver [7].

By exploiting the ideas of analog CS, we can reduce the number of filters in demodulator part of the MF receiver. Reformulation of the detection problem as a CS recovery problem is as follows. Any signal $h_\ell(t)$ can be written in the form $h_\ell(t) = H\mathbf{x}$, where $H: \mathbb{R}^N \to L_2$ is the set transformation corresponding to $\{h_i(t)\}$ and $\mathbf{x}$ is a vector containing a 1 in the ℓth position. Thus, in term of the basis defined by the transformation H, the transmitted signal is sparse. This scenario is a special case of (4) in which $k = 1$, where we consider only one symbol interval. So, we can recover $\mathbf{x}$ using $p < N$ filters chosen according to Theorem 1. For that, let $\mathbf{A}$ be an arbitrary $p \times N$ CS matrix for a 1-spares vector. Then, according to Theorem 1 the demodulator consists of filters $\{\overline{s_\ell(-t)}, 1 \le \ell \le p\}$,

$$s_\ell(t) = \sum_{m=1}^{N} \overline{\mathbf{A}_{\ell m}}\, v_m(t) \tag{15}$$

where $\{v_m(t)\}$ are the biorthogonal functions defined as

$$v_m(t) = \sum_{i=1}^{N} \emptyset_{mi} h_i(t) \tag{16}$$

with $\mathbf{\Phi} = (H^*H)^{-1}$. In operator notation, $S = V\mathbf{A}^* = H(H^*H)^{-1}\mathbf{A}^*$ and $V^*H = I$ where S, V are set transformation corresponding to $\{s_i(t)\}$ and $\{v_i(t)\}$ respectively.

Suppose a noise-free case in which $y(t) = h_\ell(t) = H\mathbf{x}$ for some index ℓ. After applying the p filters on $y(t)$, the output vector is as follows

$$\mathbf{c} = S^*y(t) = \mathbf{A}(H^*H)^{-1}H^*y(t) = \mathbf{A}\mathbf{x} \tag{17}$$

So, the problem reduces to recovery of a 1-sparse vector $\mathbf{x}$ from compressed measurement $\mathbf{c}$. From the theory of uniqueness sparse recovery, we know that for recovery of a k-sparse vector the spark of the sensing matrix must be greater than $2k$. In particular, for

uniqueness we must have that $p \geq 2k$. So, $\mathbf{A}$ may be a $2 \times N$ matrix in which no two columns are multiple of each other. The support of $\mathbf{x}$ can be recovered by choosing

$$\ell = \arg\max_i \mathcal{R}\{\langle \boldsymbol{a}_i, \mathbf{c}\rangle\} \tag{18}$$

where $\boldsymbol{a}_i$ is the ith column of $\mathbf{A}$. So, only two correlators are required to detect the transmitted signal exactly. The overall receiver is depicted in Fig. 5.

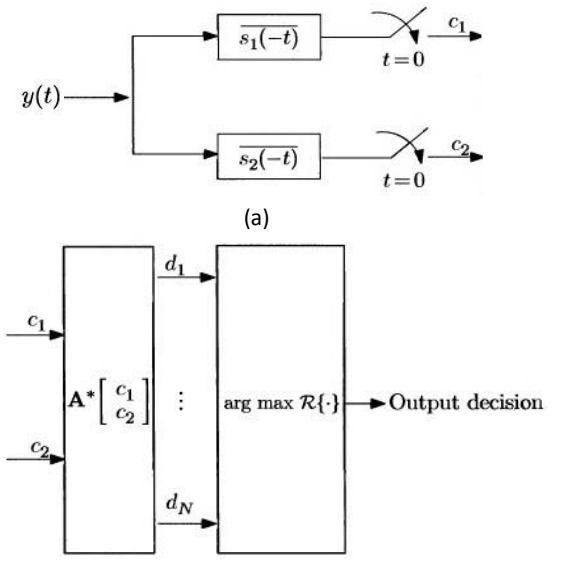


Fig. 5: A noise-free detector based on analog compressed sensing [7].

But, in noisy environment the number of correlator must be increased in order to achieve good performance. In fact we can get very good performance with $2 < p < N$ correlators. It can be shown that the selection rule is identical to the one in the noise-free case as expressed in (14).

The Proposed Method

In noisy environment, strictly maximizing the probability of correct detection will require N correlators, but as mentioned before, by using $p < N$ filters, we can get pretty good performance. With a small number of filters the performance degrades quite rapidly as a function of SNR. In fact, the CS matrix aliases all the noise components, therefore the noise increases in the compressed measurements. This effect is referred to as noise folding which is the main problem in this type of receiver. The Noise folding compensation can be accomplished by employing a sufficient number of filters. It is shown that approximately $\log N$ filters are needed to countervail this increase in noise [28]. In this paper we propose a new method based on using the weighting function embedded in the analog signal compressed sensing structure, $\mathbf{W}(e^{j\omega})$, to achieve better performance. According to Theorem 1 and in operator notation we have

$$S = V\mathbf{D}^* \tag{19}$$

where S, V are set transformation corresponding to $\{s_i(t)\}$ and $\{v_i(t)\}$ respectively and $\mathbf{D}$ is a $p \times N$ matrix with elements

$$d_{i\ell} = \mathrm{IDFT}\left(B_{i\ell}(e^{j\omega})\right) \tag{20}$$

in which $B_{i\ell}(e^{j\omega})$ are the elements of the matrix

$$\mathbf{B}(e^{j\omega}) = \overline{\mathbf{W}(e^{j\omega})\mathbf{A}} \tag{21}$$

From (9), we have that

$$\mathbf{s}(\omega) = \mathbf{B}(e^{j\omega T})\mathbf{v}(\omega) \tag{22}$$

In the time domain and in terms of $d_{i\ell}$ we can write

$$s_i(t) = \sum_{\ell=1}^{N}\sum_{n\in\mathbb{Z}} d_{i\ell}[n]\, v_\ell(t - nT) \tag{23}$$

So, we have $S = V\mathbf{D}^*$. Using this and the fact that $V^*H = I$ we have

$$\begin{aligned}\mathbf{c} = S^*y(t) &= \mathbf{D}V^*h_\ell(t) + \mathbf{D}V^*n(t) = \\ &\mathbf{D}V^*H\mathbf{x} + \mathbf{D}V^*n(t) = \mathbf{D}\mathbf{x} + \mathbf{w}\end{aligned} \tag{24}$$

where $\mathbf{w} = \mathbf{D}V^*n(t)$ is the noise component. We can write

$$\begin{aligned}&\mathrm{E}\big[\langle v_j(t), n(t)\rangle\langle v_i(t), n(t)\rangle\big] = \\ &\iint v_j(t)v_i(\tau)E[n(t)n(\tau)] = \sigma^2\langle v_j(t), v_i(t)\rangle\end{aligned} \tag{25}$$

Therefore,

$$\begin{aligned}\mathbf{R}_w = E[\mathbf{w}\mathbf{w}^*] &= \sigma^2\mathbf{D}V^*V\mathbf{D}^* \\ &= \sigma^2\mathbf{D}(H^*H)^{-1}\mathbf{D}^*\end{aligned} \tag{26}$$

In general, the noise is not white, but if the signals $h_\ell(t)$ are orthonormal and the rows of $\mathbf{D}$ are orthogonal then $\mathbf{R}_w = \kappa\sigma^2\mathbf{I}$ where κ is the squared-norm of rows of $\mathbf{D}$.

The problem is reduced to recovery of 1-spares vector $\mathbf{x}$ from noisy measurements $= \mathbf{D}\mathbf{x} + \mathbf{w}$. The standard CS algorithms can be used for solving this problem provided the $B_{i\ell}(e^{j\omega})$ are constant functions or equivalently, $d_{i\ell} = d_{i\ell}[n]\delta[n]$. Otherwise, we should develop a MAP detector [26], [30]. Assume that the $P(\ell|\mathbf{c})$ be the probability that x_ℓ, the ℓ^{th} elements of $\mathbf{x}$, is nonzero, given $\mathbf{c}$. The goal is to choose the ℓ that maximize this probability. Using the Bayes rule and the fact that $P(\ell) = 1/N$, the problem will be the maximizing $P(\mathbf{c}|\ell)$. The vector $\mathbf{c}$ is a Gaussian vector with mean $\mathbf{D}^*\mathbf{x}$ and covariance $\mathbf{R}_w = \sigma^2\mathbf{D}(H^*H)^{-1}\mathbf{D}^*$. We have

$$\begin{aligned}&\ln P(\mathbf{c}|\ell) = \\ &-\Upsilon(\mathbf{c} - \mathbf{d}_{\boldsymbol{\ell}})^*(\mathbf{D}(H^*H)^{-1}\mathbf{D}^*)^{-1}(\mathbf{c} - \mathbf{d}_{\boldsymbol{\ell}})\end{aligned} \tag{27}$$

where Υ is a constant and $\mathbf{d}_{\boldsymbol{\ell}}$ is the ℓ^{th} column of $\mathbf{D}$. Maximizing $\ln P(\mathbf{c}|\ell)$ is equivalent to minimizing the

following function

$$\Lambda(\ell) = (\mathbf{c} - \mathbf{d}_{\ell})^{*}(\mathbf{D}(H^{*}H)^{-1}\mathbf{D}^{*})^{-1}(\mathbf{c} - \mathbf{d}_{\ell}) \quad (28)$$

In the special case in which $\mathbf{R}_w = \kappa\sigma^2\mathbf{I}$, the minimization of $\Lambda(\ell)$ leads to the following selection rule

$$\ell = \arg\max_{i} \mathcal{R}\{\langle \boldsymbol{d}_i, \mathbf{c}\rangle\} \quad (29)$$

which is similar to the noise free case.

The covariance of the noise in the proposed method is similar to the one in the rival method [7], [28]. In fact, the covariance in the rival method is as

$$\mathbf{R}_w = E[\mathbf{w}\mathbf{w}^{*}] = \sigma^2\mathbf{A}V^{*}V\mathbf{A}^{*} = \sigma^2\mathbf{A}(H^{*}H)^{-1}\mathbf{A}^{*} \quad (30)$$

So, the only change is the substitution of matrix $\mathbf{A}$ in the previous method with the matrix $\mathbf{D}$. But, note that this substitution has a great impact on the variance of the noise in measurements $\mathbf{c}$. Unlike the matrix $\mathbf{A}$, the elements of matrix $\mathbf{D}$ are sequences, not scalars.

According to (23) the weighting function $\mathbf{W}(e^{j\omega})$ introduces extra freedom when designing the corresponding analog sampling filters. Meanwhile, according to (20), (21), (26), $\mathbf{W}(e^{j\omega})$ has great impact on the variance of the noise. We show this impact with a simple example.

Suppose that the signals $\{h_i(t)\}$ are orthonormal, so we have $H^{*}H = I$ and from (30), $\mathbf{R}_w = \sigma^2\mathbf{A}\mathbf{A}^{*}$. Also, suppose that $\mathbf{A}$ is chosen as random rows of a Fourier matrix. This means that $A_{i\ell} = \left(1/\sqrt{p}\right)\exp\{-j2\pi s_i\ell/N\}$ where s_i is the ith row chosen. In this case, $\mathbf{A}\mathbf{A}^{*} = (N/p)\mathbf{I}$. So, the noise variance is increased by a factor of N/p. In this example we see simply that choosing $\mathbf{W}(e^{j\omega}) = (p/N)\mathbf{I}$, results in $\mathbf{R}_w = \sigma^2$ according to (26).

But, the cases are not as simple as the previous example, and choosing the appropriate $\mathbf{W}(e^{j\omega})$ is a difficult problem. But, as we can see in the next section, we can simply bypass this problem.

The Experimental Results

In this section we will examine the ability of the proposed method in compensation of the noise folding problem. For that, we demonstrate the effect of additive noise on the proposed method and the rival method as expressed by (18) [28], [7]. There are no comparable works since 2015.

We consider a receiver with $N = 100$ different transmitted signals given by

$$h_{\ell}(t) = \begin{cases} 1, & (\ell - 1) \le t \le \ell \\ 0, & otherwise \end{cases} \qquad \ell = 1,2,\dots,N \quad (31)$$

and p correlators. For the rival method, the sensing matrix $\mathbf{A}$ is chosen to be equal to p random rows of the $N \times N$ Fourier matrix in which the columns normalized to have unit norm. For the proposed method we need a matrix

$$\mathbf{D} = \text{IDFT}\left(\overline{\mathbf{W}(e^{j\omega})\mathbf{A}}\right) \quad (32)$$

with orthogonal rows. We do not involve ourselves with the problem of selecting appropriate $\mathbf{W}(e^{j\omega})$ and $\mathbf{A}$ matrix. Instead, we simply generate synthetically a $p \times N$ matrix as the matrix $\mathbf{D}$. As mentioned before, the main difference between matrix $\mathbf{A}$ and matrix $\mathbf{D}$ is that in the former the elements of the matrix are scalars and in the latter are sequences.

We saw that when noise is present, for strictly maximizing the probability of correct detection we require $N = 100$ coorelators. Although with fewer filters we can get very good performance, the performance degrades quite rapidly as a function of SNR. This is due to noise folding phenomenon. In fact, the use of the CS matrix reduces the SNR, i.e. the CS matrix aliases all the noise component in the compressed measurement, even those corresponding to zero element in sparse vector $\mathbf{x}$, leading to a noise increase in the compressed measurements.

Fig. 6 to Fig.12 show the probability of correct detection in both methods as a function of the number of correlators for different value of the SNR. The probability is estimated by running 1000 Monte Carlo simulations. In each iteration, the transmitted signal and the noise are chosen randomly. As we can see, as long as the SNR is high enough, perfect detection is achieved in both methods using a much smaller filters compared to the MF receiver consists of 100 correlators and the proposed method needs smaller filters compared to the rival method to achieve the same performance. This in turn shows that the proposed method can remedy the noise folding problem with fewer filters than the rival method.

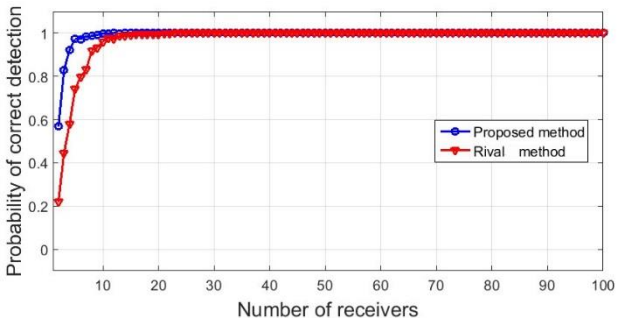


Fig. 6: Probability of correct detection as a function of the number of correlators for SNR=40.

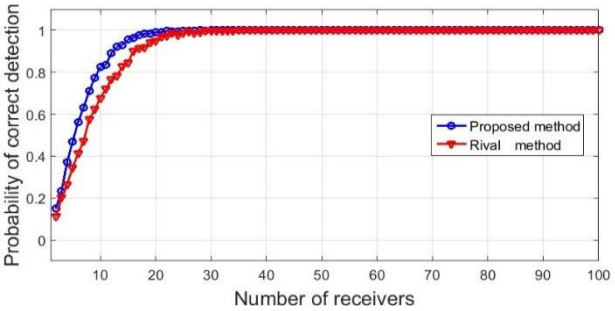


Fig. 7: Probability of correct detection as a function of the number of correlators for SNR=30.

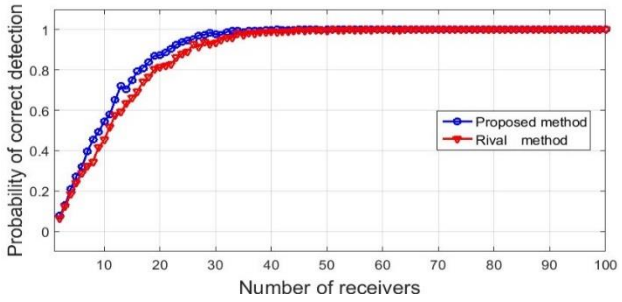


Fig. 8: Probability of correct detection as a function of the number of correlators for SNR=25.

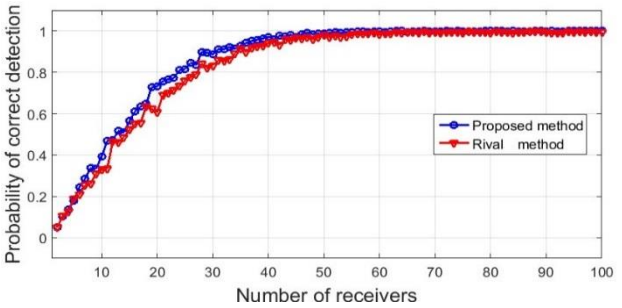


Fig. 9: Probability of correct detection as a function of the number of correlators for SNR=20.

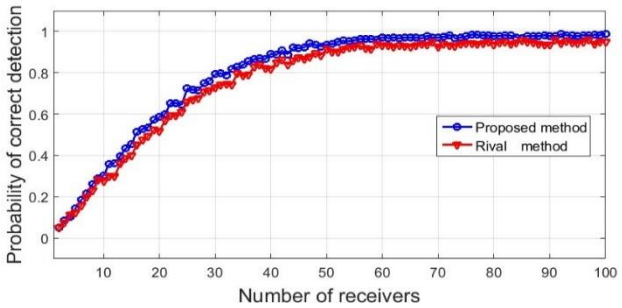


Fig. 10: Probability of correct detection as a function of the number of correlators for SNR=15.

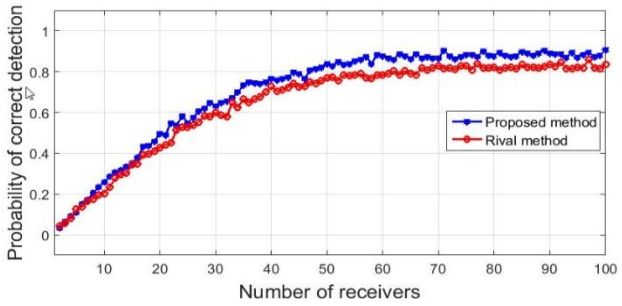


Fig. 11: Probability of correct detection as a function of the number of correlators for SNR=10.

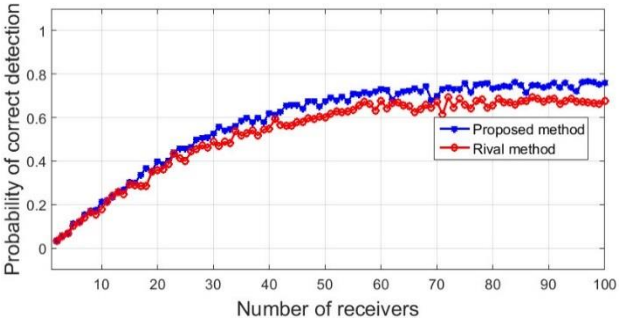


Fig. 12: Probability of correct detection as a function of the number of correlators for SNR=5.

Results and Discussion

The simulation results show that the proposed method has the ability to compensate the noise folding problem more effectively than the rival method. i.e. with fewer correlators. This is due to the fact that the elements of matrix $\mathbf{D}$ are sequences rather than scalars.

Conclusion

Noise folding is the main problem in compressed sensing based MF receiver. An approach for compensating this effect is to use sufficient number of coorelators. The proposed method achieves better performance with the same number of filters as in the previous work. This goal is achieved through the use of weighting function embedded in the analog signal compressed sensing structure. This weighting function can remedy the effect of CS matrix on the noise variance.

As stated in the previous sections, choosing the appropriate weighting functions is a difficult problem. In this paper we bypass this problem via generating synthetically a $p \times N$ matix as the matrix $\mathbf{D}$ with orthogonal rows. In this way, there is no notable difference between the proposed method and the rival's method from the point of time and space complexity. Systematically computing of matrix $\mathbf{D}$ can be suggested for future work in which the time and space complexity is a major concern because of existence of long sequences as elements of the matrix $\mathbf{D}$.

Author Contributions

M. Kalantari has written the whole paper without participation of anybody. All parts of this work have been accomplished by the author as the single author and the corresponding author of the paper.

Acknowledgments

This work was supported by Shahid Rajaee Teacher Training University under contract number 11980.

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this manuscript.

Abbreviations

CS	Compressed sensing
MF	Matched filter
SNR	Signal to noise ratio
PAM	Pulse amplitude modulation
QAM	Quadratic amplitude modulation
SI	Shift invariant
DTFT	Discrete time Fourier transform
MMV	Multiple Measurement vector
CTF	Continuous to finite block
$\mathbf{A}$	Sensing matrix
$\mathcal{R}\{.\}$	Real part of argument
$\mathbf{W}(e^{j\omega})$	Weighting matrix

$n(t)$	Noise signal
$\mathbf{R}_w$	Noise covariance matrix
MAP	Maximum a posteriori
$\mathbf{c}$	Noisy measurement vector
$\mathbf{a}_i$	ith column of $\mathbf{A}$
$x(t)$	A SI signal
$\overline{x(t)}$	Complex conjugate of $x(t)$
$h_\ell(t)$	SI generator or a known transmitted signal
$y(t)$	Received signal
$\mathbf{x}$	A sparse vector
IDFT	Inverse discrete Fourier transform

References

[1] R. Walden, "Analog-to-digital converter survey and analysis," IEEE J. Selected Areas Comm., 17(4): 539-550, 1999.

[2] D. Healy, "Analog-to-Information: Baa #05-35," 2005.

[3] D. L. Donoho, "Compressed Sensing," IEEE Trans. Inform. Theory, 52(4): 1289-1306, 2006.

[4] E. Candes, J. Romberg, T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," IEEE Trans. Inform. Theory, 52(2): 489-509, 2006.

[5] Y. C. Eldar, G. Kutynikov, Compressed Sensing: Theory and Applications, Cambridge, UK, Cambridge University Press, 2012.

[6] E. Candes, J. Romberg, T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," Comm. Pure Appl. Math., 59(8): 1207-1223, 2006.

[7] Y. C. Eldar, Sampling Theory, Beyond Bandlimited Systems, Cambridge, UK, Cambridge University Press, 2015.

[8] J. A. Tropp, M. B. Wakin, M. F. Duarte, D. Baron, R. G. Baraniuk, "Random filters for compressive sampling and reconstruction," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), 3, 2006.

[9] J. N. Laska, S. Kirolos, M. F. Durate, T. S. Ragheb, R. G. Baraniuk, Y. Masoud, "Theory and implementation of an analog-to-information converter using random demodulation," in Proc. IEEE Int. Symp. Circuits Systems (ISCAS): 1959-1962, 2007.

[10] M. Vetterli, P. Marziliano, T. Blu, "Sampling signals with finite rate of innovation," IEEE Trans. Signal Process., 50(6): 1417-1428, 2002.

[11] P. L. Dragotti, M. Vetterli, T Blu, "Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets strang fix," IEEE Trans. Signal Process., 55(5): 1741-1757, 2007.

[12] C. de Boor, R. DeVore, A. Ron, "The structure of finitely generated shift-invariant spaces in $\mathrm{L}_2(\mathrm{R}^\mathrm{d})$," J. Funct. Anal., 119(1): 37-78, 1994.

[13] J. S. Geronimo, D. P. Hardin, P. R. Massopust, "Fractal functions and wavelet expansions based on several scaling functions," J. Approx. Theory, 78(3): 373-401, 1994.

[14] O. Christansen, Y. C. Eldar, "Generalized shift-invariant systems and frames for subspaces," J. Fourier Anal. Appl., 11: 299-313, 2005.

[15] O. Christansen, Y. C. Eldar, "Oblique dual frames and shift-invariant spaces," Appl. Compute. Harmon. Anal., 17(1): 48-68, 2004.

[16] A. Aldroubi, K. Grochenig, "Non-uniform sampling and reconstruction in shift-invariant spaces," SIAM Rev., 43: 585-620, 2001.

[17] M. Unser, "Sampling, 50 years after Shannon," IEEE Proc., 88: 569-587, 2000.

[18] I. J. Schoenberg, Cardinal Spline Interpolation. Philadelphia, PA: SIAM, 1973.

[19] Y. P. Lin, P. P. Vaidyanathan, "Periodically nonuniform sampling of bandpass signals", IEEE Trans. Circuits Syst. II, 45(3): 340-351, 1998.

[20] C. Herley, P. W. Wong, "Minimum rate sampling and reconstruction of signals with arbitrary frequency support," IEEE Trans. Inf. Theory, 45(5): 1555-1564, 1999.

[21] R. Venkataramani, Y. Bresler, "Perfect reconstruction formulas and bounds on aliasing error in sub-nyquist nonuniform sampling of multiband signals," IEEE Trans. Inf. Theory, 46(6): 2173-2183, 2000.

[22] M. Mishali, Y. C. Eldar, "Blind multi-band signal reconstruction: compressed sensing for analog signals," IEEE Trans. Signal Process., 57(3): 993-1009, 2009.

[23] M. Mishali, Y. C. Eldar, "Spectrum-blind reconstruction of multi-band signals," in Proc. Int. Conf. Acoust., Speech, Signal Processing (ICASSP), Las Vegas, NV, 3365-3368, 2008.

[24] M. Mishali, Y. C. Eldar, "From theory to practice: Sub-nyquist sampling of sparse wideband analog signals," IEEE Sel. Topics Signal Process., 4(2): 357-391, 2010.

[25] J. G. Proakis, Digital Communication, 3rd edn. McGraw-Hill, 1995.

[26] S. Kay, Fundamentals of Statistical Signal Processing, Vol II: Detection Theory, Pearson, 1998.

[27] E. Arias-Castro, Y. C. Eldar, "Noise folding in compressed sensing," IEEE Signal Process Lett. 18(8): 478-481, 2011.

[28] Y. Xie, C. Eldar, A. Goldsmith, "Reduced-dimension multiuser detection", IEEE Trans. Inform. Theory, 59(6): 3858-3874, 2013.

[29] Y. C. Eldar, "Compressed sensing of analog signals in shift-invariant spaces," IEEE Trans. Signal Processing, 57(8): 2986- 2997, 2009.

[30] S. Kay, Fundamentals of Statistical Signal Processing, Vol I: Estimation Theory, Prentice Hall, 1993.

Biography

Mohammad Kalantari received B.Sc. degree in Computer Engineering from Iran University of Science and Technology (IUST), Tehran, Iran and M.Sc. and Ph.D. in Computer Engineering from Amirkabir University of Technology (AUT), Tehran, Iran in 2001 and 2009 respectively. He is currently working as Assistant Professor at Signal Processing Laboratory in Computer Engineering Department at Shahid Rajaee Teacher Training University (SRTTU), Tehran, Iran. His area of interest includes, statistical signal processing, spherical array processing, sampling theory and compressed sensing.

- Email: mkalantari@sru.ac.ir
- ORCID: 0000-0002-6852-9344
- Web of Science Researcher ID: HZJ-9229-2023
- Scopus Author ID: 55893680300
- Homepage: https://www.sru.ac.ir/kalantari/

How to cite this paper:
M. Kalantari, "Noise folding compensation in compressed sensing based matched-filter receiver," J. Electr. Comput. Eng. Innovations, 13(1): 57-64, 2025.

DOI: 10.22061/jecei.2024.10933.749

URL: https://jecei.sru.ac.ir/article_2197.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Fusion of Classifiers Using Learning Automata Algorithm

S. Mahmoodi Khah, S. H. Zahiri *, I. Behravan

Department of Electronic, Faculty of Electrical Engineering and Computer, University of Birjand, Birjand, Iran.

Article Info

Article History:
Received 25 May 2024
Reviewed 08 July 2024
Revised 10 August 2024
Accepted 24 August 2024

Keywords:
Sonar data
Reinforcement learning
Learning automata
Data classification
Analytical parameters

*Corresponding Author's Email Address: *hzahiri@birjand.ac.ir*

Abstract

Background and Objectives: Sonar data processing is used to identify and track targets whose echoes are unsteady. So that they aren't trusty identified in typical tracking methods. Recently, RLA have effectively cured the accuracy of undersea objective detection compared to conventional sonar objective cognition procedures, which have robustness and low accuracy.

Methods: In this research, a combination of classifiers has been used to improve the accuracy of sonar data classification in complex problems such as identifying marine targets. These classifiers each form their pattern on the data and store a model. Finally, a weighted vote is performed by the LA algorithm among these classifiers, and the classifier that gets the most votes is the classifier that has had the greatest impact on improving performance parameters.

Results: The results of SVM, RF, DT, XGboost, ensemble method, R-EFMD, T-EFMD, R-LFMD, T-LFMD, ANN, CNN, TIFR-DCNN+SA, and joint models have been compared with the proposed model. Considering that the objectives and databases are different, we benchmarked the average detection rate. In this comparison, Precision, Recall, F1_Score, and Accuracy parameters have been considered and investigated in order to show the superior performance of the proposed method with other methods.

Conclusion: The results obtained with the analytical parameters of Precision, Recall, F1_Score, and Accuracy compared to the latest similar research have been examined and compared, and the values are 87.71%, 88.53%, 87.8%, and 87.4% respectively for each of These parameters are obtained in the proposed method.

Introduction

Underwater objective classification and cognition technology have become a research issue due to the detection and extension of oceans and seas [1]. Sonar determines the distance and direction of underwater targets utilizing sound. Sound waves emitted from the target are detected by it and analyzed to calculate range information. It should also be said that sonar measurements are not affected by turbidity or reduced light and color, thus making it a good complement to a camera. Because of the relative ease of undersea propagation compared to other patterns of radiation, acoustic waves have been widely used for undersea discovery and other points [2], [3].

The point estimation of the target position is usually done by thresholding the normalized data and announcing the diagnosis when the threshold is crossed. Due to the sufficient and high signal-to-noise ratio (SNR) of the target echo, this approach is reliable and efficient. Because the target echoes are most likely in the survival threshold process, and connecting only the detected target positions does not require a large number of calculations. A low echo level, either due to reduced source level or low target power, makes detecting and tracking the conventional pulse active sonar less reliable.

The recognition of such targets requires a lower threshold at the cost of more false detections.

In reference [1], an identification and classification algorithm are proposed to solve this problem. This research has proposed a lightweight target detection model for small samples using the improved YOLOV4 algorithm. The improved image feature extraction network in this paper has greatly reduced the number of network parameters, and the parameters of the feature fusion module have been improved. However, this algorithm has difficulty in detecting small targets in the image and detecting targets with unusual sizes. Many researchers have enhanced and proposed various reinforcement learning algorithms (RLAs) for sonar objective classification and cognition [4]. Williams planned a feature classification and exploitation and network with only ten convolution layers for sonar objective classification. The training network proposed in this paper has increased the amount of training data images for learning. In this research, image feature integration has been used and its performance has been displayed only in the form of an AUC diagram [5]. Valdenegro-Toro et al. used a convolutional neural network (CNN) to detect the object of an undersea sonar image, and after training the network, the average detection rate in test sets reached 90% [6]. A sonar objective cognition procedure based on a shallow CNN has fault cognition and insufficient model strength. Ferguson et al proposed the use of a deep CNN to detect the sound of an undersea ship in a shallow water ambiance. In this article, a data augmentation technique is introduced, and the criterion for comparing data integration performance at the feature level is the precision parameter [7].

Huo et al. proposed a classification method for sonar target detection based on semi-synthetic data training and transfer learning for small sample sonar datasets. Experiments indicate that transfer training and semi-synthetic training can help increase model cognition accuracy [8]. In reference [9], a hybrid dragonfly algorithm is proposed to train a multi-layer perceptron (MLP) neural network to design a classifier in solving complex issues and to distinguish true targets from fake objectives in sonar applications. In this paper, by combining DA and ChoA algorithms, the researchers were able to achieve a suitable classification rate and execution time compared to the separate performance of each algorithm.

Researchers have developed many related algorithms, such as the combined probability filter algorithm, which can effectively filter the confusion in the data with the object's motion characteristics [10]. In another algorithm, fuzzy least squares regression for filtering is combined with joint probability data fusion (DF) filtering to achieve efficient target tracking [11]. Also, the new multi-sensor probabilistic hypothesis density filter algorithm can combine data from different sensors and overcome the problems of statistical information loss [12]. Environment-based performance is emphasized to obtain the most expected benefits in reinforcement learning (RL), which is one of the main branches in the field of machine learning (ML). A new data tracking and communication network structure is also developed in this field based on RL networks [13].

Various scattering mechanisms affect object-specific information in a received signal in a sonar system. Collected signals are contaminated by noise, reverberation, and confusion in the ocean environment. ML methods are traditionally used for feature exploitation and classification of active sonar data but lack interpretation. This may lead to a decrease in the confidence of the algorithm, and the reasoning of the classifier becomes unknown. Explainable artificial intelligence (AI) is a field that increases the transparency of ML algorithms by making them humans interpretable. Data fusion is the process of combining data or information to create improved estimates or predictions of the state of an entity [14]. Information obtained from a source may be unreliable or insufficient to determine accuracy. Therefore, it is necessary to use multiple data sources to increase the reliability and quality of information provided to decision-makers.

DF is especially important in many applications where a large amount of data must be fused and then intelligently combined. On the other hand, data aggregation in wireless sensor networks has a special meaning and place. So that in a wireless sensor network, a huge amount of data comes from nodes, sensors, and different input channels, and certain data must be collected before sending these data to some other nodes, outputs, sink nodes, etc. are data aggregations. Finally, sensor data fusion is done to obtain advanced quality information with appropriate integrity so that the decisions made based on these data and the integrated fused information are very reliable. Which should be more accurate about the overall situation, the target situation, the process, and the scenario of interest by reducing the uncertainty. DF should be done logically and with proper understanding of data processing methods and related methods [15].

ML is a subset of AI where a machine learns how to complete a given task without explicitly planning how to do it, by feeding it plenty of training data and building a good model to predict the true values for recent similar data. A common definition is also provided for ML. It is called a computer program that learns from experience according to a set of tasks and performance criteria. If its function in the tasks is the same as the measured

efficiency and enhances with experience [16]. In supervised ML, a dataset is given to the learning algorithm along with labels that indicate how much the correct output should be for the given data. Algorithms such as support vector machine (SVM), k-nearest neighbor (KNN), random forests (RF), and artificial neural networks (ANN) are examples of this learning.

Using six ML algorithms such as KNN, SVM, RF, decision tree (DT), extreme gradient boosting (XGboost), and ensemble methods, Krishna et al. conducted research with the help of sonar data to find sea mines. The Ensemble method is the combination of RF, XGboost, and Voting Classifier. Comparative results including Accuracy, Precision, Recall, and F1-score for all these algorithms are presented in this paper [30]. In this research, Wang et al presented a method of identifying active sonar targets based on multi-domain transformations and precision-based fusion networks. The results of the experiments show that by using multi-domain transformations, active sonar echoes can be accurately detected. Improved by 10.5% compared to single domain methods. Also, the findings show that in a high-level feature space by combining features of multiple transformations, more informative and effective results are obtained for active sonar echoes. In addition, the identification performance of different fusion models such as the early fusion model with resnet (R-EFMD) as the backbone of multi-domain attention-based feature extractor (MAFE), early fusion model with swin transformer (T-EFMD) as the backbone of MAFE, late fusion model with resnet (R-LFMD) as the backbone of single domain feature extractor (SFE) no attention-based feature extractor (AFE) module, and late fusion model with swin transformer (T-LFMD) as the backbone of SFE no AFE module has been compared [31].

Ahmed et al. investigated an underwater audio signal classification model with deep learning method. A regular neural network is also implemented to classify audio as input features. Comparing the performance of this classifier and the general results of the presented models is promising [32]. Yang et al. implemented a spatial attention deep convolutional neural network for marine mammal call detection. This method tends to use spatial attention (SA) to help the deep convolutional neural network (DCNN) to achieve better detection performance. Time-frequency image recognition-DCNN (TFIR-DCNN) is designed at the beginning of this method. Then, SA is added to the TFIR-DCNN to help the TFIR_DCNN focus on the location of call features in the time and frequency domains. Favorable marine mammal contact detection test results have been reported [33]. Tian et al. designed a collaborative learning model for underwater acoustic target recognition. In this research, firstly, a light multiscale residual deep neural network (MSRDN) is implemented using light network design techniques, where 64.18% of the parameters and 79.45% of the floating-point operations (FLOPs) from the original MSRDN are reduced in accuracy. It decreases a little. Then, a combined model of wave representation and time-frequency-based models was presented. The results of deterministic experiments prove that the performance improvement of the proposed methods from mutual deep learning has advantages such as favorable recognition accuracy [34].

Motivation and Innovation

In this work, the data fusion problem in sonar data classification is considered due to its importance in various applications such as navigation and marine surveillance. However, we must mention that the mechanism of learning automata has not been used in this field yet. We intended to check whether using mechanisms related to learning automata can be effective and efficient in data fusion at the decision level. Although data integration at the level of data, decision, and feature has been used in the problem of sonar data classification. But until now, the use of a machine learning method such as learning automata to increase the ability to classify targets has been neglected. In this article, we measured the remarkable performance of the proposed method for 5 different objectives with Precision, Recall, F1_Score, and Accuracy indicators. Noise and acoustic interferences make the act of identification difficult in the vast and diverse oceanic and marine environments. In most marine devices, target detection is done by human operators, and with the development of this method in detecting most different targets, the speed and accuracy of identification can be increased and human errors can be reduced in these cases.

Algorithms

To increase the accuracy of the classification of complex problems, it is possible to use a combination of classifications that use the same learning algorithm but with different complexities and parameters. Hybrid classifications use the fusion of several classifiers. In fact, these classifiers each build their own model on the data and save this model. Next, for the final classification, a vote will be held between these classifications, and the class that gets the most votes will be the class that has had the greatest impact on the classification. The goal of AI is to train computers to do the things that persons currently do better, and without a doubt, learning is the most significant of those targets [17].

K-Nearest Neighbor

In the KNN, an objective is classified by the majority of its neighbors' votes, and the target is classified into the class that is most general between its KNNs [18]. KNN is a classification algorithm and there are mainly two phases

in classification. The first phase is learning, in which a classification is made using the training data, and in the second phase, the evaluation of the classifier is done. Fig. 1 shows a simple KNN structure [19].

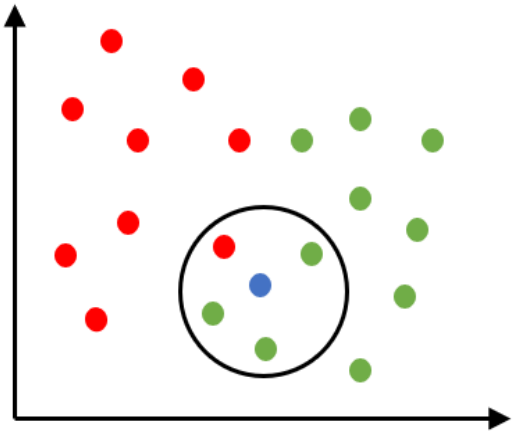


Fig. 1: A simple KNN classification.

As presented in Fig. 2. The new unlabeled data computes the distance of each of its neighbors according to the K value. Then, it specifies the class that contains the maximum number of nearest neighbors to it [20].

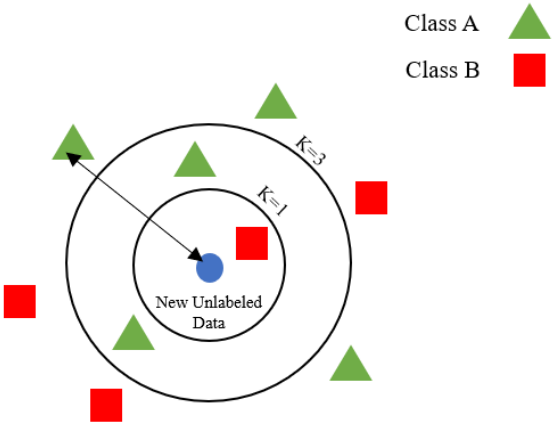


Fig. 2: New unlabeled data.

After collecting KNN, we simply select most of them to predict the training sample class. Agents that affect the operation of this algorithm are K value, Euclidean distance and parameter normalization. For a precise understanding of the algorithm's performance and according to the set of training data shown in (1), the steps are as follows.

$$\left\{\begin{matrix}(x(1),y(1)),(x(2),y(2)),\dots \\ \dots,(x(m),y(m))\end{matrix}\right\} \quad (1)$$

First, the training set is stored, and then the Euclidean distance for each new unlabeled data among two points x and y in all training data points is calculated using (2).

$$d=\sqrt{\sum_{k=1}^{N}(x_k-y_k)^2} \quad (2)$$

KNNs are determined, and the maximum number of nearest neighbors is assigned to a class. After saving the training, all the parameters should be set to normal, so that the calculations become easier. The value of K affects the algorithm because it can be used to create the boundaries of each class. The best solution is selected first by checking the data. Larger solutions of K are more accurate because they decrease the net noise, but this is not guaranteed [21].

Multi-Layer Perceptron

ANNs are structures inspired by brain performance. These networks can compute model performance estimation and manage non-linear and linear functions by learning from data generalizing and their relationships to unsighted situations. One of the most main ANNs is MLP. It is a potent modeling tool that exerts a supervised learning method using data samples with certain outputs. This method creates a non-linear function model that makes it possible to predict the output data from the given input data [22].

In order to comprehend MLP, a short description on single layer perceptron (SLP) and single neuron perceptron has been prepared. The first type is the simplest ANN and has only one output to which all inputs are linked, and the values of x_i, w_i and y are inputs, weighting of the neuron and predictive binary class respectively, which are described in Fig. 3 of the steps of weighting, summation and transfer function. Also, Fig. 4 shows its simplified model and the transfer function is calculated in (3).

Fig. 3: Perceptron steps: weighting, sum and transfer steps.

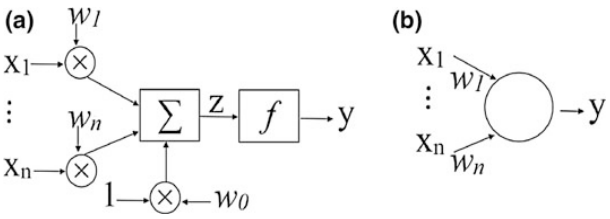


Fig. 4: Perceptron models: a) steps. b) Simplified.

$$y=f(z)\ and\ z=\sum_{i=0}^{n}w_ix_i \quad (3)$$

x_0=1, y is the output and w_0 is the bias or threshold value. The transfer function has different forms such as unit step, linear, and sigmoid. Fig. 5 shows an example of the linear and nonlinear functions, which detaches the data into two classes. A Function can be represented by

the dot product among the input and the weight vectors in (4).

$$\sum_{i=0}^{n} w_i x_i = 0 \quad (2)$$

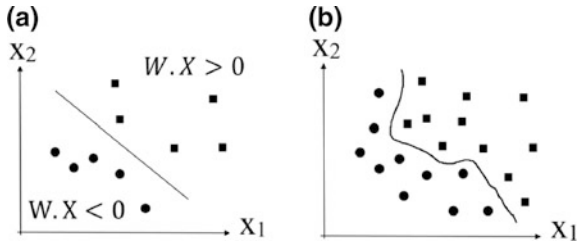


Fig. 5: Input patterns: a) linear. b) nonlinear.

Connecting many perceptrons in parallel creates a SLP structure that is used for different outputs. Fig. 6 represents an example where output and input layers are presented in a multi-class situation that can be linearly separated. The SLP does not solve separable nonlinear problems, which can be seen in Fig. 5. In this case, which is also shown in Fig. 6, a response can be found by appending any number of layers in a sequential order and making a MLP structure [23]. Any connection among neurons has its own weight, and SLP has the same activation function. Hinging on the performance, the output layer can be various functions [24].

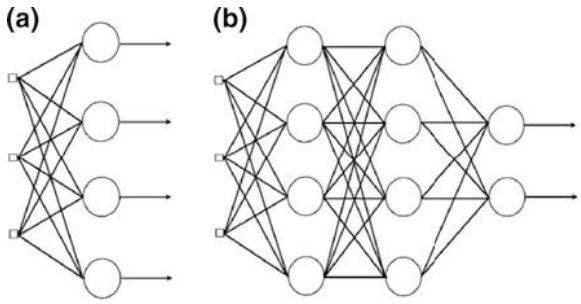


Fig. 6: Layer structure: a) SLP. b) MLP.

Learning Automata Algorithm

Automatic learning is an easy model for adaptive decision-making in anonymous stochastic ambiances. Allegedly, its performance can be supposed identically to the learning method by a living organism in such ambiances. General instances of such positions are cases where an inexperienced person learns to perform the right motions or an individual who finds the best track from home to the office. The structure efforts various operations and chooses new operations based on the response of the environment to the past acts. The structure of such adaptive selection of operations and decisions is indicated by learning automata. The learning problem the appropriate operation is complicated by the verity that ambiance responses are not entirely reliable because they are stochastic and the corresponding probability distribution is anonymous.

This model is effective in many functions related to adaptive decision-making. Hence, it would be attractive to have an algorithm that can learn appropriate selections based on some noisy evaluation of the good choice, which is consistent with the automata model. A classifier must decide on the class label of each pattern input to it in a pattern recognition problem. The law of optimal decision-making can be considered as a learning problem for choosing one of the available actions based on some random feedback about the appropriate of each selection [25].

The learning method in the field of LA is as follows. Every time it cooperates with the environment, it automatically and stochastically selects an action based on a probability distribution. After the ambiance responds to a chosen action, it automatically updates its operation probability distribution. Then, a new operation is chosen according to the updated probability distribution, and the solution of the environment is extracted for this act, and this method is rerun. The updated algorithm for the operation probability distribution is called the RLA [25].

The general method of LA, which is an unsupervised optimization procedure and one of the key parts in adaptive learning systems, is to perform an operation through cooperation with the ambience in terms of receiving a sequence of repeated evaluation cycles with selecting the highest reward compared to other operations. By learning to select the best solution, automata adapt without needing to have detailed data about the pattern of the environment [26]. The cooperation between the environment and LA is shown in Fig. 7. In [27], for the first time, the idea of automatic learning was introduced to model the mechanism of biological learning.

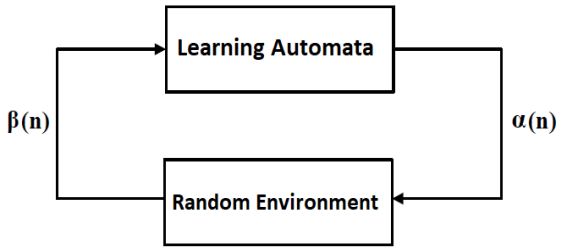


Fig. 7: The cooperation of LA with the environment.

LA is a self-organizing decision-making unit whose performance improves through repeated cooperation with a stochastic environment. A LAA learns how to select the best solution based on the response it receives from the environment. In this process and interaction, repetition number n starts when the automata select the input vector x(n) from the set $X \in R^m$ from the environment. According to the input vector, the

automata select one of its possible actions and apply it to the random environment (for example α(n) ϵ α). Then, the stochastic environment classifies the selected α(n) act in the x(n) input vector and estimates an amplification signal of β(n) ϵ β. For this purpose, the automata use the learning algorithm T, the x(n) input vector, the α(n) act, and the β(n) reinforcing signal to update its state. By repeating or continuing this process, the LA learns how to choose the optimal solution [28].

LA is more practical and effective in discovering the exact best solutions for complicated optimization issues. The dimensions of the points are equal to the number of automata used in the LAA. In other words, for the N-dimensional problem, this algorithm contains N automata [29]. Each automata is accountable for exploring one dimension and operates separately in the ambiance. The i-th LA can be defined as the model ⟨x_i, A_i, r, P_i, U⟩ where $x_i = \{x_i\}$ represents the set of possible positions in the i-th dimension. As well as, x_i is the next state in the dimension i ($x_i \in [x_{min,i}, x_{max,i}]$), the maximum and minimum amounts in dimension i are $x_{max,i}$ and $x_{min,i}$, respectively. In automatic learning, $A_i = \{a_{l,\eta}\}$ is the set of possible operations that the LA can perform in the dimension i, $a_{l,\eta}$ demonstrates that an operation is right (l=2) or left (l=1) moves and η is the length of step. Note that r is a scalar value and represents a reinforcement signal that is generated through the ambiance to demonstrate the quality of the movement x_i during the step in the selected route. As well as, P_i includes two possibilities p_1 and p_2. p_1 and p_2 respectively demonstrate the probability of choosing the right route and the left route in the i-th dimension. Suppose the right route is chosen, and the probability of selecting one cell among k cells located on the route determines the probability p_2. As well as, U is a procedure for calculating the probabilities of operations, P.

In the introduced procedure, each dimension is parted into D cells. This intends that x_i is parted into D subsets, and each subset comprises all dimensional states located in the cell. Thus, D×N cells are generated for the N-dimensional space of exploration where $\omega_{c,i}$ is a cell width in the dimension i and is computed using (5).

$$\omega_{c,i} = \frac{x_{max,i} - x_{min,i}}{D} \tag{3}$$

At the beginning of the operation exploration, it must be able to select one of two possible directions to appraise the selection of the best solution in the route. Therefore, the value of $L_2(x_i)$ is determined by the amounts of the k adjacent cells in the right route, where k is a predefined integer amount and $c_{i,j}$ is cell j in dimension i. As well as, j is computed by (6) and the amount of a route can be evaluated by (7).

$$j = floor\left(\frac{x_i - x_{min,i}}{\omega_{c,i}}\right) \tag{6}$$

$$\boldsymbol{L}_l(x_i) = (1-\lambda_1)\sum_{m=1}^{k-1}\lambda_1^{m-1}v_{l,m}^* + \lambda_1^{k-1}v_{l,k}^* \tag{4}$$

$$l = 1,2$$

where $v_{l,m}^*$ represent the variable of the vector m that is placed in the direction of l. Also, λ_1 is computed with the conditions $0 \le \lambda_1 \le 1$ and $(1-\lambda_1)\sum_{m=1}^{k-1}\lambda_1^{m-1} + \lambda_1^{k-1} = 1$, provided that the relation $(1-\lambda_1)\lambda_1^{k-2} \ge \lambda_1^{k-1}$ is established. The two probabilities p1 and p2 are obtained from (8) and (9).

$$p_1\big(L_l(x_i)\big) = \frac{e^{\frac{L_l(x_i)}{\tau}}}{\sum_{s=1}^{2} e^{\frac{L_s(x_i)}{\tau}}} \quad l = 1,2 \tag{5}$$

$$p_2\big(c_{i,j+s}\big) = \frac{e^{\frac{(V(x_i)|x_i \epsilon c_{i,j+s})}{2\tau}}}{\sum_{z=1}^{k} e^{\frac{(V(x_i)|x_i \epsilon c_{i,j+z})}{2\tau}}} \tag{6}$$

$$l = 1,2 \quad s = 1,\dots,k$$

where $V(x_i)$ is the cell value. The τ parameter makes a balance among search and utilization. With selecting a cell, the operation proceeds to the new cell with a step length that can be expressed in the act of η in (10). Thus, when L_1 is chosen, the current dimensional state of x_i changes to $x_i = x_i - \eta$ and when L_2 is selected, x_i moves to $x_i = x_i + \eta$.

$$\eta = \omega_{c,i}(\xi + \zeta) \tag{7}$$

where the distance among the former cell and the chosen cell ζ and ξ is a stochastic number (ζ ∈ (0, 1]). Next, an amplification signal is applied to investigate the next state x_i. Just after the dimensional state x_i is transferred to x_i', the i-th variable of the current state $X(x_i)$ is changed by $X(x_i')$. According to (11), the amplification signal is allocated to cell $c_{i,j}$. The amplification signal is used to update the cell value ci, j and is obtained according to (12).

$$r\big(X(x_i')\big) = \begin{cases} 1, & if\ F(X(x_i')) \le F(X_{best}) \\ 0, & otherwise \end{cases} \tag{8}$$

$$V(x_i)|_{x_i \epsilon c_{i,j}} \leftarrow r\big(X(x_i)\big) + \alpha_1 V(x_i)|_{x_i \epsilon c_{i,j}} \tag{9}$$

$$+(1-\alpha_1)\big((1-\lambda_2)L_{max}(x_i) + \lambda_2 L_{min}(x_i)\big)$$

The solution is desirable when r=1 and r=0 indicates an unfavorable answer. Also, $L_{max}(x_i) = \max\{L_1(x_i), L_2(x_i)\}$ and $L_{min}(x_i) = \min\{L_1(x_i), L_2(x_i)\}$ are two estimated path values at x_i. $L_{max}(x_i)$ has a greater impression on the cell value than $L_{min}(x_i)$. Thus, the parameter λ_2 must be given in such a way that this relation $(1-\lambda_2) > \lambda_2$ is true. The weights α_1 and $(1-\alpha_1)$ show the impression of past evaluations and route values on the new evaluation, respectively. In (13),

the relationship among X_{best} and X and is shown.

$$X_{best} \leftarrow \begin{cases} X(x_i'), X(x_i') \\ = [x_i, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_N] \\ X_{best} \end{cases} \begin{matrix} if\ r = 1 \\ otherwise \end{matrix} \quad (10)$$

Methodology

To increase the classification accuracy of complex problems, it is possible to use a combination of classifications that use the same learning algorithm but with different complexities and parameters. Hybrid classifiers use the fusion of several classifiers. In fact, these classifiers each build their own pattern on the data and save this model. Eventually, for the final classification, a vote is held between these classifications, and the class that gets the most votes will be the class that has had the greatest impact on the classification. In this work, we defined coefficients to weight the classifiers, and in order to achieve the best accuracy, we implemented voting and finding the optimal coefficients by automata learning algorithm. We proceeded with this process in five steps. Fig. 8 shows the overall process.

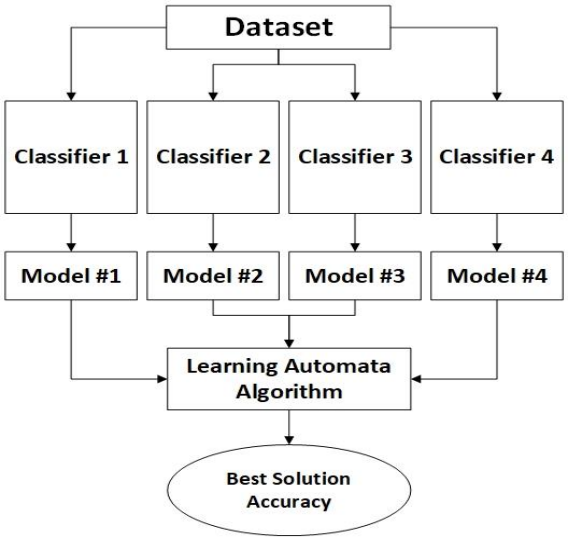


Fig. 8: The overall process of the proposed method.

In the first step, we created and stored sonar data in five classes with specific dimensions and samples.

In the second step, we loaded those data into the introduced classification training algorithm and after running the algorithm, we saved the accuracy results of each of the classification models related to the sonar data. Four classifiers (two KNN classifiers and two MLP classifiers) were used in this research.

In the third step, the stored models and data were loaded into the LA algorithm.

In the fourth step, we created and integrated functions for weighting the categories.

In the last step, to find the best accuracy answer with the majority vote, we ran the LA algorithm to find the optimal coefficients of the classifiers and saved the results.

Data and Device

In this work, a dataset of sonar targets with five different classes and dimensions of 103x129 was used. Also, these targets in different subclasses include different viewing angles and signal-to-noise. The Specifications of targets are demonstrated in Table 1.

Table 1: Specifications of objectives

Class Number	Name	Type of Application
1	MV Barzan	container carrier
2	Front Century	oil tanker
3	Harmoney of the Seas	Cruise
4	Atlas Pishro	passenger ship
5	logistic	Military

This program is implemented on a system with Intel® Core™ i7-6500U CPU (2.50-2.59) GHz processor specifications, 8 GB RAM, and MATLAB R2020b software.

Results and Discussion

In this work, we are going to investigate the improvement of the performance of combining the classifications using the automatic learning algorithm.

Also, to better check the efficiency of the used models, Accuracy, Precision, Recall, F1_Score, and AUC parameters are reported in Table 2. Also, the test charts of each model are shown in Figs 9 to 16.

In the first model, the data was trained by a KNN classifier with a nearest neighbor rate of 3. The performance of model 1 on sonar data with confusion matrix and ROC charts for 5 different classes is shown in Fig. 9 and Fig. 10.

In the second model, the data was trained by a KNN classifier with a nearest neighbor rate of 15. The performance of model 2 on sonar data with confusion matrix and ROC charts for five different classes are demonstrated in Fig. 11 and Fig. 12.

In the third model, the data was trained by an MLP classifier with an input layer of 15. The performance of model 3 on sonar data with confusion matrix and ROC charts for 5 different classes is shown in Fig. 13 and Fig. 14.

In the fourth model, the data was trained by an MLP classifier with an input layer of 15 and a hidden layer of 15. The performance of model 4 on sonar data with confusion matrix and ROC charts for 5 different classes is shown in Fig. 15 and Fig. 16.

Table 2: Machine learning Models performance results

Model Number	Precision (%)					Recall (%)					F1_Score					AUC					Accuracy (%)
	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	
1	77.27	78.94	100	100	69.56	94.44	62.5	90.47	100	80	85	69.76	95	100	74.41	0.99	0.94	0.99	1	0.95	84.46
2	100	70.58	72.72	83.33	56.25	44.44	50	76.19	100	90	61.53	58.53	74.41	90.90	69.23	0.97	0.93	0.93	1	0.91	71.84
3	100	90.47	72.72	60.60	78.26	83.33	79.16	38.09	100	90	90.90	84.44	50	75.47	83.72	0.91	0.88	0.67	0.92	0.92	77.67
4	100	78.94	72	83.33	61.53	50	62.5	85.71	100	80	66.66	69.76	78.26	90.90	69.56	0.75	0.78	0.88	0.97	0.84	75.72

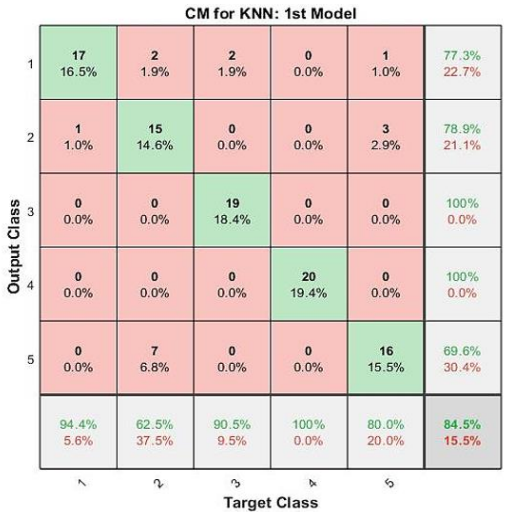


Fig. 9: Confusion matrix chart for KNN - 1st Model.

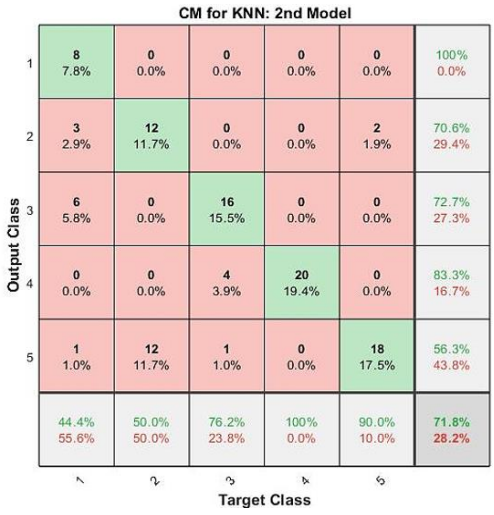


Fig. 11: Confusion matrix chart for KNN - 2nd Model.

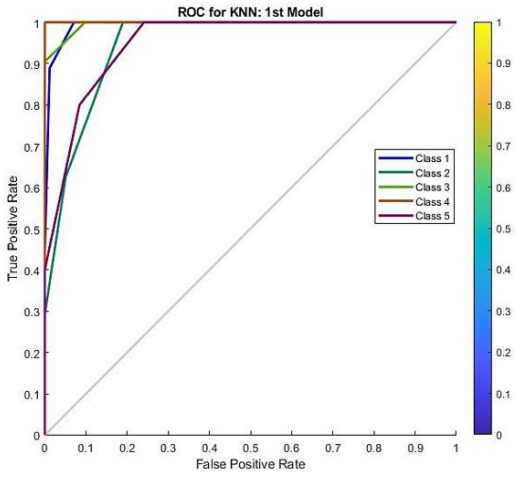


Fig. 10: ROC chart for KNN - 1st Model.

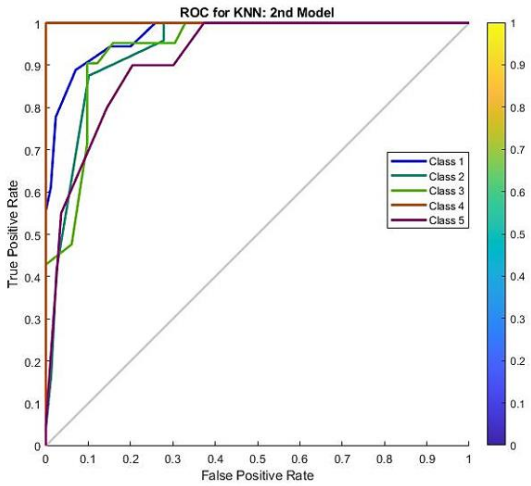


Fig. 12: Roc chart for KNN - 2nd Model.

As described in the work process in the previous sections. The stored models of each class are weighted using the LAA and weighted summation functions in the defined range. To achieve the best accuracy and decision by obtaining the best solutions for the classifications and fusion it by the LA algorithm.

Due to the fact that in this process the effective parameters in the learning automata algorithm are very effective.

The results of Accuracy, Precision, Recall, F1_Score, and AUC are reported separately for the impact of each of the K, D, and N_{femax} parameters.

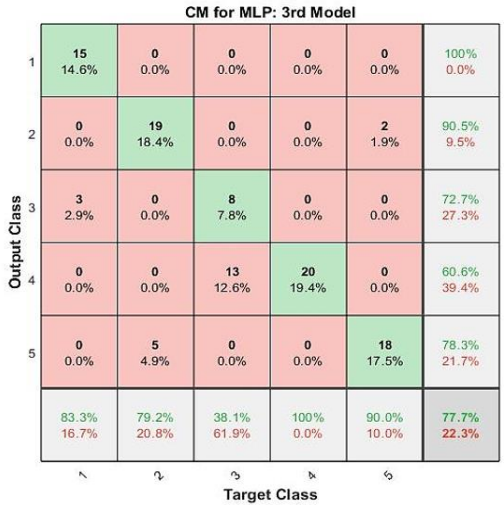


Fig. 13: Confusion matrix chart for MLP - 3rd Model.

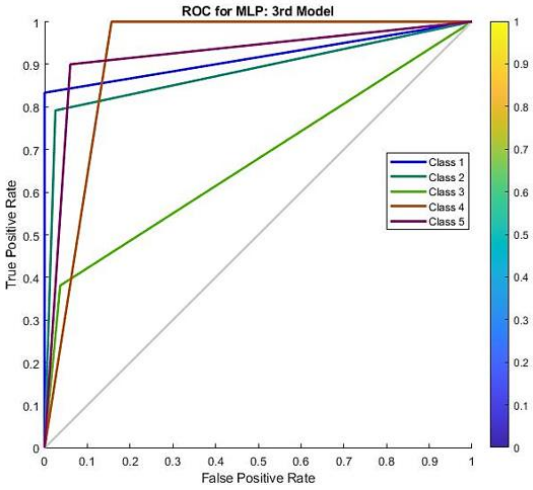


Fig. 14: ROC chart for MLP: 3rd Model.

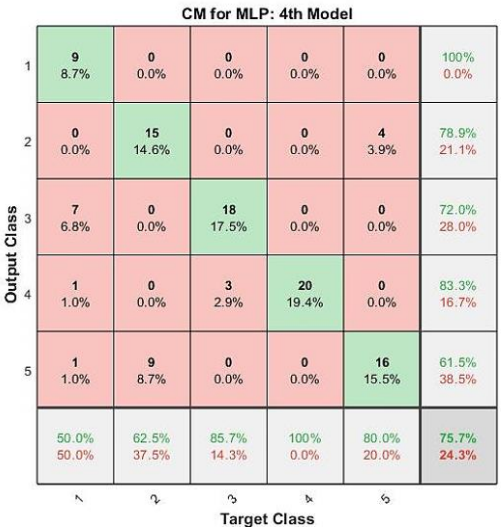


Fig. 15: Confusion matrix chart for MLP - 4th Model.

In Table 3, the parameter D = 50 is considered constant and the results are reported by changing the values of K and N_{femax} parameters. Also, the performance of sonar data fusion by LAA is shown using the confusion matrix and ROC charts in Figs 17 to 22.

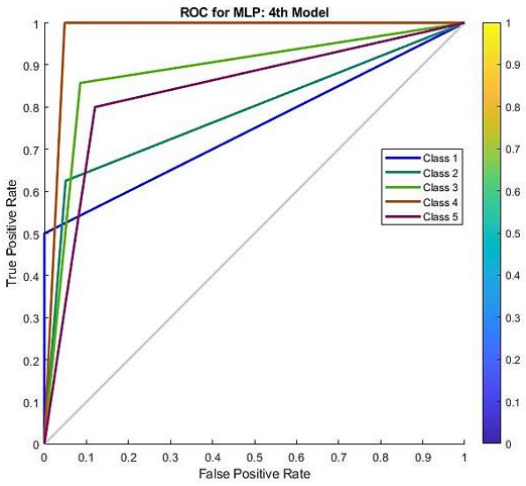


Fig. 16: ROC chart for MLP - 4th Model.

The performance of sonar data fusion by the learning automata algorithm for D = 50, K = 10, and N_{femax} = 5 values is shown in Fig. 17 and Fig. 18.

Fig. 17: Confusion matrix chart for D = 50, K = 10, and Nfemax = 5 parameters in LAA.

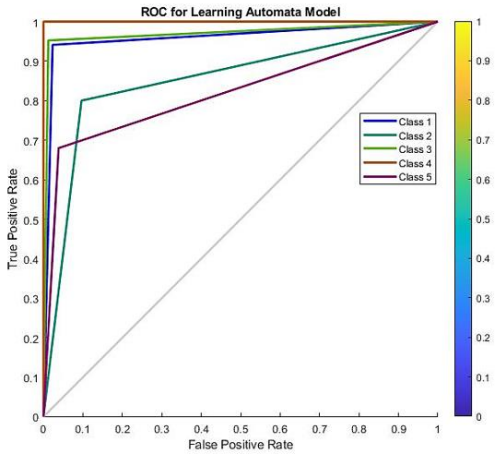


Fig. 18: ROC chart for D = 50, K = 10, and Nfemax = 5 parameters in LAA.

Table 3: LA performance results with the influence of Nfe and K parameters

The results for D=50 in LAA																									
Run Number	LAP		Best Solution				Precision (%)					Recall (%)					F1_Score					AUC					Accuracy (%)
	Nfe	K	W1	W2	W3	W4	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	
1	5	10	0	3	0	3	88.88	66.66	95	100	85	94.11	80	95.23	100	68	91.42	72.72	95.23	100	75.55	0.95	0.85	0.97	1	0.82	86.4
2	10	100	4	1	5	3	83.33	66.66	90.47	100	95	93.75	88.88	90.47	95.23	70.37	88.23	76.19	90.47	97.56	80.85	0.95	0.89	0.94	0.97	0.84	86.4
3	15	200	3	1	4	3	88.88	66.66	95	100	85	94.11	80	95.23	100	68	91.42	72.72	95.23	100	75.55	0.95	0.85	0.97	1	0.82	86.4

The performance of sonar data fusion by the learning automata algorithm for D = 50, K = 100, and N_{femax} = 10 values is shown in Fig. 19 and Fig. 20.

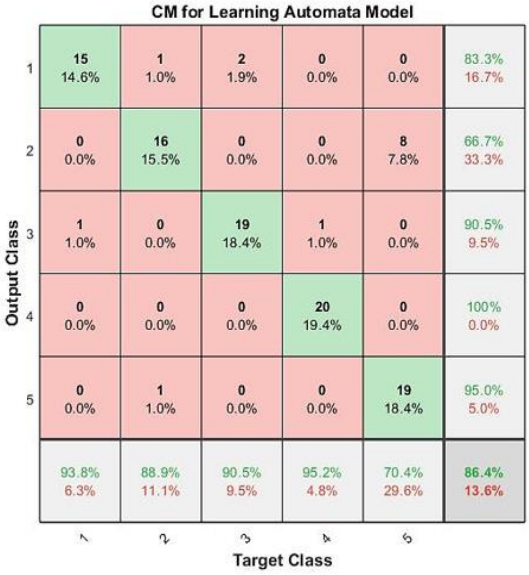


Fig. 19: Confusion matrix chart for D = 50, K = 100, and Nfemax = 10 parameters in LAA.

The performance of sonar data fusion by the learning automata algorithm for D = 50, K = 200, and N_{femax} = 15 values is shown in Fig. 21 and Fig. 22.

In Table 4, the parameter K = 100 is considered constant and the results are reported by changing the values of D and N_{femax} parameters. Also, the performance of sonar data fusion by LAA is shown using the confusion matrix and ROC charts in Figs 23 to 28.

The performance of sonar data fusion by the learning automata algorithm for K = 100, D = 10, and N_{femax} = 5 values is shown in Fig. 23 and Fig. 24.

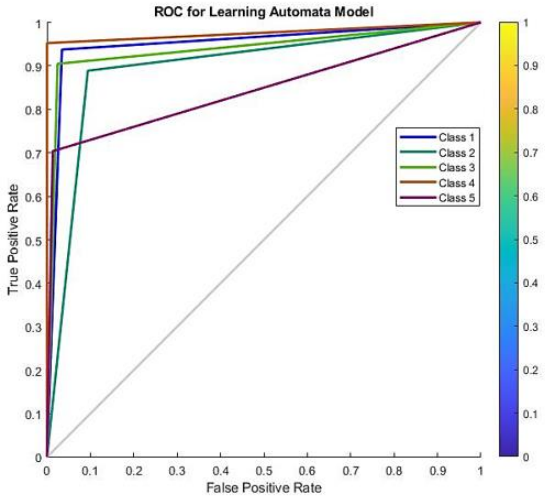


Fig. 20: ROC chart for D = 50, K = 100, and Nfemax = 10 parameters in LAA.

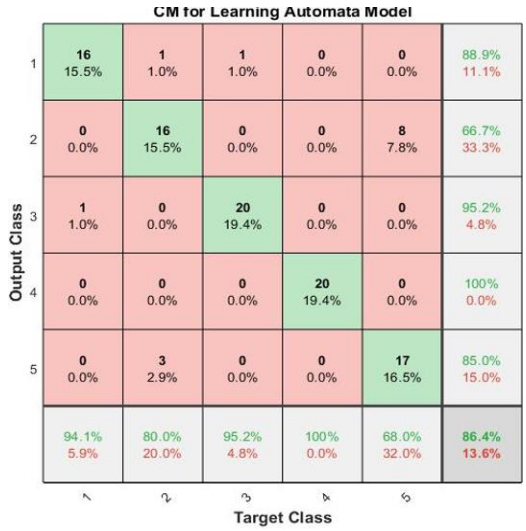


Fig. 21: Confusion matrix chart for D = 50, K = 200, and Nfemax = 15 parameters in LAA.

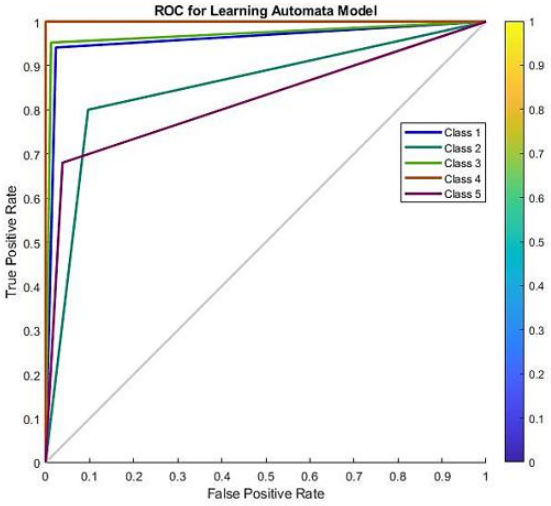


Fig. 22: ROC chart for D = 50, K = 200, and Nfemax = 15 parameters in LAA.

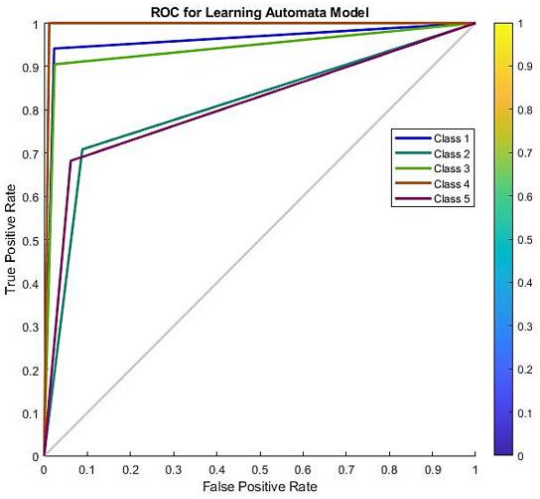


Fig. 24: ROC chart for K = 100, D = 10, and Nfemax = 5 parameters in LAA.

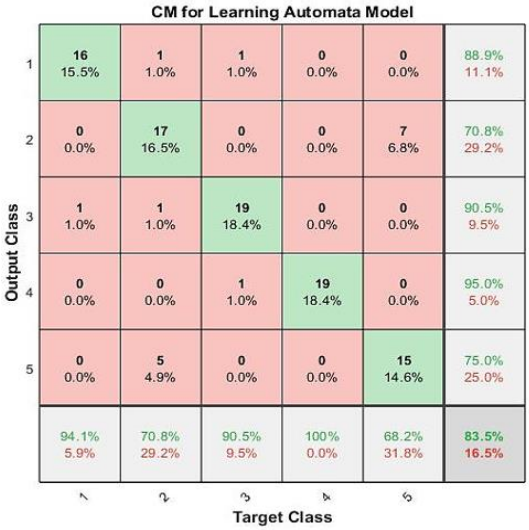


Fig. 23: Confusion matrix chart for K = 100, D = 10, and Nfemax = 5 parameters in LAA.

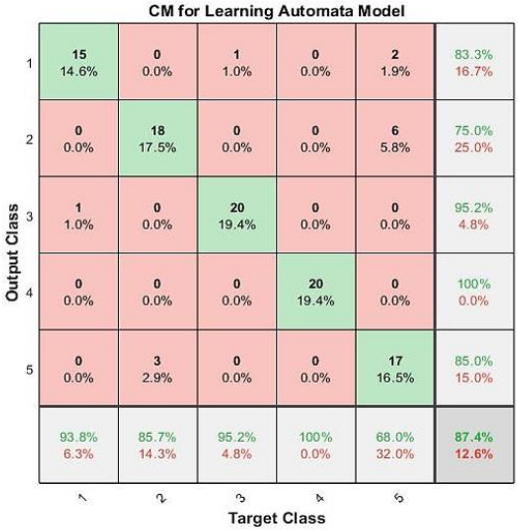


Fig. 25: Confusion matrix chart for K = 100, D = 50, and Nfemax = 10 parameters in LAA.

Table 4: LA performance results with the influence of Nfe and D parameters

The results for K=100 in LAA																												
Run Number	LAP		Best Solution				Precision (%)					Recall (%)					F1_Score					AUC					Accuracy (%)
	Nfe	D	W1	W2	W3	W4	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	
1	5	10	4	3	4	4	88.88	70.83	90.47	95	75	94.11	70.83	90.47	100	68.18	91.42	70.83	90.47	97.43	71.42	0.96	0.81	0.94	0.99	0.81	83.5
2	10	50	2	2	1	3	83.33	75	95.23	100	85	93.75	85.71	95.23	100	68	88.23	80	95.23	100	75.55	0.95	0.89	0.97	1	0.82	87.4
3	15	100	0	5	3	2	83.33	75	95.23	100	85	93.75	85.71	95.23	100	68	88.23	80	95.23	100	75.55	0.95	0.89	0.97	1	0.82	87.4

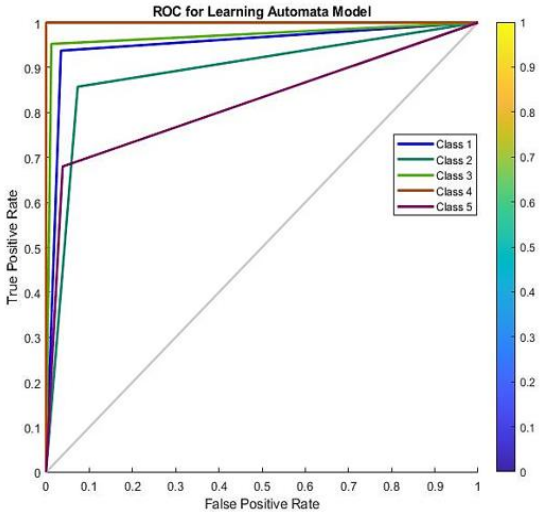


Fig. 26: ROC chart for K = 100, D = 50 and, Nfemax = 10 parameters in LAA.

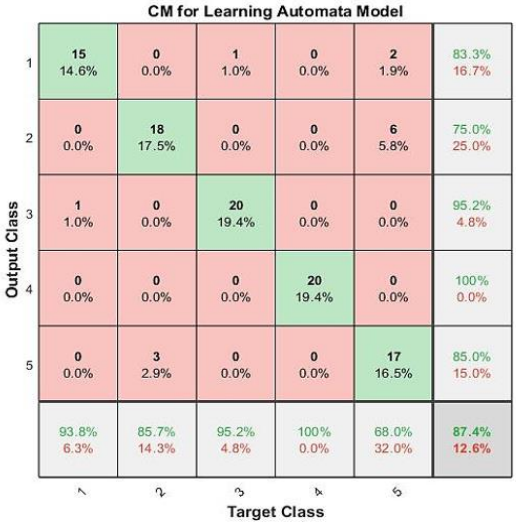


Fig. 27: Confusion matrix chart for K = 100, D = 100, and Nfemax = 15 parameters in LAA.

The performance of sonar data fusion by the learning automata algorithm for K = 100, D = 50, and N_{femax} = 10 values is shown in Fig. 25 and Fig. 26.

The performance of sonar data fusion by the learning automata algorithm for K = 100, D = 100, and N_{femax} = 15 values is shown in Fig. 27 and Fig. 28.

In Table 5, the parameter N_{femax} = 10 is considered constant and the results are reported by changing the values of K and D parameters. Also, the performance of sonar data fusion by LAA is shown using the confusion matrix and ROC charts in Figs 29 to 34.

The performance of sonar data fusion by the learning automata algorithm for N_{femax} = 10, K = 10, and D = 10 values is shown in Fig. 29 and Fig. 30.

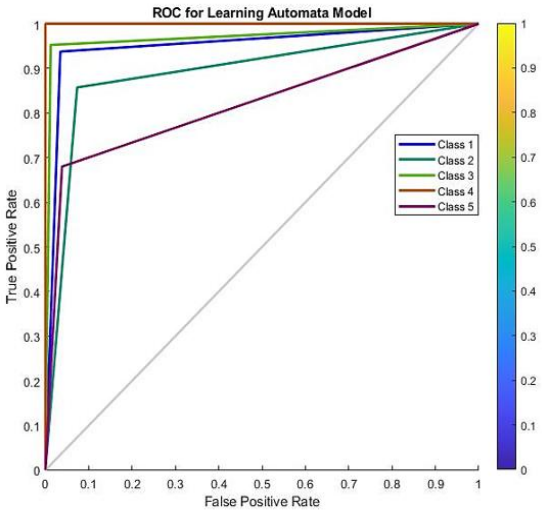


Fig. 28: ROC chart for K = 100, D = 100, and Nfemax = 15 parameters in LAA.

Fig. 29: Confusion matrix chart for Nfemax = 10, K = 10, and D = 10 parameters in LAA.

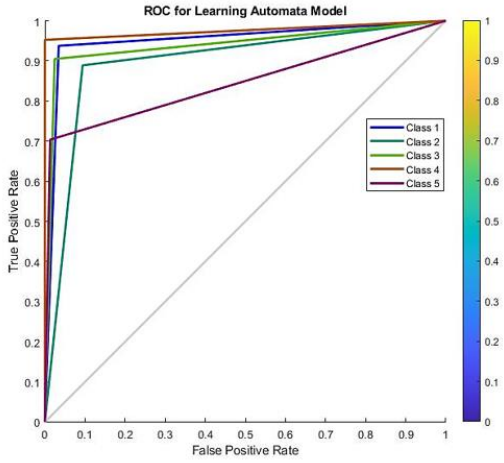


Fig. 30: ROC chart for Nfemax = 10, K = 10, and D = 10 parameters in LAA.

Table 5: LA performance results with the influence of K and D parameters

The results for Nfe=10 in LAA																											
Run Number	LAP		Best Solution				Precision (%)					Recall (%)					F1_Score					AUC					Accuracy (%)
	D	K	W1	W2	W3	W4	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	
1	10	10	1	1	2	2	83.33	66.66	90.47	100	95	93.75	88.88	90.47	95.23	70.37	88.23	76.19	90.47	97.56	80.85	0.95	0.89	0.94	0.97	0.84	86.4
2	50	100	3	5	1	1	88.88	75	80.95	100	75	88.88	72	100	95.23	68.18	88.88	73.46	89.47	97.56	71.42	0.93	0.82	0.97	0.97	0.81	83.5
3	100	200	3	2	4	4	88.88	66.66	95.23	100	85	94.11	80	95.23	100	68	91.42	72.72	95.23	100	75.55	0.95	0.85	0.97	1	0.82	86.4

Fig. 31: Confusion matrix chart for Nfemax = 10, K = 100, and D = 50 parameters in LAA.

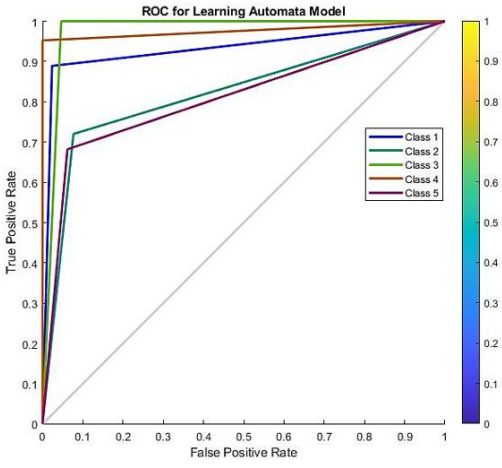


Fig. 32: ROC chart for Nfemax = 10, K = 100, and D = 50 parameters in LAA.

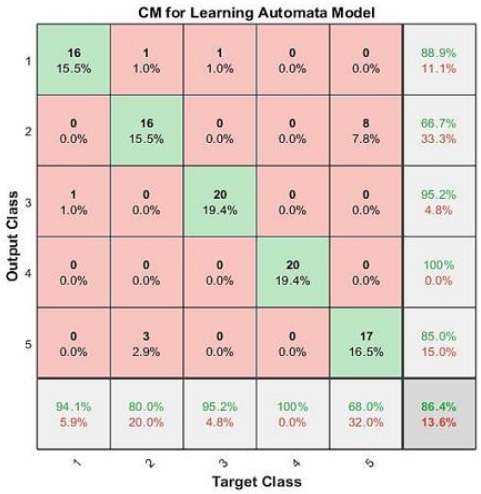


Fig. 33: Confusion matrix chart for Nfemax = 10, K = 200, and D = 100 parameters in LAA.

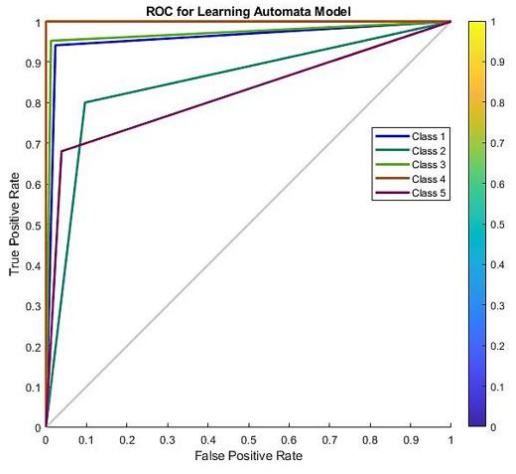


Fig. 34: ROC chart for Nfemax = 10, K = 200, and D = 100 parameters in LAA.

In addition, Table 6 shows the important parameters and computational requirements of introduced models such as SVM, RF, DT, XGboost, ensemble method, R-EFMD, T-EFMD, R-LFMD, T-LFMD, ANN, CNN, TIFR-

DCNN+SA, and joint [30-34]. The results of these models have been compared with the proposed model. Considering that the objectives and databases are different, we benchmarked the average detection rate.

In this comparison, Precision, Recall, F1_Score, and Accuracy parameters have been considered and investigated in order to show the superior performance of the proposed method with other methods. Also, in Fig. 35, the graph of this comparison is illustrated to show the results of each of the model's side by side, and the optimal performance of the data fusion method with the C algorithm is quite evident.

Table 6: Performance comparison of conventional and fused classification models

No.	Model	Precision (%)	Recall (%)	F1_Score (%)	Accuracy (%)
1	SVM	71.4	70	70	83.9
2	RF	70	77.78	73.68	76.19
3	DT	90	75	81.81	80.95
4	XGboost	80	80	80	80.95
5	Ensemble Method	60	75	66.67	71.45
6	R-EFMD	79.27	76.5	77.86	78.25
7	T-EFMD	79.51	81.5	80.49	80.25
8	R-LFMD	78.82	80	79.4	79.25
9	T-LFMD	83.17	86.5	84.8	84.5
10	ANN	63.71	64.58	64.14	65.57
11	CNN	78.47	79.39	78.92	65.57
12	TFIR-DCNN+SA	73.55	66.14	69.65	66.14
13	Joint	79.5	80.12	79.49	79.8
14	**Proposed Method**	**87.71**	**88.53**	**87.8**	**87.4**

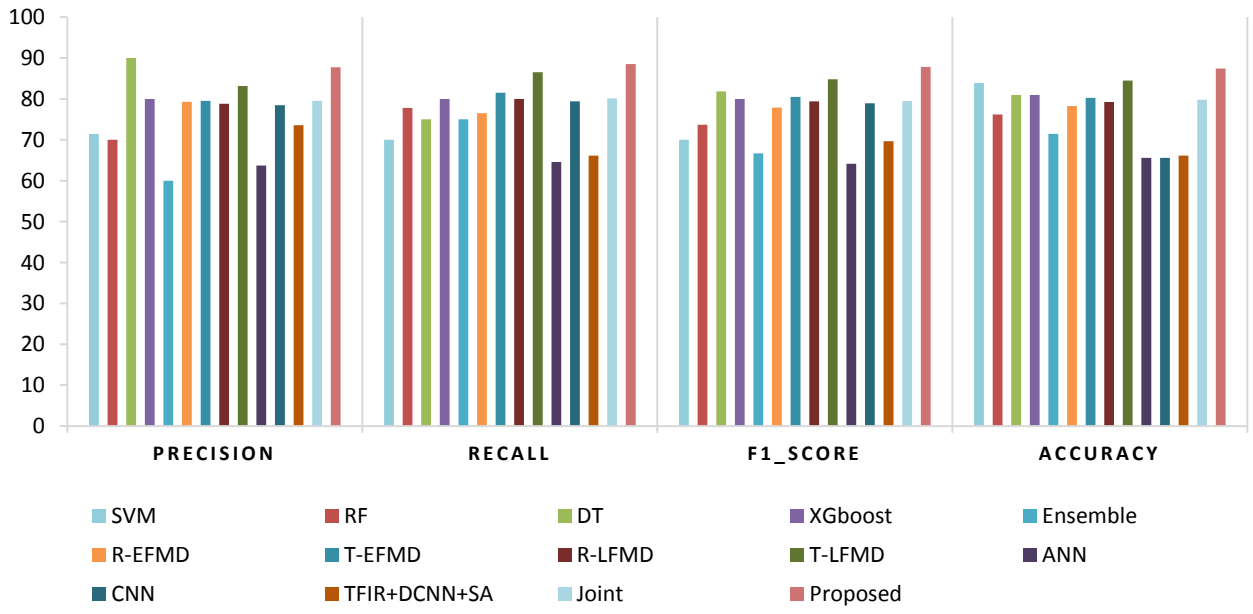


Fig. 35: Functional comparison of Precision, Recall, F1_Score, and Accuracy parameters.

Conclusion

In this article, the issue of combining classifications using LAA and sonar data is considered. The used sonar data, which includes 5 types of targets with different capabilities and specifications, were analyzed with the help of LAAs. The interference of sound waves and noise causes many problems in the detection of targets in marine environments. The classification of this data is often done in a traditional and manual way, and the possibility of target identification error is high in this

method. By using ML methods and combining them with each other, the accuracy of detecting these targets can be increased. In this research, we first used 4 classification models separately to classify sonar data. Then by combining those classifiers with learning automata algorithm to achieve the best solution and by determining the optimal coefficients for each classifier, we were able to achieve significant results compared to similar works. The results obtained with the analytical parameters of Precision, Recall, F1_Score, and Accuracy compared to the latest similar papers have been examined and compared, and the values are 87.71%, 88.53%, 87.8%, and 87.4% respectively for each of These parameters are obtained in the proposed method.

Some limitations that can be mentioned. The proper setting of learning automata parameters is the proper selection of basic classifiers and the existence of appropriate databases for training basic classifiers. In the future, it is possible to perform tasks such as fuzzing or optimizing the control parameters of learning automata for better convergence, using intelligent methods for the optimal selection of parameters, and using the proposed method in the face of incomplete and missing databases.

Author Contributions

Sajjad Mahmoudi Khah simulated the proposed method in MATLAB. Seyed Hamid Zahiri and Iman Behrvan supervised and consulted in the design, implementation and results of this research. All authors discussed important sections and contributed to the final text.

Acknowledgment

We sincerely thank the respected referees for their accurate review of this paper.

Also, we sincerely thank the ICT Research Institute and Connectivity and Communication Technologies Development Headquarters and Dr. Kharrat.

Conflict of Interest

The authors announce no potential conflict of interest regarding the publication of this paper. Also, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission and redundancy have been completely witnessed by the authors.

Abbreviations

SONAR	Sound and Range Navigation
DF	Data Fusion
RLA	Reinforcement Learning Algorithm
CNN	Convolutional Neural Network
ML	Machine Learning
SVM	Support Vector Machine
KNN	K-Nearest Neighbor
RF	Random Forest
DT	Decision Tree
XGboost	Extreme Gradient Boosting
R-EFMD	Early Fusion Model with Resnet
T-EFMD	Early Fusion Model with Swin Transformer
R-LFMD	Late Fusion Model with Resnet
T-LFMD	Late Fusion Model with Swin Transformer
SA	Spatial Attention
TFIR	Time Frequency Image Recognition
DCNN	Deep CNN
SLP	Single-Layer Perceptron
MLP	Multi-Layer Perceptron
LA	Learning Automata
LAA	Learning Automata Algorithm
LAP	Learning Automata Parameter
AUC	Area Under the ROC Curve
CM	Confusion Matrix
ROC	Receiver Operating Characteristic

References

[1] X. Fan, L. Lu, P. Shi, X. Zhang, "A novel sonar target detection and classification algorithm," Multimedia Tools Appl.s, 81(7): 10091-10106, 2022.

[2] J. U. ROBERT, Principles of underwater sound for engineers. MCGRAW-HILL., 1983.

[3] S. J. Davey, M. G. Rutten, B. Cheung, "A comparison of detection performance for several track-before-detect algorithms," EURASIP J. Adv. Signal Process., 2008(1): 1-10, 2007.

[4] M. Hassaballah, A. I. Awad, Deep learning in computer vision: principles and applications. CRC Press, 2020.

[5] D. P. Williams, "Underwater target classification in synthetic aperture sonar imagery using deep convolutional neural networks," in Proc. 2016 23rd International Conference on Pattern Recognition (ICPR): 2497-2502, 2016.

[6] M. Valdenegro-Toro, "Best practices in convolutional networks for forward-looking sonar image recognition," in OCEANS 2017-Aberdeen: 1-9, 2017.

[7] E. L. Ferguson, R. Ramakrishnan, S. B. Williams, C. T. Jin, "Convolutional neural networks for passive monitoring of a shallow water environment using a single sensor," in Proc. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 2657-2661, 2017.

[8] G. Huo, Z. Wu, J. Li, "Underwater object classification in sidescan sonar images using deep transfer learning and semisynthetic training data," IEEE Access, 8(1): 47407-47418, 2020.

[9] F. Mousavipour, M. R. Mosavi, "Sonar data classification using neural network trained by hybrid dragonfly and chimp optimization algorithms," Wireless Pers. Commun., 129(1): 191-208, 2023.

[10] Q. Li, L. Song, Y. Zhang, "Multiple extended target tracking by truncated JPDA in a clutter environment," IET Signal Process., 15(3): 207-219, 2021.

[11] E. Fan, W. Xie, J. Pei, K. Hu, X. Li, V. Podpečan, "Improved joint probabilistic data association (JPDA) filter using motion feature for multiple maneuvering targets in uncertain tracking situations," Information, 9(12): 322, 2018.

[12] T. Li, J. Prieto, H. Fan, J. M. Corchado, "A robust multi-sensor PHD filter based on multi-sensor measurement clustering," IEEE Commun. Lett., 22(10): 2064-2067, 2018.

[13] W. Xiong, X. Gu, Y. Cui, "Tracking and data association based on reinforcement learning," Electronics, 12(11): 2388, 2023.

[14] L. Snidaro, J. Garcia-Herrera, J. Llinas, E. Blasch, "Context-enhanced information fusion," Boosting Real-World Performance with Domain Knowledge, 2016.

[15] J. Raol, "Multi-Sensor Data Fusion with MATLAB. 2009," ed: CRC press.

[16] T. M. Mitchell, "Does machine learning really work?," AI magazine, 18(3): 11-11, 1997.

[17] P. Domingos, The master algorithm: How the quest for the ultimate learning machine will remake our world. Basic Books, 2015.

[18] A. K. Bathla, S. Bansal, M. Kumar, "OKC classifier: an efficient approach for classification of imbalanced dataset using hybrid methodology," Soft Comput., 26(21): 11497-11503, 2022.

[19] K. Taunk, S. De, S. Verma, A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," in Proc. 2019 International Conference on Intelligent Computing and Control Systems (ICCS): 1255-1260, 2019.

[20] D. Wettschereck, T. Dietterich, "Locally adaptive nearest neighbor algorithms," Adv. Neural Inf. Process. Syst., 6(1): 184-191, 1993.

[21] M. Sharma, S. K. Sharma, "Generalized K-nearest neighbour algorithm-a predicting tool," Int. J. Adv. Res. Comput. Sci. Software Eng., 3(11): 1-4, 2013.

[22] C. M. Bishop, Neural networks for pattern recognition. Oxford university press, 1995.

[23] H. Taud, J. Mas, "Multilayer perceptron (MLP)," in Geomatic approaches for modeling land change scenarios, Lecture Notes in Geoinformation and Cartography, Mexico City, 451-455, 2018.

[24] K. L. Du, M. N. Swamy, Neural networks and statistical learning. Springer Science & Business Media, 2013.

[25] M. A. Thathachar, P. S. Sastry, Networks of learning automata: Techniques for online stochastic optimization. Springer Science & Business Media, 2003.

[26] F. Hourfar, H. J. Bidgoly, B. Moshiri, K. Salahshoor, A. Elkamel, "A reinforcement learning approach for waterflooding optimization in petroleum reservoirs," Eng. Appl. Artif. Intell., 77(1): 98-116, 2019.

[27] M. L. v. TSetlin, "Automaton theory and modeling of biological systems," 1973.

[28] N. S. Shahraki, S. H. Zahiri, "DRLA: Dimensionality ranking in learning automata and its application on designing analog active filters," Knowledge-Based Syst., 219(1): 106886, 2021.

[29] Q. Wu, H. Liao, "Function optimisation by learning automata," Inf. Sci., 220(1): 379-398, 2013.

[30] B. N. K. Reddy, C. A. Vaithilingam, S. Kamalakkannan, "SONAR based under water mine detection using machine learning algorithms," in Proc. 4th International Conference on Innovative Practices in Technology and Management (ICIPTM): 1-4, 2024.

[31] Q. Wang, S. Du, W. Zhang, F. Wang, "Active sonar target recognition method based on multi-domain transformations and attention-based fusion network," IET Radar Sonar Navig., 2024.

[32] F. Ahmad, M. Z. Ansari, R. Anwar, B. Shahzad, A. Ikram, "Deep learning based classification of underwater acoustic signals," Procedia Comput. Sci., 235(1): 1115-1124, 2024.

[33] H. Yang, Y. Huang, Y. Liu, "Spatial attention deep convolution neural network for call recognition of marine mammal," in Proc. International Conference on Autonomous Unmanned Systems: 2725-2733, 2022.

[34] S. Z. Tian, D. B. Chen, Y. Fu, J. L. Zhou, "Joint learning model for underwater acoustic target recognition," Knowledge-Based Syst., 260(1): 110-119, 2023.

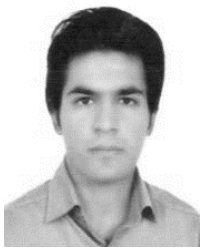
Biographies

Sajjad Mahmoudi Khah received the B.Sc. in Electrical Engineering, and Electronic from University of Sajjad in 2015 and M.Sc. in Electrical Engineering, and Electronic Integrated Circuits from University of Birjand in 2018. He is currently a doctoral student in Electrical Engineering majoring in Electronics at Birjand University, Birjand, Iran. His research interest includes soft computing, machine learning, circuit optimization, and swarm intelligence.

- Email: sajjadmahmoudikhah@birjand.ac.ir
- ORCID: 0009-0001-2184-2544
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Seyed Hamid Zahiri received the B.Sc., M.Sc., and Ph.D. degrees in Electronics Engineering respectively from Sharif University of Technology in 1993, Tarbiat Modarres University in 1995, and Ferdowsi University of Mashhad in 2005. Currently, he is Professor with the Department of Electronics Engineering, University of Birjand, Birjand, Iran. His research interests include pattern recognition, evolutionary algorithms, swarm intelligence algorithms, and soft computing.

- Email: hzahiri@birjand.ac.ir
- ORCID: 0000-0002-1280-8133
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://cv.birjand.ac.ir/zahiri/fa



Iman Behravan received the B.Sc. in Electronics Engineering from Shahid Bahonar University of Kerman, Kerman, Iran. Also, he received his M.Sc., Ph.D., and post-doctoral degrees from the University of Birjand, Birjand, Iran. His research interests include big data analytics, pattern recognition, machine learning, swarm intelligence, and soft computing.

- Email: i.behravan@birjand.ac.ir
- ORCID: 0000-0003-0319-1765
- Web of Science Researcher ID: NA
- Scopus Author ID: 5326-2017
- Homepage: https://scholar.google.com/citations?user=w9GKiVcAAAAJ&hl=en

How to cite this paper:
S. Mahmoodi Khah, S. H. Zahiri, I. Behravan, "Fusion of classifiers using learning automata algorithm," J. Electr. Comput. Eng. Innovations, 13(1): 65-80, 2025.

DOI: 10.22061/jecei.2024.10950.750

URL: https://jecei.sru.ac.ir/article_2192.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Fuzzification of Items of Media and Educational Materials and Tools

S. S. Musavian [1,*], ***A. Taghizade*** [1], ***F. Z. Ahmadi*** [2], ***S. Norouzi*** [1]

[1] *Department of Educational Sciences, Farhangian University, Tehran, Iran.*

[2] *Health and Physical Education. Dept., Organization for Educational Research and Planning, Tehran, Iran.*

Article Info

Article History:
Received 24 June 2024
Reviewed 14 July 2024
Revised 12 August 2024
Accepted 24 August 2024

Keywords:
AI assessment
Fuzzification assessment items
Fuzzy assessment
Educational media assessment
Big fuzzy rules set
Justice assessment

*Corresponding Author's Email Address: *s.musavian@cfu.ac.ir*

Abstract

Background and Objectives: The purpose of this study is to propose a solution for using large fuzzy sets in assessment tasks with a significant number of items, focusing on the assessment of media and educational tools. Ensuring fairness is crucial in evaluation tasks, especially when different evaluators assign different ratings to the same process or their ratings may even vary in different situations. Also, previous non-fuzzy assessment methods show that the mean value of assessors scores is not a good representation when the variance of scores is significant. Fuzzy evaluation methods can solve this problem by addressing the uncertainty in evaluation tasks. Although some studies have been conducted on fuzzy assessment, but their main focus is fuzzy calculations and no solution has been proposed for the problem arising when fuzzy rule set is considerably huge.

Methods: Fuzzy rules are the main key for fuzzy inference. This part of a fuzzy system often is generated by experts. In this study,15 experts were asked to create the set of fuzzy rules. Fuzzy rules relate inputs to outputs by descriptive linguistic expressions. Making these expressions is so more convenient than if we determine an exact relationship between inputs and outputs. The number of fussy rules has an exponential relationship with the number of inputs. Therefore, for a task with more than say 6 inputs, we should deal with a huge set of fuzzy rules. This paper presents a solution that enables the use of large fuzzy sets in fuzzy systems using a multi-stage hierarchical approach.

Results: Justice is always the most important issue in an assessment process. Due to its nature, a fuzzy calculation-based assessment provides an assessment in a just manner. Since many assessment tasks are often involved more than 10 items to be assessed, generating a fuzzy rule set is impossible. Results show the final score is very sensitive to slight differences in score of an item given by assessors. Besides that, assessors often are not able to consider all items simultaneously to assign a coefficient for the effect of each item on final score. This will be seriously a problem when the final score depends on many input items. In this study, we proposed a fuzzy analysis method to ensure equitable evaluation of educational media and instructional tools within the teaching process. Results of none-fuzzy scoring system show that final score has intense variations when assessment is down in different times and by different assessors. It is because of the manner that importance coefficients are calculated for each item of assessment. In fuzzy assessment no importance coefficient is used for each item.

Conclusion: In this study, a novel method was proposed to determine the score of an activity, a task, or a tool that is designed for learning purposes based on Fuzzy sets and their respective calculations. Because of the nature of fuzzy systems, approximate descriptive expressions are used to relate input items to final score instead of an exact function that is impossible to be estimated. Fuzzy method is a robust system that ensure us a fair assessment.

Introduction

Fair assessment is an important aspect in educational programs and fuzzy assessment is an approach to the fair assessment.

Simple statistical methods (such as averaging) are not so fair for an assessment task. It is because of three the following reasons:

- ✓ Often it is not so easy to assign an exact score or point to an item. In other words, when an expert is asked to set a point to an item at different times, there is a high probability that he/she will assign different points to the same item.
- ✓ An item of an activity may not be compared with other items when assigning a point to it. When assigning a point to an item, experts will focus exclusively on that specific item. Therefore, they may review the questionnaire and change the score of an item many times.
- ✓ Several experts will cause a diverse value for the point of a particular item.

In this study, we are going to deal with the concept and benefits of fuzzification of assessment items, with a focus on Media and Instructional Tools Assessment. Justice is an important component in an assessment task when different assessors may give different points to the same process. Even an assessor generally gives different points in different situations. The fuzzy assessment method will overcome this failure in assessment. Fuzzy calculations are an AI tool to deal with vague or approximate situations. It is always easier to assign an approximate value instead of the exact value to a variable.

Grade given by the teacher to a student can be optimized by using fuzzy logic [1].

Development of modern education, along with traditional learning, also requires using new assessment models (Glushkova et al, 2024) [2]. By utilizing assessments, researchers can acquire valuable data to recognize patterns, variances, and relationships within the dataset, ultimately enhancing knowledge and research development [3]. Penfield et al. (2016) [4], [5] and Andrade (2019) [6] highlighted the importance of assessment in fostering informed decision-making and augmenting research outcomes.

This method advances the interpretation and analysis within academic research by guaranteeing a more accurate depiction of subjective data and offering a flexible framework to accept varying degrees of imprecision or ambiguity.

Fuzzification, as a method, is crucial for capturing the complexity of real-world events that are inherently difficult to measure or categorize accurately (Markov et al. (2022)) [7].

The fuzzy calculations utilized here is based on Mamdani fuzzy inference [8]. Also Yunan and et al. [2020] [9] used this method of inference in their study. This is implemented by three main blocks: 1- fuzzification using membership functions, 2- aggregation fuzzy rules and 3- defuzzification using calculating the center of gravity of aggregated rules.

This work focuses on a huge fuzzy rule set that is a gap in similar previous works.

Assessing media and educational materials is so complicated and needs to be precisely down. This article discusses using fuzzification techniques to make the assessment better. Fuzzification helps us understand things that are not easy to put into categories. It works by giving different levels to words to help us grasp complexity. This method helps show data accurately and handles uncertainty well.

The integration of fuzzification techniques significantly enhances the efficacy of questionnaires in media and educational research. By adopting these methods, researchers can surpass the confines of conventional binary response formats, resulting in the collection of more refined and precise data. In media analysis, fuzzification empowers researchers to capture diverse subjective viewpoints and preferences through the inclusion of response choices with varying degrees of agreement. This approach recognizes the variability in individuals' levels of alignment or discord with a statement, leading to a more holistic comprehension of their perspectives. Likewise, within educational research, fuzzification strategies provide a deeper understanding of students' learning journeys. By offering a range of response levels that mirror varying degrees of understanding or skill, questionnaires become more accommodating to different learning preferences and competencies. The utilization of fuzzification techniques ensures that questionnaires transcend rigid binary responses, embracing the intricate nature of human experiences (Reigeluth, Honebein. 2023) [10], [11].

In a multi-item task assessment, a simple technique that quickly comes to mind is an averaging method in which we consider different coefficients for each item. In detail, we can design a questionnaire and ask some experts to determine a coefficient (say between 0 and 1) to assign to each item. In the end, we can consider the average of the given coefficients as the final coefficient for each item. Despite its simplicity, there are many serious problems with this manner due to an important concept we refer to it as "vagueness".

If some people are asked to estimate the weather temperature in degrees centigrade, they never state that the temperature is -7^{o}C, if really it is. But all of them likely say that it is "too cold". In a fuzzy assessment, assessors are asked to use approximate sentences and then the fuzzy system converts it to an exact value. One important issue that our study focuses on and was not the main subject in other similar studies is the number of Fuzzy rules. Fuzzy rules set is an essential part of Fuzzy calculations so if a Fuzzy system has incomplete rules set, results will not be reliable. Rules tell us how different items affect the final result of an assessment in an approximate verbal description. These approximate verbal descriptions are similar to these sentences: "For a

gymnast, if her/his errors are low and the time activities are finished is high, then her/his score is high", "If a student solves new math problems in a very short time, then she/he possesses a very high talent in math". A fuzzy calculation system works with these sentences. Fuzzy rules are these if-then formatted sentences.

The rules set are directly related to the items to be assessed. For example, in an assessment problem with 10 effective items in the final score, we will have a set with at least 310 different rules. Making a fuzzy system with a large rule set needs an algorithmic method to be involved.

This article explores how you can handle huge fuzzy sets in a fuzzy assessment problem.

There are some studies on fuzzy assessment tasks in which researchers pay particular attentions to different fuzzy calculations and systems. Many types of fuzzy calculations have been proposed for implementing a fuzzy assessment system. Mostly we have to perform an assessment task using more than 8 items affecting on it. This will lead to a large fuzzy set to be worked on.

A similar study has been published by the author for the assessment of Laboratory Courses in Electronics Engineering, in which the final score of a student is estimated using fuzzy calculations with three sub-activities as input parameters. These items have different contributions to the final score (Musavian, 2013) [12]. Therefore, a fuzzy rule set containing $2^3 = 8$ rules is used to calculate final scores.

Glushkova and her colleagues, (2024) [2] in their research presented a method for university teachers to evaluate their teaching performance using type-II fuzzy sets (T2 FSs). The evaluation indicator system is constructed from teaching attitude, teaching contents, teaching professionalism, teaching methods, and teaching effects. Therefore, this produces a fuzzy set including only $2^5 = 32$ fuzzy rules. Extracting a fuzzy set containing 32 rules cannot be considered a critical issue.

Also (Sheveleva et al, 2023) [13] used only 3 input variables, therefore, a total number of 8 rules were generated. This method of fuzzy calculations was sufficient only for student competency assessment and may not be developed to our task.

In Ryabko et al. (2022) [14], some graduated students were selected as samples to investigate different items affecting the quality of the education system. Therefore, rules are generated automatically. However, a weakness of this method is that these students were taught in that education system, therefore it is not a reliable criterion for assessment.

In Nurhidayah et al. (2022) [15] employees were assessed using only two variables: x_1 : the value of employee work goals. x_2: behavioural values. Therefore, making a set with 22=4 rules is a simple task. Considering only two items for assessment is not what will happen in practical cases.

Raheema (2022) [16] and Rojas (2021) [17] presented A fuzzy system for predicting student achievement throughout their education period.

Course evaluation is a critical part of undergraduate curriculum in computer science (Yan Liu 2022) [18]. In Yan Liu (2022) and et al. study only 4 fuzzy sets have been used for fuzzy inference. They used a Mamdani inference method to implement fuzzy calculations.

In Rahmanian (2021) [19], 12 items were considered for the task of assessment. One type of fuzzy calculation was performed without considering a rule set. In this manner, the fuzzy reasoning block has been omitted from the system, and the benefits of other blocks (such as the defuzzification block which is the last block of fuzzy systems to transform descriptive to quantitative values) are taken under consideration.

Also, in Sun et al. (2021) [20], similar work was performed for university teachers, a few sample rules were used (not all possible rules) merely for developing the calculations. They also used a traditional fuzzification and defuzzification blocks described in (Musavian, 2013) [21].

Antonio Cervero and et al. (2020) [22] analyzed student satisfaction with the use of virtual campuses in university teaching in order to discover the main variables influencing the overall online teaching-learning process that give quality to the virtual educational process, using a fuzzy inference system.

Higher education institutions are currently facing a competitive environment such as the increase in employers‟ demand and the challenges from Industry. Therefore, higher education institutions must ensure that students overcome the challenges in this competitive environment. In order to achieve this, student performance needs to be analyzed systematically by identifying the students‟ deficiencies and advantages. Petra (2021) [23] focused on the student performance analysis per year by using fuzzy logic evaluation methods.

In Alaa et al. (2019), [24] a total of 19 items were used to assess four English skills. Therefore, a rule set containing 219 rules was generated. Such a huge rule set is impossible to be implemented by a questionnaire.

In Namlı and Şenkal (2018) [25] only two input variables were used to estimate the final score. A maximum of $2^5 = 32\ possible$ rules can be generated with 5 input sets. These number of input variables is not so sufficient for a fair assessment and different competencies may be assessed as the same level. Here a defuzzification block can be omitted due to very small number of input variables.

In Yudono (2021) [26] was used only 3 items to do university student admission selection task, so only 8

fuzzy rules could be generated. These items were: Basic Competency, TOEFL Prediction and Interview. The last item is not so observable to assign an exact point to it. Therefore, a fuzzy method is the most suitable way to rate this item.

In Thakre and Chaudhari (2017) [27], Six effective factors for the assessment of teachers were considered with five input fuzzy sets. Therefore, a number of $5^6 = 15625$ fuzzy rules was possible to be gathered. Certainly, this number of rules is too high to be dealt with. Of course, the focus of this study was on fuzzy calculations and it was performed only using 50 (out of 15625 rules). Here some calculations similar to one that in (Musavian, 2013) are used for fuzzification and defuzzification blocks. Obviously considering only 50 rules instead of 15625 will not lead to accurate results.

In Voskoglou (2013) [28] only 3 items were presented for the assessment of students, S1: knowledge of a subject matter. S2: problem-solving related to S1. S3: the ability to adapt properly the already existing knowledge for use in analogous similar cases. Therefore, the rules set comprises up to 23=8 rules. Many student assessment procedures use more than three items to assess students in a fair manner.

In Montero et al. (2005) [29], a final score was estimated based on 5 different activities of students. Due to this number of input sets, $5^5 = 3125$ possible rules would be generated but only 6 rules were used to develop calculations. Considering only 6 items instead of 3125 rules means that the purpose of this study is developing fuzzy calculations.

In this paper we will discuss and present a hierarchal method to utilize all possible rules (in a big rule set) in a fuzzy rule base system. Most similar works with more than 5 items of assessment focused on calculations instead of dealing with a big rule set. They ignored many rules and developed their fuzzy based calculations using a very limited number of possible rules.

Traditional fuzzy calculations may vary in terms of fuzzy membership functions. many studies are focused on the effect of membership functions on final results. We showed that the type of membership function is not so critical for the task of assessment, since these membership functions are identical for all assesses.

Technical Work Preparation

Zadeh's study [30] (1965, as cited in Adeyanju et al. 2021 [31]) proposed Fuzzy sets, the core of Fuzzy logic systems in 1965. Fuzzy logic solves problems that are not handled by well-known logic systems that is crisp (either 0 or 1) logic. Judging is always unfair because of lacking in our knowledge about the universe. So, one cannot describe a complicated system in detail. For example, we aren't able to express the temperature exactly in degrees Celsius if we do have not a temperature sensor. But rather we can describe the temperature by some linguistic words, such as "the weather is very cold", "is rather cold", "is not so cold" and these kinds of expressions. Using these vague expressions, you can help someone to choose a suitable cloth on a winter day. Fuzzy sets are the key factor in understanding fuzzy systems.

- **Fuzzy Sets**

Sometimes it is not so simple to classify objects based on some of their scalar features. Suppose that students of a school are to be classified into three classes: Excellent, Good, and Poor students using the following function:

$$Student\ a\ belongs\ to \begin{cases} Excellent & if\ M > 17 \\ Good & if\ 14 < M < 17 \\ Poor & if\ M < 14 \end{cases}$$

in which M is the mean score of each student. Since each student only belongs to one set, these are referred to as crisp sets. Consider two students with mean scores equal to 13.9 and 14.1. The former is classified as Poor students and the second as Good students. However, these two students are not very differently talented. But they will be laid in completely different sets using a crisp classification method.

If we apply the crisp classification method to the task of choosing a suitable cloth in cold weather, probably we have to wear another cloth when the temperature is -1oC compared to when the temperature is +1oC.

Fuzzy methods fix the mentioned problem in decision-making tasks. In a fuzzy system, it is supposed that an object belongs to all sets. That is a student belongs to the three above sets, regardless of his/her scores. The main question is: "How we can design a system so that although we consider an element belongs to all sets, the system still can properly work?". The solution is so simple, an element is a member of all sets, with a different membership degree. Membership degree is an important parameter in fuzzy sets and has a value in the range of [0…1]. For example, a student with a mean score equal to 10 is a member of the set Excellent, but with a membership degree too close to 0, but another student with a score of 20 belongs to that set with a membership of 1. Membership degrees are determined using some mathematical functions called membership functions. Fig. 1 shows some typical membership functions for a fuzzy system with three sets.

In Fig. 1 we can see that a student whose mean score is 16, belongs to the sets Excellent, Good, and Poor with the value of membership degrees equal to 0.3, 0.3, and 0 respectively.

The membership functions of Fig. 1 are linear, but non-linear functions are also used in engineering problems such as pattern recognition. Linear functions are sufficient in problems such as the task in this study. One important aspect is the overlap of functions. The amount

that functions overlap with each other, can be suggested by experts in the field.

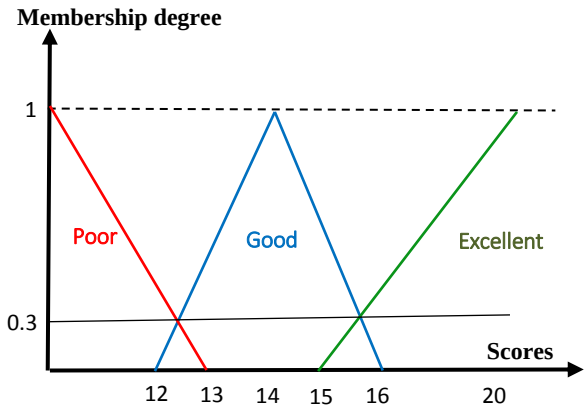


Fig. 1: a typical membership function for the student's classification problem.

There are some other mathematical membership functions such as sigmoidal functions. Results in (Musavian, 2013 [12]) shows that triangular functions shown in figures 1 and 2, not only are simpler in implementation but generate more reasonable results.

- **Fuzzy Rules**

In the two previous circumstances, only one category of sets was used. In practical problems, two categories of sets are demanded, that is input and output fuzzy sets. We need these two categories of sets to produce fuzzy rules.

We can write the task of problem-solving in mathematical language as $y = f(x_1, x_2, \dots)$ in which the input x_i becomes the output y by the system f. f cannot be made so easy by classical calculations such as statistical methods. f can be considered as a mathematical relationship, but what kind of math functions (such as sinusoidal, exponential, polynomial, etc. can be supposed to model our system? Therefore, describing an exact relation between input and output is rather impossible. Assume you are driving on a road at a speed of $70\ km/hrs$.

If an obstacle is seen at a distance of 20m, you should apply a force equal to F on the brake pedal for time T. You need to know the physical formula necessary for calculating F and T. Besides you have to know also the amount of friction coefficient between tires and the pavement.

Do you drive in such a manner? Of course not! You just know some rules from your experiences from driving: "If the speed is high AND the obstacle is so near, then push down on the pedal harder". Or like this: "If the speed is low AND the obstacle is far, then push down on the pedal gently ".

These are some uncertain but applicable rules, so a driver can control the car using these rules.

A fuzzy rule is a conditional statement that describes a decision-making guide but in a very approximate manner. It has an IF-THEN structure:

if x is A ***then*** y is B.

in which the antecedent part is "x is A" and the statement "y is B" is known as the consequent. Often Fuzzy rules have multi-part antecedents. A Fuzzy rule with a multi-part antecedent has a form as follows:

if x1 is A1 AND/OR x2 is A2 AND/OR … AND/OR xn is An ***then*** y is Bj.

For example, in the choosing warm clothes problem we have:

if(*rather cold* AND *high wind*) OR *too cold* ***then*** *thick clothes*.

In the above rule we can recognize that when it is too cold, regardless of the speed of wind, we choose a thick cloth. Therefore, this rule can be broken into two rules as follows:

if (*rather cold* AND *high wind*) ***then*** *thick clothes*.

if *too cold* ***then*** *thick clothes*.

