Wiley
Journal of Electrical and Computer Engineering
Volume 2024, Article ID 7112770, 16 pages
https://doi.org/10.1155/2024/7112770

WILEY

Research Article

New Hybrid Feature Selection Approaches Based on ANN and Novel Sparsity Norm

Khadijeh Nemati,[1] Amir Hosein Refahi Sheikhani,[1] Sohrab Kordrostami,[1] and Kamrad Khoshhal Roudposhti[2]

[1]*Department of Applied Mathematics and Computer Science, Lahijan Branch, Islamic Azad University, Lahijan, Iran*
[2]*Department of Computer Engineering, Lahijan Branch, Islamic Azad University, Lahijan, Iran*

Correspondence should be addressed to Amir Hosein Refahi Sheikhani; ah_refahi@yahoo.com

Received 12 March 2024; Revised 30 July 2024; Accepted 7 October 2024

Academic Editor: Neng Ye

Feature selection is crucial for minimizing redundancy in information and addressing the limitations of traditional classification methods when dealing with large datasets and numerous features in many machine learning applications. To improve the classification, this article introduced two hybrid methods utilizing a genetic algorithm and a gray wolf algorithm with structured dispersion norms for feature selection. These techniques involved the utilization of a genetic algorithm and a gray wolf algorithm for feature selection. The features selected by these algorithms were used in the classification process by employing a two-layer perceptron as a classifier. The novel sparse norm is employed to assess and compute classification errors in these methodologies. To assess the effectiveness of the suggested techniques, they were compared with the existing feature selection methods using various publicly accessible datasets. The results of the experiments consistently demonstrate that the proposed methods outperform other approaches.

Keywords: classification; dimensionality reduction; feature selection; genetic algorithm; gray wolf optimizer

1. Introduction

Feature selection holds significance in numerous machine learning applications as it aids in extracting meaningful features while eliminating noisy or irrelevant ones. It is a crucial component of the classification process. Reference [1] presented various feature selection methods that can be classified into three categories: filter methods, wrapper methods, and embedded methods. Filter methods rank features based on their individual scores, independent of any specific model. Features with the highest scores (top M features) or those surpassing a defined threshold value τ are selected. Wrapper methods, on the other hand, evaluate subsets of features by measuring the performance of the resulting models. These methods encompass techniques such as greedy sequential searches as well as evolutionary and swarm intelligence algorithms for feature selection. Embedded methods, the third category, require the utilization of a predictive model. Examples of such methods include least absolute shrinkage and selection operator (LASSO) regression, where feature selection is integrated into the model-building process. Reference [2] provided an overview of the current feature selection criteria used to evaluate the ability of features to preserve sample similarity. These algorithms, including ReliefF, SPEC, Fisher score, trace ratio criterion, Hilbert–Schmidt independence criterion (HSIC), similarity preserving feature selection (SPFS), SPFS-NES, robust feature selection (RFS), and minimum redundancy and maximum relevancy (mRMR), supervised and unsupervised infinite feature selection (Inf-FS), regularization techniques, and LASSO (hinged) and LASSO (unhinged) are extensively reviewed. Reference [3] described a principal component analysis (PCA) method based on the nested L_2-norm and $L_{2,p}$-norm to handle high-dimensional

data with outliers. This method employs $L_{2,p}$-norm to retain the desirable properties of PCA such as rotational invariance. The empirical results on synthetic datasets, UCI datasets, and face recognition databases validate the effectiveness of the method in processing high-dimensional data. Reference [4] proposed a variable selection method based on a fast nondominated-ranking genetic algorithm (NSGA-II) for qualitative discrimination of near-infrared (NIR) spectra. The method had two objective functions: maximizing the sum of ratios of interclass variance to intraclass variance and minimizing the sum of correlation coefficients between the selected variables. The results showed that NSGA-II could discriminate different parts of the tobacco leaves well. Reference [5] described a deep neural network and genetic algorithm (GA)–based feature selection approach for an automated diabetic retinopathy (DR) identification approach. DR is a type of eye disease that results in vision loss. Early identification of DR has the potential to prevent or delay vision loss. Convolutional neural network architectures are used to extract features, followed by the GA for feature selection and ranking features into high rank (optimal) and lower rank (unsatisfactory). Reference [6] described the efficient classification model for coronavirus protein sequences using machine learning algorithms and feature selection techniques to aid in the early detection and prediction of novel viruses. To optimize performance, we employed machine learning classification algorithms such as K-nearest neighbor (KNN) along with feature selection techniques such as GA, LASSO, and support vector machine recursive feature elimination (SVM-RFE). Reference [7] described an intelligent optimization GA to improve the classification accuracy and decrease time complexity and weight distribution based on information entropy. Information entropy of features was defined as the population labels in GA rather. Reference [8] proposed an enhanced brain MRI image classifier that has two main objectives, to achieve maximum classification accuracy and to minimize the number of features for classification. Feature selection is performed using a GA, and the random forest is used as a classifier. Reference [9] described a GA-based feature selection technique. The technique developed herein involved the use of a novel fitness function to select a combinatorial set of features. For benchmarking, the features were selected by WEKA and GA.GA-based features performed better in most cases than WEKA-based features. The GA-based features outperformed WEKA-based features in more instances. In Ref. [10], improved gray wolf optimization (IGWO) is introduced for medical diagnosis tasks. This method combines GA and GWO to enhance performance. GA is used to create initial positions in the search space, followed by GWO to update population positions in the discrete search space for optimal feature selection in medical diagnosis. This integrated approach improves the efficiency and effectiveness of feature selection in medical diagnosis applications. Reference [11] introduced a hybrid feature selection model for intrusion detection, combining GWO and particle swarm optimization algorithms. By integrating GWO and PSO, the model aims to identify key features for accurate intrusion detection, enhancing the efficiency and effectiveness of the selection process and improving system performance. Reference [12] suggested merging the popular metaheuristic population-based optimizer, the gray wolf algorithm, with the gradient descent algorithm for use in feature selection problems. The proposed hybrid gradient descent GWO motivates the weakest wolves to track down the prey following the guidance of the pack leaders. The directional information is sourced from the partial derivative of the leading wolves. Reference [13] suggested a methodology for classifying biomedical data using the feature selection technique and soft computing-based optimization algorithms. The evaluation focuses on categorizing benign and malignant tumors using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset and measures proficiency using various metrics. The proposed clinical decision support system demonstrates a highly favorable classification performance outcome, reducing the burden on expert medical practitioners.

The rest of this paper is organized as follows. Section 2 presents the literature review. In Section 3, methods and materials are illustrated. In Section 4, we illustrate the proposed method artificial neural network with GA and structured sparsity norm (GA–ANN–SSN). In Section 5, we illustrate the proposed method artificial neural network with GWO algorithm and SNN (GWO–ANN–SSN). In Section 6, experimental results are presented as computer simulation results. Finally, in Section 7, we illustrate conclusions.

2. Literature Review

Utilizing different metaheuristic methods to adjust algorithms for better results has displayed potential across several domains. The utilization of these techniques has been advantageous in enhancing machine learning algorithms via blending and acquiring heuristics. Reference [14] suggested an AI-driven computer-aided diagnosis system designed to enhance the classification of retinal images as healthy or sick using advanced machine learning features. It utilizes innovative feature selection algorithms such as emperor penguin optimization (EPO) and bacterial foraging optimization to optimize performance. The method achieves high accuracy and provides valuable support to busy medical professionals in preserving vision. Reference [15] suggested the utilization of the gravitational search optimization algorithm (GSA), EPO, and a hybrid approach that combines GSA and EPO, referred to as HGSAEPO, in order to effectively pinpoint important features while concurrently minimizing the presence of irrelevant ones. Through the application of these soft computing techniques and six machine learning classifiers, a robust framework for prognostic research is developed by categorizing data instances from the WDBC dataset. The study successfully accomplishes its goals by introducing a reliable clinical prediction system designed to assist healthcare professionals in achieving efficient diagnoses. Reference [16] suggested a method to explore the use of ant lion-based optimization for selecting a subset of features. The selected attributes were utilized to train and assess four machine learning classifiers. The research utilized three public benchmark datasets and

