Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

# Feature selection using guided population based genetic algorithm with modified crossover and parent selection

Anurup Naskar [a], [1], [*], Soumyajit Ghosh [b], [1], Mahantapas Kundu [a], Ram Sarkar [a]

[a] Department of Computer Science and Engineering, Jadavpur University, Kolkata, India
[b] A. K. Choudhury School of Information Technology, Calcutta University, Kolkata, India

ARTICLE INFO

*Keywords:*
Genetic algorithm
Modified crossover
Feature selection
Adaptive threshold based roulette
UCI data
Microarray data

ABSTRACT

In the contemporary landscape, the imperative for cost-effective solutions is paramount, especially when dealing with extensively large dimensional datasets like gene expression datasets. The use of machine learning and data mining techniques in processing these voluminous and complex datasets presents a significant challenge in terms of time and resource consumption. A notable obstacle in dataset analysis is the prevalence of extraneous features or attributes. This is particularly evident in numerous medical datasets, which are often burdened with unnecessary attributes, complicating the task of classifications or prediction algorithms in obtaining precise results. However, the application of metaheuristic optimization algorithms shows remarkable proficiency in isolating pertinent feature vectors, thus markedly improving the efficiency and cost-effectiveness of data processing endeavors. We propose a novel feature selection method using a Genetic Algorithm (GA) that enhances initial population diversity by clustering features during initialization. The paper also introduces a modified crossover technique for generating offspring and employs an adaptive threshold-based Roulette Wheel for parent selection, ensuring effective feature selection. We evaluate the proposed feature selection method on 17 UCI datasets with 3 of them having a very high number of features and the obtained results are found to be better than many state-of-the-art methods both in terms of the classification accuracy and the reduction in the number of features. We also apply our method on 5 microarray-based gene expression datasets, used for the prediction of cancer, in order to ensure scalability and robustness of our method as a feature selector in real-life scenarios. This link provides the source code of the proposed method.

## 1. Introduction

The integration of Artificial Intelligence (AI) into various domains marks a significant milestone in technological evolution. This era is defined by rapid advancements and an increase in data production, establishing AI as a major contributor in fields such as image analysis, language processing, medical imaging, and time series data analysis. AI's role extends beyond healthcare, impacting numerous sectors by offering innovative solutions for complex challenges. One prominent application of AI is in analyzing gene expression data.

The phase of data preprocessing in various applications is of paramount importance. This stage involves the reduction of data dimensions and the selection of pertinent features, both essential steps in transforming complex, raw data into useful, insightful information. Feature selection is an essential process in machine learning, aimed at identifying the most informative and relevant features for model training.

Metaheuristic optimization algorithms are essential in enhancing the efficiency of feature selection techniques in machine learning and are broadly classified into single-solution-based and population-based categories. Single-solution-based optimization algorithms, such as Simulated Annealing (SA) [1], focus on iteratively improving a single solution. These algorithms are particularly effective in scenarios where a high-quality solution can be evolved or refined over time through a sequence of iterative processes.

Population-based optimization algorithms, in contrast, involve a group of potential solutions evolving over time. Almufti et al. [2] have stated that this category includes subcategories like evolutionary algorithms and swarm intelligence algorithms. Evolutionary algorithms, such as Genetic Algorithms (GA) and Differential Evolution (DE) [3], are inspired by biological evolution, mimicking processes like mutation and selection. They are well-suited for problems, where a diverse set of potential solutions can be beneficial in exploring the search

[*] Corresponding author.
*E-mail addresses:* rup.anu2020@gmail.com (A. Naskar), soumyajitkgp2003@gmail.com (S. Ghosh), mahantapas@gmail.com (M. Kundu), raamsarkar@gmail.com (R. Sarkar).

[1] Both authors contributed equally in this paper.

https://doi.org/10.1016/j.asoc.2025.112872
Received 16 February 2024; Received in revised form 2 February 2025; Accepted 9 February 2025
Available online 21 February 2025

space more thoroughly. Swarm intelligence algorithms like Particle Swarm Optimization (PSO) [4], Ant Colony Optimization (ACO) [5], and Bee Colony Optimization [6], draw from the collective behavior of social organisms. These metaheuristic optimization techniques are frequently used in wrapper-based methods. On the other hand, filter-based methods might incorporate statistical techniques like ReliefF [7] and Chi-square [8], among others. These methodologies are crucial in streamlining the feature selection process, thereby reducing the computational burden and enhancing the overall efficiency of the system.

The efficacy of wrapper-based methods is evident in various sophisticated applications, such as augmenting image contrast [9,10] and identifying facial conditions [11]. In healthcare, the precision and trustworthiness of AI-based diagnostic tools are of the utmost importance. This necessitates a focus on ensuring the robustness, scalability, and dependability of AI systems in medical applications. The success of AI in healthcare depends heavily on targeted feature extraction, which involves removing superfluous data to emphasize the most informative elements of the dataset. The task of optimal feature selection, often categorized as NP-hard due to the evaluation of extensive potential feature combinations, has led to a growing reliance on a variety of non-deterministic optimization techniques in both research and practical applications [12,13].

Deep learning methods are widely used for selecting important features, especially in healthcare. For instance, Sharma et al. [14] have developed DeepFeature, a method based on convolutional neural network (CNN) for feature selection in non-image data. This approach transforms data for CNN modeling and identifies key features using class activation maps, showcasing its versatility in different predictive tasks. Similarly, Wang et al. [15] have introduced MOGONET, leveraging graph convolutional networks for multi-omics data integration. This technique excels in classifying diverse datasets and identifying critical features across different data types.

Our research aims to find an optimal balance between exploring and exploiting the search space in feature selection algorithms. The strategy is intended to expand the exploration and exploitation scopes, accessing broader search areas and yielding more comprehensive results. Such an approach is particularly beneficial in complex, data-rich fields like healthcare, where nuanced and detailed analysis is paramount. By broadening the search parameters, our method aims to uncover more subtle, yet crucial, data patterns that might otherwise be overlooked, thereby enhancing the capabilities and applications of AI systems in healthcare and related fields.

Traditional GA faces several challenges that can hinder their effectiveness. A significant issue is premature convergence, which arises when the population lacks diversity, limiting the exploration of the solution space. Insufficient diversification results in the generation of similar candidate solutions, making it challenging to identify optimal ones. During initialization, the repeated selection of similar sub-optimal solutions further restricts exploration. Additionally, the search for optimal solutions is computationally demanding and time-consuming, especially with datasets that have a large number of attributes. Overcoming these challenges is essential for enhancing the efficiency and accuracy of GAs in optimization tasks, including feature selection.

Our approach addresses these issues by enhancing the GA's capability to maintain genetic diversity through the Adaptive Threshold Roulette method, which employs the Shannon algorithm to improve solution robustness. To further increase diversity, we integrate K-means clustering into the population initialization process, ensuring a varied initial population and avoiding redundancy by not selecting the same types of features repeatedly. Our modified crossover technique extends traditional binary crossovers by applying arithmetic operations to binary-encoded chromosomes, allowing for more precise exploration and helping prevent premature convergence through the use of random coefficients to maintain diversity and encourage innovation in offspring. Furthermore, the pressure threshold adjusts fitness to prevent any single individual from dominating the population, promoting broader exploration. These enhancements significantly improve the efficiency and effectiveness of GA in feature selection, effectively addressing the limitations of traditional GA.

Precisely, our **contributions** are:

- The method benefits from cluster-based population initialization to improve GA's ability to explore the search space effectively.
- The proposed approach employs an Adaptive Threshold-based Roulette Wheel algorithm for performing parent selection.
- The proposed method uses a modified crossover technique for generating potentially superior offspring.

The remainder of the study is organized in the manner listed below. Section 2 discusses the related work. The application of GA in feature selection is discussed in this section. Section 3 divides the background of GA into several key components, including population initialization, selection, crossover, mutation, and fitness function. Section 4 offers comprehensive information regarding the proposed approach. Section 5 examines the outcomes attained on the UCI and gene expression datasets. The analysis of the convergence curve as well as box-plot diagrams of our proposed method is discussed in Section 6. Finally, we draw some conclusions on our work and discuss some potential future directions in Section 7.

## 2. Related work

The ability to collect vast amounts of data in today's world presents both advantages and disadvantages. On one hand, it allows for a more comprehensive analysis of various characteristics. On the other hand, managing and processing such immense volumes of data becomes challenging. To address this, dimensionality reduction techniques have become increasingly crucial as they help eliminate unnecessary features without compromising the learner's performance.

GA stands out as one of the most extensively used metaheuristic algorithms. Researchers across diverse domains, including class imbalance, medical diagnostics, image steganography, and feature selection, have extensively explored and utilized GA. Prabha et al. [16] have developed a notable technique to boost image contrast, incorporating a fuzzy intensification operator and GA. This approach primarily enhances the clarity of images by altering the intensity of information within them. Prasetio et al. [17] has employed KNN as a classifier, while GA has been used for feature selection and optimization. However the disadvantage of this work is the sensitivity of the k-nearest neighbor algorithm to the selection of the k parameter and the presence of noise or irrelevant features in medical datasets, which can degrade the algorithm's accuracy. Marjit et al. [18]have introduced a method that combines Simulated Annealing with GA to facilitate the selection of genes from microarray data. Guha et al. [19] have integrated the Great Deluge Algorithm (GDA) with GA in their approach. This amalgamation notably boosts exploitation by applying perturbations to the candidate solutions. The main shortcoming of this, as stated by the authors, is its increased computational cost compared to the basic Genetic Algorithm (GA), despite its improved accuracy and efficiency in feature selection. Kumar et al. [20] have combined GA with Support Vector Machines (SVM) to create a new algorithm. This method outperformed other advanced metaheuristic algorithms in selecting feature subsets, especially in the context of UCI datasets. Verpa et al. [21] have explored how integrating GA into the development of attribute weighting could improve the classification accuracy of attribute-weighted techniques. Their research demonstrated that using GA notably increased both the general classification accuracy and the median true positive rate for the attribute-weighted k-nearest neighbor method. The authors have identified the increased processing time as a notable shortcoming of using GA optimization in their methodology.

Vaishali et al. [22] proposed a method that applies Goldberg's GA for feature selection, reducing the Pima Indians Diabetes Dataset from eight to four features. They then utilized a multi-objective evolutionary fuzzy classifier, achieving 83.04% accuracy using NSGA II with a 70/30 train-test split. However, GA mistakenly excluded 'Plasma Glucose Concentration,' a critical feature, due to dataset outliers and missing values, highlighting areas for future improvement. Aalaei et al. [23] have employed three different classifiers: Artificial Neural Network (ANN), a PS-classifier, and a GA-classifier, to analyze datasets related to breast cancer. Too et al. [24] introduced an enhanced rival GA and a faster variant, incorporating competitive-inspired selection and crossover methods to improve global search and feature selection. Despite these advancements, they highlighted limitations, such as complex hyperparameter tuning, high computational cost, and selection randomness. Kabir et al. [25]have developed a hybrid GA with Bounded Subset Size (HGAFS) that addresses subset size challenges, preventing premature convergence and enhancing local fine-tuning. GAs, while widely applied in feature selection and areas like handwriting and human activity recognition, require carefully tuned user-specified parameters, a noted limitation of the approach. Oliviera et al. [26]have applied multi-objective GA for feature selection in handwriting recognition, combining sensitivity analysis with neural network evaluations. While effective on large datasets, their approach relies solely on the Pareto-optimal front, which may hinder generalization across databases without an additional validation set. Concurrently, Malakar et al. [27] have explored a similar domain but focused on refining a GA-based feature selection model. Their model was specifically designed to augment the identification and extraction of both local and global features from images of handwritten words, thereby enhancing the overall accuracy and efficiency of handwriting analysis. This work, while effective in reducing feature dimensions and improving recognition accuracy, is computationally expensive due to the multi-level optimization process involved in the hierarchical feature selection model.

Hossain et al. [9] have carried out a study in steganography, introducing a technique that uses GA for creating a k-digit password. This method is also computationally expensive due to the multi-level optimization process involved in the hierarchical feature selection model. Similarly, Sarkar et al. [28] put forward an innovative method in the realm of microstructural image classification, which relies on a GA-based feature selection approach. The main shortcomings stated by the authors for this method are that the extracted feature vectors may contain redundant or irrelevant features, and the application of feature selection methods in microstructural image classification is still a less explored area. Kumar et al. [29] have developed a new algorithm designed for the automatic, real-time identification of facial expressions. This method incorporates radial basis function (RBF) support in conjunction with Haar Wavelet Transform for efficient feature extraction. The classification is conducted using RBF-SVM. To improve edge detection in facial images, the algorithm integrates GA with fuzzy-C-means, thereby boosting its overall effectiveness. Xiao et al. [30] have proposed a cutting-edge method for optimizing CNN hyperparameters using a GA with adaptable length. This approach adeptly manages the complexities associated with models having diverse hyperparameters. Their findings indicate significant enhancements in performance, suggesting that optimal outcomes could be achieved with sufficient computational resources and time. Kwon et al. [31] have introduced a method that optimizes molecular structures through a fusion of evolutionary algorithms and deep learning. This technique demonstrates the capability of AI to effectively alter certain properties of molecules. The evidence of this is seen in the modification of light-absorption wavelengths among compounds in the PubChem database. Bouktif et al. [32] have developed a methodology to enhance short to medium-term electricity load forecasting, combining machine learning with long short-term memory (LSTM) neural networks. Applied to electricity consumption data from France, this method has shown a marked improvement in forecast accuracy. It accomplishes this through the optimization of feature selection, time lags, and LSTM layers, resulting in lower error rates compared to traditional machine learning models. Kalsi et al. [33] have proposed an encryption technique named DNA Deep Learning Cryptography, which merges DNA sequencing patterns with neural networks. In this method, a GA is used to transform each character of a message into a distinct DNA base sequence. The core feature of their research is the amalgamation of biological operations and deep learning into the realm of cryptography, representing a significant advancement in digital security. While DNA computing has great potential for storage capacity and computational efficiency, there are still significant challenges in realizing this potential in practical applications. Tian et al. [34] have developed a method that utilizes GA to create a framework for the selection of deep learning models. This framework is centered on selecting optimal features from existing models to address different detection challenges. Their approach capitalizes on progress in image classification, supported by comprehensive visual datasets and CNN, thereby improving the implementation of transfer learning in various fields. However, the framework's vast search space and computational demands may limit real-world efficiency and effectiveness. Bendali et al. [35] have introduced a novel method to enhance the prediction of photovoltaic energy production, employing new hybrid techniques that optimize deep learning models using GA. Their study particularly assesses the performance of LSTM, gate recurrent unit (GRU), and recurrent neural network (RNN) neural networks. The method involves using GAs to identify the most effective window sizes and the optimal number of neurons in all hidden layers, thereby improving the precision of solar irradiation output predictions. It is mentioned that the classification success rates of the studies were low, except for a few studies. Kilicarslan et al. [36] have developed innovative hybrid models, namely GA-Stacked Autoencoder (GA-SAE) and GA-CNN, which combine GAs with Stacked Autoencoders(SAE) and CNN for accurately predicting various types of nutritional anemia. These models, enhanced through GA for optimal hyperparameter tuning, have achieved a noteworthy accuracy of 98.50%, outperforming previous methods in anemia diagnosis. However, it was observed that the classification success rates of the studies were low, except for a few studies.