For computational reasons, we prefer to use multiple rules rather than one rule with multi-part antecedents.

x and y are known as input and output variables. In our task, x represents factors affecting scoring and y represents the final score. In the problem of choosing warm clothes, weather temperature and speed of wind are two factors that affect choosing suitable clothes. Therefore, some rules are expected to have multi-part antecedents. As it will be mentioned in the next section, size of a fuzzy rule set depends on input variables. The thing that will cause a rule set to be too large to be handled by fuzzy calculation, is the number of items should be assessed in a justly assessment task. Fuzzy rules are developed by experts in a fuzzy system. In an assessment task, experts are who know how much an item contributes in final score. As mentioned in section 2.6, experts were involved to generate fuzzy set in two steps.

- **Input and Output Fuzzy Sets**

Consider again two previous problems mentioned before. Each one of these two problems will describe the concept of input and output sets separately so that one describes the input and the other describes the output sets.

In the problem in which students are to be classified into three classes, the three sets of Excellent, Good, and Poor students are output fuzzy sets, as the intention is to put a student in one of the three mentioned sets. For the second problem we can define the following input sets:

Too cold, rather cold, temperate (as the first variable).

High, fairly high, gentle, not windy.

Although two example problems only have one category of sets, in practical situations, as at the present work, we have to define both categories of sets.

- **Number of Rules-Based upon Input Variables**

The number of rules in a Fuzzy system depends on the number of input variables. Effective factors in scoring are indeed input variables. As a general formula, we can write:

$$N_r = N_v^{N_s}$$

In which:

N_r: Number of Fuzzy rules

N_v: Number of input variables

N_s: Number of input Fuzzy sets

For example, in an assessment problem with 10 factors determinative of the final score, if we define 3 input sets (High, Medium, Low), then we will describe 310 different rules. This number of rules is too much to be generated by experts.

As depicted in table 2, there are 11 items to be considered for our assessment task. This will introduce more than 2000 rules and impossible to be generated by experts. There is no algorithm to sample a finite number of rules among such a huge rule set, therefore, an algorithmic method is needed to be implemented for considering the effect of all rules. In this study, a novel layering method is proposed to consider the effect of all rules.

- **Generating Fuzzy Rules in our Task**

Input and output fuzzy sets should be determined to generate fuzzy rules. As experts in the field confirmed, for an efficient assessment system, it will be sufficient for the number of input sets to be equal to three that is High, Medium, and Low. Fig. 2(a) shows membership functions for input and Fig. 2(b) for output fuzzy sets. On the other hand, more input sets, are closer to a crisp system rather than a fuzzy one. If even we consider the number of input sets to be 5, it will be a very difficult task for experts to define rules.

We consider 5 output fuzzy sets: Excellent, Very good, Good, Medium, and Poor. This number of output sets is not too large to cause confusion to experts who are asked for rule generating. Also, it is not too small to cause an injustice to be down in assessment.

Since the output of a fuzzy system is limited to a small number of sets, it will be very likely that some conditions of input sets, lead to the same result of assessment. In assessment tasks always, there are some conditions where one or more factors dominate other factors. When generating Fuzzy rules, if these dominant factors are evaluated as Excellent, then it will be very likely that we have no choice but to assign the output to the set Excellent.

Suppose the variable x_1 is a dominant variable among the other two, that is x_2 and x_3. If so, the rule:

if x_1 is *Low* ***then*** y is *Poor*.

Represents 9 different rules. In other words, this single rule reduces 27 rules to 19 rules. We proposed a grouping method to convert a large group of rules to a small one so that it can be an appropriate representative for the larger group.

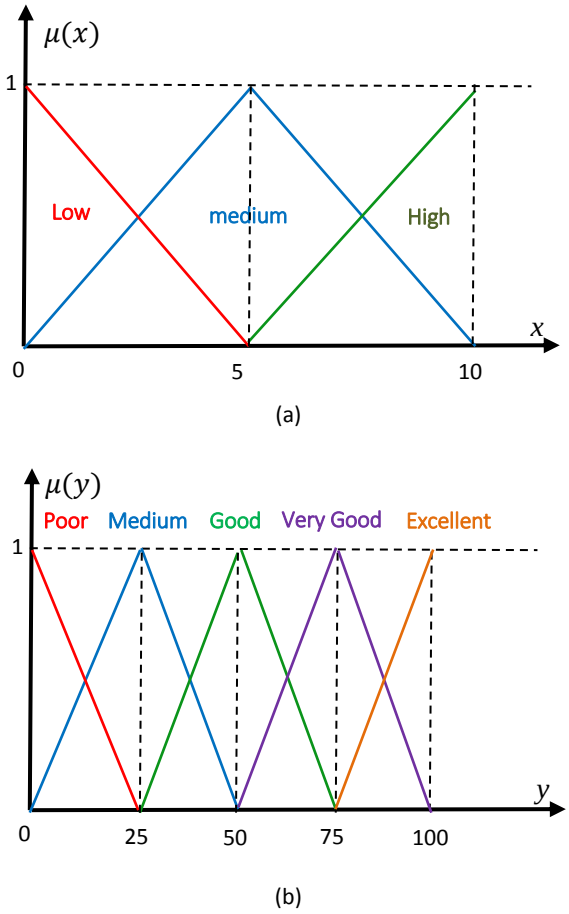


Fig. 2: (a): membership functions of input fuzzy sets used in our task of assessment, (b): membership functions of output fuzzy sets used in our task of assessment.

- **Grouping items**

Grouping items is the first step to reducing input variables. Homogeneous items are brought into one group.

Homogeneity may be defined as having the same effect on the final score.

In other words, all items with a high effect may be categorized in the same group. Fig. 3 shows an example of grouping 10 different items into three groups. All items in a specific group have the same effect on the final score. Now we change a problem including 10 items to a problem with 3 new items.

In the following, we suppose three groups are represented by sub1, sub2, and sub3.

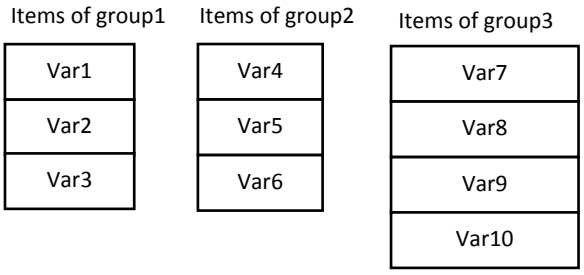


Fig. 3: 10 items are categorized into 3 groups.

The first step is to generate some rules assuming each group is considered as an item. At this step (first level of rule generation procedure), experts are asked to establish rules that relate input sets to output sets with sub1, sub2, and sub3 as input items. So now we have $3^3 = 27$ instead of $3^{10} = 59049$ rules. Now suppose experts confirm that if groups 1 and 3 are evaluated as Low, then a final fuzzy score is evaluated as Poor regardless of what group 3 may be evaluated.

We refer to these two groups as definitive items/groups. Therefore, the following single rule is representative of 27 rules:

if sub1 is *Low* AND sub3 is Low ***then*** output is *Poor*.

Also suppose groups 1 and 3 are again definitive groups if they consist of items with high effects on the final score, so we can similarly write the following rule:

if sub1 is *High* AND sub3 is *High* ***then*** output is *Excellent*.

Table 1: Items for evaluation of instructional media and tools (Assessment forms of the teaching festival at Farhangian University, 2022)

Items	
Introducing the instructional media using a certificate of authenticity (COA)	1
Ease of use, simplicity, and accessibility of instructional media and tools	2
Using creativity in the development of instructional media and tools as well as paying attention to their attractiveness	3
Appropriateness of instructional media and tools with characteristics of learners	4
Consistency, coherence, and coordination across all elements of instructional media and tools	5
Alignment of instructional media and tools with learning objectives, content, and teaching methods	6
Improving the quality of learning in different domains	7
Taking into consideration the different uses of instructional media and tools	8
Interactive instructional media and tools	9
The extent of using IT and ICT in teaching as well as introducing websites that are useful and related to the subject matter	10
Cost-effectiveness of instructional media and tools	11

It should be noted that the two above rules do not imply that the final score depends on only groups 1 and 3, even when sub1 and sub2 are both High (or both are Low).

All rules will be taken into account when mapping input sets onto output sets in a Fuzzy system.

This is why Fuzzy logic is an efficient tool for dealing with uncertain descriptions in the format of if-then statements.

An example of a rule with a three-part antecedent is as follows:

if sub1 is *Med* AND sub3 is High AND sub2 is *High* ***then*** output is *Very good*.

For the reasons mentioned above, input and output sets are assumed to be:

Output sets: Excellent, Very good, Good, Fair, Poor

Input sets: High, Med, Low

In this research, we employed the described analyses to assess an educational tool. These sets will serve as the basis for developing an assessment fuzzy system to assess instructional tools.

- **Assessment of Instructional Media**

Table 1 shows the necessary items for evaluating instructional media and tools approved by Farhangian University of Iran (teachers training center). These items are the same as input variables.

In this study, we proposed a fuzzy analysis method to ensure equitable evaluation of educational media and instructional tools within the teaching process.

Supposing three input sets (Low, Medium, and High) and according to the mathematical permutation formula, there are $3^{11} \approx 177000$ possible rules.

This extra-large group of rules makes it impossible to generate an applicable bank of fuzzy rules. Furthermore, there is no method to choose some rules among these very many rules to be an efficient representative of all possible rules.

Fifteen experts were inquired about how to categorize these eleven items into three groups (see Fig. 4).

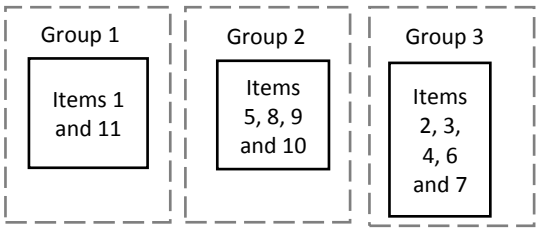


Fig. 4: Grouping of 11 items for assessing instructional media and tools.

- **Between Groups Rules**

Between groups, rules are limited to $3^3 = 27$rules (3 variables and 3 input sets). From Figure (4) we conclude

that groups 2 and 3 are definitive, so we can consider the following rules:

R1: ***if*** sub2 is *Low* AND sub3 is *Low* ***then*** output is *Poor*.

R2: ***if*** sub2 is *High* AND sub3 is *High* ***then*** output is *Excellent*.

Between-group rules are summarized in the table shown in Table 2.

- **Within Groups Rules**

Now we focus on each group and assign a descriptive value to each item inside a group. Here each item is compared with other items inside the same group. Since the number of items within each group is much less than all items, it will be much more convenient to construct rules for a sub-system having 3 or 4 items instead of 10 (compare $3^4 = 81$ with $3^{10} = 59049$).

The next step is to determine descriptive values for items within each group. Table 3 shows these values which again are obtained from experts by a questionnaire.

Tables 4(a) to 4(c) show within-group rules for groups G1, G2 and G3.

Table 2: Between groups rules

Rule Number	G_1	G_2	G_3	Fuzzy Output
R_2	×	H	H	*Excellent*
R_3	×	M	H	*Very good*
R_4	×	L	H	*Good*
R_5	H	H	M	*Very good*
R_6	L/M	H	M	*Good*
R_7	L/M	M	M	*Fair*
R_8	H	M	M	*Good*
R_9	H	H	L	*Good*
R_{10}	×	M	L	*Fair*
R_1	×	L	L	*Poor*

× stands for don't care states

Table 3: descriptive values of within-group items

G_1		G_2		G_3	
item	***Fuzzy value***	***item***	***Fuzzy value***	***item***	***Fuzzy value***
1	L	5	M	2	L
11	H	8	M	3	M
		9	H	4	H
		10	M	6	H
				7	M

- **Fuzzy inference**

Fuzzy inference is the process of relating inputs to outputs in a fuzzy manner. This block lies between fuzzification and defuzzification blocks. Fuzzification is the process of converting numerical to fuzzy inputs using fuzzy membership functions. Defuzzification is the process of converting fuzzy to numerical outputs. Center gravity method was used for the defuzzification block. Since enough straightforward to implementation, the inference method used by Musavian, (2013) [12] is used here again.

Table 4(a): within-group rules for G1

I1	**I11**	**O1**
MH	H	*Excellent*
L	H	*Very good*
H	M	*Very good*
LM	M	*Good*
H	L	*Fair*
ML	L	*Poor*

Table 4(b): within-group rules for G2

I5 I8 I10	**I9**	**O2**
all MH	H	*Excellent*
1L / 2MH	H	*Very good*
two L / one MH	H	*Good*
all L	H	*Good*
all H	M	*Very good*
all M	M	*Good*
one L / one M / one H	M	*Good*
two L / one MH	M	*Fair*
all H	L	*Good*
one L / two H	L	*Fair*
all M	L	*Fair*
one L / two LM	L	*Poor*

If scores for each item is represented by x_i, then the aggregation value for the antecedent of each rule is given by:

$$A_{antecedent} = \min(\mu(x_i)).$$

in which $\mu(x_i)$ is the value of membership function for the input x_i. For example, for the rule as bellow:

if x_1 is low and x_2 is high then …

and $(x_1,x_2)=(5,7.5)$, supposing that membership functions are as the figure 2, then:

$$A_{antecedent} = A(y) = \min\left(\mu_{low}(5), \mu_{high}(7.5)\right).$$
$$= min(0,0.5) = 0.$$

The next step of fuzzy inference is the aggregation of input and output fuzzy sets for each rule. This is accomplished by the min function again:

$$A_{ante,conse} = \min\left(A_{antecedent}, \mu(y)\right).$$

in which y is the output variable.

Defuzzification is the last step of a fuzzy system. It is the process of converting fuzzy to scalar output. Center gravity method is used for the defuzzification task:

$$scalar = \frac{\sum yA(y)}{\sum A(y)}.$$

For an assessment task with a huge fuzzy rule set, inference process is to be performed in 2 steps: the first step is to infer using within-group rules and then to infer using between-group rules. Fig. (5) shows a block diagram of these two steps. The total fuzzy inference system consists of two fuzzy inference sub-systems. The first block is the input block and it receives 11 inputs. Each input is a score given to the corresponding item by evaluators. The first block generates three outputs using rules within three groups. We can refer to these inputs as scores of groups. Then these middle scores are considered as the inputs for the output stage. The output stage uses between-groups rules to estimate the final fuzzy score.

Table 4(c): within-group rules for G3

I2	I3	I7	I4	I6	O3
×	H	H	H	H	*Excellent*
H	1M	1H	H	H	*Excellent*
LM	1M	1H	H	H	*Very good*
H	M	M	H	H	*Very good*
LM	M	M	H	H	*Good*
×	1L	1M	H	H	*Good*
×	H	H	1M	1H	*Very good*
×	1M	1H	1M	1H	*Good*
H	H	H	M	M	*Very good*
LM	H	H	M	M	*Good*
H	H	H	1L	1H	*Very good*
LM	H	H	1L	1H	*Good*
×	1M	1H	1L	1MH	*Good*
×	M	M	L	L	*Fair*
H	1L	1H	L	L	*Fair*
ML	L	L	L	L	*Poor*

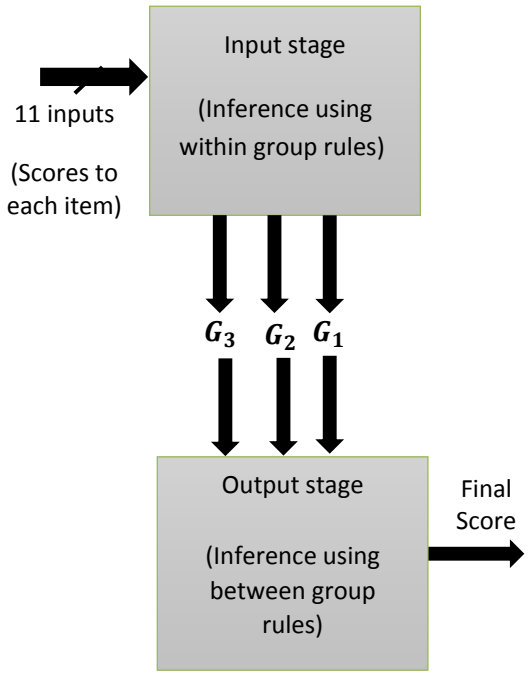


Fig. 5: inference is performed in two stages.

Results and Discussion

As a benchmark for the performance of entire the assessment system, a comparison is made between fuzzy and traditional calculations focusing only on group 1 with two items out of 11 items of assessment. Results are shown in table 6. The same 15 experts were asked to assign a coefficient to each item twice in two months interval between. Each time they were filling up a questionnaire, they assigned different coefficient to items. But when they asked to determine the linguistic descriptive relations between inputs and output (that is the fuzzy rules), the difference was so negligible. Table 6 shows assigned coefficient determined by experts in two inquiries. Also mean and variance values of coefficients for items 1 and 11 (G1) are shown.

Table 5: Mean and variance for coefficients for items 1 and 11 (G1)

Frequencies (first/second time of inquiry)	Item 1	Item 2
1/2	0.5	0.5
3/5	0.4	0.6
4/2	0.3	0.7
1/0	0.6	0.4
5/6	0.2	0.8
1/0	0.1	0.9
Mean/Var.	0.3/0.14	0.7/0.14

Comparing first and second rows in table 7, we will figure it out that a slight change in score of items due to non-uniform assessing, will slightly change the final score. See rows 2 and 3 where only item 1 has been changed. Item1 has a small effect on the final score than item11, therefore we expect a slight change on it, not to change the final score as n.f method did.

Table 6: Final scores using fuzzy and non-fuzzy methods

[item1, item2]	n.f method	Fuzzy method
[5, 5]	5	4.9
[4.7, 4.8]	4.73	4.85
[4.8, 4.8]	4.8	4.85

n.f: non-fuzzy

Conclusion

In this study, a novel method was proposed to determine the score of an activity, a task, or a tool that is designed for learning purposes based on Fuzzy sets and their respective calculations. Fuzzy assessments result in avoidance of the variety of manner of grading by different evaluators. Many studies have been conducted on fuzzy assessment tasks. Usually, an assessment will be performed based on some items and an evaluator gives scores to each of them. Final scores are calculated by aggregating these scores. A serious problem with this scoring mechanism is that giving an exact score (as a number say in the range of 0 and 10) cannot be performed in a certain manner. Therefore, evaluations always depend on evaluators and even the time that an evaluation task is being performed. Fuzzy evaluation makes it possible for an evaluator to choose one option from three or five options to give a (fuzzy) score to an item instead of giving exact numerical scores.

In a fuzzy scoring system, a serious problem will arise if there are many items to be evaluated. This leads to a very large number of fuzzy rules. The method proposed in this study fixes the problem arising from the large number of fuzzy rules.

As mentioned before, we are fixing a serious problem in fuzzy assessment tasks with large number of rules. In previous similar studies, this problem was not the main issue and main efforts was to express the benefits of fuzzy calculations in an assessment task. In our study when looking at final assessment results, and especially when we are dealing with diversity of scores, we can observe that very small variance values are presented by fuzzy method assessment. This means that an assessment by different assessors tends toward a unique score.

A limitation for the proposed method to be implemented is that items of assessment should have the property of being put in a limited number of groups. The smaller number of groups, the easier dealing with the huge rule set.

All fuzzy assessment tasks with large rule set carried out before with limited number of possible rules, are strongly recommended to apply the proposed grouping method for the whole their rule set.

Since the process of dividing rules into groups is the main key to our proposed method, a future study is to apply neuro network pattern recognizers for finding the best solution for rules grouping. This will be a very important step for fuzzy assessment tasks with many items affecting final score.

Author Contributions

The main Idea was started by S.S Musavian. She designed main blocks of fuzzy calculations and wrote the manuscript. F. Ahmadi and S. Norouzi designed experts questionnaires, gathered and prepared data for analyses. A. Taghizadeh cooperated in analyzing results.

Acknowledgment

The author gratefully acknowledges Mr. Omid Ahmadi for the implementation of fuzzy calculations in Python.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

References

[1] T. Khomeiny, T. R. Kusuma, A. N. Handayani, A. P. Wibawa, A. H. S. Irianti, "Grading system recommendations for students using fuzzy mamdani logic," in Proc. 2020 4th international conference on vocational education and training (icovet): 1-6, 2020.

[2] T. Glushkova, V. Ivanova, B. Zlatanov, "Beyond traditional assessment: A fuzzy logic-infused hybrid approach to equitable proficiency evaluation via online practice tests," Mathematics, 12(3): 371, 2024.

[3] OECD, "Evaluation and assessment frameworks for improving school outcomes: Common policy challenges summary," 2010.

[4] R. D. Penfield, "Fairness in test scoring," in Fairness in Educational Assessment and Measurement, 1st Ed., pp: 55-76, 2016.

[5] T. Penfield, M. Baker, R. Scoble, M. Wykes, "Assessment, evaluations, and definitions of research impact: A review," Res. Eval., 23(1): 21-32, 2013.

[6] H. L. Andrade, "A critical review of research on student self-assessment," Front. Educ., 4, 2019.

[7] A. Markov, Z. Seleznyova, V. Lapshin, "Credit scoring methods: Latest trends and points to consider," J. Finance Data Sci., 8: 180-201, 2022.

[8] E. H. Mamdani, "Application of fuzzy logic to approximate reasoning using linguistic synthesis," IEEE Trans. Comput., 26(12): 1182-1191, 1977.

[9] A. Yunan, M. Ali, "Study and implementation of the fuzzy Mamdani and Sugeno methods in decision making on selection of

outstanding students at the South Aceh polytechnic," Inotera, 5(2): 152-164, 2020.

[10] C. M. Reigeluth, P. C. Honebein, "Will instructional methods and media ever live in unconfounded harmony? Generating useful media research via the instructional theory framework," Educ. Technol. Res. Dev., 1-21, 2023.

[11] Z. Zhou, Y. Li, "Teaching quality evaluation of higher education based on intuitionistic fuzzy information," in Proc. 2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE): 1-7, 2023.

[12] S. S. Musavian, "Assessment of laboratory courses using fuzzy reasoning," Journal of Educational Sciences & Psychology, 3(2): 19, 2013.

[13] O. Sheveleva, V. Dobrynin, M. Mastroianni, Y. Goncharova, "A fuzzy model for assessing acquired competencies," in Proc. E3S Web of Conferences, 419: 02024, 2023.

[14] A. V. Ryabko, O. V. Zaika, R. P. Kukharchuk, T. A. Vakaliuk, V. V. Osadchyi, "Methods for predicting the assessment of the quality of educational programs and educational activities using a neuro-fuzzy approach," in CTE Workshop Proc., 9: 154-169, 2022.

[15] M. N. Raheema, A. M. Al-Khazzar, J. S. Hussain, "Prediction of students' achievements in e-learning courses based on adaptive neuro-fuzzy inference system," Int. J. Fuzzy Log. Intell. Syst., 22(2): 213-222, 2022.

[16] M. N. Raheema, A. M. Al-Khazzar, J. S. Hussain, "Prediction of students' achievements in e-learning courses based on adaptive neuro-fuzzy inference system," Int. J. Fuzzy Log. Intell. Syst., 22(2), 213-222, 2022.

[17] J. A. Rojas, H. E. Espitia, L. A. Bejarano, "Design and optimization of a fuzzy logic system for academic performance prediction," Symmetry, 13(1): 133, 2021.

[18] Y. Liu, X. Zhang, "Evaluating the undergraduate course based on a fuzzy AHP-FIS model," Int. J. Mod. Educ. Comput. Sci., 12(6), 55-66, 2020.

[19] M. Rahmanian, "A new fuzzy approach for teacher's performance evaluation," Int. Res. J. Mod. Eng. Technol. Sci., 3(01): 252-258, 2021.

[20] X. Sun, C. Cai, P. Su, N. Bao, N. Liu, "A university teachers' teaching performance evaluation method based on type-II fuzzy sets," Mathematics, 9(17): 2126, 2021.

[21] S. S. Musavian, "Assessment of laboratory courses using fuzzy reasoning," J. Educ. Sci., 12(6), 2013.

[22] A. Cervero, A. Castro-Lopez, L. Álvarez-Blanco, M. Esteban, A. Bernardo, "Evaluation of educational quality performance on virtual campuses using fuzzy inference systems," PLoS One, 15(5): e0232802, 2020.

[23] T. Z. H. T. Petra, M. J. A. Aziz, "Analysing student performance in higher education using fuzzy logic evaluation," Int. J. Sci. Technol. Res., 10(1): 322-327, 2021.

[24] M. Alaa et al., "Assessment and ranking framework for the English skills of pre-service teachers based on fuzzy Delphi and TOPSIS methods," IEEE Access, 7: 126201-126223, 2019.

[25] N. A. Namlı, O. Şenkal, "Using the fuzzy logic in assessing the programming performance of students," Int. J. Assess. Tools Educ., 5(4): 701-712, 2018.

[26] M. A. S. Yudono et al., "Fuzzy decision support system for ABC university student admission selection," in Proc. Int. Conf. Econ., Manag. Account. (ICEMAC 2021): 230-237, 2022.

[27] T. A. Thakreh, O. K. Chaudhari, N. Dhawade, "A fuzzy logic multi-criteria approach for evaluation of teachers' performance," Adv. Fuzzy Math., 12(1): 129-145, 2017.

[28] M. G. Voskoglou, "Fuzzy logic as a tool for assessing students' knowledge and skills," Educ. Sci., 3(2): 208-221, 2013.

[29] J. A. Montero, R. M. Alsina, J. A. Morán, M. Cid, "Fuzzy logic system for students' evaluation," in Proc. Lecture Notes in Computer Science: 1246-1253, 2005.

[30] L. A. Zadeh, "Fuzzy sets," Inf. Control, 8(3): 338-353, 1965.

[31] I. Adeyanju, O. O. Bello, M. A. Adegboye, "Machine learning methods for sign language recognition: A critical review and analysis," Intell. Syst. Appl., 12: 200056, 2021.

Biographies

Samaneh Sadat Musavian is an Assistant Professor at the Farhangian University, Tehran, Iran. She began her undergraduate studies in Educational Technology in 2006 at Allameh Tabataba'i University, She graduated at the top of her class in her undergraduate program. She completed her Master's degree in 2013 from Allameh Tabataba'i University and immediately started her doctoral studies in Educational Technology at Tarbiat Modares University. In 2018, she graduated with a Ph.D. degree. She has presented some scientific papers in scientific journals and conferences, and her research interest lies in qualitative research at the intersection of educational sciences and psychology. She has also gained valuable experience in the field of fuzzy analysis.

- Email: s.musavian@cfu.ac.ir
- ORCID: 0009-0008-4328-6316
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Abbas Taghizade is an Assistant Professor at the Farhangian University, Tehran, Iran. He received his Bachelor's degree in Educational Sciences in 2007 from the University of Kashan. He then ranked second in the 2008 Master's Entrance Exam in Educational Technology and entered Allameh Tabataba'i University, where he received his Master's degree in 2010. He also ranked first in the Ph.D. Entrance Exam in Educational Technology twice in 2012 and 2013, and received his Ph.D. in Educational Technology from Tarbiat Modares University in 2018. He has presented dozens of scientific papers in scientific journals and conferences, and his areas of expertise include instructional design, e-learning, and social network analysis.

- Email: a.taghizade@cfu.ac.ir
- ORCID: 0000-0002-3800-4504
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Fatemeh Zahra Ahmadi is an Assistant Professor of Curriculum Studies at the Organization for Educational Research and Planning, Tehran, Iran. Due to her work experiences as a midwife and educator in Iran and Uganda to promote women's health literacy. She designs and implements research-based dialogical curriculum for different audiences including students, teachers, and mothers. She teaches health and curriculum studies at the National Teacher Education University. Her areas of research interest include health literacy, curriculum design, curriculum implementation, and pre-service teacher education.

- Email: fatemehzahra.ahmadee@gmail.com
- ORCID: 0000-0003-3107-4961
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Sepideh Norouzi is an Assistant Professor of Curriculum Studies at the Farhangian University, Tehran, Iran. Due to her work experiences as a teacher and mathematician in Iran, she designs mathematics curriculum for student teachers in Iran, the content domain of elementary school math textbooks, and mathematics stories for young students. She teaches mathematics and curriculum studies at Farhangian University. Her areas of research interest include mathematics literacy, curriculum design and implementation, and pre-service teacher education.

- Email: Sepideh.b.noruzi@cfu.ac.ir
- ORCID: 0000-0001-7528-6451
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
S. S. Musavian, A. Taghizade, F. Z. Ahmadi, S. Norouzi, "Fuzzification of items of media and educational materials and tools," J. Electr. Comput. Eng. Innovations, 13(1): 81-92, 2025.

DOI: 10.22061/jecei.2024.11019.757

URL: https://jecei.sru.ac.ir/article_2194.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

An Intelligent Two and Three Dimensional Path Planning, Based on a Metaheuristic Method

B. Mahdipour, S. H. Zahiri *, I. Behravan

Department of Electrical Engineering, Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.

Article Info

Article History:
Received 25 May 2024
Reviewed 23 July 2024
Revised 27 August 2024
Accepted 28 August 2024

Keywords:
Particle Swarm Optimization (PSO)
Path planning
Autonomous Underwater Vehicle (AUV)
Unmanned Surface Vehicle (USV)

*Corresponding Author's Email Address: *hzahiri@birjand.ac.ir*

Abstract

Background and Objectives: Path planning is one of the most important topics related to the navigation of all kinds of moving vehicles such as airplanes, surface and subsurface vessels, cars, etc. Undoubtedly, in the process of making these tools more intelligent, detecting and crossing obstacles without encountering them by taking the shortest path is one of the most important goals of researchers. Significant success in this field can lead to significant progress in the use of these tools in a variety of applications such as industrial, military, transportation, commercial, etc. In this paper, a metaheuristic-based approach with the introduction of new fitness functions is presented for the problem of path planning for various types of surface and subsurface moving vehicles.

Methods: The proposed approach for path planning in this research is based on the metaheuristic methods, which makes use of a novel fitness function. Particle Swarm Optimization (PSO) is the metaheuristic method leveraged in this research but other types of metaheuristic methods can also be used in the proposed architecture for path planning.

Results: The efficiency of the proposed method, is tested on two synthetic environments for finding the best path between the predefined origin and destination for both surface and subsurface unmanned intelligent vessels. In both cases, the proposed method was able to find the best path or the closest answer to it.

Conclusion: In this paper, an efficient method for the path planning problem is presented. The proposed method is designed using Particle Swarm Optimization (PSO). In the proposed method, several effective fitness function have been defined so that the best path or one of the closest answers can be obtained by utilized metaheuristic algorithm. The results of implementing the proposed method on real and simulated geographic data show its good performance. Also, the obtained quantitative results (time elapsed, success rate, path cost, standard deviation) have been compared with other similar methods. In all of these measurements, the proposed algorithm outperforms other methods or is comparable to them.

Introduction

Intelligent surface and subsurface unmanned vehicles, are divided into two categories, unmanned surface vehicles (USV) and autonomous underwater vehicles (AUV). Intelligent subsurface vessels are actually unmanned submarines that are used to perform various missions. Exploration of special objects, mapping,

inspection and troubleshooting of oil or gas pipes in the deep oceans, military operations such as exploration and neutralization of sea mines, etc. These floats have a rechargeable constant voltage source and all kinds of sensors to perform different missions. These sensors are used to check the surrounding environment and measure parameters such as water salinity, pollution spread in water, navigation, depth measurement, etc. The reason for using the word automatic (intelligent) in the naming of these vessels is the absence of human intervention or any other external factor in guiding and controlling them during the mission.

They are controlled and guided by a programmed processor installed on each of them. Path planning is one of the most challenging research topics related to intelligent subsurface vessels, which has attracted the attention of many researchers. Since these vessels are unmanned and even a foreign operator is not used to guide them, determining their movement path from the origin to the destination is very vital and important. It is possible to determine the path of the float from the origin to the destination in a way before the start of the movement, and how to do this optimally has been the subject of many researches so far. While moving towards the destination, the vessel must have maneuverability to avoid collision with moving obstacles, such as other vessels or underwater creatures. In simpler words, the vessel must be able to leave the predetermined course in an emergency and then return to the course again.

Therefore, finding a suitable Path for the float to move from the origin to the destination is an important challenge in the future. However, the main goal in this research is path planning for the fleet of subsurface vessels. In many missions, for example military missions, a group of vessels must move towards the designated destination in cooperation with each other. The need to move the fleet with a specific arrangement towards the destination makes the path planning process more complicated. Because each vessel in the group must, in addition to moving towards the destination, maintain its position relative to other vessels in the considered formation. This problem becomes more complicated when due to the existence of numerous moving obstacles in the underwater space, the vessels must change their position to avoid collision with these obstacles and at the same time think about maintaining the group arrangement. For this purpose, communication between the vessels and also monitoring the surrounding environment at the moment is of high importance. Therefore, two general goals can be considered for this part of the project:

a. Designing the path planning module to determine the movement path of vessels from the origin to the designated destination. This module should be able to determine a reasonable path from the origin to the destination according to the obstacles in the underwater space.
b. Modular design to maintain the arrangement of floats while moving. At every moment of movement, this module determines the position of each vessel relative to other vessels, as well as the vessel that plays the role of the leader. Also, this module should be flexible so that when a vessel encounters an obstacle, it can change its course and return to its designated position in the group while maintaining the overall arrangement of the fleet.

Many studies have been reported on different methods of path planning for subsurface vessels. A new approach that is leveraged in research and has attracted the attention of researchers due to its ease of implementation, is the use of metaheuristic methods and biological algorithms. For further investigation, the researches related to this field are mentioned below.

In the problem of path planning in smart unmanned surface and subsurface vessels, we are faced with problems such as vessels colliding with each other, high energy consumption due to long and bad route determination, failure to maintain the relative distance of the vessels to each other, failure to maintain the overall shape of the fleet arrangement, large amount of calculations , high response time in route determination, limitations in sonar data, the unknown external disturbance, etc., which make routing scenarios more complicated. In this research a novel path planning methodology based on PSO algorithm is proposed to tackle these problems by defining new and simple fitness function.

One important issue in path planning in real environment is the unknown external disturbance, such as external waves, underwater currents and sea creatures. The proposed algorithm is designed in such a way that by saving time, it corrects the path towards not hitting the immediate obstacle. It can be considered as a factor for more reliable path planning in real environment.

Literature Review

A. An Overview 0f the Work Done in the Field of Navigation for a Subsurface Vessel

In 2001, a method based on genetic algorithm for subsurface navigation was proposed by a group of European researchers. In this method, the three-dimensional space between the origin and the destination is first gridded. That is, the first try to determine the points of the space that are between the origin and the destination. Then the genetic algorithm starts searching to find the best possible path among these points. In fact, each chromosome contains a sequence of these points

that connect them together to create a path between the origin and the destination.

In this research, the objective function is designed in such a way that the best path is the path that the float consumes the least possible energy to travel. As a result, Paths with shorter length are better Paths. Also, to simulate the static obstacles in the problem space, some points are randomly considered as obstacles, and this means that every path provided by each chromosome must be free of such points. It seems that networking the problem space is not a practical solution to find the right path. However, the proposed method can be effective for short-range path planning. Also, the results presented in this research have not been compared with any other method, which makes it difficult to reach a deep insight into the proposed method [1].

In 2007, a hybrid method was presented by a researcher named Zhang. In this method, which is a combination of genetic algorithm and Octree method, first, the problem space is divided into cells with a certain size. This work, is done with the aim of detecting empty areas and also detecting areas where there are static obstacles. Then the extracted cells are classified into three groups. The first category are cells that do not have any obstacle in their range. The second category are cells that are completely filled by a barrier, and the third category are cells that are partially filled by a barrier and are so-called half-filled. In the next step, the genetic algorithm searches the three-dimensional space to find a channel between the source and the destination among the empty and semi-empty cells. As a result, each chromosome contains a sequence of empty and semi-empty cells that can be followed from the origin to the destination. It should be noted that for half-empty cells, the Octree method divides the cell space into smaller sub-cells.

This increases the speed and accuracy of the presented method. In addition to this, other changes have been made in the genetic optimization algorithm. Including changing the intersection operator in such a way that the new chromosome produced from the intersection of two parent chromosomes contains the appropriate path. For this purpose, if two parent chromosomes contain a common cell, crossover is done from the location of this cell in the chromosomes and 4 new children are created, and if the parent chromosomes do not contain a common cell, crossover is not done. Similar to the previous research, the objective function is designed in such a way that paths with shorter length have better fitness value. Furthermore, the higher the number of empty cells in the designated Path, the better Path is considered. Despite Zhang's claim about the practicality of this method, it seems that the biggest weakness of the method is its dependence on the division of space into different cells, which makes the method unsuitable for long distances. In addition, the performance of the proposed method has not been scientifically compared with any other methods [2].

In 2011, in a joint project conducted by researchers from Hong Kong Polytechnic University and University of Calgary, Canada, a hybrid optimization method was presented to solve the subsurface floating path planning problem. In this method, which was designed by combining genetic algorithm and dynamic programming optimization method, the goal was to find the shortest and smoothest possible path to reduce the energy consumption of the float. In the presented method, since each chromosome contains a possible path between the origin and the destination, performing the normal and conventional Cross over in the GA algorithm produces unreasonable and inappropriate paths. To solve this problem and also reduce the probability of getting stuck in the local optimal point, the dynamic programming method has been used.

In the utilized model, after selecting two parent chromosomes, the DP method is used to extract reasonable paths among the parent chromosomes. The cost function or in other words the objective function whose minimum point should be found, is the weighted sum of three parameters:

a. The length of the Path provided, definitely shorter paths are better due to less energy consumption.
b. Sudden change of direction which is calculated by measuring the angle of direction change. Certainly, changing directions increases energy consumption. Therefore, the paths in which there are fewer such changes of direction are considered more suitable paths.
c. Sudden change of height, which is similar to the previous two parameters. In this case, the Paths are also desirable where the sudden change of height is less. The proposed method has been tested on two artificial environments and its performance has been compared with the conventional GA method.

The obtained results show the superiority of the proposed method over the conventional method. Despite the fact that the proposed method seems to be practical, simulating only on two environments without considering fixed obstacles, causes a little doubt in the accuracy and efficiency of this method. Also, only comparing the performance of the proposed method with another algorithm is not a suitable criterion for evaluating the proposed method [3].

In 2014, two Chinese researchers named Hui and Xiaodi proposed an interesting method called HPF (Heuristic Potential Field) to solve the problem of undersea path planning. In this method, unlike most of the presented methods, path planning is done in real time

by the float itself. Therefore, unlike other methods, here the conditions of the problem environment and the position of fixed and dynamic obstacles are not known. Rather, the float monitors the surrounding environment at any moment using various sensors, including sonar sensors, and if there is an obstacle in the visible range of the sensors, it will correct its course to avoid collision with the obstacle.

At the beginning of the work, a direct Path between the origin and the desired destination is determined as the main Path. Then the float starts moving on a straight path and checks the conditions of its surroundings at every moment. As soon as an obstacle is observed in the path ahead, the coordinates of the points of the obstacle that intersect with the path are determined and a new path is found using these coordinates. Therefore, at any moment the path of the float towards the destination is known and the float is moving on the specified path. As soon as an obstacle is observed by the sensors installed on the float, the Path correction module is activated and finds a new Path, and this process continues until reaching the destination. In spite of the fact that many details of the path planning method of the article in question are not mentioned and the method in question is presented in general terms, but it seems that the HPF method has the ability for hardware implementation due to the assumption of not knowing the geographical characteristics the problem environment is very close to reality. In addition, by using this method, it is also possible to navigate despite moving obstacles. However, the problem with this method is that the simplicity of this method may cause it to get stuck in the local optimal point or find a very long path. This problem can be solved by using optimization methods, especially meta-metaheuristic algorithm [4].

In 2015, researchers from Flinders University in Australia presented a dynamic method for subsurface navigation. In this method, at every moment of the Path, the path planning algorithm is executed and a new Path is found between the current point and the determined destination. Then, a controller module checks the new Path by checking the environmental conditions, including static and dynamic obstacles, as well as the direction of the water currents. In this research, similar to the previous researches, the path planning problem has been formulated as an optimization problem, and the Quantum behaved Particle Swarm Optimization (QPSO) has been used to solve it.

In this method, at first, the path of the float is determined from the starting point to the destination. Then the float starts moving and until reaching a certain point of the determined Path, the path planning module has the opportunity to find a new Path from the next point to the destination. Therefore, in certain time intervals, the Path of the float is updated from the point where it is present to the designated destination. Similar to other methods, the time to reach from the origin to the destination has been the most important criterion for the design of the objective function. The difference between the objective function considered in this method is that the direction of the water currents is also considered. It seems that in this method, path planning is done by dividing the three-dimensional space into different areas [5].

In 2016, a group of Australian researchers presented an interesting method for subsurface navigation. This method, which consists of two modules, first finds a general path between the origin and the destination using the genetic algorithm, which is similar to the previous methods introduced here, and the goal is to find the shortest possible path between the origin and the destination. Also, to find this path, the space between the origin and destination is gridded and the points that can be considered as positions are extracted. Therefore, the genetic algorithm finds the best possible sequence among the available points to find the path. Then the next module processes the sequence found by the genetic algorithm. In this module, the PSO algorithm searches the space between two consecutive points in the sequence to find a suitable path between these two points.

Therefore, first a general Path is determined and then other suitable paths are found between any two points of the general path. In the second module, the objective function is time. That is, the PSO algorithm searches the space with the aim of finding the shortest path. Although, the proposed method is an interesting method and it seems that this idea can be used, but it seems that gridding the problem space in the first stage is the main weakness of this method. The research team has also implemented their proposed method using the Imperialist Competitive Algorithm ICA (Imperialist Competitive Algorithm) and compared the results obtained from the implementation of the proposed method with GA and QPSO algorithms. The obtained results show that, on average, the ICA algorithm finds the optimal Path between the origin and the destination in a shorter period of time [6].

In one of the latest studies conducted in 2020, a method similar to the previous methods has been presented. The only difference is the utilization of the Ant Colony Optimization (ACO) for path planning. In this research, the ACO algorithm has only sought to find the best possible sequence among the available points between the origin and destination of the path. This method does not have any serious superiority over the other introduced methods, and the purpose of the project was only to show the power of the ACO algorithm in solving such problems [7].

As it is known, solving the path planning problem requires the use of optimization methods, which metaheuristic and meta-metaheuristic optimization algorithms are very suitable tools to solve this problem. Basically, these algorithms are designed and used to solve complex optimization problems with high dimensions. The high power and flexibility of these methods and of course their acceptable accuracy have made them popular tools in many researches in different scientific fields. However, other methods have been presented so far to solve this problem, which are briefly introduced in the Table 1. The most important weaknesses of these methods are the high probability of getting stuck in the local optimal point, the large number of calculations, and the weakness in facing the dynamic environment, which makes their use inappropriate for long distances.

B. A Review of the Work Done in the Field of Navigation for a Group of Subsurface Vessels

In 2014, a method based on genetic optimization algorithm for path planning a group of subsurface vessels was presented by a Chinese research team. This method is presented with the premise that the static obstacles in the 3D space are already detected (for example by the sonar sensors of a surface vessel). There is also a set of points that the subsurface vessel can pass through on the way to the destination. Using these points, genetic algorithm finds a path between the origin and destination of the desired float for each float. Therefore, each vessel has a specific origin and destination, and the desired method simultaneously finds the optimal. Similar to previous researches, in this method, the cost function is designed with the aim of finding the shortest possible path.

In order to prevent the collision of the floats, if the paths considered for two floats, which are specified by the chromosomes, are crossed, the time of each float's arrival at the desired point is checked. If two vessels reach the point of intersection at the same time, the Paths are corrected. Ignoring dynamic obstacles in the way is the problem of this method. In addition to this, the method proposed in this research has no fundamental difference with the methods related to the path planning of single vessels and only deals with the problem of their non-collision with each other. While one of the most important path planning challenges for the fleet of subsurface vessels is maintaining their formation while moving towards a specific destination [8].

In 2015, two Chinese researchers presented a method based on SOM (Self-Organizing map), BINM (Biological inspired neurodynamics model) and VS (Velocity Synthesis) algorithms for Task alignment and subsurface vessel group path planning. In this method, the generality of which is similar to the previous method, the SOM network is in charge of task alignment and path planning of each vessel. That is, at the beginning of the work, the SOM network determines a destination for each vessel from among the available destinations and predicts the Path to reach that destination. In other words, as in the previous method, in this method, the goal is not to move in a group and coherently towards a specific destination, but the mission is to go to several specific destinations and perform specific operations that the SOM decides which of the vessels will go to which destination. Criterion of task alignment by SOM is the lowest energy consumption.

Therefore, each float will move to a destination that is closer to its initial position. The BINM algorithm is intended to avoid the collision of the float with static obstacles. What differentiates this method from the previous method is the consideration of water currents in the navigation of vessels, which the VS algorithm is responsible for. However, the lack of modeling of dynamic obstacles and the lack of real-time processing are the main drawbacks of this method [9].

In 2020, a research team consisting of 7 Chinese researchers presented a hybrid method to solve the task division and pathing problem of a group of subsurface vessels. What differentiates this method from other methods is considering the point of release of the floats in the water and the point of their return to the water surface as the beginning and end points of the paths intended for them. Therefore, the energy used for the return path as well as the initial path should be calculated and included in the objective function. On the other hand, due to the fact that the number of destinations (target points) may be high and the distance between them is also long, the initial release points of each float should also be optimally determined. Therefore, a mission or task intended for a vessel is to go to the target point or points intended for that vessel, and to perform this mission, the vessel's path must be determined in an optimal way.

As a result, in the first part, the problem of task division and determination of initial points is done based on the divided tasks. This work is done using the differential evolution optimization algorithm (DE). In the second part, the Ant Colony Optimization (ACO) algorithm is implemented to determine the path corresponding to each mission. The objective function in this algorithm is designed in such a way that the intended path is traveled with the least possible energy consumption and without encountering obstacles in the path. In this method, the vessels operate independently of each other and each carries out the assigned mission. Therefore, there is no need to communicate between the vessels during the mission. On the other hand, the non-collision of the floats with each other means considering the dynamic obstacles in the problem [10]. In recent research conducted by a group of Indian researchers, a method for pathing a group

of subsurface vessels using the gray wolf optimization algorithm (GWO) is presented. Similar to the previous methods, the optimization algorithm determines the optimal path of each vessel towards the destination by knowing the conditions of the three-dimensional space. In simpler words, the position of static obstacles as well as the points in the three-dimensional space are already known. Therefore, the GWO algorithm performs path planning using these points and also considering the position of obstacles. The difference between this method and the previous two methods is the group movement of floats with a specific arrangement towards the destination. Therefore, the cost function is designed in such a way that in addition to finding the shortest path, the fleet arrangement is also maintained. Despite the researchers' emphasis on maintaining the arrangement of the vessels, only the relative distance of the vessels from each other has been considered. In addition to this, another big drawback of this method is not considering dynamic obstacles [11].

Many other methods have been presented so far to solve the problem of subsurface vessel pathing, surface vessel Pathing, Quadrotor motion pathing, etc, some of which are listed in the following table. Also the mentioned papers can be classified based on methodology in Table 1 format:

Table 1: A summary of the researches related to subsurface fleet Pathing

Reference Number	Year	Utilized Method	Merits & Demerits Analysis
[12]	2015	Kalman Filter	• Planning can be adjusted according to the dynamics of the environment. • Low computation cost • Only simulation results are obtained
[13]	2020	Fuzzy Logic & Ant Colony Optimization System	• Weak optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Not suitable for fast moving AUVs • Energy optimal trajectories are not obtained with collision avoidance
[14]	2017	Deep Reinforcement Learning (DRL)	• Near optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Suitable for fast moving AUVs • Energy optimal trajectories are obtained with collision avoidance
[15]	2006	Fuzzy Logic	• Weak optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Not suitable for fast moving AUVs • Energy optimal trajectories are not obtained with collision avoidance
[16]	2019	Neural Networks	• Near optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Suitable for fast moving AUVs • Energy optimal trajectories are obtained with collision avoidance
[17]	2008	Mixed Integer Linear Programming (MILP)	• Weak optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Not suitable for fast moving AUVs • Energy optimal trajectories are not obtained with collision avoidance
[18]	2019	Genetic Algorithm	• Minimizes the time expanses • Searches the solution from a large solution space • GA requires effective memory management • DE provides time optimized path in the corridor area but untimely collisions in obstruct evaluation of some good paths • The cost of computation is high

[19]	2018	Swarm Intelligence	• Near optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Suitable for fast moving AUVs • Energy optimal trajectories are obtained with collision avoidance
[20]	2014	Hybrid Genetic Algorithm & Particle Swarm Optimization	• Weak optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Not suitable for fast moving AUVs • Energy optimal trajectories are not obtained with collision avoidance
[21]	2023	Genetic Algorithm	• Minimizes the time expanses • Searches the solution from a large solution space • GA requires effective memory management • DE provides time optimized path in the corridor area but untimely collisions in obstruct evaluation of some good paths • The cost of computation is high
[22]	2023	Genetic Algorithm	• Far optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • GA requires effective memory management • Fuzzy-PID requires effective memory management • Energy optimal trajectories are obtained with collision avoidance
[23]	2023	Fuzzy Logic	• Near optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Not suitable for fast moving AUVs
[24]	2023	A Comprehensive Review of Path Planning Algorithms	• Near optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Suitable for fast moving AUVs • Energy optimal trajectories are obtained with collision avoidance
[25]	2023	Machine Learning	• Far optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Suitable for fast moving AUVs • Energy optimal trajectories are obtained with collision avoidance
[26]	2024	Artificial Potential Field Method & Multi-Algorithm Fusion	• Near optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Suitable for fast moving AUVs • Energy optimal trajectories are obtained with collision avoidance
[27]	2024	Fuzzy Logic & Simulated Annealing	• Far optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Suitable for fast moving AUVs • Energy optimal trajectories are obtained with collision avoidance
[28]	2024	Genetic Algorithm	• Near optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Suitable for fast moving AUVs • Energy optimal trajectories are obtained with collision avoidance
[29]	2024	Machine Learning	• Weak optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Not suitable for fast moving AUVs • Energy optimal trajectories are not obtained with collision avoidance

[30]	2024	Machine Learning	• Near optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Suitable for fast moving AUVs • Energy optimal trajectories are obtained with collision avoidance
[31]	2024	A Comprehensive Review of Path Planning Algorithms	• Far optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Suitable for fast moving AUVs • Energy optimal trajectories are obtained with collision avoidance
[32]	2024	Machine Learning	• Near optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Suitable for fast moving AUVs • Energy optimal trajectories are obtained with collision avoidance
[33]	2019	Particle Swarm Optimization (PSO)	• Weak optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Not suitable for fast moving AUVs • Energy optimal trajectories are not obtained with collision avoidance
[34]	2021	Particle Swarm Optimization (PSO)	• Far optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Suitable for fast moving AUVs • Energy optimal trajectories are obtained with collision avoidance
[35]	2022	Particle Swarm Optimization (PSO)	• Near optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Not suitable for fast moving AUVs
[36]	2024	Particle Swarm Optimization (PSO)	• Near optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Suitable for fast moving AUVs • Energy optimal trajectories are obtained with collision avoidance
[37]	2023	Particle Swarm Optimization (PSO)	• Weak optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Not suitable for fast moving AUVs • Energy optimal trajectories are not obtained with collision avoidance
[38]	2024	Artificial Potential Field Method & Multi-Algorithm Fusion	• Near optimal path following control for AUVs both in two- and three-dimensional environments • Computation complexity is high • Suitable for fast moving AUVs • Energy optimal trajectories are obtained with collision avoidance
[39]	2022	Machine Learning	-

In general, similar to the Pathing mode for a single vessel in the case of subsurface fleet Pathing, we are faced with an optimization problem, but more complicated, which prompts us to design the desired cost function by considering more restrictions. These restrictions, such as preventing vessels from colliding with each other, maintaining their relative distance from each other, maintaining the general shape of the fleet arrangement, etc, make the problem more complicated. As a result, solving this complex problem requires the use of powerful optimization tools that have a high ability to effectively search the response space and escape from the local optimal point. metaheuristic and heuristic optimization algorithms are among the most widely used and also the most popular optimization algorithms that have been used many times in recent years to solve complex problems that are difficult and time-consuming to analyze with other methods. The high ability of these

algorithms in finding the overall optimal point, very good convergence, high flexibility and easy implementation are among the advantages of these algorithms. In recent years, as seen, these algorithms have been used in path planning problems for all kinds of vehicles, and the results obtained show their high efficiency.

As seen, with an extensive review of the work done in Table 1, the above-mentioned contents were presented. All these papers are involved with the problem of path planning and have attempted to solve it. But what has been the main motivation for writing this paper are two very important issues, which are:

a. Calculation time required to find the best path. This problem is present in most references in such a way that it makes them out of practicality at the moment. In other words, the calculation time related to path planning in most of these researches is such that due to the type of sonar used in the vessels, practically the power of maneuvering and correcting the course is taken away from them and encountering an obstacle is inevitable.
b. Another important issue in the movement of a group of movers is the emphasis on maintaining their organization and group arrangement in most of the mentioned references. This point is somehow related to the problem mentioned in part a. That is, if a mover in a group of vessels encounters an obstacle, the time required to correct its course is often such that the possibility of returning to the previous arrangement and organization is denied.

In this research, the main motivation is to improve the two aforementioned problems, which have been tried to be solved by defining a simple fitness function.

Designing a Path Planning Module for a Group of Intelligent Subsurface Vessels

The final goal of this section is to provide a solution for the movement of the fleet of subsurface vessels towards the destination. In other words, a group of automatic subsurface vessels must move from a specific origin to a specific destination to perform a specific mission. For this purpose, two modules have been designed that operate in a hierarchical manner. The first module, which is called the path planning module, is responsible for finding a suitable Path between the origin and the destination, taking into account the presence of obstacles.

In fact, the output of this module is a set of consecutive points that determine the right path to move from the origin to the destination and are provided as input to the second module. The second module is called the path follower module. This module manages the movement of the fleet of subsurface vessels by keeping a certain formation in mind. In other words, this module determines the momentary position of each of the vessels according to the determined Path, the position of the vessel in the group and dynamic obstacles. In the following, first the Pathing module is fully described, then the functioning of the second module is explained.

Path Planning Module

Similar to many complex engineering problems, the path planning process can be turned into an optimization problem and then solved using different methods introduced to solve optimization problems. For this purpose, the first step is to design a cost function, finding its optimal point is equivalent to solving the main problem. This function is actually a kind of modeling of the desired problem space, which can be used to find a suitable answer for the problem. After designing a suitable cost function that models the problem space well, the next step is to choose a suitable method to solve the optimization problem. Various factors (such as execution speed, performance accuracy, and ease of implementation) have been considered for choosing the appropriate metaheuristic algorithm. In this paper, the PSO algorithm is used to solve the path planning problem. It is used in three-dimensional space. In the following, first the PSO algorithm is introduced, then the method of using this method in the design of the path planning module is described.

A. Particle Swarm Optimization Algorithm

The Particle Swarm Optimization (PSO) algorithm is a metaheuristic optimization algorithm inspired by the group movement of birds. In this algorithm, each particle whose position represents an answer to the problem is a member of a community that seeks to find the optimal answer collectively and inspired by the collective movement of birds. This algorithm was first proposed by Eberhart and Kennedy in 1995. The PSO algorithm uses the interaction between particles to find the optimal point. In this way, in the population created in each iteration, the particle that has the best fitness is known as the leader, and the rest of the particles tend to approach its position with the mechanism described below. Therefore, in this method, the elements cooperate to reach the optimal point. In this algorithm, after creating the initial population, each particle moves to find the optimal point in the response space. The movement of the particle is affected by two parameters:

a. The best position that the particle had from the beginning of the algorithm until the current iteration.
b. The position of the leader.

Therefore, each particle, in addition to searching the space individually, also looks at the position of the group leader and tends to approach his position. In fact, the best individual position has a function similar to the mutation operator in the genetic algorithm. In this algorithm, the movement of the particle is determined by the velocity vector. That is, after the velocity vector of the particle is

determined, by adding the velocity with the current position, the new position of the particle is obtained, which means the new answer. So, as mentioned above, each particle in the PSO algorithm, in addition to its position, has a memory where it stores the best position it has had so far. It also has a velocity vector that determines its position in the next iteration. The speed of each particle is obtained from the following equation:

$$v_i^{t+1} = \mathrm{w}.v_i^t + c_1.r_1(\, p_{leader}^t - y_i^t\,) + c_2.r_2.(\, p_{i_{best}}^t - y_i^t\,) \quad (1)$$

In this relation, v_i^{t+1} is the speed of the particle in the next iteration and v_i^t is its speed in the current iteration. p_{leader}^t is the position of the leader in the current iteration and $p_{i_{best}}^t$ is the best position that the particle had from the beginning of the algorithm execution to the t th iteration. y_i^t is the current position of the particle. The coefficients $c_1$and c_2 determine whether the particle will seek answers more individually or follow the leader of the group. As it is clear from the relationship, if c_1 is larger, the particle will seek to reach its best individual position. On the contrary, if c_1 is larger, the particle likes to approach the position of the leader, which has the best position among other particles in an iteration. c_1 is called Cognitive factor and c_2 is called Social factor. Usually, these two coefficients are chosen equal to 2. It has been found experimentally that the value of 2 for these two coefficients leads to the best performance. In (1), the constants r_1 and r_2 are two random numbers in the interval [0 1] that have been added to the relation to randomize the search for the answer space. Also, w is used to control the search process.

The largeness of this coefficient causes the search for a wider area of the response space, while its decrease prevents the scattering of particles. The value of this coefficient is high at the beginning of the algorithm execution and its value is low in the final iterations. Because in the end, it is better for the answers to converge. After the velocity of the particle is determined, its position is updated using (2):

$$y_i^{t+1} = y_i^t + v_i^{t+1} \quad (2)$$

The steps of implementing the PSO algorithm are as follows:

a. Creating a random initial population.
b. Calculating the fitness of particles and determining the leader.
c. Calculation of the speed of each particle.
d. Updating the position of particles.
e. Calculation of the best individual position of each particle.
f. If the condition is fulfilled, end the loop or repeat steps b to f otherwise.
g. Presenting the best answer found as the final answer of the algorithm.

B. Path Planning Using PSO Algorithm

In general, to solve any optimization problem using metaheuristic and heuristic algorithms, two basic points should be considered:

a. The structure of search agents, which are called particles in this algorithm.
b. Existence of a suitable objective function to calculate the fitness of particles. This objective function is clear in some cases, but in many cases, an objective function must be defined for the problem in question.

To find a suitable Path between the origin and the destination, the PSO algorithm finds a point in each run, and each point is found according to the previously found point. Therefore, in order to find the right path between the origin and the destination, the PSO algorithm must be executed in the right number so that the found points establish a suitable path between the origin and the destination. Therefore, the optimization problem is to find the next best possible point of the Path. For this purpose, we must first explain the structure of search agents (particles) and the objective function well.

B. 1. The Structure of the Search Factors

In each run, the coordinates of a point of the path must be provided as output by the PSO algorithm. This point is the best possible point according to the limitations of the problem and the considered objective function. Therefore, the particles must be three-dimensional. In other words, every time the PSO algorithm is executed, a six-dimensional vector is provided as output, which determines the location of the subsurface float. It is clear that the length, width, depth and roll and yaw angles determine the location of the float.

B. 2. Objective Function

The objective function should be designed in such a way that the next best possible point is its minimum (optimal) point. For this purpose, the limitations of the problem should be considered and based on that, a general definition of the minimum point should be provided and then the objective function should be designed. It can be said that the minimum point of this problem must have the following conditions:

a. The desired point should not be inside a three-dimensional object (obstacle). As a result, the objective function must be defined in such a way that its value is very high for points inside the range of obstacles.
b. This point should not be too close to the obstacle.
c. This point must be found on the way to the final point. In other words, this point should be located in a place that is considered a step forward and towards the destination compared to the previous point.

Therefore, in a minimization problem, we must define the objective function in such a way that if the new point is one step behind the previous point, the output of the function will be a large value so that this point is not considered a suitable point.

According to these assumptions, we consider the objective function as the weighted sum of six functions, each of which is introduced below:

Function f1: If the coordinates of a particle are inside a barrier (inside a three-dimensional object), the value of this function will be infinite, and if it is outside it, its value will be zero.

Function f2: the inverse of the distance between the point and the edge of the obstacle. In fact, adding this function is to prevent the point from getting too close to the obstacle.

Function f3: This function is defined so that if the position (coordinates) of the particle is not in the direction of reaching the final point, this particle is considered as a poor quality particle. For this purpose, the angle formed between the vector consisting of the previous point and the new point and the vector consisting of the new point and the final point must be minimized.

In the following, the said material will be further examined in the form of an example. Fig. 1 shows the concept of three functions f1, f2 and f3. To make it easier to understand the functions, this figure is designed for the two-dimensional problem. That is, it is assumed that the depth is constant and only longitude and latitude should be considered as input parameters. The goal of Pathing is the distance between the points marked with red and blue stars. The direction of movement is from the side of the red star to the blue star. In this figure, the obstacle (which is land) is marked with green color. Of course, the Path points should not fall within this range. In this image, one of the path points found by the PSO algorithm (with the mentioned settings) is drawn, and the red arrow shows the distance from the desired point to the obstacle border. In addition, the angle between the two described vectors is also indicated by the term tetta. If this angle is equal to zero, the point will be placed in the direction of the vector formed between the previous point and the final point. Definitely, from the point of view of this function, the point for which tetta is equal to zero is the optimal point. Since the found point is outside the obstacle range, the value of f1 function is zero for it. This point (particle) is the minimum point of the sum of functions f1, f2 and f3.

Function f4: This function is equal to the difference of the distance of the particle to the previous optimal point and a threshold value. The considered threshold value specifies the minimum distance between two consecutive points. In fact, this threshold value is something similar to the sampling rate.

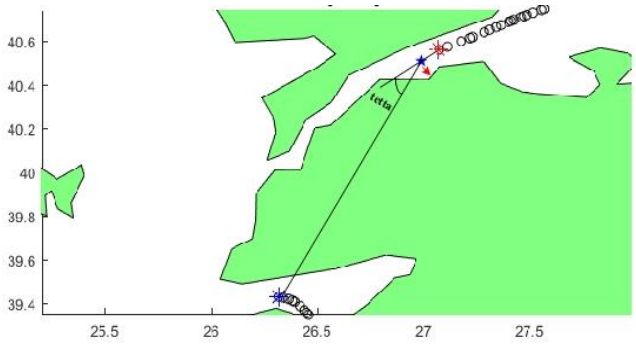


Fig. 1: Definition of functions f1, f2 and f3.

In other words, the new point must be at a minimum distance from the previous optimal point, which is determined by the threshold value entered by the user. The f4 function being zero means that the new point is at a minimal distance from the previous point.

Function f5: This function is defined to prevent the sudden change of direction of the path. In this function, the angle between the following two vectors is calculated:

a. The vector consisting of the new point and the previous point of the path (previous optimal point).
b. The vector consisting of the previous two points. If the angle between these two vectors is more than 90 degrees, the new point is not a suitable point and the f5 function value will be infinite for it. Otherwise, the f5 function value will be considered as zero.

Function f6: This function is defined to detect the optimal path leading to the destination (final point). First, a direct Path is drawn between the point under investigation and the destination. Then, among the points in this straight path, the number of points that fall within the obstacle range is considered as the value of f6 function. Certainly, the lower the f6 function value, the better the new point. Fig. 2 shows the concept of function f6 and the necessity of defining this function. Similar to the previous case, the reason for using a two-dimensional image is to better understand the conditions in which the use of the f6 function is necessary. According to this figure, it can be seen that the algorithm has found a part of the path. Now (supposedly) the next point should be selected from the points marked with red and black circles. The black circle is definitely a better choice because the value of the f6 function for it is lower than the corresponding value for the red dot. In fact, this function has been added to find the path leading to the destination and prevent getting lost.

In other words, the modeling of the problem space should be done in such a way that the value of the objective function (fitness) is lower for the blue circle. In

simpler words, this point should be a better point for the algorithm, and the reason is clear.

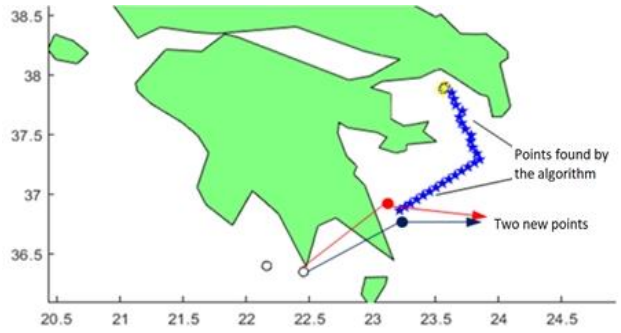


Fig. 2: The necessity of defining the f6 function in detecting the right path.

If the red circle is selected, the path to the final destination will be disturbed and the algorithm will have trouble finding the path. Because if you choose the red circle, the next points will be found in the same direction (i.e. away from the destination) and this means not finding the right path. The f6 function is intended to fix this problem and create an advantage in the blue dot over the red dot (as well as similar conditions).

Finally, the objective function is defined as the following relation:

$$objective\ function = f_1 + coeff_1 \times f_2 + coeff_2 \times f_3 + f_4 + f_5 + f_6 \quad (3)$$

The coefficients $coeff_1$ and $coeff_2$ are between zero and one and change during the execution of the process. The reason for this is the change in the importance of functions f_2 and f_3 at the beginning and end of the Pathing process. For example, at the beginning of the Pathing process, if the found point is not on the path between the previous point and the final point, there will not be much of a problem, but if it is very close to the border, the desired point is not a desirable point. On the contrary, at the end of the process, when we are close to the destination, the point must be in the direction of reaching the destination, and there is no problem if it is close to the Mazer. Therefore, at the beginning of the process, $coeff_1$ has a value close to one, and during the process, every time the PSO algorithm is executed, its value decreases, and for $coeff_2$ this procedure is considered completely opposite.

We know that the direct Path between origin and destination is the shortest Path. Therefore, if possible, the best answer for the Pathing problem is the direct Path. According to the previous explanations, we also know that the designed module seeks to find a point of the Path between the last found point and the main destination at any moment. As a result, it is possible to have a direct Path between the current origin (the last found point of the Path) and the destination. Therefore, at each stage, before the PSO algorithm is implemented, it is first checked whether it is possible to create a direct Path between the current origin and the destination or not. If such a possibility exists, the points on the direct Path between the current origin and the destination are added to the previous points and the Pathing process is terminated. Otherwise, the PSO algorithm determines the next point of the Path in the way described before. Fig. 3 shows the flowchart of the proposed method.

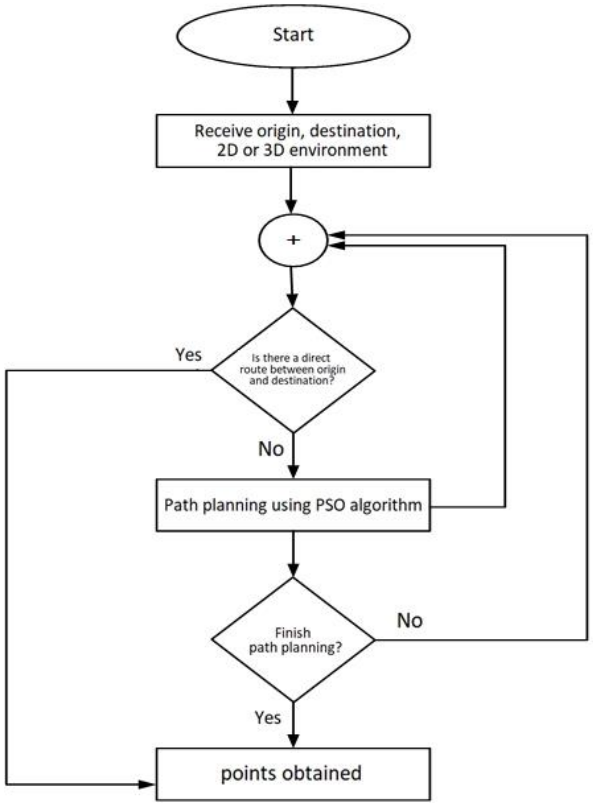


Fig. 3: Path planning module.

Similar data is necessary to evaluate the performance of the proposed module and measure its accuracy. In other words, the existence of Paths traveled by submarines or automatic subsurface vessels for comparison is inevitable. Due to the lack of access to such data and on the other hand, the existence of information related to the Paths traveled by various surface vessels on the Internet, the accuracy of the module was measured for two-dimensional space. The working method is to remove a part of the path traveled by a certain surface vessel (a ship) and find the desired path points using the designed module. Finally, the accuracy of the module is the degree of similarity between the main path traveled by the target surface float and the path completed by the designed module, which is measured using the following relationship:

$$1-\left(\frac{1}{1+e^{-dist}}-0.5\right)\times 2 \quad (4)$$

In the following, the results obtained in three different two-dimensional experiments are given. Then, in the next section, the results of using the module designed for artificial three-dimensional spaces are presented.

In this simulation, the pseudo code of PSO is shown in Table 2.

Table 2: The PSO pseudo-code

Number	the PSO pseudo-code
1	Initialize Population
2	for t=1 : maximum generation
3	for i=1 : population size
4	if $f(y_{i,d}^t(t)) < f(p_i^t(t))$ then $p_i^t = y_{i,d}^t$
5	$f(p_g^t(t)) = \min_i f(p_i^t(t))$
6	end
7	for d=1 : dimention
8	$v_i^{t+1} = w.v_i^t + c_1.r_1(p_{leader}^t - y_i^t) + c_2.r_2.(p_{i_{best}}^t - y_i^t)$
9	$y_i^{t+1} = y_i^t + v_i^{t+1}$
10	if $v_i^{t+1} > v_{max}^t$ then $v_i^{t+1} = v_{max}^t$
11	else if $v_i^{t+1} < v_{min}^t$ then $v_i^{t+1} = v_{min}^t$
12	end
13	if $y_i^{t+1} > y_{max}^t$ then $y_i^{t+1} = y_{max}^t$
14	else if $y_i^{t+1} < y_{min}^t$ then $y_i^{t+1} = y_{min}^t$
15	end
16	end
17	end
18	end

In this pseudo-code, v_i^{t+1} is the speed of the particle in the next iteration and v_i^t is its speed in the current iteration. p_{leader}^t is the position of the leader in the current iteration and $p_{i_{best}}^t$ is the best position that the particle had from the beginning of the algorithm execution to the t th iteration. y_i^t is the current position of the particle. The coefficients $c_1$and c_2 determine whether the particle will seek answers more individually or follow the leader of the group. The constants r_1 and r_2 are two random numbers in the interval [0 1].

In this simulation, the PSO configuration parameters are as described in Table 3 below.

Table 3: The PSO configuration parameters

Parameter	Value
Maximum PSO iteration	40
PSO population size	5
C1	1.5
C2	1.1
W	At first, it is 0.9, and then it decreases linearly to the value of 0.1 with respect to iteration changes.

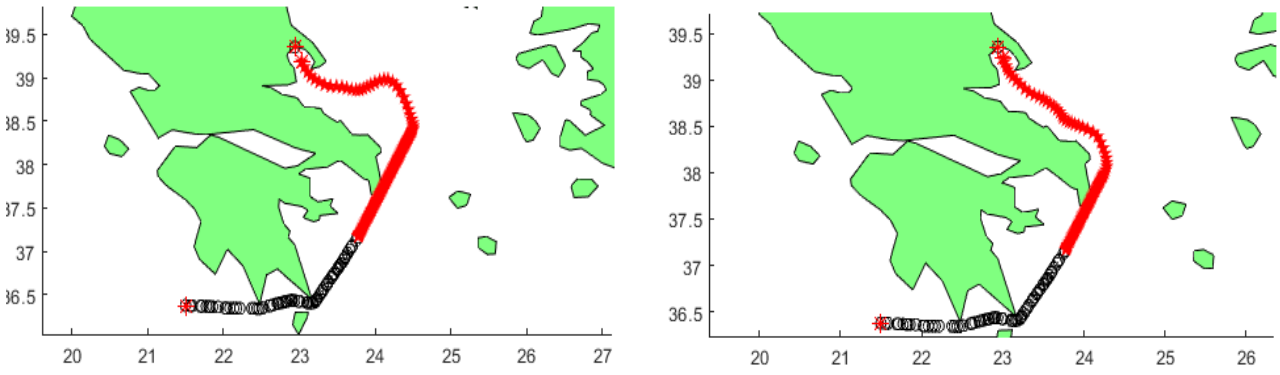
Simulation Results

A. Simulation Results on the Performance of the Module in Two-Dimensional Space

To evaluate the performance of the module, three paths were manually emptied.

A. 1. The First Path

Fig. 4 shows the main Path and the empty Path. This Path includes 192 points, from which 122 points have been removed for evaluation. In other words, 63% of it has been deleted. Fig. 5 also shows the output of the module in four executions of the Pathing module. In these images, the red five-pointed stars are the points found by the algorithm.



(a) (b)

Fig. 4: (a) & (b), the main path and the empty path, respectively.

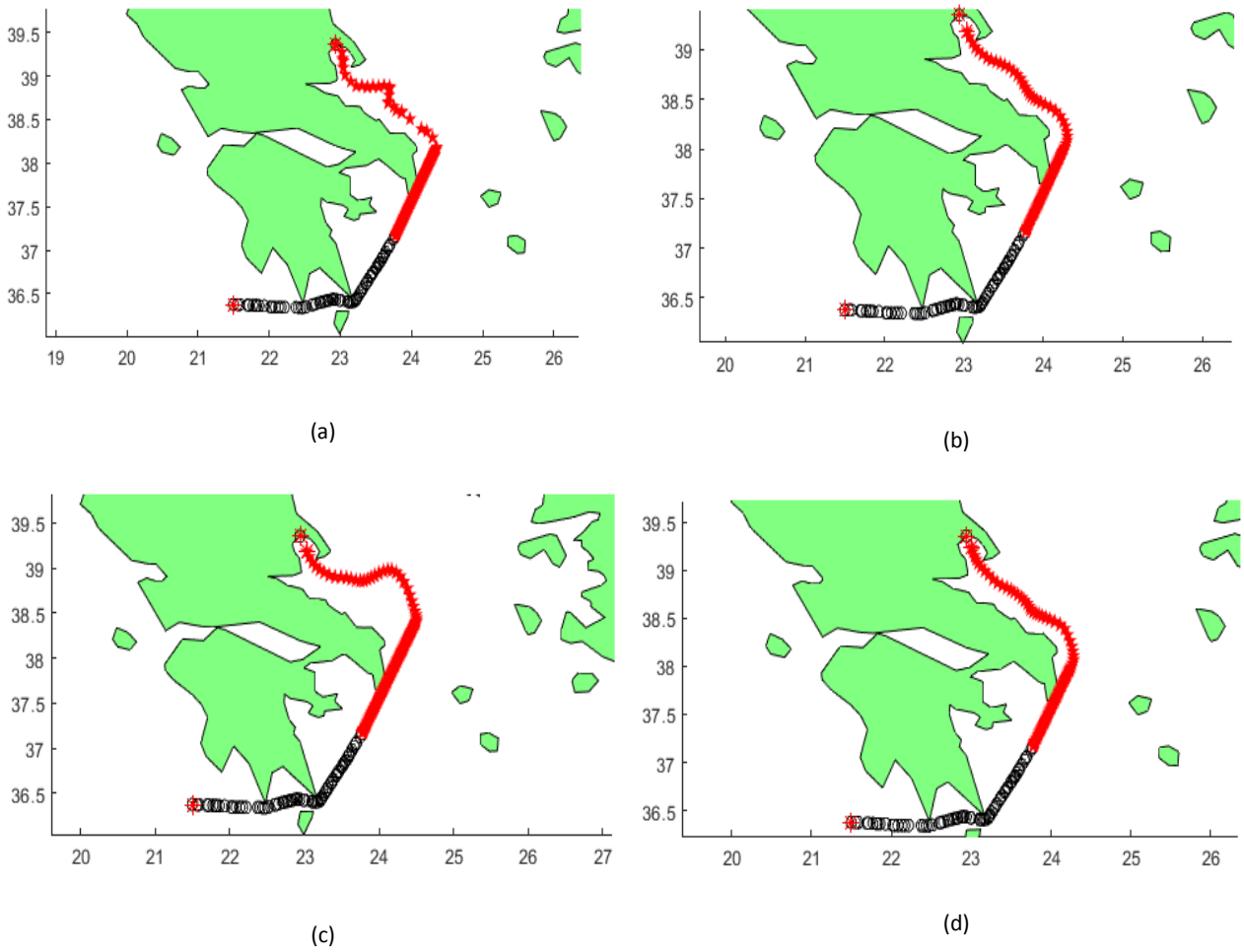


Fig. 5: Path planning module output in 4 independent experiments.

A. 2. The Second Path

95% of this Path is empty. The image of the original path and the empty path to be completed are shown in Fig. 6. Similar to the previous path, the output image of the module in four independent tests can be seen in Fig. 7.

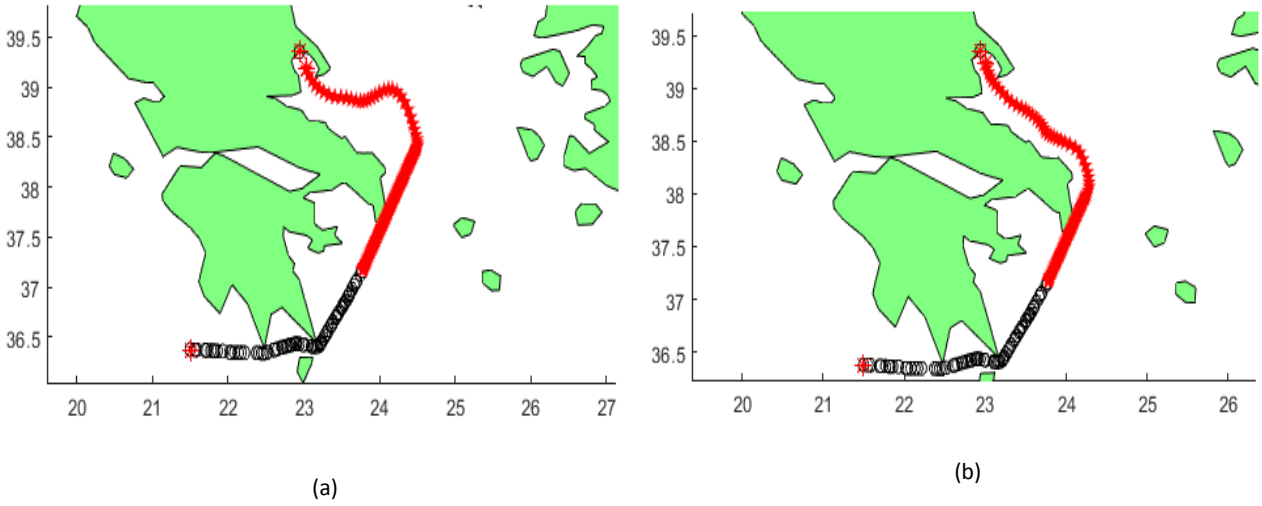


Fig. 6: (a) & (b), respectively, the main path and the empty path.

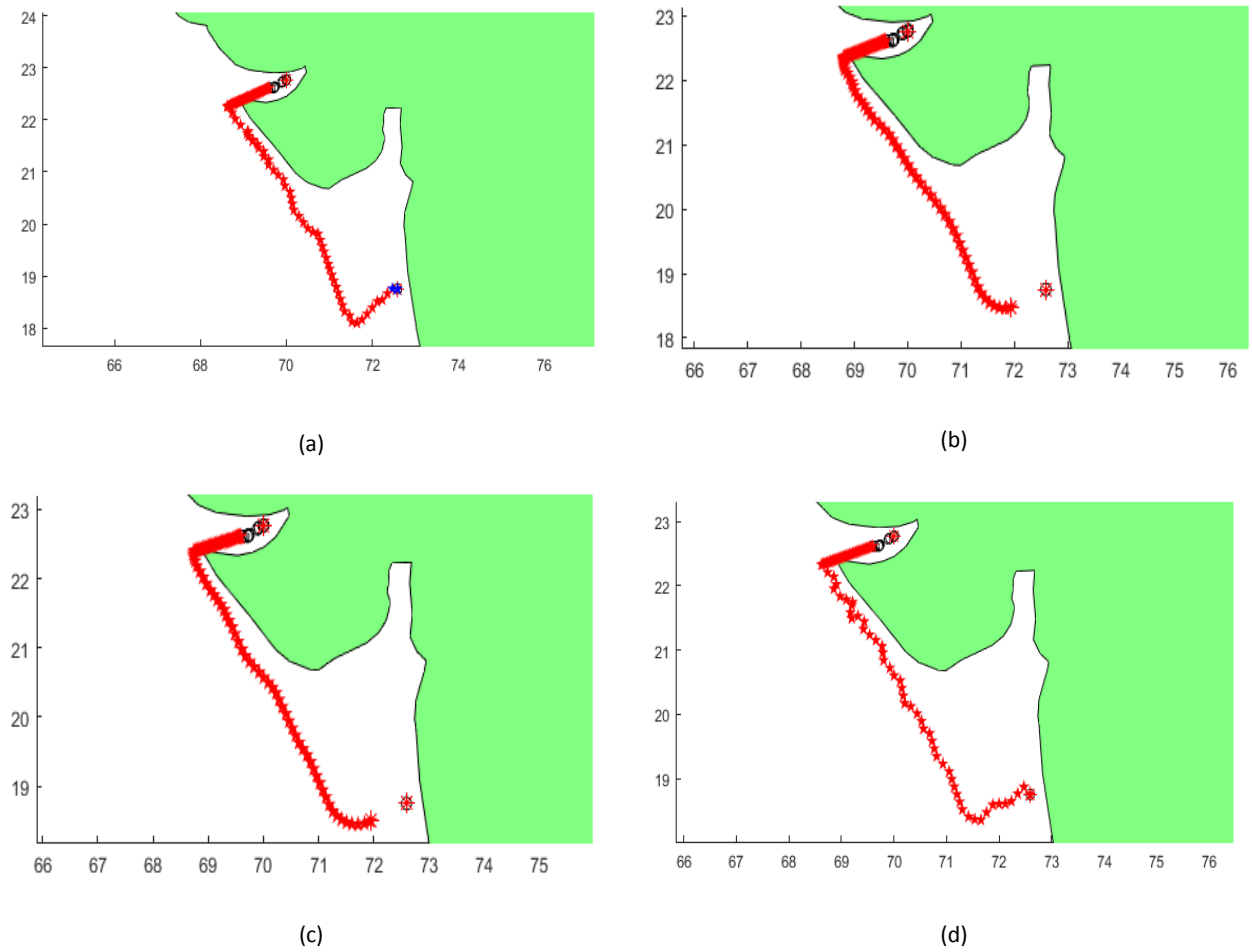


Fig. 7: Path module output in 4 independent experiments.

A. 3. The third path

Fig. 8 shows the main path and the empty path, and Fig. 9 shows the output of the module in four separate tests.

For this reason, 85% of the main path is empty.

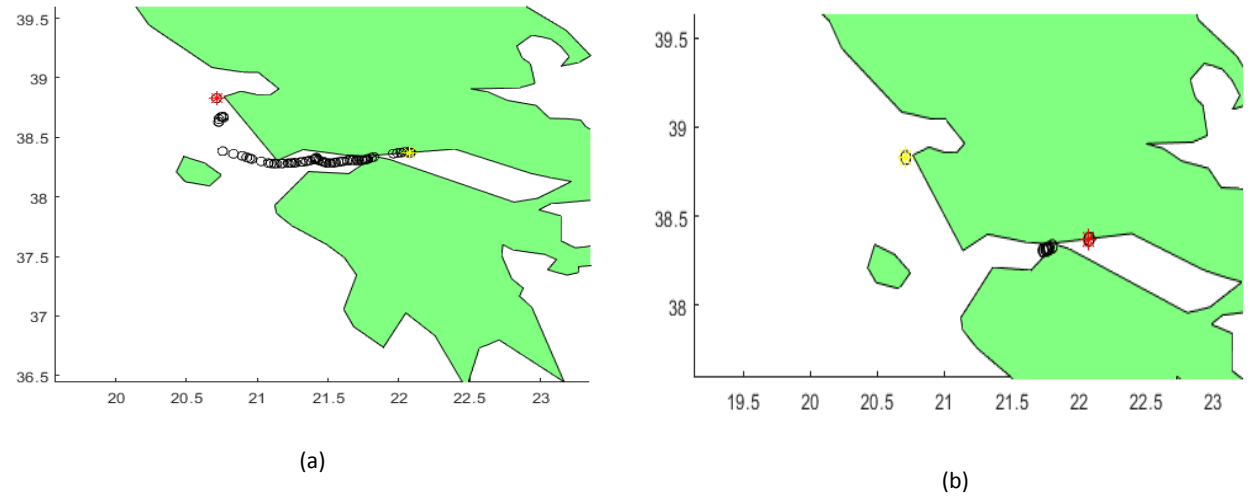


Fig. 8: (a) & (b) are the main path and the empty path respectively.

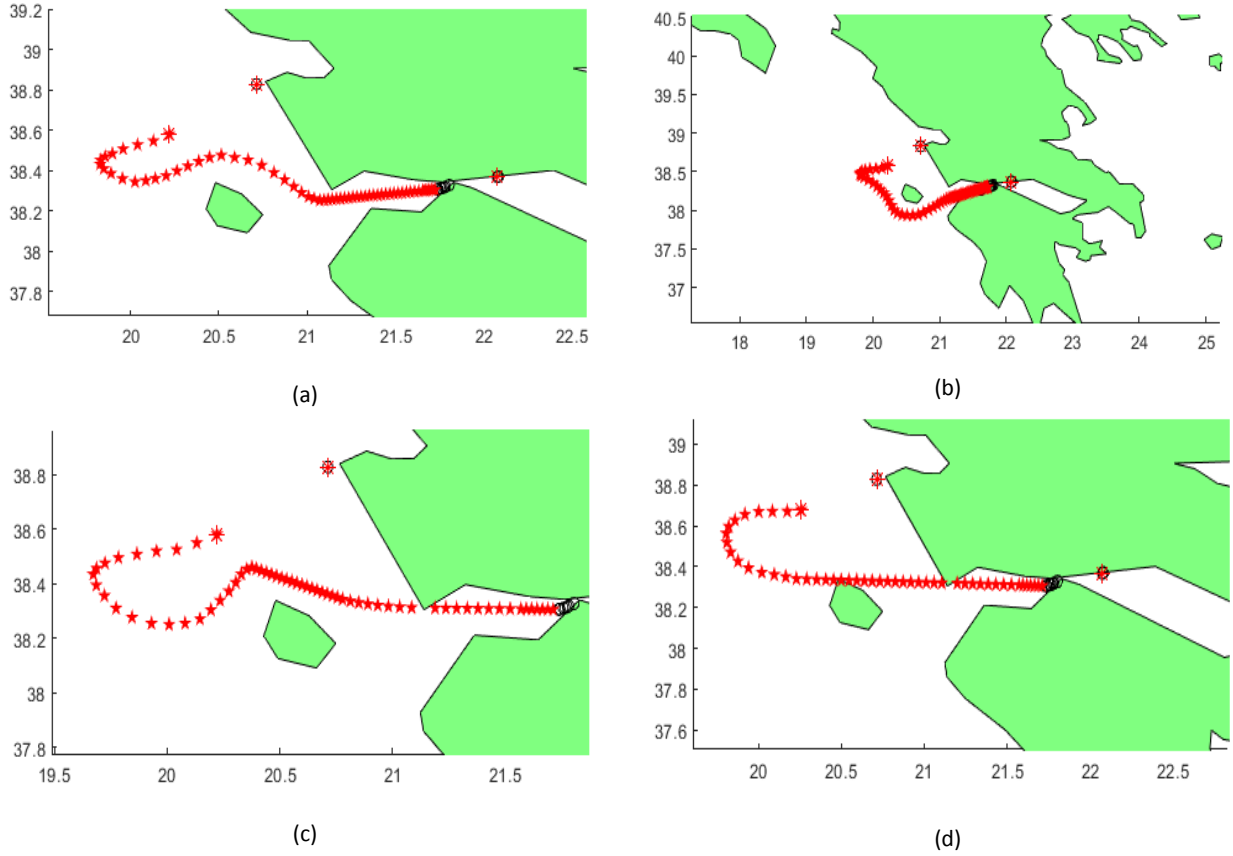


Fig. 9: Path planning module output in 4 independent experiments.

The image related to the output of the module for this test is given in Fig. 10.

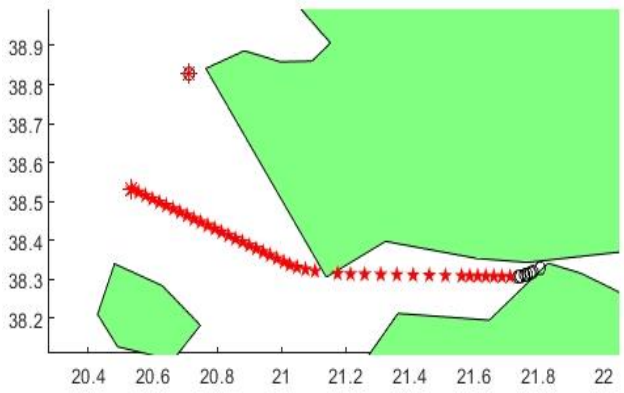


Fig. 10: The output of the module in the 12th test for the third path, for which the calculated accuracy is 83%.

According to Fig. 4 to 10, it can be seen that the proposed module provides very good performance in two-dimensional space. However, the main purpose of designing the module is to navigate in 3D space. It is easy to upgrade the designed module by adding the third dimension (which represents the depth of the sea) to the equations related to the PSO algorithm, for Pathing in the 3D space. Below is the output of the Pathing module for four experiments in different 3D spaces.

In addition to the visual detections that we had above, some quantitative parameters are presented in Table 4. In this Table, time elapsed is the time of running the proposed algorithm (in ms), success rate is the number of success in obtaining the best (or near the best) path with respect to all experiments for each scenario (12 in this paper).

Finally, standard deviation is the standard deviation value of the obtained best fitness values (for 12 experiments).

Table 4: Numerical results of experiments in two-dimensional space

Scenario	Time Elapsed (ms)	Success Rate	Standard Deviation
The first path	9.01	95%	3.2%
The second path	9.04	91%	5.1%
The third path	9.08	96%	2.7%

B. Simulation Results on the Performance of the Module in 3d Space

B. 1. The First Experiment

Fig. 11 shows the intended test environment. The scenario considered for this experiment is to move from the origin (indicated by a blue star) to the destination (indicated by a red star) in the presence of 6 obstacles (indicated as cylinders and spheres). According to this form, it is not possible to move directly between the origin and the destination. In Figures 12 to 14, the Paths found by the Pathing module are shown from three different views.

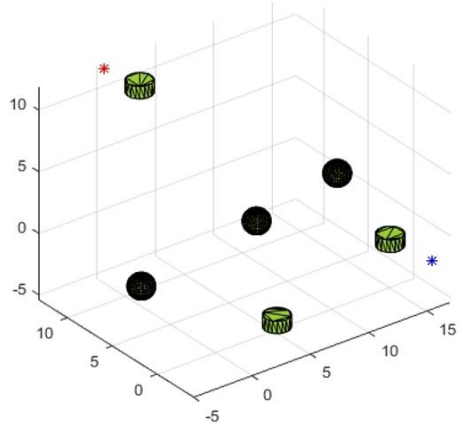


Fig. 11: 3D environment of the first experiment.

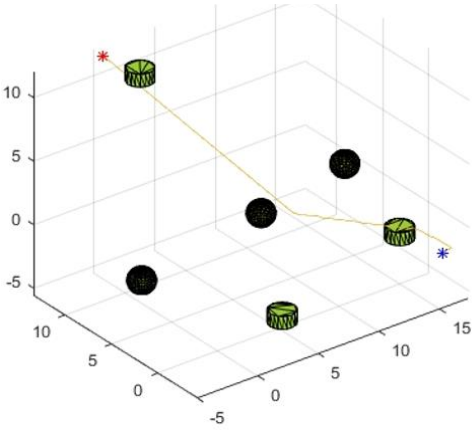


Fig. 12: The path between the origin and the destination in the first experiment from the first view.

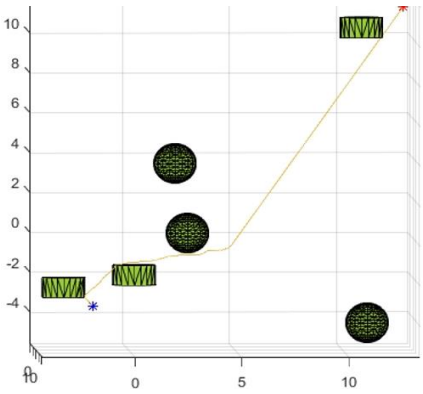


Fig. 13: The path between the origin and the destination in the first experiment from the second view.

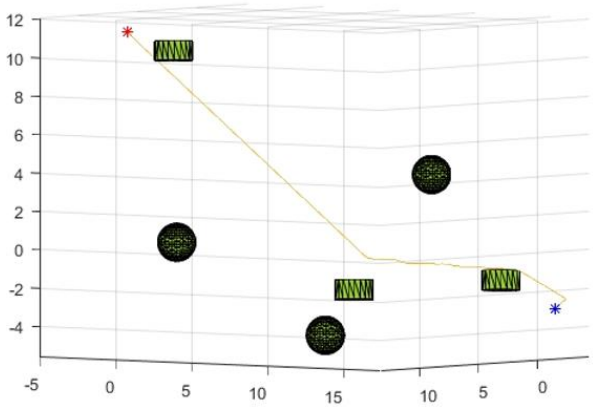


Fig. 14: The Path between the origin and the destination in the first experiment from the third view.

B. 2. The Second Experiment

In this experiment, the three-dimensional space is completely similar to the first experiment, with the difference that the place of origin and destination has been changed. Figs. 15 to 17 show the path found from three different views.

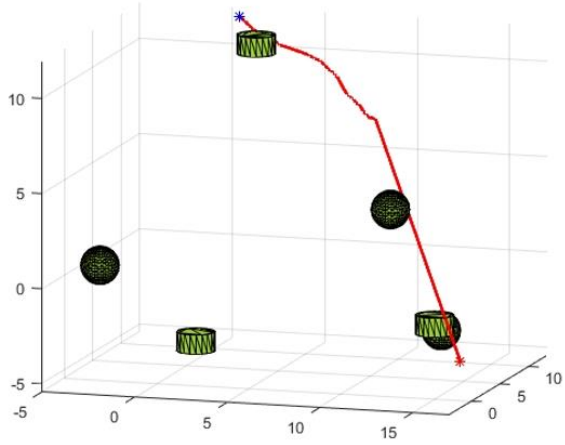


Fig. 15: The path found between the origin and the destination in the second experiment from the first view.

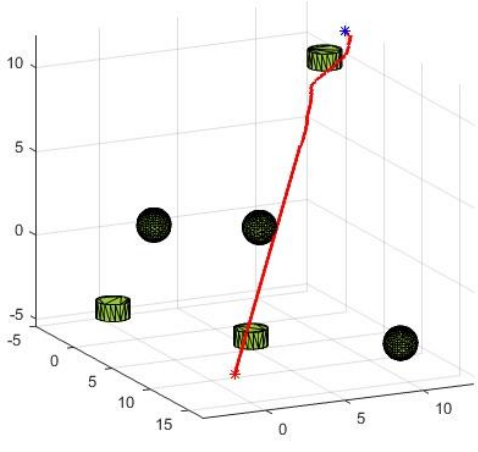


Fig. 16: The path found between the origin and the destination in the second experiment from the second view.

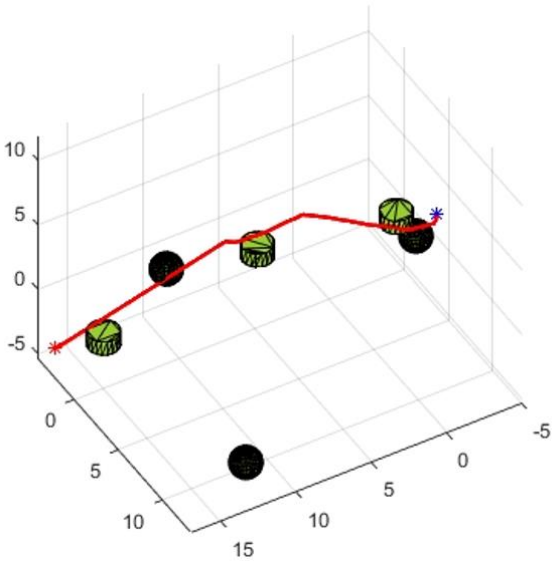


Fig. 17: The path found between the origin and the destination in the second experiment from the third view.

B. 3. The Third Experiment

The three-dimensional environment designed in this experiment is shown in Figs. 18 to 21 show the path found from three different angles.

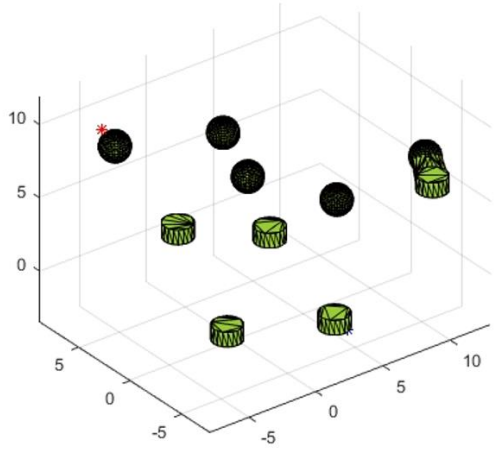
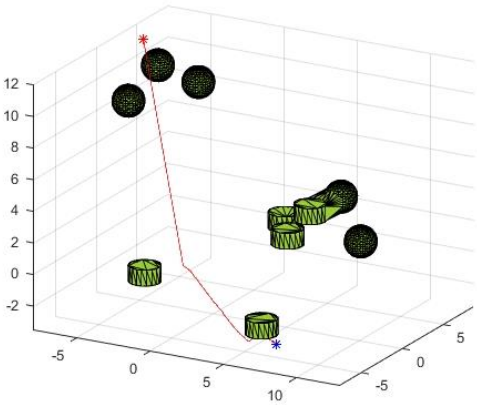


Fig. 18: 3D environment of the third experiment.



Fi. 19: The path found between the origin and the destination in the third experiment from the first view.

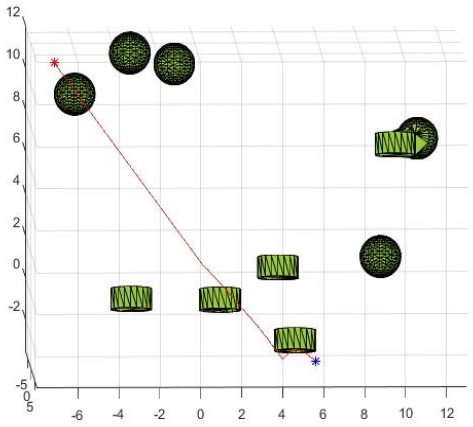


Fig. 20: The path found between the origin and the destination in the third experiment from the second view.

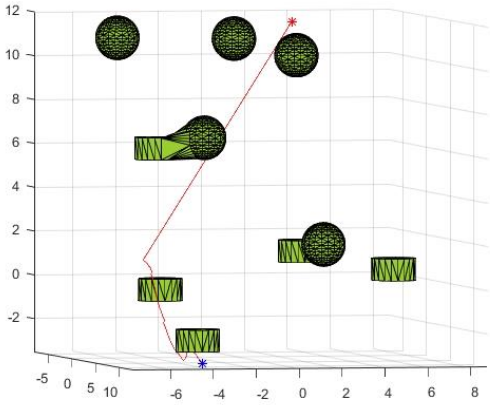


Fig. 21: The path found between the origin and the destination in the third experiment from the third view.

B. 4. The Fourth Experiment

In the fourth experiment, the environment of the problem is similar to the environment of the third experiment, with the difference that the place of origin and destination have been changed. Figs. 22 to 24 show the test output from three different views.

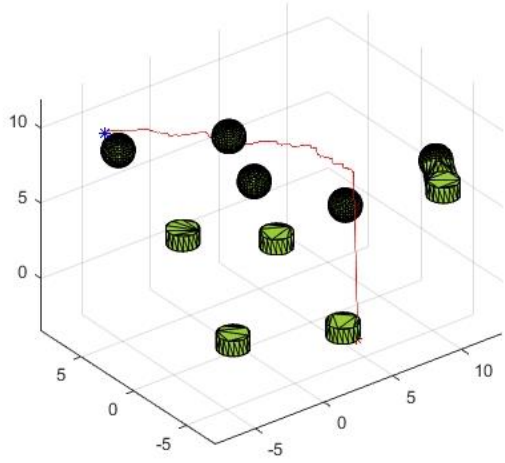


Fig. 22: The path found between the origin and the destination in the fourth experiment from the first view.

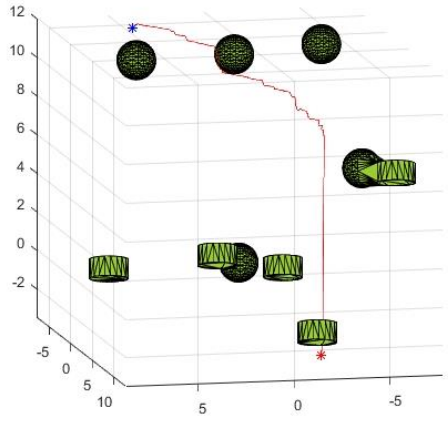


Fig. 23: The path found between the origin and the destination in the fourth experiment from the second view.

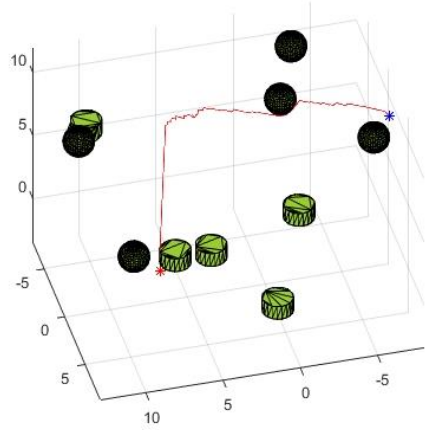


Fig. 24: The path found between the origin and the destination in the fourth experiment from the third view.

In addition to the visual detections that we had above, some quantitative parameters are presented in Table 5.

All of these parameters are defined and similar to Table 4.

Table 5: Numerical results of experiments in 3D space

Scenario	Time Elapsed (ms)	Success Rate	Standard Deviation
The first experiment	9.09	96%	6.2%
The second experiment	9.13	95%	7.5%
The third experiment	9.21	93%	8.4%
The fourth experiment	9.32	91%	10.2%

Also, regarding noise sensitivity, a numerical analysis has been done in four 3D space scenarios, in the form of Table 6. Here, we have set a tolerance of 5% in the estimation of obstacles (this tolerance can be due to sonar error, insufficient information on the map, sudden movement of obstacles, sudden underwater currents, sound noise, jamming, etc.).

We have applied this amount of tolerance randomly with a uniform distribution in each experiment (50 repetitions) to the position of the obstacles.

Table 6: Noise sensitivity analysis on the proposed Path Planning method

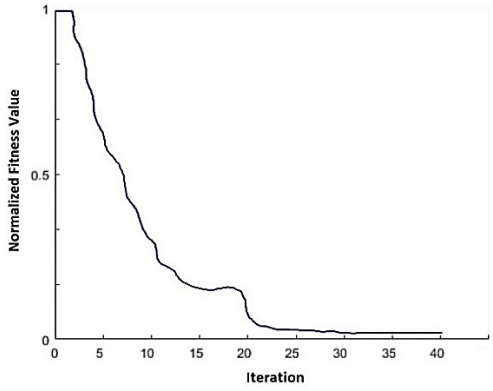
Scenario	Success Rate	Success Rate (with 5% Tolerance)
The first experiment	96%	90%
The second experiment	95%	88%
The third experiment	93%	85%
The fourth experiment	91%	82%

In the above Table 6, the impact of this tolerance on the success rate as the most important factor in the path planning performance shows that the presence of noise has been effective and has been able to reduce the success rate.

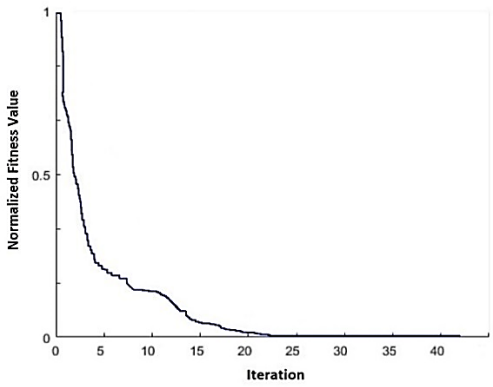
Considering that the success rate in the first experiment, is 6% on average, and in the second experiment, on average by 7%, and in the third experiment, on average by 9%, and in the fourth experiment, it has decreased by 10% on average, but still the presented method was able to find the path well.

C. The Figure of the Best Cost Function

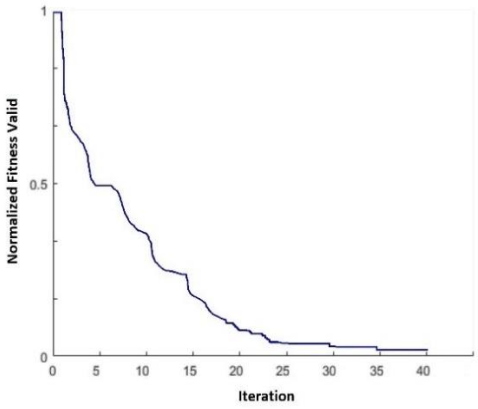
The normalized value of the fitness function is given in Fig. 25.



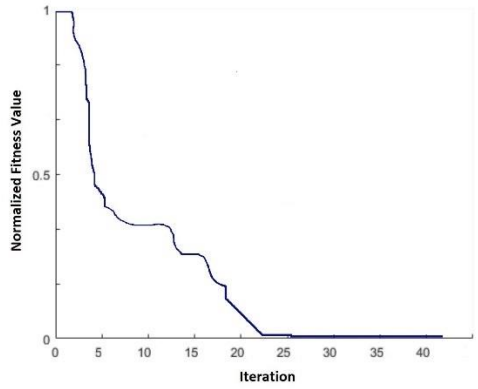
(a)



(b)



(c)



(d)

Fig. 25: The figure of the best cost function.

Path Following Module

This module manages the collective movement of subsurface vessels from origin to destination while maintaining order. In this module, after choosing a certain arrangement, the leader float and the follower floats are determined and the position of each of them is determined at each moment of the movement to the destination according to the type of arrangement. In the performed simulations, two types of arrangements are considered for the fleet of subsurface vessels:

a. Arrangement of arrowhead (^).
b. Linear arrangement.

In the first arrangement (arrowhead arrangement), the float placed at the tip of the arrow is the leader of the group and the floats placed on both sides are the follower floats. In the second arrangement (linear arrangement) the first float is the leader float and other floats are followers.

The path found by the Pathing module determines the position of the leader at each moment of the movement. Therefore, according to the position of the group leader, the position of other vessels is determined during movement.

Determining the position of the follower vessels is done according to the position of the vessel in the fleet, the position of the leader and the minimum distance from the front and side vessels.

In all the way, the main goal is to maintain the overall shape of the makeup. Therefore, at every moment of moving towards the destination, the next position of each follower vessel is calculated based on the principle of keeping the fleet formation. If the next position interferes with an obstacle, the navigation module for the desired float is activated automatically and finds the next point for the desired float. Definitely, in this situation, the overall composition of the fleet will change a little, which is inevitable.

In this way, it is considered not to encounter obstacles in this module. On the other hand, it is also necessary to mention that in the process of correcting the course of a vessel (in case of collision with an obstacle), the angular parameters related to the movement of the vessels (roll, pitch, and yaw) are also taken into account in the Pathing module to avoid collision with any obstacle. Placed. These parameters are used to model the placement of the 3D model of any subsurface vessel in the 3D space and then check whether or not the vessel collides with obstacles. In the following, pictures of group movement of floats in different tests are shown. In these tests, the fleet of subsurface vessels consists of 5 vessels that move in groups in different environments. In these pictures, subsurface vessels are marked with red (leader) and blue (follower) stars, so that in Figs. 26 to 29, the collective movement of the fleet of subsurface vessels with an arrowhead arrangement, and in Figs. 30 to 33, the movement of the group the collective fleet of subsurface vessels are shown in a linear arrangement.

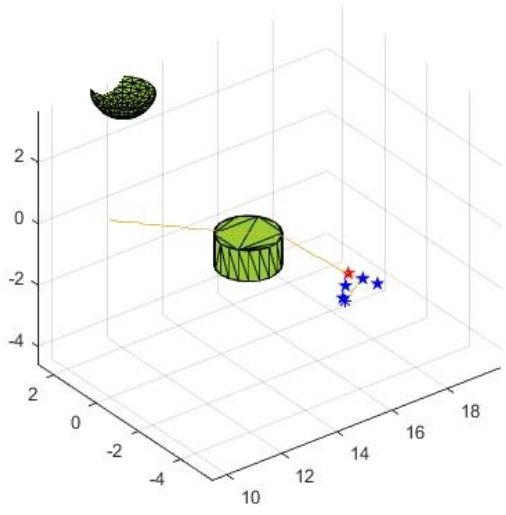


Fig. 26: Collective movement of the fleet of subsurface vessels with arrowhead formation.

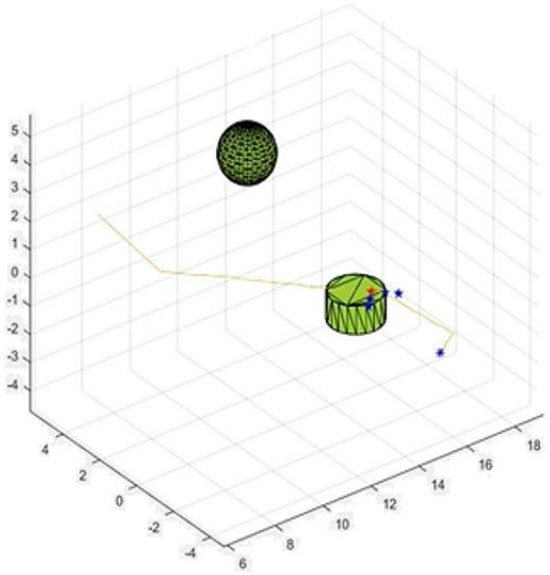


Fig. 27: Collective movement of the fleet of subsurface vessels with arrowhead formation.

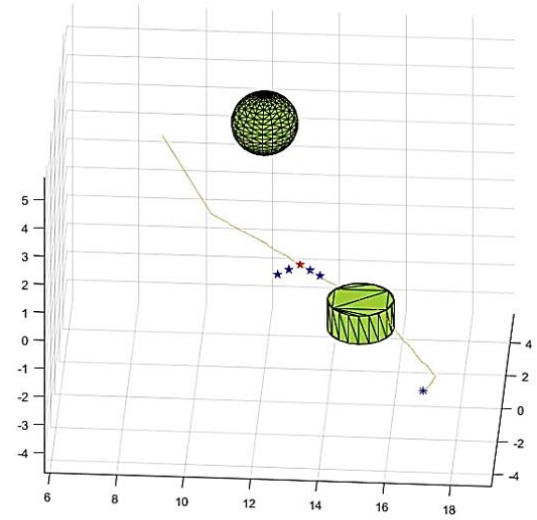


Fig. 28: Collective movement of subsurface vessels fleet with arrowhead formation.

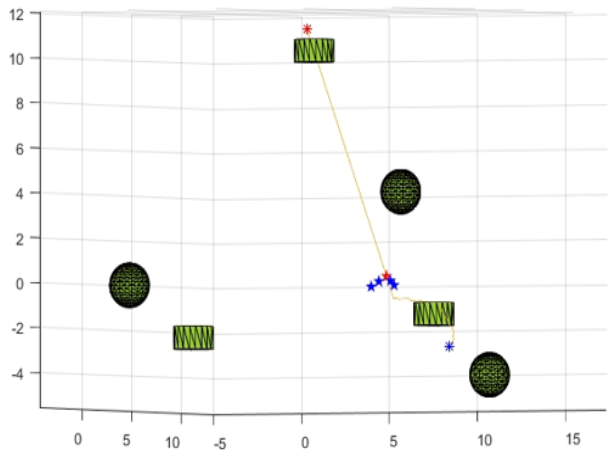


Fig. 29: Mass movement of the fleet of subsurface vessels in an arrowhead formation.

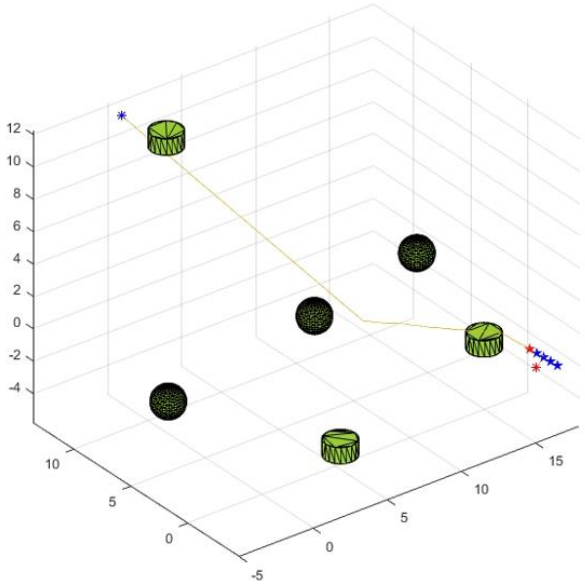


Fig. 30: Mass movement of a fleet of subsurface vessels with a linear arrangement.

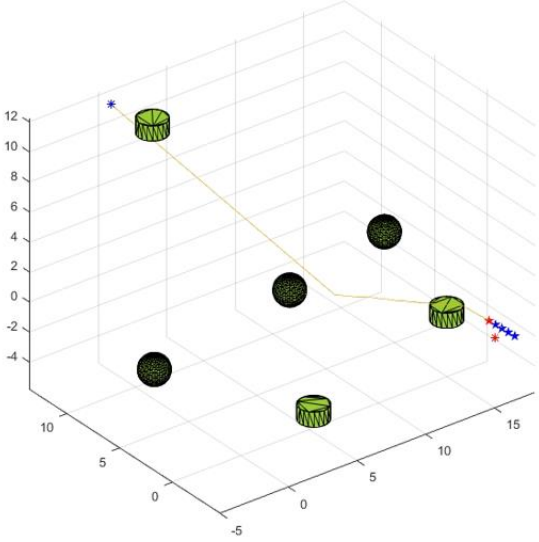


Fig. 31: Collective movement of the fleet of subsurface vessels in line formation.

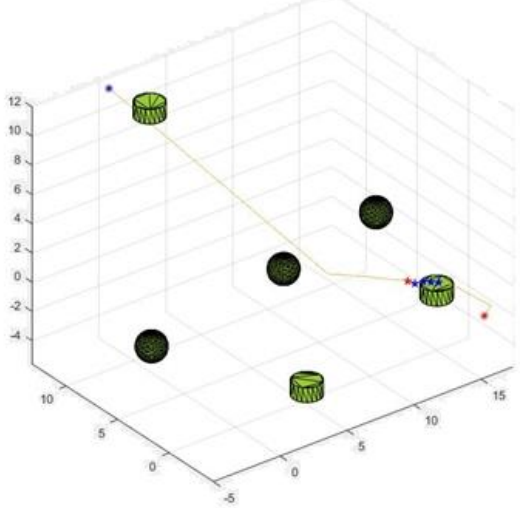


Fig. 32: Collective movement of the fleet of subsurface vessels in line formation.

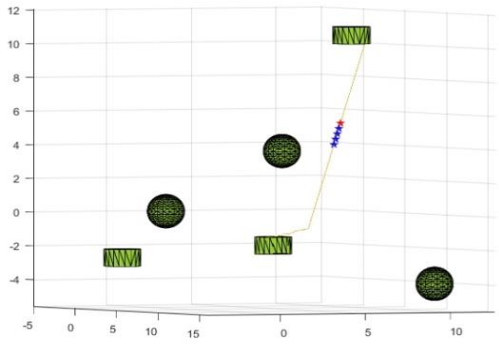


Fig. 33: Collective movement of the fleet of subsurface vessels in line formation.

By comparing the proposed method mentioned with other methods which had been utilized PSO method, it was found that our method has been able to improve the pathing speed and consequently, minimize the energy consumption of the moving group very well.

Of course, it should be noted that path changes will increase energy consumption, But naturally, in an environment that may face disturbances such as all kinds of noises, all kinds of waves and all kinds of sudden obstacles, This environment will be a random environment and because the environment is random, we must have the probability density function of the noise factors and the distribution of the types of events that cause the moving path to change and then according to these data, a theoretical research should be done in the random space to be able to provide a relatively accurate mathematical model to calculate the amount of energy consumed in that random environment.

The results of this investigation are given in the Table 7 below. The quantitative values in this table are defined for Table 4.

Table 7: Comparison with similar researches

Reference Number	Type of Path Generated	Collision/Obstacle Avoidance	Path Cost	Time of Path planning (ms)	Success Rate	Improvement Ratio (Time of Path planning)	Improvement Ratio (Success Rate)
[23]	Time optimal	Achieved	Moderate	11.53	90%	19%	05%
[33]	Time optimal	poor	High	16.34	86%	42%	09%
[34]	Energy optimal	Achieved	Low	10.29	92%	09%	02%
[35]	Time optimal	Achieved	Moderate	12.24	89%	23%	06%
[36]	Energy optimal	Achieved	Low	10.98	91%	15%	03%
[37]	Time optimal	Achieved	High	14.37	87%	35%	08%
Proposed Method	Time and energy optimal	Achieved	Low	9.32	94%	-	-

It is worth mentioning that, in this research, the maximum value of PSO repetition is considered 40, while in the mentioned papers, the maximum value of PSO repetition is considered as an average between 100 and 200. Path costs are compared as low, moderate and high. Collision and obstacle avoidance are discussed as achieved, limited and poor based on whether the algorithm focused on these issues or not.

Conclusion

In this paper, an efficient method for the path planning problem is presented. The proposed method is designed using Particle Swarm Optimization (PSO). In the proposed method, several effective fitness function have been defined so that the best path or one of the closest answers can be obtained by utilized metaheuristic algorithm. The results of implementing the proposed method on real and simulated geographic data show its fabulous performance. The achieved results are better or comparable with others method (time elapsed, success rate, Path cost, standard deviation, improvement ratio).