one custom dataset, each tailored to a specific disease. The efficacy of the proposed approach was gauged using five performance evaluation metrics, resulting in a notable enhancement in results. In Ref. [12], a population-based optimizer known as the gray wolf algorithm, along with the gradient descent algorithm, was suggested and evaluated for use in feature selection problems. The proposed optimizer was tested on six medical datasets from the UCI machine learning repository, demonstrating potential in effectively balancing the objectives of feature selection. Further improvements could be made to enhance its performance. In Ref. [17], an enhanced GWO algorithm, known as the DGWO, was introduced. This version incorporates a dynamically dimensioned search, spiral walking predation technique, and positional interaction information. Moreover, a nonlinear control parameter approach is implemented to optimize the balance between exploration and exploitation within the GWO algorithm. The DGWO outperforms other algorithms in terms of solution accuracy, robustness, and convergence speed across various scenarios. Reference [18] proposed a fitness-based GWO clustering method for identifying spam reviews. The technique employs fitness-based GWO to find optimal cluster heads. The algorithm updates gray wolves' positions in two phases: first using GWO and then adjusting positions based on fitness evaluation. The effectiveness of this approach is validated using three spam datasets: synthetic spam reviews, movie reviews, and Yelp reviews. Comparative analysis with leading metaheuristic clustering methods demonstrates that the FGWOK algorithm outperforms the existing methodologies, supported by experimental and statistical findings. Reference [19] proposed an IGWO-based feature selection model for intrusion detection systems. The proposed model consists of three primary processes: preprocessing, feature selection, and classification, followed by result evaluation. The model employs IGWOs to select a subset of input variables that minimize features to measure accuracy in the search space and obtain the best solution. The suggested hybrid IDSs, named IGWO-EMS-DHPN, were evaluated using two intrusion datasets, UNSW-NB15 and CICIDS-2017, and compared with other classifiers in terms of accuracy, precision, recall, and F1-score. The results indicate that the IGWO-EMS-DHPN classifier outperforms other classifiers, achieving maximum accuracy. Reference [20] introduced a binary iteration of the political optimizer (PO) for addressing feature selection challenges with gene expression data. Two transformation functions were employed in creating the binary PO. The initial function is derived from the sigmoid function and labeled BPO-S, whereas the second one is derived from the V-shaped function and labeled BPO-V. The effectiveness of these approaches was assessed across nine biological datasets and juxtaposed against eight established binary metaheuristic algorithms. Reference [21] introduced a mathematical framework for creating a cumulative index using the concentrations of four key pollutants SO_2, NO_2, PM2.5, and PM10 as individual factors. Additionally, they proposed a supervised learning algorithm-based classifier that utilizes SVM to categorize air quality as either good or harmful. The classifier takes the calculated CI values as inputs and is evaluated using real data from Kolkata, Delhi, and Bhopal. The results show that the classifier effectively distinguishes between different air quality levels. Reference [22] introduced a dynamic mathematical model of a simplified steam condenser using the lumped parameter modeling method. Subsequently, a pressure PI control system for the steam condenser is designed on the MATLAB simulation platform. To enhance performance, the GWO, a new metaheuristic intelligent algorithm, is employed to fine-tune the PI controller parameters. The results from simulations demonstrate that the GWO algorithm outperforms four other algorithms in terms of control performance. Reference [23] proposed the GWO, a swarm-based optimization method, to explore the feature space and identify the optimal subset of features that enhance classification accuracy. Initially, the GWO utilizes filter-based principles to identify solutions with minimal redundancy, as indicated by mutual information. Subsequently, an optimization wrapper approach is implemented to enhance classifier performance. The effectiveness of the GWO is evaluated and compared to various other metaheuristic algorithms using the NSL KDD dataset. Reference [24] proposed a modified GWO algorithm for feature selection in high-dimensional data. The algorithm incorporates the ReliefF algorithm and copula entropy during initialization to enhance the quality of the initial population. Furthermore, the modified GWO introduces two novel search strategies: a competitive guidance strategy for updating individual positions, enhancing search flexibility, and a differential evolution-based leader wolf enhancement strategy to locate and replace the leader wolf's position, thereby preventing the algorithm from getting trapped in local optima. The results on 10 high-dimensional small-sample gene expression datasets indicate that the proposed algorithm selects less than 0.67% of features, enhances classification accuracy while reducing the number of features, and achieves competitive outcomes compared to advanced feature selection methods.

Feature selection aims to reduce the number of input features in a classifier while maintaining predictive accuracy. Various existing methods utilize different approaches to select feature subsets, such as an exhaustive search evaluating all combinations. While these methods ensure optimal solutions, finding the best feature subset is an NP-hard problem. Therefore, treating feature selection as an optimization issue, we turn to metaheuristic methods for solutions. This problem involves searching for optimal or near-optimal feature subsets within the space of possibilities.

This paper introduced two hybrid methods combining a GA and a gray wolf algorithm with SSNs for feature selection to enhance classification accuracy. These techniques involve the utilization of a GA and a gray wolf algorithm for feature selection. The features selected by these algorithms are employed in the classification process using a two-layer perception. A novel sparse norm is used to evaluate and calculate classification errors in these approaches. We show that our proposed methods are capable of finding the best feature subset, which leads to obtaining higher accuracy and lower cost for each input dataset.

3. Methods and Materials

In this section, the concepts required for the feature selection problem and our proposed method and datasets will be described.

3.1. Feature Selection. In classification problems involving large datasets and numerous features, the primary focus is on achieving swift, precise, and dependable classification results. Feature selection plays a crucial role by replacing the full feature set with a smaller subset containing the most pertinent features according to predefined criteria. This technique aids in reducing the number of features and removing redundant ones, while preserving essential ones, thereby enhancing the classification model's ability to learn and generalize effectively.

3.2. GA. The GA utilizes binary coding and implements iterative optimization by mimicking the natural selection principle of survival of the fittest. This algorithm operates on a population composed of numerous individuals, each possessing unique genes. Through the processes of crossover and mutation, new offspring are generated to form the next generation of the population. Each individual represents a potential solution to the problem. Gene crossover and mutation represent alterations to the solution. Genes that are closer to the optimal solution receive higher scores under the same fitness function. During each iteration, individuals with lower scores are eliminated. By continuously applying crossover and mutation operations and setting an appropriate number of iterations, it is possible to reach a globally optimal solution.

3.3. GWO Algorithm. The GWO algorithm is inspired by the hierarchical hunting behavior of gray wolves. It mimics the roles of alpha, beta, delta, and omega wolves within a pack. Alpha wolves lead the pack and make critical decisions, while beta wolves support them. Delta wolves, older and experienced, care for pups. Omega wolves, the lowest rank, do not partake in decisions. GWO aims to find optimal solutions by replicating wolves' collaborative hunting. Wolves estimate the prey position together, and their position adjusts based on alpha, beta, and delta movements. Each wolf symbolizes a solution or a feature subset. The leader, assumed to have the best subset, guides the pack toward the optimal solution. If a member finds a superior subset, they become the leader in the next iteration. Wolves continuously adjust their positions in the hunt for the optimal feature subset.

3.4. ANN. For the classification problem, a two-layer perceptron ANN is employed. The network comprises an input layer, a hidden layer utilizing the hyperbolic tangent activation function, and an output layer employing a pure linear function. The dataset is split into 80% training data and 20% testing data to assess network performance. During training, network parameters, including feature-associated weights, are adjusted for optimization.

3.5. SSN. In this section, the SSN introduces specifically the $l_{r,p}$-norm as the loss function. For matrix $\Gamma = \{\gamma_{ij}\}$, i-th row and j-th column are denoted by γ^i and γ_j, respectively. The $L_{r,p}$-norm of the matrix $\Gamma \in \mathbb{R}^{n\times m}$ and $p > 0$ is defined as

$$\|\Gamma\|_{r,p} = \left(\sum_{i=1}^{n}\left(\sum_{j=1}^{m}\left|\gamma_{ij}\right|^r\right)^{(p/r)}\right)^{(1/p)} = \left(\sum_{i=1}^{n}\left\|\gamma^i\right\|_r^p\right)^{(1/p)}. \quad (1)$$

3.6. Dataset. In this paper, the datasets used for the experiment are public datasets. In this experiment, six datasets with different features are chosen to test. The characteristics of the datasets are shown in Table 1 [25].

4. Proposed Method 1: GA–ANN–SSN

In this section, we present an embedded feature selection approach utilizing GA–ANN–SSN to identify the optimal set of features for classifying samples into different categories. We will explain the proposed method as follows.

4.1. Initialization Parameters. In this section, we present the optimal feature selection approach called the GA–ANN–SSN method, which is based on a GA and SSN. In the feature subset selection problem, each chromosome of the population is a feature subset. Chromosomes are considered strings of 0 and 1. In each chromosome, 1 and 0 mean that the feature is selected and the feature is not selected, respectively. The string length is equal to the number of features. The initial population is considered random. We should set the GA parameters. These parameters are the size of the population, the number of generations, the rate of crossover, and the rate of mutation. To adjust these parameters, different values of the parameters were tested on the dataset. In this study, we consider $\text{population}_{\text{Size}} = 20$, $\text{Max}_{\text{Generation}} = 100$, $\text{Cross}_{\text{rate}} = 0.8$, and $\text{Mut}_{\text{rate}} = 0.2$.

4.2. Feature Selection. In this section, we outline the GA-based feature selection process. The population is chosen using a roulette wheel selection method, where chromosomes are selected based on their fitness values. Chromosomes with higher fitness have a greater chance of being chosen for the next generation. The probability selection of chromosomes is calculated as follows:

$$p_i = \frac{f_i}{\sum_{j=1}^{n} f_j}, \quad (2)$$

where p_i is the probability of selected chromosome i and f_i is the fitness value of chromosome i.

4.3. Genetic Operation. Genetic operators are employed on the selected chromosomes to generate new chromosomes in the subsequent generation. Mutation and crossover are the primary operators utilized in GAs. Crossover involves the manipulation of two parent chromosomes to produce two offspring by exchanging gene segments between the parents.

Table 1: Summary of dataset descriptions.