Bi et al. [37] have created a dual-phase deep learning approach, supported by GA, for the prediction of crop yields. This method emphasizes efficient weight initialization in neural networks and tackles challenges such as local optima and vanishing gradients. Their results demonstrate that this approach exceeds conventional gradient-based methods in achieving quicker convergence and improved accuracy in predictions. Mattioli et al. [38] have showcased the effectiveness of GA in determining the best topologies for DNNs, resulting in superior network configurations with reduced computational requirements compared to conventional techniques. Their study highlights the proficiency of evolutionary computation algorithms in streamlining intricate decision-making processes in deep learning scenarios. Slim et al. [39] have introduced an advanced method for refining deep learning architectures in human activity recognition (HAR) tasks, employing GA for the optimal tuning of parameters. This method also incorporates new statistical features alongside those extracted using standard CNN techniques. When evaluated on IoT systems and benchmark datasets like WISDM and UCI, their approach has significantly enhanced accuracy, achieving as high as 93.8% in user-dependent scenarios and 86.1% in user-independent contexts. Balaha et al. [40] have devised a novel method to advance Arabic handwritten character recognition (AHCR), addressing both segmentation and recognition challenges. Their approach utilizes CNNs for enhanced recognition capabilities, surpassing traditional machine learning techniques in automatic image feature extraction. The research includes the development of 14 unique CNN architectures, tested using the HMDB database. Additionally, it proposes a synergistic approach combining transfer learning and GA, termed as HMB-AHCR-DLGA, for the fine-tuning of training parameters, resulting in significant improvements in testing accuracy. Yoosefzadeh et al. [41] have developed an innovative approach using machine learning techniques, such as Multilayer Perceptron (MLP), (RBF, and Random Forest (RF), to forecast soybean yields based on genetic data. They identified RBF as the most accurate method. Further improvements were made by combining RBF with an ensemble bagging strategy and GA, which significantly enhanced the predictions of soybean yields and aided in the development of soybeans with greater productivity. Fernandez et al. [42] have developed a machine-learning methodology to assess variations in accuracy for electronic calculations of graphene nanoflakes across diverse theoretical models. Their innovative approach, effectively predicting Fermi level energy and band gap, has improved the screening efficiency of nanomaterial libraries. This method has the potential to be adapted to a wide range of materials and computational techniques. Gao et al. [43] have created a novel prediction method for bike-sharing demand, integrating a moment-based model with a synergy of fuzzy C-means (FCM) GA and a back propagation network (BPN). This approach proficiently classifies historical rental data using the FCM-based GA, subsequently utilizing this classification to train the BPN for precise future demand forecasting. The effectiveness of this process has been demonstrated through a practical case study employing real-world data. Wu et al. [44] have introduced an efficient and novel method for meta-atom design, combining DNNs with GA to optimize the design process. This method employs a high-accuracy DNN to effectively direct the GA, thereby minimizing the number of iterations needed. Demonstrated through the creation of sophisticated metasurfaces, such as an orbital angular momentum generator and a metalens, this approach holds significant potential for the rapid development of functional meta-devices.

Traditional GAs encounter several challenges that limit their effectiveness. A major issue is premature convergence, which occurs when the population lacks sufficient diversity, restricting the exploration of the solution space. Without proper diversification, the population tends to produce similar candidate solutions, making it difficult to identify the optimal solutions. During initialization, the repeated selection of similar sub-optimal solutions further narrows the exploration space. Additionally, searching for the optimal solution is computationally intensive and time-consuming, particularly with datasets having large attributes. Addressing these challenges is crucial for improving the efficiency of GAs in optimization tasks including feature selection.

In our approach, we enhance the GA's ability to maintain genetic diversity using the Adaptive Threshold-based Roulette Wheel method, which leverages the Shannon algorithm to improve the robustness of the solutions. To further increase diversity, we incorporate K-means clustering into the population initialization process. This ensures that we do not repeatedly select the same types of features, thereby avoiding redundancy and creating a more varied initial population. Our modified crossover technique advances traditional binary crossovers by applying arithmetic operations to binary-encoded chromosomes. This method allows for more precise exploration and helps prevent premature convergence by using random coefficients to maintain genetic diversity and foster innovation in offspring. Additionally, the pressure threshold adjusts fitness to prevent any single individual from dominating, encouraging broader exploration. These enhancements significantly improve the efficiency of GA in feature selection, effectively addressing the limitations of traditional GA.

## 3. Background

### 3.1. Genetic algorithm

GA is an optimization technique inspired by the principles of natural selection and genetics, as introduced by Darwinian evolution [45]. These algorithms are designed to iteratively improve a population of candidate solutions to approximate the global optimum of a given problem. The main components of GA include population initialization, selection, crossover, mutation, and the fitness function.

#### 3.1.1. Population initialization

The first step in a GA is to create an initial population, which consists of diverse possible solutions or chromosomes. Each chromosome represents a potential solution, typically encoded as a series of genes. The population can be generated randomly, covering a broad search space or with heuristic methods, using domain knowledge to provide a focused starting point. Effective population initialization is crucial as it enables broad exploration of the search space from the outset.

#### 3.1.2. Selection

Selection is the process of choosing parent chromosomes from the current population to produce offspring for the next generation [45]. The objective of selection is to favor the fitter chromosomes, thereby promoting the survival and propagation of superior genetic material. Several selection methods are commonly employed like Roulette Wheel Selection and Tournament Selection, but we have used Roulette Wheel based selection method.

#### 3.1.3. Crossover

Crossover, or recombination, is a genetic operator that combines the genetic material of two parent chromosomes to produce new offspring [45]. This process is inspired by biological reproduction, where offspring inherit traits from both parents. The crossover operator is essential for exploring the search space and generating new solutions that might possess better fitness. Traditional crossover techniques include Single-Point Crossover, Two-point Crossover, Uniform Crossover. In this paper, we have used the Single-Point Crossover method.

#### 3.1.4. Mutation

The mutation process in GA introduces random changes to solutions, maintaining diversity and helping the algorithm avoid focusing too narrowly on specific solutions [45]. This genetic operator prevents premature convergence and enables the GA to explore the solution space more effectively.

#### 3.1.5. Fitness function

In GA, the fitness function is a critical tool for evaluating the effectiveness of a candidate solution, essentially gauging its appropriateness for the specific problem being addressed. Within the framework of GA as a wrapper-based approach, this function depends on a learning algorithm to assess each potential solution. Such an approach enables the fitness function to accurately gauge the performance of these solutions within a given learning scenario. This assessment is crucial in aiding GA's decision-making process, particularly in refining solutions that align best with the established goals or criteria.

For the study at hand, the K-Nearest Neighbors (KNN) classifier has been employed to determine the classification accuracy of the solution vectors. In our application, the fitness function is constructed around two primary components: the accuracy of classification and the reduction of feature utilization. Both of these aspects are structured to favor higher values. This implies that any increase in classification accuracy or a decrease in the number of features used will have a positive impact on the fitness score. Higher accuracy, indicative of lower error rates, results in a higher fitness value. Similarly, an effective reduction in feature usage, which implies a more streamlined and efficient solution, also boosts the fitness score.

The fitness function, as shown in Eq. (1), evaluates a particular feature subset. In this equation, $|F|$ represents the number of features in the excluded subset of the feature vector, $|D|$ signifies the length of the feature vector in the dataset, $acc$ represents the accuracy achieved by employing the KNN classifier with the selected feature subset, and $\alpha$, a value between $[0,1]$, denotes the relative importance given to the number of features discarded in comparison to the classification accuracy. This balance ensures a comprehensive evaluation of the feature subset's effectiveness.

$$F = \alpha \times acc + (1-\alpha) \times \frac{|F|}{|D|} \tag{1}$$

For the implementation of the methodology, we have considered $\alpha$ as 0.8.

### 3.2. Feature selection

Sometimes, a classifier handles a huge dimension of the characteristics or attributes present in the datasets. However, it has been found that only a limited fraction of these features contribute to the classification or prediction purposes. Additionally, the classification accuracy is also decreased by the existence of these redundant features. To encounter this problem, a feature selection method has been proposed here. The workflow of the feature selection strategy used in this work is shown in Fig. 1.

## 4. Materials and methods

The proposed method introduces a feature selection approach using a GA, integrating the Clustering-based Population Initialization, a Modified Crossover mechanism, and an Adaptive Threshold-based Roulette method for parent selection. The method leverages the inherent structure of the data through K-means Clustering to guide Population Initialization, enhances genetic diversity through the Modified Crossover, and dynamically adjusts parent selection probabilities using the Adaptive Threshold based Roulette method.

### 4.1. Population initialization using K-means clustering algorithm

In the proposed study, a novel approach for population initialization in GA is employed (Algorithm 1), specifically tailored for feature selection. This method commences by loading and preprocessing the dataset, where it is assumed that the final column represents the class labels and the preceding columns are the features. A crucial step involves computing the cosine similarity matrix among all features, which is subsequently normalized to ensure the non-negativity of values. The optimal number of clusters is determined using the elbow method [46], which identifies the point of maximum curvature in the within-cluster sum of squares (WCSS) plot. Once the optimal number of clusters is selected, the K-means clustering algorithm is applied to the similarity matrix. Each feature, represented as a binary value (0/1), indicates the presence or absence of features in a subset or a chromosome. The initialization process involves randomly selecting features from the clusters and setting the corresponding feature's index in the chromosome to '1'. This process not only injects an element of randomness but also ensures that the initial population is influenced by the inherent clustering structure of the data. This approach leverages the clustering information to guide the feature selection, aiming to reduce dimensionality while maintaining the representativeness of the selected features for each cluster. The resulting population serves as the starting point for GA, promising a more informed search space exploration in high-dimensional datasets.

Assume we have a small dataset with three features across two instances as follows:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

*1.Compute cosine similarity:*

Calculate the cosine similarity for each pair of features. This measures the cosine of the angle between two non-zero vectors of an inner product space, which is used to measure similarity between two vectors. Similarity between Feature 1 and Feature 2:

$$\text{CosineSimilarity}(A,B) = \frac{1\times 4 + 2\times 5 + 3\times 6}{\sqrt{1^2+2^2+3^2}\times\sqrt{4^2+5^2+6^2}} = \frac{32}{\sqrt{14}\times\sqrt{77}}$$

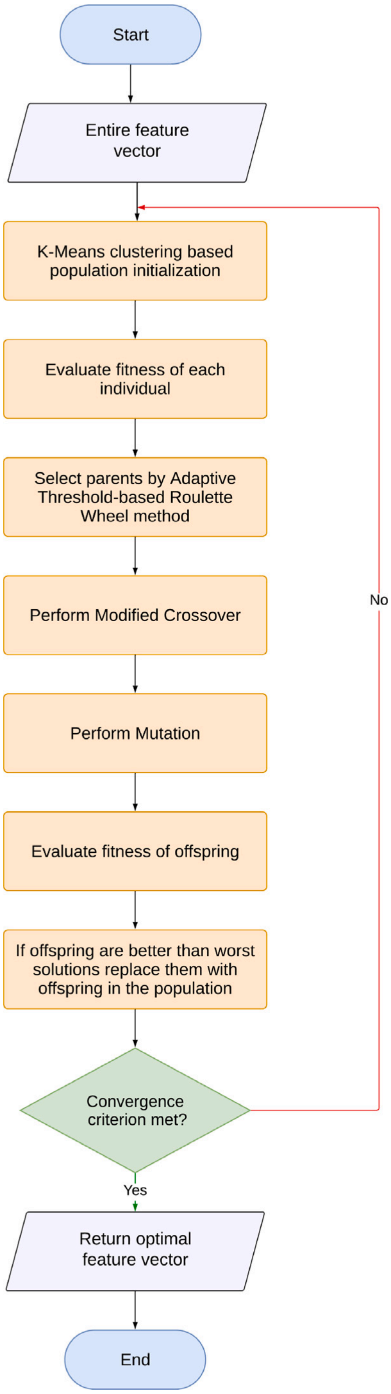
Start
Entire feature vector
K-Means clustering based population initialization
Evaluate fitness of each individual
Select parents by Adaptive Threshold-based Roulette Wheel method
Perform Modified Crossover
Perform Mutation
Evaluate fitness of offspring
If offspring are better than worst solutions replace them with offspring in the population
Convergence criterion met?
No
Yes
Return optimal feature vector
End

**Fig. 1.** The flowchart of the proposed GA based feature selection method.

Calculating this gives us a similarity score for Feature 1 and Feature 2. Repeat this calculation for each pair of features to create a full cosine similarity matrix.