Of course we must pay attention to the fact that here we are facing two situations in the matter of path planning. The first mode (completely offline): In this case, the path map, static obstacles are clear, and the probability of dynamic obstacles, disturbances and noise is zero. In this case, after the algorithm is run, the path is designed and the moving will be able to move on this path. Of course, two conditions of sonar accuracy and speed must be taken into account here the second mode (online with restrictions): In this case, in the environment, it is possible to suddenly change the map and dynamic obstacles that do not have a very high speed. So that according to the times we presented in Table 4 and Table 5, immediately after the sonar detects the obstacle, the algorithm has the ability to quickly calculate and determine the next point on the path. If these two conditions are fulfilled, according to the appropriate time cost of the algorithm, it is possible for this algorithm to determine the next point on the path in an online. But if these conditions are not taken into account, the

presented algorithm cannot quickly calculate and specify the next point on the path online. The limitations that can affect the method presented in this paper are often dependent on the response speed of the sensors used on moving parts to detect obstacles. It is natural that if the sensors used do not have the required speed, the time required to implement the proposed method and correct the path will not be possible and the pathing will not be successful. This issue is the most important challenge of the method presented in this paper.

Regarding the fields of future work, it is possible to mention the use of dynamic PSO methods and in general dynamic metaheuristic PSO to deal with environmental disturbances. Also, deriving an accurate mathematical model for the energy consumption based on the different paths, is considered as another important topic for further work.

Author Contributions

B. Mahdipour has searched for important articles in this field. Then, by checking the results and collecting the necessary data and simulated the proposed method in MATLAB, the implementation of the proposed method has been done. Dr. S. H. Zahiri and Dr. I. Behravan reviewed the results and made changes in the way of implementation and final editing of the work.

Acknowledgment

We sincerely thank the respected referees for their accurate review of this paper.

Conflict of Interest

The authors announce no potential conflict of interest regarding the publication of this paper. Also, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission and redundancy have been completely witnessed by the authors.

Abbreviations

SONAR	Sound and Range Navigation
PSO	Particle swarm optimization
AUV	Autonomous underwater vehicle
USV	Unmanned surface vehicle
LKF	Linearized Kalman Filter
EKF	Extended Kalman Filter
IMU	Inertial Measurement Unit
INS	Inertial Navigation System
HPF	Heuristic Potential Field
GA	Genetic Algorithm
QPSO	Quantum behaved Particle Swarm Optimization
ICA	Imperialist Competitive Algorithm
ACO	Ant Colony Optimization
SOM	Self-Organizing map
BINM	Biological inspired neurodynamics model
VS	Velocity Synthesis
DE	Differential Evolution
GWO	Grey Wolf Optimizer

References

[1] E. Alfaro-Cid, E. W. McGookin, D. J. Murray-Smith, "Genetic algorithm optimisation of a ship navigation system," Acta Polytech., 41(4-5): 13-18, 2001.

[2] V. Kanakakis, N. Tsourveloudis, "Evolutionary path planning and navigation of autonomous underwater vehicles," in Proc. 2007 Mediterranean Conference on Control & Automation: 1-6, 2007.

[3] K. Hao, J. Zhao, Z. Li, Y. Liu, "Dynamic path planning of underwater AUV based on an adaptive genetic algorithm," Ocean Eng., 263: 11242, 2022.

[4] M. Y. Ju, S. E. Wang, J. H. Guo, "Path planning using a hybrid evolutionary algorithm based on tree structure encoding," Sci. World J., 2014: 1-8, 2014.

[5] X. Li, S. Yu, "Three-dimensional path planning for AUVs in ocean currents environment based on an improved compression factor particle swarm optimization algorithm," Ocean Eng., 280: 114610, 2023.

[6] L. Yang, J. Qi, D. Song, J. Xiao, J. Han, Y. Xia, "survey of robot 3d path planning algorithms," J. Control Sci. Eng., 2016(7426913): 1-22, 2016.

[7] S. Singhal, S. Tanwar, Aishwarya, A. Sinha, "Ant colony optimization based routing for underwater sensor network," in Proc. 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART): 33-38, 2020.

[8] M. P. Vicmudo, E. P. Dadios, R. R. P. Vicerra, "Path planning of underwater swarm robots using genetic algorithm," 2014 International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), 2014.

[9] H. Tang, Y. Yin, H. Shen, "A model for vessel trajectory prediction based on long short-term memory neural network," J. Mar. Eng. Technol., 21(3): 136-145, 2022.

[10] S. Singhal, S. Tanwar, Aishwarya, A. Sinha, "Ant colony optimization based routing for underwater sensor network," in Proc. 9th International Conference System Modeling and Advancement in Research Trends (SMART), 2020.

[11] H. Qin, T. Meng, Y. Cao, "Fuzzy strategy grey wolf optimizer for complex multimodal optimization problems," Sensors, 22(17): 6420, 2022

[12] M. Dinc, C. Hajiyev, "Integration of navigation systems for autonomous underwater vehicles," J. Mar. Eng. Technol., 14(1): 32-43, 2015.

[13] M. Panda, B. Das, B. Subudhi, B. Bhusan Pati, "A comprehensive review of path planning algorithms for autonomous underwater vehicles," Int. J. Autom. Comput., 17(3): 321-352, 2020.

[14] Y. Wang, "Review on greedy algorithm," Theor. Nat. Sci., 14(1): 233-239, 2023.

[15] J. C. Kinsey, R. M. Eustice, L. L. Whitcomb, "A survey of underwater vehicle navigation: Recent advances and new challenges," in Proc. IFAC Conference of Manoeuvring and Control of Marine Craft, 88: 1-12, 2016.

[16] D. Li, P. Wang, L. Du, "Path planning technologies for autonomous underwater vehicles-A review," IEEE Access, 7: 9745-9768, 2019.

[17] N. K. Yilmaz, C. Evangelinos, P. F. J. Lermusiaux, N. M. Patrikalakis, "Path planning of autonomous underwater vehicles for adaptive sampling using mixed integer linear programming," IEEE J. Oceanic Eng., 33(4): 522-537, 2019.

[18] D. H. dos Santos, L. M. G. Goncalves, "A gain scheduling control strategy and short-term path optimization with genetic algorithm for autonomous navigation of a sailboat robot," Int. J. Adv. Rob. Syst., 6(1), 2020.

[19] M. Y. Ju, S. E. Wang, J. H. Guo, "Path planning using a hybrid evolutionary algorithm based on tree structure encoding," Sci. World J., 2014: 1-8, 2014.

[20] I. Behravan, S. H. Zahiri ,S. M. Razavi, R. Trasarti, "Clustering a big mobility dataset using an automatic swarm intelligence-based clustering method," J. Electr. Comput. Eng. Innovations, 6(2): 251-271, 2018.

[21] Y. Li, J. Zhao, Z. Chen, G. Xiong, S. Liu, "A robot path planning method based on improved genetic algorithm and improved dynamic window approach," Sustainability, 15(5): 4656, 2023.

[22] M. R. Razali, A. A. M. Faudzi, A. U. Shamsudin, S. Mohamaddan, "A hybrid controller method with genetic algorithm optimization to measure position and angular for mobile robot motion control," Front. Robot., 9, 2023.

[23] C. L. Pen, W. J. Chang, Y. H. Lin, "Fuzzy controller design approach for a ship's dynamic path based on ais data with the takagi–sugeno fuzzy observer model," J. Mar. Sci. Eng., 11(6): 1181, 2023.

[24] M. Reda, A. Onsy, A. Y. Haikal, A. Ghanbari, "Path planning algorithms in the autonomous driving system: A comprehensive review," Rob. Auton. Syst., 174: 104630, 2024.

[25] D. An, Y. Mu, Y. Wang, et al., "Intelligent path planning technologies of underwater vehicles: A Review," J. Intell. Rob. Syst., 107(22), 2023.

[26] R. Zhou, K. Zhou, L. Wang, et al., "An improved dynamic window path planning algorithm using multi-algorithm fusion," Int. J. Control Autom. Syst., 22: 1005-1020, 2024.

[27] L. Yu, Y. Cai, X. Feng, et al., "Parallel parking path planning and tracking control based on simulated annealing algorithm," Int. J. Autom. Technol., 25: 867-880, 2024.

[28] W. Qiu, D. Zhou, W. Hui, et al., "Terrain-shape-adaptive coverage path planning with traversability analysis," J. Intell. Rob. Syst. 110(41), 2024.

[29] X. Zhai, J. Tian, J. Li, "A real-time path planning algorithm for mobile robots based on safety distance matrix and adaptive weight adjustment strategy," Int. J. Control Autom. Syst., 22: 1385-1399, 2024.

[30] B. Zhang,P. Liu,W. Liu, et al., "Search-based path planning and receding horizon based trajectory generation for quadrotor motion planning," Int. J. Control Autom. Syst., 22: 631-647, 2024.

[31] M. Reda, A. Onsy, A. Y. Haikal, A. Ghanbari, "Path planning algorithms in the autonomous driving system: A comprehensive review," Rob. Auton. Syst., 174(104630), 2024.

[32] J. Yu, Z. Chen, Z. Zhao, et al., "A path planning method for unmanned surface vessels in dynamic environment," Int. J. Control Autom. Syst., 22: 1324–1336, 2024.

[33] L. Yujie, P. Yu, S. Yixin, Z. Huajun, Z. Danhong, S. Yong, "Ship path planning based on improved particle swarm optimization," in Proc. 2018 Chinese Automation Congress (CAC), 2019.

[34] X. Wang, K. Feng, G. Wang, Q. Wang, "Local path optimization method for unmanned ship based on particle swarm acceleration calculation and dynamic optimal control," Appl. Ocean Res., 110(102588), 2021.

[35] S. Blindheim, T. A. Johansen, "Particle swarm optimization for dynamic risk-aware path following for autonomous ships," IFAC-PapersOnLine, 55(31): 70-77, 2022.

[36] H. Feng, M. J. Liu, H. Y. Xu, "Multi-target path planning for unmanned surface vessel based on adaptive hybrid particle swarm optimization," J. Huazhong Univ. Sci. Technol. (Natural Science Edition), 46(6): 59-64, 2018.

[37] W. Ma, Y. Han, H. Tang, D. Ma, H. Zheng, Y. Zhang, "Ship route planning based on intelligent mapping swarm optimization," Comput. Ind. Eng., 176(108920), 2023.

[38] D. Mu, T. Li, X. Han, Y. Fan, F. Wang, "Global path planning for unmanned surface vehicles in complex maritime environments considering environmental interference," Ocean Eng., 310(2), 2024.

[39] J. Pak, J. Kim, Y. Park , H. Il Son, "Field evaluation of path planning algorithms for autonomous mobile robot in smart farms," IEEE Access, 10: 60253-60266, 2022.

Biographies

Behrouz Mahdipour received the B.Sc. degree in Electrical and Electronics Engineering from Yazd University, Iran, in 2008 and M.Sc. degree in Electrical and Electronics Engineering from Bojnord University, Iran, in 2019. Currently, He is a Ph.D. student in Electrical and Electronics Engineering at Birjand University, Iran, as well as an assistant professor and researcher at a scientific and research institute. His research interests include the path planning of all types of intelligent unmanned surface and subsurface vessels and swarm intelligence algorithms.

- Email: behrouzmahdipour@birjand.ac.ir
- ORCID: 0009-0007-2380-9733
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Seyed Hamid Zahiri received the B.Sc., M.Sc. and Ph.D. degrees in Electronics Engineering from Sharif University of Technology, Tehran, Tarbiat Modarres University, Tehran, and Mashhad Ferdowsi University, Mashhad, Iran, in 1993, 1995, and 2005, respectively. Currently, he is a Professor with the Department of Electronics Engineering, University of Birjand, Birjand, Iran. His research interests include pattern recognition, evolutionary algorithms, swarm intelligence algorithms, and soft computing.

- Email: hzahiri@birjand.ac.ir
- ORCID: 0000-0002-1280-8133
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Iman Behravan received his B.S.c in Electronics Engineering from Shahid Bahonar University of Kerman, Iran. Also, he received his M.Sc. and Ph.D. degrees from the University of Birjand, Iran. He also worked as a post-doctoral researcher at the University of Birjand under the supervision of Professor Seyed Mohamad Razavi for two years. Currently he is working as a senior data scientist and blockchain developer at Kara Group, Tehran, Iran. His research interests include big data analytics, pattern recognition, machine learning, swarm intelligence, and soft computing.

- Email: i.behravan@birjand.ac.ir
- ORCID: 0000-0003-0319-1765
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
B. Mahdipour, S. H. Zahiri, I. Behravan, "An intelligent two and three dimensional path planning, based on a metaheuristic method," J. Electr. Comput. Eng. Innovations, 13(1): 93-116, 2025.

DOI: 10.22061/jecei.2024.10941.751

URL: https://jecei.sru.ac.ir/article_2193.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Multistep Model Predictive Control of Diode-Clamped Multilevel Inverter

P. Hamedani *

Department of Railway Engineering and Transportation Planning, University of Isfahan, Isfahan, Iran

Article Info

Article History:
Received 16 May 2024
Reviewed 11 July 2024
Revised 05 August 2024
Accepted 06 September 2024

Keywords:
Current control
Diode-Clamped inverter
Model Predictive Control (MPC)
Multistep prediction
Weighting factor

*Corresponding Author's Email Address:
p.hamedani@eng.ui.ac.ir

Abstract

Background and Objectives: To overcome the disadvantages of the traditional two-level inverters, especially in electric drive applications, multi-level inverters (MLIs) are the widely accepted solution. Diode-Clamped Inverters (DCIs) are a well-known structure of multi-level inverters. In DCIs, the voltage balance of the DC-link capacitors and the Common Mode (CM) voltage reduction are two important criteria that should be considered.
Methods: This paper concentrates on the current control of 3-phase 4-level DCI with finite control set model predictive control (MPC) strategy. Current tracking performance, DC-link capacitor voltage balance, switching frequency minimization, and CM voltage control have been considered in the objective function of the MPC. Moreover, the multistep prediction method has been applied to improve the performance of the DCI.
Results: The effectiveness of the proposed multistep prediction control for the 4-level DCI has been evaluated with different horizon lengths. Moreover, the effect of several values of weighting factors has been studied on the system behavior.
Conclusion: Results validate the accuracy of current tracking and voltage balancing in the suggested multistep MPC for the 4-level DCI. In addition, CM voltage control and switching frequency reduction can be included in the predictive control. Decreasing the CM voltage and switching frequency will oppositely affect the dynamic behavior and voltage balancing of the DCI. Therefore, selection of weighting factors depends on the system needs and requirements.

Introduction

Nowadays, multilevel converters (MLCs) are widely utilized in medium- and high-voltage applications for generating high-quality voltage and current [1], [2]. In comparison with the conventional two-level converter, MLCs can operate at higher voltage ratings for the same switching frequency and with lower dv/dt.

Therefore, MLCs are used in high-power drives [3], active filters [4], electric transportation systems [5]-[7], and other industrial applications such as fans, blowers, and pumps.

Three well-known topologies of multi-level inverters (MLIs) are the diode-clamped inverter (DCI), flying capacitor inverter [8], and cascaded H-bridge inverter [9]-[11]. Diode-clamped inverters offer high efficiency, low number of capacitors, and low stress on power electronic switches. Therefore, they are popular in various industrial applications. This paper concentrates on the 4-level diode-clamped inverter.

Traditional control approaches for producing the switching pulses of the DCI are the linear control [12] and the space vector modulation [13], [14]. Also, modulation techniques have also been presented in low frequency applications for reducing the common mode (CM) voltage, decreasing the output THD of the inverter and

harmonics elimination, and reducing the switching frequency of the inverter [15]-[17].

In recent years, new control approaches have been studied to control the power and current of inverters. Among them is the Model Predictive Control (MPC) [18]-[20]. The MPC offers desirable advantages such as fast response dynamics and compatibility with the system nonlinearity and various restrictions [21]-[28]. The MPC uses the mathematical model of the system to predict the system's behavior in future horizons. A cost function is defined according to the desired behavior of the system. In fact, the MPC is an optimization method that obtains the optimal switching state of the inverter by minimizing the cost function. Finally, the best switching state is applied to the inverter.

Since all calculations are repeated in each sampling period, for a large number of switching states, the computation burden is high. In the practical applications of the MPC for the MLIs, it is important to reduce the number of switching states. Recently, different strategies have been proposed for moderating the computation burden in MLIs with MPC [29]-[37]. Sometimes, this issue is solved by using the offline method. For this purpose, all possible states of the system are calculated offline. Accordingly, a look-up table is prepared and given as input to the system so that they can be used in each interval instead of many calculations [38], [39].

The MPC of DCI has been investigated in the literature for different applications [40]-[47]. However, the research in this field has not yet been completed. In [41], the MPC of 4-level DCI has been investigated for wind turbine systems. Different criteria have been included in the objective function of MPC such as switching states and CM voltage. In this work, the prediction has been only carried out with a horizon of N=2. Moreover, the effect of different weighting factors has not been studied in this work. The MPC for grid-connected 4-level DCI has been evaluated in [44]. Active and reactive power, capacitor voltage balancing, and switching frequency have been considered in the objective function. The delay-compensation method has been applied. But the CM voltage has not been minimized. Furthermore, multi-step prediction has not been utilized in the control scheme. In [44], the MPC of 4-level DCI has been investigated for wind energy systems. The capacitor voltage balancing and the number of switching states have been included in the objective function of MPC. However, the CM voltage has not been considered. In addition, the MPC of a 4-level DCI has been performed with a prediction horizon of N=1. In [47], a simplified MPC has been proposed for 4-level DCIs. The suggested method yields lower computational burden and total harmonic distortion (THD) in compare to the traditional MPC. However, only the current tracking and capacitor voltage balancing have been included in the objective function of the proposed MPC. In addition, the prediction has been only carried out with a horizon of N=1. This paper proposes a multistep MPC strategy for the current control of the 3-phase 4-level DCI, considering different cost functions. Current control, DC-link capacitor voltage balance, switching frequency minimization, and CM voltage control have been considered in the prediction method. The delay compensation with the multistep prediction method has been applied to improve the performance of DCI. The main contributions of this paper are as follows:

- Presenting a multistep predictive current control for the 3-phase 4-level DCI
- Evaluating several horizon lengths in multistep prediction control of the 4-level DCI
- Including various objectives in the predictive controller such as current tracking, DC-link voltage balance, reduction of the CM voltage, and decreasing the switching frequency
- Evaluating the effect of different values of weighting factors on the system performance

Mathematical Model of 4-Level DCI

Fig. 1 illustrates the topology of the three-phase 4-level DCI. Eighteen IGBT switches with anti-parallel diodes, eighteen clamping diodes, and three capacitors are used to generate four voltage levels. The switches are placed in up and down groups and receive complementary firing pulses. A 4-level DCI has 4^3=64 switching states. Table 1 shows all feasible switching conditions of the single-phase 4-level DCI and the related voltage level.

According to Table 1, the DCI voltages can be written as [43], [44]:

$$\begin{aligned} v_{aO} &= v_{C1}.S_{a1} + v_{C2}.S_{a2} + v_{C3}.S_{a3} \\ v_{bO} &= v_{C1}.S_{b1} + v_{C2}.S_{b2} + v_{C3}.S_{b3} \\ v_{cO} &= v_{C1}.S_{c1} + v_{C2}.S_{c2} + v_{C3}.S_{c3} \end{aligned} \quad (1)$$

where v_{cj} is the voltage of j-th capacitor ($j \in \{1,2,3\}$). S_{xy} is the switching state of the y-th IGBT ($y \in \{1,2,3\}$) in phase x ($x \in \{a,b,c\}$) of the DCI (as shown in Fig. 1). In the 4-level DCI, the CM voltage can be calculated as [41]:

$$v_{nO} = v_{CM} = \frac{v_{aO} + v_{bO} + v_{cO}}{3} \quad (2)$$

where O is the negative DC-Link and n is the neutral point of the load.

The phase voltage with respect to the load neutral can be written as [41]:

$$\begin{aligned} v_{an} &= v_{aO} - v_{nO} \\ v_{bn} &= v_{bO} - v_{nO} \\ v_{cn} &= v_{cO} - v_{nO} \end{aligned} \quad (3)$$

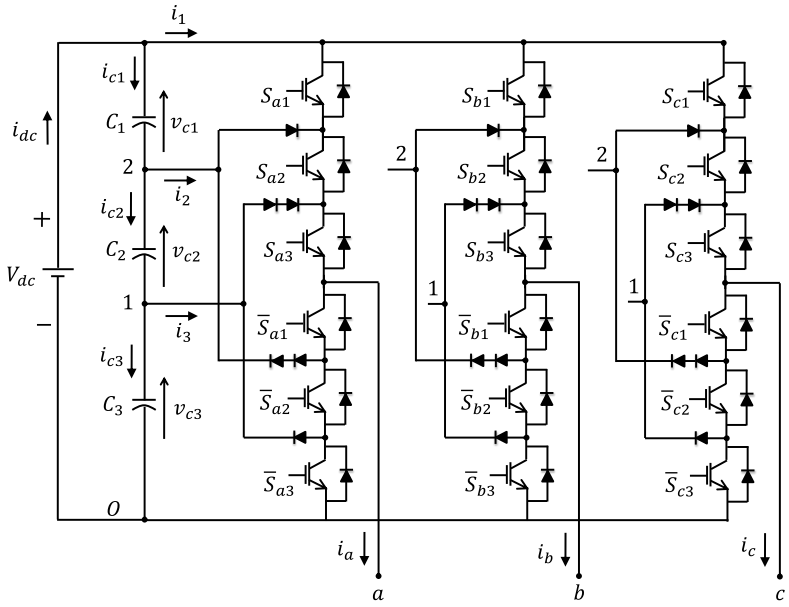


Fig. 1: Structure of 3-phase 4-level DCI [49].

Table 1: Switching states of the 4-level DCI [40]

S_x	Output Voltage Level	Switching Pulse		
		S_{x1}	S_{x2}	S_{x3}
0	0	0	0	0
1	v_{c3}	0	0	1
2	$v_{c2}+v_{c3}$	0	1	1
3	$v_{c1}+v_{c2}+v_{c3}$	1	1	1

The discrete-time model of the capacitor voltages can be written as [48]:

$$v_{c1}(k+1) = v_{c1}(k) + \frac{T_s}{C_1} i_{c1}(k) \tag{4}$$

$$v_{c2}(k+1) = v_{c2}(k) + \frac{T_s}{C_2} i_{c2}(k) \tag{5}$$

$$v_{c3}(k+1) = v_{c3}(k) + \frac{T_s}{C_3} i_{c3}(k) \tag{6}$$

in which $v_{c1}(k)$, $v_{c2}(k)$, and $v_{c1}(k)$ are the capacitor voltages. $i_{c1}(k)$, $i_{c2}(k)$, and $i_{c1}(k)$ are the capacitor currents and can be computed as [49]:

$$\begin{aligned} i_{C1}(k) &= -i_1(k) \\ i_{C2}(k) &= -i_1(k) - i_2(k) \\ i_{C3}(k) &= -i_1(k) - i_2(k) - i_3(k) \end{aligned} \tag{7}$$

where

$$\begin{aligned} i_1(k) &= K_{a1} i_a(k) + K_{b1} i_b(k) + K_{c1} i_c(k) \\ i_2(k) &= K_{a2} i_a(k) + K_{b2} i_b(k) + K_{c2} i_c(k) \\ i_3(k) &= K_{a3} i_a(k) + K_{b3} i_b(k) + K_{c3} i_c(k) \end{aligned} \tag{8}$$

where K_{xy} $(x \in \{a,b,c\}, y \in \{1,2,3\})$ can be defined as [49]:

$$\begin{aligned} K_{x1} &= S_{x1}\, S_{x2}\, S_{x3} \\ K_{x2} &= \bar{S}_{x1}\, S_{x2}\, S_{x3} \\ K_{x3} &= \bar{S}_{x1}\, \bar{S}_{x2}\, S_{x3} \end{aligned} \tag{9}$$

in which S_x is defined in Table 1.

For the resistive-inductive load, the discrete-time model of the DCI load current can be written as [48]:

$$i_a(k+1) = \left(1 - \frac{RT_s}{L}\right) i_a(k) + \frac{T_s}{L} v_{an}(k) \tag{10a}$$

$$i_b(k+1) = \left(1 - \frac{RT_s}{L}\right) i_b(k) + \frac{T_s}{L} v_{bn}(k) \tag{10b}$$

$$i_c(k+1) = \left(1 - \frac{RT_s}{L}\right) i_c(k) + \frac{T_s}{L} v_{cn}(k) \tag{10c}$$

in which k is the sampling instant and T_s represents the sampling time. R and L are the load resistance and inductance, respectively. $v_{an}(k)$, $v_{bn}(k)$, and $v_{cn}(k)$ are the phase voltages and can be obtained from (1)-(3) using the measured capacitor voltages and optimal switching states.

Single-Step Model Predictive Control of the 4-Level DCI

Fig. 2 represents the block diagram of the MPC strategy for a 4-level three-phase DCI. The main aim is to predict the load current and capacitor voltages in the next sampling instant. Accordingly, all 64 switching states of the 4-level DCI are searched to find the optimal switching state that minimizes the objective function. In the single-step MPC method and without the delay compensation, the predictions were made in the (k+1)-th sampling instant.

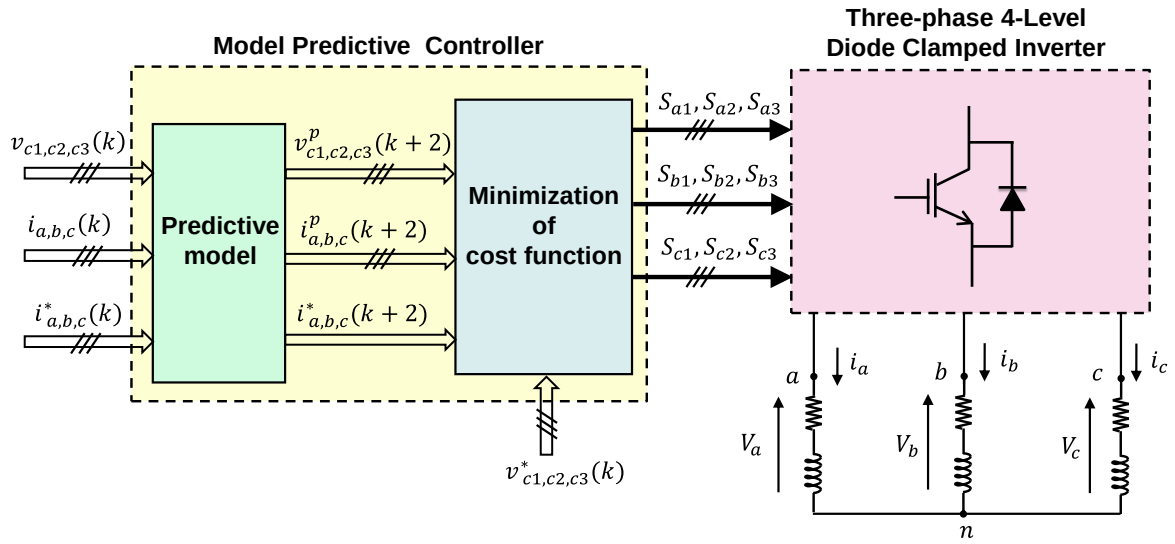


Fig. 2: MPC of the 4-level three-phase DCI with delay compensation method [40].

In practice, the computational delay due to the microprocessor's operation affects the accuracy of the prediction. Therefore, the delay compensation strategy is proposed to compensate the perdition error. In the delay compensation strategy, the predictions are made for the (*k*+2)-th sampling instant.

The prediction of load currents in the (*k*+2)-th sampling instant gives [49]:

$$i_x(k+2) = \left(1 - \frac{RT_s}{L}\right) i_x(k+1) + \frac{T_s}{L} v_{xn}(k+1) \tag{11}$$

where $x \in \{a, b, c\}$.

Furthermore, the capacitor voltages are predicted for the (*k*+2)-th sampling instant [49]:

$$\begin{aligned} v_{c1}(k+2) &= v_{c1}(k+1) + \frac{T_s}{C_1} i_{c1}(k+1) \\ v_{c2}(k+2) &= v_{c2}(k+1) + \frac{T_s}{C_2} i_{c2}(k+1) \\ v_{c3}(k+2) &= v_{c3}(k+1) + \frac{T_s}{C_3} i_{c3}(k+1) \end{aligned} \tag{12}$$

The overall objective function can be defined as:

$$g(k+2) = g_i(k+2) + \lambda_V\, g_{v_c}(k+2) + \lambda_S\, g_{sw}(k+2) + \lambda_{CM}\, g_{CM}(k+2) \tag{13}$$

where g_i, g_{v_c}, g_{sw}, and g_{CM} are the terms of the objective function to control current tracking, capacitor voltage balance, CM voltage, and switching frequency, respectively. λ_V, λ_S, and λ_{cm} are the weighting factors that adjust the capacitor voltages, CM voltage, and switching frequency, respectively.

g_i and g_{v_c} can be written as [44]:

$$g_i(k+2) = \sum_{x=a,b,c} \left(i^*_x(k+2) - i_x(k+2)\right)^2 \tag{14}$$

$$g_{v_c}(k+2) = \sum_{j=1,2,3} \left(v^*_{cj} - v_{cj}(k+2)\right)^2 \tag{15}$$

Moreover, v^*_{c1}, v^*_{c2}, and v^*_{c3} are the final capacitor voltages:

$$v^*_{c1} = v^*_{c2} = v^*_{c3} = \frac{V_{dc}}{3} \tag{16}$$

i^*_a, i^*_b, and i^*_c are the current references. The future current references of the (*k*+2)-th sampling instant can be computed using extrapolation [48]:

$$i^*_x(k+2) = 6\, i^*_x(k) - 8\, i^*_x(k-1) + 3\, i^*_x(k-2) \tag{17}$$

where $x \in \{a, b, c\}$.

Moreover, g_{sw} can be calculated as [44]:

$$g_{sw}(k+2) = \sum_{x=a,b,c} \sum_{j=1}^{3} \left|S_{xj}(k+2) - S_{xj}(k+1)\right| \tag{18}$$

Note that g_{sw} is related to the number of switching commutations that directly affect the average switching frequency of the DCI.

g_{CM} can be extracted using (2):

$$g_{CM}(k+2) = v_{CM}(k+2) \tag{19}$$

Fig. 3 illustrates the flowchart of the MPC strategy for a 4-level DCI with the prediction horizon of N=1 and delay compensation.

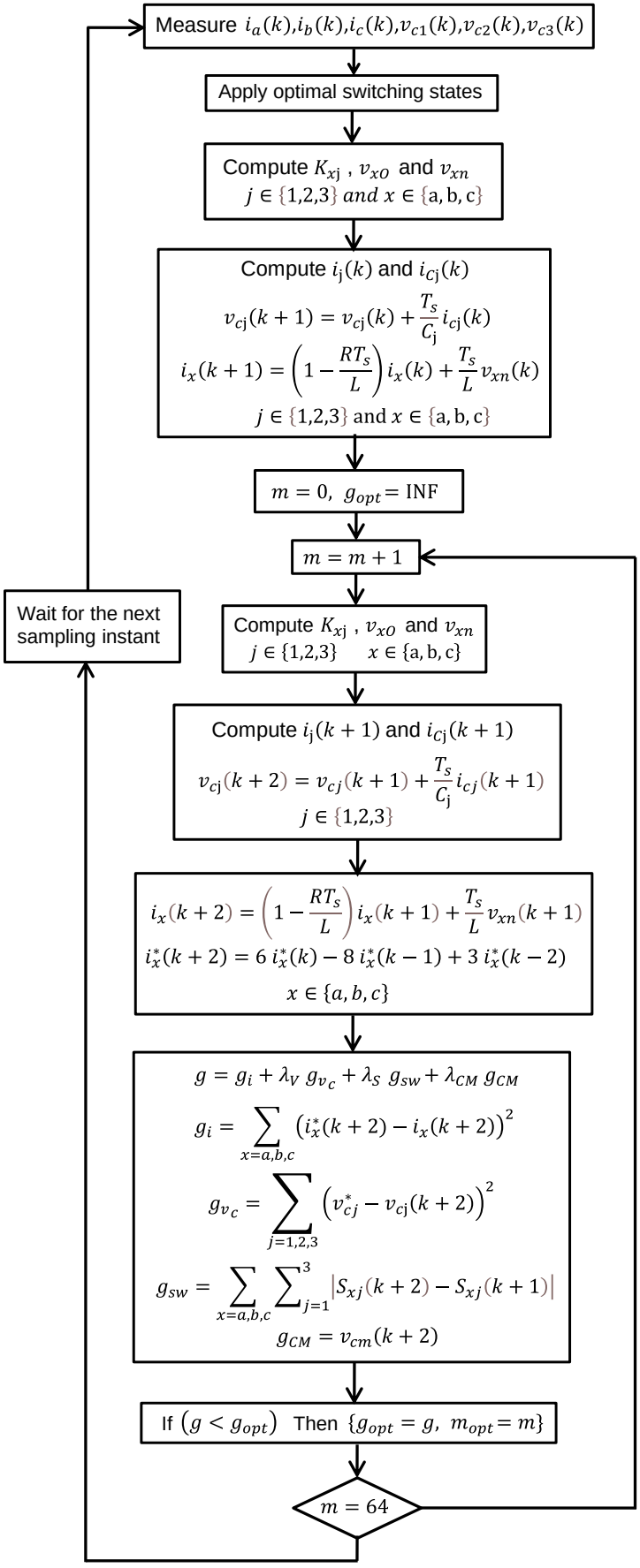


Fig. 3: Flowchart of the MPC of a 4-level DCI with a prediction horizon of N=1.

Multistep MPC in 4-Level DCI

The main aim of the multistep MPC is to predict the system behavior in more than one sampling instant. In the multistep prediction strategy, the total objective function in the horizon of N can be defined as:

$$g_{total} = \sum_{l=2}^{N+1} g(k+l) \tag{20}$$

where N is the horizon of prediction. With the prediction horizon of N=1, the prediction index is (k+2), and the objective function becomes the same as in (13).

By substituting (13)-(19) in (20), the total objective function in the horizon of N can be written as in (21). Fig. 4 represents the main part of the multistep MPC algorithm with a prediction horizon of N=3, which has the task of minimizing the objective function. All 64^3 switching possibilities will be searched.

$$g_{total} = \sum_{l=2}^{N+1} \Big(\sum_{x=a,b,c} \big(i_x^*(k+l) - i_x(k+l)\big)^2 + \lambda_V \sum_{j=1,2,3} \Big(v_{cj}^* - v_{cj}(k+l)\Big)^2 + \lambda_{sw} \sum_{x=a,b,c} \sum_{j=1}^{3} \big|S_{xj}(k+l) - S_{xj}(k+l-1)\big| + \lambda_{cm} v_{cm}(k+l) \Big) \tag{21}$$

In a specific switching condition, if the objective function g_{total} becomes lower than the optimal value g_{opt}, the switching condition will be saved as m_{opt}. The optimum switching condition m_{opt} will be applied to the DCI in the next sampling instant.

The future current references of the (k+3), (k+4), and (k+5)-th sampling instant can be calculated as [48]:

$$i_x^*(k+3) = 10\, i_x^*(k) - 15\, i_x^*(k-1) + 6\, i_x^*(k-2) \tag{22}$$

$$i_x^*(k+4) = 15\, i_x^*(k) - 24\, i_x^*(k-1) + 10\, i_x^*(k-2) \tag{23}$$

$$i_x^*(k+5) = 21\, i_x^*(k) - 35\, i_x^*(k-1) + 15\, i_x^*(k-2) \tag{24}$$

where $x \in \{a, b, c\}$.

```
for m=1:64
Compute g(k+2) = g_i(k+2) + λ_V g_vc(k+2) + λ_S g_sw(k+2) + λ_CM g_CM(k+2)
    for n=1:64
    Compute g(k+3) = g_i(k+3) + λ_V g_vc(k+3) + λ_S g_sw(k+3) + λ_CM g_CM(k+3)
        for i=1:64
        Compute g(k+4) = g_i(k+4) + λ_V g_vc(k+4) + λ_S g_sw(k+4) + λ_CM g_CM(k+4)
        g_total = g(k+2) + g(k+3) + g(k+4)
            if ( g_total < g_opt )  then (g_opt = g_total, m_opt = m)
            end if
        end for
    end for
end for
```

Fig. 4: Main part of the multistep MPC algorithm for the 4-level DCI with a prediction horizon of N=3.

Results and Discussion

The effectiveness of the suggested multistep MPC method is verified by simulating a 4-level DCI with Matlab/Simulink. The total DC-link voltage is 520 V. The DC-link capacitors are C_1=C_2=C_3=2.2 mF. The load resistance and inductance are R= 10 Ω and L= 10 mH, respectively.

Fig. 5 illustrates the simulation results in the 4-level DCI with multistep MPC for prediction horizon of N=2. The weighting factor λ_V is set to 0.5. The weighting factors λ_S and λ_{CM} are set to zero. The sampling time T_s is 50 µsec. A 50 µsec delay time has been applied to the controller for modeling the computational delay in the practical conditions. To validate the tracking performance of the multistep MPC strategy, the reference currents are changed from 10 A to 5 A at t=0.06 sec. Fig. 5(a) shows the current tracking performance in the 4-level DCI with multistep MPC. It is visible that the currents follow their references properly. Fig. 5(b) presents the line-to-line

voltage (V_{ab}) in the 4-level DCI with multistep MPC. The voltage amplitude corresponds to the step change in the load current. Fig. 5(c) shows the capacitor voltages (v_{C1}, v_{C2}, v_{C3}) in the 4-level DCI. The distortion in the voltage balance of the capacitors is low and is not affected by the step change in the reference currents. Fig. 5(d) illustrates the CM voltage (v_{cm}) in the 4-level DCI with multistep MPC for prediction horizon of N=2. The CM voltage and the switching frequency are high since they are not included in the objective function.

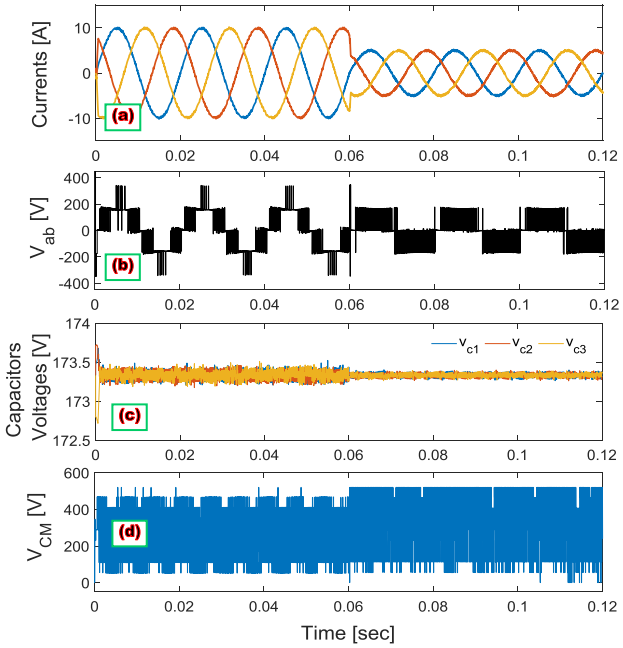


Fig. 5: Simulation results of the 4-level DCI with multistep MPC for N=2: (a) current; (b) line voltage; (c) capacitor voltages; (d) CM voltage.

Fig. 6 compares the load current, phase voltage (V_{an}), and line-to-line voltage (V_{ab}) of the MPC strategy in 2-level VSI with 4-level DCI. Simulation parameters are the same as in Fig. 5. It is obvious that the harmonic distortion is much lower in 4-level DCI than in 2-level VSI. The current THD is reduced from 15.47% in 2-level VSI to 1.82% in 4-level DCI.

Figs. 7(a)-(d) show the load current in the 4-level DCI with traditional MPC without delay compensation, single-step MPC with a horizon of N=1, and multistep MPC with a horizon of N=2 and N=3, respectively. In multistep MPC methods, the future prediction of reference currents can be obtained from (22)-(24). Simulation parameters are the same as in Fig. 5. The sampling time T_s is 100 µsec. A 100 µsec delay time has been applied to the controller. It is evident that the lowest current distortion belongs to the multistep MPC with a horizon of N=3 (as shown in Fig. 7(d)) and the highest current distortion refers to the traditional MPC without delay compensation (as shown in Fig. 7(a)).

Figs. 8(a)-(d) present the capacitor voltages in the 4-level DCI with traditional MPC without delay compensation, single-step MPC with a horizon of N=1, and multistep MPC with a horizon of N=2 and N=3, respectively. It is visible that the capacitor voltage balance is higher in the multistep MPC (as shown in Fig. 8(d)) than in the traditional MPC without delay compensation (as shown in Fig. 8(a)). The voltage balance with a horizon of N=2 (according to Fig. 8(c)) is almost similar with the horizon of N=3 (according to Fig. 8(d)). Further increase in the prediction horizon will increase the computational burden and the simulation time, while the DCI performance does not improve significantly. Thus, it is not preferable.

In the next part, the effect of different weighting factors on the system performance is investigated. The multistep MPC with a horizon of N=2 has been applied to control the 4-level DCI. The sampling time T_s is 100 µsec and a 100 µsec delay time has been applied to the controller.

Fig. 9 shows the effect of two different values of λ_V on the performance of 4-level DCI. The weighting factors λ_S and λ_{CM} are set to zero. λ_V is changed from zero to 0.5 at t=0.06 sec. Figs. 9(a)-(b) show the currents and line voltage in the 4-level DCI with multistep MPC with a horizon of N=2. The tracking performance is not affected in the steady-state condition. Fig. 9(c) presents the capacitor voltages (v_{C1}, v_{C2}, v_{C3}) in the 4-level DCI. As can be seen, increasing λ_V balances the DC link capacitor voltages. Figs. 9(d)-(e) illustrate the CM voltage and gate pulse S_{a1} in the 4-level DCI with multistep MPC. The CM voltage and the switching frequency increase significantly in higher values of the weighting factor λ_V.

Fig. 10 presents the comparative results of the system behavior with three different values of λ_S. The value of λ_S is changed from zero to 0.1 at t=0.04 sec and from 0.1 to 0.3 at t=0.08 sec. Moreover, λ_V=0.5 is selected, and λ_{CM} is set to zero. Figs. 10(a)-(b) show the currents and line voltage in the 4-level DCI with multistep MPC with a horizon of N=2. The tracking performance is not affected in different values of λ_S. Fig. 10(c) presents the capacitor voltages (v_{C1}, v_{C2}, v_{C3}) in the 4-level DCI. As can be seen, increasing λ_S results in a higher voltage unbalance. Figs. 10(d)-(e) illustrate the CM voltage and gate pulse S_{a1} in the 4-level DCI. It is visible that increasing λ_S reduces the switching frequency; however, the CM voltage is high since it is not included in the objective function.

Fig. 11 illustrates the effect of three different values of λ_{CM} on the system performance. λ_{CM} is changed from zero to 0.05 at t=0.04 sec and from 0.05 to 0.1 at t=0.08 sec. λ_V=0.5 is selected, and λ_S is set to zero. Figs. 11(a)-(b) show the currents and line voltage in the 4-level DCI with multistep MPC with a horizon of N=2. As can be seen, increasing λ_S reveals a higher harmonic distortion in the current and voltage waveforms. Fig. 11(c) presents the

capacitor voltages (v_{C1}, v_{C2}, v_{C3}) in the 4-level DCI. Reducing the CM voltage leads to a significant decrease of the voltage balance. Figs. 11(d)-(e) illustrate the CM voltage and gate pulse S_{a1} in the 4-level DCI. It is obvious from the results that increasing λ_{CM} reduces of CM voltage.

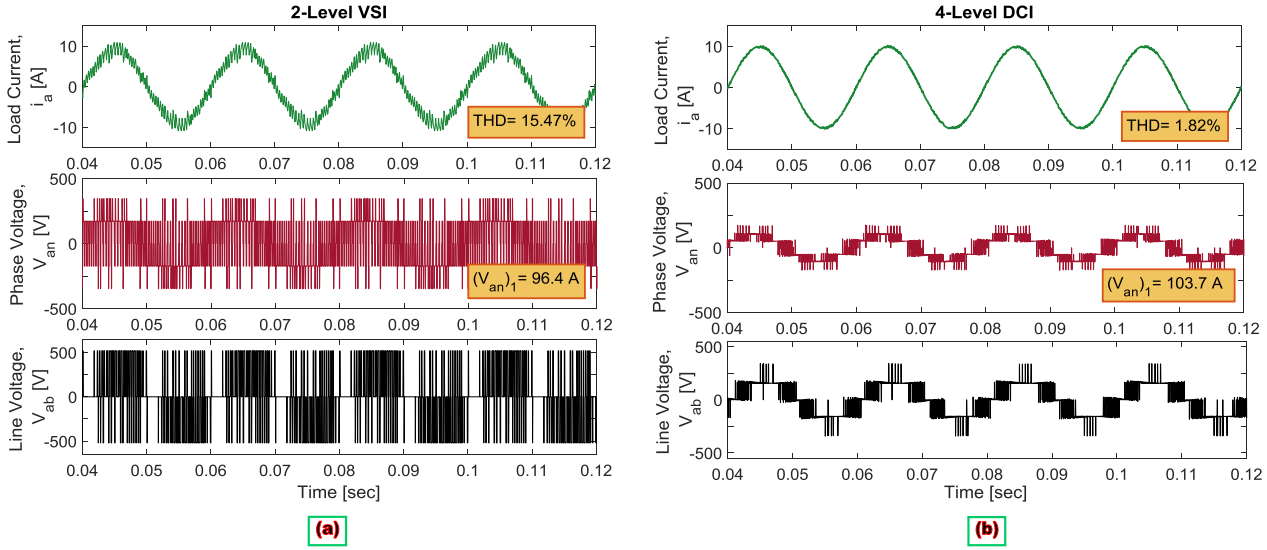


Fig. 6: Load current, phase voltage, and line voltage with MPC strategy in: (a) 2-level VSI; (b) 4-level DCI.

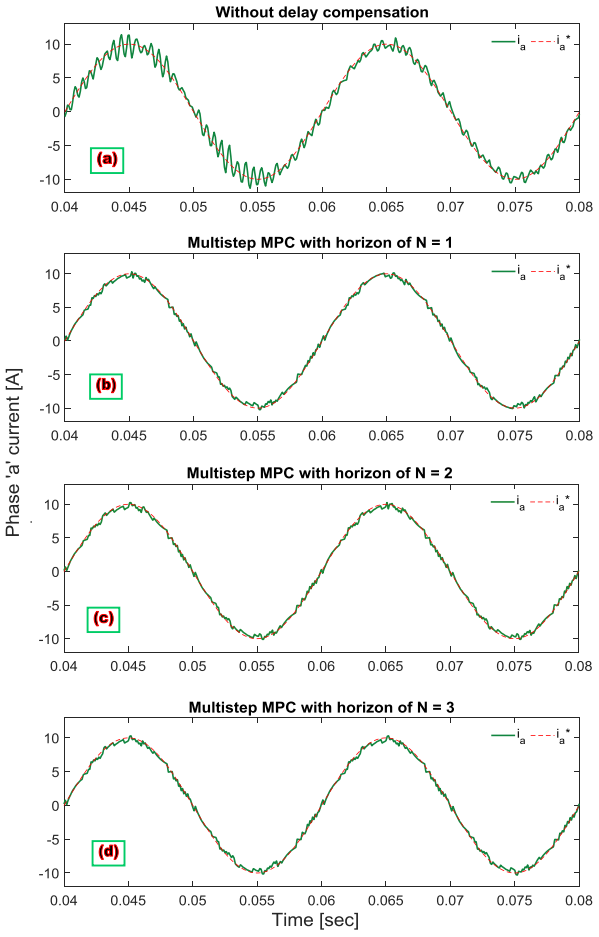


Fig. 7: Phase 'a' current in the 4-level DCI: (a) MPC without delay compensation; (b) single-step MPC with a horizon of N=1; (c) multistep MPC with a horizon of N=2; (d) multistep MPC with a horizon of N=3.

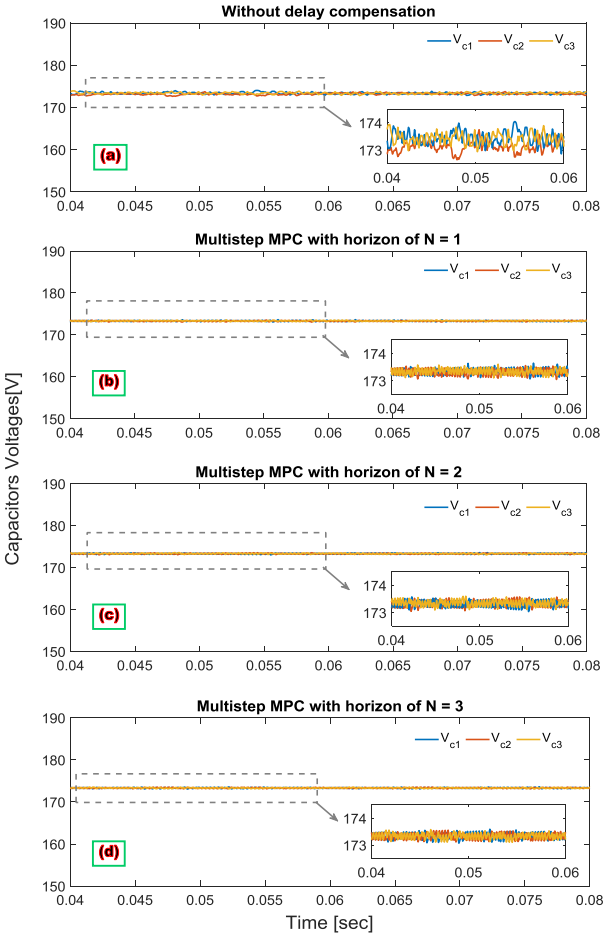


Fig. 8: Capacitor voltages in the 4-level DCI: (a) MPC without delay compensation; (b) single-step MPC with a horizon of N=1; (c) multistep MPC with a horizon of N=2; (d) multistep MPC with a horizon of N=3.

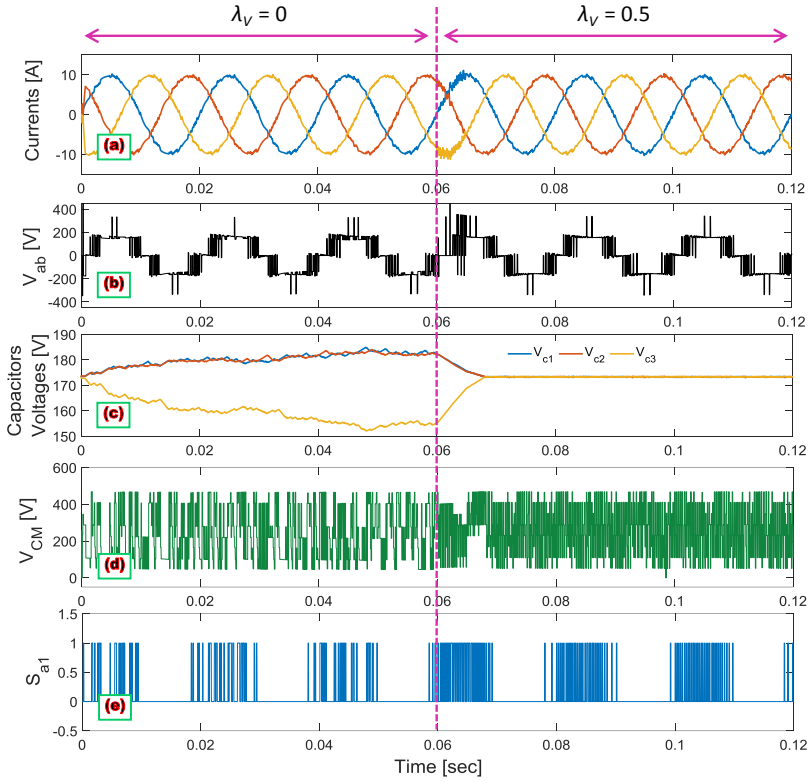


Fig. 9: Effect of weighting factor λ_V on the behavior of 4-level DCI with multistep MPC for N=2: (a) current; (b) line voltage; (c) capacitor voltages; (d) CM voltage; (e) gate pulse S_{a1}.

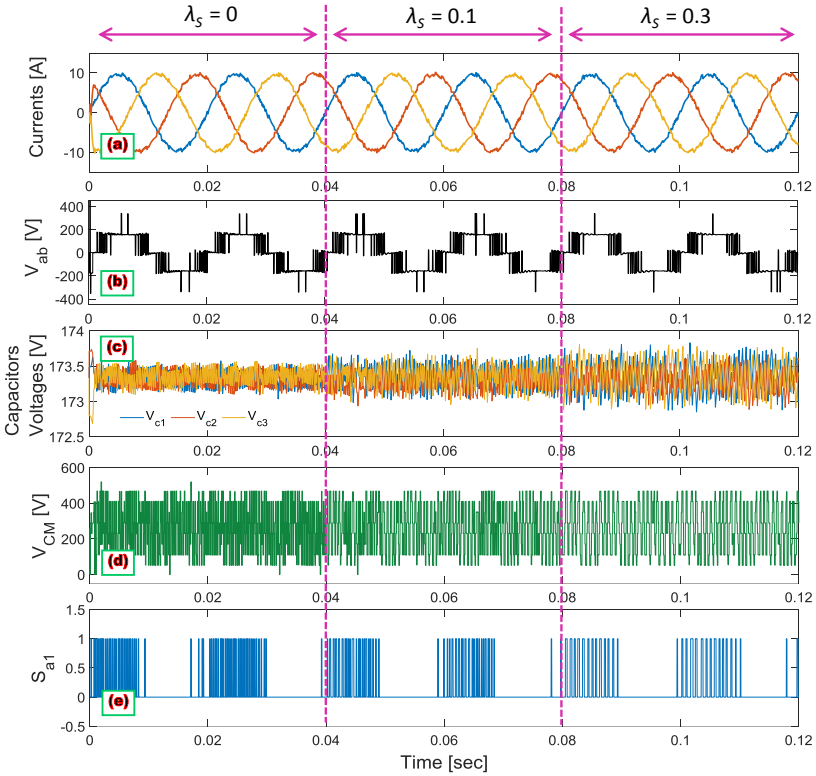


Fig. 10: Effect of weighting factor λ_S on the behavior of 4-level DCI with multistep MPC for N=2: (a) current; (b) line voltage; (c) capacitor voltages; (d) CM voltage; (e) gate pulse S_{a1}.

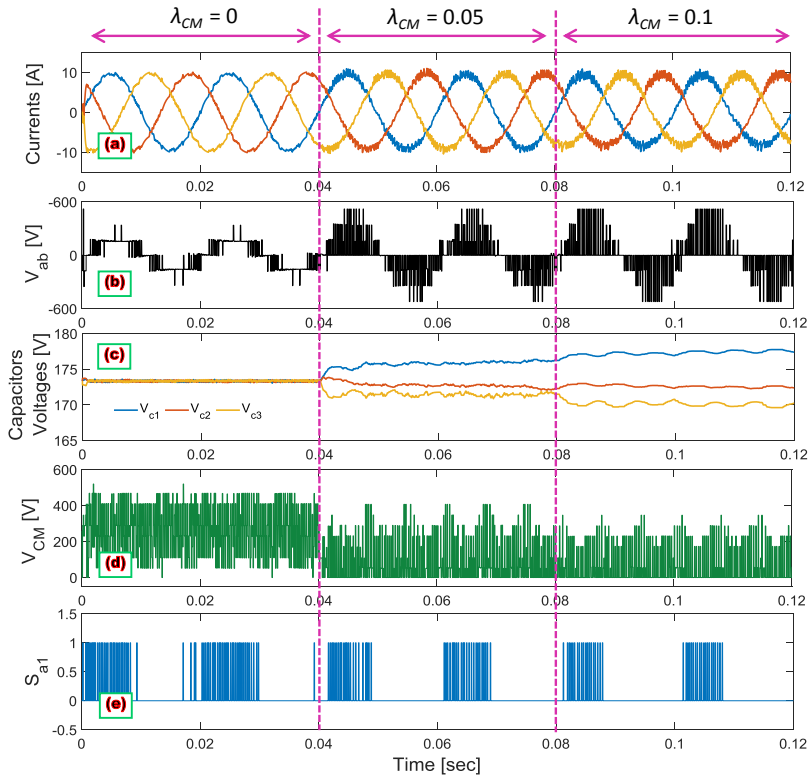


Fig. 11: Effect of weighting factor λ_{CM} on the behavior of 4-level DCI with multistep MPC for N=2: (a) current; (b) line voltage; (c) capacitor voltages; (d) CM voltage; (e) gate pulse S_{a1}.

Conclusion

This work has proposed a multistep predictive current control for the 3-phase 4-level DCI. The suggested method has succeeded in controlling the load current, while other objectives were easily included in the predictive controller. In addition to the current tracking, this paper has evaluated the DC-link voltage balance, reduction of the CM voltage, and decreasing the switching frequency in the prediction strategy. In this regard, the multistep prediction control with different horizon lengths has been applied to the 4-level DCI. Moreover, the effect of different values of weighting factors has been studied on the system performance.

Simulation results have revealed the excellent dynamic response and DC-link voltage balancing in the 4-level DCI controlled by the multistep predictive method with a horizon of N=2. However, in long prediction horizons, the simulation time and computational burden will increase. Therefore, current tracking quality and voltage balancing may not be obtained. On the other hand, decreasing the CM voltage and the switching frequency has the opposite effect on the current tracking quality and voltage balancing in the DCI. Generally, dynamic response and voltage balancing are the main requirements of the DCI. Therefore, a trade-off will be imposed when selecting the weighting factors of the objective function depending on the system requirements. The future work will focus on the multistep model predictive control of motor drives supplied with 4-level diode-clamped inverter including various objectives in the predictive controller. Furthermore, the effect of different values of weighting factors on the system performance will be investigated.

Author Contributions

P. Hamedani carried out the simulations, interpreted the results, and wrote the manuscript.

Acknowledgment

The author gratefully acknowledges the respected reviewers and the editor of JECEI for their helpful comments and accurate reviewing of this paper.

Conflict of Interest

The author declares no potential conflict of interest regarding the publication of this work.

Abbreviations

CM	Common Mode
DCI	Diode-Clamped Inverter
MLC	Multilevel Converter
MLI	Multilevel Inverter
MPC	Model Predictive Control
THD	Total Harmonic Distortion
VSI	Voltage Source Inverter

References

[1] J. Rodriguez, J. S. Lai, F. Z. Peng, "Multilevel inverters: A survey of topologies, controls, and applications," IEEE Trans. Ind. Electron., 49(4): 724-738, 2002.

[2] J. Rodriguez, L. G. Franquelo, S. Kouro, J. I. Leon, R. C. Portillo, M. A. M. Prats, M. A. Perez, "Multilevel converters: An enabling technology for high-power applications," Proc. IEEE, 97(11): 1786-1817, 2009.

[3] L. Tolbert, F. Z. Peng, T. Habetler, "Multilevel converters for large electric drives," IEEE Trans. Ind. Electron., 35(1): 36-44, 1999.

[4] H. Rudnick, J. Dixon, L. Moran, "Delivering clean and pure power," IEEE Power Energy Mag., 1(5): 32-40, 2003.

[5] S. Enyedi, "Electric cars—Challenges and trends," in Proc. IEEE 2018 International Conference on Automation, Quality and Testing, Robotics (AQTR): 1-8, 2018.

[6] H. Schefer, L. Fauth, T. H. Kopp, R. Mallwitz, J. Friebe, M. Kurrat, "Discussion on electric power supply systems for all electric aircraft," IEEE Access, 8: 84188-84216, 2020.

[7] C. Jung, "Power up with 800-V systems: The benefits of upgrading voltage power for battery-electric passenger vehicles," IEEE Electrific. Mag., 5(1): 53-58, 2017.

[8] P. Hamedani, M. Changizian, "A New hybrid predictive-PWM control for flying capacitor multilevel inverter," J. Electr. Comput. Eng. Innovations, 12(2): 353-362, 2024.

[9] J. Rodriguez, S. Bernet, B. Wu, J. O. Pontt, S. Kouro, "Multilevel voltage-source-converter topologies for industrial medium-voltage drives," IEEE Trans. Ind. Electron., 54(6): 2930-2945, 2007.

[10] P. Hamedani, A. Shoulaei, "Utilization of CHB multilevel inverter for harmonic reduction in fuzzy logic controlled multiphase LIM drives," J. Electr. Comput. Eng. Innovations, 8(1): 19-30, 2020.

[11] P. Hamedani, A. Shoulaei, "A comparative study of harmonic distortion in multicarrier based PWM switching techniques for cascaded H-bridge inverters," Adv. Electr. Comput. Eng., 16(3): 15-24, 2016.

[12] B. P. McGrath, D. G. Holmes, "Multicarrier PWM strategies for multilevel inverters," IEEE Trans. Ind. Electron., 49(4): 858-867, 2002.

[13] N. Celanovic, D. Boroyevich, "A fast space-vector modulation algorithm for multilevel three-phase converters," IEEE Trans. Ind. Appl., 37(2): 637-641, 2001.

[14] J. I. Vazquez, A. J. Watson, L. G. Franquelo, P. W. Wheeler, J. M. Carrasco, "Feed-forward space vector modulation for single-phase multilevel cascaded converters with any dc voltage ratio," IEEE Trans. Ind. Electron., 56(2): 315-325, 2009.

[15] J. Rodriguez, J. Pontt, P. Correa, P. Cortes, C. Silva, "A new modulation method to reduce common-mode voltages in multilevel inverters," IEEE Trans. Ind. Electron., 51(4): 834-839, 2004.

[16] Y. Liu, H. Hong, A. Huang, "Real-time calculation of switching angles minimizing THD for multilevel inverters with step modulation," IEEE Trans. Ind. Electron., 56(2): 285-293, 2009.

[17] Z. Du, L. M. Tolbert, J. N. Chiasson, B. Ozpineci, "Reduced switching-frequency active harmonic elimination for multilevel converters," IEEE Trans. Ind. Electron., 55(4): 1761-1770, 2008.

[18] J. Rodriguez et al., "Latest advances of model predictive control in electrical drives—part I: Basic concepts and advanced strategies," IEEE Trans. Power Electr., 37(4): 3927-3942, 2022.

[19] P. Hamedani, S. Sadr, "Model predictive control of linear induction motor drive with end effect consideration," J. Electr. Comput. Eng. Innovations, 11(2): 253-262, 2023.

[20] P. Hamedani, C. Garcia, F. Flores-Bahamonde, S. Sadr, J. Rodriguez, "Predictive control of 4-level flying capacitor inverter for electric car applications," presented at the 13th Power Electronics, Drive Systems, and Technologies Conference (PEDSTC): 224-229, 2022.

[21] J. Rodriguez et al., "Latest advances of model predictive control in electrical drives—part II: Applications and benchmarking with classical control methods," IEEE Trans. Power Electr., 37(5): 5047-5061, 2022.

[22] S. Kouro, P. Cortes, R. Vargas, U. Ammann, J. Rodriguez, "Model predictive control—a simple and powerful method to control power converters," IEEE Trans. Ind. Electr., 56(6): 1826-1838, 2009.

[23] J. Rodriguez, M. P. Kazmierkowski, J. R. Espinoza, P. Zanchetta, H. Abu-Rub, H. A. Young, C. A. Rojas, "State of the art of finite control set model predictive control in power electronics," IEEE Trans. Ind. Inf., 9(2): 1003-1016, 2013.

[24] S. Vazquez, J. Rodriguez, M. Rivera, L. G. Franquelo, M. Norambuena, "Model predictive control for power converters and drives: Advances and trends," IEEE Trans. Ind. Electr., 64(2): 935-947, 2017.

[25] P. Karamanakos, E. Liegmann, T. Geyer, R. Kennel, "Model predictive control of power electronic systems: Methods, results, and challenges," IEEE Open J. Ind. Appl., 1: 95-114, 2020.

[26] J. O. Krah, T. Schmidt, J. Holtz, "Predictive current control with synchronous optimal pulse patterns," in Proc. IEEE 2nd International Conference on Smart Grid and Renewable Energy (SGRE), 2019.

[27] T. Geyer, G. Papafotiou, M. Morari, "Model predictive direct torque control—part I: Concept, algorithm, and analysis," IEEE Trans. Ind. Electr., 56(6): 1894-1905, 2009.

[28] M. F. Elmorshedy, W. Xu, F. F. M. El-Sousy, M. R. Islam, A. A. Ahmed, "Recent achievements in model predictive control techniques for industrial motor: A comprehensive state-of-the-art," IEEE Access, 9: 58170-58191, 2021.

[29] J. O. Krah, T. Schmidt, J. Holtz, "Predictive current control with synchronous optimal pulse patterns," in Proc. IEEE 2nd International Conference on Smart Grid and Renewable Energy (SGRE), 2019.

[30] T. Geyer, G. Papafotiou, M. Morari, "Model predictive direct torque control—part I: Concept, algorithm, and analysis," IEEE Trans. Ind. Electr., 56(6): 1894-1905, 2009.

[31] M. F. Elmorshedy, W. Xu, F. F. M. El-Sousy, M. R. Islam, A. A. Ahmed, "Recent achievements in model predictive control techniques for industrial motor: A comprehensive state-of-the-art," IEEE Access, 9: 58170-58191, 2021.

[32] G. Darivianakis, T. Geyer, W. van der Merwe, "Model predictive current control of modular multilevel converters," in Proc. IEEE Energy Conversion Congress and Exposition (ECCE), 2014.

[33] M. Najjar, M. Shahparasti, R. Heydari, M. Nymand, "Model predictive controllers with capacitor voltage balancing for a single-phase five-level SiC/si based ANPC inverter," IEEE Open J. Power Electr., 2: 202-211, 2021.

[34] J. Raath, T. Mouton, T. Geyer, "Alternative sphere decoding algorithm for long-horizon model predictive control of multi-level inverters," in Proc. IEEE 21st Workshop on Control and Modeling for Power Electronics (COMPEL), 2020.

[35] K. Bandy, P. Stumpf, "Model predictive torque control for multilevel inverter fed induction machines using sorting networks," IEEE Access, 9: 13800-13813, 2021.

[36] M. Wu, H. Tian, Y. W. Li, G. Konstantinou, K. Yang, "A composite selective harmonic elimination model predictive control for seven-level hybrid-clamped inverters with optimal switching patterns," IEEE Trans. Power Electr., 36(1): 274-284, 2021.

[37] M. Aly, F. Carnielutti, M. Norambuena, S. Kouro, J. Rodriguez, "A model predictive control method for common grounded

photovoltaic multilevel inverter," in Proc. IEEE IECON 46th Annual Conference of the IEEE Industrial Electronics Society, 2020.

[38] A. G. Beccuti, S. Mariethoz, S. Cliquennois, S. Wang, M. Morari, "Explicit model predictive control of dc–dc switched-mode power supplies with extended Kalman filtering," IEEE Trans. Ind. Electron., 56(6): 1864-1874, 2009.

[39] M. Cychowski, K. Szabat, T. Orlowska-Kowalska, "Constrained model predictive control of the drive system with mechanical elasticity," IEEE Trans. Ind. Electron., 56(6): 1963-1973, 2009.

[40] P. Cortes, J. Rodriguez, S. Alepuz, S. Busquets-Monge, J. Bordonau, "Finite-states model predictive control of a four-level diode-clamped inverter," in Proc. IEEE Power Electronics Specialists Conference: 2203-2208, 2008.

[41] V. Yaramasu, B. Wu, M. Rivera, J. Rodriguez, "A new power conversion system for megawatt PMSG wind turbines using four-level converters and a simple control scheme based on two-step model predictive strategy—part I: Modeling and theoretical analysis," IEEE J. Emerging Sel. Top. Power Electr., 2(1): 3-13, 2014.

[42] V. Yaramasu, B. Wu, "Predictive control of a three-level boost converter and an NPC inverter for high-power PMSG-based medium voltage wind energy conversion systems," IEEE Trans. Power Electr., 29(10): 5308-5322, 2014.

[43] V. Yaramasu, B. Wu, "Model predictive decoupled active and reactive power control for high-power grid-connected four-level diode-clamped inverters," IEEE Trans. Ind. Electr., 61(7): 3407-3416, 2014.

[44] V. Yaramasu, B. Wu, J. Chen, "Model-predictive control of grid-tied four-level diode-clamped inverters for high-power wind energy conversion systems," IEEE Trans. Power Electr., 29(6): 2861-2873, 2014.

[45] J. G. Ordonez, D. Limon, F. Gordillo, "Multirate predictive control for diode clamped inverters with data-based learning implementation," IFAC-PapersOnLine, 56(2): 6388-6393, 2023.

[46] V. Yaramasu, A. Dekka, M. Rivera, S. Kouro, J. Rodriguez, "Multilevel inverters: Control methods and advanced power electronic applications," Academic Press, 2021.

[47] R. Atif et al., "Simplified model predictive current control of four-level nested neutral point clamped converter," Sustainability, 15(2): 955, 2023.

[48] J. Rodriguez, P. Cortes, Predictive control of power converters and electrical drives, John Wiley & Sons, 2012.

[49] E. Kabalcı, Multilevel Inverters Control Methods and Advanced Power Electronic Applications, Elsevier Science, 2021.

Biography

Pegah Hamedani was born in Isfahan, Iran, in 1985. She received B.Sc. and M.Sc. degrees from University of Isfahan, Iran, in 2007 and 2009, respectively, and the Ph.D. degree from Iran University of Science and Technology, Tehran, in 2016, all in Electrical Engineering. Her research interests include power electronics, control of electrical motor drives, supply system of the electric railway (AC and DC), linear motors & MAGLEVs, and analysis of overhead contact systems. She is currently an Assistant Professor with the Department of Railway Engineering and Transportation Planning, University of Isfahan, Isfahan, Iran. Dr. Hamedani was the recipient of the IEEE 11th Power Electronics, Drive Systems, and Technologies Conference (PEDSTC'20) best paper award in 2020.

- Email: p.hamedani@eng.ui.ac.ir
- ORCID: 0000-0002-5456-1255
- Web of Science Researcher ID: AAN-2662-2021
- Scopus Author ID: 37118674000
- Homepage: https://engold.ui.ac.ir/~p.hamedani/

How to cite this paper:
P. Hamedani, "A new hybrid predictive-PWM control for flying capacitor multilevel inverter," J. Electr. Comput. Eng. Innovations, 13(1): 117-128, 2025.

DOI: 10.22061/jecei.2024.10914.747

URL: https://jecei.sru.ac.ir/article_2196.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Electric Vehicle Battery Charging Using a Non-Isolated Bidirectional DC-DC Converter Connected to T-Type Three Level Converter

F. Sedaghati [*], ***S. A. Azimi***

Department of Electrical Engineering, Faculty of Engineering, University of Mohaghegh Ardabili, Ardabil, Iran.

Article Info

Article History:
Received 30 June 2024
Reviewed 27 July 2024
Revised 07 September 2024
Accepted 14 September 2024

Keywords:
Electric vehicle
Battery charger
Non-isolated bidirectional dc-dc converter
T-type three-level converter
Charging station

[*]Corresponding Author's Email Address:
farzad.sedaghati@uma.ac.ir

Abstract

Background and Objectives: Increasing environmental problems have led to the spread of Electric Vehicles (EVs). One of the attractive research fields of electric vehicles is the charging battery of this strategic product. Electric vehicle battery chargers often lack bidirectional power flow and the flexibility to handle a wide range of battery voltages. This study proposes a non-isolated bidirectional DC-DC converter connected to a T-type converter with a reduced number of switches to solve this limitation.

Methods: The proposed converter uses a DC-DC converter that has an interleaved structure along with a three-level T-type converter with a reduced number of switches and a common ground for the input and output terminals. Space vector pulse width modulation (SVPWM) and carrier based sinusoidal pulse width modulation (CBPWM) control the converter for Vehicle to grid (V2G) and grid to Vehicle (G2V) operation, respectively.

Results: Theoretical analysis shows 96.9% efficiency for 15.8kW output power and 3.06% THD during charging with low battery voltage ripple. In V2G mode, it achieves an efficiency of 96.5% while injecting 0.5 kW of power into the 380 V 50 Hz grid. The DC link voltage is stabilized. The proposed converter also provides good performance for a wide range of battery development.

Conclusion: The proposed converter offers high efficiency and cost reduction. It provides the possibility of charging a wide range of batteries and provides V2G and G2V power flow performance. The proposed converter is capable of being placed in the fast battery charging category. The ability to charge two batteries makes it a suitable option for charging stations.

Introduction

Electric vehicles (EV) play an important role in transportation and automotive related markets. The expansion of electric vehicles (EVs) is an ongoing trend in today's society, with the EV market growing at a very fast rate [1]. Therefore, to fully employ the potentially great number of EVs, proper charging infrastructure is a necessity [2]. This is especially crucial in terms of the ability to charge the EVs rapidly, e.g., on highways, where the utilization of highly-performant fast and ultra-fast charging stations is required. There is a large variety of approaches that can be employed to construct fast charging stations that differ in the voltage levels, the presence of additional battery energy storage, as well as the grid structure (unipolar vs. bipolar) [3], [4]. Here, an EV charging system with a bipolar DC grid with +/- 750 V and extra battery energy storage is considered, as such a system is considered advantageous compared to more conventional approaches [5]-[7]. Increasing demand in the transportation industry with electric vehicles requires a suitable charging infrastructure for this demand.

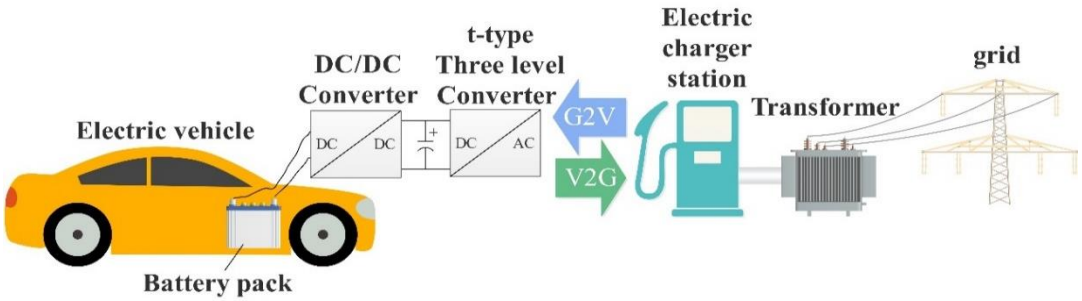


Fig. 1: Process of G2V and V2G.

However, the cost and charging time of the battery in an electric vehicle are two main issues that continue to challenge the wider application of EVs [8]. Both issues are directly related to the charging system for EVs. For example, compared to the fossil fuel station, in EV charging stations, the charging time with the battery is longer than recharging an internal combustion car, which can be balanced in the time and construction costs with the optimal design of the charger circuits. Therefore, for the development of EV applications, it is necessary to improve the charger system, especially for the off-board fast charger. In addition to solving the challenges of battery charging time and manufacturing cost, the optimal design of charger circuits should also provide power flow in two directions (V2G) and (G2V). Fig. 1 shows the process of (V2G) and (G2V). The power density can be increased by increasing the switching frequency and integrating the DC-DC converter together in an interleaved form, magnetic integration technique and introducing a single-stage approach [9]. When the battery interface converter required to connect the station's DC-link and the battery energy storage is considered, a number of topologies can be used [10], [11]. The three-level topology provides the possibility to employ well-performing off-the-shelf 1.2 kV SiC MOSFETs and obtain lower power losses compared to other approaches [12], as well as the option to balance the DC grid [13] with low general complexity. Moreover, the interleaved two phase topology shows good perspective in terms of low output ripples [14], [15], especially important for cooperating with a battery energy storage.

Some charger topologies are suitable for a small range of output voltages. However, advanced EV chargers with wide output voltage ranges have been discussed in [16].

In [17], an EV charger based on Vienna converter is presented.

The wide output voltage range is the main advantage of the topology, stated by the authors. It is noteworthy that most existing chargers, regardless of their single or two-stage power processing structures, use line frequency or high frequency transformers. However, transformers not only affect the cost and size of the charger but also, increase the voltage stress across semiconductor devices and reduce efficiency. Due to these disadvantages and motivated by the concept of integrated chargers [18], few researchers have demonstrated the configuration of transformer less chargers for EVs applications [19]. Researchers in [20]-[22] introduced some converters for wide ranges of output voltage in order to cover the wide variety of battery voltage from different car manufacturers. On-board battery chargers are generally lightweight and compact and have less power. But in electric charging stations, they are bulky chargers which have high weight. Therefore, the Off-board battery charging time is less than the On-board type [23], [24]. Electric vehicle batteries can help the reliability and stability of the power grid by storing energy in times of low demand and delivering power in times of peak load to the power grid.

Therefore, bidirectional chargers are important parts in these system [25]-[27].

The most advanced unidirectional battery chargers are focused on increasing power density, which has advantages such as high efficiency. The topology presented in [28] is suitable for a very high voltage battery where the aforementioned EV charger is unidirectional. Existing charging solutions for EVs with very low power factor sometimes work below 0.85. In addition, it makes the total harmonic distortion of the supply current worse [29]. So, at the full load of the charging stations, it

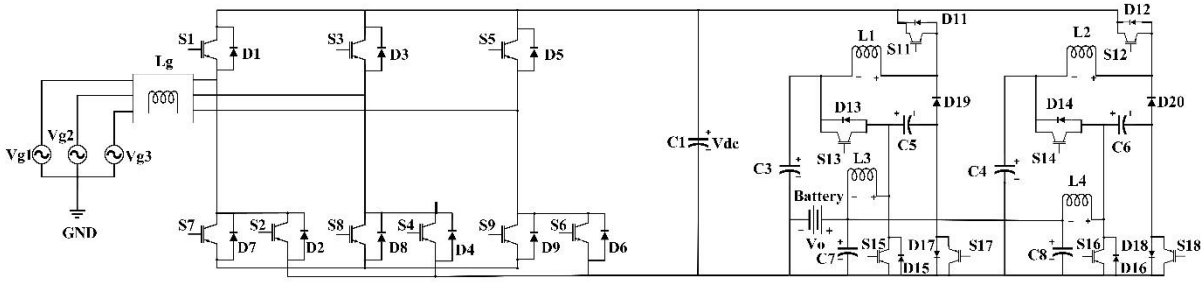


Fig. 2: Proposed charger configuration.

affects other local loads. High efficiency is required to reduce heatsink size and increase mileage per charge.

To reduce conduction losses at high load current, DC-DC converters are designed in parallel configuration for EV [30]. Due to limitation of voltage increase in the switches because of addition of leakage inductances and magnetic interference in the EMI filter and the large number of switches in the circuit, which increases the cost and reduce the efficiency of power converters. Having a common ground between the input and output ports avoids additional d*v*/d*t* problem, which is beneficial for the performance of the power converters.

This paper proposes a configuration for use in V2G and G2V applications using a bidirectional converter as an interface. The suggested configuration has good voltage gain, the low power stress in semiconductor devices, and simple structure for implementation. The converter can use switches with low rated power due to the interleaved configuration also, the proposed structure can be suitable for battery charging station and have good performance. Covering a wide range of battery voltage is one of the features of the DC-DC converter. The battery charger can charge a 380V and 40 Amp-Hour battery from 20% to 80% of State of Charge (SOC) in less than 38 minutes. The proposed configuration includes T-type three-level reduced switches converter, which is a two-way power flow AC-DC converter. In section 3, the proposed configuration is described in detail. Section 4 illustrates DC-DC converter topology and its operation. In the next sections, the control system is studied and then, the results obtained from simulations are reviewed. Finally, the last section includes conclusions.

Configuration of Proposed Charger

Fig. 2 shows the proposed charger configuration. A T-type three-level PWM converter is employed for grid-connection application. On the AC side, the three-phase voltage source is connected to the output of the three-level T-type PWM converter through an inductor filter, L_g. The AC side converter consists of nine power switches, which has three switches less than the conventional T-type three-level PWM converter. In the DC link, a capacitor whose voltage is fixed at 600 V is used. When power flows from the power grid to the DC side, switches S2, S4, and S6 are off. Also, when power flows from the DC side to the power grid, switches S7, S8 and S9 are off.

A. G2V Operation

For the power flow from the power grid to DC side, the switches of one leg are turned on and off complementary using sinusoidal pulse width modulation (SPWM) technique. Fig. 3 shows the simplified schematic of the proposed converter control, where i_{abc} and V_{abc} are the current and voltage of the three phases input to the converter, V_{ref} is the DC link reference voltage and V_{dc} is the DC link voltage. The ratio of the carrier frequency to the main frequency is larger, therefore, the main component of the output voltage changes linearly with the reference voltage. Also, the output voltage frequency is equal to the reference voltage. V_{ref} for a constant DC link voltage from the following equation:

$$v_{out}=v_{ref}\left(\sin(\omega t)+\sin\left(\omega t+\frac{2\pi}{3}\right)+\sin\left(\omega t+\frac{4\pi}{3}\right)\right) \quad (1)$$

The output voltage can be written in term of modulation index MI, as follows:

$$v_{out}=\frac{V_{dc}}{2}MI\left(\sin(\omega t)+\sin\left(\omega t+\frac{2\pi}{3}\right)+\sin\left(\omega t+\frac{4\pi}{3}\right)\right) \quad (2)$$

So, $V_{ref} < V_{dc}/2$ and $0 < MI < 1$. SPWM, which are used to stabilize the DC link voltage at 600 volts and transfer power from the grid to DC. Voltages A_n, B_n and C_n vary between two values $V_{dc}/2$ and $-V_{dc}/2$. In this mode, T-type three-level PWM converter operates like the Vienna rectifier, and diodes D1, D2, D3, D4, D5 and D6 conduct. Also, switches S1, S2, S3, S4, S5 and S6 will be turned off. It should be noted that turning on the top three switches in each leg reduces the voltage of the current passing through the top diodes of each leg.

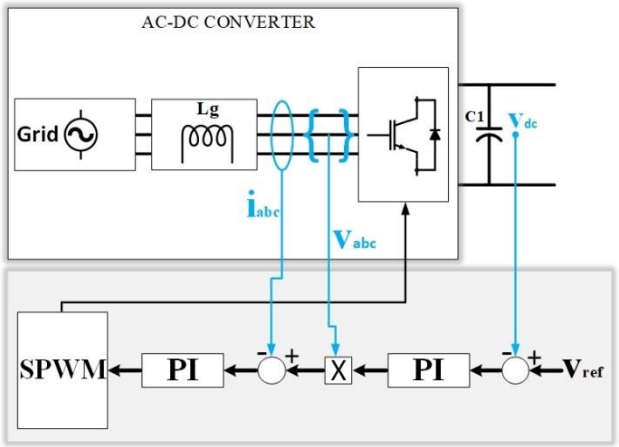


Fig. 3: Simplified schematic of AC-DC Converter with controller for G2V operation.

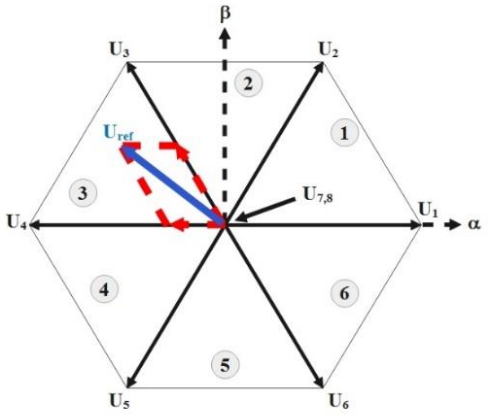


Fig. 4: Space vector diagram.

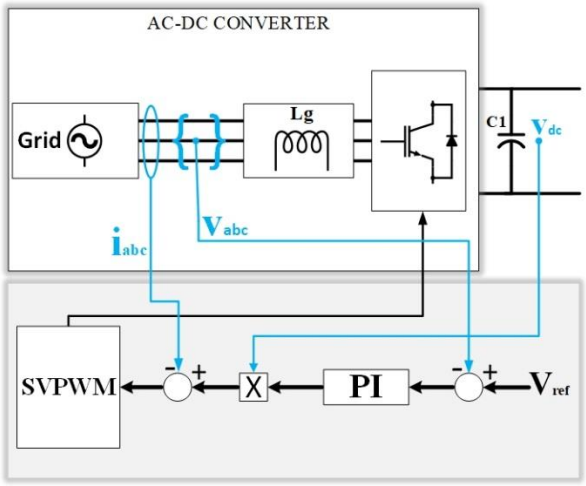


Fig. 5: Simplified schematic of AC-DC Converter with controller for V2G operation.

Table 1: State of top switches

Vector	S1	S3	S5
U1	1	0	0
U2	1	1	0
U3	0	1	0
U4	0	1	1
U5	0	0	1
U6	1	0	1
U7	0	0	0
U8	1	1	1

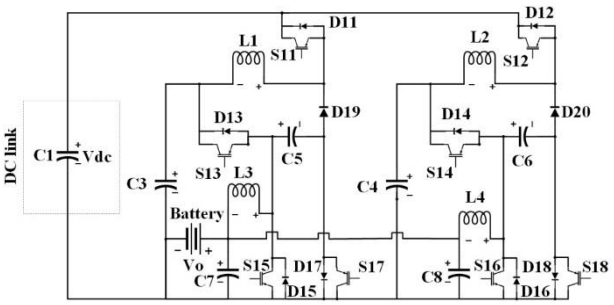


Fig. 6: Proposed DC-DC Conversion system.

B. V2G Operation

In this working mode, power flows from the electric vehicle to the grid. Therefore, the performance of the converter will be the same as a 3-phase inverter. A constant voltage to a three-phase sinusoidal voltage should be provided for the network. Switches S1-S6 are active in directing power from the DC side to the grid.

Space Vector Pulse Width Modulation (SVPWM) control method is applied to reduce switching losses, reduce harmonic distortion and also, use DC link terminal voltage properly. SVPWM can be used to generate pulses for three-phase two-level DC-AC converters. the reference vector U_{ref} is averaged using two adjacent space vectors (U3 and U4 in the Fig. 4) for a given period and a null vector (U7 or U8) for the rest of the period. Fig. 5 shows a simple schematic and controller for V2G operation of T-type converter. Table 1 shows the state of the top switches of each leg for each vector. Eight switching modes, including six active modes and two zero modes, are available. These vectors form a hexagon (Figure 4), which can be seen as consisting of six sections at 60 degrees. The reference vector representing the three-phase sinusoidal voltage is generated using SVPWM by switching between the two nearest active vectors and the zero vector. The sinusoidal reference space vector forms a circular path inside the hexagon. The largest output voltage value that can be obtained using SVPWM is the radius of the largest circle that can be recorded in the hexagon. This circle is tangent to the midpoints of the lines that join the ends of the active space vector. Finally, the model of a three-phase inverter based on space vector representation enables the proposed converter to deliver power from the battery to the grid.

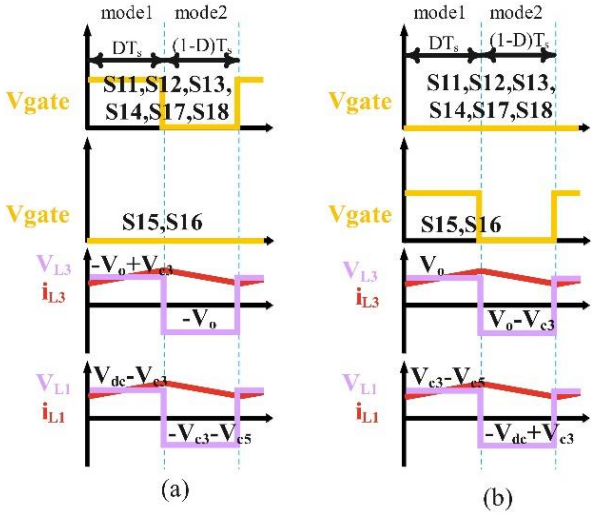


Fig. 7: Time-domain waveforms in CCM: (a) G2V, (b) V2G.

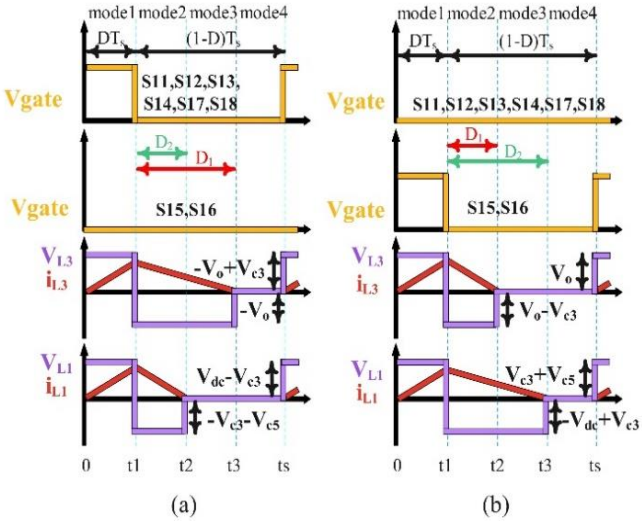


Fig. 8: Time-domain waveforms in DCM operation: (a) G2V, (b) V2G.

DC-DC Conversion

Fig. 6 shows the configuration of the proposed DC-DC conversion system. The DC-DC converter is connected to battery on one side and to the DC link on the other side. V_O and V_{dc} show the battery and the DC link voltages, respectively. Interleaved DC-DC converter is proposed to reduce the voltage and current stress and, increase the reliability of the DC-DC conversion. Therefore, we will only examine the first part of the DC-DC converter. The proposed converter consists of four inductors L1, L2, L3, and L4, four capacitors C1, C2, C3, and C4, eight switches S11, S12, S13, S14, S15, S16, S17, and S18 and ten diodes D11, D12, D13, D14, D15, D16, D17, D18, D19 and D20. Body switches and diodes conduct complementary during a complete switching period (Ts). The V2G and G2V working modes of the DC-DC converter in continuous conduction mode (CCM) and discontinuous conduction mode (DCM) are shown in Fig. 7 and Fig. 8, and the detailed analysis of each mode is given in the following.

A. G2V Operation

Two modes for CCM operation and four modes for DCM are defined as follow:

The first mode of CCM and DCM [0-t_1]: according to Fig. 9(a), switches S11, S13 and S17 are turned on and switch S15 is off. In this time interval, inductor L1 is charged by the input DC link and the energy released from capacitor C3. Therefore, the current through inductor L3 increases, while inductor L3 is energized from C5. The derived current and voltage equations according to the time-domain waveform in Fig. 8(a) are:

$$\begin{cases} v_{L1}=L_1\dfrac{di_{L1}}{dt}=V_{dc}\text{-}v_{C3}=V_{dc}\text{-}v_{C5} \\ v_{L3}=L_3\dfrac{d_{L3}}{dt}=\text{-}V_o+v_{C3}=\text{-}V_o+v_{C5} \end{cases} \quad (3)$$

the voltage of two capacitors C3 and C5 are equal.

The second mode of CCM [t1-Ts] and DCM [t1-t2]: unlike the first mode, according to Fig. 9(b), while diodes D15 and D19 conduct, switches S11, S13 and S15 are turned off. Inductor L1 gives its energy to capacitors C3 and C5. Also, the energy of inductor L3 is discharged.

$$\begin{cases} v_{L1}=L_1\dfrac{di_{L1}}{dt}=\text{-}v_{C3}\text{-}v_{C5} \\ v_{L3}=L_3\dfrac{di_{L3}}{dt}=\text{-}V_o \end{cases} \quad (4)$$

Applying volt-second balance law on inductors L1 and L2 yields:

$$D(V_{dc}\text{-}V_{C3})+(1\text{-}D)(\text{-}V_{C3}\text{-}V_{C5})=0 \quad (5)$$

$$D(\text{-}V_o+V_{C3})+(1\text{-}D)(\text{-}V_o)=0 \quad (6)$$

D stands for duty cycle. Using (6), the average voltage on the capacitors C3 and C5 are calculated as follows:

$$V_{C3}=V_{C5}=V_o/D \quad (7)$$

The voltage conversion ratio of the proposed converter during CCM operation for G2V mode is obtained from (6) and (7) as follows:

$$M_{G2V\,(CCM)}=\frac{V_o}{V_{dc}}=\frac{D^2}{2-D} \quad (8)$$

The third mode of DCM [t2-t3]: in this state, the current passing through inductor L1 in t2 and the current passing through inductor L3 in t3 reach zero.

The fourth mode of DCM [t3-Ts]: in this state, the current through the inductors reaches zero and all the switches are off. A full cycle of Ts is completed at the end of this interval. Diodes D1 and D2 can be defined as duty cycles where the current through the inductors becomes zero. Therefore, according to the Fig. 9(c), the voltages across the inductors are given as follow:

$$V_{L1}=\begin{cases} V_{dc}\text{-}v_{C3} & 0\le t<DT_s \\ \text{-}v_{C3}\text{-}v_{C5} & DT_s\le t<(D+D_2)T_s \\ 0 & (D+D_2)T_s\le t<T_s \end{cases} \quad (9)$$

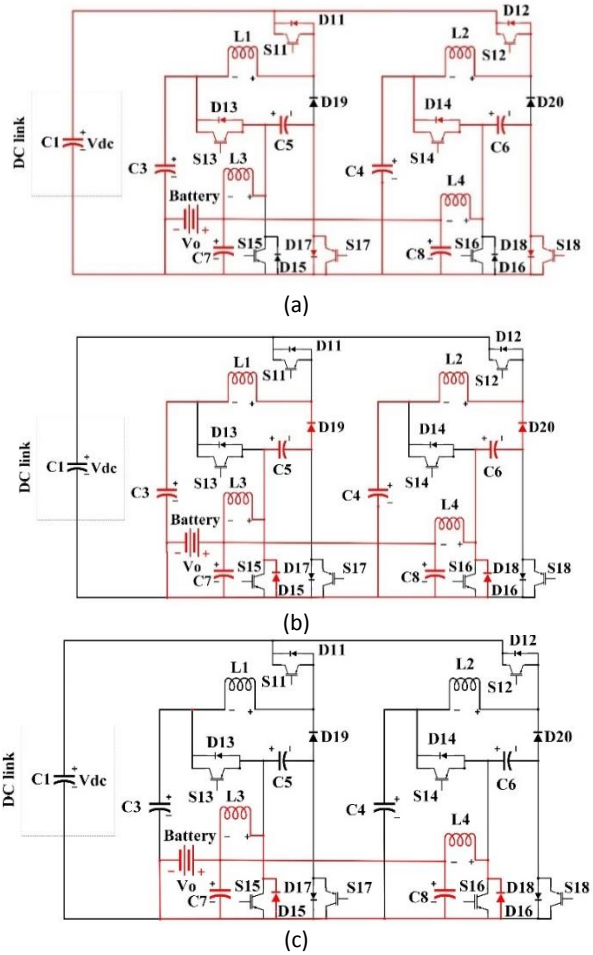


Fig. 9: Equivalent circuit of the proposed DC-DC converter in G2V operation: (a) Mode 1, (b) Mode 2, (c) Mode 3.

$$V_{L3}=\begin{cases} -V_o+v_{C3} & 0\le t<DT_s \\ -V_o & DT_s\le t<(D+D_1)T_s \\ 0 & (D+D_1)T_s\le t<T_s \end{cases} \quad (10)$$

By applying volt-second balance law on both inductors, voltage of capacitors is calculated as given in (16).

$$V_{C3}=V_{C5}=(D+D_1)V_o/D \tag{11}$$

So, during this mode, DCM voltage gain transfer ratio is calculated as follows:

$$M_{G2V(DCM)}=\frac{V_o}{V_{dc}}=\frac{D^2}{(D+D_1)(D+2D_2)} \tag{12}$$

B. V2G Operation

Unlike the G2V Operation mode, in this mode, the power flows from the battery side to the DC link. Two modes are defined for CCM operation and four modes for DCM:

First mode of CCM and DCM [0-t1]: in this interval, only switch S15 conducts, as shown in Fig. 10(a). The DC source charges inductors L3 and L1, that increases the current. The energy of capacitors C1 and C2 is discharged in inductor L1. Fig. 8(b) shows the derived current and voltage equation.

$$\begin{cases} v_{L1}=L_1\frac{di_{L1}}{dt}=v_{C3}+v_{C5} \\ v_{L3}=L_3\frac{di_{L3}}{dt}=V_o \end{cases} \tag{13}$$

The second mode, CCM [t1-Ts] and DCM [t1-t2]: according to Fig. 10(b), only three diodes D11, D13 and D17 conduct and all switches are off. Inductor L3 discharges energy to capacitor C5, and inductor L1 discharges its energy to DC link.

$$\begin{cases} v_{L1}=L_1\frac{di_{L1}}{dt}=-V_{dc}+v_{C3}=-V_{dc}+v_{C5} \\ v_{L3}=L_3\frac{di_{L3}}{dt}=V_o-v_{C3}=V_o-v_{C5} \end{cases} \tag{14}$$

It can be seen that the voltage of capacitors C3 and C5 are equal. If we apply the volt-second balance law on the inductors, we have:

$$D(V_o)+(1-D)(V_o-V_{C3})=0 \tag{15}$$

$$D(V_{C3}+V_{C5})+(1-D)(-V_{dc}+V_{C3})=0 \tag{16}$$

Using (15), the voltage of capacitors C1 and C2 can be determined as follow:

$$V_{C3}=V_{C5}=\frac{V_o}{(1-D)} \tag{17}$$

The voltage conversion ratio of the proposed converter during CCM operation for V2G is obtained from (16) and (17).

$$M_{V2G(CCM)}=\frac{V_{dc}}{V_o}=\frac{1+D}{(1-D)^2} \tag{18}$$

Third mode of DCM [T2 -T3]: current of inductor L1 reaches to zero.

Fourth mode, DCM [t3-Ts]: the current through the inductors reaches to zero and all switches are off. At the end of this interval, a full switching period of Ts is completed. As shown in Fig. 10(c), the voltages on both sides of the inductors are calculated as follow:

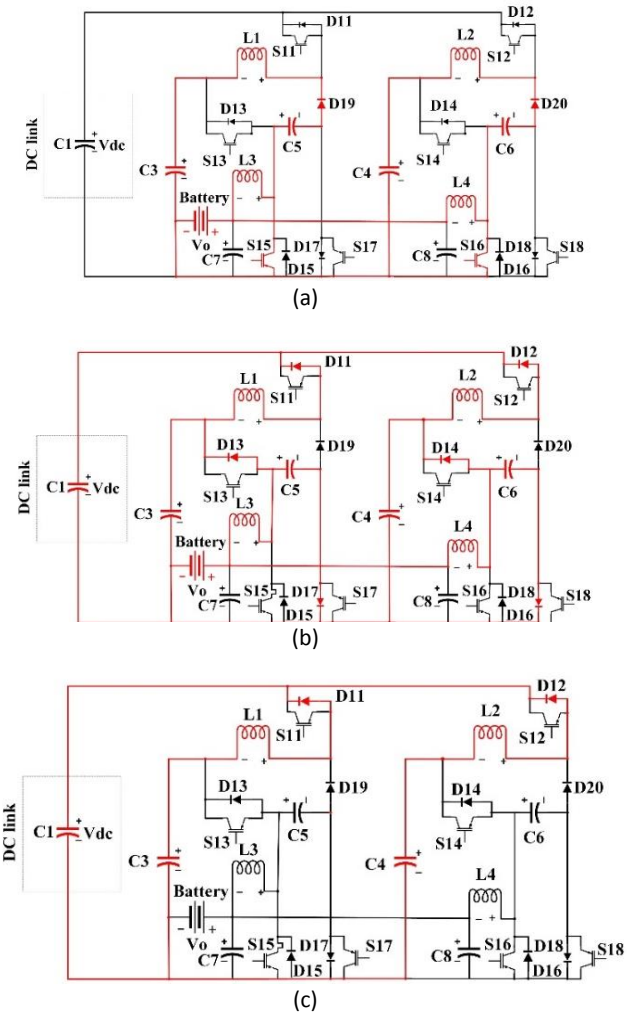


Fig. 10: Equivalent circuit of the proposed DC-DC converter in V2G operation: (a) Mode 1, (b) Mode 2, (c) Mode 3.

$$V_{L1}=\begin{cases} v_{C3}+v_{C5} & 0\le t<DT_s \\ -V_{dc}+v_{C3} & DT_s\le t<(D+D_2)T_s \\ 0 & (D+D_2)T_s\le t<T_s \end{cases} \tag{19}$$

$$V_{L3}=\begin{cases} V_O & 0\le t<DT_s \\ V_O-v_{C3} & DT_s\le t<(D+D_1)T_s \\ 0 & (D+D_1)T_s\le t<T_s \end{cases} \tag{20}$$

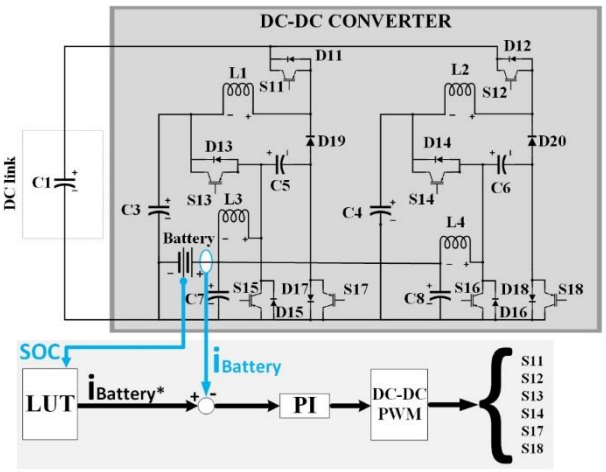
The volt-second balance law is applied to inductors L1 and L3.

Therefore, the voltages of capacitors C3 and C5 are equal to (21).

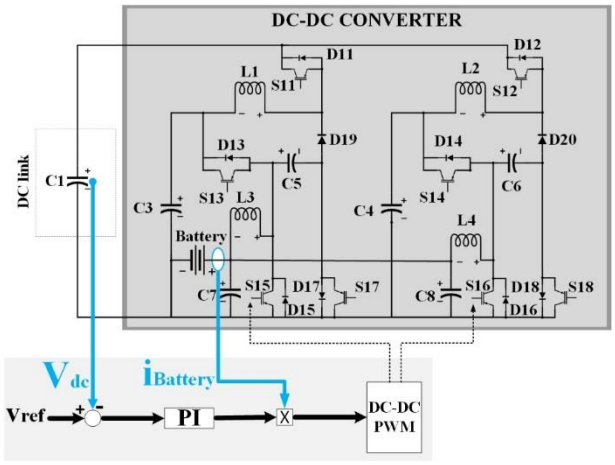
$$V_{C3}=V_{C5}=\frac{(D+D_1)V_o}{D_1} \tag{21}$$

So, the DCM voltage gain in V2G mode can be obtained as follows:

$$M_{V2G\,(DCM)}=\frac{V_{dc}}{V_o}=\frac{(D+D_1)(2D+D_2)}{D_1D_2} \tag{22}$$



(a)



(b)

Fig. 11: schematic of DC-DC Converter with controller: (a) G2V, (b) V2G.

Proposed DC-DC Converter Control System

In G2V operating mode, the difference between V_{dc} and V_{ref} is given to PI controller. The PI controller output duty cycle is compared with a sawtooth wave to apply the gate pulses to the switches S15 and S16 (fig. 11(a)). During G2V operation, switches S15 and S16 are off. SOC extracts the appropriate current of $i_{Battery}^*$ using lookup table (LUT) data. The difference between $i_{Battery}$ and $i_{Battery}^*$ is given to the PI controller. The output of the PI controller determines the duty cycle (fig. 11(b)). By using PWM Generator, the required pulses are applied to switches S11, S12, S13, S14, S17 and S18 with the desired duty cycle. A 380 Volt 40 ampere-hour battery is applied in the studied system. The proposed topology can be used for a wide range of battery voltages.

Duty Cycle for G2V and V2G Operation

Using the battery and DC link voltages, it obtained that the duty cycle in G2V and V2G operation modes are calculated as follow:

$$D_{G2V}=0.5\left(-V_o+\sqrt{V_o^2+8V_oV_{dc}}\right)/V_{dc} \quad (23)$$

$$D_{V2G}=0.5\left(2V_{dc}+V_o-\sqrt{V_o^2+8V_oV_{dc}}\right)/V_{dc} \quad (24)$$

Current ripple and average current of inductors are equal to:

$$G2V:\ \Delta i_{L3}=\frac{1-D}{L_3f_s}V_{dc};\Delta i_{L1}=\frac{2(1-D)}{L_1f_sD}V_{dc} \quad (25)$$

$$V2G:\ \Delta i_{L3}=\frac{D}{L_3f_s}V_o;\Delta i_{L1}=\frac{2D}{L_1f_s(1-D)}V_o \quad (26)$$

$$I_{L3}=I_o=\Delta i_{L3}(D+D_1)/4 \quad (27)$$

$$I_{L1}=\underbrace{\frac{D}{2-D}I_o}_{G2V}=\underbrace{\frac{1-D}{1+D}I_o}_{V2G}=\Delta i_{L1}(D+D_2)/4 \quad (28)$$

f_s is switching frequency. For DCM modes, D1 and D2 can be calculated as follow:

$$G2V:\begin{cases}D_1=\frac{2L_3f_sI_o}{(1-D)V_o}-D\\ D_2=\frac{D^2L_1f_sI_o}{(2-D)(1-D)V_o}-D\end{cases} \quad (29)$$

$$V2G:\begin{cases}D_1=\frac{2L_3f_sI_o}{DV_o}-D\\ D_2=\frac{(1-D)L_1f_sI_o}{D(1+D)V_o}-D\end{cases} \quad (30)$$

Design of Passive Elements

The values of inductance are concluded from (4) and (14), as follow:

$$G2V:\ L_3\&\geq\frac{1-D}{\Delta i_{L3}f_s}V_o;L_1\geq\frac{2(1-D)}{\Delta i_{L1}f_sD}V_o \quad (31)$$

$$V2G:\ L_3\geq\frac{D}{\Delta i_{L3}f_s}V_o;L_1\geq\frac{2D}{\Delta i_{L1}f_s(1-D)}V_o \quad (32)$$

Table 2: Values for each element in the proposed DC-DC converter

Parameters	Values
Rated power (P_{out})	17 [KW]
Battery and DC Link side voltages (V_o, V_{dc})	380 [V] and 600 [V]
3-Phase Grid side voltage (V_g)	380 [V], 50 [Hz]
Switching frequency (f_s)	1 [kHz]
Inductors L1, L3 and L_g	1 [mH], 2.2 [mH] and 1 [mH]
Capacitors C1, C3, C5 and C7	470 [µF], 330 [µF], 330 [µF] and 2.2 [µF]

In CCM operation mode, the minimum inductor current must be positive, so the critical values of inductors are obtained as follow:

$$G2V:\ L_3\&\geq\frac{(1-D)V_o}{2f_sI_o};\ L_1\geq\frac{(2-D)(1-D)}{D^2f_sI_o}V_o \quad (33)$$

$$V2G:\ L_3\geq\frac{DV_o}{2f_sI_o};L_1\geq\frac{D(1+D)V_o}{(1-D)^2f_sI_o} \quad (34)$$

Values of capacitors are determined as follow:

$$G2V:\ C_{3;5}\geq\frac{D(1-D)I_o}{\Delta v_{C3;5}f_s(2-D)};C_7\geq\frac{(1-D)V_o}{8\Delta v_{C7}f_s^2L_1} \quad (35)$$

$$\text{V2G: } C_{3;5} \geq \frac{D(1\text{-}D)I_o}{\Delta v_{C3;5} f_s (1+D)} ; C_{dc} \geq \frac{D(1\text{-}D)^2 V_o}{\Delta v_{Cdc} f_s (1+D)} \quad (36)$$

By having the allowed ripple of capacitors voltage, and considering (35) and (36), the values of capacitors are obtained. Considering the interleaved configuration of DC-DC converter, the obtained values for each element in one part is similar to the same element in the other part. Table 2 shows the values of DC-DC converter elements.

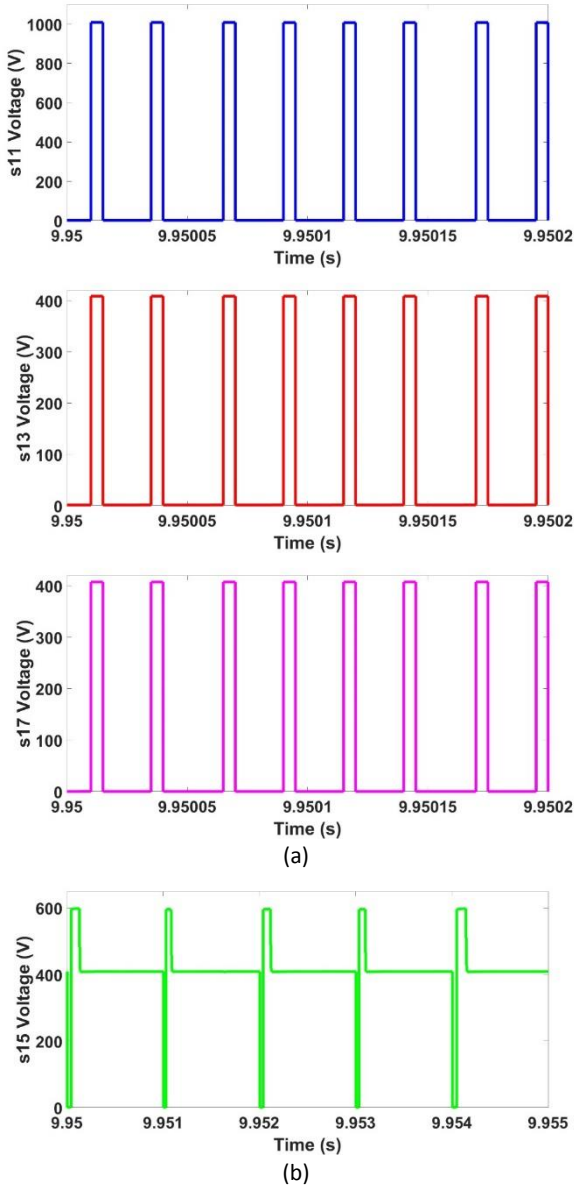


Fig. 12: The voltage of on switches: (a) G2V, (b) V2G.

Comparision

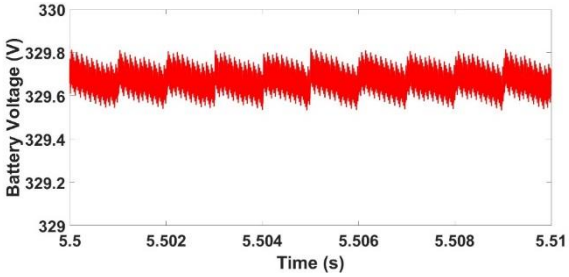
The proposed battery charger is compared with similar chargers in this section and results are given in Table. 3. As mentioned before, the proposed configuration has bidirectional power flow capability which only some of the chargers have this capability. Also, from efficiency point of view, the proposed configuration has relatively good situation. It should be noted that the ratio of the output power to the input power is used to determine the efficiency. As, in the V2G working mode the input power is the battery power and the power delivered to the network is the output power, and in the G2V working mode, the power to the battery is the output power and the converter power is considered as the input power. The given battery charger has the best battery voltage range among the compared chargers.

Table 3: Results of comparison between the proposed charger and similar configurations

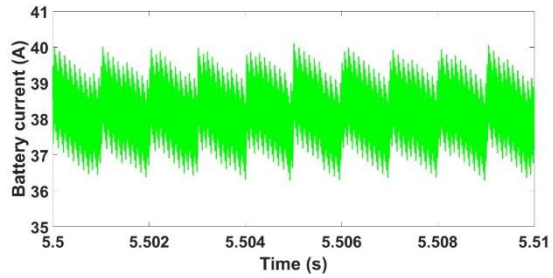
REF.	Efficiency	Bidirectional power flow	Battery voltage range	Switching frequency
[1]	99%	No	200-650 V	20-80 kHz
[2]	97.01%	No	460-800 V	150 kHz
[4]	97.5%	Yes	48 V	10 kHz
[5]	98.4%	Yes	400 V	20 kHz
[7]	98.2%	No	150-950 V	500 kHz
[8]	95.6%	Yes	48-450 V	40 kHz
[9]	97.2%	Yes	40 V	50 kHz
[12]	99.2%	No	700-900 V	50-160 kHz
[13]	98%	Yes	430-620 V	20 kHz
[15]	97.9%	No	800 V	100 kHz
[16]	92.08%	No	46-65 V	20 kHz
[17]	96.7%	No	9-16 V	260-400 kHz
Proposed	**>96%**	**Yes**	**20-600 V**	**1 kHz**

Simulation Results

It has been designed in Simulink MATLAB software to validate the control scheme and battery topology of the proposed charger. The switching frequency is considered to be 1 kHz, and the selfie filter is 1 mH is connected to a three-phase network of 380 V and 50 Hz. In this simulation, a battery with a voltage of 380 volts and a current of 40 amp-hours is used, which can be replaced with different batteries with different voltage-current ranges. Fig. 12(a) in CCM mode shows the voltage of switches S11, S13 and S17 in G2V mode, which are “on” in this mode. Fig. 12(b) shows the voltage of switch S15 in DCM mode and in V2G mode. The voltages of other switches are the same as the corresponding interleaved switch. In CCM mode and in G2V operation mode, according to Fig. 13(a), Battery voltage ripple is less than one volt. Fig. 13(b) shows the charging current of the battery in full load.



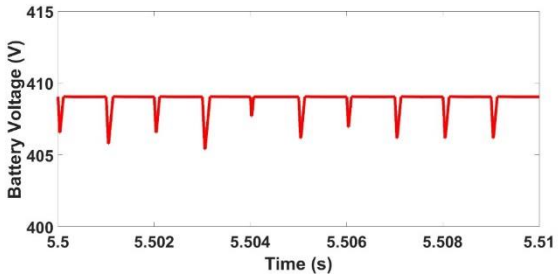
(a)



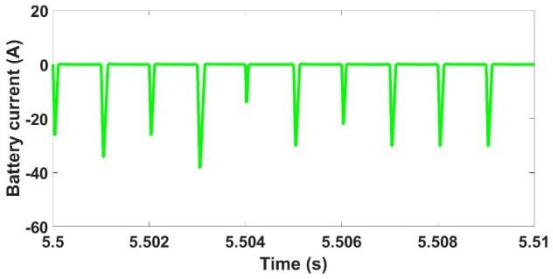
(b)

Fig. 13: CCM operation in G2V: (a) Battery voltage, (b) Battery current.

The ripple of the battery charging current is less than 4 amps. In DCM mode and V2G operation mode, battery discharge voltage and current are shown as shown in Fig. 14.

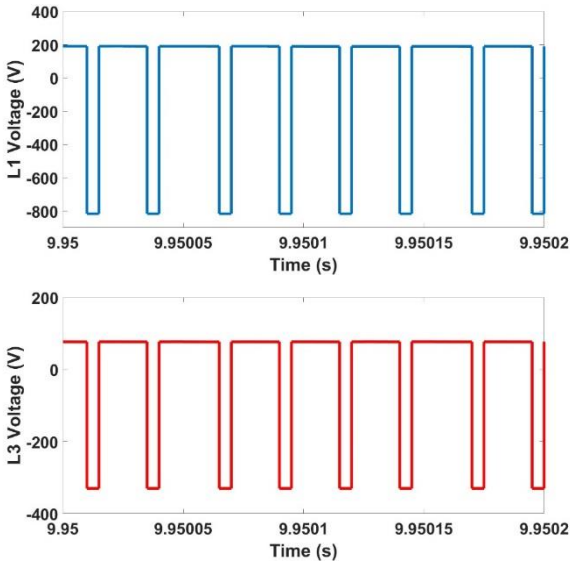


(a)

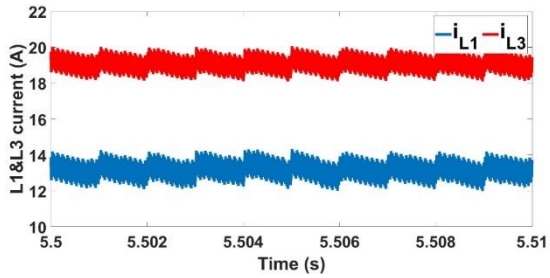


(b)

Fig. 14: DCM operation in V2G: (a) Battery voltage, (b) Battery current.

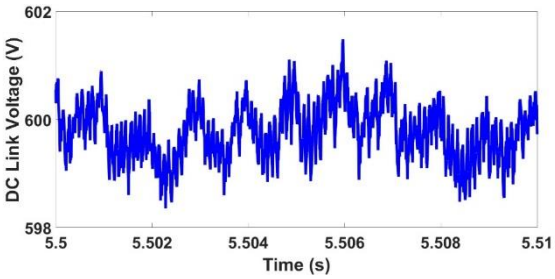


(a)

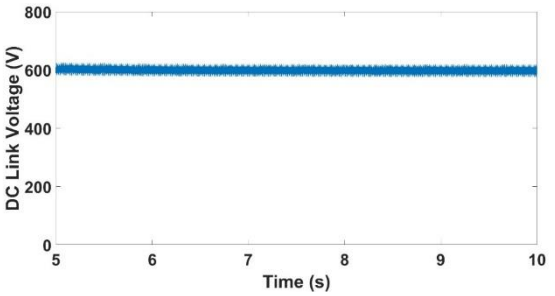


(b)

Fig. 15: Inductor voltage, and (b) inductor current in CCM mode.



(a)



(b)

Fig. 16: DC Link voltage: (a) G2V, (b) V2G.

Fig. 15 shows the voltage and current of inductors L1 and L3 in G2V operation mode. The current ripple of the inductors is about 2 amps. Also, DC link voltage in G2V and V2G operation modes are given in Fig. 16. In the V2G operation mode, the voltage and current of the inductors for 0.5kW load on the network side are shown in Fig. 17.

The reference voltage for the DC link is 600 volts, which shows a ripple of less than 4 volts for G2V operation mode. By connecting a load of 380V and 500W, the phase-to-phase voltage and the current waveform of phase A are obtained as shown in Fig. 18.

In order to check the usability of the proposed converter in charging stations, two DC-DC converters are connected to the DC link in parallel. This type of connection provides the ability to charge an electric vehicle with only one AC-DC converter, which will ultimately lead to a reduction in the number of elements and cost, but also can charge cars with different battery capacities at the same time.