Dataset	No. of features	No. of classes	No. of samples	Type of dataset
RELATHE	4322	2	1427	Text data
PCMAC	3289	2	1943	Text data
MADELON	500	2	2000	Random data
GISETTE	5000	2	7000	Handwritten digits
ALLAML	7129	2	72	Gene expression
PROSTATE-GE	5966	2	102	Gene expression

A single-point crossover method is utilized in this scenario, where genes are randomly cut at a single point on both chromosomes. The resulting chromosomes from the crossover process then undergo mutation, where a gene is randomly altered to its opposite value within a chromosome. The selection of genes for mutation is done randomly. The subsets created are utilized in the neural network for sample classification.

4.4. Classification Error Rule. Subsets are selected using the GA and training data to train the perceptron neural network and adjust its parameters. During training, each feature is assigned unique weights, initially set randomly, which are updated in subsequent iterations. Features are ranked based on their weights, with higher-weighted features retained for classification. To determine the classification error, we establish a loss function utilizing the SSN. The network's performance is evaluated using test data, and the classification error is determined. The SSNs are chosen as the loss function for their simplicity and efficient optimization. This method aims to minimize the loss function, improving the classification process by deriving optimal weights for each feature.

The loss function is as follows:

$$E = \frac{1}{2}\|\Theta - T\|_2^2. \tag{3}$$

To control variance and prevent overfitting, one or more regularization terms are added to equation (3). Here, we assumed two regularization terms $f_i(W)$ with the regularization parameters as λ_i for the weight of W. Assume $f_1(W) = \|W\|_{1,1}$ and $f_2(W) = \|W\|_{2,1}$ is the $L_{r,p}$-norm regularization term. The higher weight features in the W matrix will be more important for classifying data into distinct classes. We used the SSN ($L_{2,1}$-norm and $L_{1,1}$-norm) to calculate the feature coefficients,

$$E = \frac{1}{2}(\Gamma W - T)^t(\Gamma W - T) + \lambda_1\|W\|_{1,1} + \lambda_2\|W\|_{2,1}. \tag{4}$$

Since $\Theta = \Gamma W$, $\|W\|_{1,1} = (W^tW/|W|)$, $\|W\|_{2,1} = (W^tW/\|W\|_2)$, equation (4) will be rewritten as follows:

$$E = \frac{1}{2}\left(W^t\Gamma^t\Gamma W - W^t\Gamma^tT + T^tT\right) + \lambda_1\frac{W^tW}{|W|} + \lambda_2\frac{W^tW}{\|W\|_2}. \tag{5}$$

The gradient method has been used to calculate W, and finally, the dynamic equation system is obtained as follows:

$$\frac{dW}{dt} = \mu\frac{\partial E}{\partial W^t} = \mu\left(\Gamma^t\Gamma W - \Gamma^tT + \lambda_1\frac{W}{|W|} + \lambda_2\frac{W}{\|W\|_2}\right), \tag{6}$$

$$W = \frac{\Gamma^tT}{\Gamma^t\Gamma + (\lambda_1/|W|)I + (\lambda_2/\|W\|_2)I}. \tag{7}$$

Here, $\mu > 0$ is the learning rate of the neural network, $\lambda_i > 0$ is the regularization parameters, and I is an identical matrix. The symbols used in this study are given in Table 2.

4.5. Termination Criteria. This section presents the termination criteria for the algorithm. The termination conditions are defined as follows:

- If the maximum number of generations is reached.
- If the weights of features do not improve after 30 iterations.
- If after reorganization, the difference between the best individual (solution) and the worst individual (solution) is smaller than a predetermined value ε.

Once any of the termination conditions are met, the algorithm concludes. At this point, the optimal feature subset is returned, and the classification accuracy is determined. The termination criteria serve as stopping rules for the algorithm, ensuring that it converges to a satisfactory solution and avoids unnecessary computations.

5. Proposed Method 2: GWO–ANN–SSN

In this section, we present an embedded feature selection approach utilizing GWO–ANN–SSN to identify the optimal set of features for classifying samples into different categories. We will explain the proposed method as follows.

5.1. Initialization Parameter. In this section, we introduce the optimal feature selection method called the GWO–ANN–SSN technique, which utilizes a GWO algorithm and SSN. The parameters to be considered are the population size and the number of generations. To optimize these parameters, various values were experimented with on the dataset. For this research, we have set the population size to 20 and the maximum number of generations to 100. Each individual in the population is represented by a feature vector as a solution, and the initial population is created randomly.

Table 2: The symbols used in the paper.

Symbol	Definition
$\|W\|_{1,1}$	$L_{1,1}$-norm
$\|W\|_{2,1}$	$L_{2,1}$-norm
E	Classification error
Θ	Output of the neural network
T	The class of an observation
W	Weight vector
Γ	Output of the hidden layer
μ	Learning rate of neural network
λ_i	Regularization parameters
I	Identical matrix

5.2. Feature Selection. This segment will delineate the process of selecting a feature using the GWO's search and prey mechanism. The search commences by generating a random population of gray wolves. In each iteration, the alpha, beta, and delta wolves approximate the potential location of the prey (optimal solution). Each wolf, symbolizing a solution, adjusts its position or distance by considering the positions of other wolves in the population. This adjustment is guided by the notion that the optimal solution lies in proximity to where the alpha, beta, and delta wolves are positioned. Wolves iteratively fine-tune their positions to converge toward the optimal solution, gradually encircling it through the collective movement of the pack. The fitness value is computed for each individual, and individuals are ranked from the lowest to the highest fitness value. The top three solutions with the least fitness values are identified as the new alpha, beta, and delta positions. Given our lack of precise knowledge regarding the prey's exact location (optimal solution) within the search space, we consider the alpha position, representing the best position achieved thus far, as an approximation of the prey's location. During the hunting phase, the top three solutions obtained are preserved, while the remaining search agents update their positions based on these preserved solutions.

5.3. Classification Error Rule. To calculate the classification error, we use the relations of Section 3.4.

5.4. Termination Criteria. To terminate the execution of the algorithm, we use the termination criteria in Section 3.5.

6. Experiment Result

6.1. Limitations and Assumptions. Similar to any other approach, the suggested methods also come with limitations, including the following:

- This algorithm may not be appropriate for all scenarios, particularly those that are straightforward and already have available derived data.
- The iterative calculation of the objective function's value could be computationally costly for certain problems.
- If not implemented correctly, the randomness involved may prevent it from converging to the best solution

Regularization parameters are used in many machine learning and statistical models to prevent overfitting. Overfitting occurs when a model performs well on training data but fails to generalize to new data. Choosing the appropriate regularization parameters can be complex because a balance must be found between preventing overfitting and maintaining the model's power. However, in this study, the researchers decided to use specific values listed in Table 3 instead of searching for the optimal parameters. This table includes values that have proven useful in previous experiments or related studies. Finding suitable regularization parameters can be challenging, but the researchers in this study have overcome this challenge by using predefined values. Various values of the tuning parameters were tested for different datasets as shown in Tables 4, 5, 6, 7, 8, 9, and the values that resulted in the least cost or error over 20 executions were selected, as presented in Table 3.

To manage variance and avoid overfitting, two regularization parameters $\lambda_i > 0$ are incorporated into equations (5) and (7). Specifically, these parameters are used to compute the classification error in equation (5) and the feature coefficients in equation (7). The values of $\lambda_i > 0$ are between zero and one, and the sum of them is equal to one. These values were randomly determined during 20 running attempts with each of our proposed methods, and finally, the values that led to the highest classification accuracy for each dataset are listed in Table 3. In Table 3, last column, the number of selected features with the highest score is observed. For example, for the GISETTE dataset with values of $\lambda_1 = 0.15$ and $\lambda_2 = 0.85$ and 200 features, the accuracy of the proposed method is calculated. For the MADELON dataset with values of $\lambda_1 = 0.25$ and $\lambda_2 = 0.75$ and 200 features, the accuracy of the proposed method has been calculated. For ALLAML and PROSTATE-GE datasets with values of $\lambda_1 = 0.15$ and $\lambda_2 = 0.85$ and 20 and 80 features, the accuracy of the proposed method has been calculated. The classification accuracy results are obtained in Tables 10 and 11 by using the top 200 features, in Table 12, the accuracy is obtained by the top 20 features, and in Table 13, the accuracy is obtained by the top 80 features on each dataset. In these tables, we can observe that the proposed methods show better performance compared to other algorithms and also the GWO–ANN–SSN method achieved higher accuracy than other methods. For example, the results in Table 7 show that the proposed GWO–ANN–SSN achieves an accuracy higher than 75.0, with an average value of 79.0001 on the RELATHE dataset. This method achieves an accuracy higher than 75.0, with an averaged value of 78.5501 on the PCMAC dataset.

6.2. Time Complexity of Algorithms. The time complexity of the proposed GWO–ANN–SSN method is $O(n\text{Sample} \times n\text{Pop} \times \text{MaxIt})$, and the time complexity of the suggested GA–ANN–SSN method is $O(n\text{Sample} \times n\text{Pop} \times \text{MaxIt})$, where "$n$Sample" represents the quantity of samples, "MaxIt" signifies the number of iterations of the algorithm, and "nPop" denotes the population size in a generation.

Table 14 provides the minimum loss function value, representing the classification error and the actual execution duration, for every dataset.

Table 3: Regularization parameters of our proposed methods.