*2. Normalize cosine similarity matrix:*

Normalize the cosine similarity matrix to have values between 0 and 1. This involves scaling the similarity scores so they fit within a specific range, facilitating easier clustering. The normalization can be done using the min–max scaling approach:

$$\text{Normalized Value} = \frac{\text{Value} - \text{Min}}{\text{Max} - \text{Min}}$$

*3. Determine optimal clusters:*

Use the elbow method to determine the optimal number of clusters. This involves plotting the sum of squared distances from each point to its assigned cluster center and identifying the "elbow" point, where adding more clusters does not significantly improve the fit. Suppose, after applying the elbow method the optimal number of clusters is determined to be 2.

*4. Apply K-means clustering:*

Using the normalized cosine similarity matrix, apply K-means clustering with the determined number of clusters (e.g., 2 clusters). Features are grouped into clusters based on their similarity, ensuring similar features are assigned to the same cluster.

*5. Initialize population:*

For each chromosome in the initial population of the GA, select features according to their cluster assignments. This ensures diversity in the initial population by leveraging the underlying structure of the feature space.

**Algorithm 1** Population Initialization based on K-Means Algorithm

```
Input: $data$, $num\_chromosomes$, $num\_features$ ▷ Data for clustering, number of chromosomes, number of features
Output: $binary\_population$ ▷ Initialized population based on clustering
$cosine\_sim\_matrix \leftarrow$ CosineSimilarity($data^T$)
$X\_normalized \leftarrow$ Normalize($cosine\_sim\_matrix$)
$optimal\_clusters \leftarrow$ ElbowMethod($X\_normalized$)
$y\_kmeans \leftarrow$ KMeans(optimal_clusters, $X\_normalized$)
$binary\_population \leftarrow []$
for $i \leftarrow 1$ to $num\_chromosomes$ do
    $chromosome \leftarrow$ CreateChromosome($num\_features, y\_kmeans$)
    append $chromosome$ to $binary\_population$
end for
Return $binary\_population$
```

### 4.2. Modified crossover

The modified crossover (Algorithm 2) embodies an innovative crossover technique tailored for binary-encoded chromosomes within the realm of GA. This function ingeniously melds the principles of arithmetic crossover with binary representation, necessitating two parent chromosomes of identical length as input. The crux of this method lies in its unique transformation of binary data into decimal format, thereby facilitating arithmetic operations that would otherwise be infeasible with binary strings.

Upon this discussion, the algorithm leverages random coefficients – $\delta_1$, $\delta_2$, $\epsilon_1$, and $\epsilon_2$ – to orchestrate a sophisticated arithmetic crossover. These coefficients are derived from uniform random distributions, where $\delta_1$ and $\delta_2$ originate from a standard uniform distribution, i.e., the value $\delta$ lies between 0 to 1, and $\epsilon_1$ and $\epsilon_2$ are sourced from a uniform distribution spanning the range $[-1, 1]$. This stochastic element introduces diversity and innovation in the offspring generation process. The crossover itself is performed by combining the decimal values of the parent chromosomes in a weighted manner, incorporating the random coefficients. The use of $\epsilon$ coefficients introduces a differential element to the crossover, allowing for an exploration of the solution space that extends beyond the immediate vicinity of the parent chromosomes. This can lead to discovering novel and potentially superior genetic combinations. Post crossover, the function conscientiously ensures that the resultant offspring values remain within the valid binary range. It achieves this by clipping the offspring values to fit the range permissible by the length of the parent chromosomes, thereby preventing any overflow errors.

Finally, the offspring chromosomes are converted back to their binary format, maintaining the same length as their progenitors. This binary-to-decimal-to-binary conversion process, coupled with arithmetic crossover operations, defines the modified crossover function as a sophisticated and robust method for generating diverse and potentially superior offspring in binary-encoded GA. The crossover takes place using Eq. (2) and Eq. (3)

$$np_1 = \delta_1 \times p_1 + (1 - \delta_1) \times p_2 + \epsilon_1 \times (p_1 - p_2) \tag{2}$$

$$np_2 = \delta_2 \times p_2 + (1 - \delta_2) \times p_1 + \epsilon_2 \times (p_1 - p_2) \tag{3}$$

In this equation:

- $np_1$ represents the $newparent\_1$ after modified crossover.
- $np_2$ represents the $newparent\_2$ after modified crossover.
- $p_1$ is the $parent\_1$.
- $p_2$ is the $parent\_2$.

For instance, two parent chromosomes, $p_1 = [1, 0, 1, 1, 1, 0]$ and $p_2 = [0, 1, 1, 1, 0, 1]$, are first converted to their decimal equivalents (46 and 29). Random coefficients $\delta_1 = 0.6$, $\delta_2 = 0.4$, $\epsilon_1 = 0.3$, and $\epsilon_2 = -0.2$ are introduced to ensure diversity. Now, the offspring in terms of decimal values are calculated as follows:

$$np_1 = 0.6 \cdot 46 + (1 - 0.6) \cdot 29 + 0.3 \cdot (46 - 29) = 27.6 + 11.6 + 5.1 = 44.3$$

$$np_2 = 0.4 \cdot 29 + (1 - 0.4) \cdot 46 - 0.2 \cdot (46 - 29) = 11.6 + 27.6 - 3.4 = 35.8$$

These values are clipped into the valid decimal number, if needed, and converted back to binary: $np_1 = [1, 0, 1, 1, 0, 0]$ and $np_2 = [1, 0, 0, 1, 0, 0]$. This approach supports a slower and more deliberate exploration of the search space, ensuring diversity while avoiding the rapid and disruptive changes typical of mutation.

### 4.3. Adaptive threshold-based roulette wheel

In our research, we have introduced the *Adaptive Threshold-based Roulette Wheel* algorithm, a novel method designed to enhance selection processes in GA. This algorithm operates on a given list of individual fitness scores, with the primary objective of selecting an individual index based on adaptive probabilities. The core innovation lies in its ability to dynamically adjust selection pressure in response to the diversity of the population. Initially, the algorithm evaluates the diversity of fitness scores using Shanon's entropy-based metrics [47], effectively normalizing this diversity against the theoretical maximum. It then scales this normalized diversity, integrating it into the calculation of a pressure threshold. This threshold modulates the fitness scores, striking a balance between exploration and exploitation, a crucial aspect in evolutionary computations.

The calculation of the pressure threshold within the algorithm is a crucial step, pivotal to its performance. This threshold is derived from Eq. (4). Here, $diversity\_scaled$ represents the scaled measure of diversity within the population, reflecting the degree of variation in the fitness scores. The scaling is a function of the normalized diversity, which is then adjusted by certain algorithm-specific parameters. The exponential function $e^{-diversity\_scaled}$ ensures a non-linear response, facilitating a flexible adjustment of the selection pressure. This dynamic adjustment is key to maintaining a balance between exploring new solutions and exploiting known good solutions, based on the current state of diversity in the population. It is this adaptive nature of the pressure threshold that enhances the algorithm's ability to efficiently navigate the solution space.

A critical step involves combining the adjusted fitness with a base probability derived from the population size, ensuring that every individual has a non-zero chance of selection, thereby preventing premature convergence. The algorithm then calculates selection probabilities for each individual, facilitating a roulette wheel selection mechanism. This process ensures that individuals with higher fitness have a proportionally greater chance of being selected, while still maintaining diversity in the population. Our experiments demonstrate its efficacy in maintaining genetic diversity over generations, leading to more robust solutions in evolutionary optimization tasks.

$$pressure\_threshold = \frac{1}{1 + e^{-diversity\_scaled}} \tag{4}$$

**Algorithm 2** Modified Crossover

```
Input: $p_1$, $p_2$                                   ▷ Parent chromosomes
Output: $np_1$, $np_2$                                ▷ Offspring chromosomes
function MODIFIED_CROSSOVER($p_1$, $p_2$)
    $\delta_1, \delta_2 \leftarrow \text{Random}(0,1), \text{Random}(0,1)$
    $\epsilon_1, \epsilon_2 \leftarrow \text{RandomUniform}(-1,1), \text{RandomUniform}(-1,1)$
    $p_1 \leftarrow \text{BinaryToDecimal}(p_1)$
    $p_2 \leftarrow \text{BinaryToDecimal}(p_2)$
    $np_1 \leftarrow \delta_1 \times p_1 + (1-\delta_1) \times p_2 + \epsilon_1 \times (p_1 - p_2)$
    $np_2 \leftarrow \delta_2 \times p_2 + (1-\delta_2) \times p_1 + \epsilon_2 \times (p_2 - p_1)$
    $np_1 \leftarrow \text{DecimalToBinary}(np_1)$
    $np_2 \leftarrow \text{DecimalToBinary}(np_2)$
    return $np_1$, $np_2$
end function
```

**Algorithm 3** Adaptive Threshold-based Roulette Wheel

```
Input:
    $fitness$                      ▷ List of fitness scores for each individual
Output:
    $index$                                       ▷ Index of selected individual
Begin:
if len(set($fitness$)) == 1 then
    $diversity\_normalized \leftarrow 0$
else
    $values, counts \leftarrow$ np.unique($fitness$, return_counts=True)
    $probabilities \leftarrow counts/\text{sum}(counts)$
    $diversity \leftarrow -\text{sum}(probabilities \times \log_2(probabilities + 1e-10))$
    $max\_diversity \leftarrow \log_2(\text{len}(fitness))$
    $diversity\_normalized \leftarrow diversity/max\_diversity$
end if
$diversity\_scaled \leftarrow \text{steepness} \times (diversity\_normalized - \text{midpoint})$
$pressure\_threshold \leftarrow 1/(1 + e^{-diversity\_scaled})$
$adjusted\_fitness \leftarrow [f \times pressure\_threshold$ for $f$ in $fitness]$
$base\_probability \leftarrow 1/\text{len}(fitness)$
$combined\_fitness \leftarrow [a\_f + \text{diversity\_boost} \times base\_probability$ for $a\_f$
in $adjusted\_fitness]$
if not any($combined\_fitness$) then
    raise ValueError("The combined fitness score is 0, selection is
not possible.")
end if
$total\_fitness \leftarrow \text{sum}(combined\_fitness)$
$selection\_probs \leftarrow [f/total\_fitness$ for $f$ in $combined\_fitness]$
$index \leftarrow$ np.random.choice(len($fitness$), $p = selection\_probs$)
return $index$
End
```

## 5. Result and analysis

### 5.1. Datasets

In our research, we utilized 17 datasets from the UCI Machine Learning repository, focusing primarily on health-related analysis, is mentioned in Table 1. These include BreastCancer, HeartEW, and PenglungEW, which target breast cancer detection, heart disease risk assessment, and lung disease identification. Other datasets like Wine, Zoo, and Lymphography, examine diverse attributes related to wine quality, species characteristics, and lymph node traits. This link provides the datasets that we have used.

### 5.2. Controlling parameters

The optimization algorithms in our study are primarily driven by hyperparameters, the settings of which are usually determined through a series of trials with varied values. In this research, we employ hyperparameter settings that have been previously validated and found effective, as mentioned in [12]. For comparative analysis of our feature selection algorithm, we consider a diverse range of algorithms, including GA, Whale Optimization Algorithm (WOA), Bat Algorithm (BA), PSO, Memetic Algorithm (MA), Grey Wolf Optimizer (GWO), Harmony Search (HS), Equilibrium Optimizer (EO), Red Deer Algorithm (RDA). The specific hyperparameter settings applied for these algorithms are comprehensively detailed in Table 2.

### 5.3. Experimental results

#### 5.3.1. Comparison of the proposed method with optimization algorithms based on classification accuracy and number of features of selected

For the comparative analysis, we evaluate several other algorithms from 5.2. The parameters utilized are detailed in Table 2. Table 3 represents the comparison of our proposed method with other optimization algorithms in terms of classification accuracy and the numbers of features selected.

For Breastcancer, the proposed method achieves an accuracy of 99.28%, which is significantly higher than the other methods, except for GWO, which comes close at 99.20%. For Exactly, this method, along with GA, GWO, and PSO, achieves perfect accuracy (100%), markedly outperforming other methods like BA and HS, which have notably lower accuracies. In M-of-n dataset, it is one of the four methods (including GA, MA, and GWO) that achieve 100% accuracy, demonstrating its capability to classify this dataset. For Wine and Zoo, this method shows exceptional performance with 100% accuracy, which is matched by GA, EO, PSO, GWO, and HS for the Zoo dataset, indicating that the problem might be less complex or that these methods are particularly well-suited to these datasets. However, it does not always top the charts. For BreastEW, while our algorithm has a high accuracy of 98.24%, it is closely followed by other methods, suggesting a competitive field with a small performance gap. For PenglungEW, the proposed method has a substantial lead, although it is not the best performer. PSO matches our method's accuracy, while WOA, HS, and GWO are also strong competitors. For SpectEW, this method does well with the third-highest accuracy at 90.74%, being outperformed by PSO and HS. In summary, our approach demonstrates a strong capability to achieve high classification accuracy across various datasets. It shows particularly strong results for the BreastCancer, Exactly, M-of-n, Wine, and Zoo datasets, where it either outperforms all other methods or is part of the subset of methods that achieve the highest accuracy. This consistent performance suggests that the proposed feature selection algorithm could be a robust choice for tackling a variety of classification problems.