Fig. 19 shows the charging current of two identical batteries that have the same SOCs. The current ripple of both batteries is equal and varies from 36 to 41 amps.

Assuming that these 2 electric cars are charging, the current of phase A is shown in Fig. 20.

The peak of this current depends on the number of EVs being charged and the SOC of the batteries. According to the obtained simulation results, the proposed converter can be placed in the category of off-board fast charger, which can charge a 380 volt and 40 amp-hour battery in less than 38 minutes. The peak of this current depends on the number of EVs being charged and the SOC of the batteries.

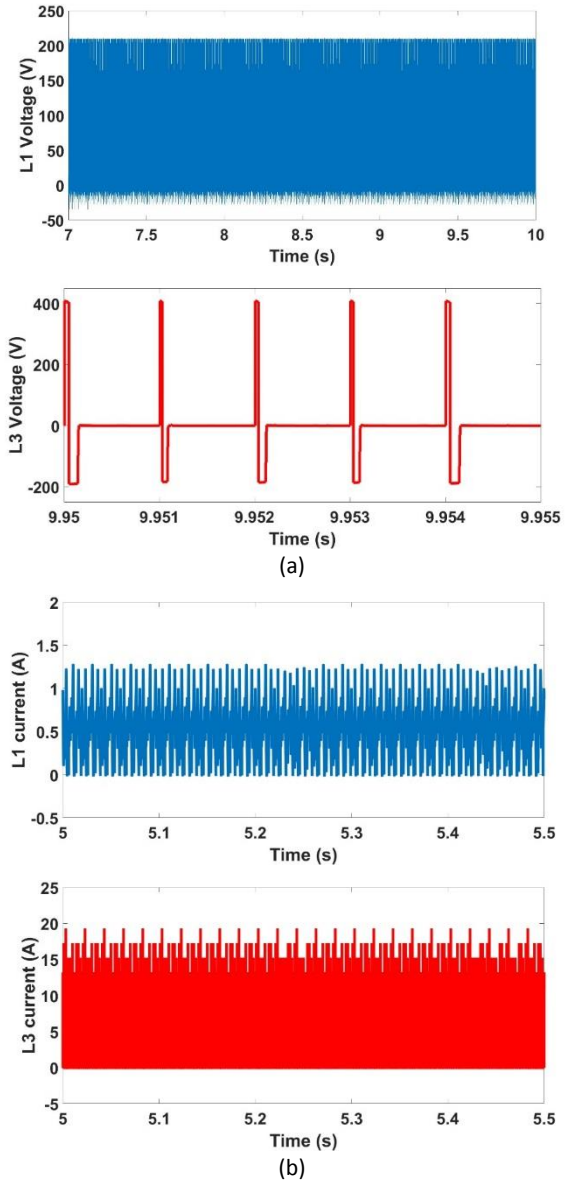


Fig. 17: (a) Inductor voltage, and (b) inductor current in DCM mode.

According to the obtained simulation results, the proposed converter can be placed in the category of off-board fast charger, which can charge a 380 volt and 40 amp-hour battery in less than 38 minutes. Fig. 21 illustrates power factor variations in the grid side of the proposed battery charger. As shown in this figure, power factor is nearly unit during the charge process.

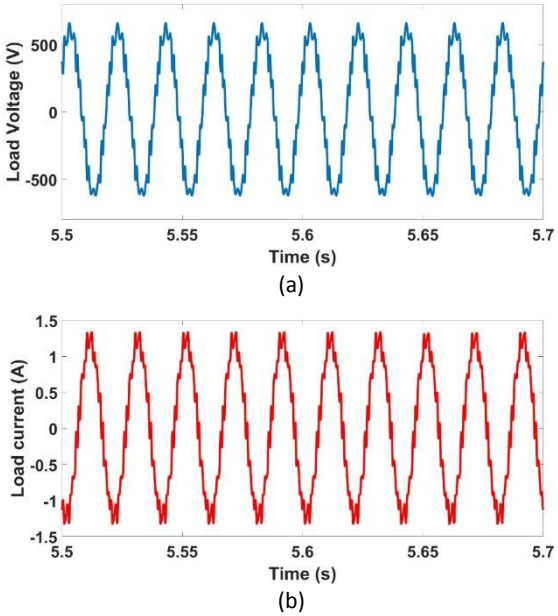


Fig. 18: Load voltage and current: (a) Phase to Phase Voltage, (b) Phase A current.

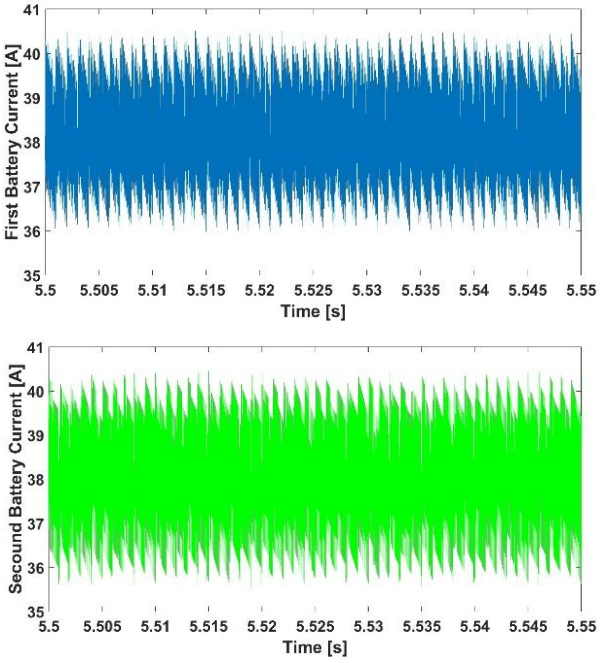


Fig. 19: The current of Batteries in the charging station.

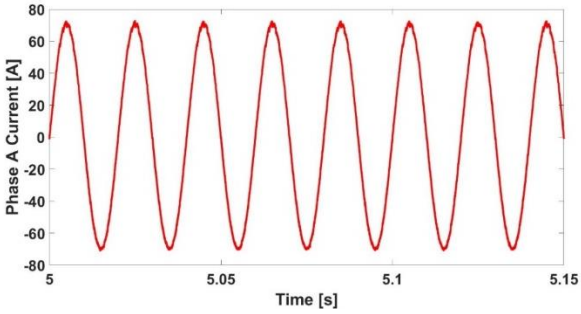


Fig. 20: Phase A current when connecting to two electric vehicles.

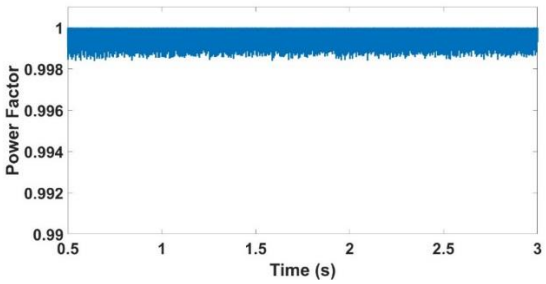


Fig. 21: Power factor variation during charge process.

Conclusions

In this paper, a non-isolated bidirectional DC-DC converter connected to a T-type three level converter for V2G and G2V applications is presented. The number of keys of the proposed T-type converter is reduced compared to the conventional T-type converter, which leads to a cost reduction in this structure. The proposed configuration has the merits of a common ground and ability of flow power in both directions. The low harmonic distortion of this structure is about 3%. In order to stabilize the DC link voltage, this structure uses SVPWM controller in V2G operation and SPWM controller in reverse power flow mode. The battery charger can charge a 380V and 40 amp-hour battery from 20-80% of SOC in less than 38 minutes that shows the charger operates as a fast charger. The proposed battery charger can be used in the charging station for a wide range of batteries with different voltages and currents. The advantages make the proposed converter one of the suitable options for battery chargers for electric vehicles.

Author Contributions

F. Sedaghati chose the field of research. S. A. Azimi collected information in this field. F. Sedaghati presented the proposed structure. S. A. Azimi simulated and controlled the proposed converter structure in MATLAB software. The authors discussed the obtained results and drew conclusions. Under the supervision of F. Sadaghti, the text of the article was prepared by S. A. Azimi. F. Sedaghati uploaded the article.

Acknowledgment

I appreciate the referees and their colleagues who helped the authors in publishing this article.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

EV	Electric Vehicle
SVPWM	Space Vector Pulse Width Modulation
CBPWM	Carrier Based Sinusoidal Pulse Width Modulation
V2G	Vehicle to grid
G2V	grid to Vehicle
THD	Total Harmonic Distortion
SOC	State of Charge
SPWM	Sinusoidal Pulse Width Modulation
CCM	Continuous Current Mode
DCM	Discontinuous Current Mode

References

[1] A. Ghosh, “Possibilities and challenges for the inclusion of the Electric Vehicle (EV) to reduce the carbon footprint in the transport sector: A review,” Energies, 13(10): 1-22, 2020.

[2] A. Hussain, V. H. Bui, J. W. Baek, H. M. Kim, “Stationary energy storage system for fast ev charging stations: Simultaneous sizing of battery and converter,” Energies, 12(23): 1-17, 2019.

[3] S. Rivera, R. L. F, S. Kouro, T. Dragičević, B. Wu, “Bipolar DC power conversion: State-of-the-art and emerging technologies,” IEEE J. Emerg. Sel. Top. Power Electron., 9(2): 1192-1204, 2020.

[4] M. A. H. Rafi, J. Bauman, “A comprehensive review of DC fast-charging stations with energy storage: Architectures, power converters, and analysis,” IEEE Trans. Transp. Electrif., 7(2): 345-368, 2021.

[5] S. Rivera, B. Wu, “Electric vehicle charging station with an energy storage stage for Split-DC bus voltage balancing,” IEEE Trans. Power Electron, 32(3): 2376-2386, 2016.

[6] I. Aghabali, J. Bauman, P. J. Kollmeyer, Y. Wang, B. Bilgin, A. Emadi, “800-V electric vehicle powertrains: Review and analysis of benefits, challenges, and future trends,” IEEE Trans. Transp. Electrif., 7(3): 927-948, 2021.

[7] Y. Tahir et al., “A state-of-the-art review on topologies and control techniques of solid-state transformers for electric vehicle extreme fast charging,” IET Power Electron., 14(9): 1560-1561, 2021.

[8] L. Zhou, M. Jahnes, M. Eull, W. Wang, M. Preindl, “Control design of a 99% efficiency transformerless EV charger providing standardized grid services,” IEEE Trans. Power Electron., 37(4): 4022-4038, 2022.

[9] H. Kim, J. Park, S. Kim, R. M. Hakim, H. Belkamel, S. Choi, “A single-stage electrolytic capacitor-less EV charger with single- and three-phase compatibility,” IEEE Trans. Power Electron., 37(6): 6780-6791, 2022.

[10] Y. Du, X. Zhou, S. Bai, S. Lukic, A. Huang, “Review of nonisolated bi-directional DC-DC converters for plug-in hybrid electric vehicle charge station application at municipal parking decks,” in Proc. 2010 Twenty-Fifth Annual IEEE Applied Power Electronics Conference and Exposition (APEC): 1145-1151, 2010.

[11] P. Falkowski, M. Korzeniewski, A. Ruszczyk, K. Kóska, "Analysis and design of high efficiency DC/DC buck converter," Przegląd Elektrotechniczny, 5: 150-156, 2016.

[12] S. Dusmez, A. Hasanzadeh, A. Khaligh, “Loss analysis of non-isolated bidirectional DC/DC converters for hybrid energy storage system in EVs,” in Proc. 2014 IEEE 23rd International Symposium on Industrial Electronics (ISIE): 543-549, 2014.

[13] L. Tan, B. Wu, V. Yaramasu, S. Rivera, X. Guo, “Effective voltage balance control for bipolar-DC-bus-fed EV charging station with three-level DC–DC fast charger,” IEEE Trans. Ind. Electron., 63(7): 4031- 4041, 2016.

[14] V. Monteiro, J. C. Ferreira, A. A. N. Meléndez, C. Couto, J. L. Afonso, “Experimental validation of a novel architecture based on a dual-stage converter for off-board fast battery chargers of electric vehicles,” IEEE Trans. Veh. Technol., 67(2): 1000-1011, 2018.

[15] S. Lu, M. Mu, Y. Jiao, F. C. Lee, Z. Zhao, "Coupled inductors in interleaved multiphase three-level DC–DC converter for high-power applications," IEEE Trans. Power Electron, 31(1): 120-134, 2016.

[16] S. Utsav, B. Singh, "A bidirectional battery charger for a wide range of electric vehicles," in Proc. 2022 IEEE Global Conference on Computing, Power and Communication Technologies (GlobConPT), 2022.

[17] A. Jain, K. K. Gupta, S. K. Jain, P. Bhatnagar, "A bidirectional five-level buck PFC rectifier with wide output range for EV charging application," IEEE Trans. Power Electron, 37(11): 13439-13455, 2022.

[18] M. Eull, L. Zhou, M. Jahnes, M. Preindl, "Bidirectional nonisolated fast charger integrated in the electric vehicle traction drivetrain," IEEE Trans. Transp. Electrif., 8(1): 180-195, 2022.

[19] J. Gupta, B. Singh, "An on-board charging system for light EVs with G2V and V2G power transfer capability," in Proc. IEEE IAS Global Conference on Emerging Technologies (GlobConET), 2022.

[20] S. Mukherjee, J. M. Ruiz, P. Barbosa, "A high power density wide range DC–DC converter for universal electric vehicle charging," IEEE Trans. Power Electron, 38(2): 1998-2012, 2023.

[21] H. Karneddi, D. Ronanki, "Universal bridgeless nonisolated battery charger with wide-output voltage range," IEEE Trans. Power Electron, 38(3): 2816-2820, 2023.

[22] H. Heydari-doostabad, T. O'Donnell, "A wide-range high-voltage-gain bidirectional DC–DC converter for V2G and G2V hybrid EV charger," IEEE Trans. Ind. Electron, 69(5): 4718-4729, 2021.

[23] D. Cittanti, M. Gregorio, E. Vico, E. Armando, R. Bojoi, "High performance digital multi-loop control of LLC resonant converters for EV fast charging with LUT-based feedforward and adaptive gain," IEEE Trans. Ind. Appl., 58(5): 6266-6285, 2022.

[24] M. A. Alharbi, A. M. Alcaide, M. Dahidah, P. Montero-Robina, S. Ethni, V. Pickert, J. I. Leon, "Rotating phase shedding for interleaved DC–DC converter-based EVs fast DC chargers," IEEE Trans. Power Electron, 38(2): 1901-1909, 2023.

[25] L. Zhou, M. Jahnes, M. Eull, W. Wang, G. Cen, M. Preindl, "Robust control design for ride-through/trip of transformerless onboard bidirectional ev charger with variable-frequency critical-soft-switching," IEEE Trans. Ind. Appl., 58(4): 4825-4837, 2022.

[26] R. Mayer, M. B. E. Kattel, S. V. G. Oliveira, "Multiphase interleaved bidirectional DC/DC converter with coupled inductor for electrified-vehicle applications," IEEE Trans. Power Electron, 36(3): 2533-2547, 2021.

[27] H. Moradisizkoohi, N. Elsayad, O. A. Mohammed, "A voltage-quadrupler interleaved bidirectional DC-DC converter with intrinsic equal current sharing characteristic for electric vehicles," IEEE Trans. Ind. Electron, 68(2): 1803-1813, 2021.

[28] M. Abbasi, K. Kanathipan, J. Lam, "An interleaved bridgeless single-stage AC/DC converter with stacked switches configurations and soft-switching operation for high voltage EV battery systems," IEEE Trans. Ind. Appl., 58(5): 5533-5545, 2022.

[29] R. Kushwaha, B. Singh, "A bridgeless isolated half bridge converter based EV charger with power factor pre-regulation," IEEE Trans. Ind. Appl, 58(3): 3967-3976, 2022.

[30] X. Zhou et al., "A high-efficiency high-power-density on-board low-voltage DC–DC converter for electric vehicles application," IEEE Trans. Power Electron, 36(11): 12781-12794, 2021.

Biographies

Farzad Sedaghati was born in Ardabil, Iran, in 1984. He received the M.S. and Ph.D. degrees both in Electrical Engineering in 2010 and 2014 from the University of Tabriz, Tabriz, Iran. In 2014, he joined the Faculty of Engineering, Mohaghegh Ardabili University, where he has been an Assistant Professor, since 2014. His current research interests include renewable energies and power electronic converters design and applications.

- Email: farzad.sedaghati@uma.ac.ir
- ORCID: 0000-0001-6974-4719
- Web of Science Researcher ID: NA
- Scopus Author ID: 35410298600
- Homepage: https://academics.uma.ac.ir/profiles?Id=617

Seyed Abbas Azimi was born in Ardabil, Iran, on February, 1996. He received his B.Sc. degree in Electrical Engineering from Islamic Azad University, Ardabil branch, Ardabil, Iran in 2018 and his M.Sc. degree in Power Electronics and electric machines engineering from the Azarbaijan Shahid Madani University, Tabriz, Iran in 2021. he is currently pursuing the Ph.D. degree in the Electrical Engineering at University of Mohaghegh Ardabili in Ardabil, Iran. His research interests include electric vehicle and power electronics design, simulation, modeling, and control of electrical machines.

- Email: abbasazimi@uma.ac.ir
- ORCID: 0009-0004-4886-2213
- Web of Science Researcher ID: NA
- Scopus Author ID: 58975611100
- Homepage: NA

How to cite this paper:
F. Sedaghati, S. A. Azimi, "Electric vehicle battery charging using a non-isolated bidirectional DC-DC converter connected to t-type three level converter," J. Electr. Comput. Eng. Innovations, 13(1): 129-140, 2025.

DOI: 10.22061/jecei.2024.10870.743

URL: https://jecei.sru.ac.ir/article_2205.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Utilizing Normalized Mutual Information as a Similarity Measure for EEG and fMRI Fusion

Z. Rabiei, H. Montazery Kordy *

Faculty of Electrical and Computer Engineering, Babol Noshirvani University of Technology, Babol, Iran.

Article Info

Article History:
Received 23 June 2024
Reviewed 11 August 2024
Revised 28 September 2024
Accepted 15 October 2024

Keywords:
Data fusion
Coupled matrix tensor factorization
Electroencephalogram (EEG)
functional Magnetic Resonance Imaging (fMRI)
Normalized mutual information (NMI)

*Corresponding Author's Email Address:
hmontazery@nit.ac.ir

Abstract

Background and Objectives: Neuroscience research can benefit greatly from the fusion of simultaneous recordings of electroencephalogram (EEG) and functional magnetic resonance imaging (fMRI) data due to their complementary properties. We can extract shared information by coupling two modalities in a symmetric data fusion.

Methods: This paper proposed an approach based on the advanced coupled matrix tensor factorization (ACMTF) method for analyzing simultaneous EEG-fMRI data. To alleviate the strict equality assumption of shared factors in the common dimension of the ACMTF, the proposed method used a similarity criterion based on normalized mutual information (NMI). This similarity criterion effectively revealed the underlying relationships between the modalities, resulting in more accurate factorization results.

Results: The suggested method was utilized on simulated data with correlation levels of 50% and 90% between the components of the two modalities. Despite different noise levels, the average match score improved by 20% compared to the ACMTF model, as demonstrated by the results.

Conclusion: By relaxing the strict equality assumption, we can identify shared components in a common mode and extract shared components with higher performance than the traditional methods. The suggested method offers a more robust and effective way to analyze multimodal data sets. The findings highlight the potential of the ACMTF method with NMI-based similarity criterion for uncovering hidden patterns in EEG and fMRI data.

Introduction

Joint analysis of neuroimaging data such as EEG and fMRI has the potential to gain a better understanding of brain functioning. The primary objective of analyzing multiple modalities is to utilize common and distinct information from complementary modalities to understand neural activities better. EEG and fMRI data fusion can provide researchers with a more comprehensive understanding of the brain's spatial and temporal functions [1], [2].

The synchronous electrical activity of brain neurons over time can be measured using EEG. While EEG has the perfect temporal resolution, this technique has poor spatial information due to the number of electrodes employed. On the other hand, blood oxygenation level-dependent (BOLD) imaging is a technique that is commonly used to measure brain activity indirectly using fMRI. Although it measures BOLD signals at a millimeter range, this technique is sluggish compared to brain activity [3], [4]. Therefore, EEG and fMRI can be fused to improve the localization of brain activity in time and space due to their complementary spatiotemporal resolutions. In recent years different types of fusion methods have been developed; the majority of them focus on the matrix factorization of EEG and fMRI into different components.

Independent component analysis (ICA) [5], principal component analysis (PCA), canonical component analysis (CCA) [6], [7], and independent vector analysis (IVA) [8] are methods for decomposing matrixes. Extensive studies have been conducted on decomposing EEG and fMRI datasets into different components using ICA and PCA methods. These methods define components using only two dimensions: time and space [5]. Whereas, EEG and fMRI datasets are typically multidimensional including time, voxels or channels, frequency, trial, and participant. One potential solution to this issue is to explore alternative methods for analyzing these datasets that preserve the interactions between different modalities. One approach uses tensor decomposition techniques designed to apply to multi-way data structures [9].

Non-physiological assumptions like orthogonality and statistical independence are the basis of matrix factorization models. However, the uniqueness of higher-order tensor decompositions is obtained by relaxed conditions (without any nonphysiological assumptions), making interpreting the extracted components easier [10]. By applying tensor decomposition algorithms such as Tucker decomposition or Parallel factorial analysis CANDECOMP/PARAFAC (CP) decomposition, researchers can extract more meaningful and interpretable components from EEG and fMRI datasets without losing important interactions between different dimensions [11], [12]. Coupled matrix tensor factorization (CMTF) is the most common method for fusing EEG and fMRI datasets using tensor decomposition methods [13]. In the CMTF model, data definition involves a third-order tensor of EEG coupled with a matrix fMRI. Gradient-based optimization algorithms are used to factorize the coupled data, and CP is utilized to model higher-order tensors. In the fusion of EEG and fMRI, the assumption is that there is one or more common modes of variation between the two modes, such as time or subject. The main disadvantage of the CMTF method is its reliance on equal shared components in the common dimension. Several methods have been introduced to alleviate this restrictive assumption. The advanced coupled matrix-tensor factorization (ACMTF) was developed in [14], identifying shared and unshared components. While the ACMTF can estimate the weights of the components and identify the factor matrices, it assumes that the shared components between the two modalities are identical. The notion of equality concerning brain signals could be confining. In [15] the CMTF model has been used to analyze the joint decomposition of EEG data at source level with fMRI along with a common spatial profile. This method can identify both common and discriminative subspaces compared to the CMTF method. A relaxed form of ACMTF was presented by the authors in [16]. This method overcomes the equality assumption of shared factors in a common dimension. This method uses the l_1-norm and l_2-norm to express similarity and then apply it to the components and their first and second derivatives. In [17] a tensor decomposition model was proposed in which a soft coupling method (Euclidean distance) was implemented for fusion EEG and fMRI. [18] has used the maximum correlation between the shared components of EEG and fMRI. Although Pearson's correlation coefficient (ρ) used in [18] is one of the most popular dependence measures with many desirable features, it only evaluates linear relationships. To assess relationships and dependencies between variables in a general sense, we need a metric, not only for linear or monotonic relationships. In contrast to the Pearson's correlation coefficient, mutual information (MI) takes into account both linear and non-linear relationships between variables, making it a more comprehensive measure of dependence. Additionally, MI can capture complex dependencies that may not be captured by the correlation coefficient alone. This makes MI a valuable tool for analyzing relationships in a wide range of fields. Overall, while the correlation coefficient is useful for measuring linear relationships, MI provides a more nuanced and flexible approach to understanding the dependencies between variables [19].

However, the estimation of MI and entropy values can be challenging. MI-based measures need appropriate estimation methods as the underlying probability distributions are unknown. The most commonly used technique for estimating MI is histogram-based density estimation [17]. Despite not always being the most accurate method, histogram-based density estimation has acceptable accuracy. Normalization of mutual information is essential because MI values can vary widely depending on the scale of the variables involved. By transforming MI into a standardized range, we can compare and interpret the information content more accurately. As a result, we used the normalized mutual information (NMI) as a similarity metric in our study [20], [21]. The contributions of the proposed method are summarized as follows:

- Using normalized mutual information as a similarity measure, our proposed method can relax the restrictive equality assumption of shared components in the ACMTF method.
- As a comprehensive approach, our method can estimate the weight of each component and identify identical and similar components with various correlation levels.
- Our proposed method, compared to other methods based on similarity criteria, can identify components that are linearly or nonlinearly related to each other.

The following is the structure of this paper. We first explain tensor decomposition, the ACMTF method, and

HRF modeling. The proposed method and the calculation of normalized mutual information are presented in the second part. Then, a simulation study is used to validate the performance of the presented method. Finally, the paper is completed with a discussion and conclusion.

Material and Methods

A. Notation

Vectors, matrices, and higher-order tensors in this study are identified using italic lower-case, italic upper-case, and italic calligraphic upper-case letters respectively. For a matrix *A*, $\bar{A}$ denotes its transpose. The symbol $\odot$ signifies the Khatri-Rao product of two matrices, A$\in \mathbb{R}^{I\times R}$ and B$\in \mathbb{R}^{J\times R}$, namely, $A\odot B = [a_1 \otimes b_1, a_2 \otimes b_2, \dots a_R \otimes b_R]$, with a_i and b_i being the ith columns of A and B respectively, and $\otimes$ denoting Kronecker prodeuct.

B. Tensor Decomposition

In mathematics, a tensor is described as a numerical array with multiple indexes, and the order of a tensor is the number of its modes or dimensions. The Canonical Polyiadic Decomposition (CP or CPD) model is briefly discussed in this section. A third-order tensor $\boldsymbol{\chi} \in \mathbb{R}^{I\times J\times K}$ with the modes of trial, frequency, and channel represents EEG data and a matrix $Y \in \mathbb{R}^{I\times L}$ (trial (scan) by voxels) indicates fMRI signal. Fig. 1 shows the EEG coupled with fMRI in the trial mode.

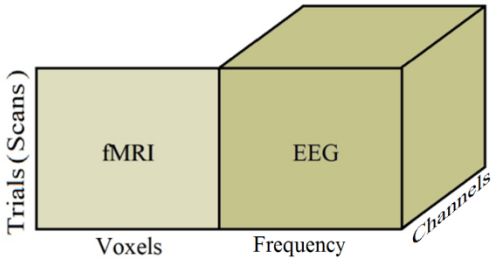


Fig. 1: A 3rd-order tensor EEG signal coupled with fMRI matrix in the trial mode.

CP is thought of as an extension of singular value decomposition (SVD) to higher-order tensors. It represents a 3rd-order tensor $\boldsymbol{\chi} \in \mathbb{R}^{I\times J\times K}$ as a linear combination of rank-one tensors:

$$\boldsymbol{\chi} = [\![\lambda; A, B, C]\!] = \sum_{r=1}^{R} \lambda_r a_r \circ b_r \circ c_r \qquad (1)$$

where $[\![.]\!]$ and $\circ$ indicate the full multilinear and vector outer product respectively. The vectors a_r, b_r and c_r are as rank-one components form the factor matrices A$\in \mathbb{R}^{R\times 1} = [a_1 \dots a_R]$, B$\in \mathbb{R}^{R\times 1} = [b_1 \dots b_R]$, and C$\in \mathbb{R}^{R\times 1} = [c_1 \dots c_R]$ respectively. The terms factor and component mention to the rank-one matrices or higher-order rank-one tensors. R signifies the number of factors and $\lambda \in \mathbb{R}^{R\times 1}$ is weights of rank-one components. The CP or PARAFAC model is one of the most popular tensor decomposition models. This model is used alongside models like Block Term Decomposition and the Tucker decomposition model [22].

C. Advanced Coupled Matrix and Tensor Factorization

We assume that the EEG data is structured as a third-order tensor to present variations across the trial, spectral, and spatial dimensions. At the same time, the fMRI matrix characterizes variations across the trial and spatial dimensions. Using the Advanced Coupled Matrix Tensor Factorization (ACMTF) model we can jointly factorize the 3rd-order tensor $\boldsymbol{\chi}$ coupled with a matrix Y in trial mode. The common mode between the EEG and fMRI is trial-to-trial (scan-to-scan) covariations in brain activity [23]. According to the definition of EEG and fMRI, a temporal relationship between EEG and fMRI data is considered in the form of hemodynamic response function (HRF) [4]-[24]. Thus, the ACMTF model [14] can be utilized to formulate an optimization problem:

$$\begin{aligned} & f(\lambda, \sigma, T_{ee}, F_{ee}, M_{ee}, M_{fm}) = \|\boldsymbol{\chi} - \\ & [\![\lambda; T_{ee}, F_{ee}, M_{ee}]\!]\| + \left\|Y - HT_{ee}\Sigma {M_{fm}}^T\right\|^2 + \\ & \beta\|\lambda\|_1 + \beta\|\sigma\|_1 \qquad (2) \\ & s.t. \|t_{eer}\| = \|f_{eer}\| = \|m_{eer}\| = \|m_{fmr}\| = \\ & 1 \quad for\ r = 1, \dots, R \end{aligned}$$

where the tensor $\boldsymbol{\chi}$ and matrix Y are decomposed based on the CANDECOMP/PARAFAC (CP) and singular value decomposition (SVD) models, respectively. The factor matrix $T_{ee} \in \mathbb{R}^{I\times R}$ (trial-to-trial variation) is common between EEG and fMRI. Moreover, hemodynamic trials of fMRI could be predicted using the convolution of T_{ee} in EEG with known HRF $h(t)$. The Toeplitz matrix H contains samples of $h(t)$ on its diagonals [4].

Also, $F_{ee} \in \mathbb{R}^{J\times R}$ and $M_{ee} \in \mathbb{R}^{K\times R}$ are factor matrices corresponding to the frequency and channel topography of the EEG signal, respectively; $M_{fm} \in \mathbb{R}^{L\times R}$ is the factor matrix corresponding to the spatial maps of voxels; and $\lambda \in \mathbb{R}^{R\times 1}$ and $\sigma \in \mathbb{R}^{R\times 1}$ are weights of rank-one components in the third-order tensor and the matrix, respectively. The $\Sigma \in \mathbb{R}^{R\times R}$ is a diagonal matrix, with σ forming its diagonal. Also, $\|.\|$ and $\|.\|_1$ represent the Frobenius norm and l_1-norm, respectively; $\beta \geq 0$ is a penalty parameter; and $t_{eer}, f_{eer}, m_{eer}$, and m_{fmr} are the r^{th} columns of T_{ee}, F_{ee} , M_{ee}, and M_{fm}, respectively. The weights λ and σ are sparsified using the l_1-norm terms. Thus, the components with significant weights in both modalities are considered shared, while unshared components have weights of almost zero in one of the datasets.

D. HRF Modeling

One method of analyzing the task-related fMRI data is to estimate the shape of the time courses corresponding to the considered stimulus. Among the various methods used to estimate these time courses, the linear

convolutional model is the most common technique. In this approach, the task-related time course can be modeled as a convolution between the stimulus function and a particular impulse response function known as the hemodynamic response function (HRF) [25]. One of the most popular techniques used to analyze fMRI signals is the general linear model (GLM). It models the BOLD signal as a linear combination of several various predictors. The GLM method requires an accurate estimate of the HRF. Several different models of the HRF are used in the analysis of fMRI signals [25], [26].

The most widely used model for the functional shape of the HRF is the double gamma distribution model, usually referred to as canonical which is used in SPM software. There are several models of canonical HRF in literature. HRF waveform shapes can be controlled by one to nine free parameters, based on the model [27]. The canonical HRF demonstrated in Fig. 2 has been used in this study. Some HRF studies use a basis function approach under the GLM framework; for example two sine basis functions or a product of a sine function and exponential function. Woolrich et al. presented constrained linear basis sets for HRF modeling using Variational Bayes. The proposed HRF was parameterized using six parameters. Several techniques exist to choose a basis set with the constraints so that the subspace spanned by the basis function creates a plausible HRF waveform [28].

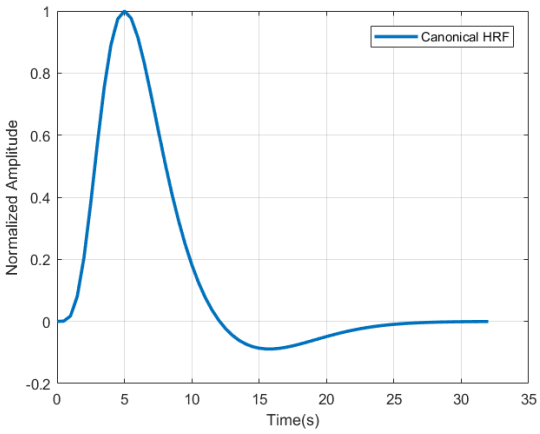


Fig. 2: Representation of the canonical HRF in SPM.

Bayesian methods are another approach for incorporating prior knowledge or uncertainty into the modeling of HRF [29]. The deconvolution method and a semiparametric approach based on finite impulse response (FIR) are other methods with more flexibility in modeling HRF [30].

Although HRF waveform varies across different subjects and different brain regions, in most studies, the HRF is considered invariant and assumed to be known. H is a convolution matrix defined as follows, where *l* is the length of the signal of T_{ee}, m is the length of HRF and the dimension of H is $(m + l - 1) \times l$.

$$H = \begin{bmatrix} h_1 & 0 & \cdots & 0 & 0 \\ h_2 & h_1 & \cdots & \vdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ h_m & h_{m-1} & \cdots & 0 & h_2 \\ 0 & h_m & \ddots & h_{m-2} & \vdots \\ 0 & 0 & \cdots & h_{m-1} & h_{m-2} \\ \vdots & \vdots & \cdots & h_m & h_{m-1} \\ 0 & 0 & 0 & \cdots & h_m \end{bmatrix} \quad (3)$$

E. Normalized Mutual Information (NMI)

Measurement of the relationship between two variables can be done using information theory (IT). A generalized criterion of dependency is mutual information (MI). MI quantifies the amount of information shared between two variables. In addition, It excels at capturing complex interdependencies that may not be adequately represented by traditional measures like Pearson's correlation coefficient. Overall, while Pearson's correlation coefficient is useful for measuring linear relationships, mutual information provides a more comprehensive understanding of the dependencies between variables. As a result, MI can be used to determine how similar or dissimilar the shared components in two datasets are.

If X and Y are two random variables, $p(x, y)$ is the joint distribution, and $p(x)$ and $p(y)$ are the marginal distributions, the mutual information is calculated as follows:

$$I(X,Y) = \sum_{x,y} p(x,y) log(\frac{p(x,y)}{p(x)p(y)}) = H(X) + H(Y) - H(X;Y) \quad (4)$$

where $I(X,Y)$ is the mutual information, $H(X)$ and $H(Y)$ are Shannon's entropies of the discrete random variables X and Y, respectively, and $H(X;Y)$ is the joint entropy [21].

Estimating the entropy and MI values requires careful consideration of the data distribution and sample size. MI-based measures require to be estimated due to the unknown underlying probability distributions. Additionally, choosing the appropriate method for estimating entropy and MI can impact the reliability and interpretability of the results. Compared to various MI estimation methods such as kernel density estimation or Gaussian mixture models, histogram-based density estimation is a simple and effective technique [20]. However, care must be taken to choose the appropriate number of bins to avoid underestimating or overestimating the true MI values. Before adding MI to the objective function, it needs to be normalized. The normalization ensures that all variables are on a consistent scale and make information more meaningful. Specifically, it is transformed into a number within the range of [0, 1]. The normalized mutual information (NMI) can be defined as

$$NMI(X,Y) = \frac{I(X;Y)}{\frac{1}{2}(H(X)+H(Y))} = \frac{2I(X;Y)}{(H(X)+H(Y))} \quad (5)$$

F. Generalized Coupled Matrix Tensor Factorization

The ACMTF has been enhanced by incorporating the NMI criteria to address the equality constraint of the shared components. Now, according to generalized coupled matrix tensor factorization (GCMTF), the cost function is as follows:

$$\begin{aligned} &f(\lambda,\sigma,T_{ee},F_{ee},M_{ee},T_{fm},M_{fm}) = \|\boldsymbol{\chi} - \\ &[\![\lambda;T_{ee},F_{ee},M_{ee}]\!]\| + \left\|Y - HT_{fm}\Sigma {M_{fm}}^T\right\|^2 + \\ &\gamma\sum_{r=1}^{R}\left(1-e^{-(\lambda_r\sigma_r)^2/\varepsilon}\right)\left(1-NMI(t_{eer},t_{fmr})\right) + \\ &\beta\sum_{r=1}^{R}\sqrt{{\lambda_r}^2+\epsilon} + \beta\sum_{r=1}^{R}\sqrt{{\sigma_r}^2+\epsilon} \end{aligned} \quad (6)$$

$$s.t.\ \|t_{eer}\| = \left\|t_{fmr}\right\| = \|f_{eer}\| = \|m_{eer}\| = \left\|m_{fmr}\right\| = 1 \quad for\ r = 1,\dots,R$$

where γ is the penalty parameter; $t_{eer}, f_{eer}, m_{eer}, t_{fmr}$, and m_{fmr} are the r^{th} columns of T_{ee}, F_{ee}, M_{ee}, T_{fm}, and M_{fm}, respectively; and NMI is the normalized mutual information between t_{eer} and t_{fmr}. By selecting a sufficiently small enough $\epsilon > 0$, the l_1-norm of λ and σ has been replaced with their differentiable equivalents. The expression $(1-e^{-(\lambda_r\sigma_r)^2/\varepsilon})$ is the smoothed l_0-norm, where ε is a tunable and small parameter to approximate l_0-norm [19]-[31]. This term is used to identify the shared components and avoid the maximization of the NMI between the unshared components.

Alternating Least Square (ALS) is the traditional approach for optimizing the objective function. In [11], non-conjugate gradient methods achieve faster convergence than ALS. In this approach, it is necessary to compute the gradients of the objective function with respect to their parameters. Hence, NMI gradients with respect to factor matrices in the objective function need to be computed. The Score Functions (SFs) defined in [20]-[32] were used to compute the NMI gradient.

If we have a bounded random vector X and a small enough Δ vector of the same dimension, it is demonstrated:

$$I(X+\Delta) - I(X) = E\{\Delta^T\beta_X(X)\} + o(\Delta) \quad (7)$$

where $\beta_X(X) = \psi_x(X) - \varphi_x(X)$ is the Score Function Difference (SFD) of X and $o(\Delta)$ represents the higher-order expressions in Δ. The terms $\psi_x(X)$ and $\varphi_x(X)$ are the Marginal Score Functions (MSFs) and Joint Score Functions (JSFs) of vector X, which are defined as follows:

$$\psi_x(X) = -\frac{d}{dx_i}lnp_{x_i}(x_i) = -\frac{\acute{p}_{x_i}(x_i)}{p_{x_i}(x_i)} \quad (8)$$

$$\varphi_x(X) = -\frac{\partial}{\partial x_i}ln\,p_X(X) = -\frac{\frac{\partial}{\partial x_i}p_X(X)}{p_X(X)} \quad (9)$$

where $p_{x_i}(x_i)$ is the marginal probability density function (PDF) of x_i and $p_X(X)$ is the joint PDF of random vector X. SFD estimation is our main concern since it is the gradient of mutual information. Histogram estimation is the preferred method among the various techniques used to estimate SFD due to its acceptable accuracy despite its simplicity [20]. The GCMTF method is graphically depicted in Fig. 3.

The cost function gradient is computed using these equations:

$$\begin{aligned}
&\partial l_0/\partial\lambda_r = (2/\varepsilon){\sigma_r}^2\lambda_r e^{-(\lambda_r\sigma_r)^2/\varepsilon}\left(1-NMI(t_{eer},t_{fmr})\right)\\
&\partial l_0/\partial\sigma_r = (2/\varepsilon){\lambda_r}^2\sigma_r e^{-(\lambda_r\sigma_r)^2/\varepsilon}\left(1-NMI(t_{eer},t_{fmr})\right)\\
&\partial f/\partial T_{ee} = (T_{(1)}-X_{(1)})(\lambda^T\odot M_{ee}\odot F_{ee}) + \alpha(T_{ee}-\bar{T}_{ee}) - \gamma\sum_{r=1}^{R}\left(1-e^{-(\lambda_r\sigma_r)^2/\varepsilon}\right)\partial NMI/\partial t_{eer}\\
&\partial f/\partial F_{ee} = (T_{(2)}-X_{(2)})(\lambda^T\odot M_{ee}\odot T_{ee}) + \alpha(F_{ee}-\bar{F}_{ee})\\
&\partial f/\partial M_{ee} = (T_{(3)}-X_{(3)})(\lambda^T\odot f_{ee}\odot T_{ee}) + \alpha(M_{ee}-\bar{M}_{ee})\\
&\partial f/\partial T_{fm} = H^THT_{fm}\Sigma M_{fm}^TM_{fm}\Sigma^T - H^TY\Sigma M_{fm} + \alpha(T_{fm}-\bar{T}_{fm}) - \gamma\sum_{r=1}^{R}\left(1-e^{-(\lambda_r\sigma_r)^2/\varepsilon}\right)\partial NMI/\partial t_{fmr}\\
&\partial f/\partial M_{fm} = M_{fm}\Sigma T_{fm}^TH^THT_{fm}\Sigma^T - Y^THT_{fm}\Sigma + \alpha(M_{fm}-\bar{M}_{fm})\\
&\partial f/\partial\lambda_r = (\tau-\chi)\times_1 t_{eer}\times_2 f_{eer}\times_3 m_{eer} + (\beta/2)\lambda_r/\sqrt{{\lambda_r}^2+\epsilon} + \partial l_0/\partial\lambda_r\\
&\partial f/\partial\sigma_r = HT_{fm}^T(HT_{fm}\Sigma M_{fm}^T - Y)M_{fm} + (\beta/2)\sigma_r/\sqrt{{\sigma_r}^2+\epsilon} + \partial l_0/\partial\sigma_r
\end{aligned} \quad (10)$$

where $\tau = [\![\lambda;T_{ee},F_{ee},M_{ee}]\!]$, $X_{(n)}$ is the tensor $\boldsymbol{\chi}$ unfolded in the nth mode, $\times_n$ defines the tensor-vector product in the nth mode, $\odot$ signifies the Khatri-Rao product and $\bar{T}_{ee}$ corresponds to T_{ee} with columns divided by their l_2-norms. Also term l_0 refers to the smoothed l_0-norm.

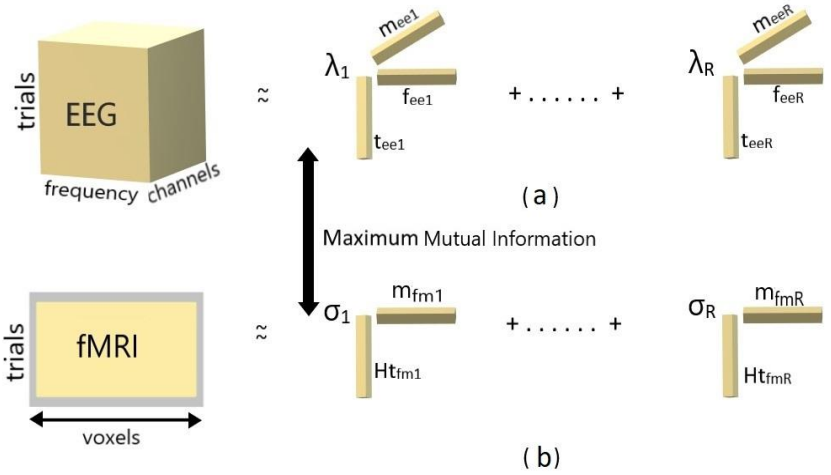


Fig. 3: Graphical representation of the GCMTF method with coupling in trial mode. (a) 3rd-order tensor EEG is represented by trials (t_{ee}), frequency (f_{ee}), and channels (m_{ee}). (b) Trials (t_{fm}) and voxels (m_{fm}) are used to index the matrix fMRI.

Results

The dataset used in this study is an oddball auditory paradigm derived from 17 healthy subjects [33]. The EEG data were denoted as a third-order tensor trial×frequency×channel and fMRI data as a matrix trial (scan) ×voxels. The EEG tensor for simulated data was generated using the same values as the real data [33]. The defined time windows (before and after the stimulus) were applied to the signals of each channel. These windows were transformed into a spectrogram using the Fourier transform. Then, based on the Canonical Polyadic Decomposition (CPD), these factors are multiplied to form the EEG tensor. To generate matrix fMRI, the trials were first convolved with canonical HRF and then multiplied with spatial factor (voxels). It is necessary to select the rank of the dataset before applying the method. Hence, using the Corcondia test, the number of components is selected to be 3. Now, it is assumed that all three components have significant values in both modalities. Thus, [1 1 1] was chosen as the values of λ and σ.

To evaluate the presented method's ability to identify shared components with different linear correlation levels, the components were selected as follows. The first two shared components have a linear correlation of 90%, the second ones have a linear correlation of 50%, and the third ones are considered to be the same. The temporal components of the two modalities with different linear correlation levels are shown in Fig. 4. The results of the GCMTF method were compared to the ACMTF method. To evaluate the robustness of our proposed method against noise, white Gaussian noise of different SNRs has been added to both the EEG and fMRI datasets. The SNR levels added to the data were selected from -15 dB to +15 dB. Although both methods have acceptable performance against different noise levels, the GCMTF method is more effective than the ACMTF method.

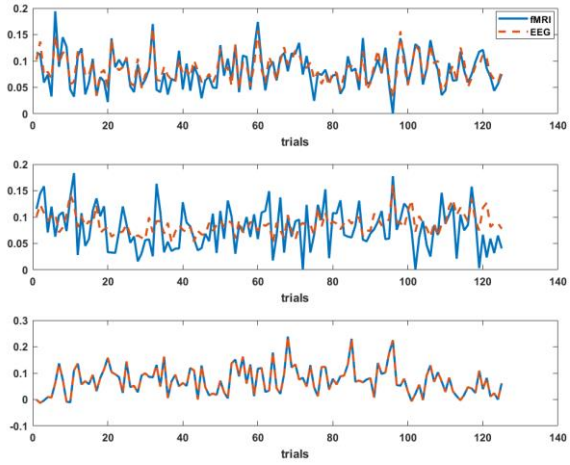


Fig. 4: From up to down, the temporal components of both modalities exhibit a linear correlation of 90%, 50%, and 100% (the same components).

Estimating the true components has decreased due to the assumption of equal components in the ACMTF method. Fig. 5 illustrates the weights of λ and σ estimated using the proposed GCMTF and ACMTF methods. Also, the shared components estimated by each method in the presence of high-level noise are depicted in Fig. 6 and Fig. 7.

In Fig. 8 the performance of the GCMTF method has been compared with the ACMTF method. The Match Score $MS = \frac{1}{R}\sum_{r=1}^{R} \hat{a}_r{}^T a_r \Big/ \|\hat{a}_r{}^T\| \|a_r\|$ is used to evaluate the results. In the MS relationship, $\hat{a}_r$ and a_r are the estimated and true values, respectively. The average match score for each simulated factor is illustrated in Fig. 8. The results indicate that the GCMTF method outperforms the ACMTF method. The average Match Score was raised by approximately 20% in the GCMTF model compared to the ACMTF model.

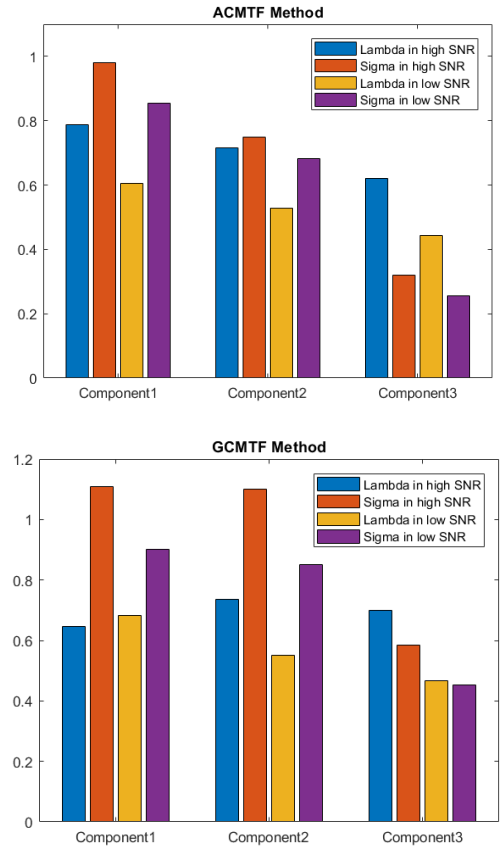


Fig. 5: The weights of λ and σ estimated using the proposed GCMTF and ACMTF methods.

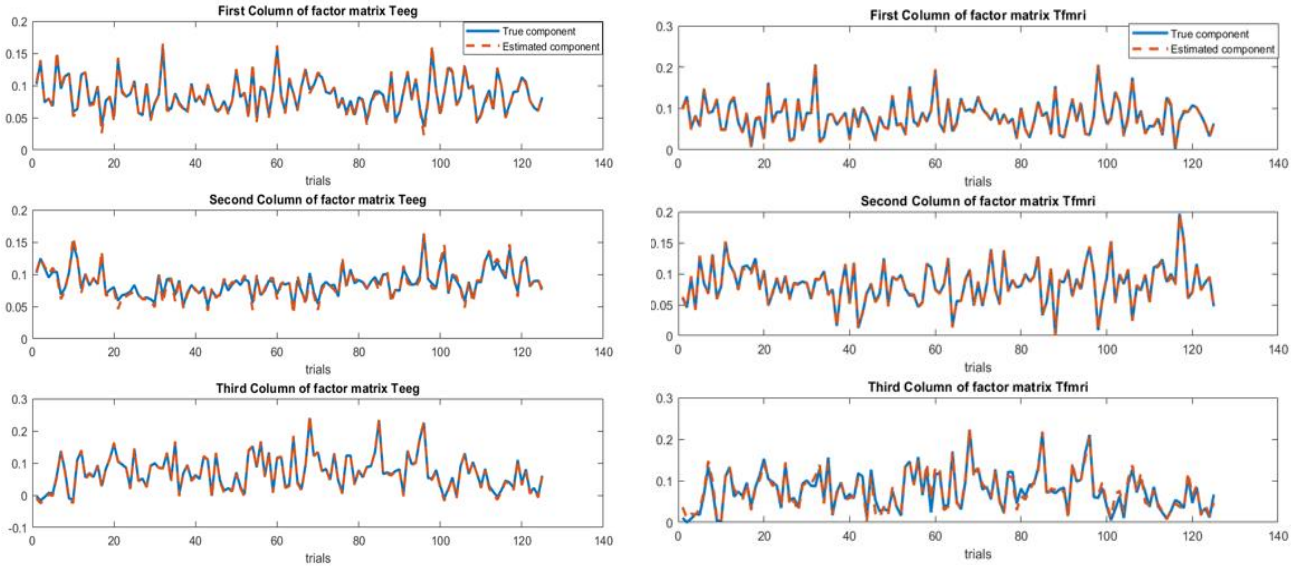
Discussion

The shared components in the common mode are assumed to be identical when fusing EEG and fMRI. However, this assumption is restrictive. Our proposed method replaced the equality assumption with a similarity measure.

Using normalized mutual information as a similarity measure, we can capture the differences between EEG and fMRI data that may not be fully accounted for by assuming identical components. This allows us to more accurately fuse information from both modalities and potentially uncover new insights that may have been overlooked with the traditional approach. Additionally, by quantifying the level of similarity between components, we can better understand the relationship between EEG and fMRI.

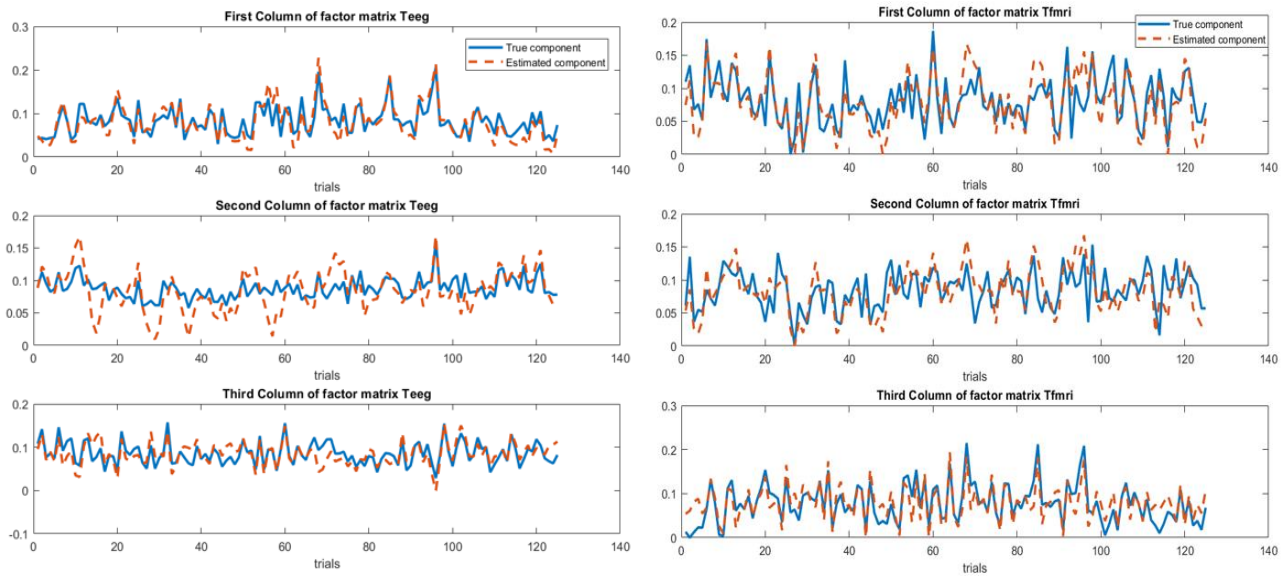
An NMI value approaching 1 indicates significant similarity, but when it nears zero, it means the opposite. Our simulations take into account three different levels of correlation.

The results indicate that our proposed GCMTF method significantly improves accurately estimating shared components with correlated temporal modes compared to existing methods like ACMTF. Moreover, the GCMTF method accurately estimates the weight values for each component corresponding to their existence in the dataset.



(a) (b)

Fig. 6: The components estimated by GCMTF method: (a) The estimated EEG components (T_{ee}) (b) The estimated fMRI components (T_{fm}).



(a) (b)

Fig. 7: The components estimated by ACMTF method: (a) The estimated EEG components (T_{ee}) (b) The estimated fMRI components (T_{fm}).

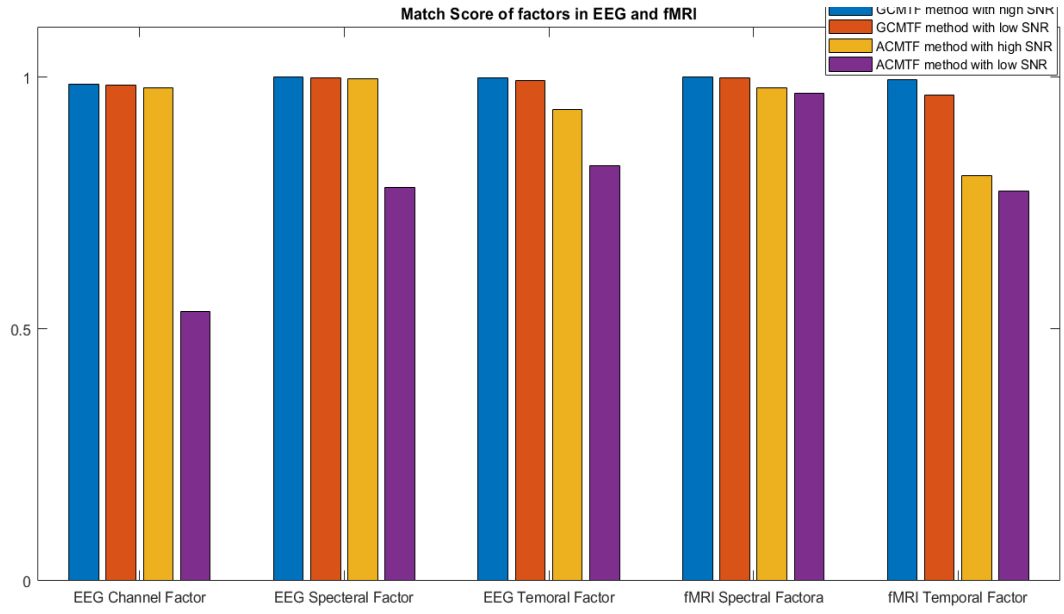


Fig. 8: Average match score (MS) between extracted factors by ACMTF and GCMTF and their ground truth at low and high SNR.

Conclusion

Overall, our proposed method offers a flexible and versatile approach to the fusion of multimodal data, providing a better understanding of the relationship between the two modalities. The application of this method can improve our capacity to study brain function in real EEG and fMRI data and opens up new possibilities for studying complex cognitive processes and neurological disorders. Although the GCMTF method is superior to the ACMTF, some modifications need to be made to improve its performance. Our method assumes that the HRF waveform is invariant for all brain voxels. However, the model can be more flexible by considering the variability in HRF across different subjects and brain regions in real data.

Furthermore, adopting other techniques to estimate MI rather than the histogram-based method may enhance the accuracy of estimating MI and entropy values. One alternative technique that could be explored

is kernel density estimation, which can provide a smoother and more continuous estimate of the underlying distribution. Additionally, advanced techniques such as neural networks or support vector machines could result in more precise and reliable estimations of mutual information and entropy in data analysis.

Author Contributions

Z. Rabiei: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data Curation, Visualization, Writing, Original Draft. H. Montazery Kordy: Conceptualization, Validation, Visualization, Editing the final version of the paper, and Supervision.

Acknowledgment

The authors would like to express their sincere gratitude to Dr. Raziyeh Mosayebi for her invaluable support, knowledge, and constructive feedback throughout this research project.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

EEG	Electroencephalogram
fMRI	functional magnetic resonance imaging
BOLD	Blood oxygenation level-dependent
HRF	Hemodynamic Response Function
GLM	General linear model
ACMTF	Advanced coupled matrix tensor factorization
NMI	Normalized mutual information
GCMTF	Generalized coupled matrix tensor factorization
MS	Match score

References

[1] Z. Jiang, Y. Liu, W. Li, Y. Dai, L. Zou, "Integration of simultaneous fMRI and EEG source localization in emotional decision problems," Behav. Brain Res., 448: 114445, 2023.

[2] D. Lahat, T. l. Adali, C. Jutten, "Multimodal data fusion: an overview of methods, challenges and prospects," Proc. IEEE, 103: 1449-1477, 2015.

[3] T. Warbrick, "Simultaneous EEG-fMRI: what have we learned and what does the future hold?," Sensors, 22: 2262, 2022.

[4] S. Van Eyndhoven, B. l. Hunyadi, L. De Lathauwer, et al., "Flexible fusion of electroencephalography and functional magnetic resonance imaging: Revealing neural-hemodynamic coupling through structured matrix-tensor factorization," in Proc. 25th European Signal Processing Conference (EUSIPCO): 26-30, 2017.

[5] H. K. Aljobouri, "Independent component analysis with functional neuroscience data analysis," J. Biomed. Phys. Eng., 13: 169, 2023.

[6] J. Sui, D. Zhi, V. D. Calhoun, "Data-driven multimodal fusion: approaches and applications in psychiatric research," Psychoradiology, 3: kkad026, 2023.

[7] L. Du, H. Wang, J. Zhang, S. Zhang, L. Guo, J. Han, A. S. D. N. Initiative, "Adaptive structured sparse multiview canonical correlation analysis for multimodal brain imaging association identification," Sci. China Inf. Sci., 66: 142106, 2023.

[8] R. F. Silva, S. M. Plis, T. l. Adali M. S. Pattichis, V. D. Calhoun, "Multidataset independent subspace analysis with application to multimodal fusion," IEEE Trans. Image Process., 30: 588-602, 2020.

[9] Y. Jonmohamadi, S. Muthukumaraswamy, J. Chen et al., "Extraction of common task features in EEG-fMRI data using coupled tensor-tensor decomposition," Brain Topogr., 33(5): 636-650, 2020.

[10] G. R. Poudel, R. D. Jones, "Multimodal neuroimaging with simultaneous fMRI and EEG," in Handbook of Neuroengineering: Springer, pp. 1-23, 2021.

[11] E. Acar, T. G. Kolda, D. M. Dunlavy, "All-at-once optimization for coupled matrix and tensor factorizations," arXiv preprint arXiv:1105.3422, 2011.

[12] E. Acar, M. A. Rasmussen, F. Savorani, et al., "Understanding data fusion within the framework of coupled matrix and tensor factorizations," Chemom. Intell. Lab. Syst., 129: 53-63, 2013.

[13] E. Acar, G. z. Gurdeniz, M. A. Rasmussen et al., "Coupled matrix factorization with sparse factors to identify potential biomarkers in metabolomics," Int. J. Knowl. Discovery Bioinf. (IJKDB), 3(3): 1-22, 2012.

[14] E. Acar, E. E. Papalexakis, G. z. Gurdeniz et al., "Structure-revealing data fusion," BMC Bioinf., 15(1): 1-17, 2014.

[15] E. Karahan, P. A. Rojas-Lopez, M. L. Bringas-Vega et al., "Tensor analysis and fusion of multimodal brain images," Proc. IEEE, 103(9): 1531-1559, 2015.

[16] B. Rivet, M. Duda, A. Guerin-Dugue, et al., "Multimodal approach to estimate the ocular movements during EEG recordings: a coupled tensor factorization method," in Proc. 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society: 6983-6986, 2015.

[17] C. Chatzichristos, E. Kofidis, L. De Lathauwer et al., "Early soft and flexible fusion of EEG and fMRI via tensor decompositions," arXiv preprint arXiv:2005.07134, 2020.

[18] R. Mosayebi, G. A. Hossein-Zadeh, "Correlated coupled matrix tensor factorization method for simultaneous EEG-fMRI data fusion," Biomed. Signal Process. Control, 62: 102071, 2020.

[19] Y. Maeda, H. Kawaguchi, H. Tezuka, "Estimation of mutual information via quantum kernel method," arXiv preprint arXiv:2310.12396, 2023.

[20] M. Babaie-Zadeh, C. Jutten, "A general approach for mutual information minimization and its application to blind source separation," Signal Process., 85: 975-995, 2005.

[21] T. O. Kvalseth, "On normalized mutual information: measure derivations and properties," Entropy, 19: 631, 2017.

[22] A. Cichocki, D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, H. A. Phan, "Tensor decompositions for signal processing applications: From two-way to multiway component analysis," IEEE Signal Process. Mag., 32: 145-163, 2015.

[23] N. M. Correa, T. Eichele, T. l. Adali, et al., "Multi-set canonical correlation analysis for the fusion of concurrent single trial ERP and functional MRI," Neuroimage, 50(4): 1438-1445, 2010.

[24] S. Van Eyndhoven, P. DuPont, S. Tousseyn, et al., "Augmenting interictal mapping with neurovascular coupling biomarkers by

structured factorization of epileptic EEG and fMRI data," Neuroimage, 228: 117652, 2021.

[25] M. Morante, "A lite parametric model for the hemodynamic response function," arXiv preprint arXiv:2004.13361, 2020.

[26] D. A. Handwerker, J. M. Ollinger, M. D'Esposito, "Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses," Neuroimage, 21: 1639-1651, 2004.

[27] Z. Y. Shan, M. J. Wright, P. M. Thompson, K. L. McMahon, G. G. Blokland, G. I. De Zubicaray, N. G. Martin, A. A. Vinkhuyzen, D. C. Reutens, "Modeling of the hemodynamic responses in block design fMRI studies," J. Cereb. Blood Flow Metab., 34: 316-324, 2014.

[28] M. W. Woolrich, T. E. Behrens, S. M. Smith, "Constrained linear basis sets for HRF modeling using Variational Bayes," Neuroimage, 21: 1748-1761, 2004.

[29] C. Gössl, L. Fahrmeir, D. P. Auer, "Bayesian modeling of the hemodynamic response function in BOLD fMRI," Neuroimage, 14: 140-148, 2001.

[30] C. Goutte, F. A. Nielsen, K. Hansen, "Modeling the hemodynamic response in fMRI using smooth FIR filters," IEEE Trans. Med. Imaging, 19: 1188-1201, 2000.

[31] H. Mohimani, M. Babaie-Zadeh, C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed l_0-norm," IEEE Trans. Signal Process., 57: 289-301, 2008.

[32] M. Babaie-Zadeh, C. Jutten, K. Nayebi, "Differential of the mutual information," IEEE Signal Process. Lett., 11: 48-51, 2004.

[33] J. M. Walz, R. I. Goldman, M. Carapezza, J. Muraskin, T. R. Brown, P. Sajda, "Simultaneous EEG-fMRI reveals a temporal cascade of task-related and default-mode activations during a simple target detection task," Neuroimage, 102: 229-239, 2014.

Biographies

Zahra Rabiei received her B.Sc. degree in Electronic Engineering from KN Toosi University of Technology,Tehran, Iran in 2001 and M.Sc. degree in Control Engineering from Ferdowsi university of Mashhad, Iran in 2005. She is currently a Ph.D. student in Biomedical Engineering in Babol Noshirvani University of Technology, Babol, Iran. The focus of her research is on biomedical signal processing, EEG, fMRI, and data fusion.

- Email: z.rabiei@stu.nit.ac.ir
- ORCID: 0000-0002-2048-8942
- Web of Science Researcher ID: NA
- Scopus Author ID:NA
- Homepage: NA

Hussain Montazery Kordy received his B.S. degree in Electronic Engineering from Mazandaran University, Babol, in 2000, the M.S. degree in Biomedical Engineering from Sharif University of Technology, in 2003 and the Ph.D. degree in Biomedical Engineering from Tarbiat Modarres University, Tehran, Iran, in 2009. Since 2010, he has been a member of the Electrical and Computer Engineering Faculty, Babol Noshirvani University of Technology, Babol, Iran. His research focuses on computer aided diagnosis, feature selection and extraction, and biomedical signal and image processing.

- Email: hmontazery@nit.ac.ir
- ORCID: 0000-0002-2010-4945
- Web of Science Researcher ID: AAD-3933-2022
- Scopus Author ID: 54386905700
- Homepage: https://ostad.nit.ac.ir/home.php?sp=389010

How to cite this paper:
Z. Rabiei, H. Montazery Kordy, "Utilizing normalized mutual information as a similarity measure for EEG and fMRI fusion," J. Electr. Comput. Eng. Innovations, 13(1): 141-150, 2025.

DOI: 10.22061/jecei.2024.10984.754

URL: https://jecei.sru.ac.ir/article_2207.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Segmentation of Skin Lesions in Dermoscopic Images Using a Combination of Wavelet Transform and Modified U-Net Architecture

S. Fooladi, H. Farsi, S. Mohamadzadeh *

Department of Electrical Engineering, Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.

Article Info

Article History:
Received 27 July 2024
Reviewed 12 September 2024
Revised 15 October 2024
Accepted 23 October 2024

Keywords:
U-Net
Segmentation
skin lesion
Deep neural networks
wavelet transform
Feature fusion
Medical image

*Corresponding Author's Email Address:
s.mohamadzadeh@birjand.ac.ir

Abstract

Background and Objectives: The increasing prevalence of skin cancer highlights the urgency for early intervention, emphasizing the need for advanced diagnostic tools. Computer-assisted diagnosis (CAD) offers a promising avenue to streamline skin cancer screening and alleviate associated costs.
Methods: This study endeavors to develop an automatic segmentation system employing deep neural networks, seamlessly integrating data manipulation into the learning process. Utilizing an encoder-decoder architecture rooted in U-Net and augmented by wavelet transform, our methodology facilitates the generation of high-resolution feature maps, thus bolstering the precision of the deep learning model.
Results: Performance evaluation metrics including sensitivity, accuracy, dice coefficient, and Jaccard similarity confirm the superior efficacy of our model compared to conventional methodologies. The results showed a accuracy of %96.89 for skin lesions in PH2 Database and %95.8 accuracy for ISIC 2017 database findings, which offers promising results compared to the results of other studies. Additionally, this research shows significant improvements in three metrics: sensitivity, Dice, and Jaccard. For the PH database, the values are 96, 96.40, and 95.40, respectively. For the ISIC database, the values are 92.85, 96.32, and 95.24, respectively.
Conclusion: In image processing and analysis, numerous solutions have emerged to aid dermatologists in their diagnostic endeavors The proposed algorithm was evaluated using two PH datasets, and the results were compared to recent studies. Impressively, the proposed algorithm demonstrated superior performance in terms of accuracy, sensitivity, Dice coefficient, and Jaccard Similarity scores when evaluated on the same database images compared to other methods.

Introduction

Over recent decades, there has been a notable rise in the incidence of skin cancer, underscoring the escalating significance of its initial treatment. Melanoma, the most lethal form of skin cancer, ranks among the most aggressive malignancies [1]. Automated segmentation of skin lesions in dermoscopic images is a crucial initial stage in utilizing computer assistance for diagnosing melanoma. Nonetheless, accurately discerning between benign and malignant skin lesions can be challenging, as there are considerable differences in lesion appearance across various patients. This ambiguity poses a diagnostic challenge even for seasoned medical professionals. Recent advancements in medical image processing have provided more effective techniques to aid dermatologists in diagnosing and classifying skin lesions. Therefore, computer-aided diagnosis (CAD) has become an indispensable tool for physicians and dermatologists,

especially when dealing with many patients in a short period [2]. The onset of this disease initiates with the impairment of skin cells, often instigated by ultraviolet radiation, resulting in mutations that prompt the rapid multiplication of skin cells, culminating in the development of cancerous growths. While typically characterized by a regulated and systematic growth pattern, specific newly generated cells may undergo unregulated proliferation, resulting in the formation of a cluster of malignant cells [3]. Early indications of melanoma often include alterations in the shape, size, and color of an individual's mole. Typically, melanomas exhibit a border that is black or blue-black in hue [3].

Automated skin lesion analysis relies heavily on segmentation, a crucial yet challenging process. Broadly, there are three main types of skin cancer:

(a) Basal Cell Carcinoma (BCC),

(b) Squamous Cell Carcinoma (SCC), and

(c) Melanoma (MM).

Segmentation essentially involves dividing an image into significant regions, with semantic segmentation specifically attributing suitable class labels to each region. In skin lesions, this typically entails two primary operations to delineate the lesion from the surrounding skin. The diagnosis of skin cancer is complex due to the diverse appearance of various skin lesions, notably Melanoma and Nevi, which pose challenges in differentiation. Despite the utilization of dermoscopy, a non-invasive diagnostic method, dermatologists' accuracy in diagnosing melanoma ranges from 75% to 84%. However, a biopsy offers a more precise diagnosis, albeit invasive and unpleasant for the patient. To prevent unnecessary biopsies, researchers have investigated various non-invasive methods for diagnosing melanoma [4]. These methods typically involve two stages: 1. Feature extraction, 2. Boundary (extent) identification of the skin lesion. Lesion segmentation is also useful as a preprocessing step when analyzing images with broad fields and multiple lesions. Effective clinical management of skin lesions relies heavily on timely diagnosis and accurate delineation of lesion boundaries to precisely identify the cancerous area for localization. Dermoscopy, employing visible light magnification, offers a more intricate skin examination compared to naked eye observation. We introduce a fully automated framework for precise detection and segmentation of lesion boundaries. This is accomplished by integrating a deep learning model with a wavelet transform map derived from specific kernel filters. The structure of this article is as follows:

Section 2 reviews existing literature in the field, highlighting significant contributions and advancements. Section 3 outlines the methodology proposed in this study, detailing the approach adopted for skin lesion segmentation. Section 4 presents the findings and results obtained through experimentation and evaluation of the proposed method. Finally, in Section 5, a comprehensive conclusion is drawn, summarizing the key insights gained from the research and outlining future directions for further investigation.

In recent years, significant advancements have been made in the field of medical imaging through various image processing techniques. One of the main challenges in this field is the precise and automatic segmentation of medical images for disease diagnosis and analysis. Medical image segmentation plays a crucial role in the early diagnosis and effective treatment of diseases. However, the accuracy and efficiency of existing algorithms still require improvement [2].

Deep neural networks, particularly the U-Net architecture, have been recognized as one of the successful architectures for medical image segmentation. U-Net, with its specific architectural design, is capable of extracting detailed and precise features from medical images. Nonetheless, it still faces limitations such as the need for a large volume of training data and high computational resource consumption [2].

One potential solution to improve U-Net's performance is the use of image preprocessing techniques such as wavelet transform. Wavelet transform, with its multi-resolution analysis capability, allows for the extraction of important and subtle features in medical images. These features can enhance segmentation accuracy and improve the performance of neural networks.

Therefore, the aim of this research is to investigate and propose a hybrid method combining wavelet transform and U-Net for improving medical image segmentation. It is expected that this approach will lead to increased accuracy and reduced segmentation errors in medical images, ultimately aiding in the timely diagnosis and treatment of diseases.

Using deep learning, it is possible to segment and detect various tumor tissues in medical images. Despite the potential difficulties, the accuracy of identifying and segmenting lesions in medical images is often accompanied by errors. An accurate and automated alternative to subjective and manual segmentation is segmentation using deep learning and computer systems. This method can achieve higher accuracy and be performed in a shorter time.

Given the deep network methods used for medical image segmentation, this research aims to leverage the strengths of various methods in the proposed model, ultimately leading to:

- Improved segmentation quality
- Reduced number of network parameters
- Reduced loss function

Related Work

In recent years, there has been notable interest among researchers in the pattern recognition and medical image processing domains towards automatic Computer-Aided Diagnosis (CAD) systems. The interest in image-based computer-aided diagnosis (CAD) systems has surged due to advancements in artificial intelligence and machine learning techniques. These developments have paved the way for creating CAD systems that utilize machine learning methods to analyze images, particularly for screening and early detection of malignant melanoma. As a result of recent technological and practical advancements, several emerging research and development areas have emerged. These areas have witnessed significant contributions from numerous researchers, resulting in a diverse range of CAD approaches and techniques. These advancements aim to assist dermatologists in automatically diagnosing melanoma from both dermoscopic and non-dermoscopic images [5], [6]. Lesion segmentation is an essential step for automatic melanoma detection. Numerous algorithms have been proposed by different researchers using various datasets employing methods such as thresholding, active contour, supervised and unsupervised techniques for segmenting dermoscopic skin lesion images. Automatic segmentation faces many challenges such as multicolor lesions, darkness at lesion boundaries, low contrast between lesion and normal skin, artifacts such as hair in the lesion area, and air bubbles due to gel applied to the skin in dermoscopy. The most commonly used methods for segmentation of skin lesions are traditional methods such as threshold based methods, clustering methods and correlated methods [7], [8]. Generally, threshold-based methods are used to extract regions of interest based on pixel intensity values. Therefore, image thresholding in grayscale transforms the image into a binary image separating the foreground from the background. Recently, research results have shown that deep learning models significantly contribute to medical image analysis for segmentation and Segmentation purposes [9], [10]. In their study [11], researchers introduced a hierarchical framework for skin lesion segmentation. Their approach involved an initial non-coherent operation, followed by passing the data to MASK RCNN for lesion segmentation. In the subsequent stage, they adapted a pre-trained DenseNet201 model and extracted features from two layers. These extracted features underwent refinement and enhancement using a combined selection block and were optimized using the salient optimization algorithm. The experimental evaluation was conducted on three dermoscopic datasets, demonstrating the enhanced performance of their proposed method. In their work [12], an intelligent framework for multi-class skin lesion segmentation was introduced. The method involved the initial segmentation of skin lesions using MASK RCNN. During this segmentation process, a 24-layer CNN model was employed, utilizing three datasets for the segmentation phase alongside the HAM10000 dataset. In their study [13], a Computer-Aided Diagnosis (CAD) system for localizing skin lesions was introduced. The process began with an initial incoherent operation by passing the data to MASK RCNN for lesion segmentation. Subsequently, they adjusted a pre-trained DenseNet201 model, extracting features from two layers. These features underwent refinement and enhancement using a combined selection block. The experimental evaluation was conducted on dermoscopic datasets, showcasing improved performance. In their paper [14], they presented a deep learning architecture tailored for skin lesion segmentation. The approach began by selecting the most optimal features to enhance the representation of lesion areas. Subsequently, an initial RCNN was deployed for the final segmentation of lesions. Dermoscopic datasets were utilized for evaluation purposes, demonstrating an enhancement in accuracy. In [15], an alternative segmentation method named "Fast Learning Artificial Neural Network" (FLANN) was introduced, serving as the foundation for an image segmentation technique. The study initiated noise reduction in the initial phase, employing a mean filter (3×3) to mitigate color distribution disparities. Following this step, pixels or neurons were converted into the R-G-B-S-V space via HSV conversion. FLANN clustering was then utilized to produce image clustering results, effectively segregating pixels of identical colors. Each image segment was allocated a distinct identifier, with close attention given to neighborhood size and tolerance effects.

In [16], a pioneering approach to skin lesion detection is presented, integrating uniform segmentation and feature selection into a cohesive strategy. This method encompasses various stages including preprocessing, lesion segmentation, feature extraction, feature selection, and final segmentation. Through a sequentially serial process, features such as color, texture, and HOG shape are extracted and combined. Subsequently, the Boltzmann entropy technique is employed for feature selection, followed by SVM classification. The effectiveness of this method is evaluated using the PH2 dataset, achieving promising results with a reported sensitivity of 97.7%, specificity of 96.7%, accuracy of 97.5%, and F-score of 97.5%.

In [17], a notable advancement in artificial hair removal was demonstrated through the fusion of deep learning and image processing techniques, yielding an accuracy of 85%. The study harnessed the Unet model for lesion segmentation, supplemented by image processing algorithms. Mainly, a Gaussian filter was applied to

diminish image noise. In contrast, [18] focused on enhancing accuracy by employing encoders like EffectNet and ResNet, achieving an accuracy of nearly 86% with the ResNet network. Furthermore, Shin et al. [19] introduced the DSM model, implementing strategies to refine segmentation accuracy by eliminating image noise and enhancing contrast.

In their study [20], researchers employed an extended U-Net network for medical image segmentation, capitalizing on the benefits of the U-Net architecture, including its compactness and skip connections. Additionally, they integrated bidirectional ConvLSTM and dense convolution mechanisms. The study focused on enhancing segmentation performance by incorporating Squeeze and Excitation modules, aiming to minimize complexity while improving results.

In their work [21], researchers introduce a multi-scale U-Net for skin lesion segmentation to address challenges like significant variations in texture and shape. This approach enhances hierarchical modeling by integrating an attention mechanism. Furthermore, it employs a bidirectional convolutional long short-term memory (BDCLSTM) structure to capture essential distinguishing features while suppressing less informative elements.

In [22], a novel segmentation method is proposed using fully convolutional networks (FCNs). This approach directly learns the full-resolution features of each pixel from the input data, eliminating the need for preprocessing or post processing operations such as artifact removal, low-contrast adjustment, or additional enhancements to improve the delineation of segmented skin lesions.

In [23], an enhanced skin lesion segmentation model based on U-Net++ is introduced to improve survival rates for melanoma patients and overcome associated challenges. A novel loss function is introduced to enhance the Jaccard segmentation index for skin lesion segmentation. Experimental results show the model's outstanding performance in segmenting the ISIC2018 I dataset, achieving an impressive Jaccard index of 84.73%. This method improves the Jaccard segmentation index for skin lesion images, aiding dermatologists in identifying and diagnosing various skin lesions while accurately delineating boundaries between lesions and normal skin.

In [18], the study presents an automated approach for segmenting lesion boundaries by combining two architectures, U-Net and ResNet, into a unified framework called Res-Unet. Moreover, image colorization eliminates unwanted hair, leading to notable enhancements in segmentation outcomes.

In [24], an exceptionally effective segmentation method is proposed to address challenges like unwanted residues (hair), uncertain boundaries, variable contrast, shape differences, and color variations in skin lesion images. The method introduces an improved FCN architecture (iFCN) tailored for segmenting high-resolution skin lesion images without needing preprocessing or post-processing. Leveraging residual structures within the FCN architecture, along with spatial information, enhances segmentation accuracy significantly.

In the study described in [25], a new CNN architecture is introduced, utilizing auxiliary information to enhance segmentation performance. Edge prediction is incorporated as an auxiliary task, running simultaneously with the main segmentation task. A cross-connection layer module is introduced, allowing intermediate feature maps from each task to influence the sub-blocks of other tasks. This approach implicitly guides the neural network to focus on the boundary area crucial for accurate segmentation.

In the study outlined in [26], two innovative end-to-end segmentation models, FBUNet-1 and FBUNet-2, are introduced. FBUNet-1 surpasses the performance of the traditional U-Net architecture by addressing spatial information loss during convolution operations. Building upon the progress of FBUNet-1, FBUNet-2 further improves accuracy by refining the loss function based on FBUNet-1's insights.

- U-Net Based Segmentation Techniques

In some articles: Introduced the U-Net architecture, which became a seminal work in the field of medical image segmentation. The U-Net's encoder-decoder structure, coupled with skip connections, enables precise localization and contextual understanding, making it effective for segmenting medical images such as MRI and CT scans. However, the model requires a large amount of annotated data and significant computational resources.

In some articles: Extended the U-Net architecture to 3D U-Net for volumetric medical image segmentation. This extension maintained the benefits of the original U-Net but adapted it to 3D data, which is essential for applications involving volumetric data such as MRI. The challenge remained in the increased computational cost and memory usage.

In some articles: Proposed a hybrid approach combining U-Net with conditional random fields (CRFs) to refine the segmentation output. The CRFs helped in capturing fine details and addressing segmentation boundaries more effectively. While the method improved accuracy, it also introduced additional complexity and computational overhead.

- Wavelet Transform in Medical Imaging

In some articles: Applied wavelet transform for ECG signal processing, showcasing its versatility beyond imaging. The study highlighted the wavelet's capability in dealing with non-stationary signals, an attribute valuable

in dynamic medical imaging contexts.

- Hybrid Approaches Combining U-Net and Other Techniques

In some articles: Introduced Attention U-Net, which incorporated attention mechanisms into the U-Net architecture to focus on relevant parts of the image. This approach aimed to enhance the model's ability to distinguish between foreground and background, improving segmentation accuracy without significantly increasing computational demands.

In some articles: Proposed UNet++ with nested and dense skip connections to address the issues of semantic gap between encoder and decoder. The model achieved better segmentation results but at the cost of increased complexity and training time.

In some articles: Combined U-Net with dilated convolutions to capture multi-scale context without reducing the resolution. This method aimed to improve the receptive field of the network, enhancing its ability to segment larger structures in medical images.

- Deep Learning for Medical Image Segmentation Beyond U-Net

In some articles: Provided a comprehensive survey of deep learning techniques for medical image analysis, including segmentation, classification, and detection. The review highlighted the dominance of CNN-based architectures and the emerging trends in integrating other techniques like GANs for improving segmentation.

In some articles: Reviewed various deep learning models for medical image segmentation, focusing on the strengths and weaknesses of different approaches. The study emphasized the importance of model robustness and generalizability across different datasets and imaging modalities.

In some articles: Discussed the use of transfer learning and multi-task learning in medical image segmentation, stressing the need for models that can leverage pre-trained knowledge and simultaneously learn related tasks for improved performance.

The reviewed literature highlights several key trends and insights in the field of medical image segmentation:

- U-Net and Its Variants: The U-Net architecture and its extensions (3D U-Net, Attention U-Net, UNet++) have proven highly effective for medical image segmentation. These models leverage skip connections and multi-scale feature extraction to achieve high accuracy. However, they require large datasets and significant computational resources, which can be a limitation in practical applications.

- Wavelet Transform: Wavelet transform offers a robust method for feature extraction and enhancement in medical imaging. Its multi-resolution analysis capability complements deep learning models by providing detailed spatial and frequency information, which is crucial for accurate segmentation.

- Hybrid Approaches: Combining U-Net with other techniques such as attention mechanisms, CRFs, and dilated convolutions can further enhance segmentation accuracy. These hybrid models address specific limitations of the original U-Net by improving feature localization, boundary detection, and multi-scale context understanding.

- Beyond U-Net: While U-Net remains a dominant architecture, other deep learning models and techniques are also being explored. The integration of transfer learning, multi-task learning, and GANs shows promise in enhancing segmentation performance and addressing challenges like data scarcity and model generalizability.

The literature suggests that while U-Net and its variants are highly effective for medical image segmentation, there is still room for improvement in terms of accuracy, computational efficiency, and robustness. Hybrid approaches that integrate wavelet transform and other techniques offer a promising direction for future research. Additionally, exploring alternative deep learning models and leveraging advanced techniques like transfer learning and multi-task learning could further advance the field and lead to more practical and reliable segmentation solutions.

Regarding the limitations of previous works, it can be noted that traditional methods were often used for feature extraction. In our research, however, we utilize deep learning-based methods for this purpose.

By leveraging the strengths of the mentioned works in this research, the use of mutual connection modules is proposed. In these modules, the intermediate feature maps of each task are fed into the sub-blocks of other tasks, which can implicitly guide the neural network to focus on the boundary region of the segmentation task.

Proposed Method

Block Diagram 2 provides an overview of the skin lesion segmentation in skin images. The proposed method initiates by extracting the feature map of the images utilizing wavelet transform, followed by selecting top features using selected feature algorithms. Subsequently, a fully automated architecture is introduced for precise lesion boundary detection and segmentation. This architecture pairs a deep learning model with the feature map derived from specific kernel filters of wavelet transform. By integrating the feature maps of wavelet transform with a deep learning U-Net model trained end-to-end, our method effectively reduces the number of trainable parameters.

The Fourier transform is used to represent the frequency properties of a signal by decomposing the

signal into different sinusoids. However, this method does not provide simultaneous signal resolution in both the time and frequency domains and gives less information for signals with varying frequency over time.

We chose Wavelet Transform for our analysis due to its ability to effectively capture both time and frequency domain characteristics of signals, making it suitable for analyzing non-stationary signals. Unlike the fixed window length constraint of Short-Time Fourier Transform (STFT), Wavelet Transform offers more flexibility, allowing for better adaptation to varying signal properties.

Wavelet transform is a fundamental mathematical tool with diverse applications across scientific fields. It addresses the limitations of the Fourier transform by excelling in analyzing non-stationary signals and dynamic systems. Unlike the Fourier transform, wavelets exhibit localization properties in both space and frequency domains. This unique characteristic enables examining spatial frequency content in signals without sacrificing positional information. Therefore, wavelet transform offers a valuable balance between pixel and Fourier space representation. The wavelet transform harnesses various essential features, including spatial localization, frequency band tuning, directionality, scale and rotational similarities, and quadrature phase. It operates through a collection of mother wavelets and a scaling equation dictating the movements of these wavelets via scaling and translation. This transformation comprises eight mother wavelets grouped into four pairs, each characterized by distinct orientations (0°, 45°, 90°, and 135°). Each pair consists of one odd-symmetric and one even-symmetric wavelet, symbolized by ϕ and ψ. The eight mother wavelets are denoted as ψ_1,...,ψ_8. Let's define a two-dimensional pulse u(x, y) [26]:

$$u(x,y)\begin{cases}1 \text{ if } 0<x\le 1,\ 0<y\le 1\\ 0 \text{ otherwise}\end{cases} \tag{1}$$

Next, we identify the four pairs of mother wavelets as follows [26]:

$$\beta_{0,e}(x,y)=\left[-u\left(x,y/3\right)+2u(x-1,y/3)-u\left(x-2,y/3\right)\right]/\sqrt{18} \tag{2}$$

$$\beta_{0,O}(x,y)=\left[-u\left(x,y/3\right)+u(x-2,y/3)\right]/\sqrt{6} \tag{3}$$

$$\beta_{45,e}\left(x,y\right)=\begin{bmatrix}-\left(u\left(x,(y-1)/2\right)+u\left(x-1,y\right)+u\left(x-1,y-2\right)+u\left(x-2,y/2\right)\right)\\ +2(u\left(x,y\right)+u\left(x-1,y-1\right)+u\left(x-2,y-2\right))\end{bmatrix}/\sqrt{18} \tag{4}$$

$$\beta_{45,o}(x,y)=\begin{bmatrix}-u(\left(x,y-2\right)+u(x-1,y)+u\left(x-2,y-1\right))+u(x,y-1)\\ +u\left(x-1,y-2\right)+u\left(x-2,y\right)\end{bmatrix}/\sqrt{6} \tag{5}$$

$$\beta_{90,o}\left(x,y\right)=\left[u\left(x/3,y\right)-u\left(x/3,y-2\right)\right]/\sqrt{6} \tag{6}$$

$$\beta_{90,e}\left(x,y\right)=\left[-u\left(x/3,y\right)+2u\left(x/3,y-1\right)-u\left(x/3,y-2\right)\right]/\sqrt{18} \tag{7}$$

$$\beta_{135,o}\left(x,y\right)=\begin{bmatrix}-(u\left(x,y\right)+u\left(x-2,y-1\right)+u\left(x-1,y-2\right))\\ +u\left(x,y-1\right)+u\left(x-1,y\right)+u\left(x-2,y-2\right)\end{bmatrix}/\sqrt{6} \tag{8}$$

$$\beta_{135,e}\left(x,y\right)=\begin{bmatrix}-(u\left(x,y/2\right)+u\left(x-1,y\right)+u\left(x-1,y-2\right))\\ +u\left(x-2,(y-1)/2\right)+2u((x,y-2)+u\left(x-,y-1\right)\\ +u\left(x-2,y\right))\end{bmatrix}/\sqrt{18} \tag{9}$$

Mother wavelets are piecewise constant functions, so that each wavelet $b_{\varphi,\phi}$ can be fully described by a 3×3 matrix.

Fig. 1 illustrates eight mother wavelets, demonstrating the transformation process. Part A showcases these wavelets in pixel space, while Part B exhibits the discrete Fourier transform (DFT) of the 3×3 mother wavelets. The selection of wavelets is such that each DFT encompasses a pair of pixels, minimizing their spatial frequency and bandwidth orientation within this space.

Furthermore, to ensure orthogonality, four pairs of wavelets are included, each oriented in one of four different directions (0°, 45°, 90°, and 135°). These pairs are composed of one wavelet with a real symmetric even DFT, corresponding to an even symmetric function in pixel space, and the other wavelet with an imaginary symmetrical odd DFT, corresponding to an odd symmetrical function in pixel space. The 3×3 DFTs allow qualitative similarity to Gabor filters. The 3×3 DFTs bear qualitative resemblance to Gabor filters. As the wavelets undergo higher-resolution decomposition, they occupy less space in the Fourier domain [26].

$b_{0,e}=\frac{1}{\sqrt{18}}\begin{bmatrix}-1&2&-1\\-1&0&-1\\-1&0&-1\end{bmatrix}$	$b_{0,o}=\frac{1}{\sqrt{6}}\begin{bmatrix}-1&-1&-1\\0&0&0\\1&1&1\end{bmatrix}$
$b_{90,o}=\frac{1}{\sqrt{6}}\begin{bmatrix}-1&-1&-1\\0&0&0\\1&1&1\end{bmatrix}$	$b_{45,o}=\frac{1}{\sqrt{18}}\begin{bmatrix}-1&1&0\\1&0&-1\\0&-1&1\end{bmatrix}$
$b_{90,e}=\frac{1}{\sqrt{18}}\begin{bmatrix}-1&-1&-1\\2&2&2\\-1&-1&-1\end{bmatrix}$	$b_{90,o}=\frac{1}{\sqrt{6}}\begin{bmatrix}-1&-1&-1\\0&0&0\\1&1&1\end{bmatrix}$
$b_{135,e}=\frac{1}{\sqrt{18}}\begin{bmatrix}2&-1&-1\\-1&2&-1\\-1&-1&2\end{bmatrix}$	$b_{135,o}=\frac{1}{\sqrt{6}}\begin{bmatrix}0&-1&1\\1&0&-1\\-1&1&0\end{bmatrix}$

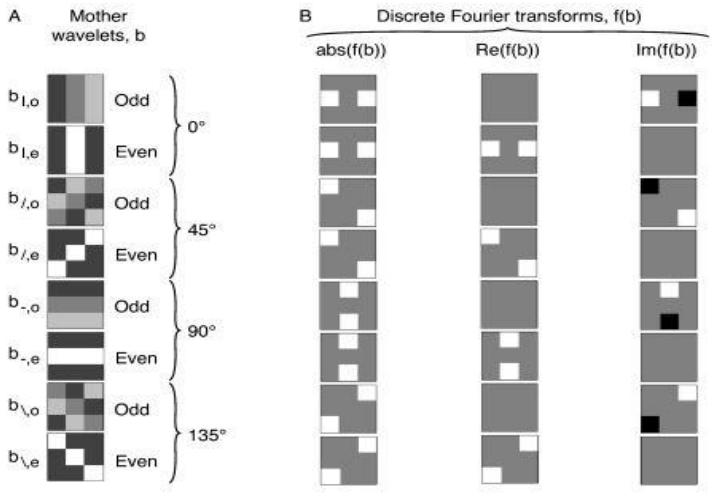


Fig. 1: Eight mother wavelets.

Feature selection involves identifying pertinent features while discarding irrelevant and redundant ones, intending to obtain a subset of features that adequately describes the problem with minimal loss of performance. Essentially, it is the process of selecting the most essential features to represent the data accurately. This task significantly improves the proposed method by removing irrelevant features and presenting the most useful ones. Some advantages of feature selection in our research include:

- Enhancing the performance of machine learning algorithms.
- Facilitating understanding of data and gaining insights into the underlying processes, aiding visualization.
- Decreasing overall data volume, thereby reducing storage requirements and potentially cutting costs.
- Streamlining the feature set, which can save resources in future data collection or utilization phases.
- Promoting simplicity and enabling simpler models, thereby enhancing speed and efficiency.

To identify a relevant feature for the problem, the study employs the following definition: a feature is deemed relevant if it holds information pertinent to the objective.

In this research, significant factors are considered to enhance the accuracy of the proposed method within the feature selection framework.

The initial stage of the proposed method involves identifying the nearest neighbors from a subset of samples randomly selected from the dataset. For each chosen sample, the feature values are compared to those of its nearest neighbors, and the scores of each feature are adjusted accordingly.

This methodology is rooted in assessing feature quality by gauging the degree of variation in their values among neighboring samples.

In the second stage of the feature selection method, correlation-based feature selection (CFS) is employed. This method deems a subset of features as favorable if, on the one hand, they exhibit a strong correlation with the target feature and, on the other hand, they are minimally correlated with each other. In this research, the merit or goodness of a feature subset is computed using the following formula [2]:

$$Merit_{s_k} = \frac{k\bar{r}_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}} \tag{10}$$

In this formula, $\bar{r}_{cf}$ the first term represents the average correlation between the target feature and $\bar{r}_{ff}$ all the features present in the dataset, while the second term represents the average pairwise correlation calculated among the features. Ultimately, the correlation-based method is formulated as follows [2]:

$$CFS = \max_{s_k}\left[\frac{r_{cf1}+r_{cf2}+\ldots\ldots+r_{cfk}}{\sqrt{k+2\left(r_{f2f1}+\ldots.+r_{fifj}+\ldots.+r_{fkf1}\right)}}\right] \tag{11}$$

In this equation, the variables r_{cfi} , r_{fifj} represent the correlation values. The correlation-based method is used for selecting the best features.

Next, the Recursive Feature Elimination (RFE) method is introduced, employing a multivariate mapping approach that iteratively constructs a model and selects the most discriminative features, whether the best or worst-performing ones.

In RFE, all voxels within a region, constituting the smallest unit of a three-dimensional image structure, are considered. Voxels that do not contribute to distinguishing features among different classes are progressively eliminated. Features are then ranked according to the order of their elimination. Thus, this method operates as a greedy optimization technique to identify the optimal subset of features.

Lastly, all features undergo decoding via statistical hypothesis testing, and each feature is ranked based on its values. Feature weights are determined utilizing the chi-square statistical method. A function evaluates the significance of a feature by computing the chi-square statistic and subsequently ranks the features accordingly, taking into account their respective classes. This ranking procedure is executed to compare the features and identify their relative importance.

Features selected through the above methods often have different ranges, and classifiers typically require normalized features because their values fall within a specific range. One of the most common normalization methods is the z-score normalization method [2]:

$$Z = (x - \mu)/\sigma \tag{12}$$

where μ denotes the mean and σ denotes the standard deviation from the mean. This method generates a range of values between 0 and 1. Below is Algorithm 1 outlining the overall process of this research in the form of pseudocode.

As can be seen in Fig. 2, the proposed method uses a number of unique methods to extract features and normalizes the features that often have different ranges to a specific range by using the score method.

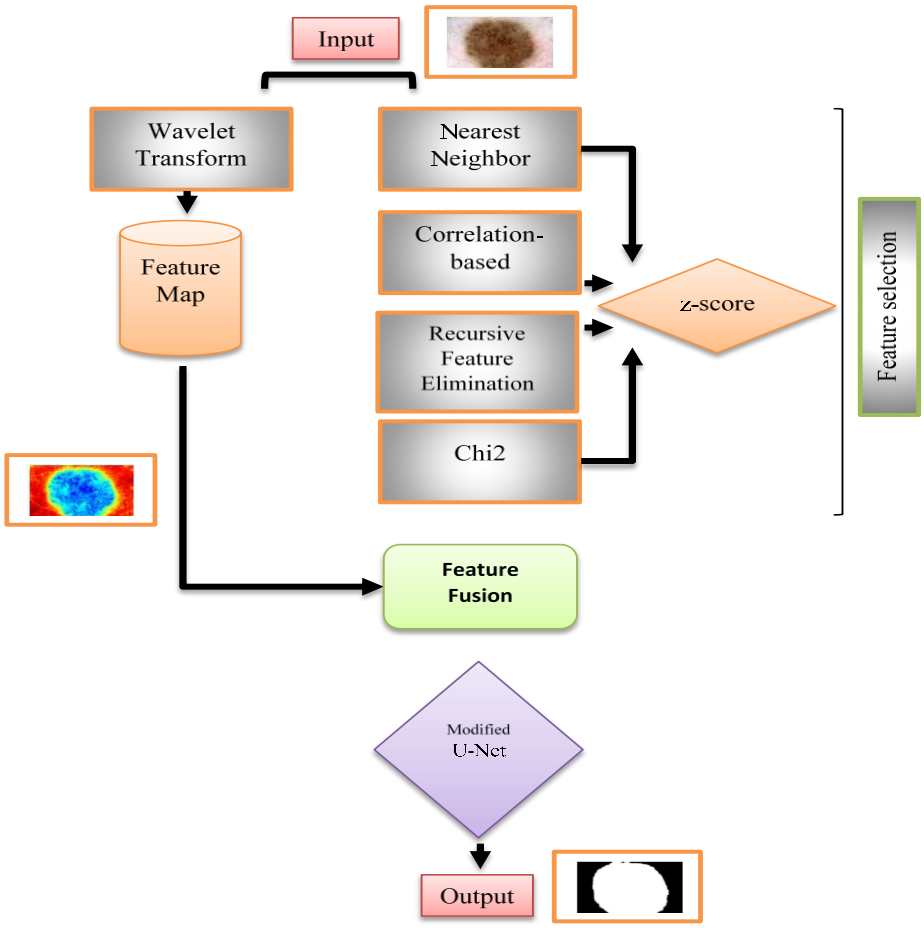


Fig. 2: Sequence of components in the proposed method.

Algorithm 1: Algorithm of the proposed method

Input:
trainImgSet: The medical images Set, with segmented skin lesion areas manually in theirs;
targets = The segmented skin lesion areas manually in *1*.trainImgSet.
get N = The number of images in trainImgSet
get wavelet = The wavelet transform according to Equation (1,2,..,9)
get NN = The Nearest Neighbor algorithm
get Cb= Correlation-based algorithm
get RFE= Recursive Feature Elimination algorithm
get ch = The Chi2 algorithm
get z_s = The z-score according to Equation (12)
get FM = The Feature Map resulting from wavelet transform
2.for i = 1 to N do:
WT[i]= wavelet(trainImgSet[i])
WT_b[i]= The Feature Map resulting from wavelet transform
Z [i] = The z-score features extracted of (NN,Cb, RFE, ch)
3.for region in the RI do :
RCI=U-Net(region)
add RCI to U-Net _Features
Selcted_features= Applying feature selection algorithm and Feature Map resulting from wavelet transform
4. Segmented_features=U-Net(Selcted_features)
Output:
Segmented_features = an image, that lesion pixels are distinguished.

Proposed Network Architecture

Fig. 3 illustrates the proposed architecture for the automatic segmentation of skin lesions. This method integrates wavelet transform feature maps and selected features via feature selection methods into a U-Net deep learning model trained end-to-end. The architecture comprises two encoder branches for abstract feature extraction and one decoder path for reconstruction.

The contraction path involves a sequence of convolutional and max-pooling layers aimed at reducing the size of the input image and extracting relevant features. Conversely, the expansion path comprises a sequence of convolutional and upsampling layers responsible for increasing the size of the feature maps obtained from the contraction path and integrating them with features from the input image to generate the final segmentation map. Skip connections allow information to bypass one or more levels within the expansion path and connect them to corresponding layers in the contraction path. These connections facilitate the transmission of high-level and low-level information from the input image to the model, thereby improving the accuracy of the final segmentation.

Broadly, the proposed architecture extracts features from the input image through the contraction path. These

features are then amalgamated with input image features via the expansion path and skip connections. Finally, the convolutional layers within the expansion path are employed to produce the final segmentation map.

The encoder architecture proposed in our study comprises seven convolutional layers and three max-pooling operations, each utilizing a stride of 2. These convolutional layers extract feature maps from the input image through 3x3 kernel convolutional operations. The study adopts a cautious approach to kernel sizes, starting with smaller kernels and gradually increasing them if necessary to avoid computational overhead and overfitting risks. To cover larger receptive fields without increasing the parameters linked to each kernel, expanded convolutional layers are utilized at every encoder level.

In addition, in order to reduce the size of the extracted feature maps, max-pooling operations are used. This process optimizes memory utilization by retaining only the pixel with the maximum value among the four neighboring pixels. However, it results in a loss of feature map resolution.

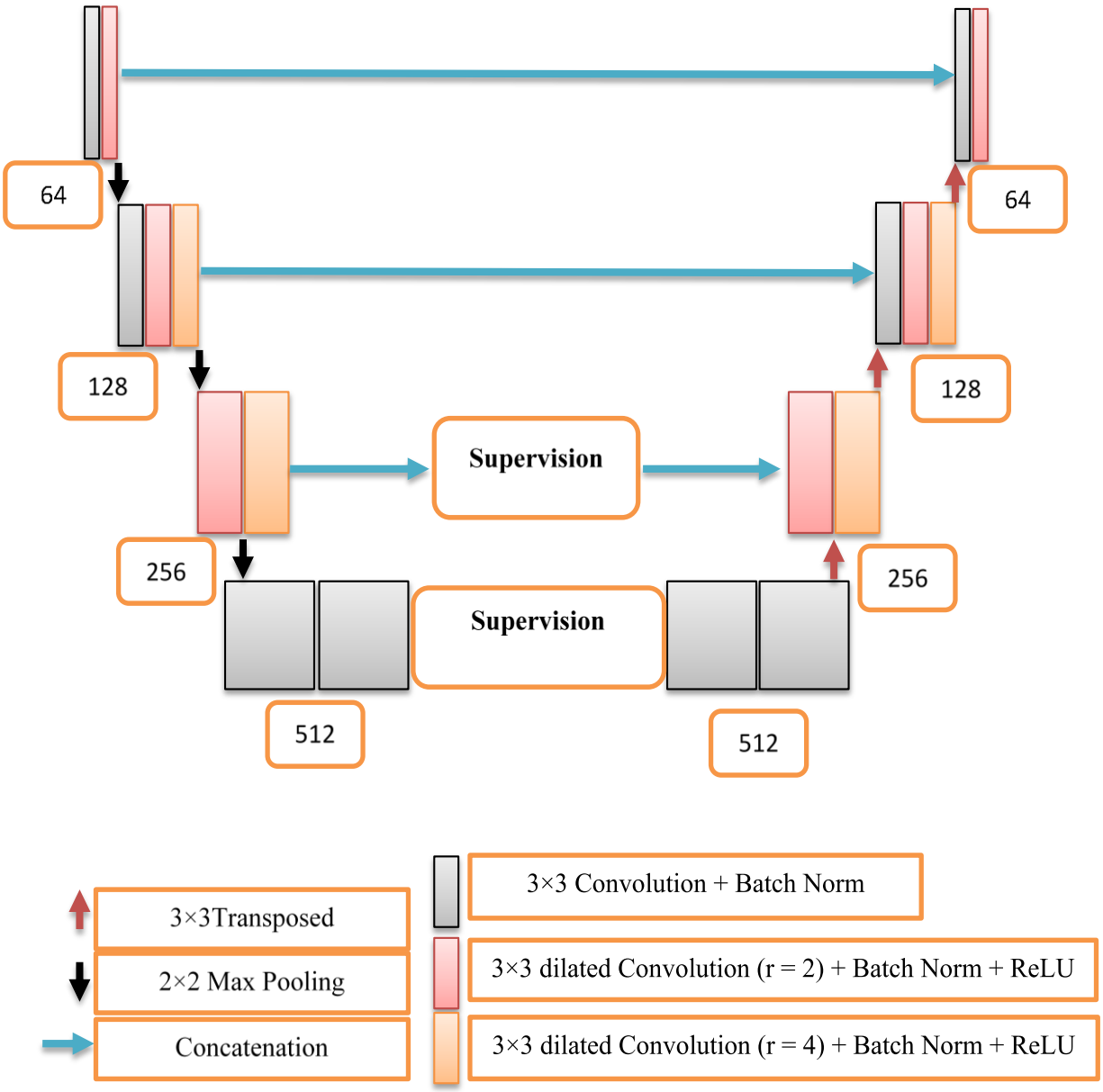


Fig. 2: Proposed architecture.

The encoder includes a convolutional block, consisting of two standardized convolutional layers in addition to the supervision layer and two other standard convolutional layers. The supervision block plays a crucial role in gathering information at various levels to incorporate more features. Meanwhile, the decoder utilizes alternating 3x3 deconvolutions to decrease the number of feature maps. Skip connections recover spatial information lost during pooling operations, following a similar approach to U-Net. Initially, the feature map is inputted to the control block and, except for the last stage, it will merge with corresponding maps from the decoder path. To generate the output segmentation map, the decoder employs a 1x1 convolution layer followed by a sigmoid layer, serving as a pixel-wise classifier. Each 3x3 convolution is accompanied by a batch normalization (BN) layer and a modified ReLU activation function to expedite the training process.

In the expanded convolution segment, drawing inspiration from reference [28], rather than downsampling feature maps at a low rate, the feature maps from the expanded convolution are harnessed for image segmentation. The receptive field of a kernel, denoted as k with a size of N×N, can be defined as follows [2]:

$$R_k = N + (N-1)(r-1) \tag{13}$$

In this equation, N = 3 (the kernel sizes are considered constant). The r represents the dilation rate, which specifies the spacing between values of the filter. In typical scenarios, r = 1 is assumed.

In this research, the supervision block is instrumental in refining localization and attaining more precise segmentation by directing attention to specific regions within an image. Within the proposed U-Net architecture, these supervision blocks aid in transferring essential features through skip connections. By leveraging contextual information, these blocks achieve a focus on specific regions. Before the concatenation operation, they effectively filter out noise and other irrelevant details from high-level features, ensuring accurate feature transmission.

Utilizing features extracted from deep networks leads to the identification of complex relationships in the image, thereby resulting in more precise segmentation results [34], [35].

The graphic view shown in Fig. 3 simply and in detail shows the proposed Unet network and its components.

Features extracted from deep networks help identify complex relationships within images, which directly lead to improved segmentation accuracy. Unlike traditional methods that may focus only on simple and distinct patterns in the data, deep networks, especially architectures like U-Net, are capable of learning features that not only capture high-level information (such as large structures) but also take into account fine details.

For example, in some studies, the use of deep networks for liver segmentation from CT scan images has demonstrated how these architectures can extract complex and precise features, providing significantly more accurate results compared to traditional methods. This improvement in accuracy is due to the ability of deep networks to learn various features across different layers, leading to segmentation results that are not only visually superior but also encompass finer and more detailed structures.

Moreover, in another study focused on enhancing the feature space using the deep network SqueezeNet, it has been observed that deep networks can lead to improved segmentation in textural images. This is particularly important in images with textural and complex patterns, as deep networks are capable of learning and recognizing intricate patterns in the images, resulting in a significant improvement in the final outcomes.

Results Database

The PH2 database results from collaborative efforts between the University of Porto, Instituto Superior Técnico Lisbon, and the Dermatology Services of Pedro Hispano Hospital in Matosinhos, Portugal. This database comprises 200 dermoscopic images, carefully curated to include 80 common nevi, 80 atypical nevi, and 40 melanomas. The selection process prioritized image quality, clarity, and the presence of dermoscopic features. Each image underwent evaluation by a dermatologist, considering parameters such as manual segmentation of the skin lesion, clinical diagnosis, histopathology (if available), and dermoscopic criteria, including asymmetry, colors, pigment network, dots/globules, streaks, and regression areas.

The ISIC2017 database, another crucial aspect of this study, consists of 2600 images designated for skin lesion analysis. Among these, 2000 images serve as training samples, while the remaining 600 images are reserved for testing purposes. The data used in this research consists of images captured with regular photographic cameras, not a specialized medical imaging camera.

The ISIC database is a public and open source for skin images. This database contains dermatological images with high quality and quality control. These images are used as a public resource for training, research and development and testing of diagnostic artificial intelligence algorithms. This database can be used to improve clinical diagnostic skills and provide support in skin cancer diagnosis. Also, the development and testing of algorithms for skin cancer triage and diagnosis also utilizes this database.

By using the database images shown in Fig. 4 and Fig. 5, the steps of the proposed method in Fig. 2 are performed in order and finally an effective division of the database images is presented as the output of the proposed network.

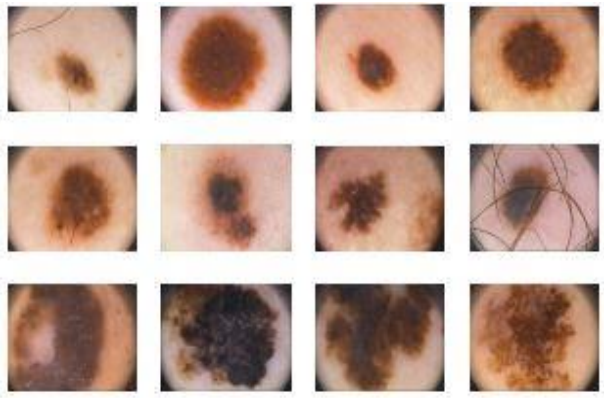


Fig. 4: A collection of images from the PH2 database, including common moles (first row), a typical moles (second row), and melanomas (third row).

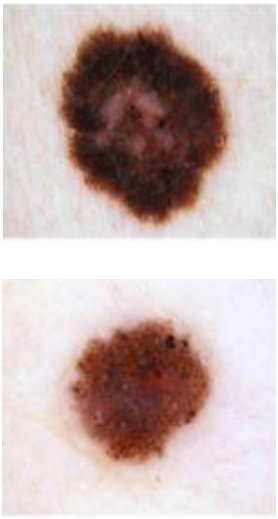


Fig. 5: A collection of images from the ISIC database.

Evaluation Criteria

In this section, the performance of the proposed method is validated by comparing the output information of deep learning with diagnostic data provided by specialized physicians in the community. The performance of the proposed skin lesion segmentation method is evaluated using various metrics including sensitivity, accuracy, dice coefficient, and Jaccard similarity.

TN (True Negatives) indicates the number of records that are actually negative, and the classification algorithm correctly identified them as negative.

TP (True Positives) represents the count of records that are truly positive, and the classification algorithm correctly classified them as positive.

FP (False Positives) represents the count of records that are actually negative, but the classification algorithm incorrectly classified them as positive.

FN (False Negatives) represents the count of records that are actually positive, but the classification algorithm incorrectly classified them as negative.

The ability to differentiate between diseased and healthy cases from other cases is referred to as accuracy. The mathematical expression of this concept is depicted in the following equation [31].

$$Accuracy = \frac{TN + TP}{TN + FN + TP + FP} \tag{14}$$

The accuracy metric does not distinguish between FN and FP. To overcome this issue, the precision metric is defined. The ability of a method to detect disease cases, lesion areas, and cancerous nuclei is called sensitivity. To calculate the sensitivity of a test, one must compute the ratio of true positive cases to the sum of true positive and false negative cases, as shown in the following mathematical expression [31].

$$sensitivity = \frac{TP}{TP + FP} \tag{15}$$

The maximum symmetric surface distance (MSSD), referred to as the Hausdorff distance, can be computed by calculating the maximum distance between the surface voxels of the predicted maps and the ground truth images, with 0 mm stands for perfect segmentation. where Ip and Iy are the predicted maps and ground truth, The shortest distance of a random voxel x to the set of surface voxels of Iy is stated as below Eq. [31].

$$d\left(x, I_y\right) = \min_{y \in I_y} \left\| x - y \right\| \tag{16}$$

It can be written as:

$$MSSD = \max \begin{pmatrix} \left(\max_{x \in I_p} d\left(x, I_d\right) \right) \\ , \left(\max_{y \in I_p} d\left(y, I_p\right) \right) \end{pmatrix} \tag{17}$$

Jaccard similarity is a frequently used metric for measuring the similarity between two objects, such as two images. It can be applied to assess the similarity between two asymmetric binary vectors or to gauge the similarity between two sets.

a. Represents the number of features equal to one for both objects i and j.

b. Indicates the number of features that are zero for object i but one for object j.

c. Denotes the number of features that are one for object i but zero for object j.

d. Refers to the number of features that are zero for both objects i and j.

$$J(i,j)=\frac{a}{a+b+c} \quad (18)$$

$$Jaccard=\frac{TP}{TP+FN+FP} \quad (19)$$

The Dice coefficient is widely employed to quantify the similarity between two images, although it can also be utilized for other data types. Essentially, it assesses the similarity and overlap between the ground truth and the predicted output.

The Dice coefficient serves as a fundamental metric for assessing the outcomes of medical image segmentation. This score quantifies the similarity between the segmentation results produced by a model and the true tissue mask.

$$dice=\frac{2TP}{2TP+FP+FN} \quad (20)$$

Output Results

In this section, the results of the proposed method are juxtaposed with those of other methods, illustrating enhancements in key parameters for image segmentation within the target database. The proposed U-Net architecture facilitates effective segmentation through the utilization of a limited number of training images, along with the integration of information obtained from both the encoder and decoder stages to generate an efficient segmentation map. Tables 1 and 2 present the outcomes of lesion segmentation using various methods, contingent upon the selecting features from the images. It is observed that the proposed method achieves superior segmentation accuracy and precision compared to alternative approaches. This superiority can be attributed to the feature extraction methodology employing deep learning networks and the amalgamation of different feature selections tailored to the structural characteristics of the database images.

Table 1: Results of skin lesion segmentation execution on the PH2 database

Methods	Accuracy	Sensitivity	Dice	Jaccard Similarity
MCGU-Net [20]	95.3	83.2	92.6	95.3
Multiscale Attention U-Net [21]	96.1	94.3	93.7	96.1
FrCN [22]	95.08	93.7	91.7	95.3
U-Net + + [23]	89.6	-	92.7	84.7
Res-Unet [18]	-	-	92.4	85.4
iFCN [24]	96.9	96.8	93.02	87.1
A novel CNN using auxiliary information [25]	94.32	88.76	-	79.46
Proposed method	**96.89**	**96**	**96.40**	**95.40**

Table 2: Results of skin lesion segmentation on the ISIC 2017 database

Methods	Accuracy	Sensitivity	Dice	Jaccard Similarity
deep residual networks [32]	93.40	80.20	84.40	76
fully convolutional-deconvolutional [33]	93.40	82.50	84.90	76.50
Proposed method	**95.98**	**92.85**	**96.32**	**95.24**

Fig. 6 illustrates the eight mother wavelets used in this study. The appropriate selection of the mother wavelet plays a key role in wavelet analysis. For example, in noise removal using wavelets, choosing the appropriate mother wavelet ensures that the most signal power is concentrated on a small number of wavelet coefficients, facilitating the separation of noisy and signal components through thresholding. Mother wavelets are categorized into different segments based on their properties. These properties include orthogonality, compact support, symmetry, and vanishing moments. The properties of mother wavelets are crucial in selecting a suitable mother wavelet. However, there often exist multiple mother wavelets with similar properties. In general, to address the challenge of selecting a mother wavelet, we need to consider the similarity between the signal and the mother

wavelet as a criterion.

The higher the similarity between the mother wavelet and a specific signal, the better signal decomposition into wavelet coefficients. There are various methods for determining the similarity between a signal and a mother wavelet based on qualitative and quantitative approaches. However, there is no single standard method for this selection.

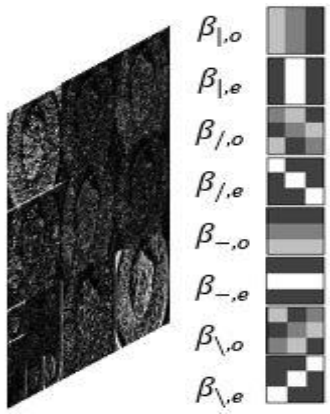


Fig. 6: a. Mother wavelets, b. Filter coefficients response to eight mother wavelets.

In the study [26], two convolutional neural network segmentation models, FBUNet-1 and FBUNet-2, based on the fusion block architecture for segmenting biomedical images, are introduced.

Tables 3 and 4 compare the results of the proposed method with these two methods, FBUNet-1 and FBUNet-2, which have been utilized for segmenting biomedical images of cells and blood vessels. The primary objective of these Tables is to showcase the superiority of the proposed method over the other two methods in handling various datasets comprising different biomedical images.

Table 3 and Table 4 not only demonstrate the superiority of the proposed method in three evaluation parameters, Sensitivity, Dice, and Jaccard Similarity, indicating the high efficiency of the proposed method in medical image segmentation, but also show a reduction in the number of parameters, leading to an increase in the speed of segmentation of the proposed method.