Dataset	λ_1	λ_2	Number of top features selected
RELATHE	0.4	0.6	200
PCMAC	0.15	0.85	200
MADELON	0.25	0.75	200
GISETTE	0.15	0.85	200
ALLAML	0.15	0.85	20, 80
PROSTATE-GE	0.15	0.85	20, 80

Table 4: Result of our proposed methods with the RELATHE dataset.

Number of runs	λ_1	λ_2	Run best cost
1	0.48	0.52	170.24
2	0.15	0.85	190.07
3	0.25	0.75	174.24
4	0.20	0.80	198.64
5	0.35	0.65	192.92
6	0.10	0.90	196.03
7	0.30	0.70	185.24
8	0.88	0.12	188.07
9	0.77	0.23	174.24
10	0.80	0.20	178.64
11	0.65	0.35	192.92
12	0.93	0.07	196.03
13	0.95	0.05	179.36
14	**0.4**	**0.6**	**165.24**
15	0.52	0.48	170.24
16	0.85	0.15	183.07
17	0.75	0.25	174.24
18	0.83	0.17	181.64
19	0.69	0.31	193.92
20	0.90	0.10	194.03

Note: The bold values indicate the lowest (or best) cost values achieved across 20 experimental runs with different parameter settings λ_1 and λ_2. These values highlight the optimal performance of our proposed method on the RELATHE dataset under specific parameter configurations, guiding the selection of the most effective parameters for subsequent experiments.

6.3. Results. In this section, the datasets used for the required parameters, their values, and the results of the proposed methods are described. To reduce the random deviation of the program, each of the algorithms is repeated 20 times and the average of the obtained answers is included in the tables. In this experiment, performance measurement models can be described by determining accuracy. Accuracy is a measure that describes how precise the model is in classifying data. Given that the majority of machine learning classification is supervised learning, precision can be described as the ratio of correctly classified test samples according to class labels. These algorithms include ReliefF, Fisher score, trace ratio criterion, HSIC, RFS, mRMR, SVM, regularization techniques, and LASSO hinged and unhinged. All algorithms are implemented in MATLAB.

The feature selection methods used in this study are given in Table 15.

Based on the accuracy obtained in each dataset, our proposed methods achieve higher accuracy on all types of data (text, random, handwritten digits, and gene expression). The results of the accuracy show that our GWO–ANN–SSN selects the superior features, thus leading to higher accuracy and lower classification costs. To evaluate the features, equations (5) and (7) have been

Table 5: Result of our proposed methods with the PCMAC dataset.

Number of runs	λ_1	λ_2	Run best cost
1	0.48	0.52	190.24
2	0.63	0.37	185.07
3	0.25	0.75	166.24
4	0.20	0.80	189.64
5	0.35	0.65	172.92
6	0.10	0.90	166.03
7	0.30	0.70	175.24
8	0.88	0.12	178.07
9	0.77	0.23	194.24
10	0.80	0.20	187.64
11	0.65	0.35	181.92
12	0.93	0.07	189.03
13	0.95	0.05	169.36
14	0.4	0.6	167.24
15	0.52	0.48	177.24
16	0.85	0.15	193.07
17	0.75	0.25	184.24
18	0.83	0.17	191.64
19	**0.15**	**0.85**	**163.07**
20	0.90	0.10	198.03

Note: The bold values indicate the lowest (or best) cost values achieved across 20 experimental runs with different parameter settings λ_1 and λ_2. These values highlight the optimal performance of our proposed method on the PCMAC dataset under specific parameter configurations, guiding the selection of the most effective parameters for subsequent experiments.

Table 6: Result of our proposed methods with the MADELON dataset.

Number of runs	λ_1	λ_2	Run best cost
1	0.48	0.52	178.24
2	0.63	0.37	182.07
3	0.15	0.85	191.24
4	**0.25**	**0.75**	**174.24**
5	0.35	0.65	182.92
6	0.10	0.90	186.03
7	0.30	0.70	175.24
8	0.88	0.12	178.07
9	0.77	0.23	194.24
10	0.80	0.20	188.64
11	0.65	0.35	195.92
12	0.93	0.07	191.03
13	0.95	0.05	189.36
14	0.40	0.60	195.24
15	0.52	0.48	180.24
16	0.85	0.15	181.07
17	0.75	0.25	179.24
18	0.83	0.17	188.64
19	0.30	0.70	193.07
20	0.90	0.10	174.03

Note: The bold values indicate the lowest (or best) cost values achieved across 20 experimental runs with different parameter settings λ_1 and λ_2. These values highlight the optimal performance of our proposed method on the MADELON dataset under specific parameter configurations, guiding the selection of the most effective parameters for subsequent experiments.

Table 7: Result of our proposed methods with the GISETTE dataset.

Number of runs	λ_1	λ_2	Run best cost
1	0.48	0.52	190.24
2	0.63	0.37	180.07
3	0.25	0.75	164.24
4	0.20	0.80	178.64
5	0.35	0.65	196.92
6	0.10	0.90	162.03
7	0.30	0.70	165.24
8	0.88	0.12	168.07
9	**0.15**	**0.85**	**158.64**
10	0.80	0.20	188.64
11	0.65	0.35	159.92
12	0.93	0.07	196.03
13	0.95	0.05	181.36
14	0.4	0.6	175.24
15	0.52	0.48	160.24
16	0.85	0.15	181.07
17	0.75	0.25	164.24
18	0.83	0.17	171.64
19	0.20	0.80	169.07
20	0.90	0.10	184.03

Note: The bold values indicate the lowest (or best) cost values achieved across 20 experimental runs with different parameter settings λ_1 and λ_2. These values highlight the optimal performance of our proposed method on the GISETTE dataset under specific parameter configurations, guiding the selection of the most effective parameters for subsequent experiments.

Table 8: Result of our proposed methods with the ALLAML dataset.

Number of runs	λ_1	λ_2	Run best cost
1	0.48	0.52	195.24
2	0.63	0.37	196.07
3	0.25	0.75	204.24
4	0.20	0.80	218.64
5	0.35	0.65	199.92
6	0.10	0.90	209.03
7	0.30	0.70	215.24
8	0.88	0.12	228.07
9	0.77	0.23	194.24
10	0.80	0.20	188.64
11	**0.15**	**0.85**	**192.92**
12	0.93	0.07	222.03
13	0.95	0.05	230.36
14	0.4	0.6	202.24
15	0.52	0.48	196.24
16	0.85	0.15	193.07
17	0.75	0.25	224.24
18	0.83	0.17	216.64
19	0.65	0.35	210.07
20	0.90	0.10	201.03

Note: The bold values indicate the lowest (or best) cost values achieved across 20 experimental runs with different parameter settings λ_1 and λ_2. These values highlight the optimal performance of our proposed method on the ALLAML dataset under specific parameter configurations, guiding the selection of the most effective parameters for subsequent experiments.

used as a loss function and weight adjustment of the features, respectively. One advantage of the proposed method is that it performs modeling during the learning

Table 9: Result of our proposed methods with the PROSTATE-GE dataset.

Number of runs	λ_1	λ_2	Run best cost
1	0.48	0.52	234.24
2	0.63	0.37	221.07
3	0.25	0.75	214.24
4	0.20	0.80	218.64
5	0.35	0.65	226.92
6	0.10	0.90	209.03
7	0.30	0.70	195.24
8	0.88	0.12	230.07
9	0.77	0.23	229.24
10	0.80	0.20	208.64
11	0.65	0.35	202.92
12	0.93	0.07	217.03
13	0.95	0.05	196.36
14	0.4	0.6	199.24
15	0.52	0.48	230.24
16	0.85	0.15	223.07
17	0.75	0.25	214.24
18	0.83	0.17	202.64
19	**0.15**	**0.85**	**194.03**
20	0.90	0.10	199.03

Note: The bold values indicate the lowest (or best) cost values achieved across 20 experimental runs with different parameter settings λ_1 and λ_2. These values highlight the optimal performance of our proposed method on the PROSTATE-GE dataset under specific parameter configurations, guiding the selection of the most effective parameters for subsequent experiments.

Table 10: Comparison of the accuracy for RELATHE and PCMAC datasets.

Feature selection method	Name of dataset	
	RELATHE	PCMAC
Fisher score	73.0	75.0
Trace ratio	73.0	75.0
HSIC	73.0	75.0
ReliefF	68.0	70.0
mRMR	75.0	75.0
SVM	73.0	74.0
ANN–SSN	75.85	76.62
GA–ANN–SSN	**76.76**	**77.31**
GWO–ANN–SSN	**79.0001**	**78.5501**

Note: The bold values represent the highest accuracy achieved for each of the two datasets, indicating the superior performance of the proposed methods compared to others. These bold values emphasize the effectiveness of the proposed approach in achieving optimal accuracy across different datasets.

operation, and it is considered one of the embedded algorithms. We define a simple robust objective function using the $L_{2,1}$_norm, which is expressed in equation (5). This method can select distinct features by optimizing the loss function. In the proposed method, it is difficult to find suitable regularization parameters. However, we used the values in Table 3.

The summary of GA–ANN–SSN is shown in Figure 1.

The summary of GWO–ANN–SSN is shown in Figure 2.