It is crucial to note that fewer features can potentially streamline the model, making it not only faster but potentially more generalizable if the correct features are retained. For the BreastCancer dataset, the proposed method reduces features to just 3, showcasing robust selection, with only RDA going further by selecting a single feature. In CongressEW, our method's choice of 4 features matches the optimal selection by RDA, outperforming others like MA, which selects nearly three times as many. For Exactly-2, our approach, while selecting 8 features, still maintains strong performance, unlike RDA and WOA's single-feature reduction, which compromises classification accuracy by discarding important features. On BreastEW, the proposed method's selection of 10 features represents a balanced approach, even though RDA

**Table 1**
Description of UCI datasets used for experimentation.

| Dataset | Number of labels | Domain | Sample size | Number of features |
|---|---|---|---|---|
| Zoo | 2 | Artificial | 101 | 16 |
| CongressEW | 2 | Politics | 435 | 16 |
| Exactly2 | 2 | Biology | 1000 | 13 |
| BreastEW | 2 | Biology | 569 | 30 |
| Wine | 3 | Chemistry | 178 | 13 |
| PenglungEW | 2 | Biology | 73 | 325 |
| M-of-n | 2 | Biology | 1000 | 13 |
| Vote | 2 | Politics | 300 | 16 |
| Exactly | 2 | Biology | 1000 | 13 |
| HeartEW | 2 | Biology | 270 | 13 |
| SpectEW | 2 | Biology | 267 | 22 |
| Lymphography | 2 | Biology | 148 | 18 |
| Ionosphere | 2 | Electromagnetic | 351 | 34 |
| BreastCancer | 2 | Biology | 699 | 9 |
| CNAE-9 | 9 | Business | 1080 | 857 |
| Parkinson's Disease Classification | 2 | Medical | 756 | 754 |
| Internet Advertisements | 2 | Computer Science | 3279 | 1558 |

**Table 2**
Hyperparameters used in various optimization algorithms.

| Optimization algorithm | Parameter(s) | Value |
|---|---|---|
| Generic Parameters | Population | 30 |
| | Weight for Accuracy ($\alpha$) | $\alpha$ = 0.8 |
| | Iterations | 60 |
| BA | Loudness | 1 |
| | Pulse Emission Rate | 0.15 |
| | maxFrequency | 2 |
| | minFrequency | 0 |
| MA | Attraction Constant (a1,a2) | a1=1, a2=1.5 |
| | Gravitational Constant | 0.98 |
| | Initial Nuptial dance coefficient | 0.1 |
| | Visibility coefficient | 2 |
| | Nupital Dance & Random Walk updating factor ($\delta$) | 0.9 |
| | Initial Random walk coefficient | 0.1 |
| EO | Constants (a1,a2) | a1 = 1, a2 = 2 |
| | Generation | 0.9 |
| | Pool Size | 4 |
| WOA | (A,C) | A and C lie in $[0, 2]$ |
| | l | l lies in $[0, 1]$ |
| | p | p lies in $[0, 1]$ |
| | Shape of spiral | 1 |
| PSO | Coefficients (r1,r2) | r1 and r2 lie in $[0, 1]$ |
| GWO | Coefficients (r1,r2) | r1 and r2 lie in $[0, 1]$ |
| | (a,c) | a and c lie in $[0, 2]$ |
| RDA | Upper Bound | 5 |
| | Beta ($\beta$) | 0.1 |
| | Gamma ($\gamma$) | 0.5 |
| | Alpha ($\alpha$) | 0.2 |
| | Lower Bound | −5 |
| GA | Mutation Probability | 0.05 |
| | Crossover Probability | 0.8 |
| | Gene Selection | Roulette Wheel |
| HS | Harmony Memory Consideration Rate (HMCR) | 0.9 |
| Proposed GA | Mutation Probability | 0.05 |
| | Crossover Probability | 0.8 |
| | Gene Selection | Adaptive Threshold-based Roulette Wheel |

achieves the minimum with just one feature. For Lymphography, our method's 8-feature selection is competitive, especially against RDA's minimal 3-feature choice. In Vote, selecting 4 features yields effective results, closely following RDA's 2-feature selection. For HeartEW, our method's 6-feature selection performs well, with RDA leading at 4 features. On M-of-n, our approach ties for the best, choosing 6 features, matching BA and HS. For Wine, selecting 3 features demonstrates a strong balance, with only MA reducing further to 2 features. For Zoo, the proposed method is among the top with 5 features, matching HS and PSO. In SpectEW, our method's 10 features achieve good performance, although RDA's minimal single-feature choice leads in terms of feature reduction.

The proposed method demonstrates superior performance in terms of classification accuracy and reduced feature selection across the datasets. For instance, it achieves a classification accuracy of 99.53% with only 173 features on the CNAE-9 dataset. In contrast, competing algorithms require more features to achieve the same accuracy. Table 4 summarizes the precision, recall, and F1-score for the datasets, highlighting the effectiveness of the proposed method, particularly on the Internet Advertisements dataset, where it achieves perfect precision, recall, and F1-score.

**Table 3**
Comparison of the proposed approach with other optimization algorithms in terms of classification accuracy (%) and number of features selected. Values in the parenthesis indicate the number of features selected.

| Dataset | Proposed | BA | WOA | PSO | MA | GA | HS | RDA | EO | GWO |
|---|---|---|---|---|---|---|---|---|---|---|
| Zoo | **100.0** (5) | 90.00 (6) | **100.0** (6) | 95.50 (7) | **100.0** (7) | **100.0** (5) | 95.00 (5) | **100.0** (6) | **100.0** (6) | **100.0** (5) |
| CongressEW | 98.85 (4) | 96.50 (4) | 94.20 (1) | 98.80 (4) | 98.80 (5) | **100.0** (5) | 98.80 (5) | 98.80 (3) | 97.70 (11) | 98.80 (4) |
| Exactly2 | **79.50** (8) | 77.50 (3) | 76.00 (1) | 76.00 (1) | 79.00 (1) | 77.50 (3) | 76.00 (3) | 76.00 (2) | 76.00 (2) | 78.50 (2) |
| BreastEW | **98.24** (6) | 95.60 (9) | 93.80 (1) | 94.70 (5) | 95.60 (8) | 95.60 (11) | 95.60 (9) | 95.60 (11) | 96.40 (10) | 96.40 (11) |
| Wine | **100.0** (4) | 97.20 (5) | **100.0** (5) | 91.67 (2) | 97.20 (7) | **100.0** (8) | **100.0** (5) | 97.20 (5) | **100.0 (3)** | **100.0** (6) |
| PenglungEW | 93.33 (75) | 73.30 (103) | 93.30 (27) | 80.00 (102) | 86.60 (140) | 86.60 (120) | **100.0** (108) | 93.30 (157) | 86.60 (236) | 93.30 (104) |
| M-of-n | **100.0 (6)** | 84.00 (6) | 80.00 (13) | 99.50 (8) | **100.0** (8) | **100.0** (7) | 83.50 (6) | 94.00 (8) | 97.50 (7) | **100.0** (7) |
| Vote | **98.33** (4) | 95.00 (4) | 96.60 (2) | 98.30 (5) | 86.60 (3) | 98.30 (4) | 95.00 (5) | 98.30 (5) | 96.60 (11) | 98.30 (3) |
| Exactly | **100.0** (6) | 68.00 (3) | 69.00 (1) | 92.50 (1) | 96.00 (2) | 80.00 (7) | 69.0 (1) | **100.0 (1)** | 95.50 (8) | **100.0** (8) |
| HeartEW | 85.10 (6) | 83.30 (6) | 81.40 (10) | 81.40 (6) | 81.40 (4) | 83.30 (7) | 85.10 (6) | 83.30 (7) | **88.80** (6) | 87.00 (6) |
| SpectEW | 90.74 (10) | 83.30 (8) | 87.00 (1) | 79.60 (8) | 90.70 (4) | 87.00 (8) | 87.00 (8) | **92.50** (8) | **92.50** (10) | 88.80 (10) |
| Lymphography | **96.66** (8) | 90.00 (5) | 80.00 (3) | 90.00 (6) | 93.30 (9) | 96.60 (6) | 93.30 (7) | 90.00 (6) | 90.00 (11) | 86.60 (7) |
| Ionosphere | **97.14** (11) | 92.80 (13) | 92.80 (12) | 92.80 (12) | 90.00 (8) | 94.20 (16) | 92.80 (13) | 92.80 (14) | 92.80 (17) | 88.50 (15) |
| BreastCancer | **99.28** (3) | 96.40 (3) | 93.50 (1) | 97.80 (3) | 97.10 (4) | 97.10 (2) | 98.50 (2) | 97.80 (4) | 97.10 (4) | 99.20 (3) |
| CNAE-9 | 99.53(173) | 99.53(256) | 99.53(223) | 99.53(300) | 99.53(616) | 99.53(270) | 99.53(274) | 99.53(**14**) | 99.53(380) | 99.53(270) |
| Parkinson’s Disease | **81.57**(236) | 78.28(224) | 75.00(174) | 76.31(292) | 69.08(529) | 73.68(246) | 75.00(233) | 75.65(**23**) | 78.28(343) | 73.02(477) |
| Internet Advertisements | **100.00**(430) | 99.78(491) | **100.00**(601) | **100.00**(600) | 93.43(1144) | 99.78(493) | **100.00**(536) | 99.78(203) | **100.00**(723) | **100.00**(506) |
| Average Rank (Features) | 3.94 | 3.82 | 3.17 | 4.17 | 6.05 | 5.64 | 4.35 | 4.64 | 3.58 | 5.35 |
| Average Rank (Accuracy) | 1.35 | 6.52 | 5.88 | 5.52 | 5.23 | 7.35 | 4.41 | 3.82 | 3.58 | 3.29 |

**Table 4**
Summary of precision, recall, and F1-score for datasets with very high dimensions.

| Dataset | Precision | Recall | F1-Score |
|---|---|---|---|
| Parkinson’s Disease | 0.81 | 0.82 | 0.79 |
| CNAE-9 | 0.99 | 1.00 | 0.99 |
| Internet Advertisements | 1.00 | 1.00 | 1.00 |

In essence, the proposed method demonstrates an effective balance between feature reduction and classification accuracy, consistently achieving high accuracy across diverse datasets while maintaining an appropriate number of features. This balance highlights its suitability for constructing models well-suited to complex classification tasks.

### *5.3.2. Comparison of the proposed method with optimization algorithms in other metrics*

The performance metrics presented in Tables 5, 6, and 7 compare the robustness of the proposed method across a variety of datasets with other optimization algorithms with respect to Precision, Recall and F1-score. As evident from the tables, our method performs consistently better compared to other optimization algorithms across all datasets.

### *5.3.3. Comparison of the proposed method with other state-of-the-art meta-heuristic algorithms based on classification accuracy and number of features of selected*

The detailed analysis presented in Table 8 meticulously examines the efficacy of the proposed method relative to Binary Grey Wolf Optimization and Particle Swarm Optimization (BGWOPSO) [48], Binary Arithmetic Optimization Algorithm (BAOA) [49], Enhanced Capuchin Search Algorithm (ECapSA) [50], Whale Optimization Algorithm with Simulated Annealing (WOASAT-2) [51], Ant Lion Optimization (ALO) using S-shaped transfer function (bALO1) [52], and ALO using V-shaped transfer function (bALO2) [52], Discrete EO combined with SA (DEOSA) [53], Population-based Binary Gaining Sharing Knowledge (PBGSK) [54], Promoted Crow Search Algorithm (PCSA) [55], Modified Binary Particle Swarm Optimization (MBPSO) [56], and Modified Grey Wolf Optimization (MGWO) [57]diverse datasets, yielding significant insights, regarding classification accuracy and numbers of features selected. Our approach has demonstrated either the highest or notably superior accuracy. For datasets like Wine and Zoo, the proposed method achieves the highest possible accuracy of 100%, matching or surpassing algorithms such as BGWOPSO and WOASAT-2 while significantly outperforming methods like bALO1 and ECapSA. For other datasets, including BreastCancer and Exactly, it consistently ranks among the top performers, achieving 99.28% and 100%, respectively. Even in complex datasets like PenglungEW and Ionosphere, the method maintains competitive performance, closely following leading algorithms such as BGWOPSO and BAOA in accuracy, demonstrating its robustness across diverse conditions.

Table 8 also provides a comprehensive comparison of the proposed method’s efficiency in feature selection against other methods like BGWOPSO, BAOA, ECapSA, WOASAT-2, bALO1, HGW0PSO, DEOSA, PBGSK_S1, PBGSK_S2, and bALO2, across various datasets. This analysis is particularly insightful in understanding how many features each method requires to achieve effective classification accuracy. The proposed method demonstrates efficient feature selection across various datasets, consistently achieving competitive performance. For example, in the Wine dataset, it requires only 3 features, matching the efficiency of BGWOPSO, DEOSA, and BAOA. In datasets like Exactly2 and BreastEW, it selects 8 and 10 features, respectively—fewer than methods like bALO1 and PBGSK_S1, yet remains competitive among other approaches. For PenglungEW, the method’s 75-feature requirement is moderate relative to approaches like WOASAT-2 and DEOSA. In datasets such as M-of-n and Vote, it selects 6 and 4 features, respectively, aligning well with other top-performing methods. Similarly, in HeartEW and Exactly, it requires only 6 features, demonstrating its effectiveness across a range of feature selection requirements.

### *5.3.4. Ablation study*

This section provides the results of an ablation study examining the performance of various modifications and combinations of modifications to a GA across different datasets. The columns in Table 9 represent different configurations of the GA: GA (the baseline GA without any modifications), GA+a (Modified Crossover), GA+b (Adaptive Threshold-based Roulette Wheel), GA+c (Cluster-based Population Initialization), GA+ab (Modified Crossover + Adaptive Threshold-based Roulette Wheel), GA+bc (Adaptive Threshold-based Roulette Wheel + Cluster-based Population Initialization), GA+ca (Cluster-based Population Initialization + Modified Crossover), and the proposed method (GA+Modified Crossover + Adaptive Threshold-based Roulette Wheel + Cluster-based Population Initialization).

The performance metric reported for each configuration is the accuracy of the algorithm, with the number of features selected in parentheses. The summary of findings reveals that the best performance for the Breastcancer dataset (99.28%) is observed with the proposed method, significantly improving over the baseline GA. For BreastEW, the highest accuracy (98.24%) is achieved by combinations including GA+b and GA+c. Given equal accuracy, the combination with fewer features is preferable.