Table 3: Results of the proposed segmentation method compared with FBUNet-1 and FBUNet-2 for cell images

Method	Sensitivity	Dice	Jaccard Similarity	Parameters
FBUNet-1	93.66	93.58	88.41	5,097,925
FBUNet-2	94.19	93.84	88.62	5,098,356
Proposed method	**96**	**96.40**	**95.40**	**1,259,523**

Table 4: Results of the proposed segmentation method compared with FBUNet-1 and FBUNet-2 for blood vessel images

Method	Sensitivity	Dice	Jaccard Similarity	Parameters
FBUNet-1	74.23	79.03	65.16	5,097,452
FBUNet-2	75.64	79.15	65.53	5,098,001
Proposed method	**96**	**96.40**	**95.40**	**1,259,523**

In Table 5, it can be seen that the proposed module improves the basic level of U-Net in terms of MSSD criteria [34].

Table 5: The results of comparing the MSSD criteria

Method	MSSD
U-Net + baseline	50.039
U-Net + CBAM	29.524
U-Net +CoT	22.886
U-Net + ECA	38.402
Proposed method	17.641

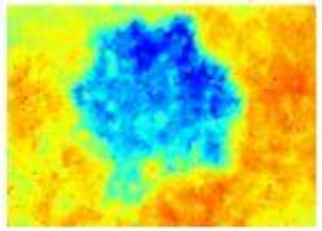
Fig. 7 displays the feature maps obtained from the wavelet transform of the proposed method.

The feature map obtained is a visual representation of the frequency content of the image.

Feature maps in this research are used for pattern recognition, anomaly detection, or feature extraction from the signal.

Fig. 8 displays some of the segmentation results obtained by the proposed method. As observed from the visual results, our network generates smooth segmentation outputs in the border region, which is clinically very useful.

It is evident that our proposed method pays more attention to the border region, and this approach creates a smooth segmentation boundary without additional noisy areas.



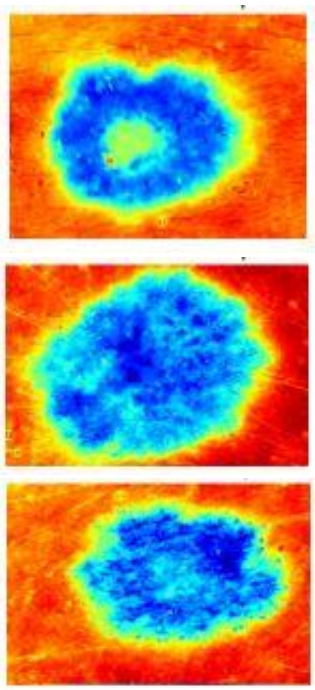


Fig. 7: Feature map obtained from wavelet transform.

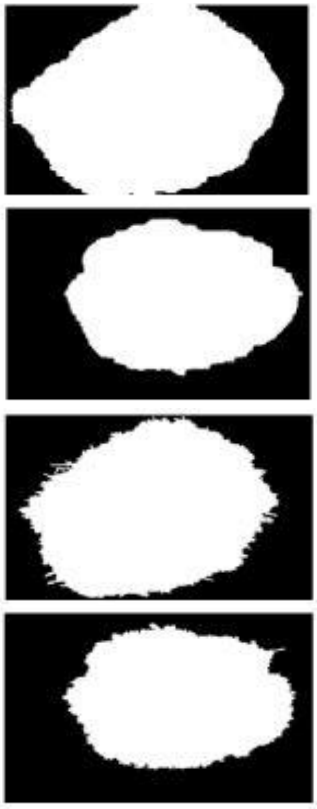


Fig. 8: First row. Ground truth second row. Predicted mask from the combination of wavelet transform and proposed architecture.

Fig. 9 outlines the boundary of the skin lesion by utilizing the output mask. Incorporating wavelet transform enhances the details of the skin lesion, thereby facilitating the development of a more refined algorithm for precise and automated segmentation of skin lesions.

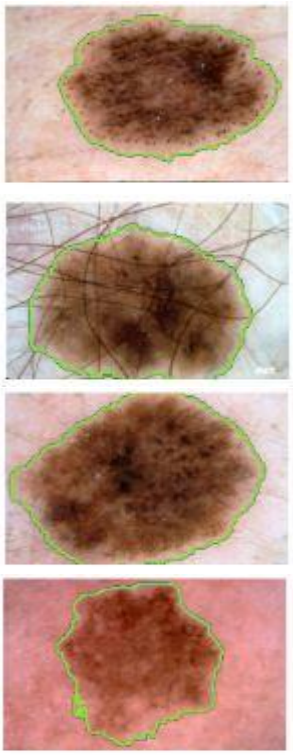


Fig. 9: Detected boundary of the skin lesion area.

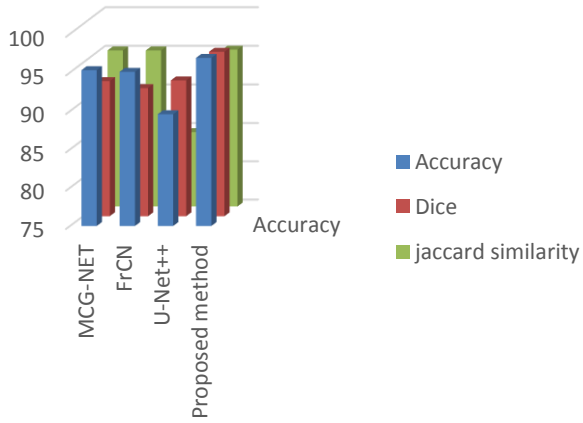


Fig. 10: Comparative plot of different methods for three evaluation metrics.

Fig. 10 illustrates a comparative plot of different methods across three important evaluation metrics in image segmentation. The detection speed in the proposed method is much higher than other methods proposed in previous research, which use low-level learning methods and human detection methods. This can be attributed to the hierarchical learning approach employed in the proposed method, leading to deep learning, feature vector reduction, and a decrease in the number of parameters.

Conclusion

In image processing and analysis, numerous solutions have emerged to aid dermatologists in their diagnostic endeavors.

This paper proposes a method for segmenting skin lesions in dermoscopic images, which tackles challenges like dense hair and gel by integrating wavelet transform and deep learning networks. Manual segmentation, a laborious task heavily reliant on operator expertise, underscores the need for fully automatic methods to delineate skin lesion extents precisely. Despite recent strides in automated algorithms for this purpose, challenges persist due to the diverse characteristics of skin lesions, including size, shape, spatial location, and appearance heterogeneity.

The proposed method significantly increases detection speed by using hierarchical learning and selecting important features, instead of low-level techniques. It improves U-Net performance with supervision blocks between encoding and decoding steps.

Suggestions for enhancing medical image segmentation with deep learning include:

- Improving Accuracy and Efficiency:

- Use multiscale neural networks and attention models.

- Enhancing Robustness:

- Implement data augmentation and transfer learning.

- Integrating Multimodal Information:

- Combine medical images with patient data and multiple models.

- Ensuring Explainability and Transparency:

- Develop methods to understand model decision-making.

- Reducing Computational Resources:

- Create lightweight and optimized models with compression techniques.

- Innovative Applications:

- Develop real-time segmentation models and detect rare diseases.

- Validation and Evaluation:

- Create standard datasets and conduct clinical studies.

These suggestions aim to develop effective and practical models for medical image segmentation, benefiting healthcare.

Author Contribution

S. Fooladi, H. Farsi, S. Mohamadzadeh designed experiments, S. Fooladi, H. Farsi and S. Mohamadzadeh have undertaken the design, execution and analysis of the study results. They also wrote the article and approved its final version.

Acknowledgment

We are grateful to all those who helped us in this research. This article has no financial resources.

Conflict of Interest

The authors state that they have no mutual interest in authoring or publishing the article.

Abbreviations

CAD	Computer-Assisted Diagnosis
BCC	Basal Cell Carcinoma
SCC	Squamous Cell Carcinoma
MM	Melanoma
CNN	Convolutional Neural Network
RFE	Recursive Feature Elimination

References

[1] A. Rehman, M. A. Butt, M. Zaman, "Attention Res-UNet: Attention residual UNet with focal Tversky loss for skin lesion segmentation," Int. J. Decis. Support Syst. Technol., 15(1): 1-17, 2023.

[2] M. K. Hasan, M. A. Ahamad, C. H. Yap, G. Yang, "A survey, review, and future trends of skin lesion segmentation and classification," Comput. Bio. Med., 106624, 2023.

[3] I. Ul Haq, J. Amin, M. Sharif, M. Almas Anjum, "Skin lesion detection using recent machine learning approaches," in Prognostic Models in Healthcare: AI and Statistical Approaches, pp. 193-211, Springer, 2022.

[4] S. Khattar, R. Kaur, "Computer assisted diagnosis of skin cancer: A survey and future recommendations," Comput. Electr. Eng., 104: 108431, 2022.

[5] S. Bakheet, S. Alsubai, A. El-Nagar, A. Alqahtani, "A multi-feature fusion framework for automatic skin cancer diagnostics," Diagnostics, 13(8): 1474, 2023.

[6] S. Fooladi, H. Farsi, S. Mohamadzadeh, "Segmenting the lesion area of brain tumor using convolutional neural networks and fuzzy k-means clustering," Int. J. Eng. Trans. B: Appl., 36(8): 1556-1568, 2023.

[7] N. Ul Huda, R. Amin, S. I. Gillani, M. Hussain, A. Ahmed, H. Aldabbas, "Skin cancer malignancy classification and segmentation using machine learning algorithms," JOM: 1-15, 2023.

[8] Y. Wu, B. Chen, A. Zeng, D. Pan, R. Wang, S. Zhao, "Skin cancer classification with deep learning: a systematic review," Front. Oncol., 12: 893972, 2022.

[9] S. Fooladi, H. Farsi, S. Mohamadzadeh, "Detection and classification of skin cancer using deep learning," J. Birjand Univ. Med. Sci., 26(1): 44-53, 2019.

[10] S. Fooladi, H. Farsi, "Segmentation of cancer cell in histopathologic images of breast cancer and lesion area in skin cancer images using convolutional neural networks," Med. J. Tabriz Univ. Med. Sci. Health Serv., 42(5): 520-528, 2020.

[11] F. Afza, M. Sharif, M. Mittal, M. A. Khan, D. J. Hemanth, "A hierarchical three-step superpixels and deep learning framework for skin lesion classification," Methods, 202: 88-102, 2022.

[12] M. A. Khan, Y. D. Zhang, M. Sharif, T. Akram, "Pixels to classes: intelligent learning framework for multiclass skin lesion localization and classification," Comput. Electr. Eng., 90: 106956, 2021.

[13] M. A. Khan, T. Akram, Y.-D. Zhang, M. Sharif, "Attributes based skin lesion detection and recognition: A mask RCNN and transfer learning-based deep learning framework," Pattern Recognit. Lett., 143: 58-66, 2021.

[14] M. Z. Alom, T. Aspiras, T. M. Taha, V. K. Asari, "Skin cancer segmentation and classification with improved deep convolutional neural network," in Proc. Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications, 11318: 291-301, SPIE, 2020.

[15] G. Nasreen, K. Haneef, M. Tamoor, A. Irshad, "A comparative study of state-of-the-art skin image segmentation techniques with CNN," Multimedia Tools Appl., 82(7): 10921-10942, 2023.

[16] M. Nasir, M. A. Khan, M. Sharif, I. U. Lali, T. Saba, T. Iqbal, "An improved strategy for skin lesion detection and classification using uniform segmentation and feature selection based approach," Microsc. Res. Tech., 81(6): 528-543, 2018.

[17] E. İ. Ünlü, A. Çinar, "Segmentation of benign and malignant lesions on skin images using U-Net," in Proc. 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT): 165-169, 2021.

[18] K. Zafar et al., "Skin lesion segmentation from dermoscopic images using convolutional neural network," Sensors, 20(6): 1601, 2020.

[19] M. Naqvi, S. Q. Gilani, T. Syed, O. Marques, H. C. Kim, "Skin cancer detection using deep learning-A review," Diagnostics, 13(11): 1911, 2023.

[20] M. Asadi-Aghbolaghi, R. Azad, M. Fathy, S. Escalera, "Multi-level context gating of embedded collective knowledge for medical image segmentation," arXiv preprint arXiv:2003.05056, 2020.

[21] M. D. Alahmadi, "Multiscale attention U-Net for skin lesion segmentation," IEEE Access, 10: 59145-59154, 2022.

[22] M. A. Al-Masni, M. A. Al-Antari, M. T. Choi, S. M. Han, T. S. Kim, "Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks," Comput. Methods Programs Biomed., 162: 221-231, 2018.

[23] C. Zhao, R. Shuai, L. Ma, W. Liu, M. Wu, "Segmentation of skin lesions image based on U-Net++," Multimedia Tools Appl., 81(6): 8691-8717, 2022.

[24] Ş. Öztürk, U. Özkaya, "Skin lesion segmentation with improved convolutional neural network," J. Digital Imaging, 33: 958-970, 2020.

[25] L. Liu, Y. Y. Tsui, M. Mandal, "Skin lesion segmentation using deep learning with auxiliary task," J. Imaging, 7(4): 67, 2021.

[26] H. Wang, J. Yang, "FBUNet: Full convolutional network based on fusion block architecture for biomedical image segmentation," J. Med. Biol. Eng., 41: 185-202, 2021.

[27] S. Mohamadzadeh, S. Pasban, J. Zeraatkar-Moghadam, A. K. Shafiei, "Parkinson's disease detection by using feature selection and sparse representation," J. Med. Biol. Eng., 41(4): 412-421, 2021.

[28] F. Yu, V. Koltun, "Multi-scale context aggregation by dilated convolutions," arXiv preprint arXiv: 1511.07122, 2015.

[29] H. K. Gajera, D. R. Nayak, M. A. Zaveri, "A comprehensive analysis of dermoscopy images for melanoma detection via deep CNN features," Biomed. Signal Process. Control, 79: 104186, 2023.

[30] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, M. H. Yap, "Analysis of the ISIC image datasets: Usage, benchmarks and recommendations," Med. Image Anal., 75: 102305, 2022.

[31] K. M. Hosny, D. Elshora, E. R. Mohamed, E. Vrochidou, G. A. Papakostas, "Deep learning and optimization-based methods for skin lesions segmentation: A review," IEEE Access, 11: 85467-85488, 2023.

[32] L. Bi, J. Kim, E. Ahn, D. Feng, "Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks," arXiv preprint arXiv:1703.04197, 2017.

[33] Z. Yuan, "Automatic skin lesion segmentation with fully convolutional-deconvolutional networks," arXiv preprint arXiv:1703.05165, 2017.

[34] J. Zhu, Z. Liu, W. Gao, Y. Fu, "CotepRes-Net: An efficient U-Net based deep learning method of liver segmentation from computed tomography images," Biomed. Signal Process. Control, 88: 105660, 2024.

[35] M. Niazi, K. Rahbar, "Entropy kernel graph cut feature space enhancement with squeezenet deep neural network for textural image segmentation," Int. J. Image Graphics, 24: 2550064, 2024.

Biographies

Saber Fooladi received a Bachelor's degree in Computer Engineering from Birjand University in Birjand center in 2015 and a Master's degree in Electrical Engineering from Birjand University in Birjand in 2017. Currently, he is a doctoral student in the field of Electrical Engineering and telecommunication Systems. Her research interests include image processing, deep learning, convolutional networks, cloud computing and machine learning.

- Email: saber.fooladi@birjand.ac.ir
- ORCID: 0000-0002-8480-2443
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Hassan Farsi received the B.Sc. and M.Sc. degrees from Sharif University of Technology, Tehran, Iran, in 1992 and 1995, respectively. Since 2000, he started his Ph.D. in the Centre of Communications Systems Research (CCSR), University of Surrey, Guildford, UK, and received the Ph.D. degree in 2004. He is interested in speech, image and video processing on wireless communications. Now, he works as Associate Professor in Communication Engineering in department of Electrical and Computer Eng., University of Birjand, Birjand, IRAN.

- Email: hfarsi@birjand.ac.ir
- ORCID: 0000-0001-6038-9757
- Web of Science Researcher ID: NA
- Scopus Author ID: 16202385600
- Homepage: https://cv.birjand.ac.ir/hasanfarsi/en

Sajad Mohamadzadeh received the B.Sc. degree in Communication Engineering from Sistan & Baloochestan, University of Zahedan, Iran, in 2010. He received the M.Sc. and Ph.D. degree in Communication Engineering from South of Khorasan, University of Birjand, Birjand, Iran, in 2012 and 2016, respectively. Now, he works as Associte Professor at department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran. His area research interests include Image and Video Processing, Deep Neural Network, Pattern recognition, Digital Signal Processing, Sparse Representation, and Deep Learning.

- Email: s.mohamadzadeh@birjand.ac.ir
- ORCID: 0000-0002-9096-8626
- Web of Science Researcher ID: NA
- Scopus Author ID: 57056477500
- Homepage: https://cv.birjand.ac.ir/mohamadzadeh/en

How to cite this paper:
S. Fooladi, H. Farsi, S. Mohamadzadeh, "Segmentation of skin lesions in dermoscopic images using a combination of wavelet transform and modified U-Net architecture," J. Electr. Comput. Eng. Innovations, 13(1): 151-168, 2025.

DOI: 10.22061/jecei.2024.10807.736

URL: https://jecei.sru.ac.ir/article_2212.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

New Distance Protection Framework in Sub-Transmission Systems through an Innovative User-defined Approach

A. Yazdaninejadi [*]***, M. Akhavan***

Power Engineering Department, Faculty of Electrical Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.

Article Info

Article History:
Received 19 August 2024
Reviewed 29 September 2024
Revised 13 ctober 2024
Accepted 23 October 2024

Keywords:
Coordination violations
Nonlinear programming model
Combinatorial protection
Selectivity
Distance protection

[*]Corresponding Author's Email Address:
a.yazdaninejadi@sru.ac.ir

Abstract

Background and Objectives: Protection of sub-transmission systems requires maintaining selectivity in the combinatorial scheme of distance and directional overcurrent relay (DOCR). This presents a complex challenge that renders the need for a robust solution. Thereby, the objective of the present study is to decrease the number of violations and minimize the tripping time of relays in this particular issue.

Methods: This study deals with this challenge by using numerical DOCRs which follow non-standard tripping characteristics without compromising the compatibility of the curves. In this process, the time-current characteristics of relays are described in such a manner that they can maintain selectivity among themselves and with distance relays. Therefore, in addition to the second zone timing of distance relays, time dial settings (TDS), and plug settings of overcurrent relays, the other coefficients of the inverse-time characteristics are also optimized. The optimization procedure is formulated as a nonlinear programming model and tackled using the particle swarm optimization (PSO) algorithm.

Results: This approach is verified by applying on two test systems and compared against conventional methods. The obtained results show that the proposed approach helps to yield selective protection scheme owing to the provided flexibility.

Conclusion: The research effectively enhanced selectivity in sub-transmission systems and minimizing relay tripping times through the innovative use of numerical DOCRs and PSO-based optimization.

Introduction

Clearing power grids faults selectivity yields a safe and reliable protection scheme. Due to the low cost and simplicity in implementation of DOCRs, deploying these relays besides distance relays is so prevalent in sub-transmission systems [1]-[4]. In order to achieve selective protection in such systems, it is necessary to fulfill three different coordination scenarios that involve: (1) a distance relay with a distance relay, (2) a DOCR with another DOCR, and (3) a DOCR with a distance relay. All of the above needs to be achieved within minimal relay tripping time. Numerous efforts have been made to address the issue of optimal solutions for relay coordination problems. Linear programming techniques such as simplex, and dual simplex are explored in [5], [6]. Requiring initial guesses based on user intuition is pointed in literatures as the drawback of these techniques [7] which may lead them towards local minima. Relays protection coordination is also formulated to tackle by heuristic optimization machines in [7]-[10]. In [10] group searching optimization is improved to solve the DOCRs coordination problem by enhancing its searching ability. Some other innovative approaches are proposed in [11]-

[15]. Aforementioned references have been concentrated on solving overcurrent relaying, and less work has been dedicated to solve the combinatorial coordination challenge. In [16], [17], as initial studies, the coordination of DOCRs beside distance relays is investigated by linear programing-based techniques. The main facing problems of these studies are first, the same tripping times are considered for second zone of all distance relays and second, there are some records of violations in selectivity constraints. In [7]-[18] the authors enhanced the coordination quality and reduced the total tripping time of relays by adjusting the second zone timing of distance relays and utilizing intelligence-based optimization techniques.

However, violation among selectivity constraints is still present between several pair relays. Another key point is the overcurrent relays are not coordinated in far-end point faults which may lead mal-operations in clearing faults along the protected lines.

Thanks to the advances in intelligent electronic devices (IEDs), numerical relays [19]-[22] are evolved as competent alternatives to overcome the existing hurdles of the conventional relays. These relays are recently commercialized by some manufacturers [23], [24] which are benefitting from a mature design and simple implementation. Users can easily preset these relays, allowing for the inclusion of arbitrary functionalities through software applications as needed. In [25], numerical DOCRs are employed to protect a radial distribution network. Piece-wised linear time-current characteristics are presented for DOCRs. In [26], based on numerical DOCRs, a new scheme is devised for protecting interconnected distribution networks in order to achieve lower tripping times for clearing probable faults. In [27], non-standard tripping characteristics are employed to coordinate dual-setting DOCRs. The need for communication links is the main drawback of deploying dual-setting DOCRs. In [28], different characteristics are provided for DOCRs to eliminate violations in combinatorial scheme of D&DOCRs. However, this method restricts the optimization space due to the limited number of characteristics available. In [29], based on numerical distance relays, non-standard tripping characteristics are considered for distance relays. However, presence of violations is still a challenging task. Although there is an enhancement in the quality of coordination metrics in abovementioned references, the presented approaches can be further extended.

As mentioned earlier, numerical relays allow users to implement arbitrary time-current characteristics through the optimized parameters. Such issues are not dealt in the preceding approach to alleviate violations in coordination problems. Thus, a non-standard coordination process is devised for coordinating DOCRs and distance relays with the aim of minimizing the number of violations and the total tripping time of relays. During the proposed process, overcurrent relaying is performed for far-end points. Therefore, an efficient coordination scheme is adopted, characterized by the following main features:

- A non-standard coordination process is proposed, optimizing the relays' characteristics intuitively within a combinatorial protection scheme;
- The Number of violations can be reduced sensibly;
- The proposed approach reduces the total tripping time of relays considering both the primary and backup relays;
- Discrimination times of relays are also sensibly diminished.

The established non-standard process for coordinating DOCRs and distance relays reveals a nonlinear optimization model. In this study, the model is addressed using PSO.

The ongoing study proceeds as follows: The proposed non-standard coordination process, which incorporates numerical DOCRs alongside distance relays, is explained in Section 2. Subsequently, it is formulated in section 3. The results of the investigated cases are discussed in section 4. The last section concludes the remarks.

The Proposed Scheme of D&DOCRs

A. Combinatorial Protection Scheme of D&DOCRs

As noted, combinatorial scheme of distance relays and DOCRs are commonly used for the protection of sub-transmission networks. The fundamental concept of this protection scheme is illustrated in Fig. 1. This scheme accommodates varying numbers of relays and relay pairs. Referring to Fig. 1, the number of relays is duplicated. In addition, instead of each relays pair, there are four pairs of relays. For the sake of clarity, consider a fault at point *F*. In this case, if the conventional scheme is adopted, R_p^{Dis} typifies primary relay which is backed up with R_b^{Dis}. This is while, in the wake of employing combinatorial scheme, (R_p^{Dis} : R_b^{Dis}), (R_p^{Dis} : R_b^{OC}), (R_p^{OC} : R_b^{Dis}), and (R_p^{OC} : R_b^{OC}) are pairs of relays. R_p^{OC} and R_b^{OC} identify primary and backup DOCRs. For such schemes, the optimal settings of relays are perused with goals of selective, fast and sensitive operation of relays during faults. With this in mind, relays must be coordinated optimally and accurately. In this way, coordination process of proposed approach has two separate parts. Initially, the impedance settings for specifying the three different zones of distance relays are calculated. Subsequently, the time settings for the second zone of distance relays (T_{z2}) as well as the settings for DOCRs, are optimally determined.

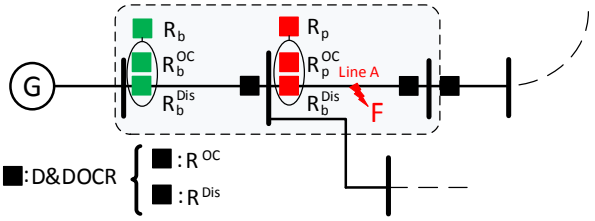


Fig. 1: Simple portion of sub-transmission network.

The impedance settings of distance relays must be calculated before initiating the optimization process. Distance relays feature different protection zones (typically three) to enable selective and sensitive operation. At least three protection zones are considered for distance relays. As shown in Fig. 2, The Z_1 is utilized for protecting the first zone of primary line. It is configured to detect probable faults on 80% of the primary line without introducing any intentional time delay. Generally, Z_2 is the second zone and it is set to cover 120% of the primary line impedance, providing a sufficient margin to accommodate potential errors in relaying. Additionally, the second protection zone acts as a backup for a portion of the adjacent lines with Z_2. The setting of the Z_3 which is the third zone, encompasses the primary protected line and the longest line from the remote bus. Conventionally, coordination between Z_2 and Z_3, and with the Z_1, is achieved through delayed trip outputs (by 15–30 cycles for Z_1 and approximately 20 cycles for Z_3.)

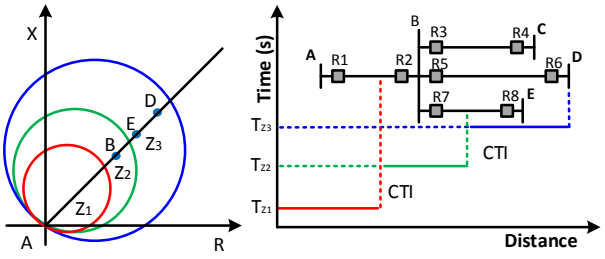


Fig. 2: Illustration of different zones of distance relay.

After specifying the zones of distance relays, T_{Z2} of distance relays must be coordinated with DOCRs. This means that the pairs of relays must satisfy selectivity constraints at five critical points, as depicted in Fig. 3. In essence, at these five critical points, backup relays must initiate operation after the primary relays in a timely fashion, adhering to the critical time interval (*CTI*), meaning they must operate at least *CTI* seconds after the primary relays. This ensures selectivity at other points along the line. *CTI* is defined as the minimum time gap between the tripping times of primary relays and those of their backups. The critical points are identified as follows: F_1 is the first critical point, and it is near the end of the primary line, F_2 is the second critical point and it is at the end of the second zone of the backup distance relay (Z_{2b}), F_3, which is the third critical point, is at the beginning of the second zone of the primary distance relay (Z_{2p}), F_4 is the fourth critical point, and it is at the far-end side of the primary line, and F_5 is the fifth critical point and it is at the beginning of the first zone of the primary distance relay. To prevent curve crossing, in addition to F_1, DOCRs must also be coordinated at F_4, as outlined in Fig. 3(a). Similarly, to fulfill the coordination requirement between R_b^{Dis} and R_p^{OC}, it is necessary to meet the *CTI* at F_2, as shown in Fig. 3(b). Likewise, to satisfy the coordination task between R_p^{OC} and R_b^{Dis}, meeting the *CTI* at F_3 and F_5 is crucial, as demonstrated in Fig. 3(c). Under these conditions, all relays will be coordinated along the protected lines.

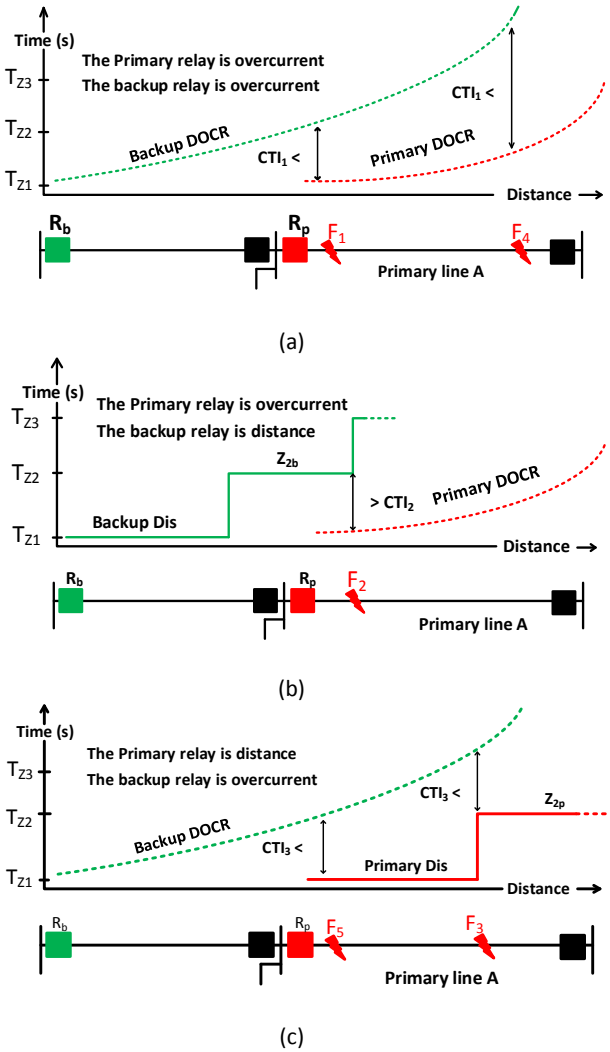


Fig. 3: Five critical points in coordination process

B. Proposed Coordination Process Based on Arbitrary Characteristics

According to IEC/IEEE/AREVA, relay operating times are derived from their characteristic curves as follows:

$$t = TDS\left(\frac{A}{\left(\frac{I_F}{I_p}\right)^B - 1} + C\right) \quad (1)$$

where I_p is the pickup current of relay. In practice, I_p is bigger than the maximum load current passing through the relay by same degree. This will guarantee the stability of the relay under normal loading condition of the network. Moreover, I_F is the magnitude of fault current which is seen by the relay. *A*, *B* and *C* are constant and depend on relays time-current characteristics. Typically, they are chosen from Table 1. For instance, standard-inverse (SI), very-inverse (VI), and extremely-inverse (EI) characteristics are depicted in Fig. 4. In the conventional coordination method, *TDS*, I_p, and I_F are adjustable, while other parameters are considered constant. In contrast, commercial numeric DOCRs offer the ability to set up arbitrary time-current characteristics in tabular form, graphically [25], and by adjusting certain constant coefficients of the relay operation function [16]. For instance, an arbitrary characteristic is displayed alongside standard characteristics in Fig 4.

Table 1: The arrangement of channels

Characteristic No.	Type of characteristic	A factor	B factor	C factor
1	Short Time Inverse	0.05	0.04	0
2	Standard Inverse	0.14	0.02	0
3	Very Inverse	13.5	1	0
4	Extremely Inverse	80	2	0
5	Long Time Inverse	120	1	0
6	Moderately Inverse	0.0515	0.02	0.114
7	Very Inverse	19.61	2	0.491
8	Extremely Inverse	28.2	2	0.1217

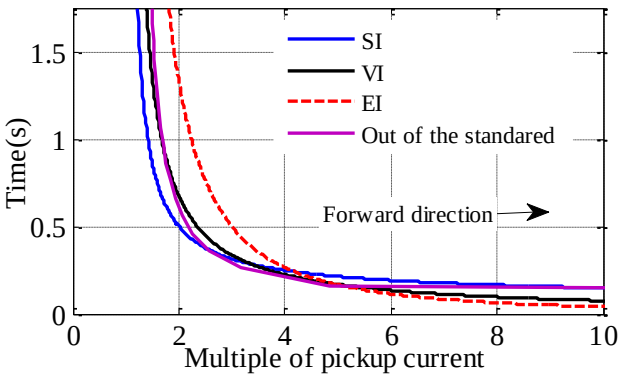


Fig. 4: An example of out of standard time-current curve.

This feature relaxes some constraints in coordination problem and enhances relaying flexibility and extensibility. Hence, in the proposed coordination process, besides *TDS*, I_p, and T_{z2}, the constant coefficients of DOCRs, *A* and *B* are also considered as optimization variables. To put it differently, in conventional protection schemes, only one standard time-current characteristic, typically SI, is used for all DOCRs. This is while, thanks to advances in numerical relays, the proposed coordination process allows for the provision of an optimal characteristic for each DOCR. Therefore, time-current characteristic of overcurrent relays and second zone timing of distance relays are not the same for all relays. Furthermore, each overcurrent relay's characteristic is designed optimally. The explained coordination process is shown in the flowchart of Fig. 5. It is crucial to consider the effects of in-feed and out-feed in the coordination process of the combinatorial scheme.

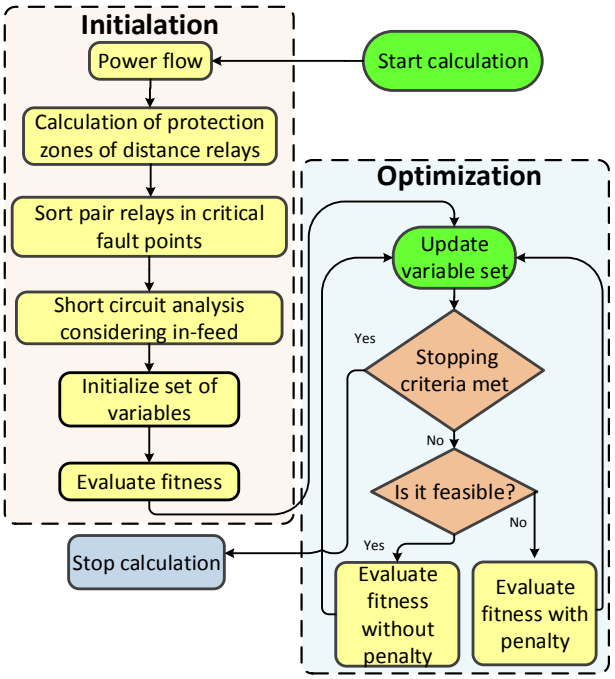


Fig. 5: Flowchart of proposed coordination process.

Problem Formulation

This paper aims to determine the values of *TDS*, I_p, *A*, and *B* for DOCRs and T_{z2} of distance relays optimally. This is done to minimize the number of violations and the total tripping time of the relays. Consequently, the objective function that requires minimization is defined as follows.

$$F^{operating-time} : Minimize \; T = \sum_{f \in F}\left(\sum_{i \in I}\left(t_p^{i,f} + \sum_{s \in S} t_b^{i,f,s}\right)\right) \quad (2)$$

where t_p is tripping time of primary relay and t_b is tripping time of backup relay. Furthermore, *f*, *i*, and *s* are the indices of fault points, all relays, and backup relays, respectively. In addition, *F*, *I*, and *S* represent the sets of fault points, all relays, and backup relays, respectively.

DOCRs are deployed in only one direction for different fault points. Consequently, the time-current characteristic of each DOCR is as follows:

$$t^{o,f} = TDS^{o}\left(\frac{A^{o}}{(IF^{o,f}/Ip^{o})^{B^{o}}-1}+C\right) \quad (3)$$

$$O \subset I \quad (4)$$

where o serves as the index for overcurrent relays, and O represents the set of overcurrent relays. $IF^{0,f}$ denotes the fault current sensed by relay o for a fault occurring at point f.

$$\Delta t^{OC/OC,k,F1} = t_b^{o,k,F1} - t_p^{o,k,F1} - CTI_1 \geq 0 \quad (5)$$

$$\Delta t^{OC/OC,k,F4} = t_b^{o,k,F4} - t_p^{o,k,F4} - CTI_1 \geq 0 \quad (6)$$

These statements are the main running constraints in coordination of DOCRs. Satisfying these constraints guarantees the coordination among over-current relays. In these constraints Δt_k is discrimination time among k-th pair relay. Denote that, primary and backup relays are over-current. The other important constraint in coordination of DOCRs and distance relays is as follows:

$$t_p^{o,k,F1} - t_p^{d,k,F1} > 0 \quad (7)$$

$$D \subset I \quad (8)$$

The constraint in (7) is designed to ensure the priority of distance relays over DOCRs within the same station. The constraints outlined in (7) must be fulfilled in F_5. d serves as the index for distance relays, and D represents the set of distance relays. Moreover, the union and closure of these sets should satisfy the following conditions.

$$O \bigcup D = I \quad (9)$$

$$O \bigcap D = 0 \quad (10)$$

Referring to Fig. 3, during faults at F_2 and F_3 the following constraint must be satisfied:

$$\Delta t^{Dis/OC,k,F2} = t_b^{d,k,F2} - t_p^{o,k,F2} - CTI_2 \geq 0 \quad (11)$$

$$\Delta t^{OC/Dis,k} = t_b^{o,k,F3} - t_p^{d,k,F3} - CTI_3 \geq 0 \quad (12)$$

Dis / OC denotes that, primary relay is distance and the backup relay is overcurrent. As well, *OC / Dis* is vice versa. The other technical constraints regarding relays coordination process are as follows:

$$TDS^{min} \leq TDS^{o} \leq TDS^{max} \quad (13)$$

$$Ip^{min} \leq Ip^{o} \leq Ip^{max} \quad (14)$$

$$Ip^{min,o} = max(I_{load_o}^{max,o}, Iset^{min}) \quad (15)$$

$$Ip^{max,o} = min(I_F^{min,o}, Iset^{max}) \quad (16)$$

$$T_{Z2}^{min} \leq T_{Z2}^{d} \leq T_{Z2}^{max} \quad (17)$$

$Iset^{min}$ and $Iset^{max}$ are the minimum range and maximum range of pickup current provided by manufacturer on relays. The minimum and maximum of I_p is dependent on system's load current and system's short-circuit capacity. Constraints (3)-(16) are the main basis for extracting the required settings of standard coordination process. In non-standard coordination process which yields non-standard time-current characteristics, parameters A and B are also included in optimization process. To do so, these coefficients are defined as the optimization variables and associated constraints are elaborated by (18), (19):

$$B^{min} \leq B^{o} \leq B^{max} \quad (18)$$

$$A^{min} \leq A^{o} \leq A^{max} \quad (19)$$

Eventually, to ensure the security and speed of the proposed protection scheme, the tripping time of each DOCR should also be capped. To do so, the constraint in the following, considers both the lowest and highest permitted times.

$$t^{o,min} \leq t_p^{o,F1} \leq t^{o,max} \quad (20)$$

It is evidently recognized that the proposed coordination scheme for the combinatorial protection scheme represents a non-linear programming problem. In this study, it is solved using the PSO algorithm where details are available in [30]. The particle swarm, established based on the proposed coordination process, is depicted in Fig. 6. It consists five variables, TDS^0, IP^0, A^0, B^0 and T_{z2}^d for both of over-current and distance relay.

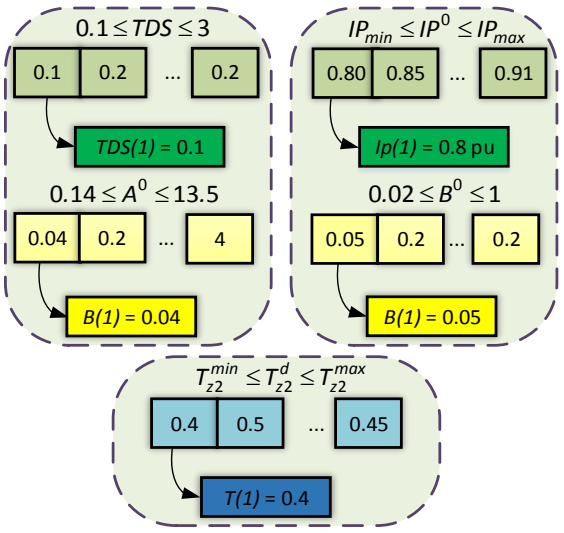


Fig. 6: The proposed particle.

Simulations Results and Discussion

To assess the efficiency and accuracy of the proposed method, it is applied to 8-bus and 9-bus testbeds in two different scenarios. The obtained results based on the proposed method are compared with method presented in [7] in the first scenarios and with [28] in the second scenario. The magnitude of maximum load currents and fault currents have been obtained through DigSILENT Power Factory 14 software. The test networks have been simulated in the mentioned software based on the presented data in [7] and [28]. In short circuit calculation bolted three phase faults are considered in critical points. The optimization process is performed in MATLAB software. It is important to note that, in interconnected power systems, due to significant in-feed (or out-feed) from the connected feeder, distance relays often face mal-operation, which threatens the selectivity of protection systems. For instance, in distance relaying, the location of the critical point F_2 depends on the in-feed (or out-feed) current from the connected feeders to the terminal, that needs to be considered when calculating of Z_2 settings and identifying the critical point F_2. In the current study, the presented method in [18] for setting the second zone of distance relays is employed and hence the effect of in-feed and out-feed on the second zone is considered.

A. First Scenario

The model which is presented in [7] are evaluated based on 8-bus system which is shown in Fig. 7.

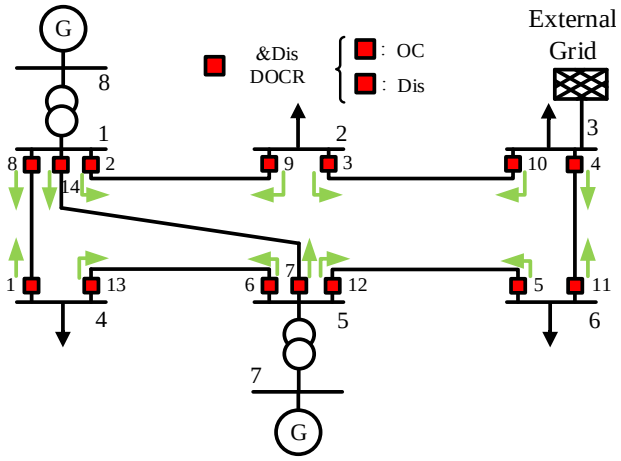


Fig. 7: Single line diagram for 8-bus test system.

The depicted arrows in this figure indicate the direction seen by the relays. *TDS*, I_p and T_{Z2} are the variables within the optimization problem. In this scenario, in order to evaluate performance of the proposed model, the best case of presented method in [7] is modeled and solved optimally. This result is then compared with the proposed model. Therefore, the tripping time of the second zone of distance relays is assumed to be between 0.2 and 0.9. For DOCRs, the value of *TDS* can be selected continuously from 0.1 to 1.1. The maximum and minimum of I_p are calculated in the same manner as presented in [31]. Herein, all the *CTIs* are assumed to be 0.2 sec. Similar to paper [7], the coordination problem is solved for critical points F_1, F_2, and F_3. Table. 2, presents the optimized settings for *TDS*, I_p and T_{Z2} for these critical points. Table. 3 presents the tripping times for faults at those three critical points. The total tripping time of primary overcurrent relays, backup overcurrent relays and second zone of distance relays are 6.4652 sec, 10.2630 sec and 9.5110 sec respectively. As well, in this table, discrimination times for all pairs of relays in all critical points are presented. It is seen that some of the primary/backup pair relays do not satisfy the respective selectivity constraints by maintaining a minimum *CTI* of 0.2 sec. The number of violations is 12, which will not yield reliable protection. The proposed coordination process, as explained in Section 3, is applied to the same test system. In this coordination process, each overcurrent relay set features four settings: *TDS*, I_p, *A*, and *B*, instead of the conventional two settings of *TDS* and I_p. For better comparison, the proposed model is solved in the same critical points considered in [7]. Here, *A* and *B* are variables; they have a minimum value of 0.14 and 0.02, and a maximum value of 1 and 13.5, respectively. Table. 4 shows Optimal results for relays in 8-bus test system based on the proposed model. Table. 5 presents the optimal setting of DOCRs and distance relays. As can be seen, the number of violations decreases with the proposed model to 1, demonstrating a significant improvement.

Table 2: Optimal results for relays in 8-bus test system based on presented model in [7]

Relay No.	Parameters		
	TDS	$I_p(A)$	$T_{Z2}(s)$
1	0.1000	446.69	0.7830
2	0.1690	701.74	0.6800
3	0.1020	831.38	0.5660
4	0.1000	931.56	0.8010
5	0.2130	154.51	0.8710
6	0.1000	923.50	0.6660
7	0.3050	116.22	0.6580
8	0.1140	764.50	0.6290
9	0.1890	166.58	0.6630
10	0.1000	780.07	0.6020
11	0.1000	809.78	0.5390
12	0.1470	847.77	0.6200
13	0.1000	483.04	0.8190
14	0.2590	147.89	0.6140

Table 3: Optimal tripping times of primary and backup relays in 8-bus test system based on presented model in [7]

Relay No./ Pair relays		Tripping times for fault @ F_1			Tripping times for fault @ F_2			Tripping times for fault @ F_3		
Primary	Backup	$t_p^{o/c}$	$t_b^{o/c}$	Δt	$t_p^{o/c}$	t_b^{Dis}	Δt	t_p^{Dis}	$t_b^{o/c}$	Δt
1	6	0.3452	0.5480	0.0028	0.3850	0.666	0.0810	0.783	1.2552	0.2722
2	1	0.5348	0.8605	0.1258	0.5585	0.783	0.0245	0.680	2.6733	1.7933
2	7	0.5348	0.7430	0.0082	0.5585	0.658	-0.1005	0.680	0.9283	0.0483
3	2	0.5408	0.7916	0.0507	0.4812	0.680	-0.0012	0.566	0.7826	0.0166
4	3	0.4879	0.6875	-0.0004	0.5030	0.566	-0.1370	0.801	1.1182	0.1172
5	4	0.528	0.7295	0.0015	0.5439	0.801	0.0571	0.871	1.2098	0.1388
6	5	0.3602	0.7594	0.1992	0.3831	0.871	0.2879	0.666	2.3276	1.4616
6	14	0.3602	0.7107	0.1505	0.3831	0.614	0.0309	0.666	1.0458	0.1798
7	5	0.5398	0.7406	0.0008	0.5650	0.871	0.1060	0.658	0.7719	-0.0861
7	13	0.5398	1.2747	0.5349	0.5650	0.819	0.0540	0.658	-	-
8	7	0.3753	0.7444	0.1691	0.3993	0.658	0.0587	0.629	1.1381	0.3091
8	9	0.3753	0.6792	0.1039	0.3993	0.663	0.0637	0.629	-	-
9	10	0.5129	0.7139	0.0009	0.4907	0.602	-0.0887	0.663	0.9343	0.0713
10	11	0.4246	0.6279	0.0033	0.4354	0.539	-0.0964	0.602	0.9422	0.1402
11	12	0.4388	0.6648	0.026	0.4511	0.620	-0.0311	0.539	0.7541	0.0151
12	13	0.5095	0.9712	0.2618	0.5325	0.819	0.0865	0.620	2.9109	2.0909
12	14	0.5095	0.7107	0.0013	0.5325	0.614	-0.1185	0.620	0.8604	0.0404
13	8	0.3752	0.5734	-0.0018	0.4026	0.629	0.0264	0.819	1.1190	0.1000
14	1	0.4921	1.0673	0.3752	0.5168	0.783	0.0662	0.614	-	-
14	9	0.4921	0.6922	0.0001	0.5168	0.663	-0.0538	0.614	0.7250	-0.0890

Table 4: Optimal results for relays in 8-bus test system based on the proposed model

Relay No.	Parameters				
	TDS	$I_p(A)$	*A*	*B*	T_{Z2} *(s)*
1	0.3311	399.3074	1.5748	0.8090	0.9000
2	0.4166	758.9497	3.2718	0.8591	0.6537
3	0.2618	787.1245	2.6239	0.7495	0.6225
4	0.2544	830.192	3.0266	0.7617	0.6919
5	0.3617	153.1538	5.4828	0.7062	0.8999
6	0.3762	914.6058	2.5562	1.0000	0.8997
7	0.5774	69.4785	5.3969	0.5561	0.4991
8	0.2275	736.9103	3.9123	0.8411	0.3728
9	0.4083	216.5528	3.3756	0.6516	0.7781
10	0.2980	795.0116	2.5574	0.7947	0.6620
11	0.3240	817.8852	2.4128	0.8630	0.5268
12	0.4276	722.4855	4.0185	0.8052	0.6887
13	0.3830	421.7058	1.3544	0.7566	0.8216
14	0.4895	165.9199	6.1853	0.7606	0.6059

The mentioned violations in Table. 3 and Table. 5 are highlighted in gray. Additionally, the total tripping time of primary overcurrent relays, backup overcurrent relays, and the second zone of distance relays in the proposed model are 2.8045 sec, 6.1459 sec, and 7.4454 sec, respectively, showing a significant reduction of approximately 56.6%, 40%, and 21.7%. This demonstrates the efficiency of the proposed scheme. These remarks confirm the satisfactory performance of the proposed approach in reducing violations and shortening relay tripping times, thereby ensuring a reliable and fast protection scheme. Fig. 8, depict comparisons of relay tripping times between the proposed approach and the conventional method at three critical points. Fig. 8d-f, demonstrate comparisons of discrimination times between the proposed approach and the conventional method. As seen in Fig. 8e, the discrimination times in the conventional approach are shorter than those in the proposed approach. However, most of the discrimination times in the conventional approach have negative values.

Table 5: Optimal tripping times of primary and backup relays in 8-bus test system based on the proposed model

Relay No./ Pair relays		Result for fault @ F_1			Result for fault @ F_2			Result for fault @ F_3		
Primary	Backup	$t_p^{o/c}$	$t_b^{o/c}$	Δt	$t_p^{o/c}$	t_b^{Dis}	Δt	t_p^{Dis}	$t_b^{o/c}$	Δt
1	6	0.1167	0.3750	0.0582	0.1431	0.8997	0.5566	0.9000	1.2684	0.1684
2	1	0.2724	0.4727	0.0003	0.2991	0.9000	0.4009	0.6537	1.4792	0.6255
2	7	0.2724	0.5884	0.1160	0.2991	0.4991	0.0000	0.6537	0.8541	0.0005
3	2	0.3889	0.5892	0.0004	0.3245	0.6537	0.1292	0.6225	0.5774	-0.245
4	3	0.3489	0.5491	0.0002	0.3656	0.6225	0.0569	0.6919	1.0160	0.1241
5	4	0.3304	0.6150	0.0847	0.3526	0.6919	0.1393	0.8999	1.1125	0.0126
6	5	0.1659	0.6790	0.3131	0.1894	0.8999	0.5105	0.8997	3.4349	2.3352
6	14	0.1659	0.5961	0.2302	0.1894	0.6059	0.2164	0.8997	1.2886	0.1890
7	5	0.3099	0.6490	0.1391	0.3427	0.8999	0.3573	0.4991	0.6991	0.0000
7	13	0.3099	0.7683	0.2583	0.3427	0.8216	0.2789	0.4991	-	-
8	7	0.1800	0.5904	0.2105	0.2041	0.4991	0.0950	0.3728	1.1511	0.5783
8	9	0.1800	0.7149	0.3349	0.2041	0.7781	0.3740	0.3728	-	-
9	10	0.4127	0.6738	0.0612	0.3760	0.6620	0.0860	0.7781	0.9787	0.0006
10	11	0.2961	0.4988	0.0028	0.3093	0.5268	0.0175	0.6620	0.8950	0.0330
11	12	0.2750	0.5963	0.1213	0.2888	0.6887	0.1999	0.5268	0.7271	0.0004
12	13	0.3734	0.5736	0.0002	0.4057	0.8216	0.2159	0.6887	1.5779	0.6892
12	14	0.3734	0.5961	0.0227	0.4057	0.6059	0.0002	0.6887	0.8895	0.0008
13	8	0.1512	0.3916	0.0404	0.1707	0.3728	0.0021	0.8216	1.0217	0.0000
14	1	0.2387	0.6100	0.1713	0.2732	0.9000	0.4268	0.6059	-	-
14	9	0.2387	0.7405	0.3018	0.2732	0.7781	0.3049	0.6059	0.8060	0.0002

Table 6: Optimal results for relays in 9-bus test system based on the presented model in [28]

Relay No.	Parameters		
	TDS	$I_P(A)$	No. of characteristic
1	0.530	226.149	8
2	0.130	236.750	6
3	0.249	105.625	6
4	1.180	71.160	4
5	0.358	186.703	8
6	0.150	181.320	6
7	0.610	108.486	1
8	1.060	93.2480	4
9	0.270	140.056	3
10	0.350	169.800	1
11	0.300	257.400	1
12	0.184	255.480	8

Table 7: Optimal results for relays in 9-bus test system based on the proposed model

Relay No.	Parameters				
	TDS	$I_P(A)$	*A*	*B*	T_{Z2}
1	2.240	242.970	0.971	1.000	0.508
2	0.061	227.280	0.140	0.020	0.490
3	0.963	101.400	0.512	0.405	0.499
4	1.724	77.090	3.866	1.000	0.505
5	1.287	200.590	1.203	1.000	0.531
6	0.826	181.320	1.014	0.812	0.594
7	1.458	103.320	0.731	0.746	0.486
8	2.483	97.760	2.038	1.000	0.522
9	1.206	149.240	2.491	1.000	0.484
10	0.776	169.800	0.741	0.651	0.508
11	0.587	257.400	0.919	0.843	0.553
12	2.253	276.770	0.507	1.000	0.538

Table 8: Comparison of discrimination times of relays

Pair relays		Δt @									
PR	BR	F_1		F_2		F_3		F_4		F_5	
		[28]	New	[28]	New	[28]	New	[28]	New	[28]	New
1	11	0	0	0.666	1.593	0.126	0.256	0.331	0.679	-0.3635	0
2	4	0.055	0.047	0.318	0.195	0.076	0.319	0	0	-0.3636	0
3	1	0	0	0.676	0.306	0.190	0.261	0.440	0.066	-0.3427	0
4	6	0	0	0.400	1.242	0.071	0.279	0.129	0.616	-0.3694	0
5	3	0.003	0	0.068	0.469	0.041	0.160	0.011	0.109	-0.3547	0
6	8	0	0	0.358	0.251	0.159	0.213	0.226	0	-0.3040	0
7	5	0	0	0.377	0.200	0.188	0.279	0.251	0.024	-0.4151	0
8	10	0	0	0.472	0.920	0.173	0.209	0.297	0.372	-0.3160	0
9	7	0	0	0.397	0.493	0.099	0.199	0.273	0.176	-0.3258	0
10	12	0	0	0.278	0.228	0.295	0.284	0.274	0.032	-0.3210	0
11	9	0	0	1.156	0.241	0.172	0.180	0.777	0	-0.2851	0
12	2	-0.31	0	-	-	-0.17	0.146	0.590	4.162	-0.3259	0
Average (for positive values)		0.01	0.004	0.437	0.511	0.346	0.232	0.490	0.520	-	0

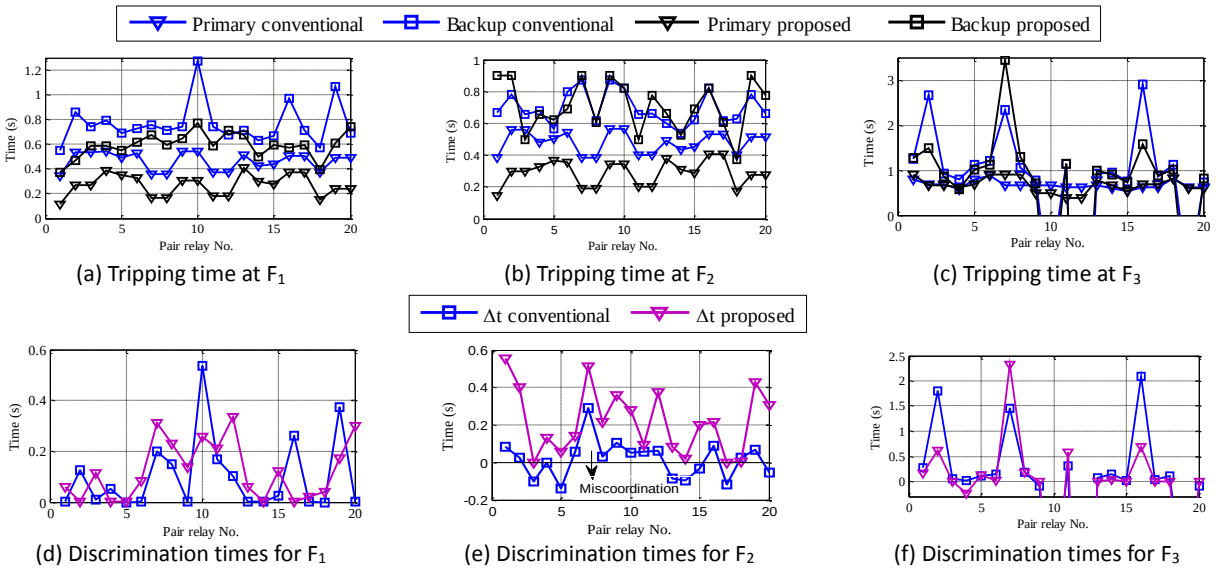


Fig. 8: Comparison of the proposed approach versus the conventional method in three critical point.

B. Second Scenario

Here, the proposed method is compared with those presented in [28]. In [28], characteristics of relays can be chosen from Table. 1.

Furthermore, the optimization process takes into account all five critical points. The 9-bus test system is considered as testbed which is depicted in Fig. 9, where detailed in [28]. Like Fig. 7, The arrows in this figure indicate the direction seen by the relays.

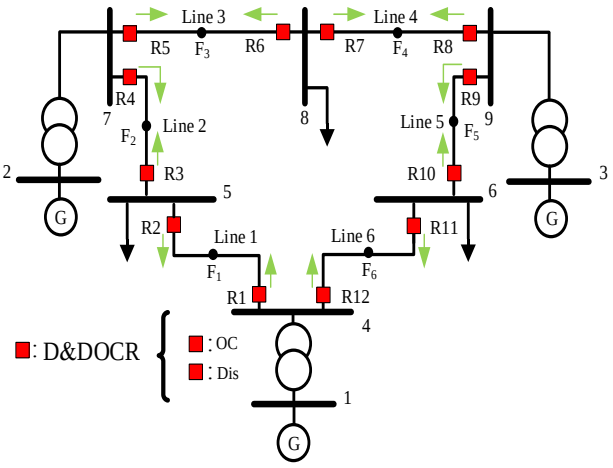


Fig. 9: Single line diagram for 9-bus test system.

TDS and I_p can range from 0.1 to 3.2 and $1.2\times I_p$ to $1.3\times I_p$ respectively. The value of T_{Z2} is set at 0.3 sec.

Optimized variable values are listed in Table. 6 and Table. 7, where Table. 6 is corresponding for conventional scheme and Table. 7 shows optimal results for relays in the 9-bus test system based on the proposed model. All the *CTIs* are assumed 0.2 sec.

Discrimination time of pair relays in five critical fault points is given in Table. 8.

As observed, 14 pairs of relays are not satisfied selectivity constraint. The proposed model is applied to the 9-bus testbed using the same objective function in [28].

Thus, besides *TDS* and I_p, the other variables *A*, *B*, and T_{Z2}, are also included in the optimization problem. The obtained solution is also given in Table. 7. As it is seen All coordination constraints are satisfied and there is no miscoordinations which yields reliable protection. The overall average discrimination time of relays, which were previously positive according to the model presented in [28], has been reduced from 1.283 sec to 1.267 sec. This study primarily focuses on the technical aspects of the combinatorial protection scheme in sub-transmission systems and does not extensively address the cost implications or economic feasibility of implementing the proposed solutions with employment of numerical relays. Future research should include a comprehensive economic analysis to evaluate the financial viability of these protection measures.

Conclusion

This study concerns relaying problem in protection scheme of over-current and distance relays in the sub-transmission system. In this paper, based on the provided flexibility by numerical relays, DOCRs was encouraged to follow user defined characteristics. Specifically, the constant coefficients of the DOCRs characteristic are considered as optimization variables, leading to a non-standard coordination process with greater flexibility. In this process relays are coordinated in five critical points to assure selectivity along the protected lines. The proposed approach is compared with conventional approaches in two different case studies. In both case studies, the discrimination time and the total tripping time of the relays have been reduced. Furthermore, in the first case study, the number of violations decreased from 12 to 1. Also, in the second case study, the number of violations has been reduced from 14 to zero. The obtained results highlight that:

- Number of pair relays satisfying selectivity constraints were increased and number of violations were decreased resulting in reliable protection scheme;
- A significant reduction in the total tripping time of relays, contributing to a faster protection system;
- Discrimination time of relays were also diminished.

Author Contributions

A. Yazdaninejadi has contributions on modeling, simulations, performing the analysis and writing the paper. M. Akhavan has contributions on modeling, simulations, performing the analysis and writing the paper.

Acknowledgment

There are not any acknowledgements to report.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

DOCR	Directional Overcurrent Relay
PSO	Particle Swarm Optimization
IED	Intelligent Electronic Device
CTI	Critical Time Interval

T_{z2}	Time Setting for The Second Zone of Distance Relay
Z_{2b}	Second Zone of The Backup Distance Relay
Z_{2p}	Second Zone of the Primary Distance Relay
TDS	Time Dial Setting
SI	Standard-Inverse
VI	Very-Inverse
EI	Extremely-Inverse

References

[1] H. Sahraei, M. Tolou Askari, "Influence of phase-shifting transformers on the distance protection of transmission lines and improve the performance of distance relay, " J. Electr. Comput. Eng. Innovations, 9(2): 173-184, 2021.

[2] T. S. Sidhu, R. K. Varma, P. K. Gangadharan, F. A. Albasri, G. R. Ortiz, "Performance of distance relays on shunt-FACTS compensated transmission lines," IEEE Trans. Power Deliv., 20(3): 1837-1845, 2005.

[3] S. M. Brahma, "Distance relay with out-of-step blocking function using wavelet transform," IEEE Trans. Power Deliv., 22(3): 1360-1366, 2007.

[4] D. Kang, R. Gokaraju, "A new method for blocking third-zone distance relays during stable power swings," IEEE Trans. Power Deliv., 31(4): 1836-1843, 2016.

[5] P. P. Bedekar, S. R. Bhide, V. S. Kale, "Optimum coordination of overcurrent relay timing using simplex method," Electr. Power Compon. Syst., 38(10): 1175-1193, 2010.

[6] P. P. Bedekar, S. R. Bhide, V. S. Kale, "Optimum coordination of overcurrent relays in distribution system using dual simplex method, " in Proc. 2009 Second International Conference on Emerging Trends in Engineering & Technology: 555-559, 2009.

[7] M. Farzinfar, M. Jazaeri, F. Razavi, "A new approach for optimal coordination of distance and directional over-current relays using multiple embedded crossover PSO," Int. J. Electr. Power Energy Syst., 61: 620-628, 2014.

[8] S. A. Ahmadi, H. Karami, M. J. Sanjari, H. Tarimoradi, G. B. Gharehpetian, "Application of hyper-spherical search algorithm for optimal coordination of overcurrent relays considering different relay characteristics," Int. J Electr. Power Energy Syst., 83: 443-449, 2016.

[9] F. A. Albasri, A. R. Alroomi, J. H. Talaq, "Optimal coordination of directional overcurrent relays using biogeography-based optimization algorithms," IEEE Trans. Power Deliv., 30(4): 1810-1820, 2015.

[10] M. Alipour, S. Teimourzadeh, H. Seyedi, "Improved group search optimization algorithm for coordination of directional overcurrent relays," Swarm Evol. Comput., 23: 40-49, 2015.

[11] D. K. Ibrahim, E. E. Din Abo El Zahab, S. Abd El Aziz Mostafa, "New coordination approach to minimize the number of re-adjusted relays when adding DGs in interconnected power systems," J. Electr. Eng. Technol., 12(2): 502-512, 2017.

[12] A. Papaspiliotopoulos, G. N. Korres, N. G. Maratos, "A novel quadratically constrained quadratic programming method for optimal coordination of directional overcurrent relays," IEEE Trans. Power Deliv., 32(1): 3-10, 2015.

[13] C. W. So, K. K. Li, "Overcurrent relay coordination by evolutionary programming," Electr. Power Syst. Res., 53(2): 83-90, 2000.

[14] M. Y. Shih, C. A. Castillo Salazar, A. C. Enríquez, "Adaptive directional overcurrent relay coordination using ant colony optimisation," IET Gener. Transm. Distrib., 9(14): 2040-2049, 2015.

[15] D. S. Alkaran, M. R. Vatani, M. J. Sanjari, G. B. Gharehpetian, M. S. Naderi, "Optimal overcurrent relay coordination in interconnected networks by using fuzzy-based GA method," IEEE Trans. Smart Grid., 9(4): 3091-3101, 2016.

[16] L. G. Perez, A. J. Urdaneta, "Optimal coordination of directional overcurrent relays considering definite time backup relaying," IEEE Trans. Power Deliv., 14(4): 1276-1284, 1999.

[17] L. G. Perez, A. J. Urdaneta, "Optimal computation of distance relays second zone timing in a mixed protection scheme with directional overcurrent relays," IEEE Trans. Power Deliv., 16(3): 385-388, 2001.

[18] Z. Moravej, M. Jazaeri, M. Gholamzadeh, "Optimal coordination of distance and over-current relays in series compensated systems based on MAPSO," Energy Conv. Manag., 56: 140-151, 2012.

[19] H. A. Darwish, M. Fikri, "Practical considerations for recursive DFT implementation in numerical relays," in Proc. 2005/2006 IEEE/PES Transmission and Distribution Conference and Exhibition: 880-887, 2006.

[20] S. Dadfar, M. Gandomkar, "Augmenting protection coordination index in interconnected distribution electrical grids: Optimal dual characteristic using numerical relays," Int. J. Electr. Power Energy Syst., 131:107107, 2021.

[21] M. A. Abdel-Salam, R. Kamel, K. Sayed, M. Khalaf, "Design and implementation of a multifunction DSP-based-numerical relay," Electr. Power Syst. Res., 143: 32-43, 2017.

[22] Y. L. Goh, A. K. Ramasamy, F. H. Nagi, A. A. Zainul Abidin, "Evaluation of DSP based numerical relay for overcurrent protection," Int. J. Syst. Appl. Eng. Dev., 5: 396-403, 2011.

[23] SIPROTEC Multi-Functional Protective Relay 7SJ62/63/64 [Online].

[24] Toshiba directional overcurrent relay GRD 140 [Online].

[25] M. Ojaghi, R. Ghahremani, "Piece-wise linear characteristic for coordinating numerical overcurrent relays," IEEE Trans. Power Deliv., 32(1): 145-151, 2016.

[26] C. A. Castillo Salazar, A. Conde Enríquez, S. E. Schaeffer, "Directional overcurrent relay coordination considering non-standardized time curves," Electric. Power Syst. Res., 122: 42-49, 2015.

[27] A. Yazdaninejadi, D. Nazarpour, S. Golshannavaz, "Dual-setting directional over-current relays: An optimal coordination in multiple source meshed distribution networks," Int. J. Electr. Power Energy Syst., 86: 163-176, 2017.

[28] S. A. Ahmadi, H. Karami, B. Gharehpetian, "Comprehensive coordination of combined directional overcurrent and distance relays considering miscoordination reduction," Int. J. Electr. Power Energy Syst., 92: 42-52, 2017.

[29] Y. Damchi, J. Sadeh, H. Rajabi Mashhadi, "Optimal coordination of distance and overcurrent relays considering a non-standard tripping characteristic for distance relays," IET Gener. Transm. Distrib., 10(6): 1448-1457, 2016.

[30] M. Hasanluo, F. Soleimanian Gharehchopogh, "Software cost estimation by a new hybrid model of particle swarm optimization and k-nearest neighbor algorithms," J. Electr. Comput. Eng. Innovations, 4(1): 49-55, 2016.

[31] P. P. Bedekar, S. R. Bhide, "Optimum coordination of directional overcurrent relays using the hybrid GA-NLP approach," IEEE Trans. Power Deliv., 26(1): 109-119, 2010.

Biographies

Amin Yazdaninejadi received the B.Sc. degree in Electrical Engineering from Urmia University, Urmia, Iran, in 2012, the M.Sc. degrees in Electrical Engineering from Amirkabir University, of Technology (AUT), Tehran, Iran, in 2014, and the Ph.D. degree in Electrical Power Engineering from Urmia University, Urmia, Iran, in 2018. He is currently an Assistant Professor with the School of Electrical and Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran. His research interests include power system protection, power system automation, smart grid technologies, machine learning and its applications, design of distribution management system (DMS), demand side management (DSM) concepts and applications, microgrid design and operation studies, design of energy management system (EMS).

- Email: a.yazdaninejadi@sru.ac.ir
- ORCID: 0000-0003-4506-2988
- Web of Science Researcher ID: NA
- Scopus Author ID: 57192178404
- Homepage: https://www.sru.ac.ir/en/faculty/school-of-electrical-engineering/amin-yazdaninejadi/

Moein Akhavan received his B.Sc. degree in Electrical Engineering from Semnan University, Semnan, Iran, in 2022. He is currently a Master's student in Department of Electrical Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran. His research interests include the modeling and simulation of power systems, power system protection, machine learning applications, smart grid and renewable energy integration.

- Email: moein.akhavan@sru.ac.ir
- ORCID: 0009-0007-5199-1851
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
A. Yazdaninejadi, M. Akhavan, "New distance protection framework in sub-transmission systems through an innovative user-defined approach," J. Electr. Comput. Eng. Innovations, 13(1): 169-180, 2025.

DOI: 10.22061/jecei.2024.11189.773

URL: https://jecei.sru.ac.ir/article_2213.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

A Machine-Learning-based Predictive Smart Healthcare System

F. Ahmed Shaban, S. Golshannavaz *

Faculty of Electrical and Computer Engineering, Urmia University, Urmia, Iran.

Article Info

Article History:
Received 22 July 2024
Reviewed 15 October 2024
Revised 16 October 2024
Accepted 31 October 2024

Keywords:
Healthcare system
Patients health monitoring
Machine-learning
Artificial intelligence
Learning algorithms

*Corresponding Author's Email Address:
s.golshannavaz@urmia.ac.ir

Abstract

Background and Objectives: In smart grid paradigm, there exist many versatile applications to be fostered such as smart home, smart buildings, smart hospitals, and so on. Smart hospitals, wherein patients are the possible consumers, are one of the recent interests within this paradigm. The Internet of Things (IoT) technology has provided a unique platform for healthcare system realization through which the patients' health-based data is provided and analyzed to launch a continuous patient monitoring and; hence, greatly improving healthcare systems.

Methods: Predictive machine learning techniques are fostered to classify health conditions of individuals. The patients' data is provided from IoT devices and electrocardiogram (ECG) data. Then, efficient data pre-processings are conducted, including data cleaning, feature engineering, ECG signal processing, and class balancing. Artificial intelligence (AI) is deployed to provide a system to learn and automate processes. Five machine learning algorithms, including Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), logistic regression, Naive Bayes, and random forest, as the AI engines, are considered to classify health status based on biometric and ECG data. Then, the output would be the most proper signals propagated to doctors' and nurses' receivers in regard of the patients providing them by initial pre-judgments for final decisions.

Results: Through the conducted analysis, it is shown that logistic regression outperforms the other AI machine learning algorithms with an F1 score, recall, precision, and accuracy of 0.91, followed by XGBoost with 0.88 across all metrics. SVM and Naive Bayes both achieved 0.85 accuracy, while random forest attained 0.86. Moreover, the Receiver Operating Characteristic Area Under Curve (ROC-AUC) scores confirm the robustness of Logistic Regression and XGBoost as apt candidates in learning the developed healthcare system.

Conclusion: The conducted study concludes a promising potential of AI-based machine learning algorithms in devising predictive healthcare systems capable of initial diagnosis and preliminary decision makings to be relied upon by the clinician. What is more, the availability of biometric data and the features of the proposed system significantly contributed to primary care assessments.

Introduction

A. Motivation

Internet of Things (IoT) technology has given momentum to adoption of healthcare applications in modern societies. The warm reception of IoT in these systems is due to several factors, including limited access to medical resources based on conventional approaches, growing number of elderly people with chronic diseases and their need for remote monitoring, rising medical costs, and the desire for telemedicine in developing countries. The IoT has made continuous patient monitoring and real-time data collection possible, which has greatly improved the

healthcare systems. In this way, a health monitoring system is generally recognized with multiple sensors and a core processor [1]. This system would monitor vital signs in real-time, collect data from various sensors related to patients, and conduct artificial intelligence (AI) analysis on the data to predict and analyze its results and determine whether the results are real and correct. People can interact with a vast array of digital and physical objects, including those used in personal healthcare, using the IoT [2]. In a recent study, patients' electrocardiogram (ECG), blood pressure, temperature, and pulse rate biometric data are analyzed using AI algorithms to ascertain their current state of health [3]. The relationship between AI and the IoT revolves around connecting things and automating processes; then, analyzing data, getting understanding, and making decisions based on data [[4], [5]]. As a complementary need, for predicting the results and reaching to decision making processes, machine learning algorithms should be trained, accurately. Such a system would provide an efficient platform of care-systems to be deployed by doctors and nurses to have a continuous and reliable monitoring on the patients' situations.

Contributing to this field, the present study is on developing an efficient healthcare system, capable of leaning and decision making on patients' health-based data gathered by biometric sensors. The patients' data is provided from IoT devices and ECG data. Then, efficient data pre-processings are conducted, including data cleaning, feature engineering, ECG signal processing, and class balancing. Then, AI is deployed to provide a system to learn and automate processes, ending to a predictive system. Five machine learning algorithms, including support vector machine (SVM), Extreme Gradient Boosting (XGBoost), logistic regression, Naive Bayes, and random forest, as the AI engines, are considered to classify health status. Then, the output would be the most proper signals propagated to doctors' and nurses' receivers in regard of the patients providing them by initial pre-judgments. The possible contributions of this study could be listed as follow:

- A cloud-based data center is developed for IoT-based biometric and ECG data;
- Data cleaning, feature engineering, and class balancing is embedded as the preliminary stage of the proposed approach;
- Predictive feature is realized by AI-based machine learning algorithms and training process;
- An automated and predictive healthcare system is developed which provides initial judgments and primary care assessments of patients.

Different machine-learning algorithms are explored to provide an overview of the proposed healthcare system. These algorithms investigate different performance metrics say as accuracy, precision, and etc. The obtained numerical results emphasize on a promising potential of machine learning paradigm in IoT-based data handling and health diagnosis which can be relied upon by clinician in significant enhancing of primary care assessments.

B. Literature Review

In [6], researchers have explored the deployment of IoT wearable electroencephalography (EEG) [7] devices and SVM for predictive analytics in epilepsy treatment. The devices used real-time brain activity data to predict seizures, allowing healthcare practitioners and caregivers to interact, efficiently. Seamless connectivity of the IoT infrastructure enables timely warnings and efficient remediation. Srinivas et.al. have shown that experimental investigations provide promising predictive accuracy and reaction time, providing individualized and proactive treatments for epilepsy patients [8].

In references [9], [10], a new approach is proposed to develop healthcare monitoring system. The researchers were interested in the methodology of combining regular medical monitoring and electronic clinical data (ECD) from complete medical records with physical data of patients as well as machine learning techniques in order to predict heart disease. The XGBoost algorithm is used as a powerful algorithm for examining large data sets effectively and extracting important features to improve prediction accuracy. The results are optimized which demonstrate that the XGBoost algorithm outperforms Naive Bayes, decision trees, and random forests and achieved a greater prediction accuracy of 99.4%. By combining IoT technologies with advanced machine learning models it would provide better results.

Authors in [11] developed some methods and algorithms to predict the health status of Coronavirus patients and classify them according to their healthy and unhealthy conditions. In this research, a comprehensive analysis of machine learning approaches was conducted in the field of diagnosing COVID-19, detecting chronic diseases in patients, and identifying symptoms of COVID-19 infection. As well, decision trees, random forest, SVM, gradient boosting, and logistic regression algorithms are used in [12]. The best results were obtained from the comparative analysis of the methods including decision tree, random forest, and gradient boosting algorithms, on the accuracy values of 1.0, 0.99, and 1.0, respectively. These results show their performance and functionality in machine learning algorithms about their goal in the field of healthcare, as well as the possibility of choosing the most appropriate one to deal with diseases [13].

Another study was concerned with collecting types of patient data that would help the doctor correctly diagnose the patient's health condition [14]. The data is analyzed by the doctor, who then confirms the disease using his medical experience and makes a diagnosis. In this study, researchers used machine learning techniques

such as Naive Bayes and random forest classification algorithms to classify several disease datasets such as heart disease, cancer, and diabetes to check whether the patient is affected by this disease or not. In [15], the results of the simulations show the effectiveness of classification techniques on the data set, as well as the nature and complexity of the deployed data set. The performance analysis for both algorithms was done and compared. Based on the results, it can be said that these algorithms are among the promising techniques in the field of disease analysis and prediction.

C. Paper Organization

This study continues as follows: In "Methodology" section, the developed healthcare methodology is outlined. The "Experimental Studies and Evaluations" section is provided for further analysis and performance validations of the developed model. Eventually, the "Conclusions and Future Works" concludes the study and provides some open future works.

Methodology

As outlined, a processor, here taken as Raspberry Pi, is used to create an application that connects the electronic system to medical sensors placed on the patient's body. These sensors measure the patient's blood pressure, heart rate, temperature, as well as ECG. Also, the inter-controller communication protocol is used to collect and transmit data from the sensors to the physician monitoring system [16]. The system software application contains the code used by the Raspberry PI controller. The medical sensors are programmed using the Python language [17]. In addition, the Java language is used to design the monitoring program that is placed near the doctor using a mobile application. The patient database is connected to sensors and takes the results from the disease state and stores them in the cloud and then transfers to the doctor's application to display the results. The database is implemented using a cloud-based storage solution. Here two possible ways are described to implement the database. The first approach is the "Cloud-based NoSQL Database". This system could use a cloud-based NoSQL database such as MongoDB [18], Cassandra, or Couchbase to store patient data. NoSQL databases are well-suited for handling large amounts of unstructured or semi-structured data, which is common in healthcare applications. The second approach is the "Relational Database". This system could use a relational database such as MySQL or PostgreSQL to store patient data. Relational databases are suited for handling structured data. The data is then collected in an Excel file for all patients with their ages and genders to begin the analysis through AI techniques and train the aforementioned algorithms for prediction. Then, the obtained results are made available to be analyzed and determine the validity of the diagnosis, as showed in Fig. 1. This figure shows how patient data is measured, transmitted, stored, and analyzed as well as displayed using applications, medical sensors, cloud storage, and AI as an overall analysis system. This figure highlights the steps and flow of data between sensors, Raspberry Pi, cloud, and physician's mobile application.

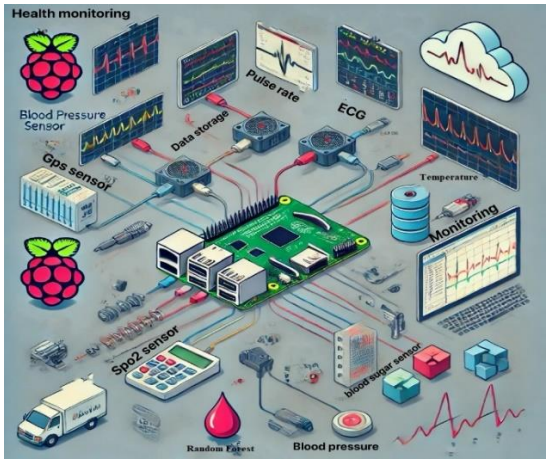


Fig. 1: System software and hardware implementation.

A. SVM Algorithm

SVM is one of the machine learning algorithms used in neuroimaging analysis, through which classification problems are addressed and provides balanced predictive performance for diagnosis in the field of brain diseases, psychiatry, and others [19]. The SVM equation and the health state are determined by evaluating the linear combination of features, as follows [20]:

$$\begin{aligned} f(x) = sign\big(&w_1 Age + w_2 Gender + w_3 Temperature \\ &+ w_4 Pulse + w_5 SpO2 + w_6 SystolicBP \\ &+ w_7 DiastolicBP + w_8 ECG1 + \ldots + w_9 \\ &+ m.ECEm + b\big) \end{aligned} \tag{1}$$

B. XGBoost Algorithm

XGBoost is a scalable, distributed, gradient-boosting decision tree (GBDT) machine-learning library [21]. It provides parallel tree boosting and is the leading machine learning library for solving regression and classification problems and is used in the healthcare field for general disease prediction as well as diagnosis and analysis [22]. The equation of XGBoost is as follows:

$$y_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i) \tag{2}$$

C. Logistic Regression Algorithm

Regression analysis is an important statistical method used to determine the relationship between several factors and disease outcomes or to identify disease-related prognostic factors through probability and prediction by estimating the occurrence of an event [23].

The equation of logistic regression is written as follows:

$$y = \frac{1}{1 + e^{-(w.x+b)}} \tag{3}$$

D. Naïve Bayes Algorithm

Naive Bayes algorithm is used to distinguish between favorable patient reviews and those that are negative. Here, it is easily understand which drugs are most beneficial and have the fewest negative effects [24]. The equation for the Naive Bayes classifier is as follows [25]:

$$P(C_k|x) = \frac{P(C_k)\prod_{i=1}^{n} P(x_i|C_k)}{P(x)} \tag{4}$$

E. Random Forest Algorithm

Leo Breiman and Adele Cutler are the trademark holders of the popular machine learning technique known as "Random Forest," which aggregates the output of several decision trees to produce a single conclusion. Its popularity has been spurred by its flexibility and ease of use, since it can handle regression and classification problems as well as healthcare [26]. The random forest algorithm combines the predictions from multiple decision trees to make a final prediction as follows:

$$y = \text{mode}\left(\{T_1(x), T_2(x), \ldots, T_B(x)\}\right) \tag{5}$$

F. Data Set of the Study

Data are collected containing biometric readings and ECG data for patients in a local hospital in Iraq, in addition to random individuals. The goal of collecting these samples and vital indicators is to evaluate them. Blood, along with a list of ECG values [27], are recorded and classified according to age, sex, and blood pressure compatibility. The number of vital signs records collected in this study is 150 people. Among these people, 80 patients are admitted to the local hospital, while 70 patients are randomly selected.

Experimental Studies and Evaluations

The conducted study proposes a hardware and Google cloud-IoT-based healthcare system being trained based on machine learning algorithms developed in Raspberry Pi in real-world implementations. Due to differences in technical specifications of the implemented hardware and the assumptions such as the volume of biometric and ECG data, system performance and decision making criteria [28], comparison of the proposed system with the existing systems would not make a right comparison platform and performance analysis. Instead, the performance of the proposed system is analyzed in-depth considering the well-known and mostly applied machine learning AI algorithms and meaningful comparisons are hence attained and discussed. By integrating Raspberry Pi with IoT sensors, a powerful health monitoring system is created that includes, blood pressure, pulse rate, ECG, temperature, and other critical indications. This integration offers a thorough and instantaneous method for monitoring and controlling multiple health metrics [29]. It would be a single, simple phone application that allows doctors to directly monitor the progress of surgeries and their effects on patients. It is also a safe, dependable, and efficient cloud storage system for transferring medical data to the doctor's application in the observation room. Furthermore, all devices and sensors are either directly connected to the patient or via a data transmission medium.

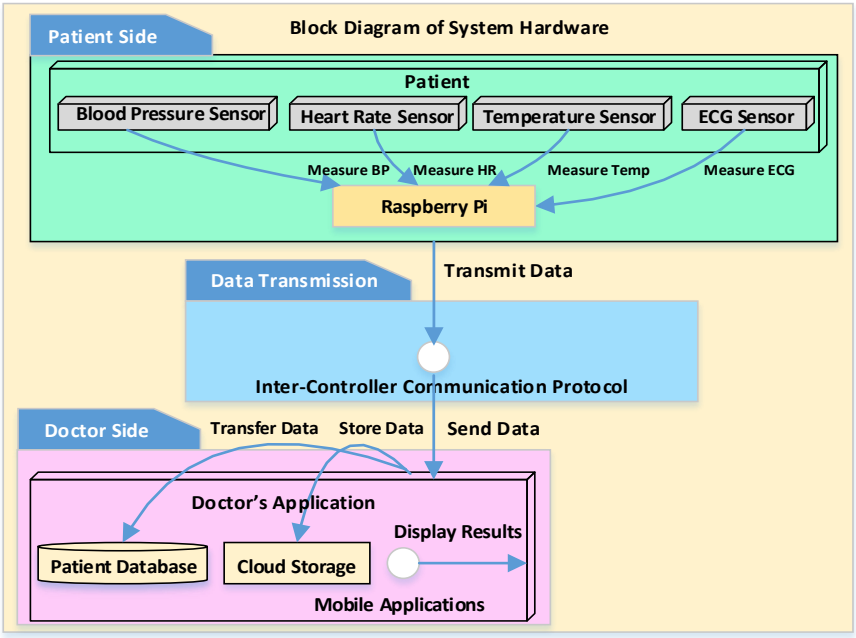


Fig. 2: Health monitoring system hardware with Raspberry Pi and IoT sensors.

To reach the general conclusion and assess the proposed system performance, the results for different machine learning algorithms including SVM, XGBoost, logistic regression, Naive Bayes [30], and random forest are obtained based on the trends reported in Fig. 2.

The selected models are supervised learning algorithms, which are trained based on expected results and evaluated on unseen data to check their validity. As mentioned earlier, models are SVM, XGBoost, random forest, and logistic regression. Table 1 shows the analysis and prediction values for different machine learning algorithms based on their performance metrics.

Table 1: Performance metric comparison

Algorithm	Accuracy	Precision	Recall	F1
SVM	0.85	0.85	0.85	0.85
XGBoost	0.88	0.88	0.88	0.88
Random Forest	0.86	0.86	0.86	0.86
Logistic Regression	0.91	0.91	0.91	0.91
Naive Bayes	0.85	0.86	0.85	0.85

As illustrated in Table 1, all of the investigated algorithms demonstrate good accuracy and ability to generalize on unseen data. The SVM model shows 82% and 85% accuracy, respectively, for training and testing parts of the dataset. The XGBoost obtained even better results compared to its precedent with a training accuracy of 90% and testing accuracy of 87.88%, while logistic regression stands out with training and testing accuracy of 86.67% and 91%, respectively.

The random forest model fulfills consistent performance in both the training and testing sets, with 85% for training and 86% for testing. Finally, Naive Bayes presents 82% training accuracy and 85% testing accuracy. As a conclusion, the logistic regression is considered as the most appropriate model to handle small to medium datasets and achieve better results compared to other models as shown in Fig. 3.

Results and Discussion

As seen, five models were trained and evaluated including SVM, XGBoost, random forest, logistic regression, and Native Bayes. Based on the obtained results and observations, it can be said that logistic regression outperforms the others by obtaining an F1 score, recall, and precision of 0.91, followed by XGBoost with 0.88 in all metrics. SVM, random forest, and Native Bayes also showed competitive results, with accuracies of 0.85, 0.86, and 0.85, respectively. The Receiver Operating Characteristic Area Under Curve (ROC-AUC) [31] results demonstrated the power of logistic regression and XGBoost, highlighting their activity, consistency with biomarkers, and training accuracy and efficiency.

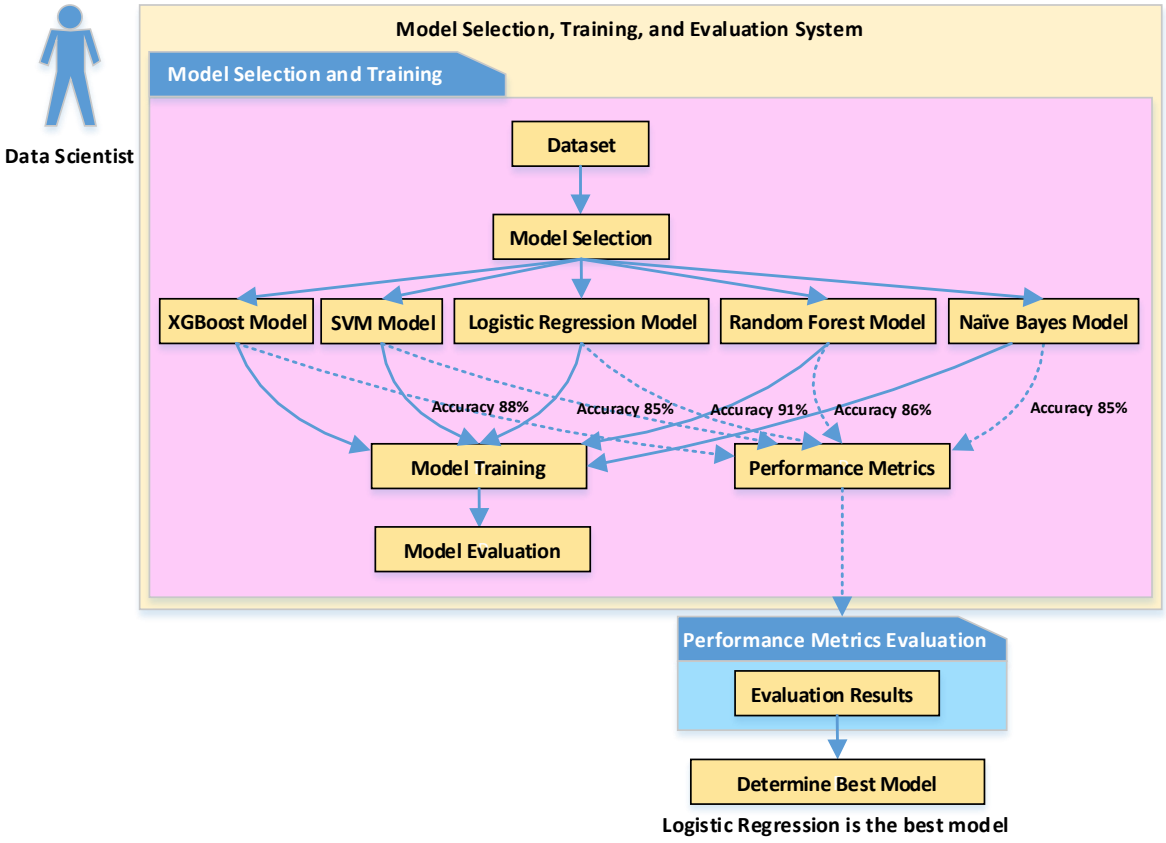


Fig. 3: Preprocessing operations.

As seen, logistic regression, a model designed for binary classification, achieved the highest F1 score equal to 0.91 compared to other models. Its strong performance is attributed to its suitability for binary classification, linear decision boundary, robustness to outliers, interpretability, and computational efficiency. Logistic regression assumes a linear relationship between input features and health condition probability, making it effective in separating classes. Its robustness reduces the impact of extreme values on the decision boundary, making it suitable for datasets with outliers. Its interpretability and feature importance make it valuable for understanding underlying factors influencing health condition classification. However, the choice of the best model depends on the dataset, the complexity of the problem, and the desired balance between interpretability, accuracy, and computational cost. Further analysis, including feature importance analysis, could provide more insights into its performance. The strong performance of logistic regression compared to XGBoost is likely due to the linear nature of the data, the relatively small dataset size, and the potential for logistic regression to be more robust and interpretable. However, as the dataset and problem complexity increase, XGBoost ability to capture nonlinear patterns and its ensemble nature may become more advantageous.

Conclusions

The results summarize the validity of the prediction of health conditions by the logistic regression algorithm, displaying an F1 score, recall, and accuracy of 0.91, indicating excellent classification performance and good diagnosis. The choice of the logistic regression depends on the dataset, the complexity of the problem, and the desired balance between interpretability, accuracy, and computational cost. The algorithm that was ranked second in terms of accuracy is XGBoost with an accuracy of 0.88, which enhances the reliability of advanced machine-learning techniques in health monitoring, its ability to capture nonlinear patterns, and its ensemble nature may become more advantageous due to the linear nature of the data. The study highlights that comprehensive pre-processing of balanced data sets and machine learning models can significantly enhance the detection and diagnosis of health problems. It can be argued from the various models and their consistent results that objective biometric data can be incorporated into primary care assessments.

In the future, this type of research must focus on a larger and more diverse data set, in addition to greater features, to improve the accuracy of the model. These developments lead to more effective care in the health field to anticipate the disease, as well as early detection and continuous health monitoring through machine learning techniques, leading to improved health outcomes and patient well-being. The proposed system has the potential to enhance its performance by incorporating additional data sources and sensors. This could include incorporating biometric sensors, such as blood glucose sensors, oxygen saturation sensors, EEG sensors, accelerometers, and gyroscopes, environmental sensors, wearable sensors, and advanced machine learning capabilities. The system could also leverage multi-modal data analysis and deep learning models to analyze complex data from multiple sources, simultaneously. The system should be designed with a modular architecture for easy integration and adaptation without significant re-engineering. The potential benefits of expansion include improved accuracy, early detection, personalized treatment, and proactive care. However, challenges such as data management, privacy, and algorithm complexity need to be addressed. The system's success depends on addressing these challenges, which include efficient data storage, processing, and analysis techniques, as well as addressing challenges like data management, privacy, and algorithm complexity. Overall, the proposed system has significant potential for expansion, enhancing accuracy, early detection, personalized treatment plans, and transitioning healthcare towards a more proactive approach.

Author Contributions

F. Ahmed Shaban and S. Golshannavaz designed the model, implemented the setup, collected the data, carried out the analysis, interpreted the results, and wrote the manuscript.

Acknowledgment

The authors would like to thank the respected referees for their accurate reviewing of this paper.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

IoT	Internet of Things
AI	Artificial Intelligence
ECG	Electrocardiogram
SVM	Support Vector Machine
XGBoost	Extreme Gradient Boosting
EEG	Electroencephalography

ECD	Electronic clinical data
GBDT	Gradient-boosting decision tree
ROC-AUC	Receiver Operating Characteristic Area Under Curve

References

[1] G. Halfacree, The official Raspberry Pi Beginner's Guide: How to use your new computer. Raspberry Pi Press, 5th ed., 2023.

[2] R. A. Mouha, "Internet of things (IoT)," J. Data Anal. Inf. Process. 9(2): 77-77, 2021.

[3] S. Maqsood, et al., "A survey: From shallow to deep machine learning approaches for blood pressure estimation using biosensors," Expert Syst. Appl., 197: 116788, 2022.

[4] V. Kaul, S. Enslin, S. A. Gross, "History of artificial intelligence in medicine," Gastrointest. Endos., 92(4): 807-812, 2020.

[5] F. J. Abdullayeva, "Internet of things-based healthcare system on patient demographic data in Health 4.0," CAAI Trans. Intell. Technol., 7(4): 644-657, 2022.

[6] K. M. Hosny, et al., "Internet of things applications using Raspberry-Pi: a survey," Int. J. Electr. Comput. Eng., 13(1): 902-910, 2023.

[7] A. Golparvar, O. Ozturk, M. K. Yapici, "Gel-free wearable electroencephalography (EEG) with soft graphene textiles," presented at the IEEE 2021 Sensors, Sydney, Australia, 2021.

[8] P. Srinivas, et al., "Support vector machines based predictive seizure care using IoT-Wearable EEG devices for proactive intervention in epilepsy," in Proc. 2024 IEEE 2nd International Conference on Computer, Communication and Control (IC4), 2024.

[9] M. Alhayani, N. Alallaq, M. Al-Khiza'ay, "Optimize one max problem by PSO and CSA," presented at the International Congress on Information and Communication Technology, 2023.

[10] S. A. Alzakari, et al., "Enhanced heart disease prediction in remote healthcare monitoring using IoT-enabled cloud-based XGBoost and Bi-LSTM," Alexandria Eng. J., 105: 280-291, 2024.

[11] S. Pokhrel, R. Chhetri, "A literature review on impact of COVID-19 pandemic on teaching and learning," Higher Educ. Future, 8(1): 133-141, 2021.

[12] Y. Izza, A. Ignatiev, J. Marques-Silva, "On explaining decision trees," arXiv preprint, arXiv:11034, 1:21, 2020.

[13] M. Alhayani, M. Al-Khiza'ay, "Analyze symmetric and asymmetric encryption techniques by securing facial recognition system," in Proc. International Conference on Networking, Intelligent Systems and Security, 2022.

[14] S. Amini, et al., "Urban land use and land cover change analysis using random forest classification of landsat time series," Remote Sens., 14(11), 2654, 2022.

[15] V. Jackins, et al., "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," J. Supercomput., 77(5): 5198-5219, 2021.

[16] L. Dai, et al., "Influence of soil properties, topography, and land cover on soil organic carbon and total nitrogen concentration: A case study in Qinghai-Tibet plateau based on random forest regression and structural equation modeling," Sci. Total Environ., 821: 153440, 2022.

[17] W. Python, Python releases for windows, 2021.

[18] S. R. Chanthati, "Second version on a centralized approach to reducing burnouts in the IT industry using work pattern monitoring using artificial intelligence using MongoDB atlas and python," World J. Adv. Eng. Technol. Sci., 13(1): 187-228, 2024.

[19] D. A. Pisner, D. M. Schnyer, Support vector machine, Machine learning: Methods and Applications to Brain Disorders, Elsevier, 101-121, 2020.

[20] T. Latchoumi, et al. "Enhancement in manufacturing systems using Grey-Fuzzy and LK-SVM approach," in Proc. 2021 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT), 2021.

[21] Q. Li, et al., "A comparative study on the most effective machine learning model for blast loading prediction: From GBDT to Transformer," Eng. Struct., 276: 115310, 2023.

[22] Y. Qiu, et al., "Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration," Eng. Comput., 38 (5): 4145-4162, 2022.

[23] P. Schober, T. R. Vetter, "Logistic regression in medical research," Anesth. Analg., 132(2): 365-366, 2021.

[24] N. Boyko, K. Boksho, Application of the Naive Bayesian Classifier in Work on Sentimental Analysis of Medical Data, IDDM, 3rd, 2020.

[25] E. M. K. Reddy, Introduction to Naive Bayes and a review on its subtypes with applications, Bayesian reasoning and Gaussian processes for machine learning applications, 1-14, Taylor& Francis, 2022.

[26] D. Tramontin, Random forest implementation for classification analysis: default predictions applied to Italian companies, Ca'Foscari University of Venice, 2020.

[27] A. E. Ulloa-Cerna, et al., "rECHOmmend: an ECG-based machine learning approach for identifying patients at increased risk of undiagnosed structural heart disease detectable by echocardiography," Circulation, 146(1): 36-47, 2022.

[28] M. Shao, et al., "A review of multi-criteria decision making applications for renewable energy site selection," Renewable Energy, 157: 377-403, 2020.

[29] G. Xu, "IoT-assisted ECG monitoring framework with secure data transmission for health care applications," IEEE Access, 8: 74586-74594, 2020.

[30] I. Wickramasinghe, H. Kalutarage, "Naive Bayes: applications, variations and vulnerabilities: A review of literature with code snippets for implementation," Soft Comput., 25(3): 2277-2293, 2021.

[31] J. Muschelli, "ROC and AUC with a binary predictor: a potentially misleading metric," J. Classif., 37(3): 696-708, 2020.

Biographies

Fahad Ahmed Shaban was born in Mosul - Iraq, in 1993. He holds a bachelor's degree in Computer Technology Engineering from the Northern Technical University, College of Engineering Technology. He ranked first in his class for the year 2015-2016, and holds a master's degree in Computer Technology Engineering from the Northern Technical University, College of Engineering Technology, Mosul, Iraq in 2018-2019. Currently, he is working as a computer engineer at the Ministry of Water Resources since 2023. He is currently a Ph.D. student in Computer Engineering, Urmia University, Urmia, Iran, in the Faculty of Electrical and Computer Engineering. His research interests are related to Internet of Things, computer network design, project management system design, database design, and neural networks.

- Email: f.ahmedshaban@urmia.ac.ir
- ORCID: 0009-0001-7971-1715
- Web of Science Researcher ID: -
- Scopus Author ID: -
- Homepage: https://scholar.google.com/citations?user=_vcr6gQAAAAJ&hl=en

Sajjad Golshannavaz was born in Urmia, Iran, in 1986. He received the B.Sc. (Honors) and M.Sc. (Honors) degrees in Electrical Engineering from Urmia University, Urmia, Iran, in 2009 and 2011, respectively. He received his Ph.D. degree in Electrical Power Engineering from School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran, in 2015. Currently, he is an Associate Professor in Electrical Engineering Department, Urmia University, Urmia, Iran. Since 2014 he has been collaborating with the smart electric grid research laboratory, Department of Industrial Engineering, University of Salerno, Salerno, Italy. His research interests are in smart distribution grid operation and planning studies, design of distribution management system (DMS), demand side management (DSM) concepts and applications, microgrid design and operation studies, design of energy management system (EMS), application of FACTS Controllers in Power systems, application of intelligent controllers in power systems.

- Email: s.golshannavaz@urmia.ac.ir
- ORCID: 0000-0003-4999-8281
- Web of Science Researcher ID: AAB-5779-2020
- Scopus Author ID: 36677225300
- Homepage: https://scholar.google.com/citations?user=YzezRFUAAAAJ&hl=en

How to cite this paper:
F. Ahmed Shaban, S. Golshannavaz, "A Machine-Learning-based Predictive Smart Healthcare System," J. Electr. Comput. Eng. Innovations, 13(1): 181-188, 2025.

DOI: 10.22061/jecei.2024.11092.765

URL: https://jecei.sru.ac.ir/article_2214.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Structure Learning for Deep Neural Networks with Competitive Synaptic Pruning

A. Ahmadi, R. Mahboobi Esfanjani *

Department of Electrical Engineering, Sahand University of Technology, Tabriz, Iran.

Article Info

Article History:
Received 26 July 2024
Reviewed 27 September 2024
Revised 16 October 2024
Accepted 31 October 2024

Keywords:
Deep neural networks
Synaptic pruning
Distillation column
PID tuning

*Corresponding Author's Email Address: mahboobi@sut.ac.ir

Abstract

Background and Objectives: A predefined structure is usually employed for deep neural networks, which results in over- or underfitting, heavy processing load, and storage overhead. Training along with pruning can decrease redundancy in deep neural networks; however, it may lead to a decrease in accuracy.

Methods: In this note, we provide a novel approach for structure optimization of deep neural networks based on competition of connections merged with brain-inspired synaptic pruning. The efficiency of each network connection is continuously assessed in the proposed scheme based on the global gradient magnitude criterion, which also considers positive scores for strong and more effective connections and negative scores for weak connections. But a connection with a weak score is not removed quickly; instead, it is eliminated when its net score reaches a predetermined threshold. Moreover, the pruning rate is obtained distinctly for each layer of the network.

Results: Applying the suggested algorithm to a neural network model of a distillation column in a noisy environment demonstrates its effectiveness and applicability.

Conclusion: The proposed method, which is inspired by connection competition and synaptic pruning in the human brain, enhances learning speed, preserves accuracy, and reduces costs due to its smaller network size. It also handles noisy data more efficiently by continuously assessing network connections.

Introduction

The parallel-distributed nature of neural networks gives them a great capacity for learning and generalization. They have therefore been used to address a variety of issues, including modelling in automatic control. During training, information is stored in the weighted connections between the neurons, just like in the human brain. In a network, the number of hidden layers and their corresponding weights determine its structure, which is a major factor in its performance. Large and small networks both have a number of disadvantages. The contrast between curve fitting and neural network training helps to explain why networks with fewer free parameters perform better in terms of generalization, as shown by theory [1] and experience [2]. Moreover, the knowledge included in the small trained networks is easier to be understood and thus the abstraction of simple rules can be facilitated [3]. Finally, small networks require very little resources to construct in any physical computer environment. Larger networks suffer from the overfitting issue and are less able to generalize. They are also quite costly and complex. Therefore, choosing the appropriate size is crucial to having an efficient and quick network.

Deep Networks and Pruning

Deep neural networks (DNNs) have been the main reason for recent improvements in machine learning.

Numerous of these networks require a large amount of memory space and computing power [4]. These features make it impossible to implement networks in situations where resources are few, like mobile phones [5]. Therefore, along with the high potential and computing power of deep networks, it is necessary to solve the limitations of these popular networks so that they can be used optimally, at a lower cost, and at a high speed in real applications. In this regard, providing a new suitable method for this problem and reaching a network with a suitable size is the main topic of this work.

Neural network pruning—the targeted removal of parameters from an existing network—is a well-liked method for lowering these resource requirements. The objective is to build a smaller network with the same degree of precision, despite the fact that the original network was huge and precise. The pruning notion has attracted much attention in the last decade, and its popularity has increased due to the emergence of deep neural networks [6]. We need to be clear about the most important thing in order to prune effectively: which connections are the best candidates for pruning? Taking into account the pruning process in the human brain provides useful solutions [7]. During learning, frequently used synapses become stronger, while rarely used synapses become weaker and more likely to disappear [8]. On the other hand, to avoid accuracy falling, consideration must be given to the values of the connections.

Therefore, the approach presented in this work solves the usual weaknesses. In this way, it uses a more accurate criterion instead of a traditional weight factor, and secondly, it examines the behavior of connections in successive stages and It looks at strong, medium and weak connections differently. Instead of immediate removal, it measures the existence of multiple warnings regarding the connection's effectiveness.

In short, the abovementioned features are combined here to develop an efficient pruning technique. In the proposed algorithm:

First, network connections are evaluated initially based on a global gradient magnitude criterion. This evaluation also lasts for the next steps, and the scores are updated continuously. This criterion gives encouraging points for effective connections and destructive points for weak connections.

Second, inspired by the brain pruning strategy, a connection with a weak score is not removed immediately; instead, it is eliminated when its net score reaches a certain number with less than a certain threshold.

Third, dividing the connections in the network into three categories: connections with high impact, medium impact and weak impact, the first category will receive rewards and their chances of survival will increase, and the third category will be penalized and their chances of being pruned will increase.

Fourth, to enhance the network's accuracy and quality, the pruning rate is explicitly determined for each specific layer of the network.

The structure of this document is as follows: Section 2 is an overview of the related prior research. Furthermore, we provide our approach in Section 3. Comparative simulation results are shown in Section 4, and we conclude the paper in Section 5.

Related Works

This section presents some relevant works on network architecture optimization. Differently from shallow networks, which have only one hidden layer, deep networks have two or more hidden layers, which help to store and organize data efficiently; namely, they serve a more precise purpose than a superficial one. Deep networks' capacity for memorization greatly aids in managing uncertainties. Training a smaller neural network to mimic the bigger model is one method of lowering the neural network's computational complexity. Network distillation is a method that Hinton et al. [9] suggested. The primary flaw with this process is the need to predefine the smaller model's structure. Pruning, which is the process of eliminating individual neurons that provide less effect on the output of a trained network, is another method for shrinking and speeding up a model.

Therefore, compared to the above two approaches, the pruning method is usually preferred. Assuming a tight relationship between weight size and significance, the traditional and simple method is to select a threshold and prune those synapses whose weights are below it [10]. Several studies have cast doubt on this tactic [11]. Actually, by employing this method, certain advantageous synapses whose weights happen to fall below the threshold may be pruned. Some studies concentrate on creating suitable standards for assessing the significance of connections and eliminating the least important ones. Molchanov et al. considered the l2-norm of the kernel weights in addition to the feature map's mean, standard deviation, and percentage activation [12]. They also compared activations and predictions using mutual information as a criterion.

Molchanov et al. introduced first-degree Taylor expansion as a tool for evaluating synaptic significance. In order to determine synaptic significance, LeCun et al. [13] and Hassibi and Stork [14] employed a diagonal Hessian matrix and concentrated on the second-order term of a Taylor expansion. In order to eliminate the most replaceable filters that include extraneous data, He et al. proposed a geometric median-based filter-cutting approach [15]. The neuron significance score (NISP) method [16] propagates the final answers' relevance

ratings to each neuron in the network, as suggested by Yu et al. The least important neurons were then eliminated in order to prune the convolutional neural network. In addition to the feature maps they connected, Li et al. eliminated the filters with relatively low weights [17]. He et al.'s gentle pruning approach allows the model to be trained utilizing the trimmed filters after the pruning [18]. Transmits the relevance scores of the final replies to every neuron in the network, as proposed by Yu et al. The convolutional neural network was then pruned by removing the least significant neurons. Li et al. removed the filters with relatively low weights in addition to the feature maps they linked [19]. During training, every neuron in dropout is probabilistically eliminated, but during inference, they are allowed to rebound. A network's complexity does not go down while using this method. By randomly changing a portion of a neural network's weights to zero, Drop Connect is used to regularize neural networks [20]. Though it occasionally surpassed dropout, it learned more slowly than both the original network and the dropout network. One of its drawbacks is that it increases the amount of time needed for training. A dropout network often requires two to three times as long to train as a neural network with the same design that is used normally. MeProp altered a relatively small portion of the settings for each back-propagation phase [21]. These methods do not introduce any fundamental changes to the structure of the network.

An architecture for a network is optimized by evolutionary approaches. Both the topology and the weights are optimized concurrently in evolving neural networks. Numerous network structure-related factors found in the genome have been refined via evolution. An evolutionary approach evaluates a network's performance using a fitness function.

Typically, one of these functions is accurate classification [22], the other is the size of the network, which comprises factors like the quantity of connections or neurons [23]. After several rounds, an artificial neural network with evolutionary capabilities can identify the optimal network architecture. An evolutionary optimization approach was proposed by Zhao et al. [24]. A network was pruned to the ideal topology using genetic algorithms [25]. These approaches focus on network design optimization to achieve the optimal trade-off between accuracy and complexity; nonetheless, there is a significant degree of unpredictability in both approaches, and significant side trips may occur due to the lengthy development process.

In summary, all of the mentioned techniques assess connections in a single phase without keeping track of connections' behavior over time or allocating different pruning rates for different layers. As explained with reasons and references, the weighted criterion is not accurate and other single-factor criteria have weaknesses. The very important point is that we classify the connections in three categories (weak, medium and strong) and in addition to gradually reducing the weaker connections scores, we also gradually strengthen the strong connections, which is not the case in previous works in this format.

Therefore, the continuous control of connections, using our proposed criteria, which is different from the usual criteria, and on the other hand, considering the positive and negative points for network connections are the main differences between the present work and previous studies. A brain-inspired competitive synaptic pruning technique is introduced in place of the traditional omission technique.

Proposed Method

A novel brain-inspired competitive pruning technique is developed in this section. Network connections are supposed to compete for survival, and the basis of this competition is based on the weighted average score that each connection gets. Like other evolutionary-based ones, the proposed algorithm starts with an initial population (here the initial at the beginning).

Minimal-value deletion prunes all synapses whose weights are below a threshold. Magnitude-based approaches are common baselines in the literature. However, this method may prune some useful synapses whose weights are incidentally below the threshold. Gradient-based methods are less common but are more accurate, simple to implement, and have recently gained popularity.

The key idea is that after considering “Weight × Gradient” as the appraisal criterion, connections are classified into three categories: down (for example, the lower 20%), top (for example, the upper 20%), and the area between them. In each step, connections in the top class get a positive score, connections in the lowest category get a negative score, and connections in the middle set get zero points. In this way, we set a reward for good connections and a penalty for poor connections. At each stage, the net score of each connection is determined by adding the current score to the previous one. Therefore, we have updated net scores for all connections. Inspired by the human brain [26], the next important fact is that whenever the net score of a connection reaches a certain threshold—in other words, it receives a certain number of warnings—that connection is removed from the network.

It is worth noting that using the gradient in the evaluation criterion, as weight × gradient adds the rate of change to it and because of its dynamic nature increases the accuracy of the method.

Fig. 1 shows how synaptic pruning works. This issue is an important part of the presented method and it is more accurate and better than the usual methods, as inspired by the human brain, the removal of network connections happens gradually and step by step. In this way, if a connection gets lower scores several times, it is concluded that its importance for the network is low and therefore it is removed. It is clear that in synaptic pruning, there are two important parameters that we need to specify. The weight threshold shows which connections could benefit from pruning. The maximum allowed warning specifies how long the associated connection will be active before being erased. After establishing a threshold value, we keep an eye on the connection scores. In Fig. 2, the procedure of the proposed pruning technique is depicted. In the case of reaching the net score of a given connection to the threshold, it is pruned.

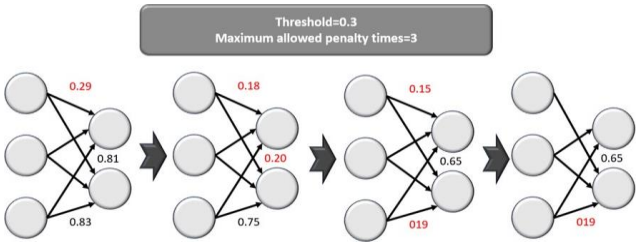


Fig. 1: Schematic of the synaptic pruning concept.

As can be seen in the flowchart, continuous evaluation of connections and obtaining a net score for each connection leads to a decision regarding removal or retention. Algorithm 1 provides a detailed presentation of the pruning pseudocode. As discussed, it is insufficient to rely just on the weighted domain, and there is a significant chance that some crucial network connections will be overlooked. Sorting the connections addressed this problem since, after the connections that may be deleted are identified, the value is not decreased all at once to complete the removal process. Actually, we warn the pruning candidates one after the other and prune them in response to these cautions. The threshold limit is chosen by trial and error. All connections at each stage are evaluated by the considered criterion and sorted based on the scores they get. The summary of the approach is that after sorting the scores of the connections, we have three categories: connections whose score are in the top 20%, as well as connections that are in the bottom 20% and connections whose scores are between these two areas. At each stage, one unit is added to the connections with them, and the connections with the lowest 20% are fined; i.e. their score are reduced by one unit. Therefore, the strong ones are strengthened and the weak ones are weakened, and they are candidate for pruning. There is a threshold value, if the connection score is less than that, a warning will be given, and after specified warnings, it will be pruned. Concisely, each link in the network is evaluated based on network error. Unless otherwise stated, we compute the errors resulting from omitting each link and, upon sorting according to the error, identify the connections with the highest number of errors. Connections are eliminated depending on the pruning rate, which is set by the designer considering factors such as layer percent.

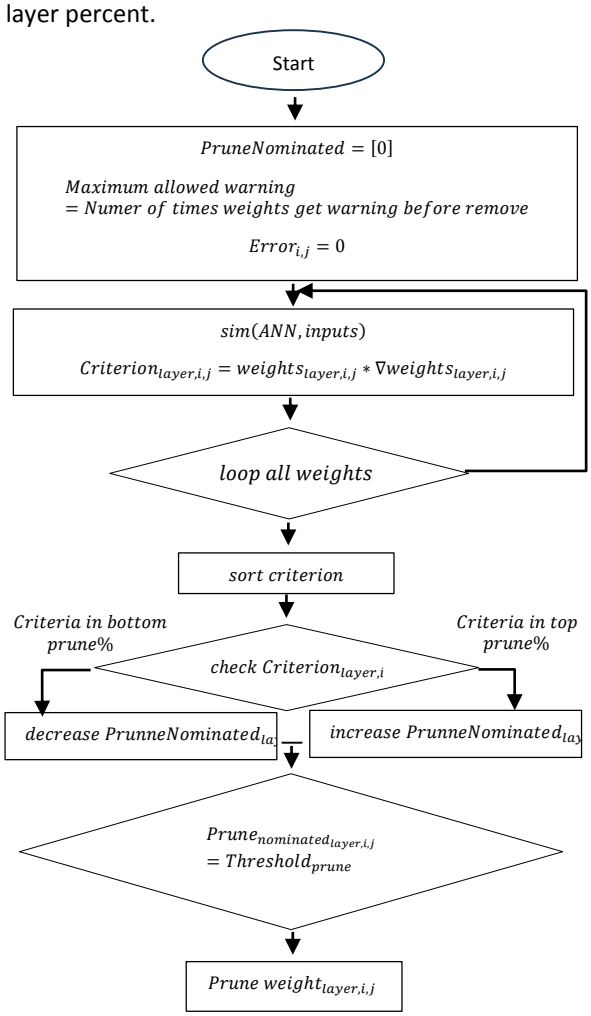


Fig. 2: Flowchart of proposed pruning.

The key idea is that, in addition to evaluation of each connection, the contribution of the layers is also taken into account in the pruning; in this way, the layer value is defined as the sum of all connection weights in the mentioned layer, and we also compute the layer percent, by dividing layer value into the total values of the network layers, as follows:

$$value_{layers} = \sum_{i=0}^{layer\ number} value_{layer_i} \quad (1)$$

$$layer\ percent_i = \frac{value_{layer_i}}{value_{layers}} \quad (2)$$

$$i = 0, \ldots, number\ of\ layers$$

Many of pruning techniques are predicated on the idea that there is a significant correlation between a weight's magnitude and significance. Recent researches have questioned this assumption and shown a notable discrepancy in the association between empirically optimum one-step decisions and weight-based pruning judgments [11].

Algorithm 1: Pruning Pseudo code

$Generate\ all\ zero\ PrunneNominated\ variable\ with$
$structure\ and\ size$
$Maximum\ allowed\ warning$
$= number\ of\ time\ we\ want\ \ a\ weight\ to\ not\ be\ cutted.$
$Initialize\ Error_{layer,i,j} = 0$
$Initialize\ Prune_{percent} = 20\%\ ,$
$shows\ percent\ of\ weights\ nominated\ for\ pruning,$
$Consider\ 20\ top\ and\ down\ for\ reward\ as\ penalty\ domain$
$Construct\ an\ evaluate\ table\ as\ below:$

$$evaluated\ table = \begin{bmatrix} error & \tau & m_{row} & m_{col} & l_{row} & l_{col} \\ \vdots & & & \vdots & & \end{bmatrix}$$

In which:

$\tau = 1$	$\tau = 2$	$\tau = 3$
$bias$	$input\ layer$	$hidden\ and\ output\ layers$
$bias\ weights\ for\ m_{row}th\ layer\ m_{col} = 1$	$weights\ of\ m_{row}th\ input\ m_{col} = 1$	$weight\ for\ layer\ m_{row}to\ m_{col}$
$row\ l_{row}of\ weight\ matrix\ l_{col} = 1$	$l_{row}\ and\ l_{col}shows\ weigh\ matrix$	$l_{row} = 1\ l_{col}shows\ weight\ matrix\ element$

$normalize\ weights\ in\ [0,1]\ range$
$loop\ weights$
$Initialize\ weight_{layer,i,j}$
$weight_{bias} = weight_{0,i,j}$
$weight_{hidden} = weight_{layer,i,j}$
$Criterion_{layer,i,j} = weights_{layer,i,j} \times \nabla weights_{layer,i,j}$
$sort\ Criterion\ \ matrix\ ascending$
$\boldsymbol{for}\ Criterion_{layer,i,j} = 1\ to\ Prune_{percent} \times count(Criterioan)$
$PrunneNominated_{layer,i,j} = PruneNominated_{layer,i,j} - 1$
$\boldsymbol{for}\ Criterion_{layer,i,j} = (1 - Prune_{percent}) \times count(Criterion) to\ count(Criterion)$
$PruneNominated_{layer,i,j} = PruneNominated_{layer,i,j} + 1$
$\boldsymbol{for}\ (layer, i, j) = PruneNominated_{layer,i,j} = maximum\ allowed\ warning$
$weight_{layer,i,j} = 0$

$$value_{layers} = \sum_{i=0}^{layer\ count} value_{laye_i}$$

$$layerpercent_i = \frac{value_{layer_i}}{value_{layers}}\ i = 0\ ..\ layer\ count$$

The use of the gradient somehow adds the rate of change to the criterion and dynamically increases the accuracy Our method, which is more accurate than other baselines, prunes the weights with the lowest absolute value of (weight * gradient), evaluated on a batch of inputs. Therefore, it is very significant and important that, firstly, our proposed criterion is much more accurate and dynamic than the simple weight criterion, and on the other hand, in addition to reducing the score of weak connections, we also give points to strong connections, and besides all this, we also use synaptic pruning in an innovative setting.