In the ROC plot of the proposed GWO–ANN–SSN, a higher value of X indicates a higher number of false positives than true negatives. In addition, a higher value of Y indicates a higher number of true-positive detections

Table 11: Comparison of the accuracy for GISETTE and MADELON datasets.

Feature selection method	Name of dataset	
	GISETTE	MADELON
LASSO unhinged	95.3	55.9
LASSO hinged	85.9	54.3
ANN–SSN	97.78	59.4
GA–ANN–SSN	**97.99**	**60.11**
GWO–ANN–SSN	**98.5001**	**61.8001**

Note: The bold values represent the highest accuracy achieved for each of the two datasets, indicating the superior performance of the proposed methods compared to others. These bold values emphasize the effectiveness of the proposed approach in achieving optimal accuracy across different datasets.

Table 12: Comparison of accuracy with 20 features [2].

Feature selection method	Name of dataset	
	ALLAML	PROSTATE-GE
Fisher score	89.11	95.09
RFS	95.89	95.09
ReliefF	90.36	92.18
mRMR	93.21	93.18
ANN–SSN	97.67	96.15
GA–ANN–SSN	**97.85**	**96.69**
GWO–ANN–SSN	**97.93**	**97.11**

Note: The bold values in Table 12 represent the highest accuracy achieved for each of the two datasets, indicating the superior performance of the proposed methods compared to others. These bolded values emphasize the effectiveness of the proposed approach in achieving optimal accuracy across different datasets.

Table 13: Comparison of accuracy with 80 features [2].

Feature selection method	Name of dataset	
	ALLAML	PROSTATE-GE
Fisher score	96.07	93.18
RFS	97.32	95.09
ReliefF	95.89	91.18
mRMR	94.46	86.36
ANN–SSN	98.07	96.15
GA–ANN–SSN	**98.53**	**97.41**
GWO–ANN–SSN	**98.55**	**98.01**

Note: The bold values in Table 13 represent the highest accuracy achieved for each of the two datasets, indicating the superior performance of the proposed methods compared to others. These values emphasize the effectiveness of the proposed approach in achieving optimal accuracy across different datasets.

than false-negative points. As observed in Figures 3, 4, 5, 6, 7, and 8, all points are located above the bisector of the plot, corresponding to a situation where the proportion of correctly classified points belonging to the positive class is greater than the proportion of incorrectly classified points belonging to the negative class. The true-positive value, true-negative value, false-positive value, and false-negative value can be observed in Table 14, which is obtained using the following equations:

Table 14: The run time and best cost for the proposed method GWO–ANN–SSN.

Dataset	Run best cost	Time (s)
RELATHE	165.24	70.489
PCMAC	163.07	93.275
GISETTE	174.24	81.101
MADELON	158.64	45.203
ALLAML	192.92	76.296
PROSTATE-GE	194.03	53.232

Table 15: The feature selection methods used in the paper.

Abbreviation name of method	Full name of method
HSIC	Hilbert–Schmidt independence criterion
mRMR	Minimum redundancy and maximum relevancy
SVM	Support vector machine
RFS	Robust feature selection
LASSO	Regularization techniques and least absolute shrinkage and selection operator
ANN	Artificial neural network
SSN	Structured sparsity norm
GA	Genetic algorithm
GWO	Gray wolf optimizer

$$\mathrm{TPR} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}, \tag{8}$$

$$\mathrm{TNR} = \frac{\text{true negative}}{\text{true negative} + \text{false positive}}, \tag{9}$$

$$\mathrm{FPR} = \frac{\text{false positive}}{\text{false positive} + \text{true negative}}, \tag{10}$$

$$\mathrm{FNR} = \frac{\text{false negative}}{\text{false negative} + \text{true positive}}. \tag{11}$$

The performance of the proposed GWO–ANN–SSN model based on correct identification rates is presented in Table 16. This table provides four key metrics: true-positive rate (TPR), true-negative rate (TNR), false-positive rate (FPR), and false-negative rate (FNR) for six datasets. As shown, the model performs exceptionally well on the GISETTE and ALLAML datasets, achieving TPR rates of 99.1% and 100%, respectively. Similarly, for the PROSTATE-GE dataset, the model achieves a TPR of 98%, with an FPR of 0%, indicating no false positives.

The confusion matrix illustrates the classification outcomes based on the available actual data. It helps in defining different criteria for category evaluation, such as accuracy, which indicates the number of correctly recognized patterns. Based on the confusion matrix results for various datasets, the effectiveness of the proposed GWO–ANN–SSN method can be observed. For example, in the PROSTATE-GE dataset, consisting of 102 data points divided into two

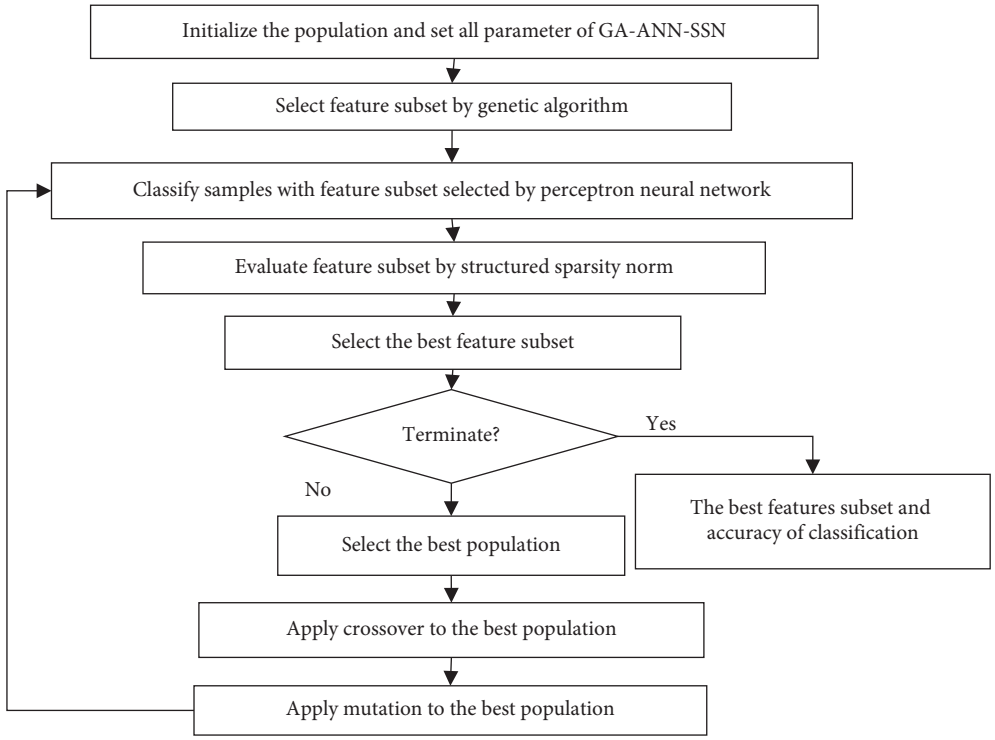

Figure 1: Proposed method diagram GA–ANN–SSN.

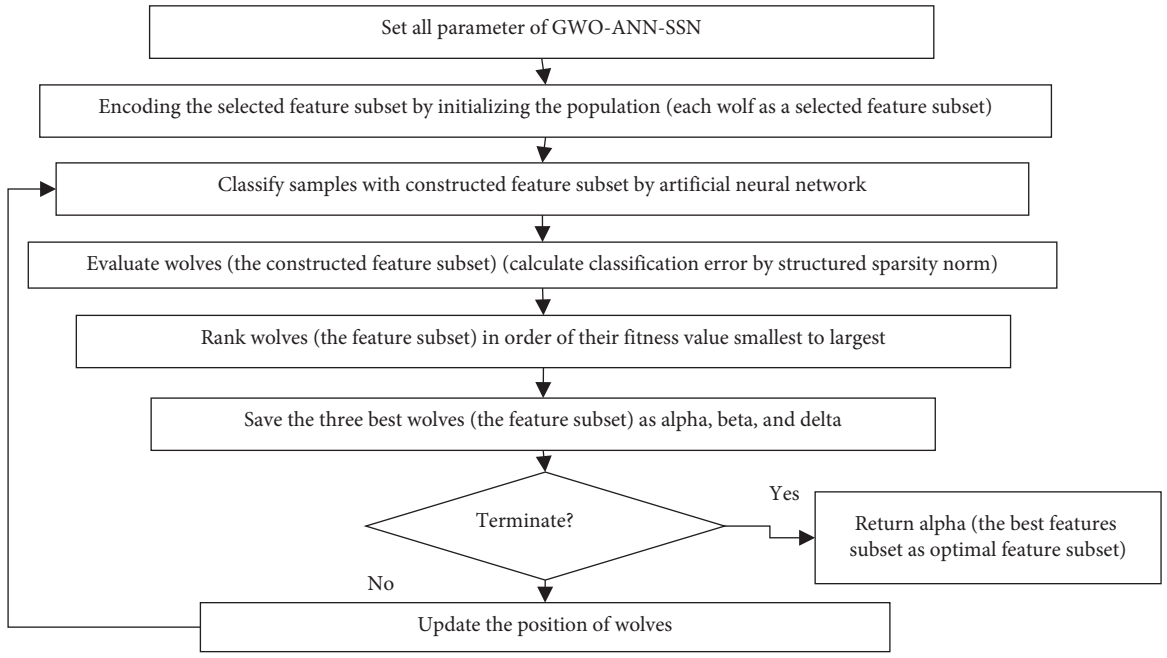

Figure 2: Proposed method diagram GWO–ANN–SSN.