The CongressEW dataset shows the baseline GA achieving perfect accuracy (100%), indicating high effectiveness without modifications. For SpectEW, the proposed method achieves the highest accuracy (90.74%), highlighting the importance of using them together. The Exactly dataset shows both GA+ab and the proposed method achieving

**Table 5**
Comparison of the proposed method with other optimization algorithms in terms of precision.

| Dataset | Proposed | BA | WOA | PSO | MA | GA | HS | RDA | EO | GWO |
|---|---|---|---|---|---|---|---|---|---|---|
| Zoo | 1.00 | 0.90 | 1.00 | 0.91 | 1.00 | 1.00 | 0.91 | 1.00 | 1.00 | 1.00 |
| CongressEW | 0.99 | 0.97 | 0.94 | 0.99 | 0.99 | 1.00 | 0.99 | 0.98 | 0.98 | 0.99 |
| Exactly2 | 0.78 | 0.76 | 0.76 | 0.58 | 0.77 | 0.75 | 0.58 | 0.75 | 0.76 | 0.77 |
| BreastEW | 0.98 | 0.96 | 0.94 | 0.95 | 0.96 | 0.96 | 0.96 | 0.96 | 0.97 | 0.96 |
| Wine | 1.00 | 0.97 | 1.00 | 0.93 | 0.97 | 1.00 | 1.00 | 0.97 | 1.00 | 1.00 |
| PenglungEW | 0.95 | 0.75 | 0.94 | 0.68 | 0.85 | 0.85 | 0.85 | 0.93 | 0.85 | 0.91 |
| M-of-n | 1.00 | 0.84 | 0.80 | 0.99 | 1.00 | 1.00 | 1.00 | 0.94 | 0.96 | 1.00 |
| Vote | 0.98 | 0.95 | 0.97 | 0.98 | 0.87 | 0.98 | 0.95 | 0.99 | 0.97 | 0.98 |
| Exactly | 1.00 | 0.68 | 0.69 | 0.93 | 0.96 | 0.80 | 0.48 | 1.00 | 0.94 | 1.00 |
| HeartEW | 0.86 | 0.84 | 0.82 | 0.98 | 0.82 | 0.84 | 0.85 | 0.84 | 0.89 | 0.87 |
| SpectEW | 0.91 | 0.83 | 0.87 | 0.63 | 0.91 | 0.87 | 0.87 | 0.91 | 0.93 | 0.89 |
| Lymphography | 0.94 | 0.90 | 0.80 | 0.87 | 0.94 | 0.97 | 0.90 | 0.90 | 0.90 | 0.87 |
| Ionosphere | 0.97 | 0.93 | 0.93 | 0.92 | 0.90 | 0.95 | 0.94 | 0.93 | 0.93 | 0.89 |
| BreastCancer | 0.99 | 0.97 | 0.92 | 0.97 | 0.97 | 0.97 | 0.99 | 0.98 | 0.97 | 0.99 |

**Table 6**
Comparison of the proposed method with other optimization algorithms in terms of recall.

| Dataset | Proposed | BA | WOA | PSO | MA | GA | HS | RDA | EO | GWO |
|---|---|---|---|---|---|---|---|---|---|---|
| Zoo | 1.00 | 0.90 | 1.00 | 0.97 | 1.00 | 1.00 | 0.95 | 1.00 | 1.00 | 1.00 |
| CongressEW | 0.99 | 0.97 | 0.95 | 0.99 | 0.99 | 1.00 | 0.99 | 0.99 | 0.98 | 0.99 |
| Exactly2 | 0.80 | 0.79 | 0.76 | 0.77 | 0.77 | 0.76 | 0.79 | 0.76 | 0.76 | 0.77 |
| BreastEW | 0.98 | 0.96 | 0.94 | 0.95 | 0.96 | 0.95 | 0.95 | 0.96 | 0.97 | 0.96 |
| Wine | 1.00 | 0.97 | 1.00 | 0.94 | 0.97 | 1.00 | 1.00 | 0.97 | 1.00 | 1.00 |
| PenglungEW | 0.95 | 0.74 | 0.93 | 0.80 | 0.87 | 0.87 | 1.00 | 0.93 | 0.87 | 0.93 |
| M-of-n | 1.00 | 0.84 | 0.80 | 0.92 | 1.00 | 1.00 | 0.83 | 0.94 | 0.98 | 1.00 |
| Vote | 0.98 | 0.95 | 0.97 | 0.98 | 0.84 | 0.98 | 0.95 | 0.98 | 0.97 | 0.98 |
| Exactly | 1.00 | 0.68 | 0.69 | 0.93 | 0.96 | 0.80 | 1.00 | 1.00 | 0.95 | 1.00 |
| HeartEW | 0.85 | 0.83 | 0.82 | 0.82 | 0.82 | 0.84 | 0.84 | 0.82 | 0.89 | 0.87 |
| SpectEW | 0.91 | 0.84 | 0.88 | 0.80 | 0.91 | 0.87 | 0.86 | 0.93 | 0.93 | 0.89 |
| Lymphography | 0.97 | 0.90 | 0.80 | 0.90 | 0.93 | 0.97 | 0.94 | 0.90 | 0.90 | 0.87 |
| Ionosphere | 0.97 | 0.93 | 0.93 | 0.93 | 0.90 | 0.94 | 0.93 | 0.93 | 0.93 | 0.89 |
| BreastCancer | 0.99 | 0.97 | 0.94 | 0.97 | 0.98 | 0.97 | 0.99 | 0.98 | 0.97 | 0.99 |

**Table 7**
Comparison of the proposed method with other optimization algorithms in terms of F1 score (%).

| Dataset | Proposed | BA | WOA | PSO | MA | GA | HS | RDA | EO | GWO |
|---|---|---|---|---|---|---|---|---|---|---|
| Zoo | 1.00 | 0.90 | 1.00 | 0.94 | 1.00 | 1.00 | 0.95 | 1.00 | 1.00 | 1.00 |
| CongressEW | 0.99 | 0.97 | 0.94 | 0.99 | 0.99 | 1.00 | 0.99 | 0.99 | 0.98 | 0.99 |
| Exactly2 | 0.76 | 0.77 | 0.76 | 0.66 | 0.77 | 0.75 | 0.67 | 0.75 | 0.76 | 0.78 |
| BreastEW | 0.98 | 0.96 | 0.94 | 0.95 | 0.96 | 0.95 | 0.95 | 0.96 | 0.96 | 0.96 |
| Wine | 1.00 | 0.97 | 1.00 | 0.93 | 0.97 | 1.00 | 1.00 | 0.97 | 1.00 | 1.00 |
| PenglungEW | 0.95 | 0.73 | 0.93 | 0.74 | 0.86 | 0.86 | 0.92 | 0.93 | 0.86 | 0.93 |
| M-of-n | 1.00 | 0.84 | 0.80 | 0.95 | 1.00 | 1.00 | 0.91 | 0.94 | 0.96 | 1.00 |
| Vote | 0.98 | 0.95 | 0.97 | 0.98 | 0.85 | 0.98 | 0.95 | 0.98 | 0.96 | 0.98 |
| Exactly | 1.00 | 0.68 | 0.69 | 0.93 | 0.96 | 0.80 | 0.65 | 1.00 | 0.94 | 1.00 |
| HeartEW | 0.85 | 0.83 | 0.82 | 0.89 | 0.82 | 0.84 | 0.84 | 0.83 | 0.89 | 0.87 |
| SpectEW | 0.91 | 0.83 | 0.87 | 0.70 | 0.91 | 0.87 | 0.86 | 0.92 | 0.93 | 0.89 |
| Lymphography | 0.95 | 0.90 | 0.80 | 0.88 | 0.93 | 0.97 | 0.92 | 0.90 | 0.90 | 0.86 |
| Ionosphere | 0.97 | 0.93 | 0.93 | 0.92 | 0.90 | 0.94 | 0.93 | 0.93 | 0.93 | 0.88 |
| BreastCancer | 0.99 | 0.96 | 0.93 | 0.97 | 0.97 | 0.97 | 0.99 | 0.98 | 0.97 | 0.99 |

perfect accuracy (100%). Given the same accuracy, the combination with fewer features is preferred.

The best accuracy for Exactly2 (79.5%) is achieved by combinations including GA+ab, with fewer features being more beneficial. The Zoo dataset indicates that the baseline GA and all modifications achieve perfect accuracy (100%), showing the dataset's ease of classification with GA. For the Wine dataset, the baseline GA achieves perfect accuracy (100%), with modifications not degrading performance, showing robustness. The M-of-n dataset shows that the baseline GA and most modifications achieve perfect accuracy (100%), indicating minimal room for improvement.

For Vote, the highest accuracy (100%) is achieved with GA+c, showing significant impact. The Lymphography dataset's highest accuracy (96.67%) is achieved by several configurations, with the least number of features preferred if accuracies are equal. For Ionosphere, the highest accuracy (97.4%) is achieved with the proposed method, demonstrating their importance. The highest accuracy for PenglungEW (93.33%) is consistently achieved by combinations including GA+b and GA+c.

**Table 8**
Comparison of the proposed approach with other State-of-the-art metaheuristic algorithms in terms of classification accuracy and the number of features selected. Values in the parenthesis indicate the number of features selected. Classification scores are in %.

| Dataset | Proposed | WOASAT-2 | ECapSA | bALO1 | BAOA | bALO2 | BGWOPSO | DEOSA | PBGSK_S1 | PBGSK_S2 | PCSA | MPSO | MGWO |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Zoo | **100.0**(5) | 97.00(5.60) | 98.51(4.063) | 88.80(7.3) | 96.00(5.05) | 98.00(5.5) | **100.0**(6.8) | **100.0(4)** | 93.30(5.8) | 93.70(6.2) | 95.71(7) | 100.00(7) | 100.00(6) |
| CongressEW | **98.85**(4) | 98.00(6.40) | 97.36(3.304) | 90.30(7.55) | 97.00(**1.26**) | 98.00(6.95) | 98.00(4.4) | 96.50(5) | 96.20(4.4) | 96.40(5) | 95.86(4) | 95.40(3) | 95.40(1) |
| Exactly2 | **79.50**(8) | 75.00(2.80) | 76.25(1.846) | 70.20(2.5) | 78.00(1.15) | 76.20(1.5) | 76.00(1.6) | 76.00(**1**) | 76.80(4) | 76.40(2.8) | 75.98(2) | 76.00(1) | 76.00(1) |
| BreastEW | **98.24**(10) | 98.00(11.6) | 96.14(**4.048**) | 93.40(16.1) | 98.00(9.65) | 97.40(14.9) | 97.00(13.6) | 96.40(8) | 95.10(12.4) | 95.10(11.2) | 94.84(11) | 94.73(6) | 96.49(5) |
| Wine | **100.0(3)** | 99.00(6.40) | 98.87(5.192) | 89.00(7.45) | 98.00(4.35) | 96.90(6.6) | **100.0**(6) | **100.0(3)** | 94.30(5.2) | 94.30(5.2) | 97.86(5) | 97.22(2) | 97.22(3) |
| PenglungEW | 93.33(75) | 94.00(127.4) | 92.85(**4.291**) | 74.50(146.9) | 96.00(15.2) | 82.60(130.15) | 96.00(130.8) | **100.00**(86) | 82.50(149) | 82.50(159.4) | 89.87(198) | 86.00(131) | **100.00(33)** |
| M-of-n | **100.0**(6) | **100.0**(6.00) | **100.0(4.615)** | 71.10(8.5) | 92.00(5.925) | 96.90(6) | **100.0**(6.0) | **100.0**(6) | 99.20(6.6) | 97.40(7) | 100.00(7) | **100.00(6)** | **100.00(6)** |
| Vote | **98.33**(4) | 97.00(5.20) | 96.15(2.688) | 90.10(7.2) | 98.00(2.63) | 96.90(6.65) | 97.00(3.4) | 96.6(**1**) | 96.20(4.8) | 96.10(7) | 95.23(5) | 96.66(4) | 96.66(3) |
| Exactly | **100.0**(6) | **100.0**(6.00) | **100.0**(4.615) | 64.50(7.7) | 84.00(**3.95**) | 98.50(6) | **100.0**(6.0) | **100.0**(7) | 91.80(7) | 95.60(6.4) | 100.00(7) | 93.00(8) | 69.00(1) |
| HeartEW | 85.10(6) | 85.00(5.40) | 83.71(**4.846**) | 74.10(8.9) | 85.00(5.00) | 83.90(7.45) | 85.00(5.8) | **90.70**(5) | **95.10**(12.4) | **95.10**(11.2) | 83.48(4) | **92.59(3)** | 90.79(3) |
| SpectEW | **90.74**(10) | 88.00(9.40) | 85.21(4.318) | 79.60(9.1) | 86.00(**3.60**) | 89.80(7.25) | 88.00(8.4) | 90.70(9) | 87.10(9.6) | 87.01(10) | 83.52(8) | **90.74(7)** | 79.62(1) |
| Lymphography | **96.66**(8) | 89.00(7.20) | 87.08(**4.778**) | 71.50(8.4) | 86.00(4.8) | 91.30(7.8) | 92.00(9.2) | 96.60(9) | 84.00(8.2) | 84.00(7.8) | 85.48(9) | 90.00(5) | 93.33(7) |
| Ionosphere | **97.14**(11) | 96.00(12.8) | 94.22(**3.655**) | 83.00(13.5) | 92.00(5.30) | 89.60(11.6) | 95.00(13) | 95.70(7) | 88.80(12.4) | 87.60(12) | 93.13(8) | 91.42(9) | 95.71(3) |
| BreastCancer | **99.28(3)** | 97.00(4.20) | 97.36(5.238) | 95.10(4.65) | 98.00(3.6) | 97.30(4.4) | 98.00(4.4) | 96.40(3) | 97.40(4.8) | 97.40(4.8) | 96.60(3) | 99.21(9) | 97.85(2) |

**Table 9**
Results of the ablation study in terms of classification accuracy and the number of features selected. Values in the parenthesis indicate the number of features selected. Classification scores are in %.