Results and Discussion

We verify the merits of the presented technique by two practical examples: identification of a refinery distillation tower and also adjusting the coefficients of a PID controller, which is the most famous and widely used controller in process industries. For the first one we use a deep feedforward neural networks as the system model, and for the second one a deep recurrent neural network as the online tuner of the controller.

Distillation Tower (Refinery)

The effectiveness of the proposed strategy is evaluated using a neural network model of the distillation tower in a refinery operation. The goal is to find out how this algorithm may improve identification accuracy and convergence speed while dealing with ideal and noisy data. Refineries are incomplete without the distillation tower, a multi-input, multi-output (MIMO) nonlinear system. One tool for separating solution components is a distillation column. Actually, the boiling point difference and volatility of the constituents of a solution are used to separate them in the distillation tower. Crude oil refining is one of the principal applications for industrial distillation towers, which are widely utilized in many process sectors. The distillation process is used in the oil business to separate various hydrocarbons according to how volatile they are. One of the most commonly utilized towers is the ethane-ethylene distillation column. The production of high-purity ethylene is necessary because of its importance. Our data comes from an identification experiment using an ethane-ethylene distillation column [27]. The data contains four series:

U_dest, Y_dest: without noise (ideal series)
U_dest_n10, Y_dest_n10:
10 percent additive white noise
U_dest_n20, Y_dest_n20:
20 percent additive white noise
U_dest_n30, Y_dest_n30:
30 percent additive white noise

There are 180 samples for neural network training. The following describes the inputs and outputs:

The inputs of the systems are: 1) the proportion between feed flow and reboiler duty; 2) the relationship between feed flow and reflux rate; 3) the proportion between the feed flow and the distillate; 4) the composition of the input ethane; and 5) the top pressure. Outputs of the system are: 1) top ethane composition; 2) bottom ethylene composition; and 3) top-bottom differential pressure. So, we employ a deep network with 90 connections, 5 inputs, and 3 outputs (Fig. 3). If we first have the proper weights training, it can be utilized the deep network's capabilities. Secondly, we can use our structural optimization approach to discover the optimal structure for the network and prevent over-fitting, which will speed up the network.

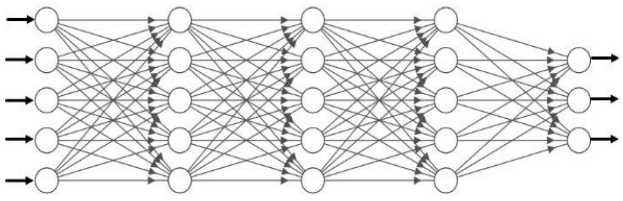


Fig. 3: Applied deep neural network.

Using the available data, we first train the network. Fig. 4 illustrates how the network's performance varies after each training epoch. Our performance function is the mean square error (MSE). Three curves with various colors are included for the test, validation, and training sets of data. The label on the horizontal axis shows the number of training cycles (epochs) of the network. The network performed its best in the 9th epoch, as evidenced by the validation data. Furthermore, Fig. 5 displays the regression curves for the test, validation, and training sets of data.

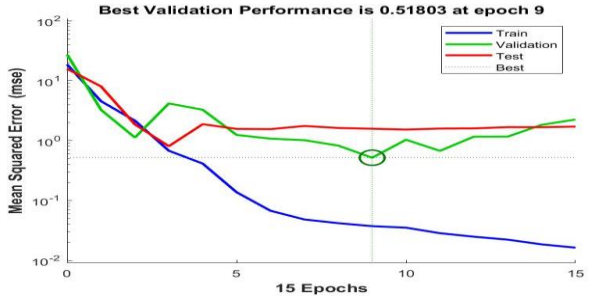


Fig. 4: Performance of the deep neural network.

More information from the training is shown in the training state visualization (Fig. 6); for instance, the "val fail" graph indicates the epoch in which the validation data evaluation was rejected. The cumulative number of failed assessments is displayed on this graph. When the network fails six consecutive assessments, training ends.

Two important measures are usually used for comparison to the simple dropout method [20]: The new size divided by the original size is defined as the compression ratio. The theoretical speedup is defined as the ratio of the initial number of multiplications and additions to the new number. Comparison statistics for two circumstances are reported in Table 1. Note that the *weight* × *gradient* is employed here.

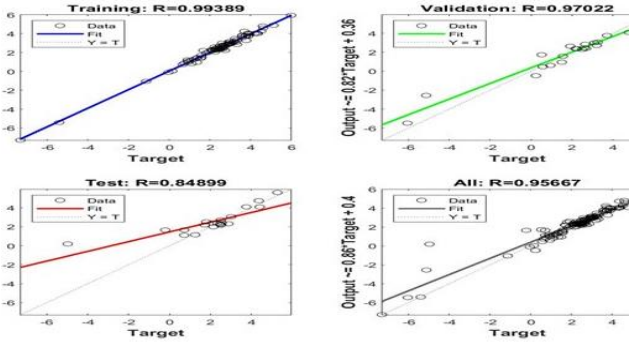


Fig. 5: Regression for the training, validation and test data.

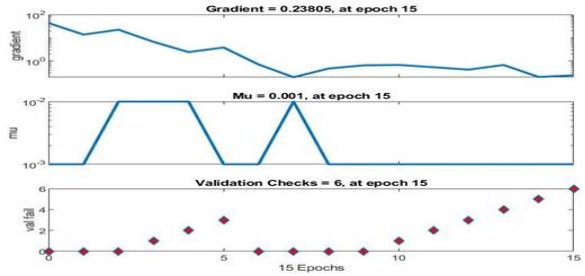


Fig. 6: Training state.

Also, in Table 2, the proposed pruning method is compared to the conventional dropout method wherein only weight is used. It verifies the superiority of the proposed scheme.

Table 1: Comparison to dropout method with weight × gradient

NN Type	Shallow (1 hidden layer)			Deep (3 hidden layers)		
	Initial	Dropout	Proposed	Initial	Dropout	Proposed
Accuracy (%)	76.6	77.1	77.9	82.6	83.8	86.3
Compression (%)	-	47	47.2	-	58	59.5
Execution Time (ms)	15	17.4	13.5	17.5	19.2	16.3

Table 2: Comparison to dropout method with only weight

NN Type	Shallow (1 hidden layer)			Deep (3 hidden layers)		
	Initial	Dropout	Proposed	Initial	Dropout	Proposed
Accuracy (%)	76.6	78.2	77.9	82.6	84.8	86.3
Compression (%)	-	48.3	47.2	-	59.4	59.5
Execution Time (ms)	15	17.3	13.5	17.5	18..2	16.3

As seen, our pruning strategy leads to a speedup in training and network performance. The suggested pruning strategy may be easily extended to other intelligent process industries. One of the most important problems in measurement and control is noisy data, which is frequently found in actual industrial settings. When working with noisy data that has 10%, 20%, and 30% noise, the outcomes of the shallow network and the deep network, which is pruned using the introduced algorithm, are compared in Fig. 7. It is clear that the suggested structure works much better, especially with noisy data.

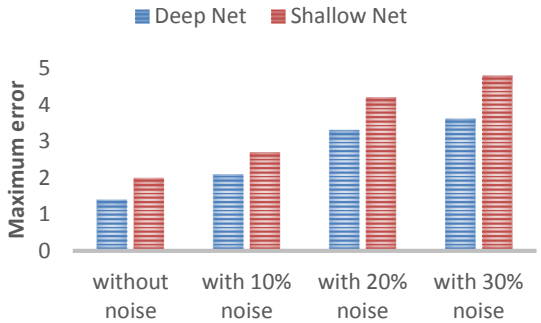


Fig. 7: Deep and shallow networks comparison in the noisy data treatment.

Briefly, the distillation tower is modelled using a deep network pruned using the suggested approach, and its effectiveness was shown in comparison to the shallow network. In order to compare the suggested model with other neural network-based models, we also compared the RMSE criteria between the model and three other structures in Table 3. The two structures that are being discussed are NARX structure-based neural networks

(which use both the Levenberg-Marquardt and the Steepest Descent algorithms) and nonlinear auto-regressive with exogenous inputs (NARX)-based ANFIS [28]-[30]. The comparison of errors clearly shows that the proposed technique is better than the other structures.

Table 3: RMSE for neural networks models, ANFIS, and the proposed CONCOMP

Outputs	Steepest Descent	Levenberg Marquardt	ANFIS	CONCOMP
Top Composition	0.639	0.2090	0.0421	0.0222
Bottom Composition	1.3127	0.4913	0.031	0.021
Pressure Difference	1.0053	0.2480	0.0189	0.0112

PID Controller

PID controllers are frequently mistuned, particularly in unreliable situations. Intelligent techniques are used recently to develop adaptive PID controllers. In order to mitigate the effects of uncertainties in the closed-loop control system, a deep dynamic neural network is used here to tune the parameters of the traditional PID controller. By using the proposed pruning method, simpler tuner is achieved and consequently the computational load is decreased.

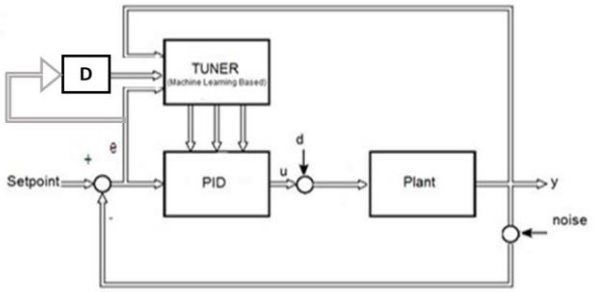


Fig. 8: Closed loop PID control with neural network tuner.

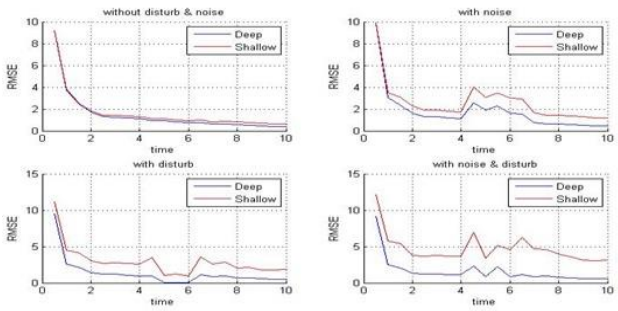


Fig. 9: RMSEs with shallow and pruned deep neural network tuner.

The transfer function of the plant in Fig. 8 is as follows:

$$H(s) = \frac{-(1.308)e^{-4.896s}}{(13.515s+1)(6.241s+1)} \quad (3)$$

The results of the Monte Carlo simulation with 100 iterations are reported in Fig. 9. At the end of the pruning process, we lost 43.5% of initial connections and reached a fast network. As can be seen, a controller with a pruned deep recurrent network tuner has superior performance compared to a shallow one. Regarding the stochastic nature of noise, deep networks better compensate for its effects.

Conclusion

This paper suggests a novel method for optimizing deep neural network topology. The weight multiplied by the gradient is employed as the criterion rather than the net weight index. Moreover, the low and high scores of connections are classified to determine the importance of the connections which compete for survival. We evaluated our method using two examples from control literature: neural network modelling of the distillation column and intelligent tuning of PID controller. We discovered that it eliminated over-fitting issues, enhanced learning speed, preserved accuracy, and reduced cost due to a smaller network size. Additionally, we demonstrated that deep neural networks in the right size and setting can handle noisy data more efficiently. This approach's main advantage over the previous ones is that a connection with a low score is not immediately terminated, and it continually assesses the effectiveness of network connections using the proposed criterion. Subsequent studies could focus on how to integrate the developed strategy with other advanced techniques like neural architecture search or automated machine learning.

Author Contributions

A. Ahmadi and R. Mahboobi Esfanjani both developed ideas, interpreted the results, wrote the manuscript, Revised and finalized it.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

DNN Depp Neural Network
MIMO Multi-Input Multi-Output
PID Proportional–Integral–Derivative

References

[1] Z. Allen-Zhu, Y. Li, Y. Liang, "Learning and overparameterized neural networks, going beyond two layers," Adv. Neural Inf. Process. Syst., 32, 2019.

[2] C. Cortes, X. Gonzalvo, V. Kuznetsov, M. Mohri, S. Yang, "AdaNet: Adaptive structural learning of artificial neural networks," in Proc. 34th International Conference on Machine Learning, 70: 874-883, 2017.

[3] O. A. Montesinos L., A. Montesinos L., J. C. Montesinos, Multivariate Statistical Machine Learning Methods for Genomic Prediction, Springer, 2022.

[4] S. Li, T. Hoefler, "Chimera: efficiently training large-scale neural networks with bidirectional pipelines," in Proc. International

Conference for High Performance Computing, Networking, Storage and Analysis, 2021.

[5] V. Sze, Y. H. Chen, T. J. Yang, J. Emer, "Efficient processing of deep neural networks: A tutorial and survey," arXiv preprint arXiv:1703.09039, 2017.

[6] S. A. Janowsky, "Pruning versus clipping in neural networks," Phys. Rev. A, 39(12): 6600-6603, 1989.

[7] G. Chechik, I. Meilijson, E. Ruppin, "Neuronal regulation: a biologically plausible mechanism for efficient synaptic pruning in development," Neurocomputing, 26-27: 633-639, 1999.

[8] M. V. Johnston et al., "Plasticity and injury in the developing brain," Dev. Neurosci., 31: 1-10, 2009.

[9] G. Hinton, O. Vinyals, J. Dean, "Distilling knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.

[10] S. Han, J. Pool, J. Tran, W. J. Dally, "Learning both weights and connections for efficient neural networks," Adv. Neural Inf. Process. Syst., 2015: 1135–1143, 2015.

[11] P. Molchanov et al., "Importance estimation for neural network pruning," arXiv preprint arXiv:1906.10771, 2019.

[12] P. Molchanov et al., "Pruning convolutional neural networks for resource-efficient inference," in Proc. ICLR 2017, 2017.

[13] Y. LeCun et al., "Optimal brain damage," Adv. Neural Inf. Process. Syst., 2, 1990.

[14] B. Hassibi, D. G. Stork, "Second-order derivatives for network pruning: Optimal Brain Surgery," Adv. Neural Inf. Process. Syst., 5: 164-171, 1993.

[15] Y. He et al., "Filter pruning via geometric median for deep convolutional neural network acceleration," in Proc. CVPR 2019: 4340-4349, 2019.

[16] R. Yu et al., "NISP: pruning networks using neuron importance score propagation," in Proc. IEEE CVPR 2018: 9194-9203, 2018.

[17] H. Li et al., "Pruning filters for efficient convolutional neural networks," arXiv preprint arXiv:1608.08710, 2016.

[18] Y. He et al., "Soft filter pruning for accelerating deep convolutional neural networks," arXiv preprint arXiv:1808.06866, 2018.

[19] N. Srivastava et al., "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., 15: 1929-1958, 2014.

[20] Z. Liu et al., "Dropout reduces underfitting," in Proc. 40th International Conference on Machine Learning, 202, 2023.

[21] X. Sun et al., "MeProp: Sparsified back propagation for accelerated deep learning with reduced overfitting," arXiv preprint arXiv:1706.06197, 2017.

[22] S. A. Mirjalili, Evolutionary Algorithms and Neural Networks, Springer, 2019.

[23] M. Kotyrba et al., "The influence of genetic algorithms on learning possibilities of artificial neural networks," Computers, 2022.

[24] F. Zhao et al., "Towards a brain-inspired developmental neural network by adaptive synaptic pruning," in Proc. ICONIP 2017, 2017.

[25] A. Ahmadi, B. Mashoufi, "A new optimized approach for artificial neural network training using genetic algorithms and parallel processing," Int. Rev. Comput. Softw., 7(5): 1828-6003, 2012.

[26] M. Morini et al., "Strategies and tools for studying microglial-mediated synapse elimination and refinement," Front. Immunol., 2021.

[27] R. P. Guidorzi et al., "The range error test in the structural identification of linear multivariable systems," IEEE Trans. Autom. Control, AC-27: 1044-1054, 1982.

[28] A. Saha and S. Patra, "Modeling and control of distillation column using ANFIS," J. Autom.Control, 51: 51-59, 2021.

[29] H. Zhao, J. Zhang, M. Wang, "Modeling and optimization of distillation column processes using neural networks and genetic algorithms," IEEE Access, 7: 76345-76354, 2019.

[30] A. Singh, P. Gupta, R. Verma, "Hybrid ANFIS and neural network-based approach for fault detection in industrial distillation processes," Appl. Soft Comput., 96, 106703.

Biographies

Aghil Ahmadi received B.Sc. degree in Electrical Engineering (Control Systems) from Sahand University of Technology, Tabriz, Iran, in 2004. From 2004 to 2009, he worked as a control systems and instrumentation supervisor engineer for several world-class projects in the oil and gas industries. He received the M.Sc. degree in Electronics Engineering from Urmia University, Urmia, Iran, in 2011. From 2009 to 2011, he was doing research at the High-Performance Computing Research Centre, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran. He has been a Ph.D. researcher at Sahand University of Technology since 2018. His current research interests include machine learning, intelligent control systems, fuzzy systems, and professional project management.

- Email: ag_ahmadi@sut.ac.ir
- ORCID: 0009-0007-6478-1685
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Reza Mahboobi Esfanjani received the B.SC. degree (Hons.) from the Department of Electrical Engineering, Sahand University of Technology, Tabriz, Iran, in 2002, and the M.Sc. and Ph.D. degrees from the Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, in 2004 and 2009, respectively. In 2010, he joined the Sahand University of Technology, where he is currently a Professor in the Department of Electrical Engineering. His current research interests include analysis and control of networked dynamical systems.

- Email: mahboobi@sut.ac.ir
- ORCID: 0000-0002-0341-8141
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
A. Ahmadi, R. Mahboobi Esfanjani, "Structure learning for deep neural networks with competitive synaptic pruning," J. Electr. Comput. Eng. Innovations, 13(1): 189-196, 2025.

DOI: 10.22061/jecei.2024.11017.758

URL: https://jecei.sru.ac.ir/article_2215.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Torque Ripple Reduction by Using Virtual Vectors in Direct Torque Control Method Using Neutral-Point-Clamped Inverter

H. Afsharirad [*]***, S. Misaghi***

Department of Electrical Engineering, Azarbaijan Shahid Madani University, Tabriz, Iran.

Article Info

Article History:
Received 06 August 2024
Reviewed 26 September 2024
Revised 25 October 2024
Accepted 14 November 2024

Keywords:
Multi-level inverter
Neutral-point-clamped inverter
Direct torque control
Permanent magnet synchronous motor
Voltage vector
Virtual vector

Corresponding Author's Email Address:
h.afsharirad@azaruniv.ac.ir

Abstract

Background and Objectives: Due to the high torque ripple and stator current harmonics in direct torque control using a two-level inverter, the use of multi-level inverters has become common to reduce these two factors. Among the multilevel inverters, the Neutral-point-Clamped Inverter has been given more attention in the industry due to its advantages. This inverter has 27 voltage vectors by which torque and flux are controlled. In order to reduce torque ripple and current harmonic as much as possible, methods such as space vector modulation methods or the use of multi-level inverters with higher levels have been considered. But the main drawback of these methods is the increase of complexity and cost.
Methods: In this article, virtual voltage vectors are used to increase the number of hysteresis controller levels. These vectors are obtained from the sum of two voltage vectors. In this way, we will have 12 voltage vectors in addition to the diode clamped inverter's voltage vectors. Therefore, we can increase the number of torque hysteresis levels from 7 levels to 11 levels.
Results: Considering that the proposed method uses virtual vectors and voltage vectors, it does not increase the cost and computational complexity. Also, one of the requirements of using this method is the use of fixed switching frequency, which solves the variable switching frequency problem of conventional methods. Therefore, the proposed control reaches an overall optimization.
Conclusion: To verify the feasibility of the proposed method and compare it with the conventional method, both of these methods are simulated in the MATLAB /Simulink environment and the simulation results represent the efficiency of the proposed control method. This method achieves less torque ripple and harmonic current without increasing the cost and computational complexity.

Introduction

Direct torque control (DTC) is one of the practical motor drive system controllers. This controller advantages include no need for other reference frame transforms, very simple structure, high transient response and its robustness. This controller directly controls flux and torque with a hysteresis controller, choosing best voltage vector in each area to reach and control flux and torque [1]-[3] However, DTC has disadvantages like high torque and flux ripple [4], [5].

One method to reduce torque ripple and current harmonics is using multilevel inverters. Multilevel inverters, besides mentioned advantages, benefit from reduced switching frequency, reduced switching losses, and reduced voltage stress on switches [6]-[8]. Multilevel inverters classify into several categories based on structure, with most important being Neutral-Point-Clamped Inverter (NPCI), flying capacitor inverters, H-

bridge inverters, and cascaded inverters.

NPCIs widely used in industry due to advantages. Its advantages include easier bidirectional power transfer capability, use at all voltage levels, and simpler structure [9]-[12]. Various methods propose to further improve these inverters' performance in reducing torque ripple. These methods include using modulation techniques in DTC, using virtual vectors, or using other control methods like model predictive controller [13]-[17]. In [18], a NPCI controls a five-phase induction motor. This paper uses a modulation method calculating switching time in switching period using two voltage vectors for switching, reducing transient states. In [19], a NPCI controls a permanent magnet synchronous motor (PMSM), where duty cycle calculation reduces torque and flux ripple. In [20], PWM uses for induction motor control, improving motor performance. However, a 60-pulse converter supplies the inverter for improved power quality. In [21], a NPCI uses for PMSM control, where duty cycle calculates for inverter switching. This method improves motor performance, reducing torque and flux ripple. In [22], switching time of two voltage vectors in switching interval adjusts, and a torque regulator replaces hysteresis controller, increasing control system efficiency. In [23], a NPCI uses for PMSM drives, employing two voltage vectors in duty cycle. Duty cycle calculation method minimally depends on motor parameters. These control methods with PMSM enhance control system advantages and stability. PMSM, due to their benefits, are main rivals to induction motors. Among these benefits are high power density, compact size, low copper losses from lacking rotor windings, and simple structure [24]-[29].

Conventional methods typically employ either direct torque control based on lookup tables or duty cycle calculation with multiple voltage vectors applied within a duty cycle period. In lookup table-based approaches, switching frequency varies, and some methods don't utilize the inverter's full capacity, using medium-amplitude voltage vectors to reduce torque ripple even when larger vectors are needed. For instance, paper [30], an example of a conventional duty cycle-based method, uses medium voltage vectors in scenarios requiring large voltage vectors. The proposed method in this paper uses fixed switching frequency for direct torque control, resolving the variable switching frequency issue.

Conversely, duty cycle calculation methods often use virtual voltage vectors without considering the sufficiency of the inverter's voltage vector. For example, reference [21] employs three voltage vectors for motor control in each duty cycle period, which is unnecessary in some intervals. The proposed method not only considers the number of voltage vectors in hysteresis bands but also maximizes inverter capability, using only one voltage vector in these bands. Additionally, in hysteresis bands with low torque ripple amplitude, an appropriate voltage vector is applied, and in some intervals, two voltage vectors are used for better torque variation response.

To prevent significant increases in switching frequency, the proposed method considers voltage vectors of previous and subsequent levels. When using a virtual voltage vector, only one voltage vector is employed to generate it. This approach effectively addresses the limitations of conventional methods while optimizing inverter performance and torque control.

The paper innovates by increasing hysteresis controller levels without adding switches to enhance voltage vectors. Traditionally, multilevel inverters reduce torque ripple in direct torque control of two-level inverters by increasing voltage vectors. A two-level inverter has 8 voltage vectors, while a diode-clamped inverter has 27, each with unique amplitudes and angles.

This method, instead of increasing voltage levels, sums voltage vectors to create new vectors with different amplitudes and angles from the main ones, applied based on torque hysteresis levels. This approach increases voltage vectors to 39, achieving more vectors and hysteresis levels without additional switches, thus reducing torque ripple and current harmonics.

The method necessitates fixed switching frequency, an advantage over conventional variable-frequency hysteresis controllers. This fixed frequency approach marks a significant improvement from traditional techniques. By combining vector summation and fixed frequency, the paper presents a novel solution for enhancing inverter performance without increasing system complexity, addressing key limitations in existing direct torque control methods.

Different sections here are: Section 2 discuss about technical work preparation. Section 3 presents simulation results, and Section 4 provides conclusions.

Technical Work Preparation

A. The Conventional Direct Torque Control Method Using a NPCI

The NPCI, due to advantages, is one of industry's most widely used inverters. Fig. 1 shows three-phase three-level inverter. As seen in this figure, each inverter leg has 4 power electronic switches and 2 diodes, with each power electronic switch connected to DC link through a diode. Each switch in this inverter represents by symbol Skfn, where k corresponds to phase number {A, B, C}, f indicates switch position {p for positive, N for negative}, and n specifies switch number.

Table 1 illustrates switching pattern for one inverter leg. As inferred from this table, if ‘Sap1’ and ‘Sap2’ turn on, voltage +vdc/2 produces at inverter output. If switch ’Sap1’ and ‘San1’ on, inverter output voltage is 0. If ‘San1’ and ‘San2’ turn on, voltage -vdc/2 produces at inverter output.

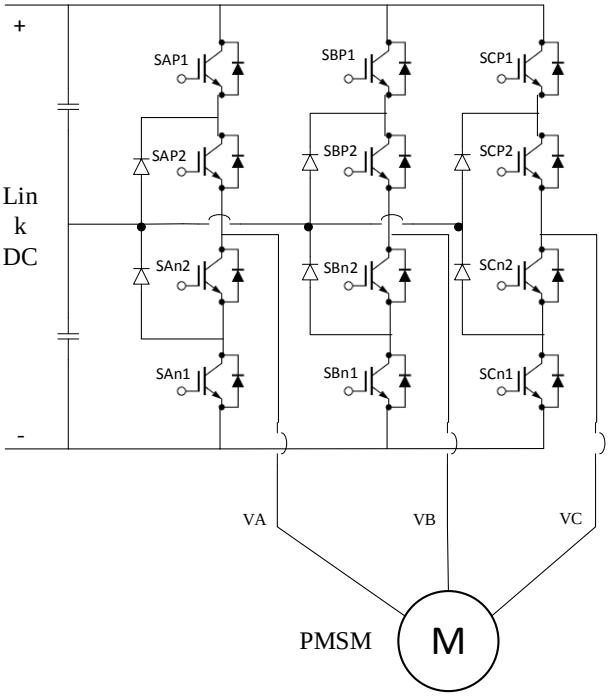


Fig. 1: Three-phase, three-level diode clamped inverter.

Table1: Switching table of NPCI

SAp1	SAp2	SAn1	SAn2	Phase Voltage
On	On	Off	Off	*Vdc*/2
Off	On	On	Off	0
Off	Off	On	On	-*Vdc*/2

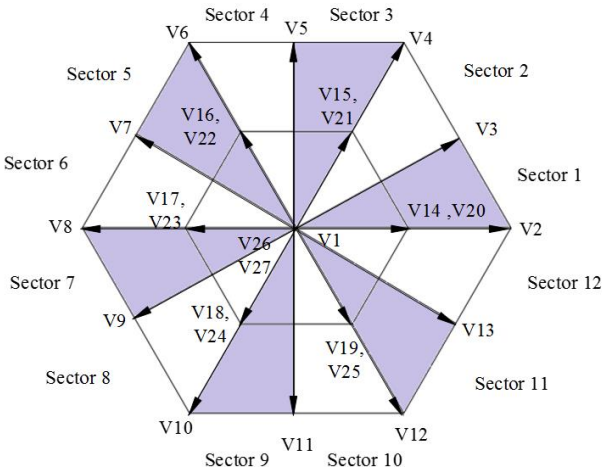


Fig. 2: Distribution of voltage vectors in the αβ plane.

Table 2 shows voltage vectors of three-phase three-level diode-clamped inverter. This table displays 27 voltage vectors resulting from this inverter. State '1' corresponds to respective switch being turned on, while state '0' corresponds to switch being turned off. Additionally, Fig. 2 illustrates distribution of voltage vectors in αβ plane.

Table 2: Diode clamped inverter's voltage vectors

Switches of inverter						VV	$\lvert\overline{V_s}\rvert$	angle
S_{Ap1}	S_{Ap2}	S_{Bp1}	S_{Bp2}	S_{Cp1}	S_{Cp2}			
0	0	0	0	0	0	$\overline{V_1}$	0	-
1	1	0	0	0	0	$\overline{V_2}$	$Vdc\sqrt{2/3}$	0
1	1	0	1	0	0	$\overline{V_3}$	$Vdc\sqrt{1/2}$	$\pi/6$
1	1	1	1	0	0	$\overline{V_4}$	$Vdc\sqrt{2/3}$	$\pi/3$
0	1	1	1	0	0	$\overline{V_5}$	$Vdc\sqrt{1/2}$	$\pi/2$
0	0	1	1	0	1	$\overline{V_6}$	$Vdc\sqrt{2/3}$	$2\pi/3$
0	0	1	1	0	1	$\overline{V_7}$	$Vdc\sqrt{1/2}$	$5\pi/6$
0	0	1	1	1	1	$\overline{V_8}$	$Vdc\sqrt{2/3}$	π
0	0	0	0	1	1	$\overline{V_9}$	$Vdc\sqrt{1/2}$	$7\pi/6$
0	0	0	0	1	1	$\overline{V_{10}}$	$Vdc\sqrt{2/3}$	$4\pi/3$
0	1	0	0	1	1	$\overline{V_{11}}$	$Vdc\sqrt{1/2}$	$3\pi/2$
1	1	0	0	1	1	$\overline{V_{12}}$	$Vdc\sqrt{2/3}$	$5\pi/3$
1	1	0	0	0	1	$\overline{V_{13}}$	$Vdc\sqrt{1/2}$	$11\pi/6$
1	1	0	1	0	1	$\overline{V_{14}}$	$Vdc\sqrt{1/6}$	0
1	1	1	1	0	1	$\overline{V_{15}}$	$Vdc\sqrt{1/6}$	$\pi/3$
0	1	1	1	0	1	$\overline{V_{16}}$	$Vdc\sqrt{1/6}$	$2\pi/3$
0	1	1	1	1	1	$\overline{V_{17}}$	$Vdc\sqrt{1/6}$	π
0	1	0	1	1	1	$\overline{V_{18}}$	$Vdc\sqrt{1/6}$	$4\pi/3$
1	1	0	1	1	1	$\overline{V_{19}}$	$Vdc\sqrt{1/6}$	$5\pi/3$
0	1	0	0	0	0	$\overline{V_{20}}$	$Vdc\sqrt{1/6}$	0
0	1	0	1	0	0	$\overline{V_{21}}$	$Vdc\sqrt{1/6}$	$\pi/3$
0	0	0	1	0	1	$\overline{V_{22}}$	$Vdc\sqrt{1/6}$	$2\pi/3$
0	0	0	1	0	1	$\overline{V_{23}}$	$Vdc\sqrt{1/6}$	π
0	0	0	0	0	1	$\overline{V_{24}}$	$Vdc\sqrt{1/6}$	$4\pi/3$
0	1	0	0	0	1	$\overline{V_{25}}$	$Vdc\sqrt{1/6}$	$5\pi/3$
0	1	0	1	0	1	$\overline{V_{26}}$	0	-
1	1	1	1	1	1	$\overline{V_{27}}$	0	-

In direct torque control, two hysteresis controllers are used to control the torque and flux, where the hysteresis levels are determined based on the available voltage vectors for control objectives. In the three-level diode-clamped inverter, there are 27 voltage vectors, and the conventional method utilizes 20 voltage vectors. To understand the operation of direct torque control, consider the (1) and (2). Equation (1) represents the torque relationship in a PMSM, and (2) is the derivative of (1) with respect to time. Equation (2) shows how torque variations affect the load angle and, consequently, how they influence the flux.

$$T_e = \frac{3P\lambda_s}{4L_dL_q}[2\lambda_m L_q \sin\delta + \lambda_s(L_d - L_q)\sin 2\delta] \quad (1)$$

$$\frac{dT_e}{dt} = \frac{3P\lambda_s}{4L_dL_q}[2\lambda_m L_q \cos\delta + \lambda_s(L_d - L_q)\cos 2\delta]\dot{\delta} \quad (2)$$

In this relationship, P represents the number of pole pairs, λ_s is the stator flux, Ld and Lq are direct- and the quadrature-axis inductances, λ_m magnitude of rotor flux linkages and δ is the angle between these two flux linkage vectors and its name is load angle.

These variations are achieved by changing the voltage vector. To further clarify this, consider Fig. 3. As seen in this figure, the voltage vector affects the flux value, load angle and consequently, the torque. Therefore, by utilizing different voltage vectors, various control states can be achieved.

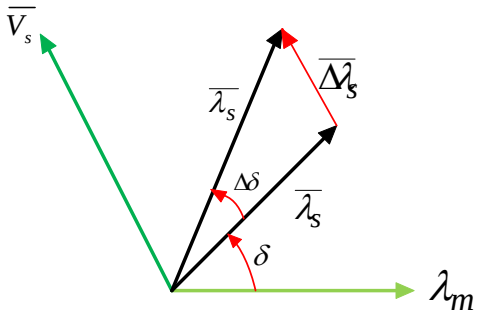


Fig. 3: Load angle and torque adjustment by voltage vectors.

According to Fig. 3, for example, it can be said that vector V4 causes torque control at its highest level, increasing both torque and flux, while vector V10 decreases the flux and controls the torque at its lowest level.

Therefore, the direct torque control table using the conventional method can be represented as shown in Table 3.

The hysteresis loops for torque and flux are given by (3) and (4).

$$\begin{cases} +3 \text{ for } 3\Delta T_e < T_{e_{error}} \\ +2 \text{ for } 2\Delta T_e \le T_{e_{error}} \le 3\Delta T_e \\ +1 \text{ for } \Delta T_e \le T_{e_{error}} \le 2\Delta T_e \\ 0 \text{ for } -\Delta T_e \le T_{e_{error}} \le \Delta T_e \\ -1 \text{ for } -2\Delta T_e \le T_{e_{error}} \le -\Delta T_e \\ -2 \text{ for } -3\Delta T_e \le T_{e_{error}} \le -2\Delta T_e \\ -3 \text{ for } T_{e_{error}} < -3\Delta T_e \end{cases} \quad (3)$$

$$\begin{cases} \varphi = 1 \text{ if } \lambda_{s_{error}} \ge \Delta\lambda/2 \\ \varphi = 0 \text{ if } \lambda_{s_{error}} < -\Delta\lambda/2 \end{cases} \quad (4)$$

B. The Proposed Method for DTC Using Voltage Vector and Virtual Voltage Vector

To further reduce the torque ripple, the number of hysteresis levels must increase.

To increase the number of hysteresis levels, the number of available voltage vectors for control in the desired regions must increase.

Table 3: Conventional look-up table of DTC

sector		1	2	3	4	5	6	7	8	9	10	11	12
λ	T												
1	+3	V4	V4	V6	V6	V8	V8	V10	V10	V12	V12	V2	V2
	+2	V3	V5	V5	V7	V7	V9	V9	V11	V11	V13	V13	V3
	+1	V15	V15	V16	V16	V17	V17	V18	V18	V19	V19	V14	V14
	0	V27	V1	V27	V1	V27	V1	V27	V1	V27	V1	V27	V1
	-1	V19	V19	V14	V14	V15	V15	V16	V16	V17	V17	V18	V18
	-2	V11	V13	V13	V3	V3	V5	V5	V7	V7	V9	V9	V11
	-3	V12	V12	V2	V2	V4	V4	V6	V6	V8	V8	V10	V10
0	+3	V6	V6	V8	V8	V10	V10	V12	V12	V2	V2	V4	V4
	+2	V5	V7	V7	V9	V9	V11	V11	V13	V13	V3	V3	V5
	+1	V16	V16	V17	V17	V18	V18	V19	V19	V14	V14	V15	V15
	0	V1	V27	V1	V27	V1	V27	V1	V27	V1	V27	V1	V27
	-1	V18	V18	V19	V19	V14	V14	V15	V15	V16	V16	V17	V17
	-2	V9	V11	V11	V13	V13	V3	V3	V5	V5	V7	V7	V9
	-3	V10	V10	V12	V12	V2	V2	V4	V4	V6	V6	V8	V8

In this paper, virtual voltages are used to increase the number of voltage vectors, such that a fixed duty cycle is defined for this controller, and in half the interval one voltage vector is used and in the other half duty cycle another voltage vector is used. To obtain the virtual vectors, the average vectors and half vectors are used, because the sum of the complete vectors has a value equal to the average vector and is collinear with it. The sum of the virtual vectors is obtained as follows:

$$\frac{1}{2}V_3+\frac{1}{2}V_5=0.61V_s\angle\frac{\pi}{3} \tag{5}$$

The virtual voltage vectors are obtained as shown in Table 4. Fig. 4 shows the distribution of these vectors in the αβ plane. Therefore, 12 voltage vectors are added to the control voltage vectors, and using these vectors, the hysteresis levels can be increased and an 11-level lookup table can be provided. The lookup table for switching the NPCI is presented as Table 5. The hysteresis levels in the proposed method are defined by (6) and (7).

$$\begin{cases}\varphi=1 \text{ if } \lambda_{s_{error}} \geq \Delta\lambda/2\\ \varphi=0 \text{ if } \lambda_{s_{error}} < -\Delta\lambda/2\end{cases} \tag{6}$$

$$\begin{cases}+5 \text{ for } T_{e_{error}} \geq 5\Delta T_e\\ +4 \text{ for } 4\Delta T_e \leq T_{e_{error}} \leq 5\Delta T_e\\ +3 \text{ for } 3\Delta T_e \leq T_{e_{error}} \leq 4\Delta T_e\\ +2 \text{ for } 2\Delta T_e \leq T_{e_{error}} \leq 3\Delta T_e\\ +1 \text{ for } \Delta T_e \leq T_{e_{error}} \leq 2\Delta T_e\\ 0 \text{ for } -\Delta T_e \leq T_{e_{error}} \leq \Delta T_e\\ -1 \text{ for } -2\Delta T_e \leq T_{e_{error}} \leq -\Delta T_e\\ -2 \text{ for } -3\Delta T_e \leq T_{e_{error}} \leq -2\Delta T_e\\ -3 \text{ for } -4\Delta T_e \leq T_{e_{error}} \leq -3\Delta T_e\\ -4 \text{ for } -5\Delta T_e \leq T_{e_{error}} \leq -4\Delta T_e\\ -5 \text{ for } T_{e_{error}} \leq -5\Delta T_e\end{cases} \tag{7}$$

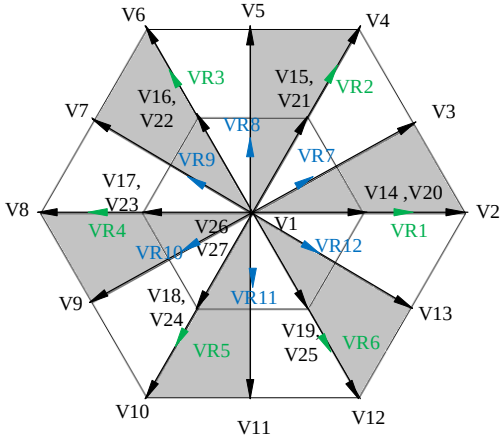


Fig. 4: Distribution of voltage vectors and virtual voltage vectors in the αβ plane.

DTC block diagram is shown in Fig. 5. In the next section, the simulation results will be examined in detail.

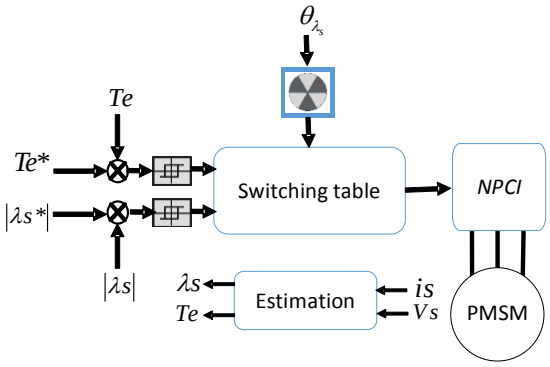


Fig. 5: Block diagram of DTC with proposed look-up table.

Table 4: Virtual vectors synthesis table

Virtual vector	Virtual vector	Sum of two voltage vector
VR1	V_3 , V_{13}	$0.61V_s\angle0$
VR2	V_3 , V_5	$0.61V_s\angle\frac{\pi}{3}$
VR3	V_5 , V_7	$0.61V_s\angle2\frac{\pi}{3}$
VR4	V_7 , V_9	$0.61V_s\angle\pi$
VR5	V_9 , V_{11}	$0.61V_s\angle-2\frac{\pi}{3}$
VR6	V_{11} , V_{13}	$0.61V_s\angle-\frac{\pi}{3}$
VR7	V20 , V21	$0.35V_s\angle\frac{\pi}{6}$
VR8	V21 , V22	$0.35V_s\angle\frac{\pi}{2}$
VR9	V22 , V23	$0.35V_s\angle5\frac{\pi}{6}$
VR10	V23, V24	$0.35V_s\angle7\frac{\pi}{6}$
VR11	V24 , V25	$0.35V_s\angle3\frac{\pi}{2}$
VR12	V25 , V20	$0.35V_s\angle11\frac{\pi}{6}$

C. *Smooth Vector Switching and Neutral-Point Voltage Balancing Control*

According to paper [30], whenever voltage vectors of large and small magnitudes are applied to the motor, neutral point oscillations occur. To prevent these oscillations, switching should be limited to adjacent vectors, adhering to the smooth voltage vector switching criterion. The hysteresis controller for flux control is a

two-level controller that oscillates between +1 and -1. Therefore, if the torque is continuously changing within the positive hysteresis band, for example, if the flux vector is also in the first region, according to the voltage vector switching table, the voltage vectors are adjacent to each other, and the voltage vector transition is from one vector to its adjacent vector. This approach ensures that the switching occurs between neighboring vectors, maintaining the smooth voltage vector switching principle and minimizing neutral point oscillations.

Furthermore, paper [26] states that the charge balance of DC link capacitors depends on the balance of the neutral point voltage. Therefore, if the neutral point voltage is zero, according to the following equations, the DC link capacitors are charged and discharged in a balanced manner:

$$i_o = i_{c1} - i_{c2} \tag{8}$$

$$\begin{cases} i_{c1} = c_1 \dfrac{dv_{c1}}{dt} \\ i_{c2} = c_2 \dfrac{dv_{c2}}{dt} \end{cases} \tag{9}$$

$$\begin{cases} V_{c1} = \dfrac{V_{dc}}{2} - v_o \\ V_{c2} = \dfrac{V_{dc}}{2} + v_o \end{cases} \tag{10}$$

$$i_o = -2C \frac{dv_o}{dt} \tag{11}$$

$$v_o = \frac{-1}{2C} \int i_o dt \tag{12}$$

Therefore, if the neutral point voltage equals zero, the neutral point balance is maintained, and the DC link capacitors are charged in a balanced manner.

Table 4: proposed look-up table of DTC

sector		1	2	3	4	5	6	7	8	9	10	11	12
λ	T												
	+5	V4	V4	V6	V6	V8	V8	V10	V10	V12	V12	V2	V2
	+4	V3	V5	V5	V7	V7	V9	V9	V11	V11	V13	V13	V3
	+3	V3+V5	V3+V5	V5+V7	V7+V5	V7+V9	V7+V9	V9+V11	V9+V11	V11+V13	V11+V13	V13+V3	V13+V3
	+2	V15	V15	V16	V16	V17	V17	V18	V18	V19	V19	V14	V14
	+1	V20+V21	V21+V22	V21+V22	V22+23	V22+23	V23+V24	V23+V24	V24+V25	V24+V25	V25+V20	V25+V20	V20+V21
1	0	V27	V1	V27	V1	V27	V1	V27	V1	V27	V1	V27	V1
	-1	V24+V25	V25+V20	V25+V20	V21+V20	V21+V20	V21+V22	V21+V22	V22+V23	V22+V23	V23+V24	V23+V24	V24+V25
	-2	V19	V19	V14	V14	V15	V15	V16	V16	V17	V17	V18	V18
	-3	V11+V13	V11+V13	V3+V13	V3+V13	V3+V5	V3+V5	V5+V7	V5+V7	V7+V9	V7+V9	V9+V11	V9+V11
	-4	V11	V13	V13	V3	V3	V5	V5	V7	V7	V9	V9	V11
	-5	V12	V12	V2	V2	V4	V4	V6	V6	V8	V8	V10	V10
	+5	V6	V6	V8	V8	V10	V10	V12	V12	V2	V2	V4	V4
	+4	V5	V7	V7	V9	V9	V11	V11	V13	V13	V3	V3	V5
	+3	V5+V7	V5+V7	V7+V9	V7+V9	V9+V11	V9+V11	V13+V11	V13+V11	V13+V3	V13+V3	V3+V5	V3+V5
	+2	V16	V16	V17	V17	V18	V18	V19	V19	V14	V14	V15	V15
	+1	V21+V22	V23+V22	V23+V22	V23+V24	V23+V24	V24+V25	V24+V25	V25+V20	V25+V20	V20+V21	V20+V21	V21+V22
0	0	V1	V27	V1	V27	V1	V27	V1	V27	V1	V27	V1	V27
	-1	V23+V24	V24+V25	V24+V25	V20+V25	V20+V25	V21+V20	V21+V20	V21+V22	V21+V22	V22+V23	V22+V23	V23+V24
	-2	V18	V18	V19	V19	V14	V14	V15	V15	V16	V16	V17	V17
	-3	V9+V11	V9+V11	V11+V13	V11+V13	V13+V3	V13+V3	V5+V3	V5+V3	V5+V7	V5+V7	V7+V9	V7+V9
	-4	V9	V11	V11	V13	V13	V3	V3	V5	V5	V7	V7	V9
	-5	V10	V10	V12	V12	V2	V2	V4	V4	V6	V6	V8	V8

Results and Discussion

To evaluate the feasibility of the proposed method, the conventional method and the proposed method were simulated in the MATLAB/Simulink environment, and the results are discussed in detail in this section. The parameters of the studied PMSM are observable in Table 6. It should be noted that both methods under study were simulated under the same conditions and with identical

motors. In this section, first, the steady-state results are studied, and then experiments entitled transient and no-load experiments are performed.

In the transient experiment, different speeds and loads were applied to the motor. In the no-load experiment, the motor started its operation from the unloaded condition, and then load variations were applied at a constant speed.

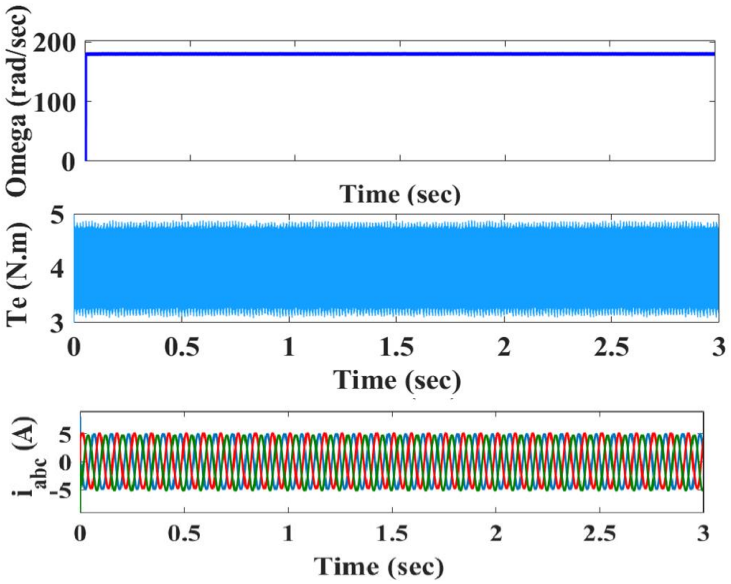
Fig. 6 shows the steady-state experiment. In this experiment, the motor was running at 180 rad/sec, and a torque of 4 N.m was applied to it.

Fig. 6(a) corresponds to the conventional method, and Fig. 6(b) corresponds to the proposed method. As observed, in the proposed method, the torque ripple is significantly lower than the conventional method, and the steady-state responses are as good as the conventional method.

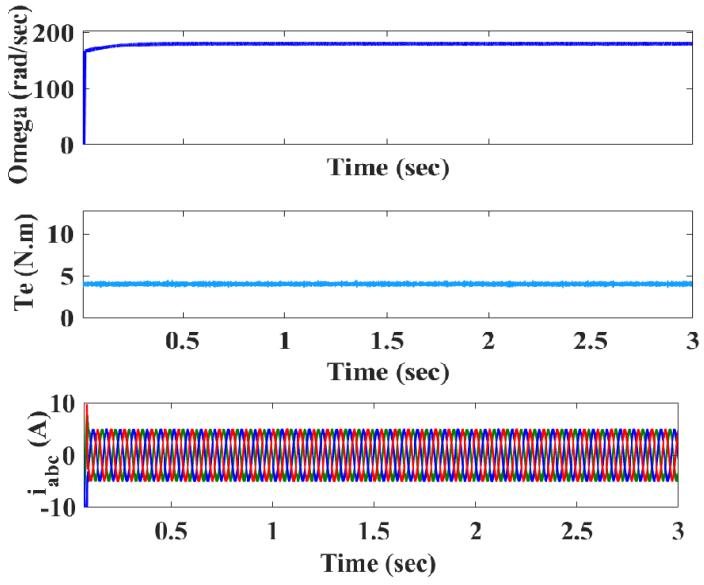
Table 6: parameters of PMSM

Pole pair: Pn	4
Stator resistance: R_s	0.57Ω
d-axis inductance	8.72 mH
q-axis inductance	28.8 mH
Magnet flux linkage: $\boldsymbol{\lambda_m}$	0.108 wb
Vdc	310v

Fig. 7 shows the FFT analysis of the stator current, where Fig. 7(a) corresponds to the conventional method, and Fig. 7(b) corresponds to the proposed method. As observed, the stator current harmonics are significantly reduced compared to the conventional method, which is one of the advantages of the proposed method.

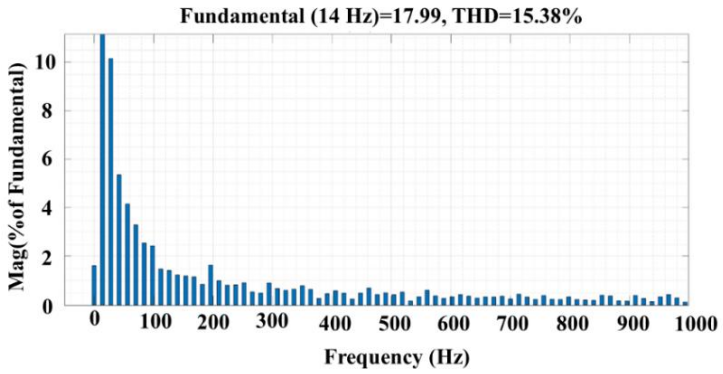


(a) Conventional methode

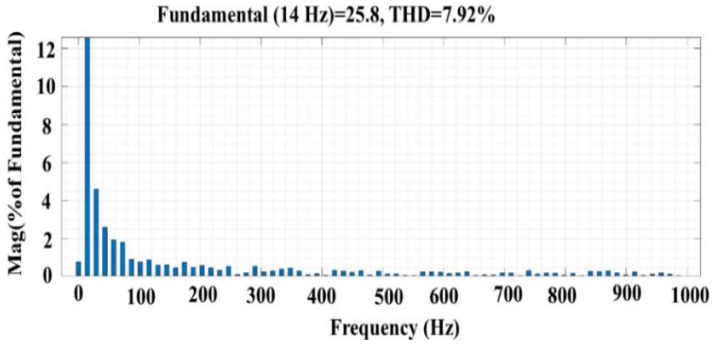


(b) Proposed methode

Fig. 6: Steady state performance. From top: speed, Torque & stator current.



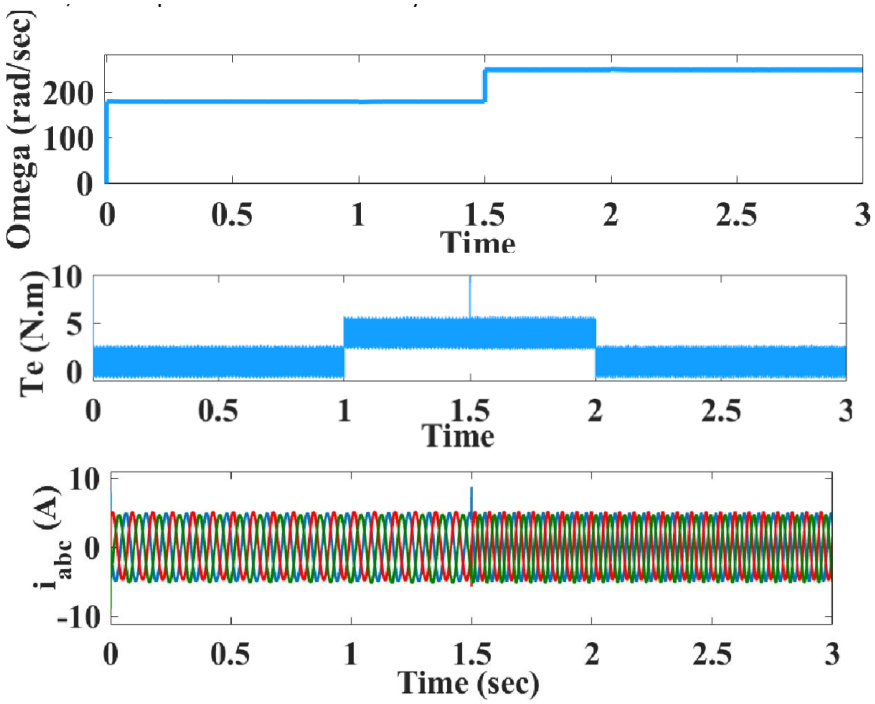
(a) Conventional methode



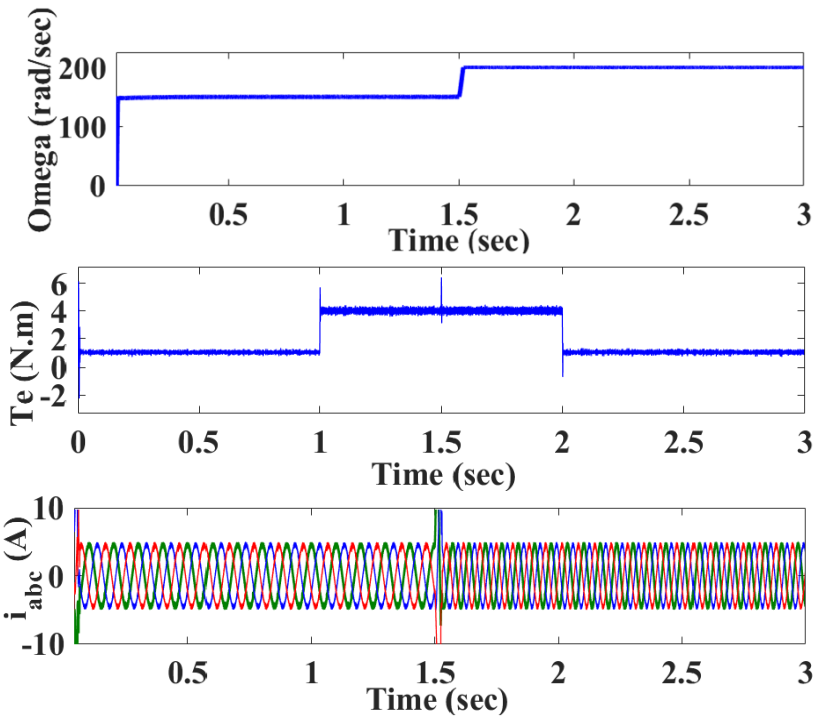
(b) Proposed methode

Fig. 7: FFT results of stator current.

Fig. 8 displays the motor experiment under transient conditions. In this experiment, the initial speed value was 188 rad/sec, and then it increased to 250 rad/sec in 1.5 sec. In this experiment, the torque value was also initially 1 N.m, which increased to 4 N.m in 1 sec and then decreased to 1 N.m in 2 sec. As observed, the transient results of the proposed method are as good as the conventional method.



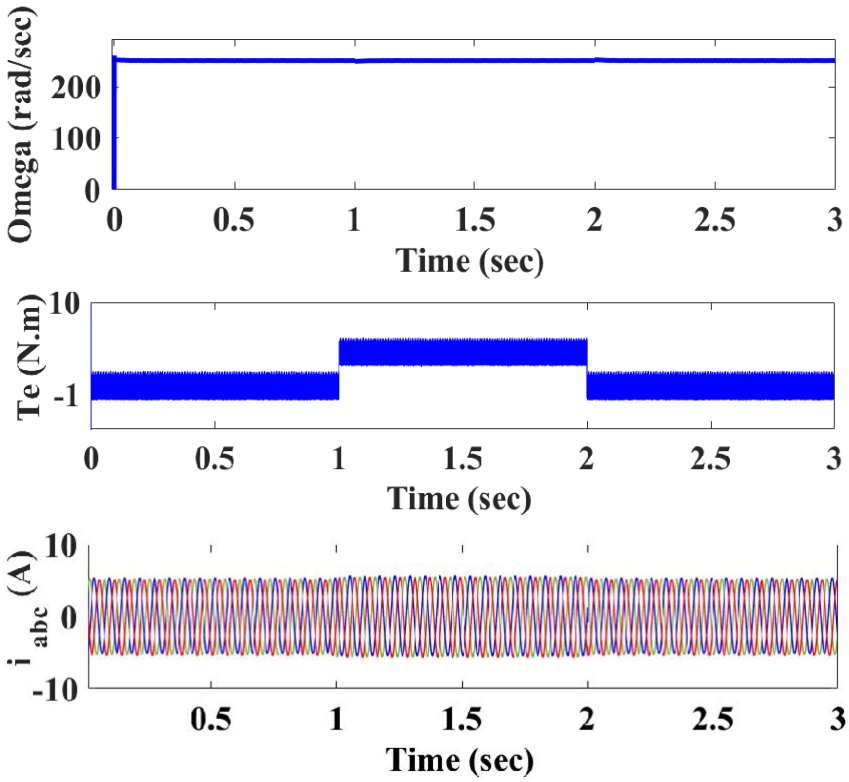
(a) Conventional methode



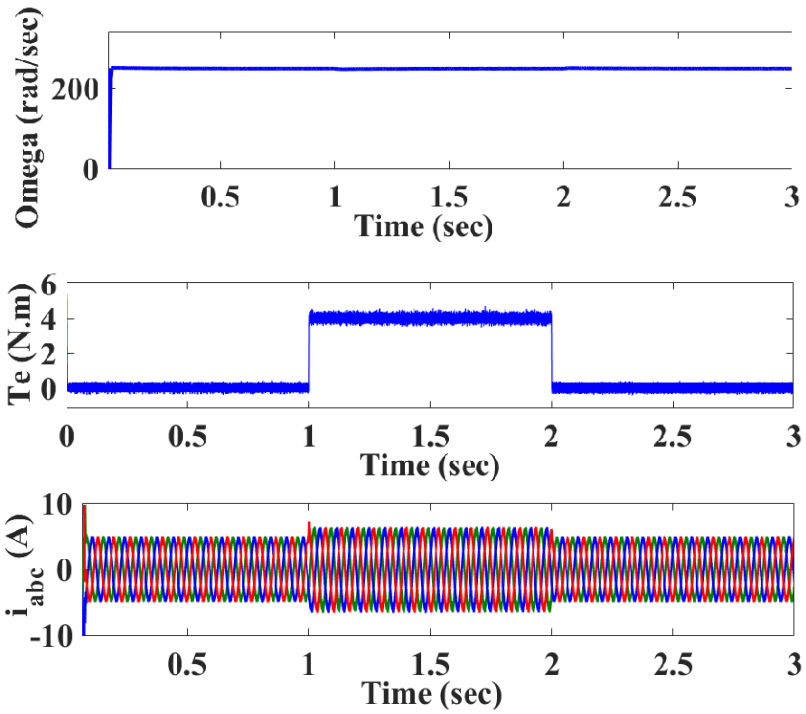
(b) Proposed methode

Fig. 8: Dynamic performance. From top: speed, Torque & stator current.

Fig. 9 shows the no-load test of the motor in simulation. In this test, the motor speed remains constant at 250 rad/sec, and the load torque increases from 0 N.m to 4 N.m and then returns to 0 N.m. In this experiment, it is observed that under load torque changes, the motor maintains its stability and follows the reference values.



(a) Conventional methode



(b) Proposed methode

Fig. 9: Load disturbance. From top: speed, Torque & stator current.

Conclusion

In this paper, the DTC method is used for controlling a PMSM. In this controller, the motor is powered by a NPCI. NPCIs have 27 voltage vectors, but in the conventional method, only 20 voltage vectors are used for switching, and a maximum of 7 hysteresis levels can be defined. In the proposed method, virtual vectors are formed to increase the number of levels to reduce torque ripple and current harmonics. By forming virtual vectors, the problem of variable switching in the conventional method is solved, and the hysteresis levels are increased to 11 levels. The constant switching frequency leads to reduced losses and increased system efficiency. In the proposed method, the amount of torque ripple has decreased significantly, and current harmonics have also decreased. Another advantage of this method compared to modulation methods is the simplicity of the controller. Therefore, the proposed system provides an overall improvement over conventional method.

Author Contributions

This article is the result of a course project. H. Afsharirad provided critical supervision and expert guidance throughout the project, overseeing the research direction and ensuring the technical accuracy of the work. S.Misaghi was responsible for performing the simulations and drafting the initial manuscript. Both H.Afsharirad and S.Misaghi collaborated on analyzing the simulation results. The final manuscript was refined under the close supervision of H.Afsharirad.

Acknowledgment

The authors would like to thank the editor and anonymous reviewers.

Conflict of interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

NPCI	Neutral-Point-Clamped Inverter
DTC	Direct Torque Control
PMSM	Permanent magnet synchronous motor
PWM	Pulse Width Modulation

Reference

[1] J. Hyoung Ryu, K. W. Lee, J. S. Lee, "A unified flux and torque control method for DTC-based induction-motor drives," IEEE Trans. Power Electron., 21(1): 234-242, 2006.

[2] X. Chen, Z. Zhang, L. Yu, Z. Bian, "An improved direct instantaneous torque control of doubly salient electromagnetic machine for torque ripple reduction," IEEE Trans. Ind. Electron., 68(8): 6481-6492, 2021.

[3] P. Naganathan, S. Srinivas, "Direct torque control techniques of three-level h-bridge inverter fed induction motor for torque ripple

reduction at low-speed operations," IEEE Trans. Ind. Electron., 67(10): 8262-8270, 2020.

[4] K. B. Lee, S. H. Huh, J. Y. Yoo, F. Blaabjerg, "Performance improvement of DTC for induction motor-fed by three-level inverter with an uncertainty observer using RBFN," IEEE Trans. Energy Convers., 20(2): 276-283, 2005.

[5] E. P. Sarika, R. S. P. Raj, "Performance comparison of direct torque control of two level and three level neutral point clamped inverter fed three phase induction motor," in Proc. 2014 International Conference on Advances in Green Energy (ICAGE): 179-183, 2014.

[6] K. Tian, B. Wu, M. Narimani, D. Xu, Z. Cheng, N. Reza Zargari, "A capacitor voltage-balancing method for Nested Neutral Point Clamped (NNPC) inverter," IEEE Trans. Power Electron., 31(3): 2575-2583, 2016..

[7] S. E. Daoudi, L. Lazrak, M. A. Lafkih, "Sliding mode approach applied to sensorless direct torque control of cage asynchronous motor via multi-level inverter," Prot. Control Mod. Power Syst., 5(2): 1-10, 2020.

[8] G. H. B. Foo, T. Ngo, X. Zhang, M. F. Rahman, "SVM direct torque and flux control of three-level simplified neutral point clamped inverter fed interior PM synchronous motor drives," IEEE/ASME Trans. Mech., 24(3): 1376-1385, 2019.

[9] O. Chandra Sekhar, K. Chandra sekhar, "A novel five-level inverter topology for DTC induction motor drive," in Proc. 2012 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT): 392-396, 2012.

[10] Z. Wang, X. Wang, J. Cao, M. Cheng, Y. Hu, "Direct torque control of T-NPC inverters-Fed double-stator-winding PMSM drives with SVM," IEEE Trans. Power Electron., 33(2): 1541-1553, 2018.

[11] P. Naganathan, S. Srinivas, "Direct torque control techniques of three-level h-bridge inverter fed induction motor for torque ripple reduction at low-speed operations," IEEE Trans. Ind. Electron., 67(10): 8262-8270, 2020.

[12] N. Babu A, P. Agarwal, "Nearest and non-nearest three vector modulations of NPCI using two-level space vector diagram—a novel approach," IEEE Trans. Ind. Appl., 54(3): 2400-2415, 2018.

[13] F. Faraji, A. A. M. Birjandi, S. M. Mousavi G, J. Zhang, B. Wang, X. Guo, "An improved multilevel inverter for single-phase transformerless PV system," IEEE Trans. Energy Convers., 36(1): 281-290, 2021.

[14] X. Wang, Z. Wang, M. Cheng, Y. Hu, "Remedial strategies of T-NPC three-level asymmetric six-phase PMSM drives based on SVM-DTC," IEEE Trans. Ind. Electron., 64(9): 6841-6853, 2017.

[15] S. Kouro et al., "Recent advances and industrial applications of multilevel converters," IEEE Trans. Ind. Electron., 57(8): 2553-2580, 2010.

[16] T. Geyer, S. Mastellone, "Model predictive direct torque control of a five-level anpc converter drive system," IEEE Trans. Ind. Appl., 48(5): 1565-1575, 2012.

[17] Z. Wang, X. Wang, J. Cao, M. Cheng, Y. Hu, "Direct torque control of T-NPC inverters-fed double-stator-winding PMSM drives with SVM," IEEE Trans. Power Electron., 33(2): 1541-1553, 2018.

[18] S. Payami, R. K. Behera, A. Iqbal, "DTC of three-level NPC inverter fed five-phase induction motor drive with novel neutral point voltage balancing scheme," IEEE Trans. Power Electron., 33(2): 1487-1500, 2018.

[19] D. Mohan, X. Zhang, G. H. B. Foo, "A simple duty cycle control strategy to reduce torque ripples and improve low-speed performance of a three-level inverter fed DTC IPMSM drive," IEEE Trans. Ind. Electron., 64(4): 2709-2721, 2017.

[20] P. Kant, B. Singh, "A sensorless DTC scheme for 60-pulse AC-DC converter fed 5-level six-leg NPC inverter based medium voltage induction motor drive," IEEE Trans. Energy Convers., 35(4): 1916-1925, 2020.

[21] A. Ramprasad, D. Giribabu, S. K. Kakodia, A. K. Panda, "Performance analysis of three-level NPC inverter Fed PMSM drives," in Proc. 2022 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS): 1-6, 2022.

[22] D. Mohan, X. Zhang, G. H. B. Foo, "Three-level inverter-fed direct torque control of IPMSM with torque and capacitor voltage ripple reduction," IEEE Trans. Energy Convers., 31(4): 1559-1569, 2016.

[23] I. M. Alsofyani, K. B. Lee, "Improved predictive torque control with unidirectional voltage vector selection of PMSM fed by three-level neutral-point-clamped inverter," in Proc. 2023 IEEE International Symposium on Sensorless Control for Electrical Drives (SLED): 1-6, 2023.

[24] X. Jiang, Y. Wang, J. Dong, "Speed regulation method using genetic algorithm for dual three-phase permanent magnet synchronous motors," CES Trans. Electr. Mach. Syst., 7(2): 171-178, 2023.

[25] J. Shen, X. Wang, D. Xiao, Z. Wang, Y. Mao, M. He, "Online switching strategy between dual three-phase PMSM and open-winding PMSM," IEEE Trans. Transp. Electrif., 10(1): 1519-1529, 2024.

[26] O. Sandre Hernandez, J. Rangel Magdaleno, R. Morales Caporal, E. Bonilla Huerta, "HIL simulation of the DTC for a three-level inverter fed a PMSM with neutral-point balancing control based on FPGA," Electr. Eng., 100: 1441-1454, 2018.

[27] S. G. Petkar, V. K. Thippiripati, "A novel duty-controlled DTC of a surface PMSM drive with reduced torque and flux ripples," IEEE Trans. Ind. Electron., 70(4): 3373-3383, 2023.

[28] Y. Liu et al., "Direct torque control schemes for dual three-phase PMSM considering unbalanced DC-Link voltages," IEEE Trans. Energy Convers., 39(1): 229-242, 2024.

[29] T. Yuan, D. Wang, X. Wang, X. Wang, Z. Sun, "High-precision servo control of industrial robot driven by PMSM-DTC utilizing composite active vectors," IEEE Access, 7: 7577-7587, 2019.

[30] T. A. Huynh, Y. T. Nguyen Le, Z. Lee, M. C. Tsai, P. W. Huang, M. F. Hsieh, "Influence of Flux barriers and permanent magnet arrangements on performance of high-speed flux-intensifying IPM Motor," IEEE Trans. Magn., 59(11): 1-6, 2023.

Biographies

Hadi Afsharirad was born in Abhar, Iran, in 1985. He received the B.Sc. degree from the Zanjan University, Iran and M.Sc. and Ph.D. degrees, from the University of Tabriz, Tabriz, Iran, in 2008, 2010, and 2018, respectively, all in Electrical Engineering. He is an Assistant Professor with the Department of Electrical Engineering, Azarbaijan Shahid Madani University, Tabriz, which he joined in 2020. His main research interests include the electric and hybrid electric vehicles, renewable energy, linear electric machines and electrical drives.

- Email: h.afsharirad@azaruniv.ac.ir
- ORCID: 0009-0005-6259-7855
- Web of Science Researcher ID: LNR-5960-2024
- Scopus Author ID: 36670802500
- Homepage: hppts://pajouhesh.azaruniv.ac.ir/_Pages/Researcher.aspx?ID=8492

Sara Misaghi was born in Iran, in 1997. She received the B.Sc. and M.Sc. degree from the Azarbaijan Shahid Madani University, Iran. Her main research interests include electric machines and Electrical Drives.

- Email: sara.misaghi@azaruniv.ac.ir
- ORCID: 0009-0002-0145-4461
- Web of Science Researcher ID: LNR-6161-2024
- Scopus Author ID: NA
- Homepage: https://pajouhesh.azaruniv.ac.ir/_Pages/Researcher.aspx?ID=12999

How to cite this paper:
H. Afsharirad, S. Misaghi, "Torque ripple reduction by using virtual vectors in direct torque control method using neutral-point-clamped inverter," J. Electr. Comput. Eng. Innovations, 13(1): 197-208, 2025.

DOI: 10.22061/jecei.2024.10894.745

URL: https://jecei.sru.ac.ir/article_2227.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

A Fast and Accurate Tree-based Approach for Anomaly Detection in Streaming Data

K. Moeenfar, V. Kiani *, A. Soltani, R. Ravanifard

Computer Engineering Department, Faculty of Engineering, University of Bojnord, Bojnord, Iran.

Article Info

Article History:
Received 26 July 2024
Reviewed 17 September 2024
Revised 28 October 2024
Accepted 17 November 2024

Keywords:
Anomaly detection
Data streams
Concept drift
Sliding window
Isolation tree

*Corresponding Author's Email Address: *v.kiani@ub.ac.ir*

Abstract

Background and Objectives: In this paper, a novel and efficient unsupervised machine learning algorithm named EiForestASD is proposed for distinguishing anomalies from normal data in data streams. The proposed algorithm leverages a forest of isolation trees to detect anomaly data instances.
Methods: The proposed method EiForestASD incorporates an isolation forest as an adaptable detector model that adjusts to new data over time. To handle concept drifts in the data stream, a window-based concept drift detection is employed that discards only those isolation trees that are incompatible with the new concept. The proposed method is implemented using the Python programming language and the Scikit-Multiflow library.
Results: Experimental evaluations were conducted on six real-world and two synthetic data streams. Results reveal that the proposed method EiForestASD reduces computation time by 19% and enhances anomaly detection rate by 9% compared to the baseline method iForestASD. These results highlight the efficacy and efficiency of the EiForestASD in the context of anomaly detection in data streams.
Conclusion: The EiForestASD method handles concept change using an intelligent strategy where only those trees from the detector model incompatible with the new concept are removed and reconstructed. This modification of the concept drift handling mechanism in the EiForestASD significantly reduces computation time and improves anomaly detection accuracy.

Introduction

The detection of anomalies in data streams has become an increasingly significant research area, driven by the exponential growth in the volume and velocity of streaming data across diverse domains such as finance, healthcare, Internet of Things (IOT), and computer networks [1]-[3]. Anomalies, which are data instances that deviate significantly from the norm, can provide valuable insights into abnormal events, fraud, or potential risks in various applications [4]-[6]. However, traditional anomaly detection methods designed for static datasets are ill-suited for streaming data due to the dynamic nature of data streams and their inherent challenges.

Anomaly detection is a classification task encompassing supervised, semi-supervised, and unsupervised learning approaches [7]. Supervised learning methods are constrained in their ability to detect new anomalies and can only identify those resembling previously encountered data anomalies. Conversely, unsupervised methods offer the advantage of discovering novel anomalies without the need for training labels, which is particularly valuable considering the cost associated with acquiring and labeling training data. Given this rationale, the primary focus of this research lies in the identification of anomalies within data streams through the unsupervised learning methods.

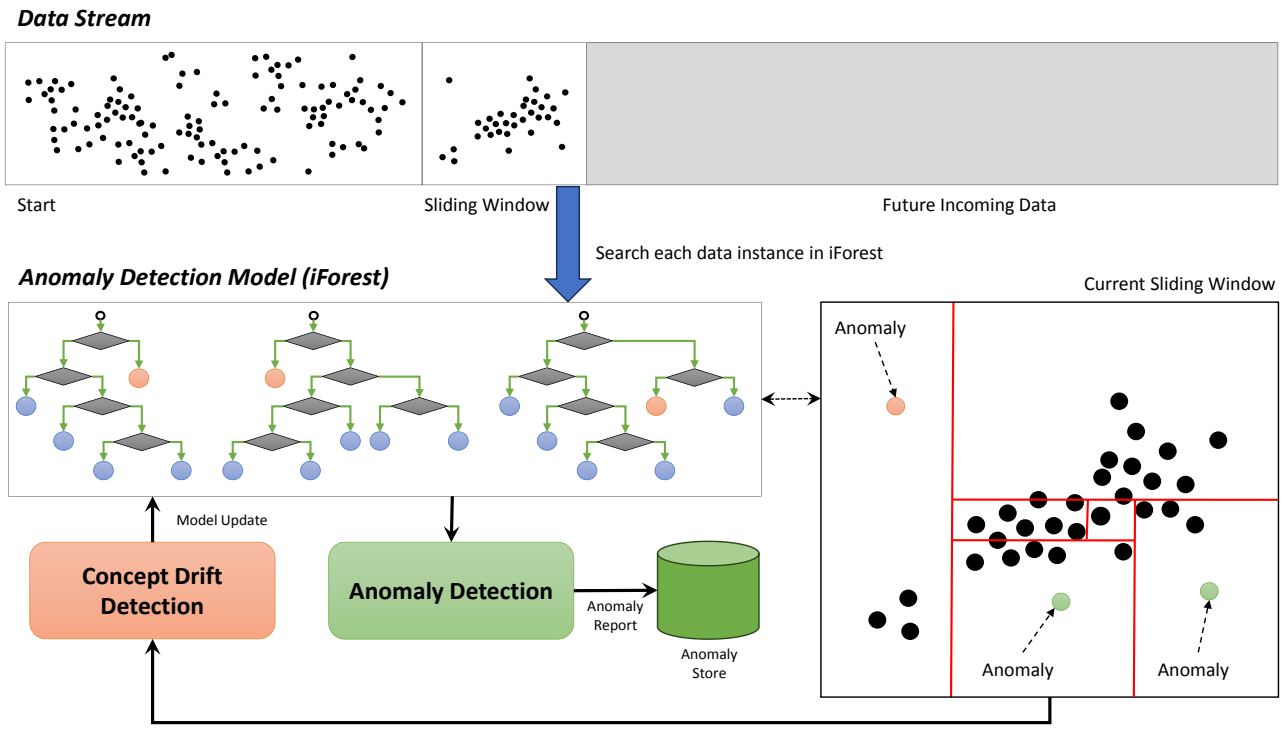


Fig. 1: Conceptual framework of the proposed method for anomaly detection in data streams.

A data stream is a massive, continuous, unbounded and ordered sequence of incoming data at high speed [8]-[10]. In the context of data streams, anomaly data refers to a data instance that significantly deviates from the expected or normal behavior of the data stream [11]. It is an observation that stands out from the majority of the data points and exhibits characteristics that are unusual or unexpected. Anomalies can represent abnormalities, irregularities, or rare events in the data stream, which may hold valuable information or indicate potential issues in the underlying process generating the data [3]. While outliers are often identified using statistical methods, anomaly detection techniques can be more advanced and include methods like machine learning, time-series analysis, and clustering, which may consider temporal patterns and relationships in data streams.

An ideal anomaly detection method for data streams must possess several key characteristics [9]. Firstly, it should be capable of processing massive and never-ending streams of data in real time, as the volume and velocity of streaming data necessitate efficient and timely analysis. Secondly, the method should be adaptable to potential changes in the underlying data distribution over time, as data streams often exhibit concept drift, where the statistical properties of the data may evolve. Thirdly, the method should exhibit high accuracy in identifying anomaly data instances, as the consequences of missing or misclassifying anomalies can be significant in critical applications.

To address these challenges, this research paper introduces a novel and efficient method called Enhanced Isolation Forest Adapted for Streaming Data (EiForestASD). Conceptual framework of the proposed anomaly detection method EiForestASD is shown in Fig. 1. The proposed method leverages a forest of isolation trees, a popular technique in anomaly detection, to effectively distinguish anomalies from normal data instances in streaming environments. In addition, EiForestASD integrates an adaptable detector model that dynamically adjusts to accommodate new data over time. Rather than completely discarding the detector model when encountering a concept drift, the proposed method employs a manipulation strategy to effectively manage the concept change. Specifically, only invalid and weak detectors are removed instead of entirely eliminating the set of detectors. This approach facilitates the high-speed processing of data streams and enhances the accuracy of anomaly detection within the data stream. The contributions of this article are as follows:

1. Introducing a novel and efficient method named EiForestASD for anomaly detection in streaming data.
2. Addressing the challenges posed by streaming data by developing a method that is fast, capable of processing massive and endless data streams, adaptable to potential changes in data distribution over time, and exhibits high accuracy in identifying anomaly data instances.
3. Incorporating an adaptable detector model that

adjusts to new data over time, discarding only those isolation trees that are incompatible with the new concept in the event of concept drift.

4. Conducting experimental evaluations on both real-world and synthetic data streams.

Overall, this research contributes to the field of anomaly detection in data streams by introducing a novel method that addresses the main challenges posed by streaming data. The findings highlight the efficacy and efficiency of EiForestASD in handling massive and perpetual data streams with high speed, adapting to concept drift, and accurately identifying anomalies. The proposed method holds great promise for various applications that require real-time and accurate anomaly detection in streaming data.

The remainder of this article is organized as follows: Initially a comprehensive literature review of unsupervised methods for anomaly detection in data streams is presented. Subsequently, an extensive review of related works is conducted, concentrating exclusively on methodologies that utilize isolation trees and isolation forests for unsupervised anomaly detection within data streams, given that our proposed method is founded on isolation forests. Then the notion of isolation trees, the proposed method EiForestASD, and its mechanism for handling concept drift are explained. After that, reports and analyzes of the experimental results on the benchmark datasets is provided. Finally, the article is concluded and directions for future research are introduced.

Literature Review

Anomaly detection is defined as the process of identifying patterns within data that deviate from expected behavior. Unlike static datasets, which typically contain labeled examples suitable for model training, real-time data streams often lack such annotations, thereby rendering unsupervised methods particularly advantageous. This literature review emphasizes unsupervised techniques that utilize the inherent structure of the data to identify anomalies in the absence of prior labeling. Prior research on unsupervised anomaly detection in data streams can be categorized into seven primary groups: statistical methods, distance-based methods, density-based methods, clustering methods, tree-based methods, deep learning methods, and hybrid approaches.

The statistical or parametric approach to anomaly detection assumes that data instances in a data stream conform to a specific statistical distribution, with significant deviations from this distribution identified as anomalies. Various methodologies illustrate this approach. In [12], the outlier score for each data point is calculated based on the Gaussian mixture model (GMM). In [13], correlation of two or more correlated features is computed and an ellipsoidal boundary around these features is constructed as the model of normal data. In [14], the effects of seasonality and trend is first removed from the data stream, and then RobustSTL and RobustScaler methods are used to detect anomalies based on mean, variance, median, and interquartile range of data. In [15], the kernel density estimation (KDE) is used to generate and continually update a real-time statistical model of the data stream, and the likelihood estimates are then used to detect anomalies. In [16], considering high-dimensional medical data streams, the information entropy and an efficient pruning technique are combined in a novel sliding window model to judge whether the data is anomalous or not.

The distance-based approach considers the distance of each data point to its nearest neighbors. For example, considering k and R parameters, a data is known as an outlier if less than k data in the input data are within R distance from this data point. Exact-Storm [17] and Direct-Update [18] algorithms are two common algorithms in the group of distance-based methods. Recently, most research works in this group are focused on the efficient computation of distance-based methods. In [19], data points at similar locations are grouped and the detection of outliers or inliers is handled at the set level. In [20], micro-clustering technique is integrated with adaptive thresholding of Thresh-LEAP algorithm. In [21], a grid-based index is proposed to effectively manage summary information of streaming data, and a min-heap algorithm is employed to efficiently calculate the distance bounds between objects and their k nearest neighbors. In [22], a method based on subspaces is proposed for explaining anomalies and describing relevant dimensions in the unsupervised distance-based outlier detection.

In the density-based approach to outlier detection, the local density of each data point serves as a fundamental criterion for identifying anomalies [11]. A prominent method within this framework is the Local Outlier Factor (LOF), which introduces the notion of comparing the local densities of neighboring data points with that of the target data point [23]. Data points exhibiting a high LOF are deemed outliers. The concept of LOF is then integrated with a sliding window approach to effectively manage data streams. This integration has established a fundamental principle that underpins various subsequent research endeavors, including iLOF [24], DiLOF [25], CLOF [26], and GiLOF [27]. Recently, LOF method is enhanced by leveraging ensemble techniques and GPU acceleration on data streams [28]. In [29], LOF is combined with PCA-based dimensionality reduction to infer data stream anomalies in real time. In [30], using information entropy for feature selection, clustering for

memory reduction, and data insertion for density computation, the detection accuracy of LOF is enhanced while its memory requirements is reduced for high-dimensional data streams.

Clustering techniques have proven effective for detecting anomalies by identifying groups of similar data points. Methods like k-means and DBSCAN can be adapted for anomaly detection in data streams. Among preliminary algorithms that fall into this category, we can mention DenStream [31], DBStream [32], and EvoStream [33]. Several recent methods for anomaly detection in data streams are based on clustering. For example, in [34], a clustering technique is used to summarize data before applying anomaly detection methods on the summary. In [35], several clustering algorithms including K-means clustering, Mixture of Gaussian models, density-based clustering, and self-organizing maps are employed on stream data for online anomaly detection and monitoring of ship machinery systems. In [36], a streaming sliding window local outlier factor coreset clustering algorithm (SSWLOFCC) is proposed, which integrates local outlier factor, agglomerative clustering, and PCA for efficient outlier detection in large datasets. In [37], a dynamic micro-clustering scheme is proposed, generating macro-clusters from interconnected micro-clusters to identify anomalies by assessing both global and local density perspectives.

Tree-based methods are commonly employed for unsupervised anomaly detection in both static datasets and data streams [38]. In a tree-based anomaly detection method for data streams, an ensemble of tree-based anomaly detection models such as half-space tree, random-space tree, or isolation tree is often combined with a sliding window mechanism to detect anomalies in stream data and update anomaly detector continually [39], [40]. For example, in [41], an ensemble of random half-space trees called Streaming HS-Trees method is employed to detect anomalies in stream data. It offers several advantages, including constant amortized time complexity, constant memory requirements, and favorable detection accuracy. In [42], to leverage the benefits of fully randomized-space trees, the RS-Forest utilized multiple fully RS-Trees to create a density estimator that is both fast and accurate. Then, the incoming instances in a data stream are scored based on the density estimates averaged over all trees in the forest and the anomalies are identified. In [43], an ensemble of isolation-Trees known as Isolation Forest (iForest) is proposed for detecting anomalies in static datasets. To compute the anomaly score for a particular data point, the path lengths of the trees containing that point are averaged. In [44], the isolation forest technique was extended to effectively handle the unique characteristics of streaming data by incorporating sliding windows mechanism. Some recent extensions of isolation forest to streaming data include applying ADWIN to iForestASD [45], Historical Isolated Forest (HIF) [46], and Bilateral-Weighted Online Adaptive Isolation Forest (BWOAIF) [47].

Deep learning has opened new avenues for anomaly detection in data streams. In [48], the LSTM networks is used as a predictor of future data, and anomalies are detected by comparing the predicted value and actual value of current data point. A similar scheme is employed in [49] where Temporal Convolutional Neural Networks (TCN) provided higher accuracy than LSTM and GRU models. Another approach is to employ deep learning models in an auto-encoder network to reconstruct the output based on previous sequence of inputs, and detect anomalies based on reconstruction loss. In this regard, in [50] an LSTM-based auto-encoder network is designed for anomaly detection in vibration data of wind turbines. In [51], an auto-encoder anomaly detector is equipped with concept drift detection module based on the Mann-Whitney U Test to adapt nonstationary environments. In [52], to reduce network complexity and computational requirements, the encoder network is constructed from LSTM layers, while the decoder network is comprised from fully connected layers. Lastly, Generative Adversarial Networks (GANs) have been applied to anomaly detection by generating normal data samples for comparison with observed data [53]-[56].

Hybrid approaches combine multiple methods to improve anomaly detection accuracy and robustness [57]. For example, a hybrid Model of One-class SVM and Isolation Forest (HMOI) has been proposed in [58] for wireless sensor data, where isolation forest is employed for anomaly labeling of unlabeled data, and one-class SVM is utilized for final classification of anomalies. In [59], to detect anomaly in surveillance videos, a Convolutional Neural Network (CNN) is employed to extract spatial information, combined with a vision transformer to learn long-term temporal relationships. In [60], a hybrid approach based on deep learning is proposed that combines CNN and LSTM models in the encoder and decoder parts of an auto-encoder model to detect anomalies in spatio-temporal data.

In conclusion, the domain of unsupervised anomaly detection in data streams presents a rich and diverse landscape of methodologies, each tailored to address the unique challenges posed by the absence of labeled data and the dynamic nature of real-time information. By categorizing existing techniques into seven distinct groups—statistical, distance-based, density-based, clustering, tree-based, deep learning, and hybrid approaches—we can appreciate the breadth of strategies developed to tackle this complex problem. Advances in these areas continue to enhance detection accuracy and

computational efficiency, paving the way for real-time applications across various fields, including finance, healthcare, and cybersecurity.

Related Works on Isolation Forests

Similar to our research work, numerous other scholars have employed the concepts of isolation trees and isolation forest for the identification of anomalous data in data streams. Accordingly, in this section a detailed review of the related works to application of isolation forest and isolation trees for anomaly detection in data streams is presented.

In the realm of streaming data analysis, Ding et al. [44] extended the Isolation Forest technique to effectively handle the unique characteristics of streaming data, such as high speed, large volume, and concept drift. Their proposed method, known as iForest Adapted for Streaming Data (iForestASD), incorporates sliding windows to cope with the continuous flow of data and adapt to concept drift. By employing bootstrap sampling, an initial anomaly detection model is constructed for the streaming dataset, and iTrees are built based on the randomly sampled data. The trained iForest model is continuously updated as new data arrives, ensuring the detection of evolving anomalies. Moreover, iForestASD is equipped to detect and handle concept drifts by monitoring the anomaly rate within a sliding window. If the anomaly rate exceeds a predefined threshold, concept drift is identified, and a new iForest model is constructed to accommodate the latest data window. With the increasing popularity of the Python programming language in the data science, Togbe et al. [61] implemented the iForestASD method under the Python programming language and the Scikit-Multiflow machine learning framework.

Togbe et al. [45] extended the iForestASD method to handle drifting data by introducing three new algorithms. These algorithms utilize two primary drift detection methods: ADWIN and KSWIN. By calculating and analyzing the average statistics in two sub-windows, ADWIN identifies concept drift. Similarly, the KSWIN method employs the Kolmogorov-Smirnov test to identify changes in data distribution. In addition, Togbe et al. introduced N-Dimensional KSWIN (NDKSWIN) to adapt KSWIN for multidimensional data streams, declaring a drift if a change is detected in at least one dimension.

Madkour et al. [46] enhanced the existing IForestASD methodology by introducing a Historical Isolated Forest (HIF) framework and reusing previously constructed iForests. Their proposed method retains previously constructed isolation forests and utilizes the isolation forest most analogous to the current concept drift distribution as its operational model. Additionally, it maintains the mean and standard deviation of the training data chunk alongside each isolation forest within the ensemble pool to facilitate the assessment of similarity between the current concept drift distribution and earlier data distributions. Evaluations revealed that while the HIF approach achieved reduced computational times compared to IForestASD, it often did not improve and sometimes decreased anomaly detection accuracy.

Hannak et al. [47] improved the IForestASD by adding timestamps for each isolation tree (iTree) and using a bilateral weighting mechanism for calculating anomaly scores. Their approach, called the Bilateral-Weighted Online Adaptive Isolation Forest (BWOAIF), assigns weights to iTrees to reduce the impact of outdated trees in changing data distributions. The anomaly score calculation employs bilateral weighting, where one component mitigates the influence of iTrees built from differing distributions, while the other emphasizes more recent trees. Empirical results showed that BWOAIF effectively adapts to various concept drift situations, including slow and fast shifts, splits, and the emergence or disappearance of concepts.

Yang et al. [62] proposed ASTREAM method which integrates Locality-Sensitive Hashing (LSH) into isolation Forest (iForest) to achieve better anomaly detection performance. The underlying model used in ASTREAM is called LSHiForest. ASTREAM addresses the limitations of existing approaches by incorporating sliding window, model updates, and change detection strategies into LSHiForest. The sliding window mechanism effectively handles the continuous flow of data streams, while Principal Component Analysis (PCA) considers the correlations between different dimensions and transforms a set of relevant dimensions to a set of irrelevant dimensions. Extensive experiments conducted on the KDDCUP99 dataset have demonstrated the superior performance of ASTREAM in terms of accuracy, efficiency, and scalability compared to baseline methods.

Another study by Yang et al. [63] introduces DLSHiForest, which combines Locality-Sensitive Hashing (LSH), Isolation Forest, and the time window technique to achieve accurate and efficient anomaly detection in data streams from wireless sensors. DLSHiForest takes into account correlations between different dimensions and detects anomaly based on Locality-Sensitive Hashing. Each streaming data point is treated as a multidimensional vector, and the hash functions consider all the dimension information while hashing the data points, thereby accounting for the cross-correlation among different dimensions. The hashing process involves the dot product of two vectors, which represents the comprehensive consideration of all dimensional information of the data point. The efficient partitioning of data points and the detection of anomaly in DSLHiForest heavily rely on the hashing technique.

Li et al. [64] introduced an innovative human-machine

interactive streaming anomaly detection approach, referred to as ISPForest, which is capable of being adaptively updated in real time through the integration of human feedback. In their framework, the feedback mechanism plays a critical role in recalibrating both the computation of anomaly scores and the architecture of the detector itself, thereby enhancing the accuracy of future anomaly score assessments. The empirical findings from their study indicate that the inclusion of feedback significantly improves the performance of anomaly detection systems with minimal human intervention.

A detailed review of existing literature in the domain of tree-based unsupervised anomaly detection using isolation forests has revealed that isolation forest exhibits notable efficiency and scalability in the detection of anomaly within data streams. In contrast to the majority of such tree-based methods that fail to adequately address the challenge of concept drift in data, the iForestASD algorithm employs continual monitoring of the anomaly rate to identify concept drift and subsequently reconstructs the entire detector model upon detecting a change in the concept. The approach employed by newer tree-based methods, which utilize isolation forest, closely mirrors that of iForestASD in terms of identifying and handling concept drift. Nevertheless, the iForestASD method is impeded by the time-intensive process of rebuilding the detector model, resulting in significant algorithmic slowdown and delays in identifying anomaly data. Hence, it is imperative to explore alternative methods that offer more efficient management of concept drift. To address this objective, in this research work, the EiForestASD technique is introduced which integrates a mechanism to manage concept drift without discarding the entire detector model. This approach requires less time to update the model, making it highly adaptable to data changes and particularly suitable for resource-constrained devices.

EiForestASD Method

This research aims to develop a method that can detect anomalies in data streams and update the detector model efficiently when the concept drifts occur. The proposed method, named Enhanced Isolation Forest Algorithm for Stream Data (EiForestASD), is an enhancement of the iForestASD algorithm proposed by [44], [61], but with more smartness in handling concept drifts and updating the detector model. The EiForestASD is a partitioning-based anomaly detection method for data streams, which employs a forest of isolation trees (iTrees) to isolate anomalies. The EiForestASD continuously updates the forest of iTrees on the data stream and uses it to detect anomalies in each window of data. The steps of the proposed method EiForestASD are depicted in Fig. 2.

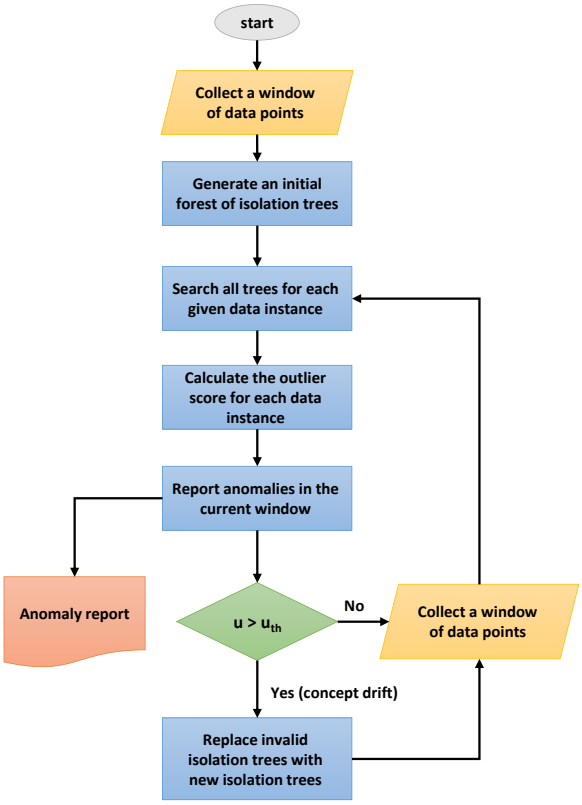


Fig. 2: Steps of the proposed method EiForestASD for anomaly detection in data streams.

The proposed method EiForestASD begins by receiving a window of data and constructing the initial iForest detector model. Subsequently, for each incoming data point, the algorithm searches the input data point in all isolation trees and uses the isolation forest to compute anomaly score of the input data point and report anomaly data points. After reporting anomaly data instances, EiForestASD computes the anomaly rate in the last window of data and compares it with a predefined threshold to determine if there is a concept drift. If a concept drift is detected, the EiForestASD identifies weak and obsolete isolation trees in the current detector model and replaces them with new ones constructed on the last window of data. The following sections will explain the details of the proposed method.

A. Isolation Tree

An isolation tree (iTree) is a binary tree that recursively partitions the data space in a hierarchical manner. Each node of the iTree represents a subset of the data, and each branch represents a split of the subset into two smaller subsets. The construction of an iTree is done randomly. In the randomization process, to expand an arbitrary node of the tree, a feature is randomly selected from the data, and then a split point is randomly selected from the range of values of that feature. Then, the selected feature and the split point are used to split the

data of the current node into two subsets. Data points that have higher values than the split point for the selected feature form the right child of the current node, and those that have lower values form the left child. The construction of an iTree starts from the root node, which contains the whole dataset. The randomization process is applied to the current node to split it into two children, and this process is repeated on the children. This process continues until a leaf node is created that contains a small number of data points or a maximum depth is reached. Fig. 3 illustrates an example of partitioning a dataset using an iTree, where this sample dataset has only two features and five data points. In Fig. 3, the data instance "a" is isolated from other data instances at the first level of the iTree, and would be a proper candidate for anomaly.

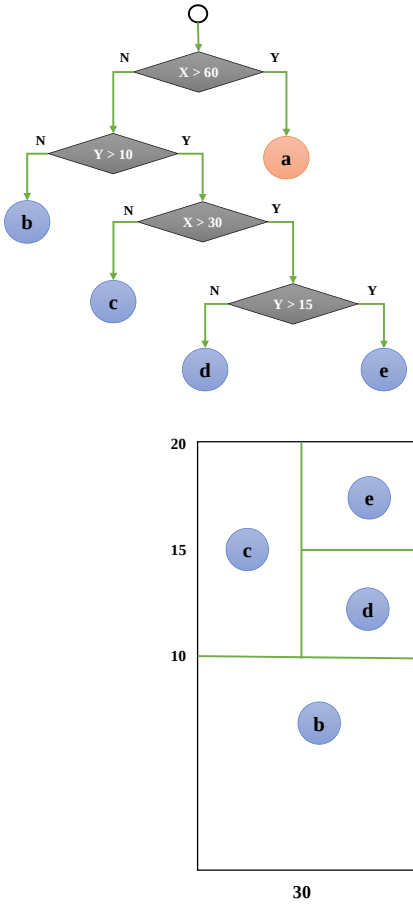


Fig. 3: Using an isolation tree to partition dataset and separate data points.

B. Anomaly Detection

An isolation forest, also known as an iForest, is a group of isolation trees. The concept of isolating anomalies instead of profiling normal instances is introduced in the isolation forest [43], [65], resulting in a more efficient sub-sampling method and a linear time complexity algorithm with low memory requirements. The isolation forest constructs a collection of isolation trees, with each tree randomly selecting a subset of instances and creating splits based on randomly selected attributes. The anomaly score of an instance is measured by its average path length in the isolation trees, with shorter path lengths indicating higher anomaly scores.

In an iTree, the leaves that have smaller depths are isolated from the rest of the data with only a few partitioning steps. Therefore, the leaves that are located at a lower depth are likely to be anomalous. After an iTree is constructed, for each new data point, we search it in the tree to reach a leaf. The depth of that leaf node determines the anomaly score of the new data point. The deeper the leaf node, the lower the anomaly score should be. Fig. 4 demonstrates the difference in leaf node depth and the number of partitioning steps for anomaly and non-anomaly data points.

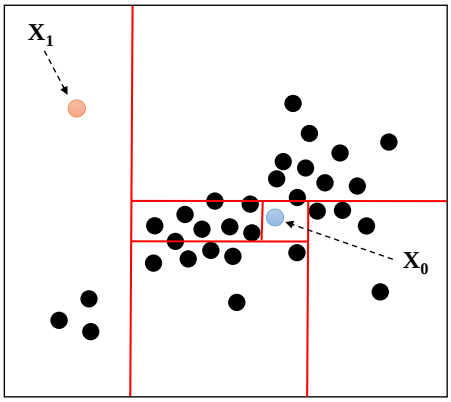


Fig. 4: It takes five steps to separate the inlier data X_0, while the anomaly data X_1 is separated much faster in just two steps.

In the proposed method EiForestASD, a forest of iTrees is used as a detector model. The iForest is built from the first observed data window, and updated as subsequent windows arrive. For each new data point, its anomaly score must be computed by the current detector model. To this end, the new data point is searched in all the trees of the iForest, and based on its depth in the trees, an anomaly score $s(x)$ is calculated for the new data point. Data points that have an anomaly score higher than the anomaly threshold, denoted by s_{th}, are reported as anomalous.

Assume that the size of the sliding window is equal to M samples. Also, suppose that the number of iTrees in the iForest detector model is N trees. In this case, the anomaly score $s(x)$ of the data point x is calculated with the following equation:

$$s(x) = 2^{-\frac{E(h(x))}{c(M)}} \quad (1)$$

where the symbol $c(M)$ represents the expected average value of the path length $h(x)$ for all data points in the current window. If the window length M is greater than 2, the value of $c(M)$ is calculated according to the following equation:

$$c(M) = 2H(M-1) - (M-1)/M \quad (2)$$

where $H(M)$ represents the harmonic number, which can be estimated by the relation $H(M) = ln(M) + \gamma$ and the value of $\gamma = 0.5772156649$. Also, the symbol $E(h(x))$ shows the average depth of the leaf node containing x in the isolation trees of the current iForest detector model, which is calculated according to the following equation:

$$E\big(h(x)\big) = \frac{1}{N}\sum_{i=1}^{N} h_i(x) \quad (3)$$

C. Concept Drift Detection and Handling

The EiForestASD method handles the concept drifts in the data streams by monitoring a typical anomaly rate. The anomaly rate in each data window is computed by calculating the ratio of the number of data points detected as anomalous to the total number of data points in the window. If the anomaly rate of the window exceeds the concept drift threshold, then a concept drift has occurred. In case of a concept drift, the baseline algorithm iForestASD would discard its current forest of isolation trees and start building a new forest using all the data points. However, in the proposed method EiForestASD, we use a more intelligent approach. To reduce the computation time of the algorithm, in the EiForestASD method, when a concept drift occurs, instead of removing all the isolation trees, we remove only weak iTrees, the trees that classify most of the data points of the current window as anomalies. The procedure of concept drift detection and handling in the proposed method EiForestASD is outlined by Algorithm 1.

Subsequent to processing the most recent window of data points, Algorithm 1 is employed to identify and address any occurrence of concept drift. Initially, in lines 1 to 4, the anomaly status of all data points within the last window of data, denoted as W, is examined, enabling the computation of the anomaly rate associated with W. Subsequently, in line 5, the concept drift is determined by comparing the anomaly rate, referred to as u, of the window W of most recent data points with the predefined threshold for concept drift, known as u_{th}. If the anomaly rate surpasses the concept drift threshold, it implies the occurrence of concept drift, necessitating the execution of appropriate steps outlined in lines 6 to 15.

To effectively address the concept drift, a for loop is initiated in line 7, inspecting each isolation tree within the current ensemble model for obsoleteness. In the event that a tree is deemed invalid, it is replaced with a newly generated isolation tree model, crafted based on the latest data points obtained from window W. The obsoleteness checking process is carried out in lines 8 to 11, involving the computation of the anomaly rate associated with the current isolation tree t in relation to the data points present within the latest window W. Subsequently, this anomaly rate is compared with the specified anomaly threshold, u_{th}. If the anomaly rate of isolation tree t exceeds the anomaly threshold, the tree t is considered to be obsolete and invalid, consequently requiring substitution with a newly created isolation tree as delineated in lines 11 to 14.

Algorithm 1 Concept drift detection and handling in EiForestASD

Inputs: W – window of latest data points, F – ensemble of iTrees, s_{th} – anomaly threshold, u_{th} – concept drift threshold
Outputs: F – updated ensemble of isolation trees

1: Search each data point **x ∈ W** in all isolation trees **t ∈ F**
2: Compute anomaly score **s(x)** for each data point **x ∈ W** using all isolation trees **t ∈ F**
3: Count the number of anomaly points in the last window of data **W**
4: Compute anomaly rate **u** of the last window of data **W**
5: if **$u > u_{th}$**
6: // Concept drift happened
7: for each isolation tree **t ∈ F**
8: Compute the anomaly score for each data point **x ∈ W** using only one isolation tree **t**
9: Count the number of anomaly points detected by isolation tree **t** in the last window of data **W**
10: Compute anomaly rate of isolation tree **t** as **u_{tree}**
11: if **$u_{tree} > u_{th}$**
12: // This tree is not valid and should be replaced
13: replace isolation tree **t ∈ F** with a new iTree trained on the last window of data **P**
14: end if
15: end for
16: end if
17: return updated ensemble of isolation trees **F**

In this way, in the proposed method EiForestASD, only obsolete iTrees are removed from the current detector model. The removed iTrees are replaced with new iTrees constructed based on all the data points in the current window.

Evaluation and Results

In this section, the performance of the proposed method EiForestASD in identifying anomalies in the data stream will be evaluated and compared with the baseline iForestASD method in [44]. For this purpose, the proposed method was implemented using Python programming language and with the help of the Scikit-Multiflow library and compared with the Python implementation of the iForestASD method in this library [61].

In the experiments of this section, the actual anomaly rate of each data set was used to value the anomaly rate threshold parameter in the concept drift detection section. The value of parameter M, which determines the size of the window, was considered equal to 100. For the parameter N, which determines the number of isolation trees in the detector model, values of 30, 50, and 100 were tested. The anomaly threshold s_{th} for each data point was set to 0.5. In this case, a data point is considered an anomaly by the detector model if its average depth in

the detector trees is less than half of the expected value for the depth of the leaf nodes. The concept drift threshold u_{th} was considered equal to the actual anomaly rate of each data set. In other words, we assume that the anomaly rate in each window is the same as the anomaly rate of the entire dataset.

A. Evaluation Metrics

This research will compare anomaly detection methods in terms of computation time and accuracy. The main objective of the proposed method EiForestASD is to reduce the computation time of the algorithm by intelligently managing the concept drifts. Therefore, the computation time will be the primary and most important evaluation criterion. To compare the computation time, the amount of CPU time spent by each algorithm on each dataset will be measured and reported in seconds.

Besides the computation time, the accuracy of anomaly detection is also crucial for each algorithm. To evaluate the accuracy of anomaly detection by each algorithm, the $\mathrm{F1}$ score will be used as a performance measure. Specifically, the anomaly detection problem will be treated as a binary classification problem. Anomalies will be considered as the positive class and non-anomalies as the negative class. Let P denote the precision and R denote the recall. Then, the $\mathrm{F1}$ score, which is a suitable metric for imbalanced classification problems, will be computed using the following equation:

$$\mathrm{F1} = \frac{2 \times P \times R}{P + R} \tag{4}$$

B. Benchmark Datasets

The proposed method and the competing algorithm were evaluated using two sets of real and synthetic data sets, which are commonly used as benchmarks for anomaly data identification in streaming data. The real datasets were obtained from the Anomaly Detection Datasets (ODDS)[1] library. Synthetic datasets were generated by the Scikit-MultiFlow library using different data generators. Table 1 summarizes the characteristics of the benchmark datasets, which were treated as data streams.

The synthetic data streams included Mulcross, which followed a multivariate normal distribution, and SEA, which had four blocks and abrupt concept drifts between them. The real data streams included HTTP and SMTP, which involved computer networks attack detection and computer networks intrusion prediction tasks, respectively; Forest Cover, which involved vegetation classification based on soil information; Shuttle, which involved deciding how to land a spacecraft; Sat Image, which involved pixel classification of satellite images; and MNIST, which involved image classification of English handwritten digits. The consecutive samples of each dataset were considered as a data stream.

Table 1: Characteristics of the benchmark datasets

Dataset	Number of Instances	Number of Attributes	Anomaly Rate
SEA	10000	3	0.10 %
Mul Cross	262144	4	10 %
HTTP	567498	3	0.39 %
SMTP	976175	3	0.03 %
Forest Cover	286048	10	0.96 %
Shuttle	49097	9	7 %
Sat Image	7603	100	9 %
MNIST	5803	36	1.22 %

C. Evaluation of Computation Time

The primary objective of the EiForestASD method is to enhance the efficiency of anomaly detection in data streams by reducing the required computation time. To assess the effectiveness of the proposed method in achieving this goal, an experiment was conducted to compare the computation time of the proposed method EiForestASD with the baseline iForestASD method across different datasets.

The EiForestASD method incorporates two crucial parameters: the sliding window size and the ensemble size. Therefore, the experiment was performed with varying window sizes of 50, 100, and 500, and different numbers of trees including 30, 50, and 100.

The results of the experiment are presented in Table 2, Table 3, and Table 4 for window size of 50, 100, and 500 data instances, respectively. The CPU time of each method is reported in seconds. In Table 2, for each dataset, CPU time of the proposed method EiForestASD and the baseline method iForestASD is reported for window size of 50 data instances.

For each dataset, the experiment was carried out for 30, 50, and 100 iTrees in the anomaly detector model and the results are reported. The column Ratio indicates the ratio of the CPU time of EiForestASD with respect to the CPU time of iForestASD.

These findings indicate that the EiForestASD method consistently outperforms the baseline iForestASD algorithm in terms of computation time across all datasets and for different number of iTrees. For window size of 50 data instances, in average, the proposed method EiForestASD achived a reduction of 19% in computation time with respect to the baseline method iForestASD.

[1] http://odds.cs.stonybrook.edu/about-odds/

Table 2: Reduction of computation time (seconds) by the proposed method EiForestASD for window size of 50 data points

Dataset	# Trees	iForestASD	EiForestASD	Ratio
SEA	30	173	132	0.76
	50	183	150	0.82
	100	226	184	0.82
MulCross	30	4527	3851	0.85
	50	7563	6632	0.88
	100	14287	12387	0.87
HTTP	30	657	300	0.46
	50	777	663	0.85
	100	2640	2037	0.77
SMTP	30	453	384	0.85
	50	498	396	0.80
	100	898	729	0.81
ForestCover	30	4884	4103	0.84
	50	7238	6195	0.86
	100	8921	7834	0.88
Shuttle	30	6849	5540	0.81
	50	7853	6628	0.84
	100	11187	9249	0.83
SatImage	30	1531	1246	0.81
	50	2502	2131	0.85
	100	5121	4321	0.84
MNIST	30	1838	1371	0.75
	50	3002	2344	0.78
	100	6145	4753	0.77
Average		4165	3482	0.81

Table 3: Reduction of computation time by the proposed method EiForestASD for window size of 100 data points

Dataset	# Trees	iForestASD	EiForestASD	Ratio
SEA	30	172	132	0.77
	50	191	142	0.74
	100	243	198	0.82
MulCross	30	4655	3908	0.84
	50	7885	6818	0.86
	100	16131	14292	0.89
HTTP	30	1269	1128	0.89
	50	7505	6291	0.84
	100	11978	10344	0.86
SMTP	30	1051	900	0.86
	50	2103	1945	0.92
	100	4897	4507	0.92
ForestCover	30	5775	4895	0.85
	50	7824	6975	0.89
	100	10647	9233	0.87
Shuttle	30	8948	7368	0.82
	50	12381	10374	0.84
	100	15254	12531	0.82
SatImage	30	3161	2610	0.83
	50	5776	4975	0.86
	100	10595	9022	0.85
MNIST	30	3793	2871	0.76
	50	6932	5472	0.79
	100	12714	9924	0.78
Average		6745	5702	0.84

To determine the effect of window size on the performance of the algorithms under evaluation, the same experiment was repeated for window size of 100 data instances and the results are reported in Table 3. Similarly, the proposed method EiForestASD consistently outperforms the baseline iForestASD algorithm in terms of CPU time. For window size of 100 data instances, in average, the proposed method EiForestASD achived a reduction of 16% in computation time. Table 4 presents the results of the same experiment for window size of 500 data instances. While the computation time of the algorithms is significantly increased with respect to window size of 100 and 50 data instances, the EiForestASD still consistently outperforms the baseline iForestASD algorithm in terms of CPU time. For window size of 500 data instances, in average, the proposed method EiForestASD achived a reduction of 15% in computation time. These findings demonstrate that the proposed method EiForestASD is capable of processing input data streams more efficiently than the baseline method iForestASD.

Table 4: Reduction of computation time by the proposed method EiForestASD for window size of 500 data points

Dataset	# Trees	iForestASD	EiForestASD	Ratio
SEA	30	266	174	0.66
	50	274	200	0.73
	100	305	278	0.91
MulCross	30	9112	7800	0.86
	50	16927	14579	0.86
	100	36584	32001	0.87
HTTP	30	26850	24843	0.93
	50	30234	28950	0.96
	100	38981	37275	0.96
SMTP	30	9321	8425	0.90
	50	13803	12972	0.94
	100	26842	25840	0.96
ForestCover	30	9035	7959	0.88
	50	16535	14113	0.85
	100	30005	25825	0.86
Shuttle	30	14811	12176	0.82
	50	19868	16253	0.82
	100	31793	26081	0.82
SatImage	30	12087	10122	0.84
	50	19854	17069	0.86
	100	37401	32215	0.86
MNIST	30	14504	11134	0.77
	50	23824	18776	0.79
	100	44881	35437	0.79
Average		20171	17521	0.85

The effect of the number of trees {30, 50, 100} in the detector model on the computation time of each of the EiForestASD and iForestASD algorithms for different window sizes is illustrated in Fig. 5. The computation time increased for both algorithms as the number of trees increased, and the rate and pattern of this increase were similar for both algorithms. Moreover, EiForestASD

reduced the computation time for all datasets compared to the baseline method iForestASD. Therefore, the proposed method EiForestASD achieved a strong saving in computation time.

D. Evaluation of Anomaly Detection Accuracy

The accuracy of anomaly detection in data streams serves as a crucial criterion for evaluating the performance of anomaly detection algorithms. In this experiment, the anomaly detection accuracy of the proposed method EiForestASD was compared with the basic iForestASD algorithm across various datasets. The experiment encompassed different window sizes of 50, 100, and 500, as well as varying numbers of trees including 30, 50, and 100. The results obtained from this experiment are presented in Table 5.

Table 5: Improvement of anomaly detection accuracy by the proposed method EiForestASD compared with the baseline method iForestASD according to the F1 criterion

		Win Size = 50		Win Size = 100		Win Size = 500	
Dataset	# Trees	iForestASD	iForestASD	iForestASD	iForestASD	iForestASD	iForestASD
SEA	30	0.37	0.41	0.39	0.45	0.42	0.51
	50	0.37	0.41	0.39	0.47	0.43	0.51
	100	0.38	0.44	0.39	0.48	0.49	0.56
MulCross	30	0.63	0.69	0.69	0.75	0.76	0.82
	50	0.64	0.71	0.69	0.78	0.78	0.82
	100	0.65	0.71	0.69	0.78	0.78	0.84
HTTP	30	0.19	0.25	0.26	0.33	0.31	0.41
	50	0.20	0.26	0.29	0.36	0.31	0.44
	100	0.18	0.29	0.29	0.38	0.31	0.45
SMTP	30	0.39	0.43	0.40	0.46	0.42	0.49
	50	0.39	0.43	0.40	0.47	0.43	0.51
	100	0.40	0.44	0.41	0.47	0.43	0.52
ForestCover	30	0.24	0.30	0.31	0.36	0.52	0.59
	50	0.24	0.31	0.32	0.37	0.53	0.61
	100	0.25	0.31	0.32	0.39	0.55	0.61
Shuttle	30	0.67	0.72	0.73	0.76	0.81	0.86
	50	0.67	0.73	0.73	0.78	0.82	0.86
	100	0.68	0.73	0.74	0.78	0.84	0.88
SatImage	30	0.22	0.29	0.24	0.33	0.28	0.40
	50	0.23	0.30	0.24	0.33	0.28	0.40
	100	0.22	0.31	0.24	0.34	0.29	0.40
MNIST	30	0.38	0.49	0.41	0.55	0.44	0.59
	50	0.39	0.52	0.42	0.55	0.45	0.62
	100	0.39	0.54	0.44	0.57	0.47	0.62
Average		0.39	0.46	0.43	0.51	0.51	0.60

The values reported in the Table 5 demonstrate that the proposed method EiForestASD exhibits a substantial improvement in anomaly detection accuracy compared to the baseline algorithm. Unlike the basic iForestASD method that discards all isolation trees in the event of a concept change, the proposed method EiForestASD retains the isolation trees that remain compatible with the new concept. This smart approach contributes to the increased accuracy of the proposed method EiForestASD. On average, the proposed method achieved an improvement in anomaly detection accuracy of approximately 7% for a window size of 50, 8% for a window size of 100, and 9% for a window size of 500.

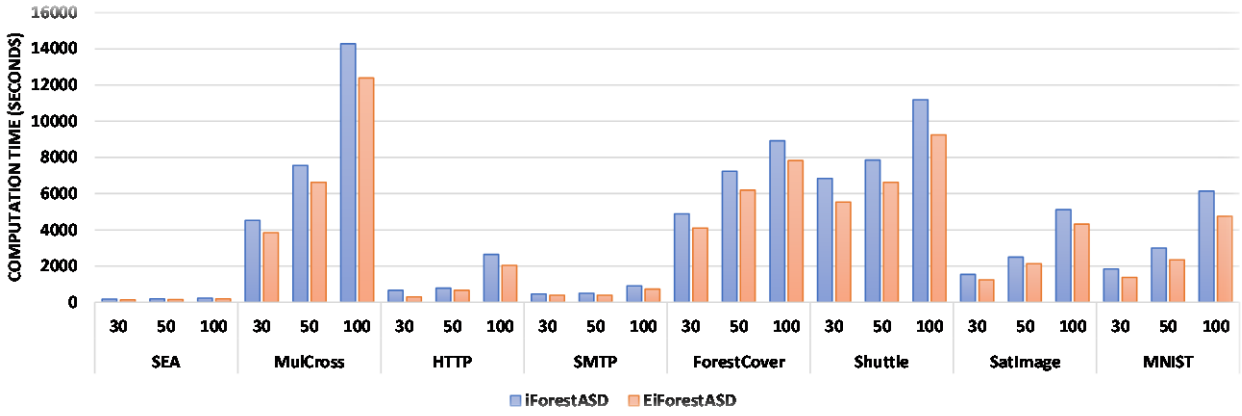
E. Limitations and Future Works

This research has several limitations that could be a basis for future investigations. Firstly, in the proposed methodology, the threshold value (u_{th}) for each data stream was established based on the pre-determined anomalous data rate. However, such a rate is typically unknown in real-world scenarios. Addressing this issue may involve developing a method to compute the threshold value of u_{th} based on the statistical distribution and inherent characteristics of the data, with the capacity for dynamic updates over time. Secondly, the proposed approach employed a fixed window size. A more adaptive strategy, where the window size is calibrated according to the properties of the data stream and adjusted periodically, could potentially enhance the accuracy of anomalous data detection. Similarly, the number of trees in the anomaly detection model can also computed atomically and updated dynamically. In this study, the iTree was chosen as the base model for anomaly detection; however, investigating alternative base models may yield valuable insights for future research. Lastly, considering that data streams represent an unbounded sequence of data points, summarizing the previous data points and the aggregation and analysis of these data summaries could significantly contribute to improving the efficacy of anomaly detection outcomes in a hybrid method.