classes, the GWO–ANN–SSN method accurately identified all 50 Class 1 data and correctly classified them, while 51 out of 52 Class 2 data were also accurately recognized and classified. This resulted in a 99% overall accuracy for all data. In addition, this performance was observed across various datasets. For example, in the ALLAML dataset comprising 72 data divided into two classes, with 47 in Class 1 and 25 in Class 2, the proposed GWO–ANN–SSN method accurately classified 46 out of 47 Class 1 data and correctly identified all 25 Class 2 data based on the confusion matrix. The GWO–ANN–SSN method demonstrated an accuracy of 98.6% for this dataset. Thus, in Figure 9, for the ALLAML

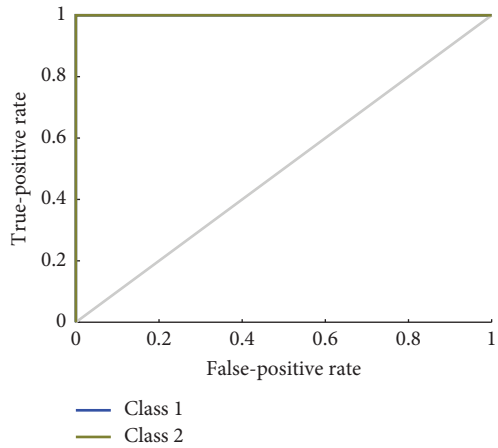

Figure 3: ROC plot for the ALLAML dataset.

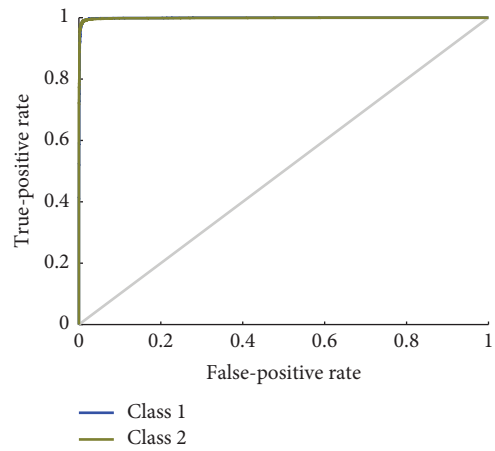

Figure 4: ROC plot for the GISETTE dataset.

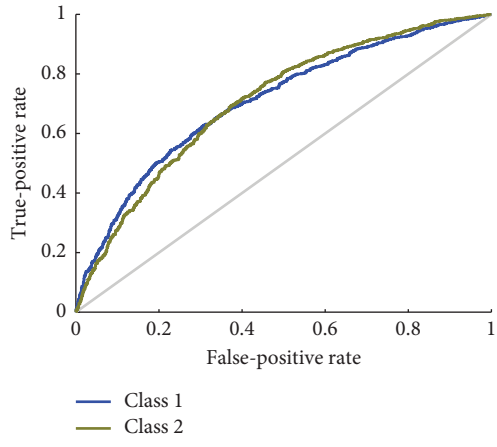

Figure 5: ROC plot for the MADELON dataset.

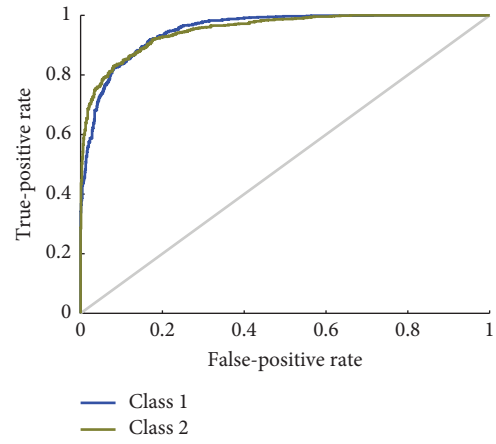

Figure 6: ROC plot for the PCMAC dataset.

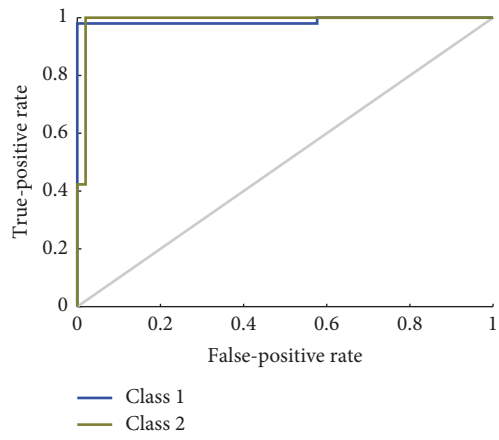

Figure 7: ROC plot for the PROSTATE-GE dataset.

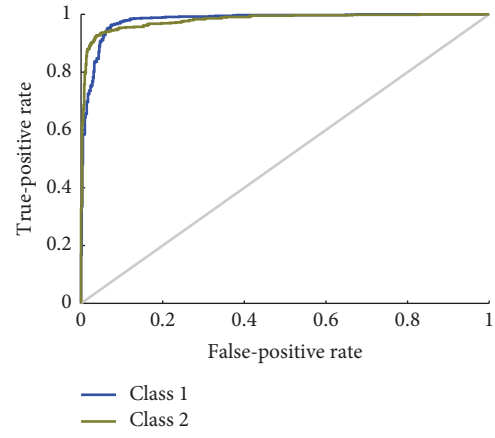

Figure 8: ROC plot for the RELATHE dataset.

dataset, 98.6% of the data classification was correctly recognized. In Figure 10, for the GISETTE dataset, 99% of the data classification was correctly recognized. In Figure 11, for the MADELON dataset, 62.1% of the data classification was correctly recognized. In Figure 12, for the PCMAC dataset, 67.3% of the data classification was correctly recognized. In Figure 13, for the PROSTATE-GE dataset, 99% of the data classification was correctly recognized. In Figure 14, for the RELATHE dataset, 88.9% of the data

Table 16: The correct rate identification indices of the proposed method GWO–ANN–SSN.

Dataset	TPR (%)	TNR (%)	FPR (%)	FNR (%)
RELATHE	90.2	87.5	12.5	9.8
PCMAC	86.8	87.8	12.2	13.2
GISETTE	99.1	98.9	1.1	0.9
MADELON	65.9	59.8	40.2	34.1
ALLAML	100	96.2	3.8	0
PROSTATE-GE	98.0	100	0	2.0

Figure 9: Confusion matrix plot for the ALLAML dataset.

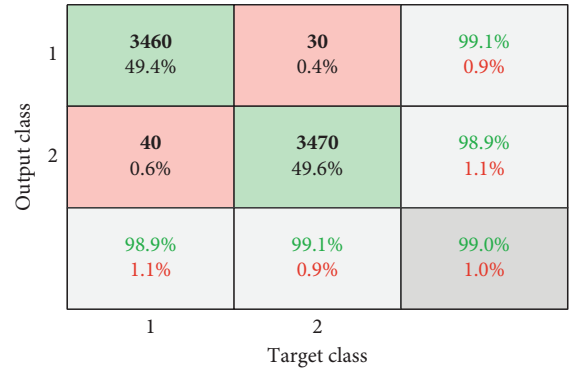

Figure 10: Confusion matrix plot for the GISETTE dataset.

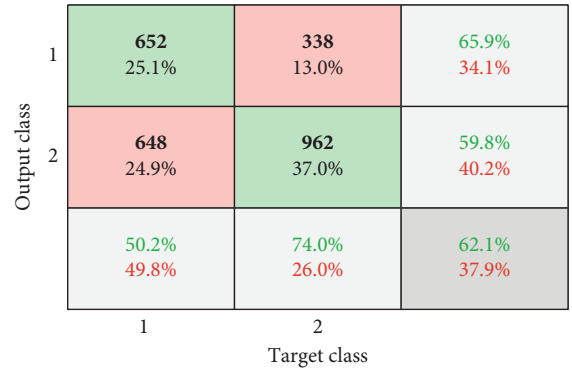

Figure 11: Confusion matrix plot for the MADELON dataset.

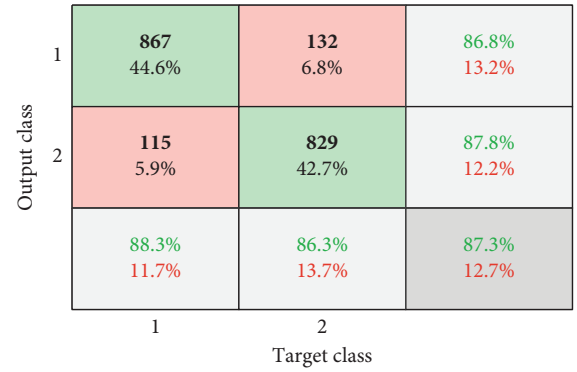

Figure 12: Confusion matrix plot for the PCMAC dataset.

Figure 13: Confusion matrix plot for the PROSTATE-GE dataset.