| Dataset | GA | GA+a | GA+b | GA+c | GA+ab | GA+bc | GA+ca | Proposed GA |
|---|---|---|---|---|---|---|---|---|
| Breastcancer | 97.10(4) | 96.82(2) | 97.14(3) | 98.57(3) | 97.85(2) | 99.28(5) | 99.28(5) | 99.28(3) |
| BreastEW | 95.60(8) | 92.10(7) | 96.49(12) | 98.24(11) | 98.24(11) | 98.24(11) | 98.24(11) | 98.24(10) |
| CongressEW | 100.00(5) | 95.40(4) | 96.55(5) | 97.70(2) | 97.70(5) | 97.70(2) | 95.61(6) | 98.85(4) |
| SpectEW | 87.00(4) | 88.88(13) | 88.88(10) | 88.88(11) | 88.88(7) | 88.88(8) | 88.88(10) | 90.74(10) |
| Exactly | 80.00(2) | 87.00(8) | 98.00(7) | 88.50(8) | 98.50(7) | 100.00(7) | 97.50(7) | 100.00(6) |
| Exactly2 | 77.50(1) | 76.00(1) | 78.00(5) | 78.50(6) | 79.50(9) | 79.00(4) | 78.00(5) | 79.50(8) |
| Zoo | 100.00(7) | 100.00(6) | 100.00(6) | 100.00(6) | 100.00(6) | 100.00(6) | 100.00(6) | 100.00(5) |
| Wine | 100.00(7) | 100.00(6) | 100.00(6) | 100.00(5) | 100.00(5) | 100.00(4) | 100.00(4) | 100.00(3) |
| M-of-n | 100.00(8) | 100.00(7) | 100.00(7) | 100.00(7) | 100.00(6) | 100.00(7) | 100.00(7) | 100.00(6) |
| Vote | 98.30(3) | 98.33(7) | 100.00(7) | 98.33(6) | 98.33(6) | 98.33(4) | 96.67(1) | 98.33(4) |
| Lymphography | 96.66(9) | 96.66(8) | 96.67(7) | 96.67(7) | 96.67(8) | 96.67(5) | 96.67(6) | 96.67(8) |
| Ionosphere | 94.2(8) | 94.28(7) | 94.28(8) | 95.714(12) | 92.85(5) | 95.71(10) | 95.71(12) | 97.4(11) |
| PenglungEW | 86.60(140) | 88.66(77) | 86.66(102) | 93.33(93) | 93.33(86) | 93.33(90) | 93.33(87) | 93.33(75) |
| HeartEW | 85.18(13) | 85.15(3) | 85.18(6) | 85.18(7) | 87.03(6) | 85.18(6) | 83.33(4) | 85.10(6) |

For HeartEW, the highest accuracy (87.03%) is achieved with the GA+b modification, indicating its effectiveness. In conclusion, combinations of modifications, particularly the proposed method, often result in the highest accuracies, suggesting synergistic effects. GA+b and GA+c frequently contribute to improved performance. When modifications yield the same accuracy, the configuration with fewer features is more advantageous, highlighting the importance of feature reduction alongside accuracy improvements. The impact of modifications varies across datasets, underscoring the need for dataset-specific tuning. This study demonstrates the potential of tailored modifications to enhance the effectiveness of GAs in various classification tasks.

### 5.3.5. *Microarray datasets*

In previous literature, we examine the effectiveness of our proposed feature selection technique against various optimization algorithms using 14 standard UCI datasets. This section focuses on evaluating our system's performance on high-dimensional datasets, specifically microarray-based gene expression datasets. Microarray data analysis plays a crucial role in various domains, including pharmacology, medicine, and biology. The data collection is crucial for predicting cancer based on gene expression data.

Table 10 provides a detailed overview of the microarray datasets utilized in our experiments. The DLBCL dataset is integral in researching the predominant variant of non-Hodgkin lymphoma, delving into the gene expression profiles of affected cells to better comprehend the disease's mechanisms and to forge specific therapeutic approaches. The AMLGSE2191 dataset focuses on Acute Myeloid Leukemia, playing a vital role in identifying unique genetic markers that are essential for crafting personalized treatment plans. The SRBCT dataset encompasses, covering a range of childhood cancers characterized by distinct cellular features, with its microarray analysis being crucial for accurate diagnosis and effective treatment strategies. In the realm of bladder cancer, the bladderGSE89 dataset provides critical insights into the genetic underpinnings of the disease, paving the way for novel therapeutic interventions. Furthermore, the Leukemia dataset, while bearing similarities to the AMLGSE2191 in its leukemia focus, offers broader insights into the disease's various forms, aiding in the understanding of its progression and the complexities of treatment resistance. Collectively, these microarray datasets are invaluable in the field of oncology, significantly enhancing diagnostic accuracy, prognostic assessments, and the customization of treatment modalities.

It is important to note the contrast in these datasets: they have a high number of features but a relatively low number of samples. The effectiveness in terms of accuracy and the number of relevant features selected by our hybrid feature selection system are detailed in Table 11. The results showcased in these tables underscore the system's capability to select pertinent features with high accuracy from datasets with very high dimensionality.

Microarray datasets, characterized by their high feature count and low sample size, present unique challenges in feature selection. As previously noted, it is impractical to consider every possible combination of features, given that the number of potential feature subsets is $2^N - 1$, with $N$ being the total number of features. This situation demands the exclusion of extraneous genes to ensure both efficiency and efficacy in the analysis. In comparison to standard UCI datasets, microarray datasets are often burdened with a larger number of irrelevant features, emphasizing the importance of their removal.

The main goal in analyzing microarray data is to uncover the intricate interrelations among genes, particularly to identify gene associations linked to specific cancer types. Notably, these datasets contain certain biomarkers that are key to predicting cancer. The evidence and discussions presented in the tables we referenced earlier affirm the effectiveness of our hybrid feature selection system in managing and extracting meaningful insights from high-dimensional data, particularly in identifying and focusing on the most relevant features for cancer prediction.

**Table 10**
Overview of the microarray data sets employed in the experimental studies.

| Dataset | Classes | Samples | Attributes | Use |
|---|---|---|---|---|
| DLBCL | 2 | 77 | 7070 | Detection of diffused large B-cell lymphoma |
| AMLGSE2191 | 2 | 54 | 12 626 | Acute Myeloid Leukemia detection |
| SRBCT | 4 | 83 | 2309 | Analysis of small-round-blue-cell tumors |
| bladderGSE89 | 3 | 40 | 5725 | bladder cancer detection |
| Leukemia | 2 | 72 | 5148 | Identification of leukemia. |

**Table 11**
Comparison of the proposed method with other optimization algorithms in terms of classification accuracy (%) and number of features selected. Values in the parenthesis indicate the number of features selected.

| Dataset | Proposed | WOA | MA | PSO | EO | RDA | GWO | BA | HS | GA |
|---|---|---|---|---|---|---|---|---|---|---|
| Leukemia | 100.0 (8) | 100.0 (89) | 93.30 (219) | 100.0 (117) | 93.30 (138) | 100.0 (10) | 100.0 (168) | 100.0 (126) | 86.60 (170) | 100.0 (128) |
| DLBCL | 100.0 (160) | 93.75 (150) | 100.0 (159) | 100.0 (149) | 93.75 (139) | 93.75 (51) | 93.75 (185) | 93.75 (92) | 93.71 (91) | 93.75 (95) |
| bladderGSE89 | 100.0 (7) | 100.0 (104) | 100.0 (159) | 100.0 (104) | 100.0 (123) | 100.0 (113) | 100.0 (170) | 100.0 (108) | 87.50 (104) | 100.0 (145) |
| AMLGSE2191 | 100.0 (8) | 100.0 (159) | 81.81 (208) | 72.72 (131) | 100.0 (169) | 90.90 (120) | 90.90 (181) | 72.72 (128) | 72.72 (98) | 100.0 (158) |
| SRBCT | 100.0 (138) | 100.0 (75) | 82.30 (210) | 100.0 (104) | 100.0 (140) | 94.10 (122) | 100.0 (173) | 100.0 (92) | 100.0 (151) | 100.0 (124) |

#### 5.3.6. Comparison with other optimization algorithms for microarray datasets

From Table 11 it is clear that our proposed method has achieved classification accuracy of 100% outperforming all the other metaheuristic nature-based wrappers.

In our study, the method we have developed demonstrates a remarkable efficiency in feature selection across a range of datasets, markedly outshining other algorithms. In the SRBCT dataset, the proposed method's selection of 138 features further exemplifies its adeptness at striking a middle ground, navigating between the higher feature counts of MA and the more conservative selections by WOA. The precision of our approach is even more evident in the Leukemia and AMLGSE2191 datasets. By selecting only 8 features in each, our method significantly surpasses its competitors. Similarly, in the bladderGSE89 dataset, the efficiency and effectiveness of our method are once again highlighted as it chose the least number of features — only 7. This is significantly lower than other algorithms.

Overall, the results from these datasets collectively highlight the reliability of our method. It not only consistently matches but often exceeds the performance of established algorithms, affirming its position as a powerful tool for data analysis, particularly in scenarios with medium to high feature ranges. The consistency in performance across datasets of varying feature counts further cements its standing as a dependable and adaptable data analysis methodology.

#### 5.3.7. Statistical test

Table 12 in our study details the outcomes of a statistical assessment designed to measure the significance of our proposed method against other existing methods. Before initiating this test, we have accumulated statistical evidence to verify the significance of the provided data. Our initial hypothesis posits that "The results achieved by the proposed algorithm are on par with those obtained by existing algorithms". The Wilcoxon Rank-Sum Test, aimed at contrasting each method against our proposed one, serves to challenge this hypothesis. This analysis includes 14 distinct runs per method on each dataset to ascertain the $p$-value, which is then juxtaposed with our method. We document the accuracy of classification for each run. A $p$-value below <0.05 (5%) leads us to discard the hypothesis, indicating inadequate evidence to uphold the original assertion.

Analysis of the data in Table 12 demonstrates that in 108 of the 126 scenarios, the hypothesis was not supported. This statistically significant result reflects the algorithm's robustness across different data splits, indicating that our algorithm consistently delivers reliable results, regardless of how the data is partitioned into training and test sets. This statistical validation emphasizes the algorithm's dependability across diverse data configurations. From these findings, we infer that our methodology exhibits a statistically improved performance in comparison to existing algorithms.

## 6. Discussion

In this section, we illustrate the convergence trends of our advanced hybrid feature selection algorithm through Fig. 2. This algorithm, characterized by its unique integration of GAs, cluster-based population initialization, modified crossover, and parent selection using an adaptive threshold-based roulette wheel method, demonstrates consistent convergence across iterative processes. The convergence graphs highlight the effectiveness of our approach, where the mean fitness of candidate solutions is utilized as a key metric. The balanced approach, combining exploration and exploitation through innovative techniques, as mentioned earlier, ensures a thorough and effective search process, crucial for feature selection tasks in complex datasets. The empirical results, as depicted in the convergence graphs, underscore the efficacy of our proposed method in navigating and optimizing the feature selection landscape.

The convergence trends of various methods considered in this study are depicted in Fig. 3 for the purpose of comparison. In the majority of instances, it is noticeable that the mean fitness gradually converges towards an optimal solution with the advancement of iterations. The figures reveal that most algorithms quickly converge to optimal solutions, maintaining stability in later iterations. In contrast, our method demonstrates a sequential convergence towards its optimal solution for most of the datasets as the iterations advance. Additionally, it is evident that for specific datasets such as M-of-n, Exactly, and Ionosphere, some algorithms reach their optimal solutions in the early iterations, showing minimal improvement throughout the remaining iterations. Notably, our base algorithm, GA, although it converges to an optimal solution, achieves relatively lower fitness values compared to our proposed method. In this context, it is evident that employing cluster-based population initialization has conferred a substantial advantage in improving the overall quality of solutions.

In Fig. 4, we have provided box plots illustrating the classification accuracies from 14 separate runs for each method. The median is represented by the yellow line, while the lower boundary of the box indicates the first quartile of the distribution, and the upper boundary represents the second quartile. Across most scenarios, it is noticeable that the box size for our proposed method is consistently smaller compared to that of the other optimization algorithms. This observation highlights the robustness of our method in consistently generating similar results across various data splits. Additionally, we observe that the distribution of results does not exhibit significant skewness towards one end, as indicated by the relative position of the yellow line with respect to the box. Moreover, a reduced number of outliers in our results can be observed. These findings collectively contribute to the overall stability of each method. This stability is attributed to our method's ability to effectively balance exploration and exploitation, resulting in consistent and reliable outcomes. It is also important to