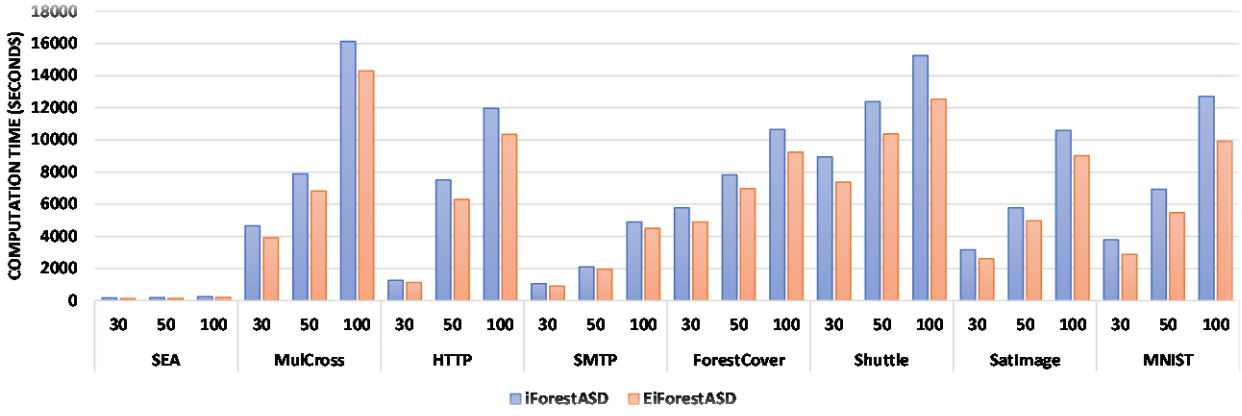
Discussion

The findings of this study highlight the effectiveness of the proposed EiForestASD algorithm in the realm of anomaly detection within data streams. By employing a specialized adaptive detection mechanism that discards only those isolation trees incompatible with new concepts, EiForestASD not only reduces computational overhead but also enhances detection accuracy. This distinguishes the proposed method from traditional algorithms such as the baseline iForestASD, which blindly eliminates all isolation trees upon encountering concept drift. In examining the computational efficiency of EiForestASD, the results indicate a consistent 19% improvement in computation time across varied datasets and configurations. By utilizing a window-based approach that maintains only relevant isolation trees, the algorithm adapts comprehensively to concept drift while sustaining high processing speeds. Furthermore, the evaluated datasets confirm that the proposed method significantly

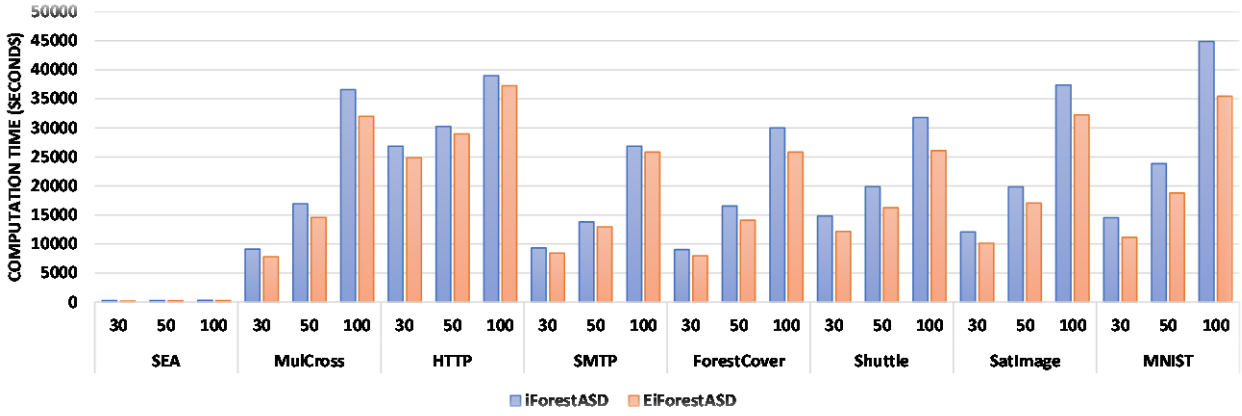
surpasses iForestASD in both computation time and anomaly detection accuracy, achieving up to a 9% improvement in the latter. The nuanced handling of concept changes contributes to this achievement, supporting the hypothesis that targeted tree removal is more efficient than blind destruction.



(a) window size of 50 samples



(b) window size of 100 samples



(c) window size of 500 samples

Fig. 5: Reduction of computation time by the proposed method EiForestASD compared with the baseline method iForestASD for: (a) window size of 50 samples, (b) window size of 100 samples, and (c) window size of 500 samples.

While the results are encouraging, several limitations necessitate further investigation. First, the reliance on a predetermined anomalous data rate to establish the threshold value (u_{th}) for anomaly detection poses

questions regarding the algorithm's applicability to real-world scenarios where such information is often unavailable. Future work should focus on devising a more dynamic thresholding system that can adaptively compute u_{th} based on the statistical characteristics of incoming data streams. Additionally, the static window size employed within this study may not optimally capture the differences of all data streams. An adaptive strategy that permits adjustment of window sizes based on real-time data analysis could yield substantial enhancements in detection performance. The proposed EiForestASD framework can be further enhanced to accommodate the detection of both incremental and recurring concept drift, enabling it to adopt distinct behaviors in response to each type of drift. In addition, development of a noise-resilient version of the EiForestASD algorithm is crucial, particularly for applications in dynamic environments where data quality can fluctuate considerably.

Conclusion

In this paper, the EiForestASD method is introduced as a means of identifying anomalies in data streams using a forest of isolation trees over time. The algorithm detects anomalies in the current window of data and updates the detector model, the forest of isolation trees, with each new window of data. The EiForestASD method handles concept change by removing and reconstructing only those trees from the detector model that are incompatible with the new concept, labeling most of the current window data as anomalies. This approach is more intelligent than the baseline iForestASD method, which discards all isolation trees when faced with concept change. The modification of the concept drift handling mechanism in the EiForestASD not only reduced computation time of the anomaly detection, but also improved anomaly detection accuracy. Since various types of concept drift exist in data streams, future research should focus on extending the proposed method to address gradual, recurring, and incremental drifts more effectively. The algorithm's robustness against these drift types could be evaluated using simulated drifts in synthetic datasets. Furthermore, the application of our proposed algorithm or its enhanced variants to real-world scenarios, such as human activity monitoring, presents an interesting area of research for exploration. Another critical challenge in data stream analysis is the presence of noise. Future versions of the proposed method should aim to enhance the resilience of both anomaly detection and concept drift detection mechanisms against noise, thereby improving overall performance in dynamic, real-world environments.

Author Contributions

V. Kiani supervised the research and did the necessary work to achieve the research goals; he sketched the research framework and the roadmap, analyzed the results, wrote the manuscript, and prepared revisions. K. Moeenfar implemented the main idea, performed experiments, and prepared experimental results. A. Soltani was research advisor and analyzed the results. R. Ravanifard analyzed the results and revised the manuscript. All authors read and approved the final version.

Acknowledgment

The authors would like to thank the editor and anonymous reviewers.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

ADWIN	Adaptive Window
BWOAIF	Bilateral-Weighted Online Adaptive Isolation Forest
CLOF	Composite Local Outlier Factor
CNN	Convolutional Neural Network
DLSHiForest	Dynamic Anomaly Detection based on Locality-Sensitive Hashing Isolation Forest
DiLOF	Density Summarizing Incremental LOF
EiForestASD	Enhanced iForestASD
GAN	Generative Adversarial Networks
GiLOF	Genetic-based incremental LOF
GMM	Gaussian Mixture Model
HIF	Historical Isolated Forest
HMOI	Hybrid Model of One-class SVM and Isolation Forest
HS-Tree	Half-space Tree
iForest	Isolation Forest
iForestASD	iForest Algorithm for Stream Data
iLOF	Incremental Local Outlier Factor
ISPForest	Interactive Space Partitioning Forest
KDE	Kernel Density Estimation
KSWIN	Kolmogorov–Smirnov Window
LOF	Local Outlier Factor

LSH	Locality-Sensitive Hashing
LSTM	Long Short-term Memory Network
ODDS	Outlier Detection Data Sets
PCA	Principal Component Analysis
RS-Tree	Randomized-space Tree
SSWLOFCC	Streaming Sliding Window LOF Coreset Clustering
TCN	Temporal Convolutional Neural Network

References

[1] R. Al-amri, R. K. Murugesan, M. Man, A. F. Abdulateef, M. A. Al-Sharafi, A. A. Alkahtani, "A review of machine learning and deep learning techniques for anomaly detection in IoT data," Appl. Sci., 11(12): 5320, 2021.

[2] R. A. Ariyaluran Habeeb, F. Nasaruddin, A. Gani, I. A. Targio Hashem, E. Ahmed, M. Imran, "Real-time big data processing for anomaly detection: A Survey," Int. J. Inf. Manag., 45: 289-307, 2019.

[3] M. Hosseini Shirvani, A. Akbarifar, "A survey study on intrusion detection system in wireless sensor network: Challenges and considerations," J. Electr. Comput. Eng. Innovations, 12(2): 449-474, 2024.

[4] A. A. Cook, G. Mısırlı, Z. Fan, "Anomaly detection for IoT time-series data: A survey," IEEE Internet Things J., 7(7): 6481-6494, 2020.

[5] L. Qi, Y. Yang, X. Zhou, W. Rafique, J. Ma, "Fast anomaly identification based on multiaspect data streams for intelligent intrusion detection toward secure industry 4.0," IEEE Trans. Ind. Inform., 18(9): 6503-6511, 2022.

[6] B. Steenwinckel, D. D. Paepe, S. V. Hautte, P. Heyvaert, M. Bentefrit, P. Moens, A. Dimou, B. V. D. Bossche, F. D. Turck, S. V. Hoecke, F. Ongenae, "FLAGS: A methodology for adaptive anomaly detection and root cause analysis on sensor data streams by fusing expert knowledge with machine learning," Future Gener. Comput. Syst., 116: 30-48, 2021.

[7] M. E. Villa-Pérez, M. Á. Álvarez-Carmona, O. Loyola-González, M. A. Medina-Pérez, J. C. Velazco-Rossell, K. K. R. Choo, "Semi-supervised anomaly detection algorithms: A comparative summary and future research directions," Knowl. Based Syst., 218: 106878, 2021.

[8] A. Oloomi, H. Khanmirza, "Fault tolerance of RTMP protocol for live video streaming applications in hybrid software-defined networks," J. Electr. Comput. Eng. Innovatons, 7(2): 241-250, 2019.

[9] T. Lu, L. Wang, X. Zhao, "Review of anomaly detection algorithms for data streams," Appl. Sci., 13(10): 6353, 2023.

[10] Z. Nouri, V. Kiani, H. Fadishei, "Rarity updated ensemble with oversampling: An ensemble approach to classification of imbalanced data streams," Stat. Anal. Data Min. ASA Data Sci. J., 17(1): e11662, 2024.

[11] I. Souiden, M. N. Omri, Z. Brahmi, "A survey of outlier detection in high dimensional data streams," Comput. Sci. Rev., 44: 100463, 2022.

[12] K. Yamanishi, J. Takeuchi, G. Williams, P. Milne, "On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms," Data Min. Knowl. Discov., 8(3): 275-300, 2004.

[13] F. Rollo, C. Bachechi, L. Po, "Anomaly detection and repairing for improving air quality monitoring," Sensors, 23(2): 640, 2023.

[14] C. Bachechi, F. Rollo, L. Po, "Detection and classification of sensor anomalies for simulating urban traffic scenarios," Clust. Comput., 25: 2793-2817, 2022.

[15] A. Shylendra, P. Shukla, S. Mukhopadhyay, S. Bhunia, A. R. Trivedi, "Low power unsupervised anomaly detection by nonparametric modeling of sensor statistics," IEEE Trans. Very Large Scale Integr. VLSI Syst., 28(8): 1833-1843, 2020.

[16] Y. Yang, C. Fan, L. Chen, H. Xiong, "IPMOD: An efficient outlier detection model for high-dimensional medical data streams," Expert Syst. Appl., 191: 116212, 2022.

[17] F. Angiulli, F. Fassetti, "Detecting distance-based outliers in streams of data," in Proc. ACM Conference on Information and Knowledge Management: 811-820, 2007.

[18] M. Kontaki, A. Gounaris, A. N. Papadopoulos, K. Tsichlas, Y. Manolopoulos, "Continuous monitoring of distance-based outliers over data streams," in Proc. IEEE 27th International Conference on Data Engineering: 135-146, 2011.

[19] S. Yoon, J. G. Lee, B. S. Lee, "NETS: extremely fast outlier detection from a data stream via set-based processing," in Proc. VLDB Endow., 12(11): 1303-1315, 2019.

[20] M. J. Bah, H. Wang, M. Hammad, F. Zeshan, H. Aljuaid, "An effective minimal probing approach with micro-cluster for distance-based outlier detection in data streams," IEEE Access, 7: 154922-154934, 2019.

[21] R. Zhu, X. Ji, D. Yu, Z. Tan, L. Zhao, J. Li, X. Xia, "KNN-based approximate outlier detection algorithm over IoT streaming data," IEEE Access, 8: 42749-42759, 2020.

[22] T. Toliopoulos, A. Gounaris, "Explainable distance-based outlier detection in data streams," IEEE Access, 10: 47921-47936, 2022.

[23] M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander, "LOF: identifying density-based local outliers," in Proc. ACM SIGMOD International Conference on Management of Data: 93-104, 2000.

[24] D. Pokrajac, A. Lazarevic, L. J. Latecki, "Incremental local outlier detection for data streams," in Proc. IEEE Symposium on Computational Intelligence and Data Mining: 504-515, 2007.

[25] G. S. Na, D. Kim, H. Yu, "DILOF: Effective and memory efficient local outlier detection in data streams," in Proc. ACM SIGKDD International Conference on Knowledge Discovery & Data Mining: 1993-2002, 2018.

[26] H. Yao, X. Fu, Y. Yang, O. Postolache, "An incremental local outlier detection method in the data stream," Appl. Sci., 8(8): 1248, 2018.

[27] O. Alghushairy, R. Alsini, X. Ma, T. Soule, "A genetic-based incremental local outlier factor algorithm for efficient data stream processing," in Proc. ACM International Conference on Compute and Data Analysis: 38-49, 2020.

[28] D. Barrish, J. Vuuren, "Enhancement of the local outlier factor algorithm for anomaly detection in time series," Easy Chair Preprint: 14238, 2024.

[29] D. Apoji, K. Soga, "Soil clustering and anomaly detection based on epbm data using principal component analysis and local outlier factor," in Proc. Geo-Risk 2023: 1-11, 2023.

[30] L. Chen, W. Wang, Y. Yang, "CELOF: Effective and fast memory efficient local outlier detection in high-dimensional data streams," Appl. Soft Comput., 102: 107079, 2021.

[31] L. Wan, W. K. Ng, X. H. Dang, P. S. Yu, K. Zhang, "Density-based clustering of data streams at multiple resolutions," ACM Trans. Knowl. Discov. Data, 3(3): 14, 2009.

[32] A. Bär, P. Casas, L. Golab, A. Finamore, "DBStream: An online aggregation, filtering and processing system for network traffic monitoring," in Proc. International Wireless Communications and Mobile Computing Conference (IWCMC): 611-616, 2014.

[33] N. A. Supardi, S. J. Abdulkadir, N. Aziz, "An evolutionary stream clustering technique for outlier detection," in Proc. International Conference on Computational Intelligence (ICCI): 299-304, 2020.

[34] C. Yin, S. Zhang, Z. Yin, J. Wang, "Anomaly detection model based on data stream clustering," Clust. Comput., 22(1): 1729-1738, 2019.

[35] E. Vanem, A. Brandsæter, "Unsupervised anomaly detection based on clustering methods and sensor data on a marine diesel engine," J. Mar. Eng. Technol., 20(4): 217-234, 2021.

[36] R. A. A. Habeeb, F. Nasaruddin, A. Gani, M. A. Amanullah, I. A. T. Hashem, E. Ahmed, M. Imran, "Clustering-based real-time anomaly detection—A breakthrough in big data technologies," Trans. Emerg. Telecommun. Technol., 33(8): e3647, 2022.

[37] X. Wang, M. M. Ahmed, M. N. Husen, H. Tao, Q. Zhao, "Dynamic micro-cluster-based streaming data clustering method for anomaly detection," in Proc. International Conference on Soft Computing in Data Science: 61-75, 2023.

[38] C. H. Park, "Outlier and anomaly pattern detection on data streams," J. Supercomput., 75(9): 6118-6128, 2019.

[39] K. Gokcesu, M. M. Neyshabouri, H. Gokcesu, S. S. Kozat, "Sequential outlier detection based on incremental decision trees," IEEE Trans. Signal Process., 67(4): 993-1005, 2019.

[40] T. Barbariol, F. D. Chiara, D. Marcato, G. A. Susto, "A review of tree-based approaches for anomaly detection," in Control Charts and Machine Learning for Anomaly Detection in Manufacturing, Springer, pp: 149-185, 2022.

[41] S. C. Tan, K. M. Ting, T. F. Liu, "Fast anomaly detection for streaming data," in Proc. International Joint Conference on Artificial Intelligence (IJCAI): 1511-1516, 2011.

[42] K. Wu, K. Zhang, W. Fan, A. Edwards, P. S. Yu, "RS-Forest: A rapid density estimator for streaming anomaly detection," in Proc. IEEE International Conference on Data Mining: 600-609, 2014.

[43] F. T. Liu, K. M. Ting, Z. H. Zhou, "Isolation-based anomaly detection," ACM Trans. Knowl. Discov. Data, 6(1): 3, 2012.

[44] Z. Ding, M. Fei, "an anomaly detection approach based on isolation forest algorithm for streaming data using sliding window," IFAC Proc., 46(20): 12-17, 2013.

[45] M. U. Togbe, Y. Chabchoub, A. Boly, M. Barry, R. Chiky, M. Bahri, "Anomalies detection using isolation in concept-drifting data streams," Computers, 10(1): 13, 2021.

[46] A. H. Madkour, A. Elsayed, H. Abdel-Kader, "Historical isolated forest for detecting and adaptation concept drifts in nonstationary data streaming," Int. J. Comput. Inf., 10(2): 16-27, 2023.

[47] G. Hannák, G. Horváth, A. Kádár, M. D. Szalai, "Bilateral-Weighted online adaptive isolation forest for anomaly detection in streaming data," Stat. Anal. Data Min. ASA Data Sci. J., 16(3): 215-223, 2023.

[48] Y. Liu, C. Liu, J. Li, Y. Sun, "Anomaly detection of streaming data based on deep learning," in Proc. International Conference on Internet of Things, Communication and Intelligent Technology: 459-465, 2024.

[49] M. E. A. Azz, A. Aljasmi, A, E. F. Seghrouchni, W. Benzarti, P. Chopin, F. Barbaresco, R. A. Zitar, "ADS-B data anomaly detection with machine learning methods," in Proc. International Workshop on Metrology for AeroSpace: 94-99, 2024.

[50] Y. Lee, C. Park, N. Kim, J. Ahn, J. Jeong, "LSTM-Autoencoder based anomaly detection using vibration data of wind turbines," Sensors, 24(9): 2833, 2024.

[51] J. Li, K. Malialis, M. M. Polycarpou, "Autoencoder-based Anomaly Detection in Streaming Data with Incremental Learning and Concept Drift Adaptation," in Proc. International Joint Conference on Neural Networks (IJCNN): 1-8, 2023.

[52] M. Molan, A. Borghesi, D. Cesarini, L. Benini, A. Bartolini, "RUAD: Unsupervised anomaly detection in HPC systems," Future Gener. Comput. Syst., 141: 542-554, 2023.

[53] M. Pourreza, B. Mohammadi, M. Khaki, S. Bouindour, H. Snoussi, M. Sabokrou, "G2D: Generate to detect anomaly," in Proc. IEEE Winter Conference on Applications of Computer Vision (WACV): 2002-2011, 2021.

[54] P. Jiao, T. Li, Y. Xie, Y. Wang, W. Wang, D. He, H. Wu, "Generative evolutionary anomaly detection in dynamic networks," IEEE Trans. Knowl. Data Eng., 35(12): 12234-12248, 2023.

[55] T. Yang, Y. Hu, Y. Li, W. Hu, Q. Pan, "A standardized ICS network data processing flow with generative model in anomaly detection," IEEE Access, 8: 4255-4264, 2020.

[56] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in Proc. IEEE International Conference on Image Processing (ICIP): 1577-1581, 2017.

[57] J. Wang, J. Liu, J. Pu, Q. Yang, Z. Miao, J. Gao, Y. Song, "An anomaly prediction framework for financial IT systems using hybrid machine learning methods," J. Ambient Intell. Humaniz. Comput., 14(11): 15277-15286, 2023.

[58] A. Srivastava, M. R. Bharti, "Hybrid machine learning model for anomaly detection in unlabelled data of wireless sensor networks," Wirel. Pers. Commun., 129(4): 2693-2710, 2023.

[59] W. Ullah, T. Hussain, F. U. M. Ullah, M. Y. Lee, S. W. Baik, "TransCNN: Hybrid CNN and transformer mechanism for surveillance anomaly detection," Eng. Appl. Artif. Intell., 123(A): 106173, 2023.

[60] Y. Karadayı, M. N. Aydin, A. S. Öğrenci, "A hybrid deep learning framework for unsupervised anomaly detection in multivariate spatio-temporal data," Appl. Sci., 10(15): 5191, 2020.

[61] M. U. Togbe et al., "Anomaly detection for data streams based on isolation forest using scikit-multiflow," in Proc. Computational Science and Its Applications (ICCSA): 15-30, 2020.

[62] Y. Yang, X. Yang, M. Heidari, M. A. Khan, G. Srivastava, M. R. Khosravi, L. Qi, "ASTREAM: Data-Stream-Driven scalable anomaly detection with accuracy guarantee in IIoT environment," IEEE Trans. Netw. Sci. Eng., 10(5): 3007-3016, 2022.

[63] Y. Yang, S. Ding, Y. Liu, S. Meng, X. Chi, R. Ma, C. Yan, "Fast wireless sensor for anomaly detection based on data stream in an edge-computing-enabled smart greenhouse," Digit. Commun. Netw., 8(4): 498-507, 2022.

[64] Q. Li, Z. Yu, H. Xu, B. Guo, "Human-machine interactive streaming anomaly detection by online self-adaptive forest," Front. Comput. Sci., 17(2): 172317, 2022.

[65] F. T. Liu, K. M. Ting, Z. H. Zhou, "Isolation forest," in Proc. IEEE International Conference on Data Mining: 413-422, 2008.

Biographies

Khadije Moeenfar was born in Bojnord, Iran. She received B.Sc. degree in Computer Science and M.Sc. degree in Computer Engineering from University of Bojnord, Bojnord, Iran, in 2015 and 2023, respectively. Her research interests include machine learning and data mining.

- Email: moeenfar2014@gmail.com
- ORCID: NA
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Vahid Kiani received M.S. and Ph.D. degrees in Computer Engineering from Ferdowsi University of Mashhad (FUM), Mashhad, Iran, in 2011 and 2016, respectively. In 2017, he joined University of Bojnord as an Assistant Professor in the Department of Computer Engineering. His research interests include machine learning, data mining, and digital image processing.

- Email: v.kiani@ub.ac.ir
- ORCID: 0000-0002-8248-9262
- Web of Science Researcher ID: AAD-4191-2019
- Scopus Author ID: 54973793600
- Homepage: https://ub.ac.ir/en/~v.kiani

Azadeh Soltani received the B.S., M.S, and Ph.D. degrees in Computer Engineering from Ferdowsi University of Mashhad, Mashhad, Iran, in 2001, 2004, and 2014, respectively. She was a lecturer at Azad University of Bojnord from 2004 to 2006. She is currently an assistant professor in the Department of Computer Engineering at University of Bojnord, Bojnord, Iran. Her current research interests include machine learning, data mining, and evolutionary algorithms.

- Email: a.soltani@ub.ac.ir
- ORCID: 0000-0003-3090-7992
- Web of Science Researcher ID: AAA-6000-2022
- Scopus Author ID: 14123895600
- Homepage: https://ub.ac.ir/~a.soltani

Rabeh Ravanifard received B.Sc., M.Sc. and Ph.D. in 2002, 2007 and 2020 from Isfahan University of Technology, Amirkabir University of Technology, and Isfahan University of Technology, respectively. She is currently an Assistant Professor of Computer Engineering at University of Bojnord. Her research interests include machine learning and soft computing.

- Email: ravanifard@ub.ac.ir
- ORCID: 0000-0002-9472-0011
- Web of Science Researcher ID: HSH-9074-2023
- Scopus Author ID: 57212561977
- Homepage: https://ub.ac.ir/~ravanifard

How to cite this paper:
K. Moeenfar, V. Kiani, A. Soltani, R. Ravanifard, "A fast and accurate tree-based approach for anomaly detection in streaming data," J. Electr. Comput. Eng. Innovations, 13(1): 209-224, 2025.

DOI: 10.22061/jecei.2024.11110.767

URL: https://jecei.sru.ac.ir/article_2228.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research Paper

A Robust Concurrent Multi-Agent Deep Reinforcement Learning based Stock Recommender System

S. Khonsha [1], *M. A. Sarram* [2], *R. Sheikhpour* [3,*]

[1] *Computer Engineering Department, Zarghan Branch, Islamic Azad University, Zarghan, Iran.*

[2] *Computer Engineering Department, Yazd University, Yazd, Iran.*

[3] *Department of Computer Engineering, Faculty of Engineering, Ardakan University, P.O. Box 184, Ardakan, Iran.*

Article Info

Article History:
Received 21 August 2024
Reviewed 06 October 2024
Revised 04 November 2024
Accepted 17 November 2024

Keywords:
Multi-agent
Concurrent learning
Deep reinforcement learning stock recommender system

*Corresponding Author's Email Address:
rsheikhpour@ardakan.ac.ir

Abstract

Background and Objectives: Stock recommender system (SRS) based on deep reinforcement learning (DRL) has garnered significant attention within the financial research community. A robust DRL agent aims to consistently allocate some amount of cash to the combination of high-risk and low-risk stocks with the ultimate objective of maximizing returns and balancing risk. However, existing DRL-based SRSs focus on one or, at most, two sequential trading agents that operate within the same or shared environment, and often make mistakes in volatile or variable market conditions. In this paper, a robust Concurrent Multiagent Deep Reinforcement Learning-based Stock Recommender System (CMSRS) is proposed.

Methods: The proposed system introduces a multi-layered architecture that includes feature extraction at the data layer to construct multiple trading environments, so that different feed DRL agents would robustly recommend assets for trading layer. The proposed CMSRS uses a variety of data sources, including Google stock trends, fundamental data and technical indicators along with historical price data, for the selection and recommendation suitable stocks to buy or sell concurrently by multiple agents. To optimize hyperparameters during the validation phase, we employ Sharpe ratio as a risk adjusted return measure. Additionally, we address liquidity requirements by defining a precise reward function that dynamically manages cash reserves. We also penalize the model for failing to maintain a reserve of cash.

Results: The empirical results on the real U.S. stock market data show the superiority of our CMSRS, especially in volatile markets and out-of-sample data.

Conclusion: The proposed CMSRS demonstrates significant advancements in stock recommendation by effectively leveraging multiple trading agents and diverse data sources. The empirical results underscore its robustness and superior performance, particularly in volatile market conditions. This multi-layered approach not only optimizes returns but also efficiently manages risks and liquidity, offering a compelling solution for dynamic and uncertain financial environments. Future work could further refine the model's adaptability to other market conditions and explore its applicability across different asset classes.

Introduction

Choosing the appropriate share for investment and accurately identifying the time to buy and sell shares is considered a challenging task. To become a professional stock trader and make successful transactions, investor must have significant experience and be able to recognize the trend of share price changes. Prediction tools for detecting market trends and forecasting stock

movements have been popular for several years. Various techniques and models are used to predict stock prices efficiently.

Linear regression [1] is introduced to predict continuous price values. Time series models such as the ARIMA (AutoRegressive Integrated Moving Average) [2] have also been proposed to model historical stock price data. Machine learning algorithms like LSTM (Long Short-Term Memory) [3], [4], RNN (Recurrent Neural Networks) [5] are also provided for stock price prediction and market trend detection. The variable, non-linear and fluctuating nature of the stock market has prevented the proposed models and algorithms from predicting well and being able to perform well in highly volatile markets and crashes. However, another class of techniques called reinforcement learning, which have worked well in computer games [6], [7] and performed as well as or better than humans, can be used to predict stock prices. Reinforcement learning in finance and stock trading involves training intelligent algorithms to make trading decisions by interacting with financial markets. These algorithms learn optimal strategies through trial and error, adapting to market dynamics to maximize returns and manage risks and the reinforcement agent is encouraged with any action that makes reaching the goal with more points, and is punished on the contrary. Accordingly, some researchers have designed trading strategies based on deep reinforcement algorithms [8]-[12], leveraging the power of neural networks to capture intricate market patterns and make informed decisions. These approaches aim to enhance portfolio management, risk assessment, and adaptive trading in the dynamic and complex landscape of financial markets.

Because the financial market is constantly changing and very complex, it is not convenient to learn the optimal trading policy using only an DRL agent [13]. So, in recent researches for automatic trading strategies [13]-[18], different multiagent deep reinforcement learning models have been used to extract features in order to display the environment observations of the reinforcement agent. One of the challenges that these systems face is how to accurately represent the agent's environment, which can give the agents a correct perspective for correct action. All the Multiagent SRS studies that have been done, trading agents use shared data source for learning. In the complex and volatile stock market environment, various distributed and decentralized data sources reflect market changes from different perspectives. A key challenge is obtaining the temporal characteristics of these data types and feeding them into agents differently to provide a deeper understanding of the stock market environment. In order to mitigate this challenge, we use various data sources such as Google stock trends and fundamental data and technical indicators along with historical price data to select and recommend suitable stocks to buy or sell by multiple agents concurrently.

Besides that, due to the different behavior of distinct stocks in financial markets, the presented approaches face unsolved challenges yet. Considerable, all parts of the RL environment only reproduce common historical price data to train trading agents for all assets which makes the efficiency of the algorithm not acceptable for some out-of-sample stocks. While some stocks have fundamental behavior, some are price driven and some of them follow the overall movement of the market. To solve the presented challenge, nonidentical from the research done, we hypothesize that the treatment of stock selection for buy or sell trade specially in the time of volatile market in the form of the stock-based feature selection and learning the trading behavior of each stock independently and the cooperation of agents in choosing the final decision is useful to make robust profitable trading decisions.

Another challenge that automated trading systems face is that in highly volatile markets, they face a lack of liquidity to reduce the average share purchase price. Therefore, we define the novel reward function in such a way that the agent always has adequate liquidity in order to avoid excessive losses in fluctuating markets. The summarized contributions of this paper are as follows:

- We proposed a Concurrent Multiagent Stock Recommender System (CMSRS) to generate collaborative recommendations.
- We used diverse data to co-train multiple concurrent DRLs to robustly detect market trends.
- For different stocks, the Concurrent RL trading agents have a custom-built environment for training. More precisely, effective features are extracted for different stocks using the dvlw state formation, and RL agents are trained using these features as states separately and update shared policy.
- To mitigate losses in bear markets, we defined a novel liquidity-based reward function. This reward function gives points to the reinforcement agent based on the current amount of cash so that the agent can always maintain cash at a suitable level.

The organization of this article is divided as follows: Next section provides a bibliometric-based review of previous studies on DRL and Multiagent RL trading which was extracted on August 20, 2024. Then we present the preliminaries and background concepts needed to define the RL framework for recommender systems as well as the classification of different reinforcement learning algorithms. After that, we describe the proposed method to solve the raised challenges, including the complete Multiagent framework based on multi-layer

recommender structure optimization and the trading agent training process.

Final Section shows the experimental results, including the parameters optimization results for different stocks and the results of the tests and methods applied to compare the performance of the algorithms and presents the trading performance of a trained CMSRS in a real environment. Finally, we concludes the paper with conclusions from this study and provides directions for future research.

Related Work

In recent years, recommender systems [19]-[21] and reinforcement learning (RL) methods have experienced significant advancements, leading to widespread adoption across various complex problem domains. RL, in particular, has led to an increase in the adoption of its algorithms to solve many problems, even those that seemed difficult to solve in the past. In the field of stock trading, these methods have recently received more attention.

Fig. 1 demonstrates annual scientific production trend in this field.

Fig. 1 shows since 2018, researchers have paid more attention to the use of reinforcement learning algorithms in stock trading.

Especially, the attention of individual and institutional investors and financial researchers has also been drawn to DRL algorithms. Many investors are looking for algorithms that can provide reliable investment recommendations by taking into account the turbulent, changing and dynamic conditions of financial markets and considering all aspects affecting these markets. Table 1 shows a systematic comparative review using bibliometrics on the research done in the field of stock trading using RL techniques and Multiagent RL (MARL). According to Table 1, the analysis from 1998-2024 shows that 60 out of 264 RL-related documents contain the keyword Multiagent (from 2002-2024). These documents cover various applications, including but not limited to Stock Recommender Systems.

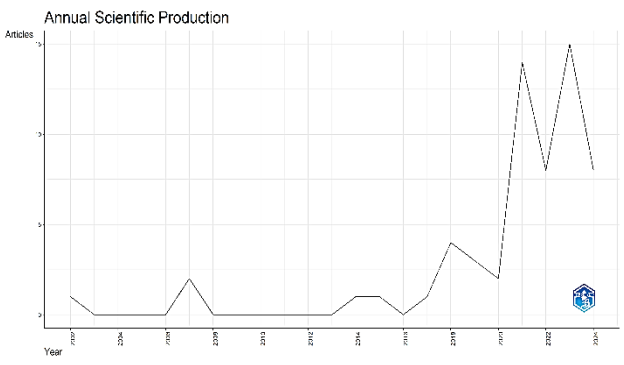


Fig. 1: Annual scientific production.

Table 1 indicates that no book chapters on Multiagent topics have been published. Additionally, only one conference review and one journal review on techniques related to Multiagent Reinforcement Learning (MARL) have been published.

Table 2 summarizes recent research related to 'reinforcement learning in stock trading' using specific keywords.

Table 1: Descriptive statistics of the studies conducted in 1998 to 2024: Reinforcement learning stock trading systematic review

Description	RL Results	MARL Results
MAIN INFORMATION ABOUT DATA		
Timespan	1998:2024	2002:2024
Sources (Journals, Books, etc)	148	44
Documents	264	60
Annual Growth Rate %	9.19	9.91
Document Average Age	5.02	4.06
Average citations per doc	13.7	14.53
References	6781	1132
DOCUMENT CONTENTS		
Keywords Plus (ID)	1293	268
Author's Keywords (DE)	470	102
DOCUMENT TYPES		
Article	99	14
Book chapter	8	0
Conference paper	125	20
Conference review	7	1
Review	5	1

In Fig. 2, the network between the top research sources, researcher countries and keywords are presented, the left side is cited sources, the middle is the country names and on the left side the keywords are specified.

Table 2: Very recent research with keywords “reinforcement learning stock trading”

Reference	Journal		Remarks
[22]	Expert systems with applications (2024)	Contributions	A Cascaded LSTM (CLSTM-PPO) model is utilized. Initially, LSTM is applied to extract time-series features from daily stock data. Additionally, another LSTM model is employed within the RL strategy functions for further training.
		Disadvantages	Instabilities during training.
[18]	Information Sciences (2023)	Contributions	An RRL algorithmic trading model by using self-attention to extract hidden temporal representation of series with hybrid loss is introduced.
		Disadvantages	High computational complexity due to sequential training of model
[22]	Knowledge based systems (2023)	Contributions	DRL-UTrans model is proposed that uses architecture of U-Net and transformer layers combined to RL for trading of single stock.
		Disadvantages	It does not support multi-stock trading and portfolio construction
[24]	Applied soft computing (2023)	Contributions	A multi-agent model is introduced that multiple generative adversarial networks cooperate to regenerate historical price of stocks to resolve generalization issue in stock trading
		Disadvantages	It does not support multi-stock trading and portfolio construction
[13]	Neural Computing and Applications (2023)	Contributions	A Multi-agent DRL is proposed that formulate trend consistency factor into reward function as a regularization term for portfolio construction
		Disadvantages	Accurate trend consistency/inconsistency calculation is a challenge
[25]	Advances in Transdisciplinary Engineering (2022)	Contributions	Three actor critic RL models (SAC, TD3, A2C) is employed to construct an ensemble strategy to automate stock trading
		Disadvantages	Unstable result to choose an agent with best Sharpe ratio
[26]	Expert systems with applications (2022)	Contributions	A deep reinforcement learning model for asset-specific trading rules is investigated that uses different feature extraction modules
		Disadvantages	It does not support multi-stock trading and portfolio construction
[27]	Expert systems with applications (2022)	Contributions	The ResNet-LSTM actor model for crypto currency trading rules investigated that uses ResNet architecture
		Disadvantages	It does not use reinforcement learning but use classification approach

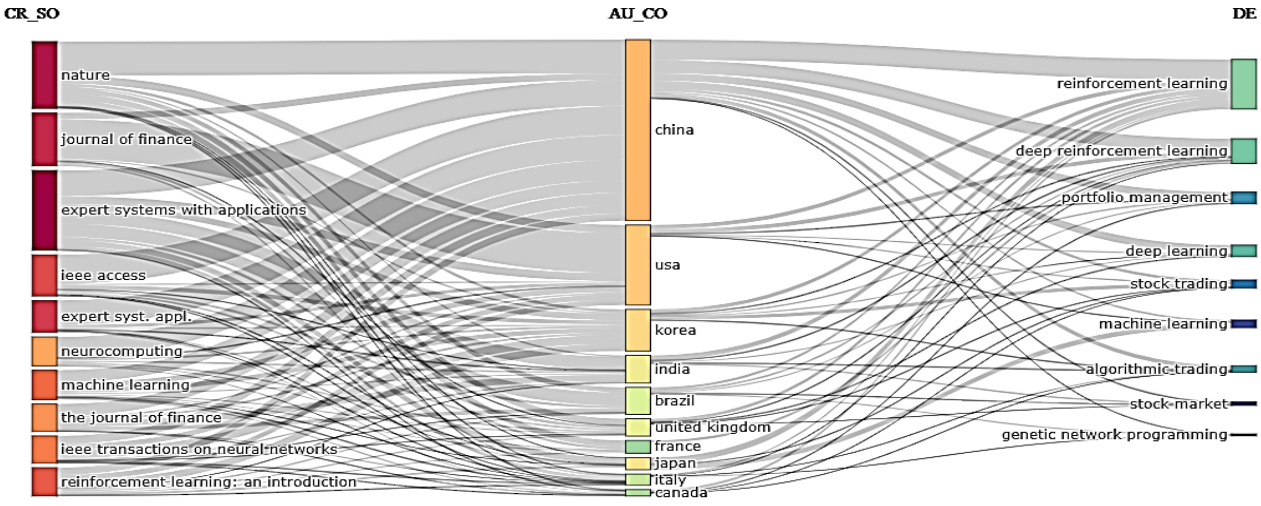


Fig. 2: Cited Sources (CR_SO), Countries (AU_CO) and keywords (DE).

Fig. 3 and Table 3 show word cloud of the most relevant words used in documents.

Fig. 3: World cloud.

Table 3: Most frequent words

Terms	Frequency
reinforcement learning	200
Commerce	164
financial markets	140
electronic trading	115
deep learning	97
Investments	84
learning systems	65
learning algorithms	56
reinforcement learnings	50
Profitability	45
trading strategies	42
stock trading	40
decision making	35
algorithmic trading	24
portfolio managements	23

Preliminaries and Backgrounds

A. Single Agent Reinforcement Learning

In decision-making problems, the Markov decision process (MDP) serves as a framework where outcomes are partially random and partially influenced by the decision maker. MDPs are commonly used to describe the environment in RL, with RL models being a type of state-based model that leverages MDPs. In essence, RL involves training an agent through a system of rewards and punishments. The RL agent observes the current state, performs an action in the environment, receives a reward for that action, and this action transitions the environment to the next state. Fig. 4 schematically shows MDP process in RL.

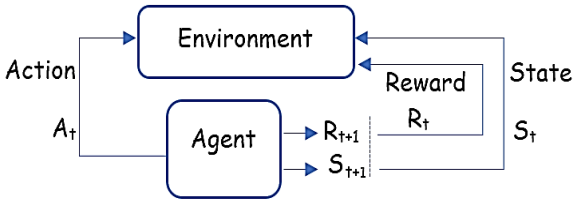


Fig. 4: MDP process in RL [28].

In a trading Reinforcement Learning algorithm:

- At time t, the agent (reinforcement algorithm) assesses the current state (s_t) of the environment, which encompasses various factors such as cash balance, stock prices in the portfolio, the quantity of each share, the time since a share was purchased, technical indicators, fundamental parameters of the share, the share board details (including the number of individual and institutional buyers and sellers), the volume of shares bought by individual and institutional traders, and other relevant features that define the current state of the environment..
- The agent selects the optimal action (a_t) from the available options (buy/sell/hold).
- The environment transitions to a new state (s_{t+1}).
- The environment generates a reward (r_t), which reflects the change in the portfolio's value (increase/decrease/no change).

The process of selecting an action based on the current state is governed by the policy function ($\pi(s_t) = a_t$), which maps states to actions. In reinforcement learning, the system always consists of an environment with a set of states, actions, a policy function (which guides transitions between states), and rewards expressed as numerical values. The reinforcement learning agent continuously observes the current state, uses the policy function to determine the best action, and receives a reward for that action.

This cycle repeats until an end state is reached. The agent's goal is to maximize the reward, with an optimal policy being one that achieves the highest possible reward. Reinforcement learning algorithms vary widely, and their classification in Fig. 5 is based on the specific components they employ to construct the workflow outlined in Fig. 4, ultimately aiming to achieve the optimal policy. In recent years, multi-agent models [13]-[17], [24], ensemble models [8], [25], [29], and models incorporating autoencoders [30] have also been introduced for portfolio optimization.

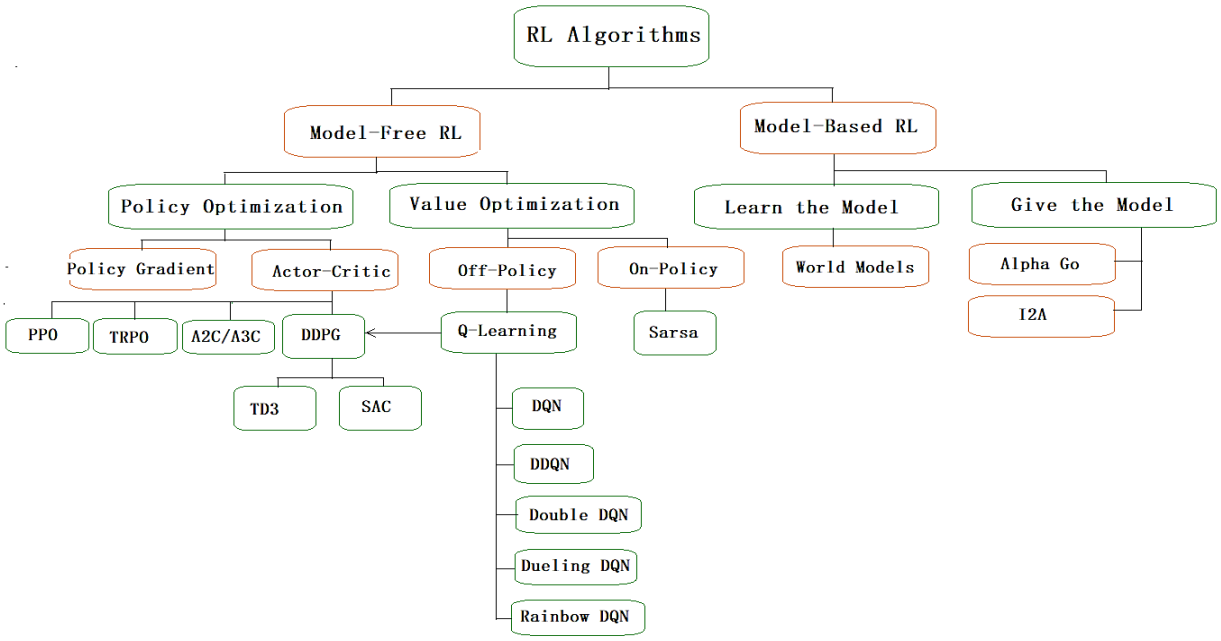


Fig. 5: RL Algorithms [34].

The workflow of all reinforcement learning algorithms typically includes the following steps:

1. Initialize the policy (π) with random parameters.
2. Using the current policy, select the action (a) with the highest probability and store the obtained reward (r) along with the states before (s) and after (s_(t+1)) the action in the experience memory (D).
3. Choose a model to refine the policy.
4. Repeat step 2 to gather more experience with the improved policy and continue refining the policy.

In other words, a common approach to finding an optimal policy that maximizes the expected cumulative discounted reward for each state is policy iteration. This method is particularly useful when faced with multiple options, each with its own distinct rewards and risks. Policy iteration involves a two-stage process that alternates between policy evaluation and policy improvement.

In the policy evaluation stage, we intend to find the exact value function for our current policy. To achieve this goal, we iteratively apply the Bellman equation defined as (1) until we reach convergence.

$$\mathcal{V}_{\pi}(\mathcal{s}) = \sum_{\mathcal{s}',r} \mathrm{p}\,(\mathcal{s}', \mathrm{r} | \mathcal{s}, \pi(\mathcal{s}))[\,\mathrm{r} + \gamma \mathcal{V}_{\pi}(\mathcal{s}')] \qquad (1)$$

where $\mathcal{s}'$ represents the next state, and $\pi(\mathcal{s})$ is the action taken from state $\mathcal{s}$ under policy π. The transition probability, denoted as p, is the likelihood of moving from state $\mathcal{s}$ to the next state $\mathcal{s}'$ when performing action $\pi(\mathcal{s})$ and receiving reward r. The discount factor $\gamma \in [0,1]$ accounts for the time value of rewards.

In essence, rewards may not be immediately received by the agent. Early rewards are generally more predictable and likely, so they are prioritized over potential long-term rewards. In sequences, even larger rewards are discounted if they are further in the future, as the agent is uncertain about receiving them. The discount factor (γ) is used to adjust the value of future rewards. A higher γ means that the agent places more importance on long-term rewards, while a lower γ indicates a greater focus on short-term rewards.

In the policy improvement stage, as shown in (2), the process involves repeatedly applying the Bellman optimality operator.

$$\pi'(\mathcal{s}) = \mathrm{argmax}_{a} \sum_{\mathcal{s}',r} \mathrm{p}\,(\mathcal{s}', \mathrm{r} \mid \mathcal{s}, a)[\,\mathrm{r} + \gamma \mathcal{V}_{\pi}(\mathcal{s}')] \qquad (2)$$

Similarly, the value of choosing action a in state $\mathcal{s}$ under policy π is denoted as Q_{π}(s,a). This represents the expected cumulative reward of taking action a in state $\mathcal{s}$, with all subsequent actions being determined by the policy π, as expressed in (3).

$$\mathrm{Q}_{\pi}(\mathcal{s}, a) = \sum_{\mathcal{s}',r} p\,(\mathcal{s}', r | \mathcal{s}, a)[\,r + \gamma \mathrm{Q}_{\pi}(\mathcal{s}', a')] \qquad (3)$$

The Q_{π} is called the value-action function for the policy π. The value function of $\mathcal{V}_{\pi}$ and Q_{π} can be estimated with repetitive experiments. For example, if an agent follows a policy and averages the amount it receives from experiences for each situation, after an infinite number of repetitions, the value of $\mathcal{V}_{\pi}(\mathcal{s})$ will converge to the real value. Now, if this average is kept for each state-action pair separately, then $\mathrm{Q}_{\pi}(\mathcal{s}, a)$ will be estimated and stored in the table. Such estimation methods are called Monte Carlo methods, which include averaging over a large number of random samples of the real return reward.

For complex and dynamic problems like stock trading and portfolio optimization, which involve high-

dimensional and continuous state-action spaces, finding the exact optimal solution using lookup table-based methods is often impractical. Instead, rough approximators such as neural networks are used. A neural network comprises several layers, where the input layer receives the state vector s, and the output layer determines the action a.

Fig. 1 illustrates the training process of an agent based on a Q-Network with experience replay memory. This architecture consists of three main components:

- Q-network Q(s, a;θ) where θ determines the agent's behavioral policy,
- Q-target network Q(s', a' ;θ′), which is used to obtain the Q values for the error part of the Deep Q-Network (DQN) and
- Experience Replay Memory, which the agent uses to randomly transfer samples to train the Q network.

The replay memory is used to address the issue of high correlation between consecutive examples in the problem, which can slow down convergence when used for training a neural network. To mitigate this, transitions—comprising the state, action, resulting next state, and associated reward—are stored in a replay memory. These transitions are then randomly sampled from the memory for training the network. By doing so, the network can learn from a more diverse set of experiences, reducing the impact of correlation and improving the stability of the learning process. Additionally, because these experiences are valuable, the replay memory allows them to be reused multiple times for more efficient training. The target network, which shares the same structure as the main network, is periodically updated by copying the weights from the main network to the target network θ' after a fixed number of steps. This approach helps to reduce the negative effects of network fluctuations, leading to more stable training and faster convergence.

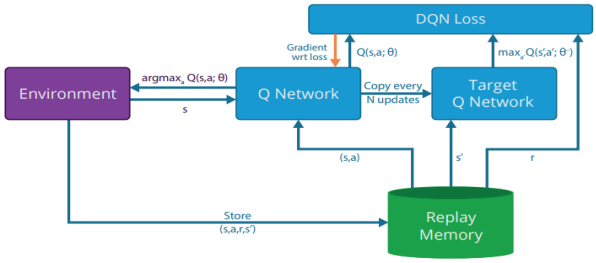


Fig. 6: DQN architecture [31].

B. Multi Agent Reinforcement Learning

In MARL, multiple agents interact and learn from each other in order to better coordinate their actions in the environment to maximize target long-term reward [30]. This coordination is achieved through a process known as cooperative learning, where agents share their experiences with each other and learn from each other's experiences. This allows agents to learn from each other and improve their policies. Specially in the case of the RL agent that uses the neural network approximator, the direct use of single-agent methods in multi-agent frameworks violates the Markov assumptions required for convergence because other agents are considered as part of the environment, and the environment from the perspective of each agent seems to be non-stationary [32]. As described in [33], if we know the actions that are taken according to the local observations of each agent, even with changing policies, the environment has the property of being stationary.

Algorithm 1 Multiagent RL with N **Agent**

Create experience memory D with M size
Create Q function by θ random weights
Create $\hat{Q}$ target function by $\theta' = \theta$ weights
for episode from 1 to MaxEpisode do
  Initialize a random process N for action exploration
  Receive sequence s_1 and preprocessed sequence $\phi_1 = \phi(s_1)$
  for t from 1 to MaxTimeStep do
    $\forall\, agent\ i \in \{1, \dots, N\}\ a^i_t = \mu^i_\theta(o^i) + N_t$
    Execute all N actions a_t and observe reward r_t and state s_{t+1}
    Set $s_{t+1} = s_t$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$
    Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in D
    for agent i = 1 to N do
      Sample random mini-batch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$from D
      $y_j = r^i_j + \gamma Q(\phi(s_t), a'; \theta')$
      Set $y_j = \begin{cases} r_j, & for\ terminal\ \phi_{j+1} \\ r_j + \gamma \max_{a'} Q(\phi(s_t), a'; \theta'), & for\ non-terminal\ \phi_{j+1} \end{cases}$
      Do a gradient descent step on $(y_j - Q(\phi_j), a_j; \theta))^2$ respect to the θ parameters
      Each C timesteps assign $\hat{Q}$ = Q
    end for
end for

Mathematically, MARL is an MDPs generalization for multi-agent reinforcement learning and can be defined as $(N, S, A_{1:N}, T, R_{1:N}, \gamma)$ tuple, where N is the number of RL agents, S denotes states set, $A_{1:N}$ indicates actions of N agents, T describe probability transition function from states and actions to [0,1] and $R_{1:N}$ is average reward received by N agents, depending on the type of multi-agent MDP, the reward function of the agents can be the same or different. In the multi-agent case, each agent i, in the t-th iteration, only updates the value of $Q(s_t,i,a_t,i)$ and leaves the other entries to the Q function unchanged. Algorithm 1 shows the multi-agent learning process in details.

Proposed Concurrent MultiAgent Stock Recommender System

We propose a multi-layer multi-agent stock recommender system based on deep reinforcement algorithm. In the proposed multi-layer CMSRS, we propose centralized value function estimator and decentralized policy networks of RL agents to diminish explained non-stationary issue and stabilize RL agent training in the DRL layer. The CMSRS architecture consists of four distinct layers, each serving a specific purpose in generating stock recommendations for users. These layers typically include:

Data Layer: This layer involves gathering and aggregating various data sources and extracting relevant features from the preprocessed data.

Environment Layer: In this layer, the new risk aversion reward function is proposed to reduce asset variance and decrease maximum percentage loss.

DRL Layer: This layer includes Multi agent DRL model that concurrently train on the environment to reach optimal policy.

Trading Layer: Finally, the contributions made in the above three layers will lead to more robust recommendations to profitable stock trading in the trading layer.

In the data layer, preprocessing and feature extraction is done on the various data sources such as Google stock trends [34], [35] and fundamental data and technical indicators along with historical price data to construct multiple trading environments, so that different feed DRL agents in the DRL layer can have different observations. Fig. 7 shows the proposed multi-layer CMSRS system. The details of each layer are explained in the following subsections.

A. Data Layer

The data layer in a stock recommender system plays a crucial role in gathering and preparing the various data sources required to make informed recommendations. In the proposed architecture, various data sources including Google Trends, fundamental data, and historical daily stock price data (OHLCV) augmented with technical indicators (MACD, RSI, CCI and IDX) is used to form each agent observation separately.

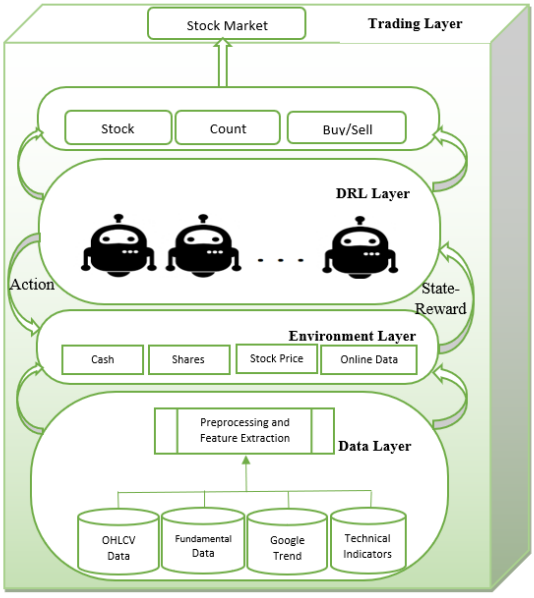


Fig. 7: Proposed CMSRS architecture.

Google trends as a proxy for market sentiment analysis, as analyzed in [35] can improve the Sharpe ratio of trading. Therefore, this feature has been used to feed agents. A normalization preprocess and missing data dealing along with feature extraction is done on the raw data in the data layer. The data layer includes the following tasks in details:

OHLCV Data: This refers to historical price and trading volume data for stocks. It includes the opening, highest, lowest, and closing prices of a stock on a specific day, as well as the trading volume. This data provides insights into price trends, volatility, and trading activity over time. The data layer collects and preprocesses this information, ensuring it is clean, consistent, and ready for analysis.

Fundamental Data: Fundamental indicators are key financial metrics that provide insights into a company's financial health and performance. A fundamental risk aversion indicator could be derived from metrics such as debt-to-equity ratio, earnings per share (EPS), and other relevant financial ratios. This indicator helps assess the financial stability and risk profile of a company. The data layer gathers these fundamental indicators for the stocks under consideration.

Google Trends Data: Google Trends provides information about the popularity of search terms over time. In the context of a stock recommender system, Google Trends data can be used to gauge public interest and sentiment towards specific stocks or sectors. The data layer collects Google Trends data related to search terms relevant to the stocks being analyzed.

The data collected from these sources is often in disparate formats and may require preprocessing to align timestamps, handle missing values, and normalize the data. Once prepared, the data can be integrated into a unified dataset for further analysis.

Overall, the role of data layer in a stock recommender system involves collecting, preprocessing, and integrating diverse data sources to create a comprehensive dataset that captures both historical market trends and external factors affecting stock performance. This integrated dataset serves as the foundation for building CMSRS models that take into account various dimensions of stock behavior and market sentiment.

B. Environment Layer

As mentioned in the preliminaries section, in RL-based learning, the trader agent comes to gain experience by interacting with the market environment through trial-and-error procedure to maximize the reward function. The data on which the agent's observations rely the sensors that provide input to the deep reinforcement algorithm. We assume that the quality and quantity of this data is effective on the amount of reward that the agent can achieve. In addition, the way of rewarding the reinforcement agent is very effective in the convergence of the model. In this layer, the precise definition of the state construction and reward function is proposed, which is explained in detail below.

State Formulation

Defining the state structure in complex environments such as stock trading needs some expertise information. According to our latest information, other researches have used the simple structure of the time window to construct the state. We use novel multi-source n-dimensional vector to represent state. First, we define the *difference vector* in the *t*th timestep for *f*th feature of data, dv_t^f, as the element-wise subtraction of d_{t-1}^f from d_t^f: $dv_t^f := d_t^f \ominus d_{t-1}^f = (d_t^f - d_{t-1}^f)$. Then, *dv* is calculated for a rolling lookback window (*w*) that is selected to the current time and automatically shifts forward with the timesteps, Fig. 8 shows $state_t$ formation for one window. This valuable information is used to construct observations of agents. In experiments this process called difference vector lookback window (*dvlw*) and compared versus simple lookback window (slw).

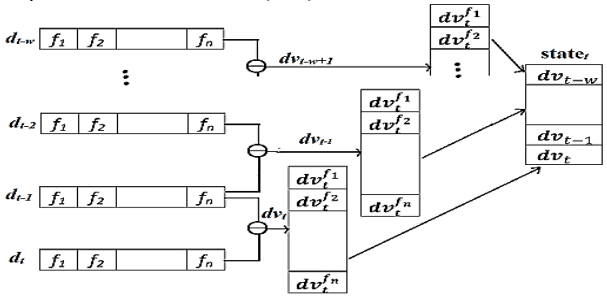


Fig. 8: State formation.

Risk Aversion Reward Function

Designing a reward function for a multi-agent RL that aims to maximize returns, avoid risks, and reduce maximum drawdowns is a complex task that requires careful consideration of various factors. The common definition of the reward in trading agents is the amount of change in the value of the portfolio after the execution of the action that is not risk averse. But in practice, a trader does not prefer his capital balance to be unstable. In other words, high profit along with high loss is not desirable for investors. To model this risk aversion and comprise liquidity requirement, we define precise reward function to dynamically manage cash reserves and penalize the model for not maintaining a reserve of cash. We add a penalty term to the reward function, which aims to reduce the capital variance. This new function enables the model to execute transactions with high confidence and manage cash reserves. Accordingly, we propose the following reward function and compare its effect empirically in real experiments. The immediate reward of ith agent at timestep t, after executing action a_t^i (Sell/Buy shares from jth stock) in state s_t^i and transition to state $s_{t+1}^{'i}$ defined by:

$$r_t^i(s_t^i, a_t^i, s_{t+1}^i) = \left(\mathcal{C}_{t+1}^i + h_{t+1,j}^i * p_{t+1,j}^i\right) - \left(\mathcal{C}_t^i + h_{t,j}^i * p_{t,j}^i\right) - \mathcal{P}_{t+1}^i - \mathcal{C}_0^i \quad (4)$$

where $\mathcal{C}$ denotes cash value and $\mathcal{P}_{t+1}^i$ is cash penalty term as:

$$\mathcal{P}_{t+1}^i = MAX\left(0, \sum\left(\mathcal{C}_t^i + h_{t,j}^i * p_{t,j}^i\right) * \mathcal{P} - \mathcal{C}_t^i\right) \quad (5)$$

where $\mathcal{P}$ is a hyperparameter which determines the liquidity percentage of the portfolio. The cooperative goal of the agents in CMSRS is to maximize the team average cumulative discounted reward obtained by all agents:

$$Q_\pi(s_t, a_t) = \mathbb{E}_{s_{t+1}}[r_t(s_t, a_t, s_{t+1}) + \gamma \mathbb{E}_{a_{t+1\sim\pi(s_{t+1})}}[Q_\pi(s_{t+1}, a_{t+1})] \quad (6)$$

C. DRL Layer

In the DRL layer, we use concurrent multiprocessing training via various observations of the local environment to improve the performance of DRL trading agents. Each trading agent *i* interacts with a market environment to produce transitions independently in the form of $\{s_i, a_i, r_i, s'_i\}$ that respectively are state, action, reward and next state. Then, collection of experience transitions from all the RL agents are stored in a shared replay memory to update a learner. Fig. 9 demonstrates the policy optimization process of the proposed CMSRS. the optimal policy of DRL model is learned by using gradient descent on the loss function: $\mathcal{L}_\theta = \mathbb{E}[(\mathcal{Y}_i - Q(\phi_i, a_i; \theta))^2]$, where θ is policy network`s parameter, ϕ_i is the preprocessed state and $\mathcal{Y}_i = r_i + \gamma \max_{a'} \widehat{Q}(\phi(s_i), a'; \theta')$ is target value.

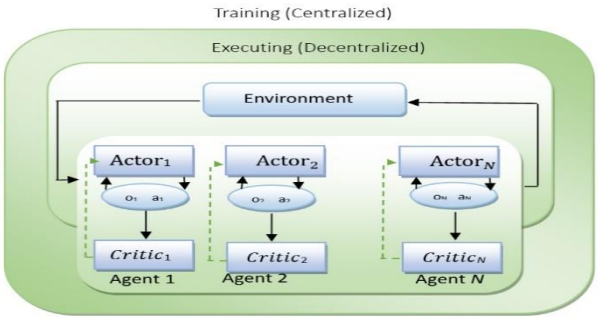


Fig. 9: Multiagent architecture of DRL layer.

So far, several methods have been proposed for RL with neural network approximators, including those based on policy gradients. Proximal Policy Optimization (PPO) [36] has been shown to provide better stability among other RL algorithms. PPO, as the name suggests, seeks to find a proximal policy that uses advantage function ($\mathcal{A}$) as the difference between the future discounted sum of rewards on a certain state and action, and the value function of that policy and thus avoids large policies update. Let the ratio $\mathcal{R}(\theta) = \frac{\pi_\theta'(a_t|s_t)}{\pi_\theta(a_t|s_t)}$, loss function of PPO is:

$$\mathcal{L}_\theta^c = \mathbb{E}[(\min(\mathcal{R}(\theta)\,\mathcal{A}, c\,(\mathcal{R}(\theta),1-\epsilon,1+\epsilon)\,\mathcal{A})] \quad (7)$$

where c denotes clipping operator and ϵ is the bound threshold hyperparameter.

We investigate the concurrent Multiagent PPO algorithm to learn a shared decentralized policy by leveraging team experience from all the PPO agents. At the first of training phase, the parameters of policy of all agents are set to an initial value. Then, for each episode, *N* agents in each timestep sample an action (Buy/Sell) using own deep neural network. After that, the agent executes the action in the trading environment and observes the reward and transfers to the next state. All of team experiences store in the experience replay memory. After collecting samples for an episode, *M* epochs of updating are performed with a small batch of transitions sampled from memory *D* on the loss function of (7) using SGD (stochastic gradient descent). In this architecture, all agents work cooperatively as a team to maximize the team-average cumulative discounted reward.

D. Trading *Layer*

The trading layer in a stock recommender system plays a pivotal role in executing the recommendations provided by a DRL agent. This layer bridges the gap between the recommendations generated by the DRL agent and the actual execution of trades in the financial market. The motivation to create a very robust trading system is achieved by cooperating with several robust models to maximize the cumulative reward and let them trade based on the DRL layer output. Here is an explanation of the key functions of the trading layer:

1. Trade Execution: Once the DRL agent generates stock recommendations based on its learned policy, the trading layer is responsible for executing these recommendations. It converts the agent recommendations into actionable buy or sell orders in the market.

2. Risk Management: The trading layer incorporates risk management strategies to control potential losses. This involves setting limits on the size of trades, diversification across different stocks or asset classes, and implementing stop-loss and take-profit mechanisms to manage trade outcomes.

3. Transaction Costs: The trading layer takes transaction costs into account, including brokerage fees, taxes, and spreads. It aims to optimize trade execution to minimize these costs and enhance overall trading performance.

Experiments

All the experiments are carried out on a computer having a 16 GB RAM with CPU Intel Core i7-10750H and GPU Nvidia GeForce GTX 1080 8 GB dedicated memory, 80 GB of virtual memory has been used to optimize the parameters. Data collection consists of three parts. Historical price data, fundamental data and historical data of Google Trends.

In all experiments, the initial amount of cash balance is 10,000$. We incorporate the transaction cost to reflect market friction, e.g., 0.1% of each buy or sell trade. To control risk during market crash situations the volatility index (VIX) is used that is a real-time U.S. stock market index representing the market's expectations for volatility over the coming 30 days.

In our experiments, we select seven most active stocks from United States stock market due to the high market liquidity, including TSLA, AAPL, AMZN, MSFT, GOOG, META and IBM to evaluate the proposed CMSRS. The first six (META, GOOG, TSLA, MSFT, AAPL, AMZN) have been used for training and evaluation and generalization testing, the last one (IBM) not utilized in training, used to test the robustness of the model and its efficiency. The time period used is from January 1, 2013 to July 1, 2023. One last year, from August 1, 2022 to August 1, 2023, has been used for the trading phase Fig. 10, shows the price plot of six train datasets. We use the following widely used metrics in both research and practice to evaluate the proposed CMSRS:

- Cumulative Return (CR): reflects the overall effect of the trading strategy in a certain period of time
- Sharpe Ratio (SR): returns the earned per unit of volatility, which is a widely used measure of an investment performance.
- Maximum DrawDown (MDD): shows the maximum percentage loss during the trading period.

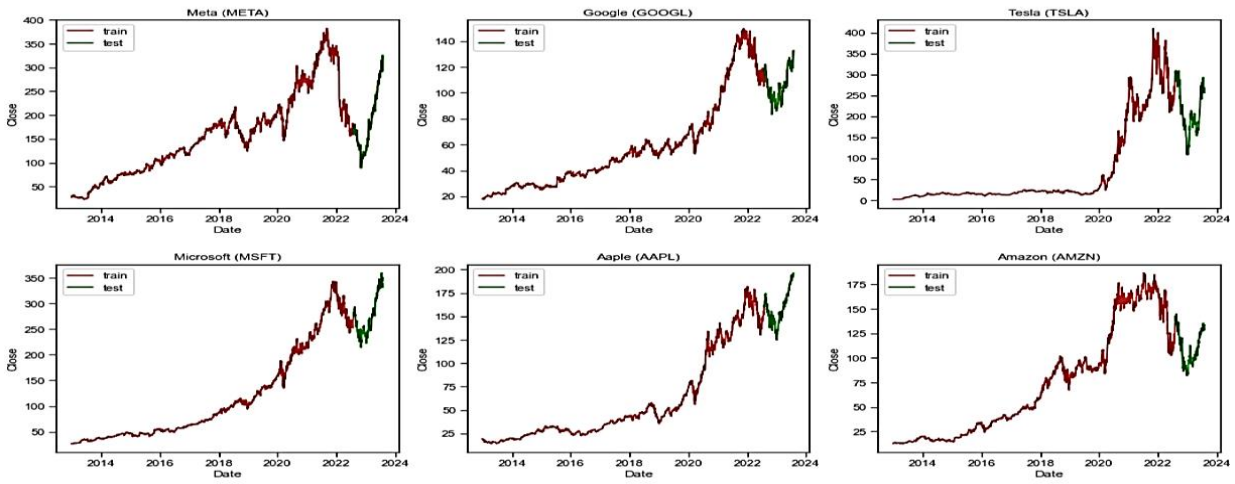


Fig. 10: Stock prices, the training set (red color): January 1, 2013 to August 1, 2022 and trading set (green color): from July 1, 2022 to August 1, 2023- From top left to bottom right: META, GOOGL, TSLA, MSFT, AAPL and AMZN.

Profit and Loss (P&L): presents the amount of profit or loss of the algorithm in the desired time period.

Reinforcement learning algorithms are very sensitive to hyperparameter values, and one of the most time-consuming processes of reinforcement learning is the optimization phase of hyperparameters. To optimize the parameters, we must define a search space in which the valid values of the hyperparameters are specified. Values can be sampled in two ways: normal distribution and uniform distribution.

The selection of hyperparameters can be done both randomly and in a grid manner. We perform Bayesian optimization algorithm for search and use Sharpe ratio as risk adjusted return measure for hyperparameter optimization in validation phase. Fig. 11 shows the effect of the number of episodes on the convergence of the reinforcement agent in the training phase. The convergence of the algorithm is clearly seen in episode 10,000, but in episode 500, the output of the agent's reward fluctuates a lot.

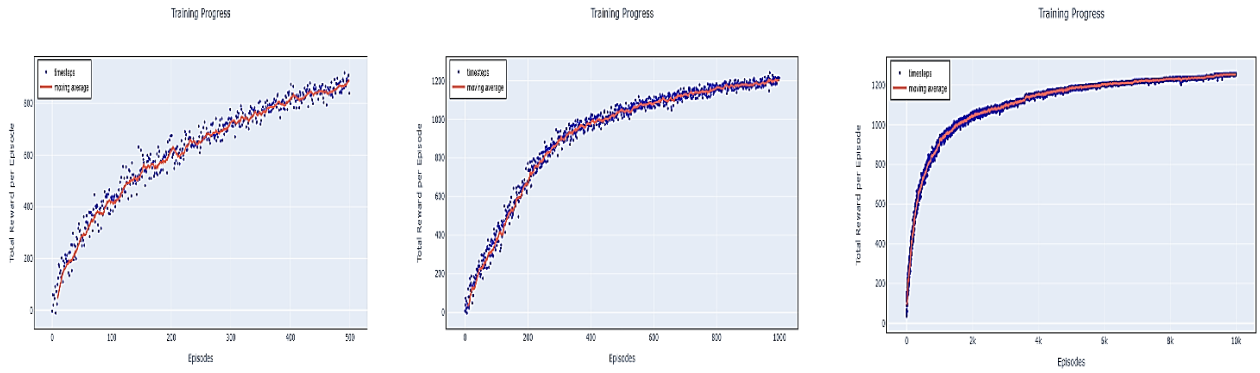


Fig. 11: Effect of the number of episodes on the convergence of the reinforcement agents in the training phase (Left to right: 500, 1000 and 10,000 episodes).

Results and Discussion

A. Results of the Proposed State Formulation

To evaluate the impact of proposed state formation on the convergence speed and the returned reward, the experiment setup involving three agents with random stocks is implemented.

The result as shown in Fig. 12 confirms that the proposed state construction process significantly converges to higher rewards in less time.

Fig. 12: Training reward using proposed difference vector lookback window (dvlw) and simple lookback window(slw), w=32.

B. Results of the Risk Aversion Reward Function

Fig. 13 illustrates the trading actions and outcomes achieved through the utilization of the proposed risk aversion reward function, specifically concerning the concept of Maximum DrawDown. By maintaining a predetermined cash reserve level and implementing a trading reward function that penalizes the RL agent when the cash level falls below a specified threshold, the potential for enhancing the Maximum DrawDown metric becomes apparent. Maximum DrawDown (MDD) represents a widely used risk assessment tool within the realms of trading and investment. It gauges the utmost loss suffered by an investment or trading strategy from its peak to its lowest point before reaching a new peak (ovals in Fig. 13(a) In this context, smoothed increasing gained reward (Fig. 13(b)) versus fluctuated reward curve (Fig. 13(a)) demonstrates that by imposing penalties on the RL agent for sustaining a cash level beneath a designated threshold, a proactive encouragement is established for the agent to uphold a more substantial cash balance. Furthermore, the findings in Fig. 13(c) validate the results of the trade action distribution chart, showcasing a reduced number of long positions and an increased occurrence of short positions, accompanied by instances of holding flat positions. This configuration can be interpreted as a strategy for managing risks. This approach contributes to the mitigation of drawdown severity by deterring the agent from assuming overly risky positions that might otherwise lead to substantial losses. The tabular CR results of training using risk aversion cash penalty (RAPW) and no limit on cash (NLOC) is given in Table 4.

(a) (b) (c)

Fig. 13: a) Trading simulation by no limit on cash reserving and high volatile return. b) Trading simulation by using risk aversion cash penalty. c) Action distribution of RAPW of stocks.

Table 4: CR of stock in training process using risk aversion cash penalty (RAPW) and no limit on cash (NLOC)

Stock / Reward function	MSFT	AAPL	GOOGL	AMZN	META	TSLA
RAPW	2193	997	1063	1788	2755	1762
NLOC	2030	976	950	1487	2402	1695

C. Results of the Multiagent vs. Single agent

This experiment kicked off with the development of a comprehensive set of trading strategies for both the multiagent and single agent systems, taking into account various technical, fundamental, and sentiment-based factors. Prior to launching the experiment, extensive backtesting was carried out to fine-tune the parameters of each trading system, ensuring optimal strategy execution and reducing the potential for overfitting. Fig. 14 shows the result of this experiment in terms of gained total reward per episode.

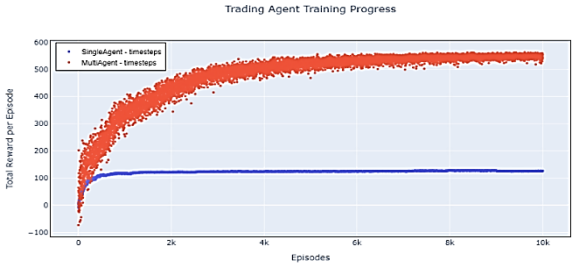


Fig. 14: Total reward per Episode of proposed MultiAgent system and single agent trading system, w=64.

In summary, the empirical evidence provides compelling support for the superiority of the multiagent trading system over the single-agent system in terms of returns. Through dedicating additional time to the model convergence process and refining policies, more favorable outcomes can be realized. This undertaking will not only enhance the model quality and precision, but also markedly elevate final efficiency. Thus, the accurate fusion of patience and focused parameter adjustments and policy optimizations has ultimately culminated in attaining superior performance and greater value compared to the invested time and efforts.

D. Results of the Robustness

To thoroughly challenge the CMSRS adaptability and robustness, the experiment ventured into uncharted territory by subjecting it to non-trained stock [IBM]. Analysis of trading performance metrics such as Sharpe ratio (SR), maximum drawdown (MDD), and average trade profit and loss (P&L) is shown in Table 5. The initial investment has set to 10,000$.

Table 5: Trading performance of proposed MultiAgent system on non-trained IBM stock

Period	January 1, 2013 - July 1, 2023		August 1, 2022 - August 1, 2023	
Measures	Proposed model	Buy and Hold	Proposed model	Buy and Hold
SR	0.1276	-1.3430	0.0057	-1.5821
MDD	-0.3592	-0.4372	-0.101	-0.1761
P&L	26456	-38.58	2512.04	1345.44

E. Comparison with Baselines

In this section, we provide the performance comparison results of our proposed CMSRS against other approaches during both the training and backtesting phases.

We compare our CMSRS against Buy and Hold baseline. Besides that, we employ state of the art Multi Agent DQN (MADQN) RL algorithm (Table 6).

Table 6: CMSRS backtesting results

Stock	Period	January 1, 2013 - July 1, 2023			August 1, 2022 - August 1, 2023		
	Measures	CMSRS	MADQN	B&H	CMSRS	MADQN	B&H
META	SR	**0.4325**	0.087	-0.8224	**0.0326**	0.0076	-0.4849
	MDD	**-0.24**	-0.432	-0.5922	**0.05831**	-0.311	-0.5085
	P&L	**53754.3**	50213	46798.70	**15467**	11098	10179.76
GOOG	SR	**0.3708**	0.0097	-1.1431	**0.05831**	0.0102	-0.8647
	MDD	**-0.21**	-0.298	-0.3087	**-0.09**	-0.203	-0.3166
	P&L	**73217.9**	57605.9	54208.64	**7521**	2314	1534.46
TSLA	SR	**0.1570**	0.0145	-0.4834	**0.1203**	0.0021	-0.5223
	MDD	**-0.3203**	-0.4509	-0.6063	**-0.34**	-0.43	-0.6505
	P&L	**2134012**	1348905	1250510	**3210**	23	-1207.48
MSFT	SR	**0.3773**	0.05	-1.1231	**0.0778**	0.0101	-0.9483
	MDD	**-0.1983**	-0.2809	-0.2908	**-0.12**	-0.311	-0.2684
	P&L	**123709**	113900	112621	**7525**	3715	2181.32
AAPL	SR	**0.3597**	0.1348	-1.0466	**0.09381**	0.0176	-1.0458
	MDD	**-0.2521**	-0.3	-0.3852	**-0.092**	-0.211	-0.2826
	P&L	**97654**	91212	85957.96	**4614.9**	1441	2141.46
AMZN	SR	**0.3464**	0.09	-0.9348	**0.06134**	0.0076	-0.7986
	MDD	**-0.2709**	-0.398	-0.4516	**-0.1689**	-0.271	-0.4349
	P&L	**108306**	97201	94856.15	**3251.65**	1258	-348.67

Fig. 15 shows learning curve of the proposed multiagent RL and baseline multi DQN. As evident from Fig. 15, the learning process of DQN exhibits notable variance. In contrast, the presented model not only outperforms DQN in terms of achieving superior rewards but also excels in learning speed and training efficiency, requiring significantly less time-approximately one-tenth of the time invested by DQN.

Fig. 15: Learning curve of the proposed multiagent RL and baseline multi DQN.

Conclusion

This study has introduced a significant advancement in the realm of stock recommender systems through the development of a Concurrent Multiagent Deep Reinforcement Learning-based Stock Recommender System (CMSRS).

While previous systems focused on a limited number of sequential trading agents within the same environment, often leading to errors in volatile market conditions, the proposed CMSRS represents a robust solution by leveraging concurrent multi-layer architecture. The CMSRS framework is designed with meticulous consideration, encompassing feature extraction in the data layer to construct diverse trading environments.

This innovative approach enables multiple feed Deep Reinforcement Learning (DRL) agents to make recommendations robustly within the trading layer. The system effectively integrates various data sources, incorporating Google stock trends, fundamental data, technical indicators, and historical price data. This comprehensive dataset empowers the concurrent agents to collaboratively select and recommend stocks for buying or selling.

To further enhance the effectiveness of the system, the Sharpe ratio is employed as a risk-adjusted return measure, facilitating the optimization of hyperparameters during the validation phase. Additionally, the introduced reward function ensures dynamic management of cash reserves, thereby addressing liquidity requirements and penalizing deviations from maintaining an adequate cash reserve. Empirical results obtained from real U.S. stock market data corroborate the supremacy of the Concurrent Multiagent SRS (CMSRS), particularly evident in volatile market conditions and out-of-sample scenarios. The CMSRS not only demonstrates its ability to navigate challenging market dynamics but also exhibits robustness and superior performance in comparison to prior systems. By advancing the capabilities of stock recommender systems in the domain of deep reinforcement learning, this research contributes significantly to the field of financial technology and investment strategies.

Author Contributions

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

References

[1] M. Z. Asghar, F. Rahman, F. M. Kundi, S. Ahmad, "Development of stock market trend prediction system using multiple regression," Comput. Math. Organ. Theory, 25(2019): 271-301, 2019.

[2] A. A. Ariyo, A. O. Adewumi, C. K. Ayo, "Stock price prediction using the ARIMA model," in Proc. 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation: 106-112, 2014.

[3] Y. Wang, Y. Liu, M. Wang, R. Liu, "LSTM model optimization on stock price forecasting," in Proc. 2018 17th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES): 173-177, 2018.

[4] S. Banik, N. Sharma, M. Mangla, S. N. Mohanty, S. Shitharth, "LSTM based decision support system for swing trading in stock market," Knowl. Based Syst., 239: 107994, 2022.

[5] S. Selvin, R. Vinayakumar, E. A. Gopalakrishnan, V. K. Menon, K. P. Soman, "Stock price prediction using LSTM, RNN and CNN-sliding window model," in Proc. 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI): 1643-1647, 2017.

[6] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness et al., "Human-level control through deep reinforcement learning," Nature, 518(7540): 529-533, 2015.

[7] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifreet al., "Mastering the game of Go with deep neural networks and tree search," Nature, 529 (7587): 484-489, 2016.

[8] A. R. Azhikodan, A. G. Bhat, M. V. Jadhav, "Stock trading bot using deep reinforcement learning," in Innovations in Computer Science and Engineering, Proc. the Fifth ICICSE 2017: 41-49, Springer, 2019.

[9] X. Wu, H. Chen, J. Wang, L. Troiano, V. Loia, H. Fujita, " Adaptive stock trading strategies with deep reinforcement learning methods," Inf. Sci., 538: 142-158, 2020.

[10] S. Carta, A. Corriga, A. Ferreira, A. S. Podda, D. R. Recupero, "A multi-layer and multi-ensemble stock trader using deep learning and deep reinforcement learning," Appl. Intell., 51: 889-905, 2021.

[11] X. Y. Liu, H. Yang, Q. Chen, R. Zhang, L. Yang, B. Xiao, C. D. Wang, "FinRL: A deep reinforcement learning library for automated stock trading in quantitative finance," arXiv preprint arXiv:2011.09607, 2020.

[12] S. Yang, "Deep reinforcement learning for portfolio management," Knowl. Based Syst., 278: 110905, 2023.

[13] C. Ma, J. Zhang, Z. Li, S. Xu, "Multi-agent deep reinforcement learning algorithm with trend consistency regularization for portfolio management," Neural Comput. Appl., 35(9): 6589-6601, 2023.

[14] Z. Huang, F. Tanaka, "MSPM: A modularized and scalable multi-agent reinforcement learning-based system for financial portfolio management," Plos one, 17(2): e0263689, 2022.

[15] J. Lussange, I. Lazarevich, S. Bourgeois-Gironde, S. Palminteri, B. Gutkin, "Modelling stock markets by multi-agent reinforcement learning," Comput. Econ., 57: 113-147, 2021.

[16] J. Lee, R. Kim, S. W. Yi, J. Kang, "MAPS: Multi-agent reinforcement learning-based portfolio management system," arXiv preprint arXiv:2007.05402, 2020.

[17] P. Koratamaddi, K. Wadhwani, M. Gupta, D. S. G. Sanjeevi, "A multi-agent reinforcement learning approach for stock portfolio allocation," in Proc. the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD): 410-410, 2021.

[18] D. Kwak, S. Choi, W. Chang, "Self-attention based deep direct recurrent reinforcement learning with hybrid loss for trading signal generation," Inf. Sci., 623: 592-606, 2023.

[19] S. Forouzandeh, K. Berahmand, R. Sheikhpour, Y. Li, "A new method for recommendation based on embedding spectral clustering in heterogeneous networks (RESCHet)," Expert Syst. Appl., 231: 120699, 2023.

[20] S. Forouzandeh, M. Rostami, K. Berahmand, R. Sheikhpour, "Health-aware food recommendation system with dual attention in heterogeneous graphs," Comput. Biol. Med., 169: 107882, 2024.

[21] M. Nourahmadi, A. Rahimi, H. Sadeqi, "Designing a stock recommender system using the collaborative filtering algorithm for the Tehran stock exchange," Financ. Res. J., 26(2): 302-330, 2024.

[22] B. Yang, T. Liang, J. Xiong, C. Zhong, "Deep reinforcement learning based on transformer and U-Net framework for stock trading," Knowl. Based Syst., 262: 110211, 2023.

[23] J. Zou, J. Lou, B. Wang, S. Liu, "A novel deep reinforcement learning based automated stock trading system using cascaded lstm networks," Expert Syst. Appl., 242: 122801, 2024.

[24] F. F. He, C. T. Chen, S. H. Huang, "A multi-agent virtual market model for generalization in reinforcement learning based trading strategies," Appl. Soft Comput., 134, 109985, 2023.

[25] S. Singh, V. Goyal, S. Goel, H. C. Taneja, "Deep reinforcement learning models for automated stock trading," in Advanced Production and Industrial Engineering, 27: 175, 2022.

[26] M. Taghian, A. Asadi, R. Safabakhsh, "Learning financial asset-specific trading rules via deep reinforcement learning," Expert Syst. Appl., 195: 116523, 2022.

[27] L. K. Felizardo, F. C. L. Paiva, C. de Vita Graves, E. Y. Matsumoto, A. H. R. Costa, E. Del-Moral-Hernandez, P. Brandimarte, "Outperforming algorithmic trading reinforcement learning systems: A supervised approach to the cryptocurrency market," Expert Syst. Appl., 202: 117259, 2022.

[28] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, MIT press, 2018.

[29] T. Faturohman, T. Nugraha, "Islamic stock portfolio optimization using deep reinforcement learning," J. Islamic Monetary Econ. Finance, 8(2): 181-200, 2022.

[30] H. Yue, J. Liu, D. Tian, Q. Zhang, "A novel anti-risk method for portfolio trading using deep reinforcement learning," Electronics, 11(9): 1506, 2022.

[31] A. Nair, P. Srinivasan, S. Blackwell, C. Alcicek, R. Fearon, A. De Maria et al., "Massively parallel methods for deep reinforcement learning," arXiv preprint arXiv:1507.04296, 2015.

[32] Y. Shoham, K. Leyton-Brown, "Multiagent systems: Algorithmic, game-theoretic, and logical foundations," Cambridge University Press, 2008.

[33] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, I. Mordatch, "Multi-agent actor-critic for mixed cooperative-ompetitive environments," Adv. Neural Inf. Processing Syst., 30, 2017.

[34] S. Khonsha, M. A. Sarram, R. Sheikhpour, "A profitable portfolio allocation strategy based on money net-flow adjusted deep reinforcement learning," Iran. J. Finance, 7(4): 59-89, 2023.

[35] H. Hu, L. Tang, S. Zhang, H. Wang, "Predicting the direction of stock markets using optimized neural networks with Google Trends," Neurocomput., 285: 188-195, 2018.

[36] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2018.

Biographies

Samira Khonsha is a faculty member of the department of computer engineering, Zarghan branch, Islamic Azad University. She holds a Bachelor's degree in Software Engineering from the Shiraz University. She also has a Master's degree in in Software Engineering from Shiraz University and a Ph.D. in Software Engineering from the Yazd University. Her areas of expertise include reinforcement learning, and financial markets.

- Email: khonsha.samira@gmail.com
- ORCID: 0000-0001-9301-760X
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://scholar.google.com/citations?user=4wEfZaAAAAAJ&hl=en&oi=ao

Mehdi Agha Sarram is an Associate Professor at the department of Computer Engineering in Yazd University, Yazd, Iran. He received his Ph.D. degree from University of Wales, Cardiff, U.K. in 1979. He is Member of Australian Institute of Control and Instrumentation and also Member of Steering Committee on IT standards (ISIRI-ITTC). He has been Casual Lecturer in Australian Universities such as SIBT Macquarie University, University of Western Sydney Macarthur and SWIC University of Western Sydney from 2000 to 2003. His research interests include Machine learning, Data mining, Network coding and Wireless sensor networks.

- Email: mehdi.sarram@yazd.ac.ir
- ORCID: 0000-0002-1872-6155
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://scholar.google.com/citations?user=Hx1_SDYAAAAJ&hl=en

Razieh Sheikhpour received her Ph.D. in Computer Engineering from Yazd University, Yazd, Iran, in 2017. Currently, she is an Associate Professor at the department of Computer Engineering at Ardakan University, Ardakan, Iran. Her research interests include machine learning, semi-supervised feature selection and bioinformatics.

- Email: rsheikhpour@ardakan.ac.ir
- ORCID: 0000-0002-3119-3349
- Web of Science Researcher ID: N-3816-2017
- Scopus Author ID: 55321804800
- Homepage: https://scholar.google.com/citations?user=SyIdF_4AAAAJ&hl=en

How to cite this paper:
S. Khonsha, M. A. Sarram, R. Sheikhpour, "A robust concurrent multi-agent deep reinforcement learning based stock recommender system," J. Electr. Comput. Eng. Innovations, 13(1): 225-240, 2025.

DOI: 10.22061/jecei.2024.11193.775

URL: https://jecei.sru.ac.ir/article_2229.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research Paper

Cross-correlation based Approach for Counting Nodes of Undersea Communications Network Considering Limited Bandwidth

M. Zillur Rahman [1], ***J. E Giti*** [1,*], ***S. Ariful Hoque Chowdhury*** [2], ***M. Shamim Anower*** [1]

[1]*Department of Electrical & Electronic Engineering, Faculty of Electrical and Computer Engineering, Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh.*

[2]*Department of Electronics & Telecommunication Engineering, Faculty of Electrical and Computer Engineering, Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh.*

Article Info

Article History:
Received 04 September 2024
Reviewed 10 October 2024
Revised 02 November 2024
Accepted 14 November 2024

Keywords:
Bandwidth (BW)
Coefficient of Variation (CV)
Cross-Correlation (CC)
Node counting
Scaling factor (S_F)
Undersea Acoustic Sensor Network (UASN)

*Corresponding Author's Email Address:
jishan.e.giti@gmail.com

Abstract

Background and Objectives: Node counting is undoubtedly an essential task since it is one of the important parameters to maintain proper functionality of any wireless communications network including undersea acoustic sensor networks (UASNs). In undersea communications networks, protocol-based node counting techniques suffer from poor performance due to the unique propagation characteristics of the medium. To solve the issue of counting nodes of an undersea network, an approach based on cross-correlation (CC) of Gaussian signals has been previously introduced. However, the limited bandwidth (BW) of undersea communication presents a significant challenge to the node counting technique based on CC, which traditionally uses Gaussian signals with infinite BW. This article aims to investigate this limitation.

Methods: To tackle the infinite BW issue, a band-limited Gaussian signal is employed for counting nodes, impacting the cross-correlation function (CCF) and the derived estimation parameters. To correlate the estimation parameters for finite and infinite BW scenarios, a scaling factor (S_F) is determined for a specific BW by averaging their ratios across different node counts.

Results: Error-free estimation in a band-limited condition is reported in this work if the S_F for that BW is known. Given the typical undersea BW range of 1–15 kHz, it is also important to establish a relationship between the S_F and BW. This relationship, derived and validated through simulation, allows for determining the S_F and achieving accurate node count under any band-limited condition within the 1–15 kHz range. Furthermore, an evaluation of node counting performance in terms of a statistical parameter called the coefficient of variation (CV) is performed for finite BW scenarios. As a side contribution, the effect of noise on the CC-based undersea node counting approach is also explored.

Conclusion: This research reveals that successful node counting can be achieved using the CC-based technique in the presence of finite undersea BW constraints.

Introduction

In addition to extensive environmental monitoring, undersea acoustic sensor networks (UASNs) are utilized for tasks such as predicting seismic and volcanic activity, offshore exploration, deep-sea archaeology, tactical surveillance, and monitoring oil and gas spills. Ensuring each node in the UASN functions correctly is crucial for these operations. Therefore, counting the operational

nodes in a UASN is required for detecting faulty nodes and maintaining effective network operations, including routing [1] and medium access [2].

The active node number of a UASN can be counted through the cross-correlation (CC) of the Gaussian signals coming from each node. These Gaussian signals are collected by multiple probing nodes called sensors. It is already known that a Gaussian signal has infinite bandwidth (BW). Since no BW constraint is applied to the Gaussian signals coming from each node, the superposition of infinite BW Gaussian signals received at each sensor location is used for node counting. Consequently, this limits the viability of the CC-based scheme in the BW-constrained undersea environment.

The primary objective of this paper is to examine the impact of finite BW on the CC-based technique. This exploration shows error-free node counting through the evaluation of the scaling factor (S_F) for a particular BW. The contributions of this paper are given as follows:

- A band-limited Gaussian signal is employed for counting nodes to investigate the corresponding effect on the cross-correlation function (CCF) and the derived estimation parameters.
- Successful estimation in a BW-constrained condition is demonstrated if the S_F for that BW is known.
- A mathematical expression is derived and validated through simulation for determining the S_F and achieving accurate node count under any finite band condition within the undersea BW range.
- An analysis of node counting error in terms of a statistical parameter called the coefficient of variation (CV) is conducted for varying BW.
- The noise impact on the CC-based undersea node counting method is reported.

A content summary of each subsequent section of the remaining article is given as follows: An in-depth review of existing node counting approaches can be found in the *Related Works* Section. After that, the *Research Gap* Section mentions the limitations of existing works to highlight further research scopes. Then, a brief overview of CC-based node counting methods for background context is provided in the *Background on Estimation Schemes Using CC* Section. The next two consecutive sections titled *BW Impact* and *Relation Between S_F and BW* present explorations of CC-based schemes in finite BW conditions through a mathematical relationship between S_F and BW. The succeeding *Performance Metric* Section discusses the calculation process to determine the counting error parameter and the dependency of that parameter on BW. All mathematical relationships derived throughout the paper are verified in the *Results and Discussion* Section through simulation. Finally, the *Conclusion* Section summarizes the findings of the paper with future directions.

Related Works

To justify the need for a CC based node counting technique, an overview of various estimation procedures for different types of networks is provided in this section. One such early research article by Varagnolo et al. [3] explored distributed anonymous strategies for estimating network cardinality. However, the feasibility of this strategy in wireless networks is yet to be investigated. Later, several algorithms for node estimation in different wireless networks, including wireless sensor and heterogeneous wireless networks, have been developed [4]-[7] without considering the dynamic behavior of the network. In contrast, Cattani et al. [8] introduced Estreme, a neighborhood cardinality estimator designed for dynamic wireless networks.

Apart from the dynamic behavior, network anonymity is another important factor to consider while counting nodes. With this in mind, some researchers work on size estimation methods for anonymous networks based on consensus [9]-[11]. On the other hand, Manaseer et al. [12] conducted a recent study to count the nodes using the mean number of hops required for each exchanged message in mobile ad-hoc networks. Another recent research work by Chatterjee et al. [13] investigated the issue of node number estimation in sparse networks with Byzantine nodes. Nonetheless, the applicability of these terrestrial communication-based node counting methods mentioned so far in undersea communications as well as radio frequency identification (RFID) networks requires further investigation.

RFID technology offers an affordable and flexible solution for object identification. The applications of this technology include the localization and tracking of objects in the supply chain, animal identification, ensuring secure operations in dangerous environments, facilitating electronic payments, and production control. In an RFID system, there are two main parts: a large number of tags for each object and several readers to identify those tags. Tag estimation of an RFID network is equivalent to the node counting in dynamic wireless networks. Since tag counting is a well-studied topic, numerous protocols and schemes for this task can be found in the literature. Recent protocols and schemes for tag counting include the single slot reuse protocol [14], the reliable missing tag estimation protocol [15], the coloring graph-based estimation scheme [16], and the cell averaging constant false alarm rate scheme [17].

RFID is also widely used in the Internet of Things (IoT) where communication overhead is one of the main challenges for active node counting. To address this, an algorithm known as approximate cardinality estimation [18] is utilized for large-scale IoT networks. In

another attempt to boost the node counting scalability for direct-to-satellite IoT networks, Parra et al. [19] proposed an optimistic collision information-based estimator. But, these solutions require significant effort to be developed and further improved. For this reason, machine learning classifiers and artificial neural networks have been introduced recently to tackle the cardinality estimation problem in RFID and IoT networks [19]-[23] which offers competitive performance with reduced design effort. However, all the abovementioned techniques lack the consideration of distinct aspects of undersea acoustic channel (UAC) such as significant capture effect, high path loss, and variable propagation delays, making them unsuitable for undersea environments [24].

A few investigations attempted to overcome the issues of UAC including capture effect. The capture effect refers to the phenomenon where one signal with a received power higher than those of interfering signals by a threshold amount is correctly received by the receiver. Thus, weak signals are not received in the presence of stronger interfering signals. With this in mind, Nemati et al. [25] accounted for the capture effect in tag estimation but did not address other challenges of UAC. A more comprehensive solution considering long propagation delays on top of the capture effect is provided by Howlader et al. [26]-[28] for estimating underwater network size. Blouin [29] also explored a size and structure estimation method based on node-to-node intermissions, enabling distributed computation in underwater networks. The protocol dependency of these methods [26]-[29] adds complexity to undersea network node estimation, making practical implementation difficult. To tackle the problem of protocol complexity in estimating underwater network size, Anower et al. [30]-[33] and Chowdhury et al. [34], [35] propose a new CC-based scheme utilizing two and three probing nodes (sensors), respectively. These schemes employ a straightforward probing protocol and are unaffected by the capture effect. Several assumptions underpin these techniques, including same received power (SRP) from each node, unity signal strength, infinite signal length, and ideal channel conditions (infinite bandwidth, no multipath propagation, and zero Doppler shift). While the SRP can be achieved through the probing technique, practical challenges in UAC such as finite bandwidth, multipath propagation delay, and Doppler shift require further examination. Previous studies have already explored the impacts of signal length [36]-[38], signal strength [39], multipath propagation delay [39]-[42], and dispersion coefficient [43].

Research Gap

Initially, these techniques [30]-[35] use CC of infinite bandwidth Gaussian signals. Later, further research [44], [45] shows the suitability of band-limited (10kHz and 5kHz) Gaussian signals. Scaling factors for 5kHz and 10kHz bandwidths are derived in [44] and [45] for efficient estimation only in these two limited bandwidth scenarios. However, a more general solution is required for any band-limited condition. To achieve this, scaling factors are derived in this paper for the entire underwater bandwidth, considering the finite bandwidth of UAC with two and three sensors to establish a relationship between S_F and bandwidth.

Background on Estimation Schemes Using CC

So far, three estimation schemes have been investigated using CC. They are two-sensor scheme [30]-[33], three-sensor schemes with SL (sensors in line) approach [34] and TS (triangular sensors) approach [35].

System models of these estimation schemes are shown in Fig. 1, where *N* nodes are uniformly distributed across 3D spherical regions underwater but sensor arrangements are different for each case.

In Fig. 1(a), the sensors are located with separation distance, d_{DBS} for two-sensor scheme such that, the distances between the centre of the sphere and the sensors are equal. In SL case, the middle sensor (H_2) is placed at the sphere centre and the other two sensors (H_1 and H_3) are positioned along a line with H_2 such that, $d_{\mathrm{DBS}_{12}}$ (distance between H_1 and H_2) = $d_{\mathrm{DBS}_{23}}$ (distance between H_2 and H_3) = dDBS which is obvious from Fig. 1(b).

In TS scheme, three sensors are positioned such that, $d_{\mathrm{DBS}_{12}} = d_{\mathrm{DBS}_{23}} = d_{\mathrm{DBS}_{31}}$ (distance between *H3* and *H1*) = dDBS to form an equilateral triangle, where the centroid of the triangle lies at the sphere centre which can be visualized in Fig. 1(c). Please note that it is also possible to perform node estimation with random placement of the sensors and different or unequal spacing between the sensors [46]-[48].

The estimation procedure initiates as sensors emit probe requests to *N* nearby nodes. These *N* nodes are treated as acoustic signal emitters capable of transmitting Gaussian signals in reply. When these signals reach the sensors, they arrive in various delayed and attenuated forms and are combined at each sensor location, resulting in mixed Gaussian signals. Through the operation of CC between these signals, CC functions (CCFs) are computed, which manifest as a series of delta functions [30]. For the two-sensor scheme, one CCF is obtained from the CC of the two mixed Gaussian signals received at the two sensor locations. For the SL scheme, two CCFs are derived from the CC of the two signals received by H_1 and H_2 sensors, and, H_2 and H_3 sensors. In the TS scheme, three CCFs are produced from the CC of the two signals received at three pairs of equidistant sensor locations.

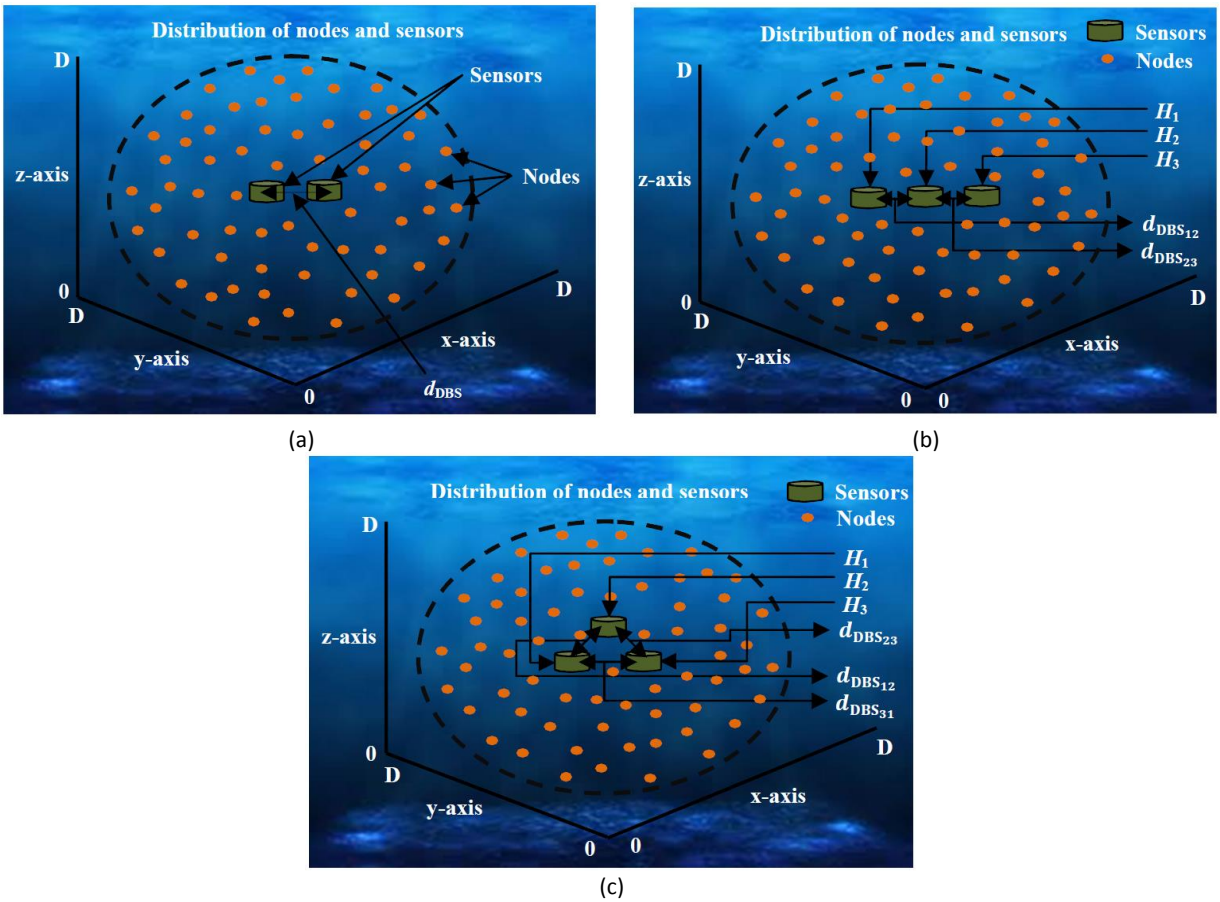


Fig. 1: System models with *N* transmitting nodes for counting the nodes of undersea wireless sensor network: (a) two-sensor method; (b) SL method; and (c) TS method

Fig. 2 illustrates a CCF derived from a network of *N* (=1000) nodes. In this context, bins, denoted as *b*, represent areas where deltas with identical delay variances are situated within a space twice as wide as the sensor spacing. The arrangement of deltas within these bins is dictated by the difference in signal delay experienced by the sensors. Number of bins, *b* is written as follows [49]:

$$b = \frac{2 \times d_{DBS} \times S_R}{S_P} - 1 \tag{1}$$

where, S_P is the speed of acoustic wave propagation and S_R is the sampling rate.

According to [49] and [50], the most preferred parameter for determining the node number using CC based schemes is calculated by taking the ratio between the standard deviation (σ (and the mean (μ (of a CCF. For scenarios with multiple CCFs, such as in the SL and TS cases, several parameters can be acquired, and the ultimate parameter to count the nodes is computed through the averaging of these values. The complex procedure for statistical computation of σ and μ of the CCF can be simplified by treating the CC formulation problem as a probabilistic problem [31].

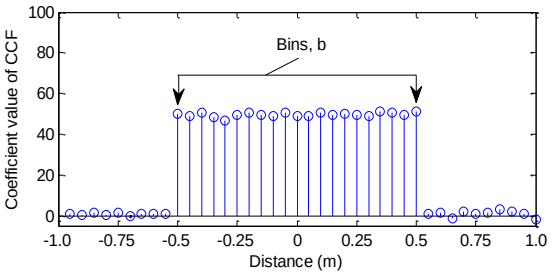


Fig. 2: Bins, *b* of the CCF.

This remodeling is derived from the fact that the bin number of a CCF follows a binomial probability distribution. Now, considering the infinite BW of Gaussian signals, the node counting parameters of two-sensor scheme, SL scheme and TS scheme is written after reformulation from [30], [34] and [35] as:

$$R^{1CCF}_{infiniteBW} = \frac{\sigma}{\mu} = \sqrt{\frac{(b-1)}{N}} \tag{2}$$

$$R^{2CCF}_{infiniteBW} = \frac{R_{12} + R_{23}}{2} = \sqrt{\frac{(b-1)}{N}} \tag{3}$$

and

$$R_{\text{infiniteBW}}^{\text{3CCF}} = \frac{R_{12} + R_{23} + R_{31}}{3} = \sqrt{\frac{(b-1)}{N}} \quad (4)$$

respectively, where R_{12} and R_{23} are the two node counting parameters of SL scheme and R_{12}, R_{23} and $R_{31}$are the three node counting parameters of TS scheme.

Node number, N can be calculated from (2), (3) and (4) for all three methods since estimation parameters are computed using the CCFs and b is determined from S_R, d_{DBS} and S_P using (1).

For 10kHz BW, estimation parameters for two-sensor case and SL case can be expressed from [44] and [45] as:

$$R_{\text{finiteBW}}^{\text{1CCF}} = 0.8093 \times R_{\text{infiniteBW}}^{\text{1CCF}} \quad (5)$$

and

$$R_{\text{finiteBW}}^{\text{2CCF}} = 0.8151 \times R_{\text{infiniteBW}}^{\text{2CCF}} \quad (6)$$

respectively, where 0.8093and 0.8151are the scaling factors. These values are obtained by taking the average of the ratios of R_{finiteBW} for 10kHz to the $R_{\text{infiniteBW}}$ for different N.

From (5) and (6), generalized expressions of estimation parameters for the three estimation schemes in finite BW conditions can be obtained using (2), (3) and (4) as follows:

$$R_{\text{finiteBW}}^{\text{1CCF}} = S_{\text{F}}^{\text{1CCF}} \times \sqrt{\frac{(b-1)}{N}} \quad (7)$$

$$R_{\text{finiteBW}}^{\text{2CCF}} = S_{\text{F}}^{\text{2CCF}} \times \sqrt{\frac{(b-1)}{N}} \quad (8)$$

$$R_{\text{finiteBW}}^{\text{3CCF}} = S_{\text{F}}^{\text{3CCF}} \times \sqrt{\frac{(b-1)}{N}} \quad (9)$$

where, $S_{\text{F}}^{\text{1CCF}}$, $S_{\text{F}}^{\text{2CCF}}$and $S_{\text{F}}^{\text{3CCF}}$ are the scaling factors of two-sensor approach, SL approach and TS approach, respectively.

Relation between S_F and BW

To investigate the dependency of S_F on BW in two-sensor scheme, values of $S_{\text{F}}^{\text{1CCF}}$ for different BW considering the BW range of undersea acoustic communication (15–1kHz) are obtained with b) 19 = S_R = 30kSa/s and d_{DBS}0.5 = m) as shown in Table 1. These values of $S_{\text{F}}^{\text{1CCF}}$ are plotted against BW using linear and log-log scale in Fig. 3(a) and 3(b), respectively.

A straight line equivalence of the $S_{\text{F}}^{\text{1CCF}}$ versus BW curve is presented in Fig. 3(b) where the approximate value of the slant of that straight line is .0.5221Since. Fig. 3(b) is a logarithmic plot, the $S_{\text{F}}^{\text{1CCF}}$ is written as:

$$\log_{10}(S_{\text{F}}^{\text{1CCF}}) = 0.5221 \times \log_{10}(\text{BW}) + k$$

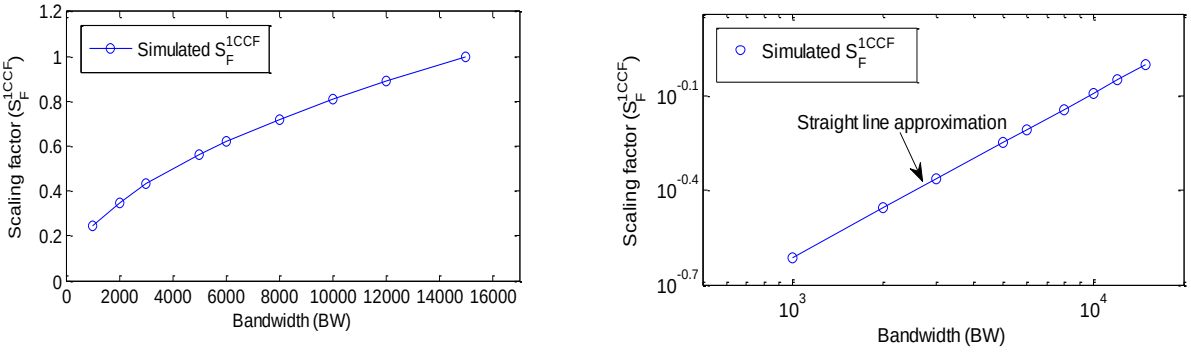
$$\Rightarrow \log_{10}(S_{\text{F}}^{\text{1CCF}}) = \log_{10}(\text{BW})^{0.5221} + \log_{10}(k_3)$$

$$\Rightarrow S_{\text{F}}^{\text{1CCF}} = k_3 \times \text{BW}^{0.5221} \quad (10)$$

Here k and k_3 are constants. Their relationship is given as $k = \log_{10}(k_3)$. The constant k_3 is determined by putting the values of a point from Fig. 3(a) into (10). Thus, the approximate value of k_3 is obtained as 0.0066.

Consequently, the final expression relating S_F and BW for two sensor approach is rewritten using (10) as:

$$S_{\text{F}}^{\text{1CCF}} = 0.0066 \times \text{BW}^{0.5221} \quad (11)$$



(a) (b)

Fig. 3: Linear (a) and logarithmic (b) plot of $S_{\text{F}}^{\text{1CCF}}$with respect to BW for two-sensor scheme.

Table 1: Scaling factor, $S_{\text{F}}^{\text{1CCF}}$ of two-sensor scheme

BW	1kHz	2kHz	3kHz	5kHz	6kHz	8kHz	10kHz	12kHz	15kHz
$S_{\text{F}}^{\text{1CCF}}$	0.2431	0.3491	0.4315	0.5633	0.6196	0.7200	0.8093	0.8898	1.000

Similarly, Table 2 and 3 contain the values of S_F^{2CCF} and S_F^{3CCF} for SL and TS cases, respectively, with different BW using b .19 = These values are plotted in Fig. 4, where Fig. 4(a) and (b) represent the linear and logarithmic plot of S_F^{2CCF} versus BW, respectively, and Fig. 4(c) and (d)4 represent the linear and logarithmic plot of S_F^{3CCF} versus BW, respectively.

From the straight line approximations as shown in Fig. 4(b) and 4(d), the slopes of the lines are obtained approximately as 0.4956and 0.4608and the values of the intercepts are 0.0085and 0.0119, respectively.

Therefore, S_F^{2CCF} of SL case and S_F^{3CCF} of TS case can be expressed as:

$$S_F^{2CCF} = 0.0085\times BW^{0.4956} \quad (12)$$

and

$$S_F^{3CCF} = 0.0119\times BW^{0.4608} \quad (13)$$

respectively.

Now, putting the expressions of S_F^{1CCF} ,S_F^{2CCF} and S_F^{3CCF} from (11), (12) and (13) into (7), (8) and (9), respectively, the estimation parameters, $R_{\text{finiteBW}}^{1CCF}$ of two sensor

Table 2: Scaling factor, S_F^{2CCF} of SL scheme

BW	1kHz	2kHz	3kHz	5kHz	6kHz	8kHz	10kHz	12kHz	15kHz
S_F^{2CCF}	0.2607	0.3676	0.4494	0.5789	0.6337	0.7308	0.8162	0.8934	1.000

Table 3: Scaling factor, S_F^{3CCF} of TS scheme

BW	1kHz	2kHz	3kHz	5kHz	6kHz	8kHz	10kHz	12kHz	15kHz
S_F^{3CCF}	0.2870	0.3951	0.4762	0.6026	0.6554	0.7483	0.8294	0.9021	1.000

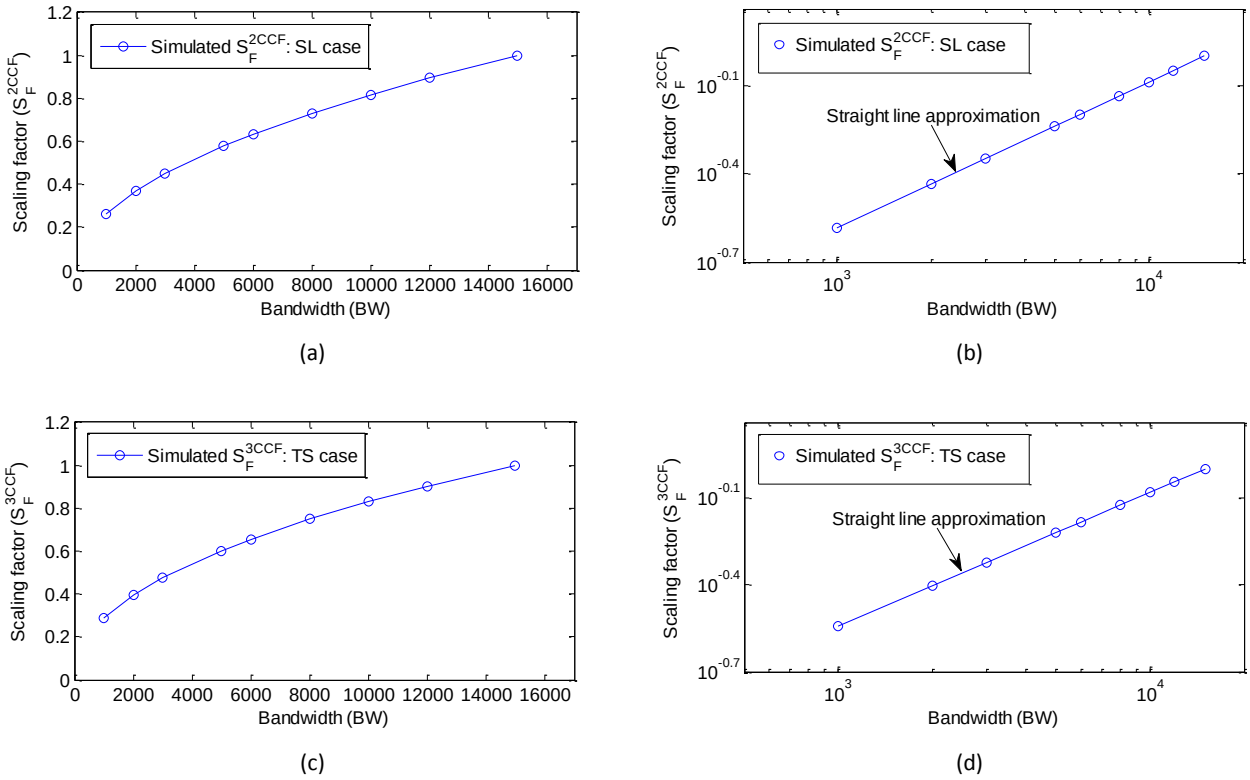


Fig. 4 : S_F^{2CCF} versus BW plot for SL approach in (a) normal; and) b) logarithmic scale and S_F^{3CCF} versus BW plot for TS approach in (c) normal; and (d) logarithmic scale.

scheme, $R_{\text{finiteBW}}^{\text{2CCF}}$ of SL scheme and $R_{\text{finiteBW}}^{\text{3CCF}}$ of TS scheme in finite BW conditions can be expressed as:

$$R_{\text{finiteBW}}^{\text{1CCF}} = 0.0066 \times \text{BW}^{0.5221} \times \sqrt{\frac{(b-1)}{N}} \tag{14}$$

$$R_{\text{finiteBW}}^{\text{2CCF}} = 0.0085 \times \text{BW}^{0.4956} \times \sqrt{\frac{(b-1)}{N}} \tag{15}$$

and

$$R_{\text{finiteBW}}^{\text{3CCF}} = 0.0119 \times \text{BW}^{0.4608} \times \sqrt{\frac{(b-1)}{N}} \tag{16}$$

respectively. Using these relationships, N can be estimated in the three estimation schemes with different BW conditions.

Performance Metric

Due to the statistical nature of the CC-based scheme, a statistical error parameter called coefficient of variation (CV) [51] is used as the performance indicator. CV is calculated by taking the ratio between the standard deviation (σ (and the mean (μ from obtained (several estimated N.