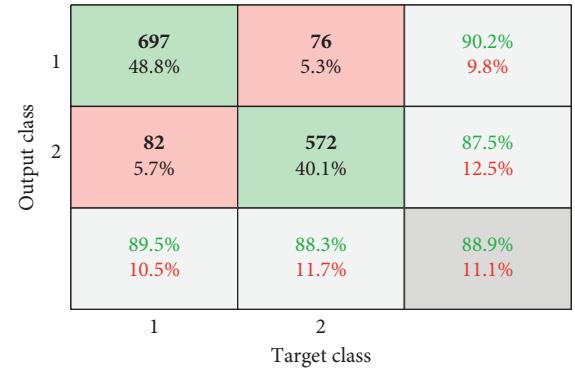

Figure 14: Confusion matrix plot for the RELATHE dataset.

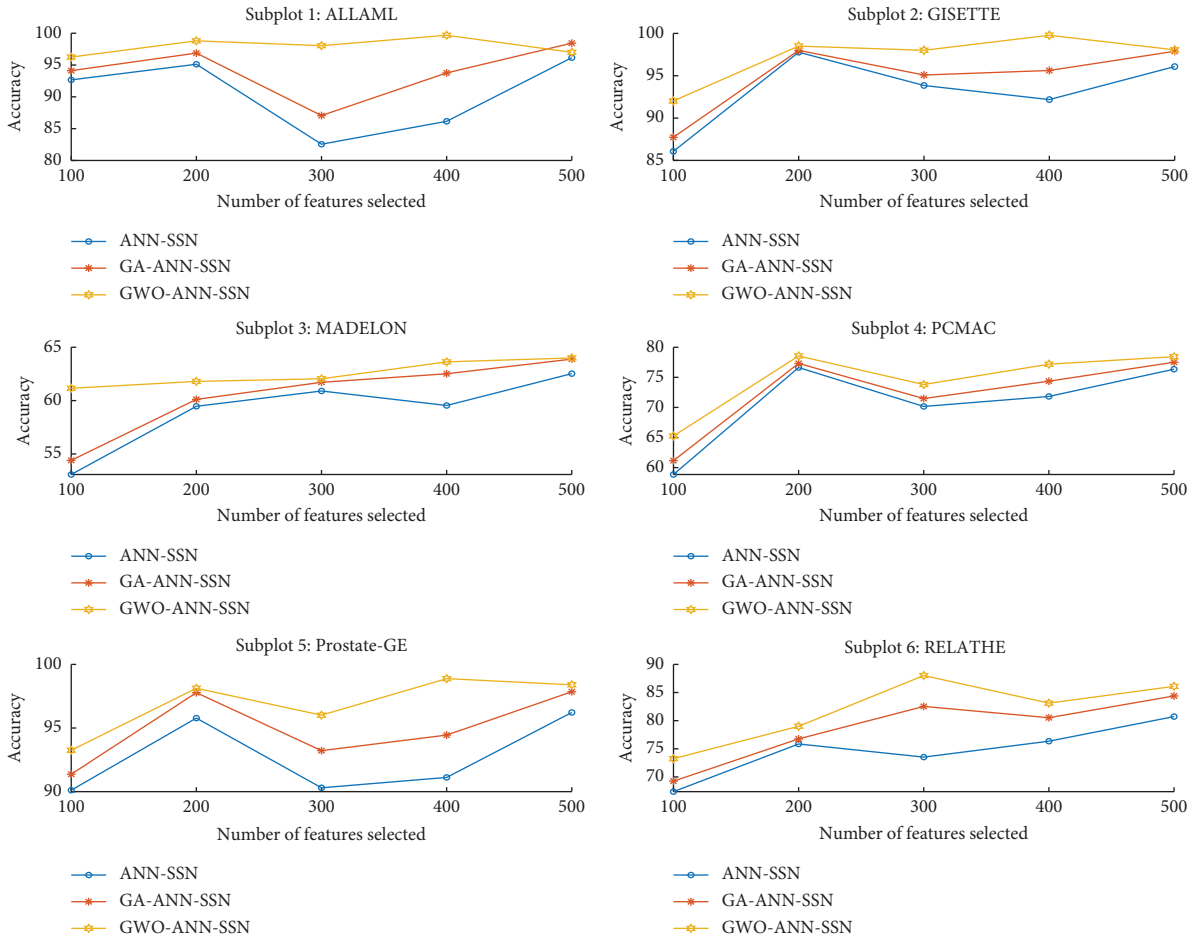

Figure 15: Comparison of accuracy of our proposed methods for each of the datasets with 100–500 features selected.

classification was correctly recognized. In Figure 15, for instance, in the case of the GISETTE dataset, the accuracy of the ANN–SSN, GA–ANN–SSN, and GWO–ANN–SSN models is similar when utilizing 200 features. However, the GWO–ANN–SSN model achieved the highest accuracy at 98.5001. Furthermore, the accuracy results for the GWO–ANN–SSN model were comparable across different feature counts, such as 200, 300, and 400. Therefore, reducing the number of features to 200 can lead to decreased memory usage and computational complexity while maintaining accuracy. In Figure 15, we can compare the accuracy achieved by the GWO–ANN–SSN algorithm (represented by the orange line), the GA–ANN–SSN algorithm (illustrated by the red line), and the ANN–SSN algorithm (shown by the blue line) across various datasets based on the number of selected features. Upon observation, it is evident that the GWO–ANN–SSN algorithm outperforms the others in all plots. The utilization of the least squares error with SSNs as the loss function transforms the problem into a convex one, ensuring convergence to the optimal global solution. Moreover, employing this norm guarantees sparsity in the results by focusing on the components' content rather than their quantity. Additionally, the imposition of the sparse condition through the loss function ensures that samples

Table 17: The unimodal benchmark functions.

Benchmark functions
$F1 = \sum_{i=1}^{n} x_i^2$
$F2 = \sum_{i=1}^{n} (\sum_{j-1}^{i} x_j)^2$
$F3 = \max_i \{\lvert x_i \rvert, 1 \ll i \ll n\}$
$F4 = \sum_{i=1}^{n-1} [100 (x_{i+1} - x_i^2)^2 + (x_i - 1)^2]$
$F5 = \sum_{i=1}^{n} ([x_i + 0.5])^2$
$F6 = \sum_{i=1}^{n} i x_i^4 + \text{random}[0, 1)$
$F7 = \sum_{i=1}^{n} [x_i^2 - 10 \cos(2\pi x_i) + 10]$
$F8 = \sum_{i=1}^{D} \lvert x_i \sin(x_i) + 0.1 x_i \rvert$

are effectively classified into distinct classes. According to the results in Tables 10, 11, 12, and 13, it is evident that the accuracy of the novel approaches, particularly the GWO–ANN–SSN method, surpasses that of previously established methods. The performance of the suggested GWO–ANN–SSN technique is demonstrated in Figure 9 through Figure 14, utilizing confusion matrices for individual datasets. Based on the data derived from these matrices, it is evident that the novel proposed approach exhibits exceptional proficiency in categorizing data into distinct classes and accurately recognizing the class of each data category.

Table 18: The comparison of accuracy of the proposed methods with the benchmark functions.

	RELATHE	PCMAC	GISETTE	MADELON	ALLAML	PROSTATE-GE
GWO–ANN–SSN	**79.0001**	**78.5501**	**98.5001**	**61.8001**	**98.55**	**98.01**
GA–ANN–SSN	76.76	77.31	97.99	60.11	98.53	97.41
F1	73.3929	65.4824	92.7567	58.0846	93.75	62.5
F2	69.9766	66.9912	90.6741	59.1193	84.7222	87.3303
F3	66.7414	67.7246	88.5782	56.6843	90.9091	83.4842
F4	65.2578	75.3036	92.4292	57.5345	88.8889	87.5
F5	70.7431	67.1072	92.4344	57.0768	95.8333	86.6667
F6	65.8641	67.6961	92.2685	57.0621	91.6667	93.75
F7	75.8341	93.2361	93.1185	50.0921	89.3367	80.711
F8	85.9631	81.6163	80.1685	53.1121	85.6117	73.733

Note: The bold values represent the highest accuracy achieved by the proposed methods compared to the benchmark functions. These values emphasize the notable improvement in performance, indicating that the proposed methods consistently outperform the benchmarks. Statistical analysis confirms the significance of these improvements, validating the effectiveness of the proposed approaches.

To assess the effectiveness of the suggested approaches, they were evaluated against various benchmark functions in Table 17, with the outcomes presented in Table 18. The results indicate that the accuracy achieved with the suggested methods surpasses that of the benchmark functions across all datasets. Notably, the GWO–ANN–SSN method exhibited the highest accuracy among the proposed methods.

This study introduces two embedded techniques for feature selection. The first proposed approach leverages an ANN integrated with a GA and SSN, while the second method employs an ANN in conjunction with a GWO algorithm and SSN. These methods uphold classification accuracy while utilizing a compact and effective feature subset. By employing the least squares error function, the feature selection problem is transformed into a convex problem, guaranteeing convergence toward the global optimum. The proposed methods allocate distinct weights to features, reflecting each feature's capacity to categorize samples into distinct classes. These approaches offer fault tolerance and noise resilience, combining global and local search strategies to identify the optimal feature subset. By reducing the input features in a classifier without compromising predictive efficacy, these methods exhibit notable advantages. The derivation of individual feature weights through equation (7) and the selection of features with higher weights result in comparable or enhanced classification accuracy compared to utilizing all features. There is an aspiration to broaden the scope of our proposed method to diverse domains requiring swift classification. Despite the encouraging outcomes, our approach does have limitations. While it may not decrease training and optimization time, it notably enhances classification accuracy.