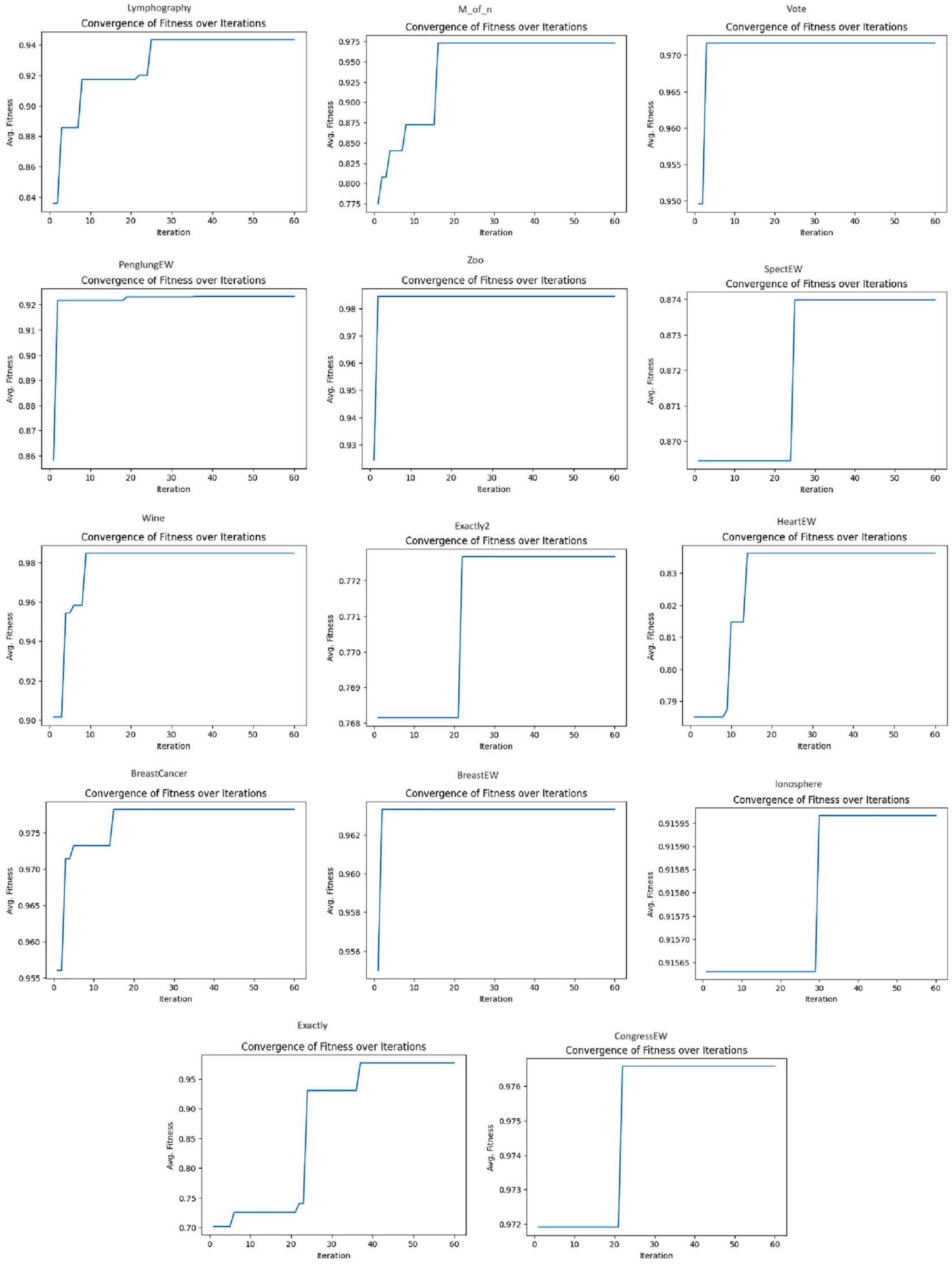

**Fig. 2.** Convergence charts for all datasets utilizing the modified GA proposed. The $x$-axis represents the advancement in iterations, while the $y$-axis denotes the average fitness value considering all agents.

emphasize that the introduction of the modified crossover mechanism, as well as the Adaptive Threshold-based Roulette Wheel method, serves to enhance the results without compromising the original outcomes.

## 7. Conclusion and future scope

Feature selection is essential in removing superfluous data from primary datasets. In our research, we introduce an improved version of the GA tailored for this purpose. This refinement includes a population initialization based on K-means clustering, which enhances the ability to explore the search space more efficiently. Moreover, the integration of a modified crossover technique and an adaptive method for parent selection, inspired by roulette strategies strikes a dynamic balance between exploration and exploitation. We have tested our proposed feature selection methodology on a range of UCI datasets and some gene expression datasets based on microarray technology, where it outperforms many existing metaheuristic-based feature selection algorithms. We have conducted a thorough examination of our findings through multiple established techniques, such as evaluating sensitivity, assessing convergence curves, and boxplot analysis, to achieve a more profound understanding of our research. The robust nature of this approach presents vast potential for application across a diverse range

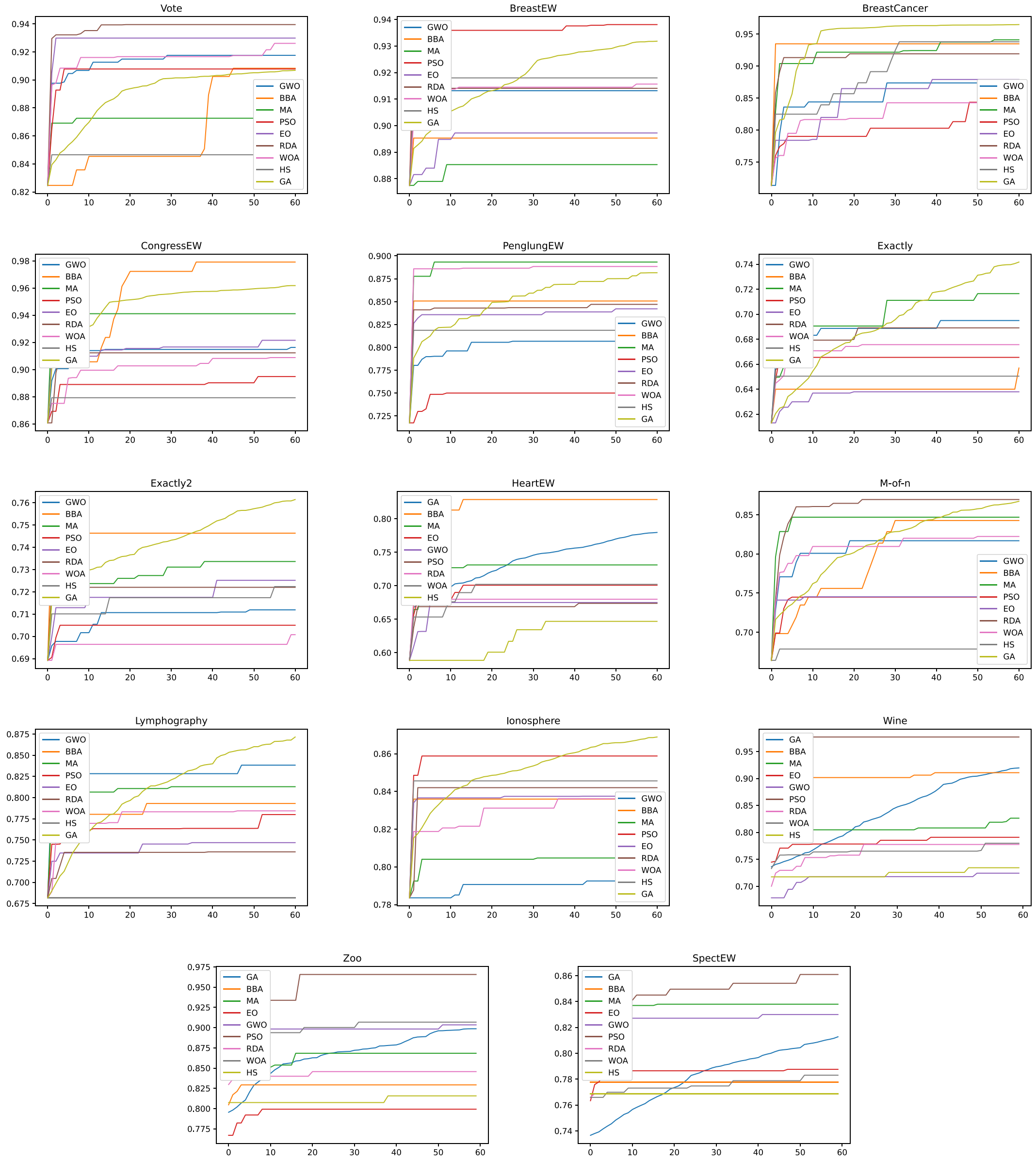

**Fig. 3.** Convergence charts pertaining to all datasets when using various optimization algorithms. Here, the $x$-axis indicates the sequence of iterations, while the $y$-axis reflects the average fitness value, accounting for all agents involved.

**Table 12**
Comparison of proposed approach with other optimization methods in terms of p-values (statistical test).

| Dataset | MA | GA | RDA | PSO | HS | GWO | EO | BA | WOA |
|---|---|---|---|---|---|---|---|---|---|
| Zoo | 8.3E−02 | 1.5E−01 | 1.4E−02 | 6.1E−05 | 1.6E−03 | 4.5E−02 | 2.5E−02 | 1.2E−03 | 1.5E−01 |
| CongressEW | 6.1E−05 | 1.8E−04 | 6.1E−05 | 3.4E−02 | 6.1E−05 | 2.0E−03 | 4.2E−03 | 4.2E−04 | 1.1E−03 |
| Exactly2 | 6.1E−05 | 6.1E−05 | 6.1E−05 | 2.7E−04 | 6.1E−05 | 6.1E−05 | 6.1E−05 | 6.1E−05 | 1.2E−04 |
| BreastEW | 1.5E−03 | 6.1E−05 | 6.1E−05 | 6.1E−05 | 6.1E−05 | 2.1E−03 | 6.1E−05 | 6.1E−05 | 6.1E−05 |
| Wine | 2.5E−02 | 6.1E−05 | 5.7E−03 | 1.1E−02 | 1.2E−03 | 1.7E−01 | 5.8E−02 | 6.4E−03 | 3.8E−02 |
| PenglungEW | 6.1E−05 | 6.1E−05 | 4.2E−04 | 5.6E−01 | 6.1E−05 | 4.5E−01 | 6.1E−05 | 8.5E−04 | 9.4E−02 |
| M-of-n | 9.4E−03 | 5.4E−04 | 9.5E−04 | 3.2E−03 | 6.1E−05 | 8.5E−01 | 1.4E−03 | 4.6E−04 | 6.1E−05 |
| Vote | 1.1E−03 | 6.1E−05 | 1.0E−02 | 8.3E−03 | 4.2E−03 | 8.3E−03 | 4.2E−03 | 6.1E−05 | 3.5E−02 |
| Exactly | 1.4E−03 | 6.1E−05 | 6.1E−05 | 2.2E−03 | 6.1E−05 | 2.1E−03 | 3.3E−03 | 8.7E−04 | 1.4E−03 |
| HeartEW | 7.5E−01 | 6.1E−05 | 6.1E−05 | 3.3E−01 | 9.2E−03 | 2.2E−02 | 7.0E−03 | 4.2E−01 | 6.4E−02 |
| SpectEW | 2.2E−01 | 6.1E−05 | 6.7E−01 | 4.8E−01 | 2.5E−02 | 9.4E−02 | 6.3E−01 | 4.1E−02 | 3.3E−03 |
| Lymphography | 6.1E−05 | 6.1E−05 | 6.1E−05 | 1.8E−03 | 6.1E−05 | 6.1E−05 | 6.1E−05 | 6.1E−05 | 6.1E−05 |
| Ionosphere | 6.1E−05 | 6.1E−05 | 6.1E−05 | 6.1E−05 | 6.1E−05 | 6.1E−05 | 6.1E−05 | 6.1E−05 | 6.1E−05 |
| BreastCancer | 6.1E−05 | 6.1E−05 | 6.1E−05 | 6.1E−05 | 6.1E−05 | 1.3E−03 | 3.0E−04 | 6.1E−05 | 3.0E−03 |

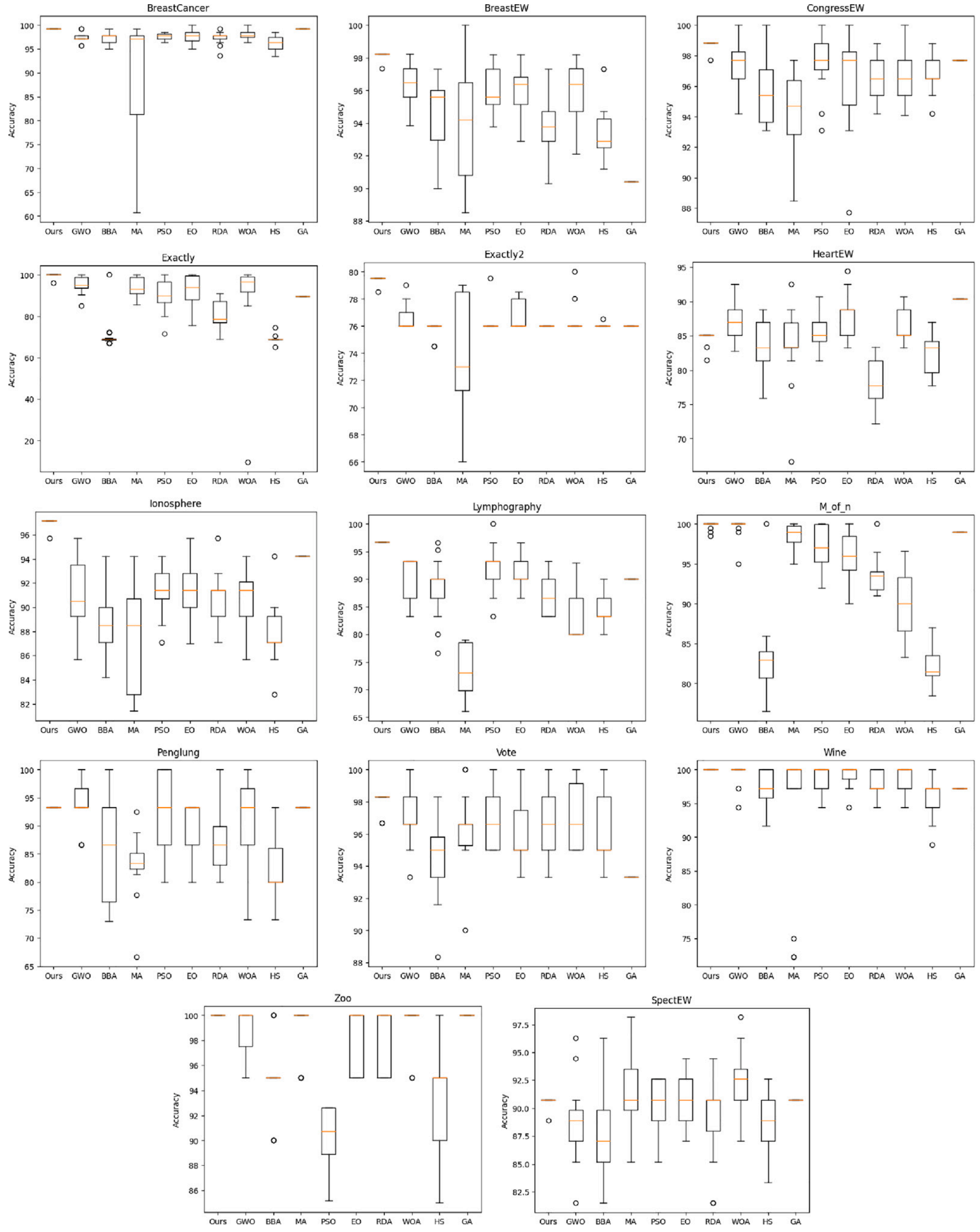

**Fig. 4.** Box-plots illustrating different feature selection algorithms, inclusive of the proposed method. These plots are based on the classification accuracy from 15 separate trials on each dataset. The *x*-axis shows the optimization algorithm, and the *y*-axis represents the accuracy.

of research fields, encompassing data mining, image processing, and data analysis. In the future, the focus could be on developing advanced strategies that more precisely modulate the reduction of diversity over time. Additionally, we have not yet analyzed datasets with a large number of classes due to the scarcity of such datasets. In the context of image datasets, it is necessary to extract features before performing feature selection. Extracting features using deep learning techniques, such as deep neural networks or employing machine learning-based methods with feature engineering can be considered in the future. We will also explore methods for adapting our approach to more complex datasets. This may involve collaborations to create suitable datasets or incorporating additional feature extraction methods.