The expression of CV corresponding to the first iteration can be written as follows [52], [53]:

$$\text{CV}_1(N) = \frac{\sigma_1(N)}{\mu_1(N)} \tag{17}$$

Since the standard deviation of a set of estimated N decreases after each iteration of CV calculation, the $\text{CV}_\text{u}(N)$ corresponding to uth iteration is $1/\sqrt{u}$ times smaller than that of the first iteration [54], [55]. Now, we can rearrange (7), (8) and (9) as:

$$N = (b-1) \times \left(\frac{S_\text{F}^{\text{1CCF}}}{R_{\text{finiteBW}}^{\text{1CCF}}}\right)^2 \tag{18}$$

$$N = (b-1) \times \left(\frac{S_\text{F}^{\text{2CCF}}}{R_{\text{finiteBW}}^{\text{2CCF}}}\right)^2 \tag{19}$$

$$N = (b-1) \times \left(\frac{S_\text{F}^{\text{3CCF}}}{R_{\text{finiteBW}}^{\text{3CCF}}}\right)^2 \tag{20}$$

to obtain the expressions to determine N for two-sensor approach, SL approach and TS approach, respectively. By putting these expressions of N into (17), the CVs (after u iterations) for the three estimation schemes can be written as:

$$\text{CV}_{\text{finiteBW}}^{\text{1CCF}}(N) = \frac{1}{\sqrt{u}} \frac{\sigma_\text{u}\left((b-1)\left(\frac{S_\text{F}^{\text{1CCF}}}{R_{\text{finiteBW}}^{\text{1CCF}}}\right)^2\right)}{\mu_\text{u}\left((b-1)\left(\frac{S_\text{F}^{\text{1CCF}}}{R_{\text{finiteBW}}^{\text{1CCF}}}\right)^2\right)} \tag{21}$$

$$\text{CV}_{\text{finiteBW}}^{\text{2CCF}}(N) = \frac{1}{\sqrt{u}} \frac{\sigma_\text{u}\left((b-1)\left(\frac{S_\text{F}^{\text{2CCF}}}{R_{\text{finiteBW}}^{\text{2CCF}}}\right)^2\right)}{\mu_\text{u}\left((b-1)\left(\frac{S_\text{F}^{\text{2CCF}}}{R_{\text{finiteBW}}^{\text{2CCF}}}\right)^2\right)} \tag{22}$$

$$\text{CV}_{\text{finiteBW}}^{\text{3CCF}}(N) = \frac{1}{\sqrt{u}} \frac{\sigma_\text{u}\left((b-1)\left(\frac{S_\text{F}^{\text{3CCF}}}{R_{\text{finiteBW}}^{\text{3CCF}}}\right)^2\right)}{\mu_\text{u}\left((b-1)\left(\frac{S_\text{F}^{\text{3CCF}}}{R_{\text{finiteBW}}^{\text{3CCF}}}\right)^2\right)} \tag{23}$$

where, the CVs corresponding to two-sensor scheme, SL scheme and TS scheme are represented by $\text{CV}_{\text{finiteBW}}^{\text{1CCF}}(N)$, $\text{CV}_{\text{finiteBW}}^{\text{2CCF}}(N)$ and $\text{CV}_{\text{finiteBW}}^{\text{3CCF}}(N)$, respectively. It is obvious from (21), (22) and (23) that the CVs vary with b and scaling factors as well as BW.

Results and Discussion

This section contains all results related to node counting and performance comparison in three subsections. The results of the first two subsections correspond to node counting in the presence of noise and limited BW conditions. The third subsection reports counting errors in terms of CV.

Node counting in the presence of noise: To show the effect of noise, the internal noise of the receivers (sensors) is added to the node counting process. At first, the effect of noise on the two-sensor scheme is shown considering additive white Gaussian noise (AWGN) as the internal noise of a receiver. Simulations are conducted using the MATLAB programming tool, with varying signal ,lengthN_S (varies from 10^3 to 10^6samples) and signal-to-noise ratio, SNR, (varies from 10^{-5} to 10^5) of the receivers for a certain number (32 in this case) of nodes.

Other parameters used in the simulations (throughout the work unless otherwise mentioned) are: sphere dimension, 200D = 0m; sampling rate, S_R = 60kSa/s (because underwater acoustic bandwidth is around 15 kHz, this sampling rate is considered without violating the sampling theorem); speed of propagation, S_P = 1500m/s (typical underwater sound velocity); distance between sensors, d_{DBS} = 0.5m (so that the estimation sensors remain at the one node); absorption coefficient, a = 1 and dispersion factor, k = 1.5 (these are typical values for underwater acoustic communication). Results are plotted in Fig. 5, which presents the surface plots of N, SNR and N_S.

It can be seen in the results that, for a particular signal length (for example 100,000 samples) up to a certain SNR (<=0.05), the estimation is constant at the worst possible value but then improves with increases in SNR (up to SNR = 1).

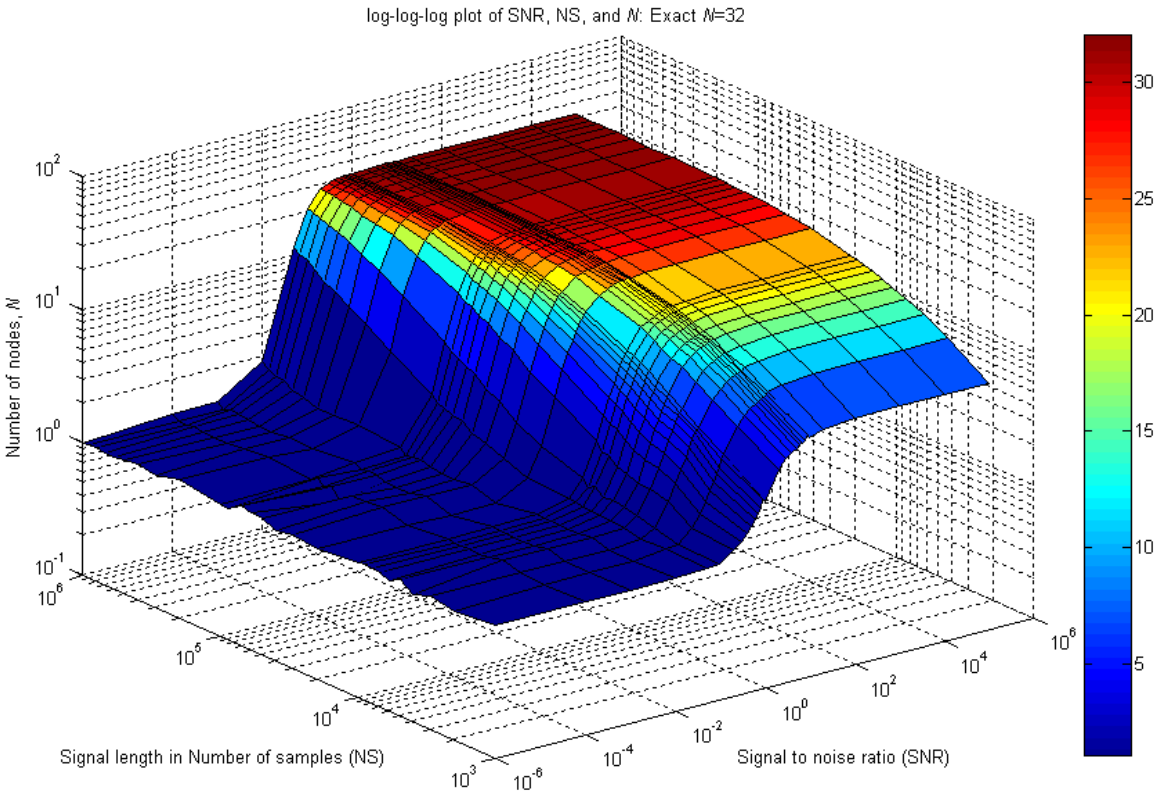


Fig. 5: Surface plot of SNR (10^{-5} to 10^{5}), N_S (10^3 to 10^6samples) and N (ideally 32) in logarithmic scale for two-sensor scheme.

Finally, the estimation becomes constant again at the maximum value, which corresponds to the case without noise. That is: when the SNR is less than 1, although the noise dominates over the signal, there are some signals that are strong enough to count; and, although we cannot estimate the appropriate number, we receive a reduced number of the signal sources, i.e., nodes. It can also be seen from Fig. 5 that, there is a transition zone between the worst and best possible values in which the estimation is varied with the SNR whose start and end points are varied with the signal length, i.e., it will start earlier with a greater signal length and later with a shorter signal length.

In a noisy environment, the transmitted signal must have sufficient power such that a suitable SNR is achieved. The effect of noise also varies with signal length (which determines the integration time of the cross-correlation process), the greater the length, the less is the effect of noise. This investigation shows that, if the signal strength and length are chosen properly, the estimation performance is similar to that of the ideal (without noise) case. It also shows that a SNR of 20 dB is sufficient to receive the signal with no errors as well as to neglect the noise effect in the estimation process.

Similar conclusions can be drawn for SL and TS schemes TS schemes and as the effect of noise will be similar for all three schemes.

Node counting with finite BW: The theoretical relationship between SF and BW, developed in one of the previous sections, is verified here through simulation work. Simulation results of $R^{1CCF}_{\text{finiteBW}}$,$R^{2CCF}_{\text{finiteBW}}$ and $R^{3CCF}_{\text{finiteBW}}$ for two-sensor, SL and TS approaches with corresponding theoretical results using (14), (15) and (16) are shown in Fig. 6(a), 6(b) and 6(c), respectively, using b = 39; and in Fig. 7(a), 7(b) and 7(c), respectively, using b = 89 with 12kHz and 3kHz BW.

In Fig. 6 and 7, the matching results from simulation and theory prove the usefulness of the expressions formulated in the *Relation between* S_F and *BW* Section. The additional simulation parameters are: signal length, Ns = 10^6 samples; signal to-noise ratio, SNR = 20 dB.

To demonstrate the significance of the S_F, Figs. 8, 9 and 10 show plots of the estimated (from simulation by averaging over 500 iterations) versus the exact node number for the two-sensor, SL, and TS approaches, respectively.

The plots include simulations both with and without S_F, along with the theoretically estimated node count using different BW and b values. The values of BW and b are: BW = 12kHz and b = 39 (S_R = 60kSa/s and d_{DBS}= 0.5m) in Fig. 8(a), 9(a), and 10(a); BW = 12kHz and b =89 (S_R = 45kSa/s and d_{DBS} = 1.5m) in Fig. 8(b), 9(b), and 10(b); BW = 3kHz and b = 39 in Fig. 8(c), 9(c), and 10(c) and BW = 3kHz and b = 89 in Fig. 8(d), 9(d) and 10(d).

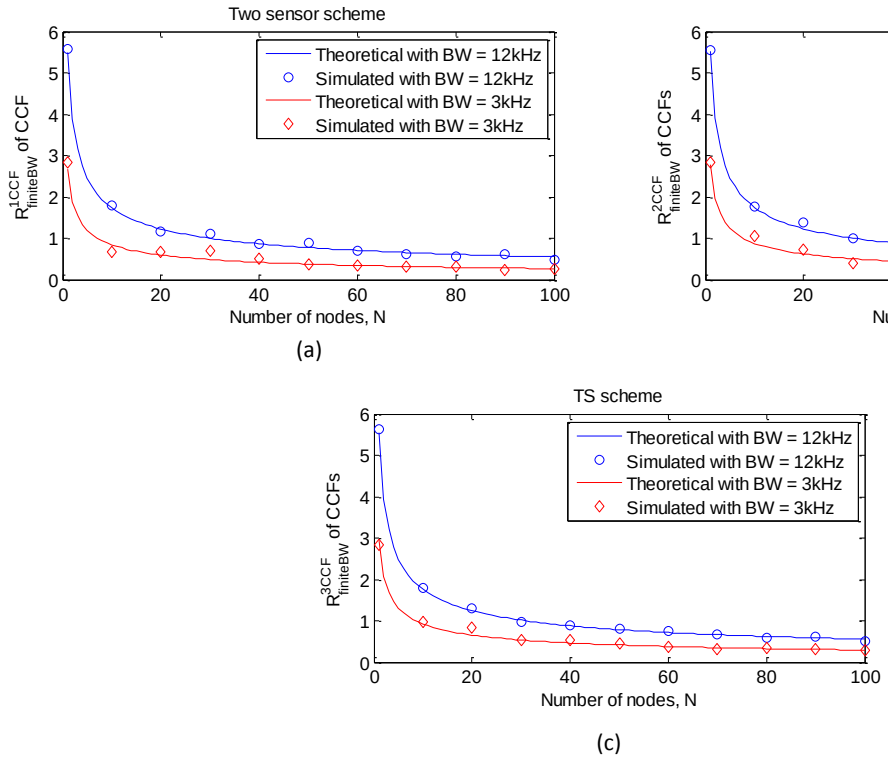


Fig. 6: Estimation parameter: (a) $R^{1CCF}_{finiteBW}$ of two-sensor scheme; (b) $R^{2CCF}_{finiteBW}$ of SL scheme; and (c) $R^{3CCF}_{finiteBW}$ of TS scheme versus *N* plot with BW = 12kHz and BW = 3kHz for b = 39 (d_{DBS} = 0.5m and S_R = 60kSa/s).

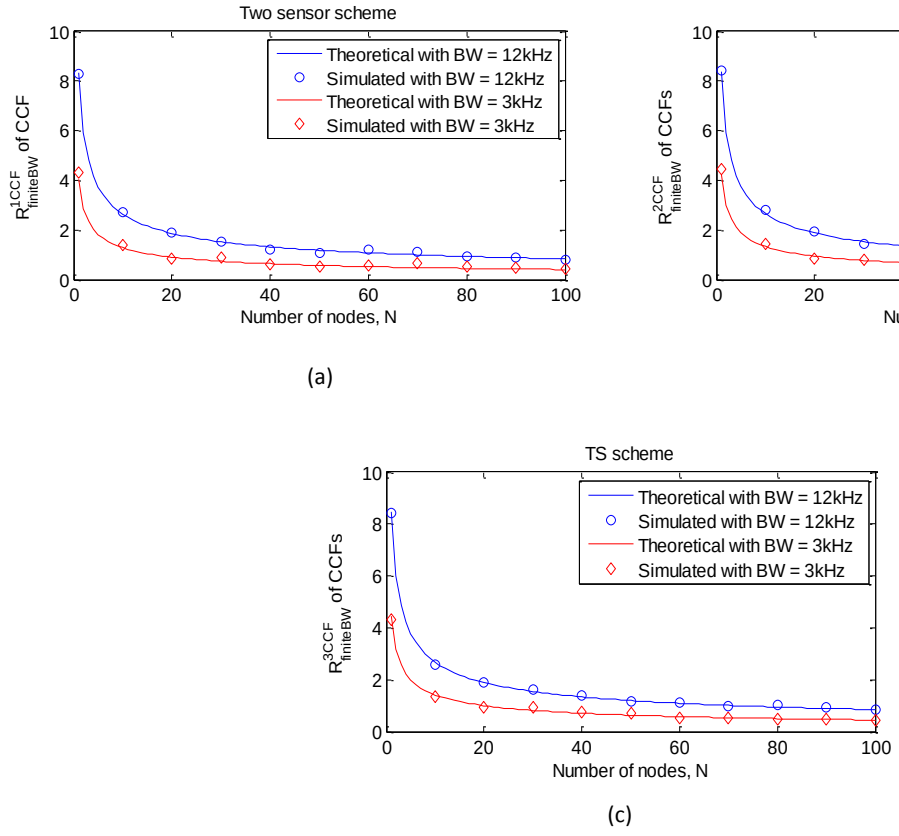


Fig. 7: Estimation parameter: (a) $R^{1CCF}_{finiteBW}$ of two-sensor scheme; (b) $R^{2CCF}_{finiteBW}$ of SL scheme; and (c) $R^{3CCF}_{finiteBW}$ of TS scheme versus *N* plot with BW = 12kHz and BW = 3kHz for b = 89 (d_{DBS} = 1.5m and S_R = 45kSa/s).

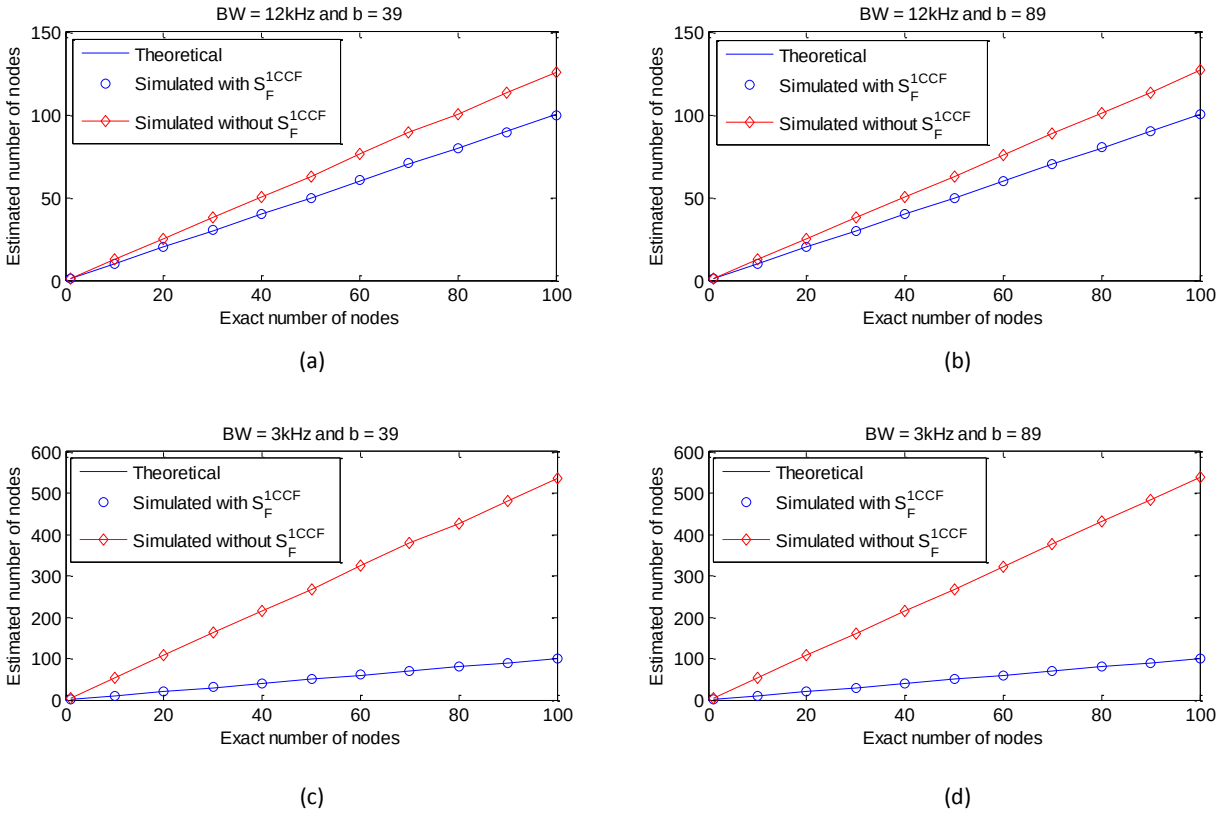


Fig. 8: Comparative analysis of estimated node number obtained from simulation for two-sensor scheme with and without S_F using different bandwidth (BW) and number of bins (b).

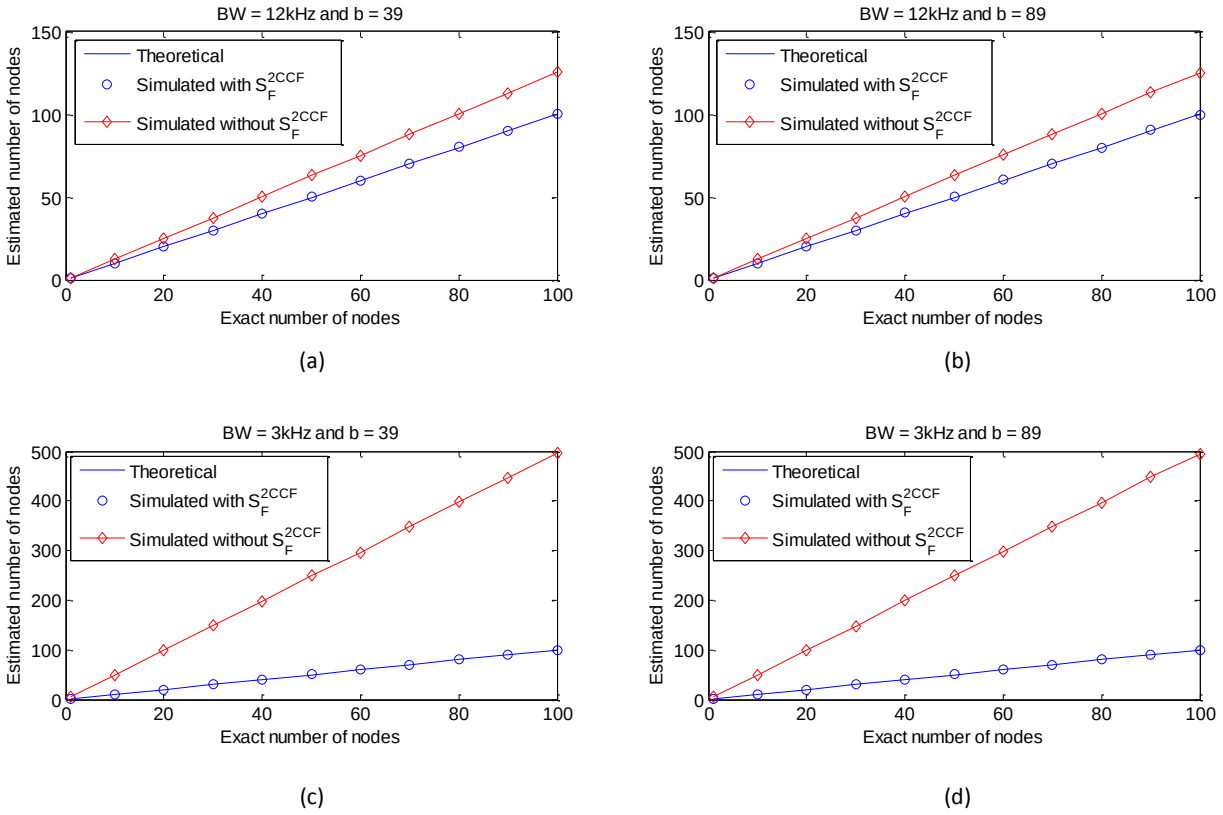


Fig. 9: Comparative analysis of estimated node number obtained from simulation for SL scheme with and without S_F using different bandwidth (BW) and number of bins (b).

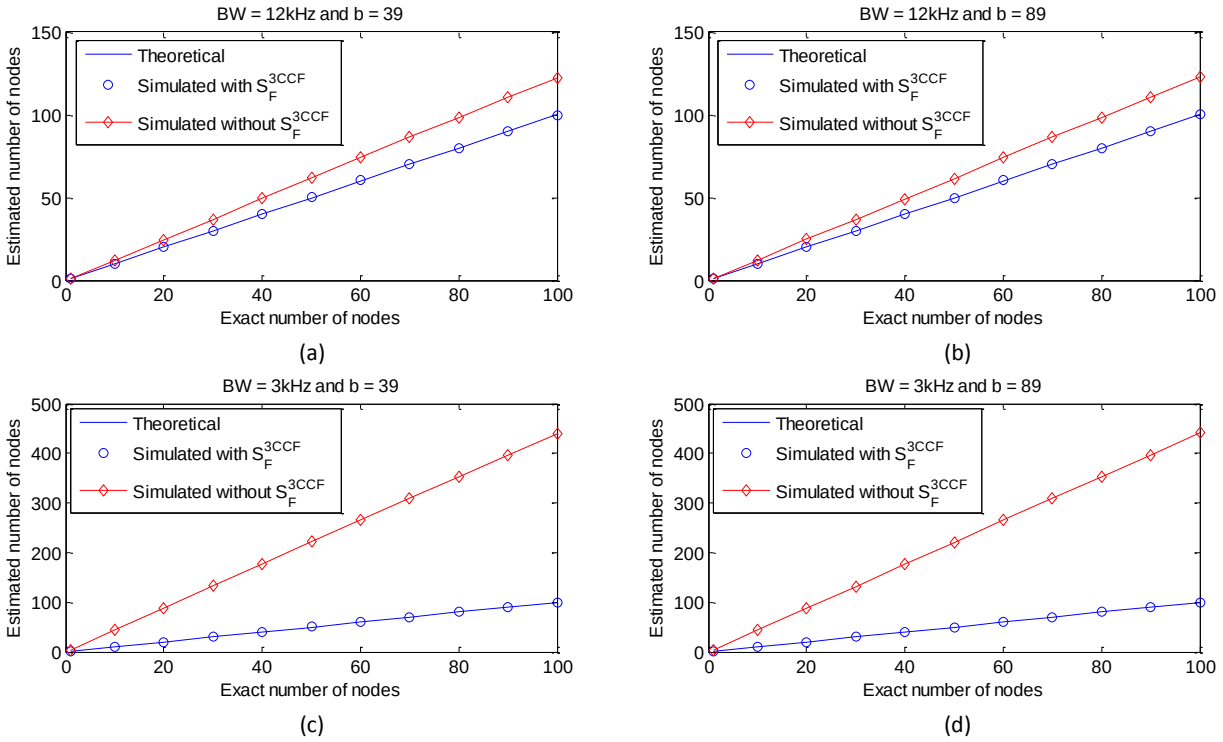


Fig. 10: Comparative analysis of estimated node number obtained from simulation for TS scheme with and without S_F using different bandwidth (BW) and number of bins (b).

It is obvious from all the results presented by Fig. 8, 9 and 10 that, satisfactory estimation can be achieved with S_F in different band-limited conditions using various b. In these plots, simulated node counting results (as shown by blue markers) match the corresponding theoretical results (as shown by blue lines), displaying the adequacy of S_F. It is also clear from these figures that, erroneous estimation (as shown by overlapping red markers and lines) is obtained without S_F in the same band-limited conditions using the same values of b, which indicates the importance of S_F.

Nonetheless, the difference between blue and red lines (as well as markers) in each plot of Fig. 8, 9 and 10 indicates the node counting error caused by limited BW. For without S_F case, this difference and consequently, the estimation error is smaller in the top two plots of each of these figures compared to those of the bottom two plots. Moreover, the top row plots (Fig. 8(a), 8(b), 9(a), 9(b), 10(a) and 10(b)) correspond to a higher BW than that of the bottom row plots (Fig. 8(c), 8(d), 9(c), 9(d), 10(c) and 10(d)). According to these findings, we can say that narrower BW conditions affect the CCFs as well as the estimation parameters derived from those CCFs more significantly than the wider BW cases leading to higher node counting errors. Therefore, the higher the BW, the lower the estimation error, and vice versa.

For further investigation, percentage relative estimation errors (e_r) without using S_F for two-sensor method, SL method and TS method are shown in Fig. 11 using different BW and b to compare the effect of BW on estimation accuracy of these three estimation methods. In Fig. 11, the values of BW and b corresponding to each plot are the same as those of the previous three figures (Fig. 8, 9 and 10). It can be seen from Fig. 11 that, the two-sensor approach shows the maximum e_r and TS approach shows the minimum e_r among the three schemes. Therefore, the SL and TS approaches are less impacted by finite BW than the two-sensor approach. However, the effect of BW is more pronounced in the SL approach compared to the TS approach. Fig. 11 also shows that, e_r increases with the decrease of BW for all three methods. In Fig. 11, fluctuating values of e_r indicate that, it is preferable to calculate statistical error for these schemes in terms of CV.

Node counting error: The performance of CC-based schemes is measured using the CV which provides a statistical node counting error. Due to the inverse relationship between CV and node counting accuracy, higher CV indicates lower accuracy and vice versa. To demonstrate the impact of finite BW on CV, simulation results of $\mathrm{CV}_{\mathrm{finiteBW}}^{1\mathrm{CCF}}(N)$, $\mathrm{CV}_{\mathrm{finiteBW}}^{2\mathrm{CCF}}(N)$ and $\mathrm{CV}_{\mathrm{finiteBW}}^{3\mathrm{CCF}}(N)$ corresponding to 100th iteration for two-sensor, SL and TS approaches are shown in Fig. 12. The plots include simulation results both with and without S_F, using different BW and b values to emphasize the further importance of scaling factors. The chosen values of BW and b of Fig. 12(a), 12(b), 12(c) and 12(d) are the same as those of Fig. 11(a), 11(b), 11(c) and 11(d), respectively, so that a more comprehensive and conclusive analysis of BW impact can be conducted.

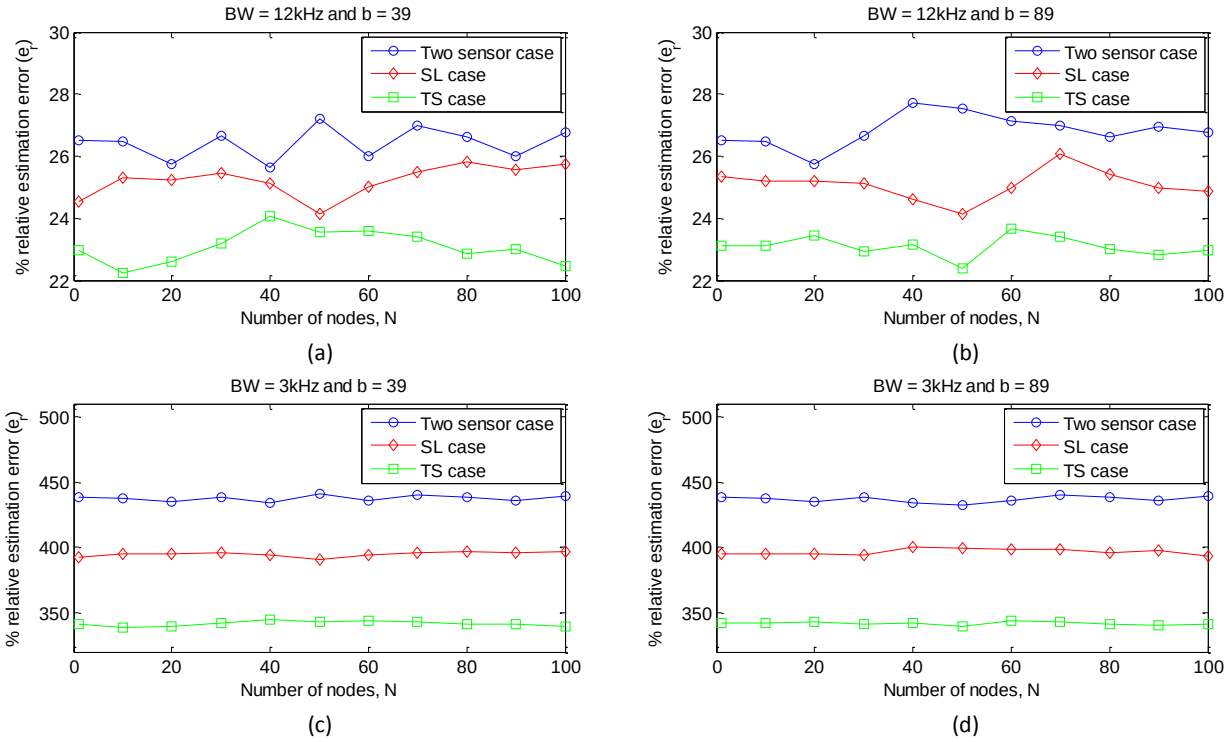


Fig. 11: Comparisons of er in two-sensor, SL and TS methods without using S_F for: (a) BW = 12kHz and b = 39 (dDBS = 0.5m and S_R = 60kSa/s); (b) BW = 12kHz and b = 89 (d_{DBS} = 1.5m and S_R = 45kSa/s); (c) BW = 3kHz and b = 39; and (d) BW = 3kHz and b = 89.

The CVs of Fig. 12 for different schemes and BW scenarios show similar traits as those of the percentage relative estimation errors, e_r of Fig. 11. Similar to e_r, CV also decreases with the increase of BW for all three schemes, and vice versa according to Fig. 12. This similar characteristics of CV and e_r with respect to BW is expected since both of these metrics have a similar dependency on the node counting parameters derived from the inaccurately formed CCFs due to the impact of finite BW condition.

It can be seen from Fig. 12 that, the TS approach shows the minimum CV whereas the two-sensor approach shows the maximum CV among the three methods. The underlying reason for the best node counting performance (corresponding to the lowest CV) achieved by the TS scheme is the use of more (in this case three) estimation parameters. This is because node counting using three (or u number of) parameters is equivalent to counting the nodes three (or u number of) times (or iterations) using a single parameter and since the accuracy of CC-based methods increases with the number of iterations used in the estimation process. Therefore, the performance of two-sensor and SL approaches in terms of CV is more impacted by limited BW conditions than the TS approach. Similarly, the SL approach is more robust compared to the two-sensor approach since two parameters are averaged by the SL scheme while a single node counting parameter is used by the two-sensor scheme. Fig. 12 also shows that the CVs obtained with S_F are lower than those of the corresponding scenarios without S_F. This is quite similar to Fig. 8, 9, and 10 where the lack of using S_F in finite BW conditions leads to inaccurate node counting results. These erroneous results are expected to have higher values of CV compared to those of the corresponding results obtained with S_F. Moreover, higher CVs as a consequence of not using S_F are more noticeable in smaller BW scenarios than in broader BW conditions. Hence, the use of S_F is more critical in lower BW cases.

Conclusion

This work addresses the issue of undersea bandwidth constraints on CC-based node counting schemes. It derives the relationship between scaling factors (S_F) and BW to provide generalized expressions for estimation parameters across three CC-based schemes in finite BW conditions, supported by simulations using various BW and b values. The study demonstrates that efficient estimation is achievable using S_F in limited BW conditions. However, without S_F, significant estimation errors occur which has been evaluated in terms of a statistical parameter called the coefficient of variation. Additionally, it shows that the estimation results of the TS approach are less affected by limited BW compared to the SL and two-sensor approaches. Future goals include analyzing the consequence of various environmental factors such as temperature, salinity, and underwater currents on the estimation performance of these three methods and eliminating other assumptions of CC-based schemes, such as the network's spherical shape and uniform node distribution.

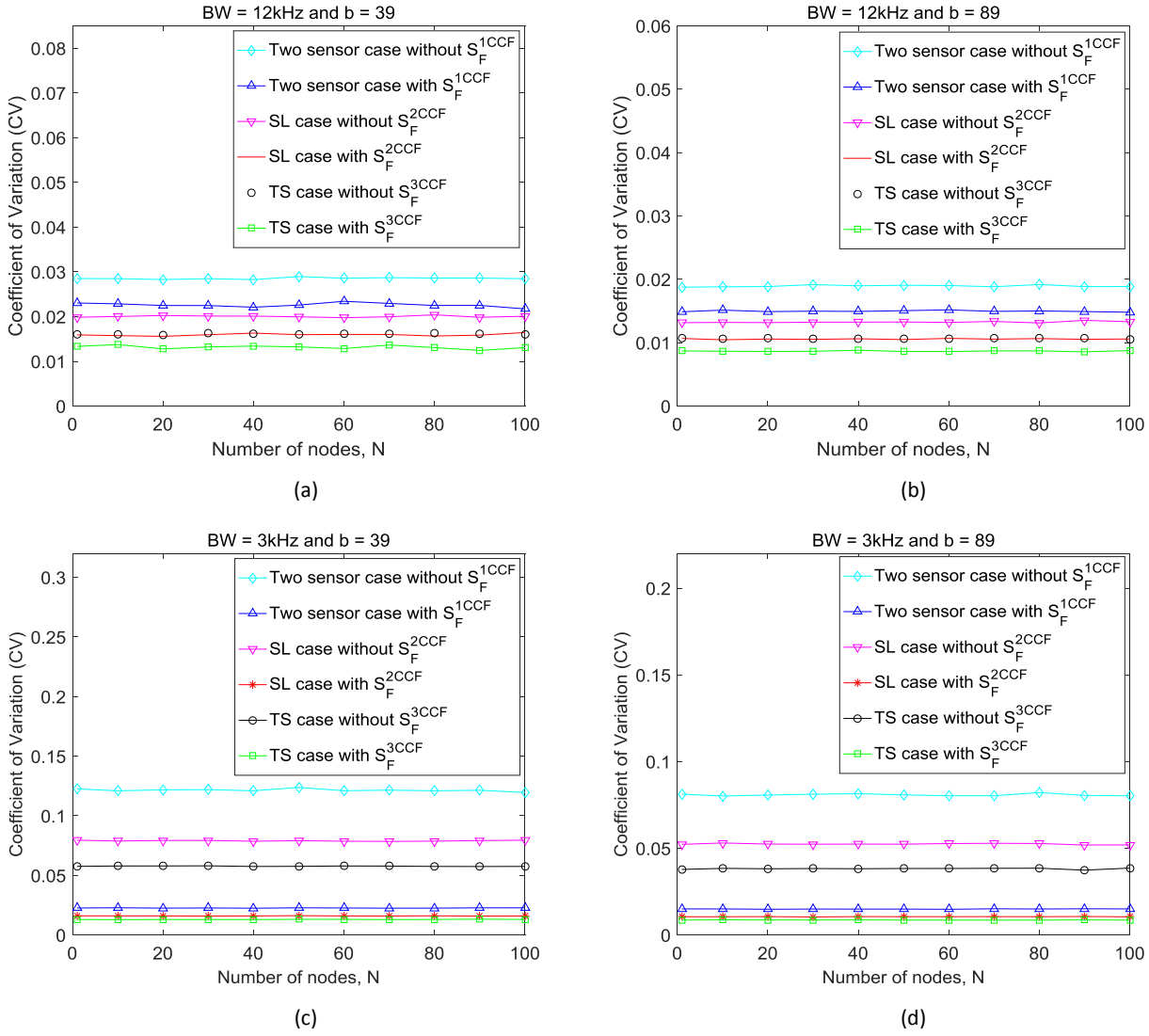


Fig. 12: Node counting performance evaluation in terms of CV of two-sensor, SL and TS methods with and without using S_F for: (a) BW = 12kHz and b = 39 (d_{DBS} = 0.5m and S_R = 60kSa/s); (b) BW = 12kHz and b = 89 (d_{DBS} = 1.5m and S_R = 45kSa/s); (c) BW = 3kHz and b = 39 (d_{DBS} = 0.5m and S_R = 60kSa/s); and (d) BW = 3kHz and b = 89 (d_{DBS} = 1.5m and S_R = 45kSa/s).

Author Contributions

Conceptualization: Shah Ariful Hoque Chowdhury, Md. Shamim Anower; Methodology: Md. Zillur Rahman; Formal analysis and investigation: Md. Zillur Rahman, Jishan E Giti; Writing - original draft preparation: Md. Zillur Rahman, Jishan E Giti; Writing - review and editing: Md. Zillur Rahman, Jishan E Giti; Supervision: Shah Ariful Hoque Chowdhury, Md. Shamim Anower.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

BW	*Bandwidth*
CV	*Coefficient of Variation*
CC	*Cross-correlation*
CCF	*Cross-correlation Function*
SL	*Sensors in Line*
TS	*Triangular Sensors*
UAC	*Undersea Acoustic Channel*
UASN	*Undersea Acoustic Sensor Network*

References

[1] S. Ashraf, M. Gao, Z. Chen, H. Naeem, T. Ahmed," CED-OR based opportunistic routing mechanism for underwater wireless sensor networks," Wireless Pers. Commun., 125: 487-511, 2022.

[2] G. Cario, A. Casavola, F. Torchiaro, "Medium access control in underwater sensor networks: a comparison between the standard JANUS and Ad-Hoc energy-efficient MAC protocols," in Proc. of the

10th Convention of the European Acoustics Association Forum (Acusticum): 5071-5078, 2023.

[3] D. Varagnolo, G. Pillonetto, L. Schenato, "Distributed Cardinality Estimation in Anonymous Networks," IEEE Trans. Autom. Control, 59(3): 645-659, 2014 .

[4] S. Chen, Y. Qiao, S. Chen, J. Li, "Estimating the cardinlaity of a mobile peer-to-peer network," IEEE J. Sel. Areas Commun., 31(9): 359-368, 2013.

[5] O. Sluciak, M. Rupp, "Network size estimation using distributed orthogonalization," IEEE Signal Processing Lett., 20(4): 347-350, 2013.

[6] A. Douik, S. A. Aly, T. Y. Al-Naffouri, M. S. Alouini, "Cardinality estimation algorithm in large-scale anonymous wireless sensor networks," in Proc. International Conference on Advanced Intelligent System and Informatics: 569-578, 2018.

[7] S. Kadam, K. S. Bhargao, G. S. Kasbekar, "Node cardinality estimation in a heterogeneous wireless network deployed over a large region using a mobile base station," J. Network Comput. Appl., 221: 103779, 2024.

[8] M. Cattani, M. Zuniga, A. Loukas, K. Langendoen, "Lightweight neighborhood cardinality estimation in dynamic wireless networks," in Proc. 13th Int. Symposium Information Processing in Sensor Networks: 179-189 , 2014.

[9] D. Varagnolo, G. Pillonetto, L. Schenato, "Consensus based estimation of anonymous networks size using Bernoulli trials," in Proc. American Control Conf.: 2196-2201, 2012.

[10] G. Luna, R. Baldoni, S. Bonomi, I. Chatzigiannakis, "Conscious and unconscious counting on anonymous dynamic networks," in Proc. International Conference on Distributed Computing and Networking: 257-271, 2014.

[11] R. Lucchese, D. Varagnolo, J. C. Delvenne, J. Hendrickx, "Network cardinality estimation using max consensus: The case of Bernoulli trials," in Proc. 54th IEEE Conference on Decision and Control (CDC): 895-901, 2015.

[12] S. Manaseer, I. Alhabash, "Number of node estimation in mobile Ad Hoc networks," Int. J. Interac. Mob. Technol. (iJIM), 11(6): 65-72, 2017.

[13] S. Chatterjee, G. Pandurangan, P. Robinson, "Network size estimation in small-world networks under byzantine faults," in Proc .IEEE International Parallel and Distributed Processing Symposium (IPDPS): 855-865, 2019.

[14] Z. Xi, X. Liu, J. Luo, S. Zhang, S. Guo, "Fast and reliable dynamic tag estimation in large-scale RFID systems," IEEE Internet Things J., 8(3): 1651-1661, 2021.

[15] B. Wang, G. Duan, "A reliable cardinality estimation for missing tags over a noisy channel," Comput. Commun., 188: 125-132, 2022.

[16] Z. He, "Reader scheduling for tag population estimation in multicategory and multireader RFID systems," Wireless Commun. Mob. Comput., 2021.

[17] A. Frahtia, M. Benssalah, A. Kifouche, K. Drouiche, "Improved tag estimation method for TDMA anticollision protocols using CA-CFAR technique," Frequenz, 78(11-12): 697-708, 2024.

[18] Q. Cao, Y. Feng, Z. Lu, H. Qi, L. M. Tolbert, L. Wan, Z. Wang, W. Zhou, "Approximate Cardinality Estimation (ACE) in large-scale Internet of Things deployments," Ad Hoc Networks, 66: 52-63, 2017.

[19] P. I. Parra, S. M. Sánchez, J. A. Fraire, R. D. Souza, S. Céspedes, "Network size estimation for direct-to-satellite IoT," IEEE Internet Things J., 10(7): 6111-6125, 2023.

[20] X. Jie, L. Haoliang, D. Wei, J. Ao, "Network host cardinality estimation based on artificial neural network," Secur. Commun. Netw., 2022.

[21] L. D. Rodić, I. Stančić, K. Zovko, T. Perković, P. Šolić, "Tag estimation method for ALOHA RFID system based on machine learning classifiers," Electronics, 11(16): 2605, 2022.

[22] P. S. Page, A. S. Siyote, V. S. Borkar, G. S. Kasbekar, "Node cardinality estimation in the internet of things using privileged feature distillation," IEEE Trans. Mach. Learn. Commun. Netw., 2: 1229-1247, 2024.

[23] S. A. Alhuthali, M. Murad, I. A. Tasadduq, M. H. Awedh, A. M. Rushdi, S. Alotaibi, "An effective tag estimation method based upon artificial neural networks and signal strength for anticollision in radio frequency identification systems," Int. J. Comput. Intell. Syst., 17(200), 2024.

[24] S. Climent, A. Sanchez, J. V. Capella, N. Meratnia, J. J. Serrano , "Underwater acoustic wireless sensor networks: Advances and future trends in physical, MAC and routing layers," Sensors (Basel Switzerland), 14(1): 795-833, 2014.

[25] M. Nemati, H. Takshi, V. Shah-Mansouri, "Tag estimation in RFID systems with capture effect," in Proc. 23rd Iranian Conference on Electrical Engineering: 368-373, 2015.

[26] M. S. A. Howlader, M. R. Frater, M. J. Ryan, "Estimation in underwater sensor networks taking into account capture," in Proc. IEEE Oceans'07, Aberdeen, Scotland: 1-6, 2007.

[27] M. S. A. Howlader, M. R. Frater, M. J. Ryan, "Estimating the number of neighbours and their distribution in an underwater communication network," in Proc. Second Int. Conf. Sensor Technologies and Applications, 2007.

[28] M. S. A. Howlader, M. R. Frater, M. J. Ryan, "Delay-insensitive identification of neighbors using unslotted and slotted protocols," Wirel. Commun. Mob. Comput., 2012.

[29] S. Blouin, "Intermission-based adaptive structure estimation of wireless underwater networks," in Proc. 10th IEEE Int. Conf. Networking, Sensing and Control: 146-151, 2013.

[30] S. Anower, M. R. Frater, M. J. Ryan, "Estimation by cross-correlation of the number of nodes in underwater networks," in Proc. Australasian Telecommunication Networks and Applications Conf.: 1-6, 2009.

[31] M. S. Anower, M. A. Motin, A. S. M. Sayem, S. A. H. Chowdhury, "A node estimation technique in underwater wireless sensor network," in Proc. Int .Conf. Informatics, Electronics & Vision: 1–6, 2013.

[32] S. Hossain, A. Mallik, M. A. Arefin, "A signal processing approach to estimate underwater network cardinalities with lower complexity," J. Electr. Comput. Eng. Innovations (JECEI), 5(2): 131-138, 2017.

[33] N. Afrin, S. Anower, M. Islam, "Dimensionality determination of unknown deployed underwater sensor network (UWSN) using cost function," Int. J. Commun. Syst., 33 (15): e4537, 2020.

[34] S. A. H. Chowdhury, M. S. Anower, J. E. Giti, "A signal processing approach of underwater network node estimation with three sensors," in Proc. 1st Int. Conf. Electrical Engineering and Information & Commun. Technology: 1-6, 2014.

[35] S. A. H. Chowdhury, M. S. Anower, J. E. Giti, "Effect of sensor number and location in cross-correlation based node estimation technique for underwater communications network," in Proc. 3rd Int. Conf .Informatics, Electronics & Vision: 1-6, 2014.

[36] M. A. Hossen, S. A. H. Chowdhury, M. S. Anower, S. Hossen, M. F. Pervej, M. M. Hasan, "Effect of signal length in cross-correlation based underwater network size estimation," in Proc. International Conference on Electrical Engineering and Information Communication Technology (ICEEICT): 1-6, 2015.

[37] B. K. Dash, H. H. Raton, S. A. H. Chowdhury, S. A. Rahman, "Performance analysis of cross-correlation based underwater network node estimation technique by varying signal length," in Proc. International Conference on Advancement in Electrical and Electronic Engineering: 1-4, 2018.

[38] S. A. H. Chowdhury, J. E. Giti, M. S. Anower, "Transmit energy calculation in cross-correlation based underwater network cardinality estimation," in Proc. IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE): 362-365, 2015.

[39] S. A. H. Chowdhury, M. S. Anower, J. E. Giti, M. I. Haque, "Effect of signal strength on different parameters of cross-correlation function in underwater network cardinality," in Proc. 2014 17th International Conference on Computer and Information Technology (ICCIT), 2014.

[40] M. S. Anower, S. A. H. Chowdhury, J. E. Giti, "Mitigating the effect of multipath using cross-correlation: application to underwater network cardinality estimation," Int. J. Syst., Control Commun., 7(3): 197-220, 2016.

[41] M. S. Anower, S. A. H. Chowdhury, J. E. Giti, "A robust signal processing approach of underwater network size estimation taking multipath propagation effects into account," Adv. in Netw., 3(3): 22-32, 2015.

[42] H. Sarker, M. Oli-Uz-Zaman, I. H. Chowdhury, S. A. H. Chowdhury, M. R. Islam, "Node estimation approach of underwater communication networks using cross-correlation for direct and multi-path propagation," in Proc. International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST): 286-291, 2019.

[43] M. K. Hossain, M. S. Anower, M. M. Rahman, S. M. N. Siraj, "Effect of dispersion coefficient on underwater network size estimation," in Proc. International Conference on Electrical Engineering and Information Communication Technology (ICEEICT): 1-4, 2015.

[44] M. S. Anower, S. A. H. Chowdhury, J. E. Giti, M. I. , Haque"Effect of correlation based underwater network size -bandwidth in cross ,estimation" in .Proc8th International Conference on Electrical and Computer Engineering: 413-416, 2014.

[45] S. K. Bain, S. A. H. Chowdhury, A. H. M. Asif, M. S. Anower, M. F. Pervej, S. S. Haque, "Impact of underwater bandwidth on cross-correlation based node estimation technique," in Proc. 2014 17th International Conference on Computer and Information Technology (ICCIT), 2014.

[46] H. H. Raton, S. A. H. Chowdhury, M. J. Rana, M. S. Anower, S. A. Hossain, M. I. Sarkar, "Cross-correlation based approach of underwater network cardinality estimation with random placement of sensors," in Proc. IEEE International Conference on Telecommunications and Photonics (ICTP): 1-5, 2015.

[47] D. K. Mondal, S. A. H. Chowdhury, Q. N. Ahmed, M. S. Anower, "Cross-correlation based approach of underwater network size estimation with unequal sensor separation," in Proc. International Conference on Computer and Information Engineering (ICCIE): 99-102, 2015.

[48] B. K. Dash, S. A. H. Chowdhury, A. H. M. M. Kamal, M. S. Anower, A. Halder, "Underwater network cardinality estimation using cross-correlation: Effect of unequal sensor spacing," in Proc. International Workshop on Computational Intelligence (IWCI): 181-186, 2016.

[49] M. S. Anower, "Estimation using cross-correlation in a communications network," Ph.D. dissertation, SEIT, University of New South Wales at Australian Defense Force Academy, Canberra, 2011.

[50] M. S. , AnowerS. A. H. , ChowdhuryJ. E. , GitiA. S. M. , SayemM. I. ,Haque"Underwater network size estimation estimation using ,correlation: selection of estimation parameter-cross" in Proc. 9th 9th International Forum on Strategic Technology (IFOST), 2014.

[51] S. W. Smith, "Chapter 2: Statistics, probability and noise," in The Scientist and Engineer's Guide to Digital Signal Processing, California Technical Publishing, San Diego, CA, 1999.

[52] S. A. H. ,ChowdhuryM. S. , AnowerJ. E. ,Giti"Performance comparison of underwater network size estimation techniques," Int. J. Syst., Control Commun., 7 (1): 16-34, 2016.

[53] S. A. H. Chowdhury, J. E. Giti, M. S. Anower, "Optimization between estimation error and transmit energy in cross-correlation based underwater network cardinality estimation," Wireless Pers. Commun., 97: 5797-5816, 2017.

[54] M. S. A. Howlader, "Estimation and identification of neighbours in wireless networks considering the capture effect and long delay," Ph.D. dissertation, SEIT, University of New South Wales at Australian Defense Force Academy, Canberra, 2009.

[55] B. A. Barry, "Errors in practical measurement in science, engineering, and technology," John Wiley & Sons, Hoboken, New Jersey, 1978.

Biographies

Md. Zillur Rahman received his B.Sc. degree in Electrical & Electronic Engineering from the Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh, in 2012. His research interests are Signal Processing and Underwater Communication. Currently, he is pursuing M.Sc. degree in the department of Electrical & Electronic Engineering at the Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh.

- Email: zillur.eee07@gmail.com
- ORCID: NA
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Jishan E Giti obtained her Ph.D. degree from the Monash University, Australia in 2020. She received her B.Sc. and M.Sc. degrees in Electrical & Electronic Engineering from the Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh, in 2011 and 2014, respectively. Her major research interests are Physical Layer Security, Wireless Communication and Networking. Currently, she is serving as an Associate Professor in the department of Electrical & Electronic Engineering at the Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh.

- Email: jishan.e.giti@gmail.com, jishan@eee.ruet.ac.bd
- ORCID: 0000-0001-5286-5450
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://www.ruet.ac.bd/jishan

Shah Ariful Hoque Chowdhury obtained his Ph.D. degree from the Australian National University, Australia in 2021**.** He received his bachelor's degree in Electronics & Telecommunication Engineering and master's degree in Electrical & Electronic Engineering from the Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh, in 2011 and 2014, respectively. His research fields are Underwater Communication, Wireless Communication and Computer Vision. Currently, he is serving as a Professor in the department of Electronics & Telecommunication Engineering at the Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh.

- Email: arif.1968.ruet@gmail.com, ariful.hoque@ete.ruet.ac.bd
- ORCID: 0000-0003-2597-156X
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://www.ruet.ac.bd/ariful

Md. Shamim Anower was born on 5th October 1977 in Bangladesh. He obtained his Ph.D. degree from the University of New South Wales, Australia in 2012**.** He received his B.Sc. and M.Sc. degrees in Electrical & Electronic Engineering from the Rajshahi University of Engineering & Technology in 2002 and 2007, respectively. His major research interests are Underwater Wireless Communication, Power Line Communication, Signal Processing for Communications, Power System Analysis and Stability Enhancement. He published more than 100 international and national journal and conference articles so far. Currently, he is serving as a Professor in the department of Electrical & Electronic Engineering at the Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh.

- Email: md.shamimanower@yahoo.com, msanower@eee.ruet.ac.bd
- ORCID: 0000-0001-6986-6847
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://www.ruet.ac.bd/Shamim

How to cite this paper:
M. Zillur Rahman, J. E Giti, S. Ariful Hoque Chowdhury, M. Shamim Anower, "Cross-correlation based approach for counting nodes of undersea communications network considering limited bandwidth," J. Electr. Comput. Eng. Innovations, 13(1): 241-256, 2025.

DOI: 10.22061/jecei.2024.11252.783

URL: https://jecei.sru.ac.ir/article_2230.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Modified Topologies for Single Source Switched-Capacitor Multilevel Inverters

F. Sedaghati [1,*], S. Ebrahimzadeh [1], H. Dolati [2]

[1]*Department of Electrical Engineering, Faculty of Engineering, University of Mohaghegh Ardabili, Ardabil, Iran.*

[3] *Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran.*

Article Info

Article History:
Received 01 September 2024
Reviewed 03 October 2024
Revised 15 November 2024
Accepted 17 November 2024

Keywords:
Multilevel inverter
Switched-capacitor
Capacitor charging
Longest discharge period

*Corresponding Author's Email Address:
farzad.sedaghati@uma.ac.ir

Abstract

Background and Objectives: Increasing environmental problems and challenges have led to increased use of renewable energy sources such as photovoltaic or PV system. One of the attractive research fields is power electronic converters as interfaces for renewable energy sources. Multilevel inverters can operate as such interfaces. This paper introduces modified topologies of switched-capacitor multilevel inverters, designed to overcome constraints of low voltage renewable energy sources such as PV.

Methods: Configuration of topologies utilize a single DC source with series or parallel connection of capacitors to produce 7-level, 9-level, and 11-level voltage in the converter load side. The paper presents the converter operation principle, elements voltage stress analysis, and capacitor sizing calculations. Also, operation analysis of suggested inverter topologies is validated using implemented set up.

Results: Comprehensive comparative analysis reveals that the proposed topologies have merits and superior performance compared to existing solutions regarding component number, voltage boost factor, and voltage stress. The experimental measurement results confirm the accuracy of multilevel output voltage waveforms and the self-balancing of capacitor voltages, as predicted by theoretical analysis.

Conclusion: The suggested switched-capacitor multilevel inverters, moreover the superiority over previously presented topologies, show great potential for application in photovoltaic systems and electric vehicle battery banks.

Introduction

In order to achieve zero carbon emissions, renewable energy sources have gained noteworthy regard due to their dependable performance, minimal ecological footprint, cost efficiency, and adaptability within systems. Increasing adoption of renewable energy sources and electric vehicles has led to a growing need for enhanced voltage boost capability. Nevertheless, most of DC sources such as solar cells and EV batteries have a restricted capacity for boosting power, indicating the need for immediate improvements in this field [1]-[3].

Commercial solutions in this field typically use two-stage power conversion, utilizing a step-up converter. In order to get the highest possible output voltage, the components utilized in these converters operate at high switching frequencies, which increase power losses and prices [4]. The impedance source inverter, which employs LC units, incurs additional system volume and expense. In addition, the converters' two-level setup results in high total harmonic distortion (THD) and power losses. The high rate of voltage variation (d*v*/d*t*) in switches increases power losses during switching, and also, increases voltage

stress and electromagnetic interference (EMI), and impaired reliability. Ultimately, this can lead to reduction in lifespan of electric vehicles or motors [5]-[8].

Industry commonly utilizes multilevel inverters, including traditional diode-clamp, flying-capacitor, and cascade H-bridge topologies. These inverters offer several benefits, such as low output voltage THD, less switching stress, and lower operating frequency. Traditional configurations of multilevel inverters cannot enhance the input voltage [5]-[8].

An innovative multilevel inverter utilizing switched capacitor (SC) approach was provided as a solution to address the abovementioned limitations [9]-[11]. SC multilevel inverters offer the following benefits:

- Capacitor voltage self-balance: This feature guarantees that the voltage across every capacitor in the inverter is similar, hence preserving the stability of the inverter [12], [13].
- The inverter does not utilize significant magnetic components and transformers, resulting in lighter and smaller sizes, enhancing its portability and compactness [12], [14], [15].
- Offer a high degree of flexibility as they can be quickly adjusted to match the individual needs of various applications [16], [17].
- SC multilevel inverters provide the ability to boost the input voltage, which distinguishes them from traditional multilevel inverters [12], [18], [19].

SC multilevel inverters have some demerits, such as employing high count of active and passive elements, and capacitor voltage balancing concerns [20]. However, they are still highly regarded for their merits and have been proposed for use in high-power photovoltaic systems [21].

A multilevel inverter including SC cells was introduced in [22]. This inverter can be constructed in symmetric or asymmetric configurations and possesses the capability to expand up to greater levels. Its main advantage is low voltage stress on its switches. Nevertheless, the primary disadvantages of this system consist in its extensive utilization of numerous switches and the inclusion of unidirectional switches. A topology capable of producing a voltage with seven distinct levels and a voltage boost factor of 3 has been suggested in [23]. This is accomplished by utilizing only four high-frequency switches, which are secured by low-voltage capacitors. Nevertheless, this configuration experience significant voltage stress on the switches and also, necessitate additional passive components. Consequently, size and weight of the converter are increased. A seven-level inverter design was suggested in [24] for use in medium-voltage scenarios, particularly for high-power applications. The configuration includes eight operational switches, two internal flying-capacitor units, and two diodes. However, this topology has two primary disadvantages: the use of two unidirectional current switches and the imposition of considerable voltage stress on the switch.

In [25], a generalized boost multilevel inverter that can be utilized in applications with low-voltage input sources.

By controlling the parallel and series connection of capacitors and DC sources, this configuration can produce high voltage levels.

The important demerit of this topology is utilizing high count of passive elements, which increases the converter cost, size, and weight. The single-phase SC MLI proposed in [26] can produce a nine-level AC voltage with a voltage gain of 4 in the output. This 9-level topology is achieved by modifying the switching algorithm of the 13-level configuration introduced in [26].

However, this topology has essential downside as it requires a large number of switches, which in turn increases the need for gate drivers, and subsequently, the cost and size of the converter. A 9-level inverter was introduced in [27], which offers the significant advantage of zero current switching for charging capacitors. However, its primary drawback is utilization of high count of switches, diodes, capacitors, and switches with very high total standing voltage (TSV). According to the configuration presented in [28], 11-level switched-capacitor multilevel inverter can produce an output voltage waveform using 14 switches, 3 capacitors, and 2 diodes. However, this topology has a major drawback of using a lot of switches and a low boost factor. A single-phase switched-capacitor based 11-level inverter topology is presented in [29].

This configuration offers increased levels, a quintuple voltage boost factor, and natural capacitors voltage-balancing as its main features. However, due to the inclusion of more passive elements and high TSV, power loss, cost, and volume are increased. The 11-level topology that uses the SC technique, described in [30], offers several benefits including capacitor voltage self-balancing and high boost efficiency. However, because of large number of power switches and high TSV, the converter experiences high power losses that consequently reduce efficiency.

The rest of the paper is organized as follows: In section 2, proposed multilevel inverter topology and operation are described. Section 3 presents extended topology of suggested converter. Capacitor sizing calculations is given in section 4, and converter power losses are computed in section 5. Section 6 presents comprehensive comparison of the introduced converter with similar topologies. Experimental test results are illustrated in section 7, and section 8 concludes the paper.

Proposed 7-Level Inverter Topology

A. Circuit Description

Circuit schematic of proposed boost multilevel inverter is illustrated in Fig. 1. Suggested 7-level topology contains a DC voltage source, V_{in}, two capacitors, C_1 and C_2 along with the voltage source with two power diodes, and eight power semiconductor switches to produce 7-level voltage in the output. Capacitors C_1 and C_2 have equal capacity and are charged in the same manner. Proposed MLI generates output voltage with levels of $\pm V_{in}$, $\pm 2V_{in}$, and $\pm 3V_{in}$ by H-bridge inverter. Different switching modes for the suggested topology to produce 7-level output voltage are shown in Table 1. All switches used in this configuration have the ability to facilitate bidirectional current so, the converter supports the inductive load with reverse current flow.

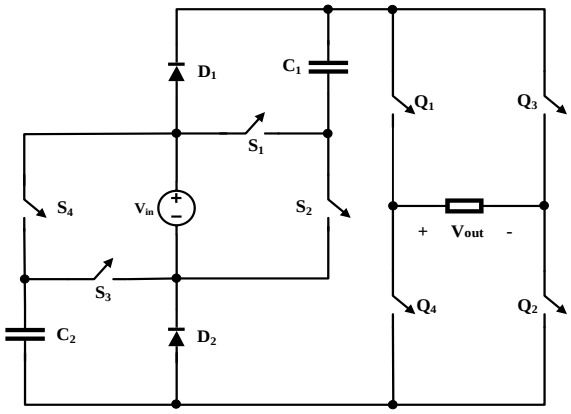


Fig. 1: Configuration of proposed 7-level switched-capacitor inverter.

Table 1: Circuit elements states for 7-level topology operation.

$\frac{V_o}{V_{in}}$	Switch and diode state										Capacitor state	
	S_1	S_2	S_3	S_4	Q_1	Q_2	Q_3	Q_4	D_1	D_2	C_1	C_2
+3	1	0	1	0	1	1	0	0	R	R	D	D
+2	1	0	0	1	1	1	0	0	R	F	D	C
+1	0	1	0	1	1	1	0	0	F	F	C	C
+0	0	1	0	1	1	0	1	0	F	F	C	C
-0	0	1	0	1	0	1	0	1	F	F	C	C
-1	0	1	0	1	0	0	1	1	F	F	C	C
-2	1	0	0	1	0	0	1	1	R	F	D	C
-3	1	0	1	0	0	0	1	1	R	R	D	D

B. Operation Principle

Operation of the introduced 7-level inverter with charge and discharge cycles of capacitors at each generated voltage level is given in Fig. 2. Switching operation of switches S_1, S_2, and switches S_3, S_4 are complementary which simple control. The switches in the flow path are marked in red colour in Fig. 2.

Mode 1: In this mode, both switches S_2 and S_4 are on, causing the capacitors to be charged equally using the source as the current flows through the diodes. The output voltage at the "*E*" terminal equals the source voltage level (Fig. 2(a)).

Mode2: Switches S_1 and S_4 are on in this mode that keeps capacitor C_2 charged. Due to the capacitor C_1 being in series with the V_{in}, voltage level of $2V_{in}$ is generated in terminal *E* (Fig. 2(b)).

Mode 3: This mode involves connecting capacitors C_1 and C_2 in series with V_{in} and discharging them to generate voltage level $3V_{in}$ in terminal by turning on switches S_1 and S_3 (Fig. 2(c)).

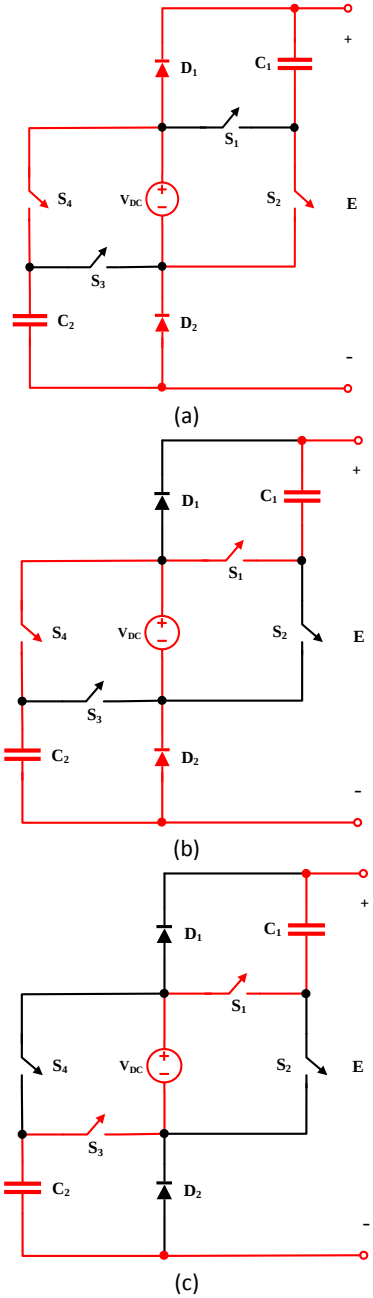


Fig. 2: Equivalent circuits of proposed inverter for each output voltage level, (a) mode 1, (b) mode 2, (c) mode 3.

C. Voltage Stress Analysis

Table 2 presents voltage stress of semiconductor elements applied in the 7-level configuration. Also, Fig. 3 illustrates TSV values of the diodes and switches. It indicates that switches S_1-S_4 have the same voltage stress, equals to V_{dc}. On the other hand, switches Q_1-Q_4, which forms the H-bridge in the output, bears varying voltage stress in each step, with the maximum voltage stress of $3V_{dc}$. TSV of the proposed 7-level configuration is obtained as given in the following:

$$TSV = 4 \times V_{dc} + 4 \times 3V_{dc} = 16V_{dc} \quad (1)$$

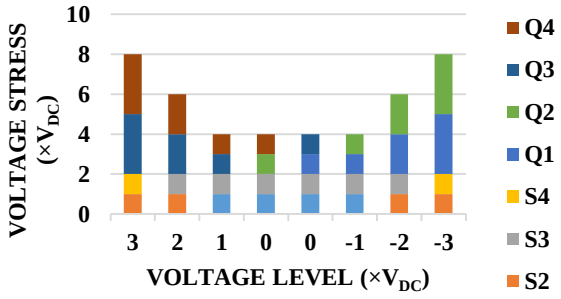


Fig. 3: Voltage stress of switches in the suggested 7-level topology.

Table 2: Circuit elements states for 9-level topology operation

$\frac{V_o}{V_{in}}$	Switch and diode state													Capacitor state		
	S_1	S_2	S_3	S_4	S_5	S_6	Q_1	Q_2	Q_3	Q_4	D_1	D_2	D_3	C_1	C_2	C_3
+4	1	0	1	0	0	1	1	1	0	0	R	R	R	D	D	D
+3	1	0	0	1	1	0	1	1	0	0	R	R	F	D	D	C
+2	1	0	0	1	0	1	1	1	0	0	R	F	F	D	C	C
+1	0	1	0	1	0	1	1	1	0	0	F	F	F	C	C	C
+0	0	1	0	1	0	1	1	0	1	0	F	F	F	C	C	C
-0	0	1	0	1	0	1	0	1	0	1	F	F	F	C	C	C
-1	0	1	0	1	0	1	0	0	1	1	F	F	F	C	C	C
-2	1	0	0	1	0	1	0	0	1	1	R	F	F	D	C	C
-3	1	0	0	1	1	0	0	0	1	1	R	R	F	D	D	C
-4	1	0	1	0	0	1	0	0	1	1	R	R	R	D	D	D

Extended Topology

A. Extended 9-Level Topology

As mentioned before, using multiple DC sources to get more levels in the output voltage is one of demerits of the multilevel inverters. However, in the extended topology of the suggested multilevel inverter, increasing the output voltage levels to 9 with only one DC sources is provided as demonstrated in Fig. 4. The introduced topology uses only one more capacitor, one more diode, and two additional switches compared to 7-Level topology, which increases the output voltage up to 4 times the input voltage. The switching modes for 9-Level topology are described in Table 2.

B. Extended 11-Level Topology

Fig. 5 illustrates 11-level topology of suggested multilevel inverter that is constructed by adding capacitor C_4, diodes D_4 and D_5, and switch S_7 compare to 9-level topology. This configuration increases the output voltage up to 5 times the input voltage with only one DC voltage source. The 11-level topology switching algorithm is listed in Table 3.

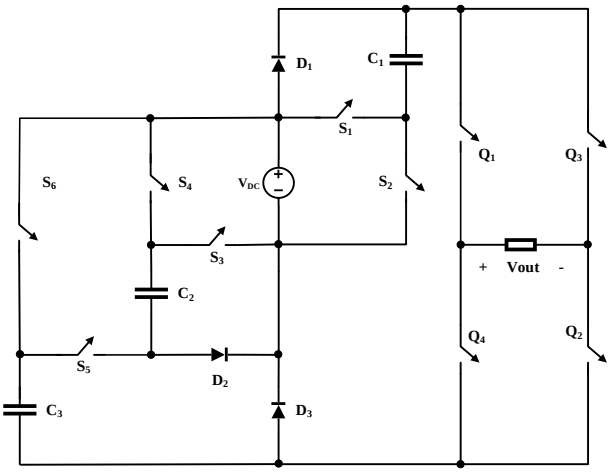


Fig. 4: Extended 9-level topology.

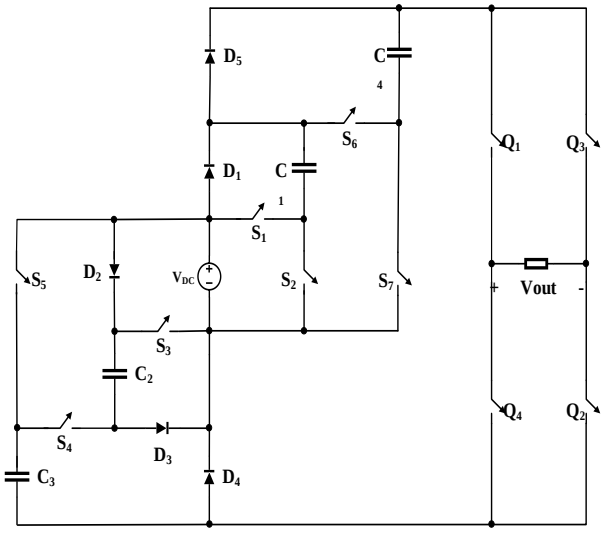


Fig. 5: Extended 11-level topology.

Capacitors Size Calculations

To determine the optimum value of capacitors, it is necessary to calculate the longest discharge period (LDP) for a given capacitor per switching cycle. During the LDP, the reserved energy within the capacitors is released and transferred to the load, resulting in generation of a specific voltage level. The LDP amounts for capacitors C_1 and C_2 are presented in Fig. 6, as indicated by the data given in Table 1.

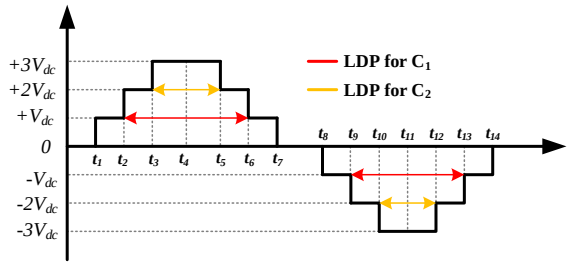


Fig. 6: LDP of capacitors.

Table 3: Circuit elements states for 11-level topology operation

$\frac{V_o}{V_{in}}$	Capacitor state																			
	S_1	S_2	S_3	S_4	S_5	S_6	S_7	Q_1	Q_2	Q_3	Q_4	D_1	D_2	D_3	D_4	D_5	C_1	C_2	C_3	C_4
+5	1	0	1	1	0	1	0	1	1	0	0	R	R	R	R	R	D	D	D	D
+4	1	0	0	0	1	1	0	1	1	0	0	R	R	R	R	F	D	D	D	C
+3	1	0	0	1	0	1	0	1	1	0	0	R	R	R	F	F	D	D	C	C
+2	0	1	0	1	0	1	0	1	1	0	0	R	F	F	F	F	D	C	C	C
+1	0	1	0	1	0	0	1	1	1	0	0	F	F	F	F	F	C	C	C	C
+0	0	1	0	1	0	0	1	1	0	1	0	F	F	F	F	F	C	C	C	C
-0	0	1	0	1	0	0	1	0	1	0	1	F	F	F	F	F	C	C	C	C
-1	0	1	0	1	0	0	1	0	0	1	1	F	F	F	F	F	C	C	C	C
-2	0	1	0	1	0	1	0	0	0	1	1	R	F	F	F	F	D	C	C	C
-3	1	0	0	1	0	1	0	0	0	1	1	R	R	R	F	F	D	D	C	C
-4	1	0	0	0	1	1	0	0	0	1	1	R	R	R	R	F	D	D	D	C
-5	1	0	1	1	0	1	0	0	0	1	1	R	R	R	R	R	D	D	D	D

Time t_i is the transit time between two levels with different values.

According to Fig. 6, it can be concluded that LDP for capacitor C_1 is equal to (t_2-t_6) or (t_9-t_{13}), and for capacitor C_2 is equal to (t_3-t_5) or (t_{10}-t_{12}).

During LDP, the amount of charge transferred from capacitors C_1 and C_2 (Q_{C1}, Q_{C2}) is calculated as given in (2).

$$Q_{C_1} = 2 \times \int_{t_2}^{t_4} i_o(t)dt \quad ; Q_{C_2} = 2 \times \int_{t_3}^{t_4} i_o(t)dt \qquad (2)$$

where i_O is the load current or the capacitor discharge current during the LDP. Considering the specific capacitor ripple voltage (σ), the optimum capacitance for C_1 and C_2 is obtained from (3).

$$C_1 \geq \frac{Q_{C_1}}{\sigma \times V_{dc}} \quad ; C_{12} \geq \frac{Q_{C_2}}{\sigma \times V_{dc}} \qquad (3)$$

For resistive load, i_O in LDP is determined as given in the following:

$$i_o(t) = \begin{cases} 2\frac{V_{dc}}{R}; & for \ \ t_2 \leq t < t_3 \\ 3\frac{V_{dc}}{R} \ ; & for \ \ t_3 \leq t < t_4 \end{cases} \qquad (4)$$

Transit times t_2, t_3, and t_4 are obtained from (5) considering modulation with fundamental switching frequency and unit modulation index.

$$t_2 = \frac{\sin^{-1}(1/2)}{2\times\pi\times f};\ \ t_3 = \frac{\sin^{-1}(5/6)}{2\times\pi\times f};\ \ t_4 = T \quad (5)$$

where f is the switching frequency, and T represents the periodicity of the output voltage, by using (2)-(5). The optimum value of capacitance is calculated as determined in the following:

$$C_1 \geq \frac{2.67}{\pi\times f\times\sigma\times R};\ \ C_2 \geq \frac{1.76}{\pi\times f\times\sigma\times R} \quad (6)$$

Power Loss Calculations

The dominant power losses related to the introduced inverter are; a) conductive losses of switches, diodes, and capacitors; and b) voltage ripple losses of capacitors.

As mentioned before, the suggested inverter is modulated with a low switching frequency so, the switching loss is negligible. However, the conductive loss is calculated by considering on-state resistance of the switches, r_{on}, and diode, r_d, and the equivalent series resistance of capacitors, r_c. Conduction loss, $P_{cond.i}$, for the i^{th} voltage level is obtained as given in (7).

$$P_{cond.i} = r_{eq.i} \times i_{o.i}^2 \quad (7)$$

where $r_{eq.i}$ is the equivalent series resistance in the charging path of the output current and i_{oi} is the load current for the i^{th} voltage level. Because the i^{th} voltage level is repeated 4 times in each switching cycle, the average of conduction loss per cycle is calculated as given in the following:

$$P_{cond.ave.i} = \frac{4(t_{i+1}-t_i)}{T} P_{cond.i} \quad (8)$$

where (t_i+1 - t_i) is the time interval of the i^{th} voltage level. Similarly, for each voltage level, the average conduction loss of the inverter switches should be calculated.

The summation of conduction losses serves as an estimation for the overall power loss incurred by the inverter.

Ripple losses of capacitor manifest during the charge phase of the capacitor. The magnitude of the loss is contingent upon the variation between the input voltage and the instantaneous voltage across the capacitor during the charging process, as well as the capacitance value.

The equation representing the losses for the j^{th} capacitor is depicted in (9). The expression (t_j+1 - t_j) represents the duration of the charging time, while $iC_j(t)$ denotes the charging current of the j^{th} capacitor. The total ripple losses of the inverter are determined by the summation of the ripple losses of all capacitors that are utilized.

$$P_{rip.j} = \frac{1}{2T} C_j \left[\Delta V_{C_j}\right]^2 = \frac{1}{2T} C_j \left[\int_{t_j}^{t_{j+1}} i_{C_j}(t)dt\right]^2 \quad (9)$$

Therefore, total power loss of the inverter is obtained as given in the following:

$$P_{total} = P_{cond.ave} + P_{rip.j} \quad (10)$$

Comparative Analysis

This section presents a comprehensive comparative study among the proposed topologies and several recently discovered SC-based multilevel inverters. The study highlights the distinctive advantages and disadvantages of the proposed topologies, providing robust evidence to support the superiority of the suggested topologies over other competing alternatives. In Table 4, a comparison is performed by considering items such as number of voltage sources (N_{dc}), diodes (N_D), capacitors (N_C), and switches (N_{SW}), as well as the boost factor (BF) and TSV of all switches. In addition, a general comparison of the topologies introduced in [22]-[30] has been included in Fig. 7.

According to comparison results, proposed 7-level topology, despite having a boost factor equal to [22], requires fewer switches due to the reduced number of gate drivers. As a result, it costs less and is considered superior to other topologies.

In the configuration proposed in [23], there are more passive elements, and TSV on the switches is also high in compared with suggested 7-level topology. The values of N_D, N_C, and BF are the same in both proposed 7-level and [24] configurations. Additionally, the voltage stress on the switches in [24] is higher than proposed 7-level topology.

Compared to the topologies introduced in [25]-[27], the suggested 9-level topology has less voltage stress and fewer passive elements than [25]-[27] topologies.

Table 4: Comparative analysis of proposed topologies with similar SC-MLI

Top	N_L	N_{dc}	N_{SW}	N_D	N_C	BF	TSV
[22]	7	1	10	-	2	3	4.33
[23]	7	1	6	4	4	3	6
[24]	7	1	8	2	2	3	6
[25]	9	1	8	6	3	4	6
[26]	9	1	13	3	3	4	5.25
[27]	9	1	17	4	4	4	12.25
[28]	11	1	14	2	3	2.5	3
[29]	11	1	11	5	4	5	6.6
[30]	11	1	17	1	5	5	7.14
[7-Level topology]	7	1	8	2	2	3	5.33
[9-Level topology]	9	1	10	3	3	4	5.5
[11-Level topology]	11	1	11	5	4	5	5.4

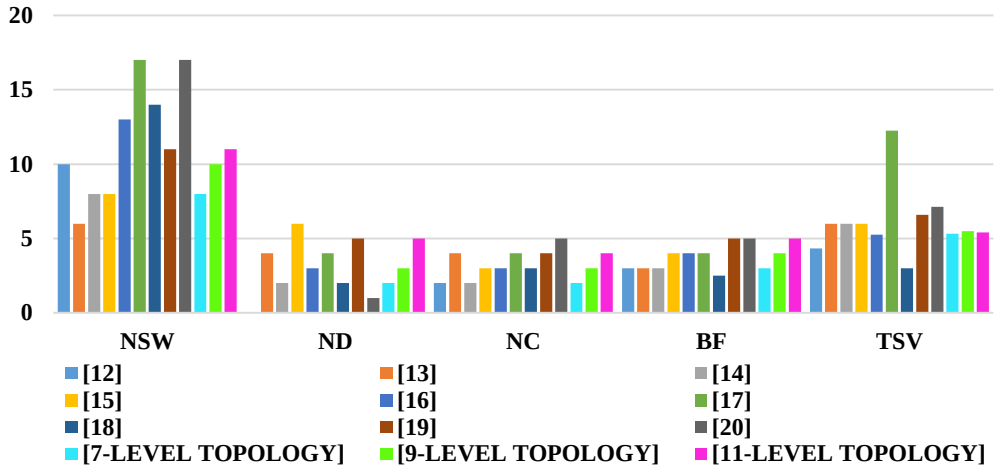


Fig. 7: Comparative analysis chart.

All these factors make proposed 9-level topology superior to the others. The topology given in [26] has lower voltage stress than proposed 9-level topology however, its demerit is applying more switches.

Although, the topology suggested in [28] has more switches than proposed 11-level topology, it has lower boost factor.

On the other hand, the topology introduced in [29] has same switches and diodes in compared to proposed 11-level topology, but it has more TSV. Despite, the combination of lower N_{SW}, N_C and TSV make the given 11-level topology superior to the topology suggested in [30].

Results and Discussion

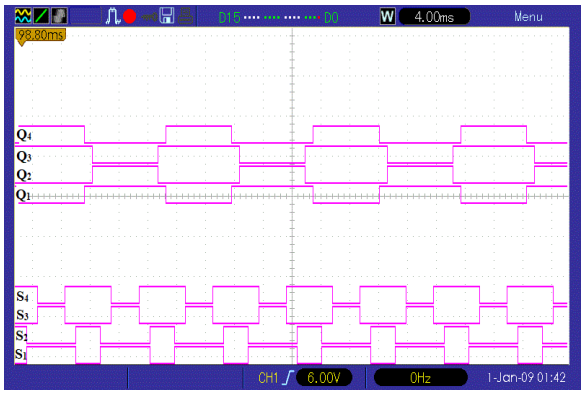
To verify practicability of the introduced topologies, a 7-level topology experimental prototype is implemented as depicted in Fig. 8. Table 5 lists the elements were employed in the laboratory prototype.

Fig. 8: Hardware setup

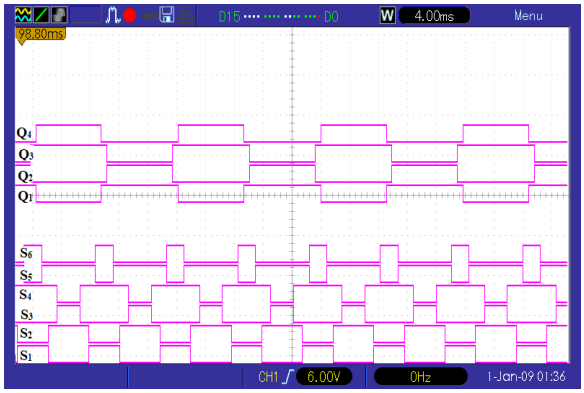
Table 5 Prototype circuit elements values

Parameter	value
Input DC sources	V_{dc}=30 V
MOSFETs type	IRFP260n
Output frequency	f_o=50 Hz
Opto-coupler	TLP250
Capacitances	4700µf
Load	R=100Ω, L=50mH
Microcontroller	Atmega8a

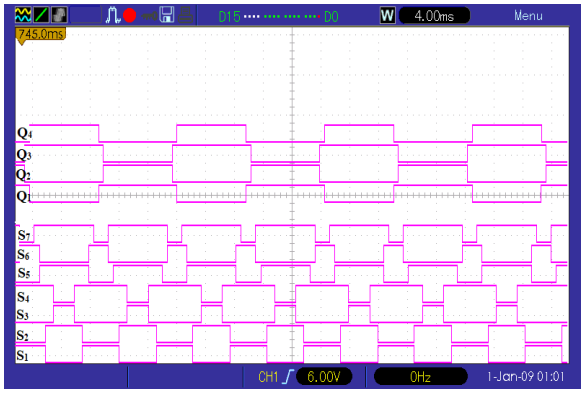
Atmega8a microcontroller is utilized to generate the gate signals of the implemented inverter switches that is shown in Fig. 9. Fig. 10(a)-(c) display the voltage and current waveforms of the resistive-inductive load (100Ω - 50 mH), respectively. The DC source has the value of 30 volts, consequently, the output voltage range is some lower than theoretical value due to losses incurred by the switches and diodes. The output current in Fig. 10 exhibits a sinusoidal waveform due to the presence of an inductive load within the circuit. Fig. 11 shows dynamic operation of the 11-level topology during load change from resistive-inductive to resistive. As shown in this figure, the suggested topology response to the load change is without performance failure. Among the proposed topologies, it is noteworthy that capacitor C_1 in 7, 9, and 11-level configurations holds the highest level of significance due to its LDP. Therefore, it is crucial to consider this aspect while deciding on the final configuration. Fig. 12 depicts the maximum voltage of capacitors C_1 in the suggested 11-level topology. The output voltage of the capacitor has a self-balancing profile, which indicates that the capacitors experience minimal ripple.



(a)



(b)



(c)

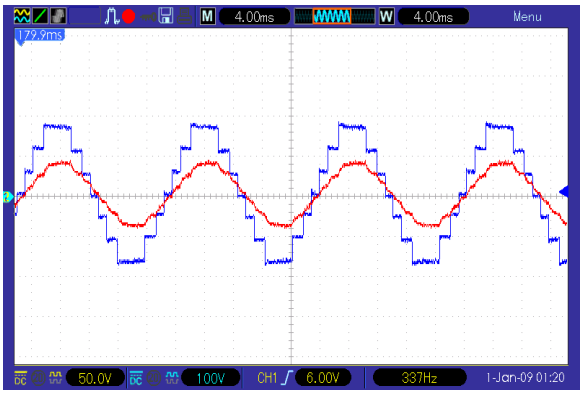
Fig. 9: Gate signals of proposed (a) 7-level, (b) 9-level, and (c) 11-level topologies.

Conclusions

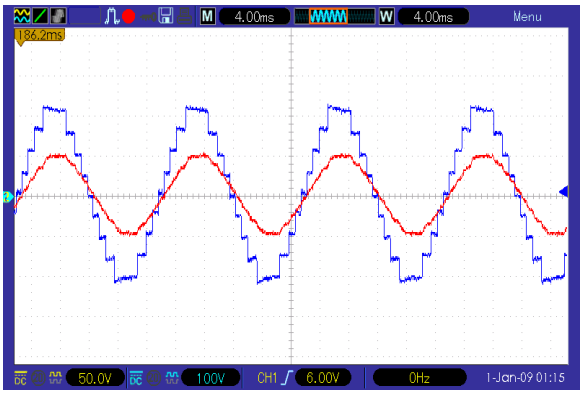
This paper has presented modified switch-capacitor multilevel inverter topologies with 7, 9, and 11 levels in output voltage. All topologies have the potential to be used in renewable energy integration and electric vehicles.

Comprehensive examinations of circuit performance, voltage stress, capacitor sizing, and power loss were presented. The comparative assessments demonstrated that the suggested configurations achieve a greater voltage boost while utilizing fewer elements in comparison to the most advanced solutions available. The experimental test results confirmed the accuracy of the theoretical statements regarding the multilevel output voltage waveforms and the ability to balance capacitor voltages.

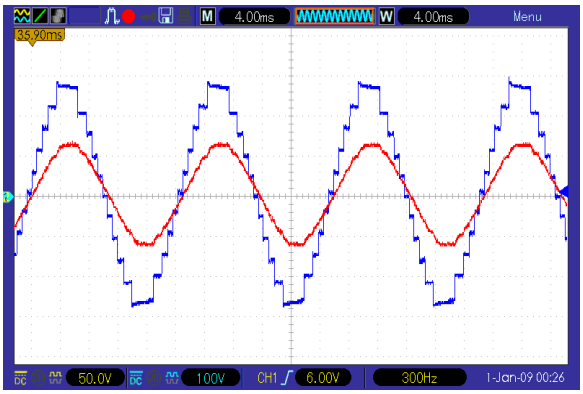
Therefore, the suggested SC multilevel inverters show great potential as solutions for applications such as solar photovoltaic systems and electric vehicle battery banks. Subsequent research will highlight the optimization of topologies for particular power ratings and the execution of efficiency tests.



(a)



(b)



(c)

Fig. 10: Experimental test results; voltage and current waveforms of resistive-inductive load for (a) 7-Level inverter (b) 9-level inverter and (c) 11-level inverter.

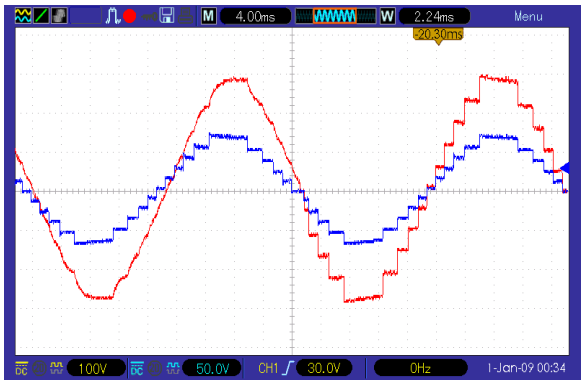
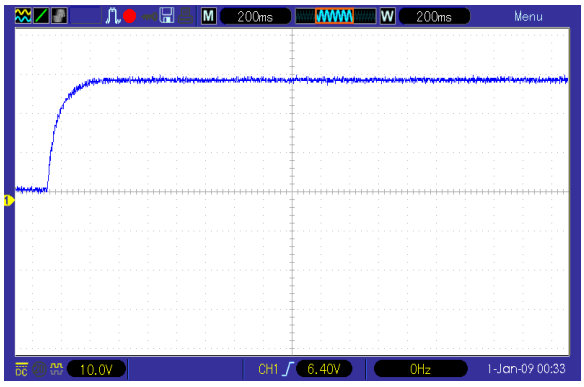
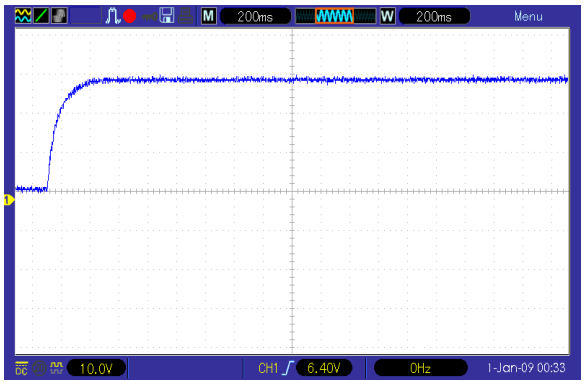


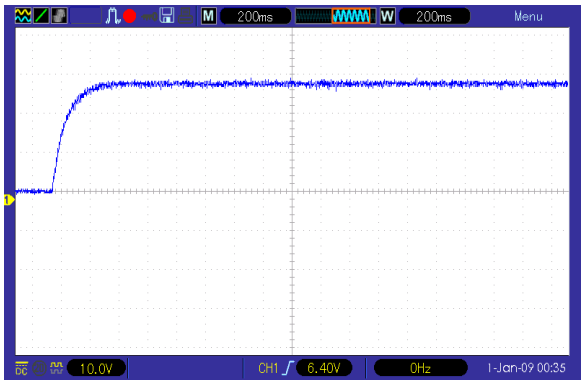
Fig. 11: Experimental test results of voltage and current waveforms for load change from resistive-inductive to resistive.



(a)



(b)



(c)

Fig. 12: Experimental test results of voltage waveforms for Capacitors C1 (a) 7-level topology, (b) 9-level topology, (c) 11-level topology.

Author Contributions

F. Sedaghati chose the field of research. S. Ebrahimzadeh and H. Dolati collected information in this field. F. Sedaghati presented the proposed topology. S. Ebrahimzadeh and H. Dolati simulated and fabricated the proposed converter converter. The authors discussed the obtained results and drew conclusions. Under the supervision of F. Sadaghti, the text of the article was prepared by S. Ebrahimzadeh and H. Dolati. F. Sedaghati submitted the manuscript.

Acknowledgment

I appreciate the referees and their colleagues who helped the authors in publishing this article.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

PV	Photovoltaic
EV	Electric Vehicle
PWM	Pulse Width Modulation
SC	Switched Capacitor
MLI	Multilevel inverter
THD	Total Harmonic Distortion
TSV	Total Standing Voltage
LDP	Longest Discharge Period
BF	Boost Factor

References

[1] B. Saumya, Y. Zong, Sh You, L. Mihet-Popa, J. Xiao, “Technical and economic analysis of one-stop charging stations for battery and fuel cell EV with renewable energy sources,” Energies, 13(11): 2855, 2020.

[2] M. A. Velasquez, J. Barreiro-Gomez, N. Quijano, A. I. Cadena, M. Shahidehpour, “Distributed model predictive control for economic dispatch of power systems with high penetration of renewable energy resources,” Int. J. Electr. Power Energy Syst., 113: 607-617, 2019.

[3] L. He, J. Sun, Z. Lin, B. Cheng, “Capacitor-voltage self-balance seven-level inverter with unequal amplitude carrier-based APODPWM,” IEEE Trans. Power Electron., 36(12): 14002-14013, 2021.

[4] S. B. Kjaer, J. K. Pedersen, F. Blaabjerg, “A review of single-phase grid-connected inverters for photovoltaic modules,” IEEE Trans. Ind. Appl., 41(5): 1292-1306, 2005.

[5] T. Jin, X. Yan, H. Li, J. Lin, Y. Weng, Y. Zhang, "A new three-winding coupled inductor high step-up DC–DC converter integrating with switched-capacitor technique," IEEE Trans. Power Electron., 38(11): 14236-14248, 2023.

[6] S. J. Salehi, M. A. Shmasi-Nejad, H. R. Najafi, “A new generalized step-up multilevel inverter topology based on combined t-type and cross capacitor modules,” Int. J. Eng., 36(7): 1354-1368, 2023.

[7] K. M. Nagabushanam, T. Mahto, S. V. Tewari, R. R. Udumula, M. A. Alotaibi, H. Malik, F. P. G. Márquez, “Development of high-gain switched-capacitor based bi-directional converter for electric vehicle applications,” J. Energy Storage, 82: 110602, 2024.

[8] S. Chen, Y. Ye, S. Chen, T. Hua, X. Wang, “Three-phase boost multilevel inverter based on coupled-structure switched-capacitor

and V2SVM," IEEE J. Emerg. Sel. Top. Power Electron., 11(1): 679-690, 2023.

[9] H. Dolati, E. Babaei, S. Ebrahimzade, "Reduced switch count single-source seven-level switched-capacitor boost multilevel inverter with extendibility," Iran. J. Sci. Technol. Trans. Electr. Eng., 48: 1313-1321,2024.

[10] S. K. Baksi, R. K. Behera, U. R. Muduli, "Optimized 9-level switched-capacitor inverter for grid-connected photovoltaic systems," IEEE Trans. Ind. Appl., 60(2): 3284-3296, 2024.

[11] K. Jena, D. Kumar, B. H. Kumar, K. Janardhan, A. R. Singh, R. Naidoo, R. C. Bansal, "A single DC source generalized switched capacitors multilevel inverter with minimal component count," Int. Trans. Electr. Energy Syst., 3945160, 7: 1-12, 2023.

[12] R. Barzegarkhoo, M. Forouzesh, S. S. Lee, F. Blaabjerg, Y. P. Siwakoti, "Switched-capacitor multilevel inverters: A comprehensive review," IEEE Trans. Power Electron., 37(9): 11209-11243, 2022.

[13] M. V. Patel, M. L. Patel, N. D. Shah, "A review of Switched Capacitor (SC) circuits," Int. J. Sci. Res., 2(1): 81-83, 2012.

[14] B. M. Varghese, B. M. Jos, "Switched capacitor multilevel inverter with different modulation techniques," in Proc. International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS): 1-6, 2017.

[15] D. Singh, N. Sandeep, "Switched-capacitor-based multi-source multilevel inverter with reduced part count," IEEE J. Emerg. Sel. Top. Ind. Electron., 4(3): 718-724, 2023.

[16] M. Ali, M. Tayyab, A. Sarwar, M. Khalid, "A low switch count 13-level switched-capacitor inverter with hexad voltage-boosting for renewable energy integration," IEEE Access, 11: 36300-36308, 2023.

[17] K. Jena, D. Kumar, K. Janardhan, B. H. Kumar, A. R. Singh, S. Nikolovski, M.Bajaj, "A novel three-phase switched-capacitor five-level multilevel inverter with reduced components and self-balancing ability," Appl. Sci., 13(3): 1713, 2022.

[18] P. Kumari et al, "Self-balanced high gain switched-capacitor boosting inverter with lower cost function," Int. J. Electron., 111(8): 1301–1318, 2023.

[19] Y. Wang, J. Ye, R. Ku et al., "A modular switched-capacitor multilevel inverter featuring voltage gain ability," J. Power Electron. 23: 11-22, 2023.

[20] M. N. H. Khan, R. Barzegarkhoo, Y. P. Siwakoti, S. A. Khan, L. Li, F. Blaabjerg, "A new switched-capacitor multilevel inverter with soft start and quasi resonant charging capabilities," Int. J. Electr. Power Energy Syst., 135: 107412, 2022.

[21] M. A. Hosseinzadeh, M. Sarebanzadeh, C. Garcia, E. Babaei, J. Rodriguez, "Efficient switched-capacitor multilevel inverters for high-power solar photovoltaic systems," IET Renewable Power Gener., 16(11): 2248-2266, 2022.

[22] F. Sedaghati, S. Ebrahimzadeh, H. Dolati, H. Shayeghi, "A modified switched capacitor multilevel inverter with symmetric and asymmetric extendable configurations," J. Oper. Autom. Power Eng., 13(1): 20-27, 2025.

[23] M. Chen, P. C. Loh, Y. Yang, F. Blaabjerg, "A six-switch seven-level triple-boost inverter," IEEE Trans. Power Electron., 36(2): 1225 - 1230, 2020.

[24] Y. P. Siwakoti, A. Mahajan, D. J. Rogers, F. Blaabjerg, "A novel seven-level active neutral-point-clamped converter with reduced active switching devices and DC-link voltage," IEEE Trans. Power Electron., 34(11): 10492-10508, 2019.

[25] Y. Wang, K. Wang, G. Li, F. Wu, K. Wang, J. Liang, "Generalized switched-capacitor step-up multilevel inverter employing single DC Source," CSEE J. Power Energy Syst., 8(2): 439-451, 2022.

[26] S. Islam, M. D. Siddique, A. Iqbal, S. Mekhilef, "A 9- and 13-level switched-capacitor-based multilevel inverter with enhanced self-balanced capacitor voltage capability," IEEE J. Emerg. Sel. Top. Power Electron., 10(6): 7225-7237,2022.

[27] H. Khoun Jahan, M. Abapour, K. Zare, "Switched-capacitor-based single-source cascaded h-bridge multilevel inverter featuring boosting ability," IEEE Trans. Power Electron., 34(2): 1113-1124, 2019.

[28] M. R. Hussan, A. Sarwar, I. Khan, M. Tariq, M. Tayyab, W. Alhosaini, "An eleven-level switched-capacitor inverter with boosting capability," Electronics, 10(18): 2262, 2021.

[29] S. Deliri, K. Varesi, S. Padmanaban, "An extendable single-input reduced-switch 11-level switched-capacitor inverter with quintuple boosting factor," IET Gener. Transm. Distrib., 17(3): 621-631, 2023.

[30] M. N. H. Khan, M. Forouzesh, Y. P. Siwakoti, L. Li, F. Blaabjerg, "Switched capacitor integrated (2n + 1)-level step-up single-phase inverter," IEEE Trans. Power Electron., 35(8): 8248-8260, 2020.

Biographies

Farzad Sedaghati was born in Ardabil, Iran, in 1984. He received the M.S. and Ph.D. degrees both in Electrical Engineering in 2010 and 2014 from the University of Tabriz, Tabriz, Iran. In 2014, he joined the Faculty of Engineering, University of Mohaghegh Ardabili, where he has been an Assistant Professor, since 2014. Also, he is Associate Professor since 2019. His current research interests include power electronic converters design and applications and renewable energies.

- Email: farzad.sedaghati@uma.ac.ir
- ORCID: 0000-0001-6974-4719
- Web of Science Researcher ID: NA
- Scopus Author ID: 35410298600
- Homepage: https://academics.uma.ac.ir/profiles?Id=617

Soghra Ebrahimzadeh was born in Ardabil, Iran in 1994. She obtained her B.Sc. and M.Sc. degrees in Power Electrical Engineering University of Mohaghegh Ardabili in 2016 and 2019, respectively. Currently, she is pursuing a Ph.D. degree in Electrical Engineering at the University of Mohaghegh Ardabili, Ardabil, Iran. Her current research interests include switched-capacitor multilevel inverters and grid-tied multilevel inverters.

- Email: soghraebrahimzade2@gmail.com
- ORCID: 0009-0008-6726-1519
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: N/A

Hadi Dolati was born in Ardabil, Iran, in 1997. He received his B.Sc. degree in Electronic Engineering from the University of Mohaghegh Ardabili, Ardabil, Iran, and the M.Sc. degree in electrical engineering from the University of Tabriz, Tabriz, Iran, in 2019, and 2024, respectively. His current research interest include design, control, and applications of power electronics converters.

- Email: hadidolati1997@gmail.com
- ORCID: 0009-0001-4245-7311
- Web of Science Researcher ID: NA
- Scopus Author ID: N/A
- Homepage: N/A

How to cite this paper:
F. Sedaghati, S. Ebrahimzadeh, H. Dolati, "Modified topologies for single source switched-capacitor multilevel inverters," J. Electr. Comput. Eng. Innovations, 13(1): 257-266, 2025.

DOI: 10.22061/jecei.2024.11234.780

URL: https://jecei.sru.ac.ir/article_2231.html

SRTTU

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

JECEI

PAPER TYPE? (Research paper, short paper, Review paper *et al.*)

Instructions and Formatting Rules for Authors of Journal of Electrical and Computer Engineering Innovations, JECEI

F. Author, S. Author, T. Author

Affiliations of the Authors: (Department, Faculty, University (Institution), City, Country)

Article Info

Article History:
Received
Reviewed
Revised
Accepted

Keywords:
The author(s) shall provide up to 6 keywords to help identify the major topics of the paper

*Corresponding Author's Email Address:

Abstract

Background and Objectives: This section should be the shortest part of the abstract and should very briefly outline the following information: 1-What is already known about the subject, related to the paper in question. 2- What is not known about the subject and hence what the study intended to examine (or what the paper seeks to present). In most cases, the background can be framed in just 2–3 sentences, with each sentence describing a different aspect of the information referred to above; sometimes, even a single sentence may suffice. The purpose of the background, as the word itself indicates, is to provide the reader with a background to the study, and hence to smoothly lead into a description of the methods employed in the investigation.
Methods: The methods section is usually the second-longest section in the abstract. It should contain enough information to enable the reader to understand what was done, and how.
Results: The results section is the most important part of the abstract and nothing should compromise its range and quality. This is because readers who peruse an abstract do so to learn about the findings of the study. The results section should therefore be the longest part of the abstract and should contain as much detail about the findings as the journal word count permits.
Conclusion: This section should contain the most important take-home message of the study, expressed in a few precisely worded sentences. Usually, the finding highlighted here relates to the primary outcome measure; however, other important or unexpected findings should also be mentioned. It is also customary, but not essential, for the authors to express an opinion about the theoretical or practical implications of the findings, or the importance of their findings for the field. Thus, the conclusions may contain three elements: 1- The primary take-home message 2-The additional findings of importance 3-The perspective.

Introduction

This document provides an example of the desired layout for JECEI paper and can be used as a template for Microsoft Word versions 2003 and later. It contains information regarding desktop publishing format, type sizes, and typefaces. Style rules are provided to explain how to handle equations, units, figures, tables, abbreviations, and acronyms. Sections are also devoted to the preparation of appendixes, acknowledgments, references, and authors' biographies. For additional information including electronic file requirements for text and graphics, please refer to www.autjournal.com.

Technical Work Preparation

Please use automatic hyphenation and check your spelling. Additionally, be sure your sentences are complete and that there is continuity within your paragraphs. Check the numbering of your graphics and make sure that all appropriate references are included.

Doi:

A. Template

This document may be used as a template for preparing your technical work.

B. Format

If you choose not to use this document as a template, prepare your technical work in single-spaced, double-column format, on paper 21.6×27.9 centimeters (8.5×11 inches or 51×66 picas). Set top and bottom margins to 25 millimeters (0.98 inch) and left and right margins to about 20 millimeters (0.79 inch). Do not violate margins (i.e., text, tables, figures, and equations may not extend into the margins). The column width is 82 millimeters (3.2 inches). The space between the two columns is 6 millimeters (0.24 inch). Paragraph indentation is 4.2 millimeters (0.17 inch). Use full justification. Use either one or two spaces between sections, and between text and tables or figures, to adjust the column length.

C. Typefaces and Sizes

Please use a proportional serif typeface such as Calibri and embed all fonts. Table 1 provides samples of the appropriate type sizes and styles to use.

D. Section Headings

A primary section heading is enumerated by a Roman numeral followed by a period and is centered above the text. A primary heading should be in capital letters.

A secondary section heading is enumerated by a capital letter followed by a period and is flush left above the section. The first letter of each important word is capitalized and the heading is italicized.

A tertiary section heading is enumerated by an Arabic numeral followed by a parenthesis. It is indented and is followed by a colon. The first letter of each important word is capitalized and the heading is italicized.

A quaternary section heading is rarely necessary, but is perfectly acceptable if required. It is enumerated by a lowercase letter followed by a parenthesis. It is indented and is followed by a colon. Only the first letter of the heading is capitalized and the heading is italicized.

E. Figures and Tables

Figure axis labels are often a source of confusion. Try to use words rather than symbols. As an example, write the quantity "Magnetization," or "Magnetization, *M*," not just "*M*." Put units in parentheses. Do not label axes only with units. As in Fig. 1, write "Magnetization (kA/m)" or "Magnetization (kA·m^{-1})," not just "kA/m." Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)," not "Temperature/K." Figure labels should be legible, approximately 8- to 10-point type.

Large figures and tables may span both columns, but may not extend into the page margins. Arrange these one column figures and tables at either top or end of a page, or at the end of the paper right before the references. Figure captions should be below the figures; table captions should be above the tables. Do not put captions in "text boxes" linked to the figures. Do not put borders around your figures. Use Insert | Reference | Caption to number your tables and figures, and use Insert | Reference | Cross- reference to refer to their numbers.

Table 1: Samples of Calibri sizes and styles used for formatting a pes technical work

Point Size	Purpose in Paper	Special Appearance
9	Table text, figure text footnotes, subscripts, superscripts, references, bio, Figure caption, keywords	Table Title
10	Body text, equations, author affiliation, abstract	*Subheadings*
11		**Section Titles**
12	***Author Name***	

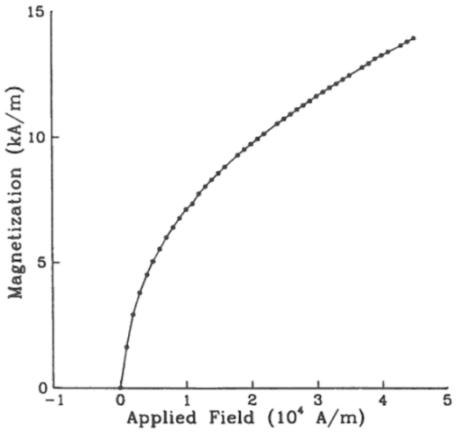


Fig. 1: Magnetization as a function of applied field. (Note that there is a colon after the figure number followed by two spaces.)

All figures and tables must appear near, but not before, their first mention in the text. Use the abbreviation "Fig. 1," even at the beginning of a sentence.

To insert images in Word, use Insert | Picture | From File.

F. Numbering

Number reference citations consecutively in square brackets [1]. The sentence punctuation follows the brackets [2]. Multiple references [2], [3] are each numbered with separate brackets [1][1]-[2]. Refer simply to the reference number, as in [2]. Do not use "Ref. [2]" or "reference [2]" except at the beginning of a sentence: "Reference [2] shows…."

Number footnotes separately with superscripts

(Insert | Footnote). Place the actual footnote at the bottom of the column in which it is cited. Do not put footnotes in the reference list. Use letters for table footnotes.

Use Arabic numerals for figures and Roman numerals for tables. Appendix figures and tables should be numbered consecutively with the figures and tables appearing in the rest of the paper. They should not have their own numbering system.

G. Units

Metric units are preferred for use in IEEE publications in light of their global readership and the inherent convenience of these units in many fields. In particular, the use of the International System of Units is advocated. This system includes a subsystem of units based on the meter, kilogram, second, and ampere (MKSA). British units may be used as secondary units (in parentheses). An exception is when British units are used as identifiers in trade, such as 3.5-inch disk drive.

H. Math and Equations

Number equations consecutively with equation numbers in parentheses flush with the right margin, as in (1). First use the equation editor to create the equation. Then select the "Equation" markup style. Write the equation number in parentheses using Insert | Caption.

Use the Microsoft Equation Editor for all math objects in your paper (Insert | Object | Create New | Microsoft Equation *or* MathType Equation). "Float over text" should *not* be selected.

To make your equations more compact, you may use the slash (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Use parentheses to avoid ambiguities in denominators. Number equations consecutively with equation numbers in parentheses flush with the right margin, as in (1). Be sure that the symbols in your equation have been defined before the equation appears or immediately following. Italicize symbols (*T* might refer to temperature, but T is the unit Tesla).

Use Insert | Reference | Caption to number equations. Refer to "(1)," not "Eq. 1" or "equation (1)," except at the beginning of a sentence: "Equation (1) is ...". Punctuate equations when they are part of a sentence, as in

$$\int_0^{r_2} F(r,\varphi)\, dr\, d\varphi = [\sigma r_2 / (2\mu_0)] \cdot \int_0^{\infty} \exp(-\lambda\,|\,z_j - z_i\,|)\,\lambda^{-1}\, J_1(\lambda r_2)\, J_0(\lambda r_i)\, d\lambda \quad (1)$$

Use two column tables to locate equations and their numbers properly in one line, as follows:

$$I_F = I_B = -I_C = A^2 I_{A1} + A I_{A2} + I_{A0} = \frac{-J\sqrt{3}E_A}{Z_1 + Z_2} \quad (2)$$

where I_F is the fault current. Be sure that the border is off.

Results and Discussion

The Results section should briefly present the experimental data in text, tables or figures. Tables and figures should not be described extensively in the text.

The Discussion should focus on the interpretation and the significance of the findings with concise objective comments that describe their relation to other work in the area. It should not repeat information in the results. The final paragraph should highlight the main conclusion(s), and provide some indication of the direction future research should take.

Conclusion

As the Conclusion section is the most important element of a manuscript, so it must be more expanded scientifically and contently at least half a page length.

Example:

In this study, a forecast model was developed to determine the generation of MSW in the municipalities of the CCS, Chiapas State, Mexico. A MLR was used to obtain the forecast model with social and demographic explanatory variables. Two forecast models were presented and analyzed, with variables that met the multicollinearity test. The most important variables to predict the rate of MSW generation in the study area were the population of each municipality (XPop), the population born in another municipality (XPbam) and the population density (XPd). XPop is the most influential explanatory variable of waste generation, particularly it is related in a positive way. XPbam is less related to waste generation. XPd is the variable that least influences waste generation prediction; in addition, it can present problems of correlation with other explanatory variables. Although other variables, such as daily per capita income (XDpi) and average schooling (XAs), are very important, they do not seem to have an effect on the response variable in this study. The user of this forecast model should use model 2, since it is the one with the highest parsimony (it uses fewer variables); R^2_{adj}, MAPE, MAD and RMSE values indicated high influence on the explained phenomenon and high forecasting capacity. Additionally, it is important to mention that when using the models proposed for forecasting purposes, it is necessary to make a transformation in the explanatory and response variables (use inverse of natural logarithm). The inferences made on the municipalities of the study area showed that, except in some municipalities, the MSW generation rate usually presented a gradual increase

with respect to population growth and with respect to the number of inhabitants that were born in another entity (migration). Finally, this study can be a solid basis for comparison for future research in the area of study. It is possible to use different mathematical models such as artificial neural network, principal component analysis, time-series analysis, etc., and compare the response variable or the predictors.

Author Contributions

Each author role in the research participation must be mentioned clearly.

Example:

A. Mahboobi, B. Bagheri, and C. Ahmdi designed the experiments. A. Mahboobi collected the data. A. Mahboobi carried out the data analysis. A. Mahboobi, B. Bagheri, and C. Ahmdi interpreted the results and wrote the manuscript.

Acknowledgment

The following is an example of an acknowledgment. (Please note that financial support should be acknowledged in the unnumbered footnote on the title page.)

The author gratefully acknowledges the IEEE I. X. Austan, A. H. Burgmeyer, C. J. Essel, and S. H. Gold for their work on the original version of this document.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

Define less common abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title unless they are unavoidable.

Example:

MS	Multispectral
SMF	Spectral Matched Filter
SAM	Spectral Angle Mapper
MSD	Matched Subspace Detector
OSP	Orthogonal Subspace Projection
CEM	Constrained Energy Minimization
ASD	Adaptive Subspace Detector
STD	Sparsity Based Target Detector
KSAM	Kernel Based SAM
DTD	Difference Based Target Detection
AP-CR	Attribute Profile Based Collaborative Representation
ROC	Receiver Operating Characteristic

MS	Multispectral
SMF	Spectral Matched Filter
SAM	Spectral Angle Mapper
MSD	Matched Subspace Detector
OSP	Orthogonal Subspace Projection
CEM	Constrained Energy Minimization
ASD	Adaptive Subspace Detector
STD	Sparsity Based Target Detector
KSAM	Kernel Based SAM
DTD	Difference Based Target Detection
AP-CR	Attribute Profile Based Collaborative Representation
ROC	Receiver Operating Characteristic

References

References are important to the reader; therefore, each citation must be complete and correct. There is no editorial check on references; therefore, an incomplete or wrong reference will be published unless caught by a reviewer or discusser and will detract from the authority and value of the paper. References should be readily available publications. List only one reference per reference number. If a reference is available from two sources, each should be listed as a separate reference. Give all authors' names; do not use *et al.*

Samples of the correct formats for various types of references are given below.

Periodicals:

[1] J. F. Fuller, E. F. Fuchs, K. J. Roesler, "Influence of harmonics on power distribution system protection," IEEE Trans. Power Deliv., 3(2): 549-557, 1988.

Books:

[2] E. Clarke, Circuit Analysis of AC Power Systems, vol. I. New York: Wiley: 81, 1950.

Technical Reports:

[3] E. E. Reber, R. L. Mitchell, C. J. Carter, "Oxygen absorption in the Earth's atmosphere," Aerospace Corp., Los Angeles, CA, Tech. Rep. TR-0200 (4230-46)-3, Nov. 1968.

[4] S. L. Talleen. (1996, Apr.). The Intranet Architecture: Managing information in the new paradigm. Amdahl Corp., Sunnyvale, CA.

Papers Presented at Conferences (Unpublished):

[5] D. Ebehard, E. Voges, "Digital single sideband detection for interferometric sensors," presented at the 2nd Int. Conf. Optical Fiber Sensors, Stuttgart, Germany, 1984.

[6] Process Corp., Framingham, MA. Intranets: Internet technologies deployed behind the firewall for corporate productivity. Presented at INET96 Annu. Meeting.

Papers from Conference Proceedings (Published):

[7] J. L. Alqueres, J. C. Praca, "The Brazilian power system and the challenge of the Amazon transmission," in Proc. IEEE Power Engineering Society Transmission and Distribution: 315-320, 1991.

Dissertations:

[8] S. Hwang, "Frequency domain system identification of helicopter rotor dynamics incorporating models with time periodic coefficients," Ph.D. dissertation, Dept. Aerosp. Eng., Univ. Maryland, College Park, 1997.

Standards:

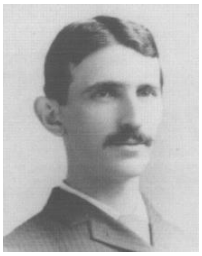
[9] IEEE Guide for Application of Power Apparatus Bushings, IEEE Standard C57.19.100-1995, Aug. 1995.

Patents:

[10] G. Brandli and M. Dick, "Alternating current fed power supply," U.S. Patent 4 084 217, Nov. 4, 1978.

Biographies

A technical biography for each author may be included, but without any title, as it is seen herein. It should begin with the author's name (as it appears in the byline). A photograph and an electronic file of the photo should also be included for each author. The photo should be black and white, glossy, and 3.0 centimeters (1.18 inches) wide by 3.8 centimeters (1.5 inches) high. The head and shoulders should be centered, and the photo should be flush with the left margin. The following is an example of the text of a technical biography:



Nikola Tesla (M'1888, F'17) was born in Smiljan in the Austro-Hungarian Empire, on July 9, 1856. He graduated from the Austrian Polytechnic School, Graz, and studied at the University of Prague. His employment experience included the American Telephone Company, Budapest, the Edison Machine Works, Westinghouse Electric Company, and Nikola Tesla Laboratories. His special fields of interest included high frequency. Tesla received honorary degrees from institutions of higher learning including Columbia University, Yale University, University of Belgrade, and the University of Zagreb. He received the Elliott Cresson Medal of the Franklin Institute and the Edison Medal of the IEEE. In 1956, the term "tesla" (T) was adopted as the unit of magnetic flux density in the MKSA system. In 1975, the Power Engineering Society established the Nikola Tesla Award in his honor. Tesla died on January 7, 1943.

- Email:
- ORCID:
- Web of Science Researcher ID:
- Scopus Author ID
- Homepage:

How to cite this paper:
F. Author, S. Author, T. Author," Instructions and Formatting Rules for Authors of Journal of Electrical and Computer Engineering Innovations, JECEI," J. Electr. Comput. Eng. Innovations, ×(×): ×××-×××, ××××.

DOI:

URL: http://jecei.srttu.edu/journal/authors.note