7. Conclusion

In this study, we introduce two integrated methodologies, GA–ANN–SSN and GWO–ANN–SSN, for feature selection. The first approach combines an ANN with a GA and SSN, while the second method integrates an ANN with a GWO algorithm and SSN. These methods maintain classification accuracy while utilizing a concise and efficient feature subset. By employing the least squares error function, the feature selection problem is converted into a convex problem, ensuring convergence toward the global optimum (refer to Figures 1 and 2). The objective is to select features that excel in classifying data samples into distinct classes. In the GA–ANN–SSN method, the GA is employed to select the feature subset. Subsequently, the perceptron neural network is trained using the selected features and training data, with adjustments made to network parameters. The neural network's capability to classify data samples is then assessed using test data. Similarly, in the GWO–ANN–SSN method, the GWO algorithm is utilized for feature selection, followed by training the neural network and evaluating its classification performance using test data. The classification error is gauged using SSNs (equation (9)) in both methodologies. Feature weights are assigned to each feature in these methods, initially set randomly and updated in subsequent iterations using equation (9). Opting for features with higher weights leads to comparable or enhanced accuracy compared to using all features for classification. The adoption of this new sparsity norm offers several advantages. The structured norm reduces computation time, transforms the problem into a convex one ensuring convergence to the global optimum, and guarantees sparsity in the results by focusing on component content rather than quantity. The proposed approach is resilient to faults and noise, employing a blend of global and local search techniques to identify the optimal feature subset. Our methods effectively reduce the number of input features in a classifier without compromising predictive power. To validate the efficacy of our approach, we compare it against the existing feature selection methods across various publicly available datasets, including text data, random data, handwritten digits, and gene expression discovery. Across all experimental results presented in Tables 10, 13, 14, and 15 and Figures 3, 9, 10, 11, 12, 13, and 14, our method consistently outperforms other relevant approaches. The examination of features selected through our method aligns with the findings presented in this paper, supporting our results.

8. Future Work

For future work, regularization parameters are an important factor that can affect the generalization ability of our method.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author on request. The data are not publicly available due to privacy or ethical restrictions.

Ethics Statement

This article does not contain any studies with human participants or animals performed by any of the authors.

Consent

This research does not involve any human participants or animals.

Disclosure

A preprint was already published by Nemati et al. [2] whose data were used to compare the results and evaluate the efficiency of the proposed method of this article.

Conflicts of Interest

The authors declare no conflicts of interest.

Author Contributions

All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by all authors. The first draft of the manuscript was written by Amir Hosein Refahi Sheikhani, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

No funds, grants, or other support was received.

References

[1] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang, "Benchmark for Filter Methods for Feature Selection in High-Dimensional Classification Data," *Computational Statistics & Data Analysis* 143 (2020): 106839, https://doi.org/10.1016/j.csda.2019.106839.

[2] K. Nemati, A. Refahi Sheikhani, S. Kordrostami, and K. Khoshhal Roudposhti, "The Embedded Feature Selection Method Based on Artificial Neural Network Using Gray Wolf Optimizer and Structured Sparsity Norms," *Preprint in Research Square* 4 (2023).

[3] Y. Gao, Z. Xu, and C. Cao, "Avoiding Optimal Mean Calculation Robust PCA Based on the Nested LP-Norm and $L_{2,p}$-Norm," in *Annual International Conference on Network and Information Systems for Computers (ICNISC)* (Xiamen, China, October 2023), 497–502.

[4] H. Liu, N. Liu, Y. Yuan, C. Zhang, L. Zhao, and J. Li, "A Variable Selection Method Based on Fast Nondominated Sorting Genetic Algorithm for Qualitative Discrimination of Near Infrared Spectroscopy," *Journal of Spectroscopy* 1 (2022).

[5] N. Ullah, M. I. Mohmand, K. Ullah, et al., "Diabetic Retinopathy Detection Using Genetic Algorithm-Based CNN Features and Error Correction Output Code SVM Framework Classification Model," *Wireless Communications and Mobile Computing* 13 (2022).

[6] S. Aminah, G. Ardaneswari, M. Husnah, G. Deori, and H. Prasetyo, "Detection of COVID-19 Using Protein Sequence Data via Machine Learning Classification Approach," *Journal of Applied Mathematics* 1 (2023).

[7] Y. Wu, "An Information Entropy Embedding Feature Selection Based on Genetic Algorithm," *Security and Communication Networks* 2022 (2022): 1–10, https://doi.org/10.1155/2022/7111034.

[8] S. MaryJoans and J. Sandhiya, "A Genetic Algorithm Based Feature Selection for Classification of Brain MRI Scan Images Using Random Forest Classifier," *International Journal of Advanced Engineering Research and Science (IJAERS)* 21 (2017): 124–130.

[9] O. Babatunde, L. Armstrong, J. Leng, and D. Diepeveen, "A Genetic Algorithm-Based Feature Selection," *International Journal of Electronics, Communications and Computer Engineering* 15 (2015): 899–905.

[10] Q. Li, H. Chen, H. Huang, et al., "An Enhanced Grey Wolf Optimization Based Feature Selection Wrapped Kernel Extreme Learning Machine for Medical Diagnosis," *Computational and Mathematical Methods in Medicine* 15 (2017).

[11] A. Mohammad, "Intrusion Detection Using a New Hybrid Feature Selection Model," *Intelligent Automation and Soft Computing* 140 (2021).

[12] P. M. Kitonyi and D. R. Segera, "Hybrid Gradient Descent Grey Wolf Optimizer for Optimal Feature Selection," *BioMed Research International* 2021 (2021): 1–33, https://doi.org/10.1155/2021/2555622.

[13] M. khanna, L. K. Singh, K. Shrivastava, and R. singh, "An Enhanced and Efficient Approach for Feature Selection for Chronic Human Disease Prediction: A Breast Cancer Study," *Heliyon* 10, no. 5 (2024): e26799, https://doi.org/10.1016/j.heliyon.2024.e26799.

[14] L. KumarSingh, M. khanna, and Rekhasingh, *Feature Subset Selection Through Nature Inspired Computing for Efficient Glaucoma Classification From Fundus Images* (Berlin, Germany: ResearchGate, 2024).

[15] L. KumarSingh, M. khanna, and Rekhasingh, *An Enhanced Soft-Computing Based Strategy for Efficient Feature Selection for Timely Breast Cancer Prediction: Wisconsin Diagnostic Breast Cancer Dataset Case* (Berlin, Germany: Springer, 2024).

[16] M. Rekhasingh, L. KumarSingh, and H. Garg, *A Novel Approach for Human Diseases Prediction Using Nature Inspired Computing and Machine Learning Approach* (Berlin, Germany: Springer, 2023).

[17] F. Yan, J. Xu, and K. Yun, "Dynamically Dimensioned Search Grey Wolf Optimizer Based on Positional Interaction Information," *Complexity* 2019, no. 1 (2019): https://doi.org/10.1155/2019/7189653.

[18] S. Shringi, H. Sharma, and D. L. Suthar, "Fitness-Based Grey Wolf Optimizer Clustering Method for Spam Review Detection," *Mathematical Problems in Engineering* 3 (2022).

[19] M. Yuvaraja, S. Arunkumar, P. VinodhKumar, and L. MaryImmaculateSheela, "Improved Grey Wolf Optimization Based Feature Selection on Multiview Features and Enhanced Multimodal Sequential Network Intrusion Detection Approach," *Communications and Mobile Computing* 3 (2023).

[20] G. Manita and Q. Korbaa, "Binary Political Optimizer for Feature Selection Using Gene Expression Data," *Computational Intelligence and Neuroscience* 18 (2020).

[21] A. Saxena and S. Shekhawat, "Ambient Air Quality Classification by Grey Wolf Optimizer Based Support Vector Machine," *Journal of Environmental and Public Health* 2017 (2017): 1–12, https://doi.org/10.1155/2017/3131083.

[22] S. XiaLi and J. ShengWang, "Dynamic Modeling of Steam Condenser and Design of PI Controller Based on Grey Wolf Optimizer," *Mathematical Problems in Engineering* 1519 (2015).

[23] E. M. R. Devi and R. C. Suganthe, "Feature Selection in Intrusion Detection Grey Wolf Optimizer," *Asian Journal of Research in Social Sciences and Humanities* 7, no. 3 (2017): 671, https://doi.org/10.5958/2249-7315.2017.00197.6.

[24] H. Pan, S. X. Chen, and H. Xiong, "A High-Dimensional Feature Selection Method Based on Modified Gray Wolf Optimization," *Applied Soft Computing* 135 (2023): 110031, https://doi.org/10.1016/j.asoc.2023.110031.

[25] G. Roffo, S. Melzi, U. Castellani, and A. Vinciarelli, "Infinite Feature Selection: A Graph-Based Feature Filtering Approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12 (2020).