## CRediT authorship contribution statement

**Anurup Naskar:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Soumyajit Ghosh:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Mahantapas Kundu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ram Sarkar:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

The authors extend their gratitude to the Center for Microprocessor Applications for Training Education and Research (CMATER), a research laboratory at the Department of Computer Science and Engineering, Jadavpur University, Kolkata, India, for their invaluable infrastructural support that contributed to this research.

## Data availability

Data will be made available on request.

## References

[1] D. Bertsimas, J. Tsitsiklis, Simulated annealing, Statist. Sci. 8 (1) (1993) 10–15.

[2] S.M. Almufti, R.B. Marqas, P.S. Othman, A.B. Sallow, Single-based and population-based metaheuristics for solving NP-hard problems, Iraqi J. Sci. (2021).

[3] D. Karaboğa, S. Ökdem, A simple and global optimization algorithm for engineering problems: differential evolution algorithm, Turk. J. Electr. Eng. Comput. Sci. 12 (1) (2004) 53–60.

[4] Y. Shi, Particle swarm optimization, IEEE Connect. 2 (1) (2004) 8–13.

[5] M. Dorigo, M. Birattari, T. Stutzle, Ant colony optimization, IEEE Comput. Intell. Mag. 1 (4) (2006) 28–39.

[6] D. Teodorovic, P. Lucic, G. Markovic, M. Dell'Orco, Bee colony optimization: principles and applications, in: 2006 8th Seminar on Neural Network Applications in Electrical Engineering, IEEE, 2006, pp. 151–156.

[7] N. Spolaôr, E.A. Cherman, M.C. Monard, H.D. Lee, Relieff for multi-label feature selection, in: 2013 Brazilian Conference on Intelligent Systems, IEEE, 2013, pp. 6–11.

[8] M.L. McHugh, The chi-square test of independence, Biochem. Medica 23 (2) (2013) 143–149.

[9] S. Hossain, S. Mukhopadhyay, B. Ray, S.K. Ghosal, R. Sarkar, A secured image steganography method based on ballot transform and genetic algorithm, Multimedia Tools Appl. 81 (27) (2022) 38429–38458.

[10] S. Mukhopadhyay, S. Hossain, S. Malakar, E. Cuevas, R. Sarkar, Image contrast improvement through a metaheuristic scheme, Soft Comput. (2022).

[11] S. Saha, M. Ghosh, S. Ghosh, S. Sen, P.K. Singh, Z.W. Geem, R. Sarkar, Feature selection for facial emotion recognition using cosine similarity-based harmony search algorithm, Appl. Sci. 10 (8) (2020) 2816.

[12] T. Bhattacharyya, B. Chatterjee, P.K. Singh, J.H. Yoon, Z.W. Geem, R. Sarkar, Mayfly in harmony: A new hybrid meta-heuristic feature selection algorithm, IEEE Access 8 (2020) 195929–195945.

[13] K.K. Ghosh, P.K. Singh, J. Hong, Z.W. Geem, R. Sarkar, Binary social mimic optimization algorithm with X-shaped transfer function for feature selection, IEEE Access 8 (2020) 97890–97906.

[14] A. Sharma, A. Lysenko, K.A. Boroevich, E. Vans, T. Tsunoda, DeepFeature: feature selection in nonimage data using convolutional neural network, Brief. Bioinform. 22 (6) (2021).

[15] T. Wang, W. Shao, Z. Huang, H. Tang, J. Zhang, Z. Ding, K. Huang, MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification, Nat. Commun. 12 (1) (2021) 3445.

[16] D. Surya Prabha, J. Satheesh Kumar, An efficient image contrast enhancement algorithm using genetic algorithm and fuzzy intensification operator, Wirel. Pers. Commun. 93 (2017) 223–244.

[17] R.T. Prasetio, Genetic algorithm to optimize k-nearest neighbor parameter for benchmarked medical datasets classification, J. Online Inform. (2020) 153–160.

[18] S. Marjit, T. Bhattacharyya, B. Chatterjee, R. Sarkar, Simulated annealing aided genetic algorithm for gene selection from microarray data, Comput. Biol. Med. 158 (2023) 106854.

[19] R. Guha, M. Ghosh, S. Kapri, S. Shaw, S. Mutsuddi, V. Bhateja, R. Sarkar, Deluge based genetic algorithm for feature selection, Evol. Intell. 14 (2021) 1–11, http://dx.doi.org/10.1007/s12065-019-00218-5.

[20] G. Ravi Kumar, G. Ramachandra, K. Nagamani, An efficient feature selection system to integrating SVM with genetic algorithm for large medical datasets, Int. J. Adv. Res. Comput. Sci. Softw. Eng. 4 (2014) 272–277.

[21] K. Varpa, K. Iltanen, M. Juhola, Genetic algorithm based approach in attribute weighting for a medical data set, J. Comput. Med. 2014 (2014) 1–11, http://dx.doi.org/10.1155/2014/526801.

[22] R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, S. Nalluri, Genetic algorithm based feature selection and MOE fuzzy classification algorithm on Pima Indians diabetes dataset, in: 2017 International Conference on Computing Networking and Informatics, ICCNI, IEEE, 2017, pp. 1–5.

[23] S. Aalaei, H. Shahraki, A. Rowhanimanesh, S. Eslami, Feature selection using genetic algorithm for breast cancer diagnosis: Experiment on three different datasets, Iran. J. Basic Med. Sci. 19 (2016) 476–482.

[24] J. Too, A.R. Abdullah, A new and fast rival genetic algorithm for feature selection, J. Supercomput. 77 (2021) 2844–2874.

[25] M.M. Kabir, M. Shahjahan, K. Murase, A new local search based hybrid genetic algorithm for feature selection, Neurocomputing 74 (17) (2011) 2914–2928.

[26] L.S. Oliveira, R. Sabourin, F. Bortolozzi, C.Y. Suen, Feature selection using multi-objective genetic algorithms for handwritten digit recognition, in: 2002 International Conference on Pattern Recognition, vol. 1, IEEE, 2002, pp. 568–571.

[27] S. Malakar, M. Ghosh, S. Bhowmik, R. Sarkar, M. Nasipuri, A GA based hierarchical feature selection approach for handwritten word recognition, Neural Comput. Appl. 32 (2020) 2533–2552.

[28] A.H. Khan, S.S. Sarkar, K. Mali, R. Sarkar, A genetic algorithm based feature selection approach for microstructural image classification, Exp. Tech. (2022) 1–13.

[29] S. Kumar, S. Singh, J. Kumar, Automatic live facial expression detection using genetic algorithm with haar wavelet features and SVM, Wirel. Pers. Commun. 103 (2018) 2435–2453, URL https://api.semanticscholar.org/CorpusID:53769719.

[30] X. Xiao, M. Yan, S. Basodi, C. Ji, Y. Pan, Efficient hyperparameter optimization in deep learning using a variable length genetic algorithm, 2020, arXiv preprint arXiv:2006.12703.

[31] Y. Kwon, S. Kang, Y.-S. Choi, I. Kim, Evolutionary design of molecules based on deep learning and a genetic algorithm, Sci. Rep. 11 (1) (2021) 17304.

[32] S. Bouktif, A. Fiaz, A. Ouni, M.A. Serhani, Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches, Energies 11 (7) (2018) 1636.

[33] S. Kalsi, H. Kaur, V. Chang, DNA cryptography and deep learning using genetic algorithm with NW algorithm for key generation, J. Med. Syst. 42 (2018) 1–12.

[34] H. Tian, S.-C. Chen, M.-L. Shyu, Genetic algorithm based deep learning model selection for visual data classification, in: 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science, IRI, IEEE, 2019, pp. 127–134.

[35] W. Bendali, I. Saber, B. Bourachdi, M. Boussetta, Y. Mourad, Deep learning using genetic algorithm optimization for short term solar irradiance forecasting, in: 2020 Fourth International Conference on Intelligent Computing in Data Sciences, ICDS, IEEE, 2020, pp. 1–8.

[36] S. Kilicarslan, M. Celik, Ş. Sahin, Hybrid models based on genetic algorithm and deep learning algorithms for nutritional Anemia disease classification, Biomed. Signal Process. Control. 63 (2021) 102231.

[37] L. Bi, G. Hu, A genetic algorithm-assisted deep learning approach for crop yield prediction, Soft Comput. 25 (2021) 10617–10628.

[38] F. Mattioli, D. Caetano, A. Cardoso, E. Naves, E. Lamounier, et al., An experiment on the use of genetic algorithms for topology selection in deep learning, J. Electr. Comput. Eng. 2019 (2019).

[39] S.O. Slim, M. Elfattah, A. Atia, M.-S.M. Mostafa, IoT system based on parameter optimization of deep learning using genetic algorithm, Int. J. Intell. Eng. Syst. 14 (2) (2021).

[40] H.M. Balaha, H.A. Ali, E.K. Youssef, A.E. Elsayed, R.A. Samak, M.S. Abdelhaleem, M.M. Tolba, M.R. Shehata, M.R. Mahmoud, M.M. Abdelhameed, et al., Recognizing arabic handwritten characters using deep learning and genetic algorithms, Multimedia Tools Appl. 80 (2021) 32473–32509.

[41] M. Yoosefzadeh-Najafabadi, D. Tulpan, M. Eskandari, Application of machine learning and genetic optimization algorithms for modeling and optimizing soybean yield using its component traits, PLoS One 16 (4) (2021) e0250665.

[42] M. Fernandez, A. Bilić, A.S. Barnard, Machine learning and genetic algorithm prediction of energy differences between electronic calculations of graphene nanoflakes, Nanotechnology 28 (38) (2017) 38LT03.

[43] X. Gao, G.M. Lee, Moment-based rental prediction for bicycle-sharing transportation systems using a hybrid genetic algorithm and machine learning, Comput. Ind. Eng. 128 (2019) 60–69.

[44] G. Wu, L. Si, H. Xu, R. Niu, Y. Zhuang, H. Sun, J. Ding, Phase-to-pattern inverse design for a fast realization of a functional metasurface by combining a deep neural network and a genetic algorithm, Opt. Express 30 (25) (2022) 45612–45623.

[45] J.H. Holland, Genetic algorithms, Sci. Am. 267 (1) (1992) 66–73, URL http://www.jstor.org/stable/24939139.

[46] M. Syakur, B. Khotimah, E. Rochman, B.D. Satoto, Integration k-means clustering method and elbow method for identification of the best customer profile cluster, in: IOP Conference Series: Materials Science and Engineering, vol. 336, IOP Publishing, 2018, 012017.

[47] A. Lesne, Shannon entropy: a rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics, Math. Structures Comput. Sci. 24 (3) (2014) e240311.

[48] Q. Al-Tashi, S.J.A. Kadir, H.M. Rais, S. Mirjalili, H. Alhussian, Binary optimization using hybrid grey wolf optimization for feature selection, Ieee Access 7 (2019) 39496–39508.

[49] N. Khodadadi, E. Khodadadii, Q. Al-Tashi, E.-S.M. El-Kenawy, L. Abualigah, S.J. Abdulkadir, A. Alqushaibi, S. Mirjalili, BAOA: binary arithmetic optimization algorithm with K-nearest neighbor classifier for feature selection, IEEE Access (2023).

[50] M. Abd Elaziz, S. Ouadfel, R.A. Ibrahim, Boosting capuchin search with stochastic learning strategy for feature selection, Neural Comput. Appl. 35 (19) (2023) 14061–14080.

[51] M.M. Mafarja, S. Mirjalili, Hybrid whale optimization algorithm with simulated annealing for feature selection, Neurocomputing 260 (2017) 302–312.

[52] M. Mafarja, D. Eleyan, S. Abdullah, S. Mirjalili, S-shaped vs. V-shaped transfer functions for ant lion optimization algorithm in feature selection problem, in: Proceedings of the International Conference on Future Networks and Distributed Systems, 2017, pp. 1–7.

[53] R. Guha, K.K. Ghosh, S.K. Bera, R. Sarkar, S. Mirjalili, Discrete equilibrium optimizer combined with simulated annealing for feature selection, J. Comput. Sci. (2023) 101942.

[54] P. Agrawal, T. Ganesh, D. Oliva, A.W. Mohamed, S-shaped and v-shaped gaining-sharing knowledge-based algorithm for feature selection, Appl. Intell. (2022) 1–32.

[55] B. Samieiyan, P. MohammadiNasab, M.A. Mollaei, F. Hajizadeh, M. Kangavari, Solving dimension reduction problems for classification using promoted crow search algorithm (PCSA), Computing 104 (6) (2022) 1255–1284.

[56] R. Ramaswamy, P. Kandhasamy, S. Palaniswamy, Feature selection for Alzheimer's gene expression data using modified binary particle swarm optimization, IETE J. Res. 69 (1) (2023) 9–20.

[57] H. Pan, S. Chen, H. Xiong, A high-dimensional feature selection method based on modified Gray Wolf Optimization, Appl. Soft Comput. 135 (2023) 110031.