International Journal of Electrical and Computer Engineering (IJECE)
Vol. 13, No. 2, April 2023, pp. 2052~2067
ISSN: 2088-8708, DOI: 10.11591/ijece.v13i2.pp2052-2067

Optimizing Alzheimer's disease prediction using the nomadic people algorithm

Shaymaa Taha Ahmed[1], Suhad Malallah Kadhem[2]
[1]Department of Computer Science, College of Basic Education, University of Diyala, Diyala, Iraq
[2]Department of Computer Science, University of Technology, Baghdad, Iraq

Article Info

Article history:

Received May 16, 2022
Revised Sep 30, 2022
Accepted Oct 25, 2022

Keywords:

Alzheimer's disease
Gene selection
Microarray technology
Nomadic people optimizer
Support vector machine

ABSTRACT

The problem with using microarray technology to detect diseases is that not each is analytically necessary. The presence of non-essential gene data adds a computing load to the detection method. Therefore, the purpose of this study is to reduce the high-dimensional data size by determining the most critical genes involved in Alzheimer's disease progression. A study also aims to predict patients with a subset of genes that cause Alzheimer's disease. This paper uses feature selection techniques like information gain (IG) and a novel metaheuristic optimization technique based on a swarm's algorithm derived from nomadic people's behavior (NPO). This suggested method matches the structure of these individuals' lives movements and the search for new food sources. The method is mostly based on a multi-swarm method; there are several clans, each seeking the best foraging opportunities. Prediction is carried out after selecting the informative genes of the support vector machine (SVM), frequently used in a variety of prediction tasks. The accuracy of the prediction was used to evaluate the suggested system's performance. Its results indicate that the NPO algorithm with the SVM model returns high accuracy based on the gene subset from IG and NPO methods.

Corresponding Author:

Shaymaa Taha Ahmed
Department of Computer Science, College of Basic Education, Diyala University
Diyala, Iraq
Email: mrs.sh.ta.ah@gmail.com

1. INTRODUCTION

Alzheimer's disease is a progressive illness that causes memory and cognition to deteriorate. Damage to the brain's nerve cells can lead to issues with language and memory. Symptoms of Alzheimer's disease may occur after 65 years of age and grow more prevalent as you become older. In the United States alone, it is estimated that by 2050 there will be between 11 and 16 million people with Alzheimer's disease (AD) [1]. This will cost the global economy $2 trillion by 2030. Unfortunately, there is no reliable way to identify Alzheimer's disease before it manifests symptoms, this may be the sole opportunity to intervene in the disease's course [2]. However, new technology has been developed that uses microarray to define the gene that causes AD. This is important because it could lead to early detection and treatment of the disease. Additionally, this technology may also help identify people who are at risk for developing AD [3].

Biologists use microarray technologies to monitor gene expression levels in an organism. The main problem with analyzing microarray data is that there are many genes and small samples [3], [4]. This can result in a reduction in prediction accuracy as well as an enhancement in over-fitting issues. Gene selection is a way of resolving this issue, that only takes the best set of genes for constructing a classification model [5], [6]. Gene selection is a strategy for selecting a small sample from a larger number of informative genes [7].

This article discusses how researchers have used a specific group of genes to understand genetic diseases better. The use of this subset of genes allows for a reduction in the cost of computation and an increase in the classifier efficiency for Alzheimer's disease. A variety of algorithms can be used to choose genes, such as information gain (IG), and nomadic people optimizer (NPO). These techniques are common unsupervised approaches for analyzing gene expression microarray data. This can provide an overall framework of the dataset being examined. This has recently been used to produce low-dimensional gene expression data before categorization on large datasets. Microarray data classification is a difficult task. The bioinformatics community uses multiple approaches machine learning techniques were used to diagnose and categorize microarray data. However, using a specific subset of genes can be beneficial in this process. Therefore, more research should be conducted to gain a better understanding of how these genes can be used to improve the classification of microarray data.

2. RELATED WORKS

Recent studies by [8] have shown that it is possible to predict Alzheimer's disease by employing the NPO, a novel algorithm. This algorithm is designed for standard and large-scale optimization problems. In addition, [9] used the Rhinoceros search algorithm as a criterion for feature selection tool for improving the accuracy of predictions. This best result in terms of predicting Alzheimer's disease was yielded by applying a deep learning algorithm. However, traditional machine learning algorithms may also be used to achieve good results.

AD can be predicted through a novel method for large-scale and typical optimization issues: the NPO. This algorithm is more efficient in differentiating between AD and normal genes. Additionally, it was shown that the multi-layer perceptron neural network (MLPNN) model was more effective in defining the differences between AD and normal genes. In a study by [10], machine learning was used to find the most important genes; support vector machine (SVM) was used as an effective gene categorization method. Ultimately, this information can be used to improve our understanding of AD and help develop treatments for this disease. The experiments were carried out in July-August 2020 and involved two microarray datasets for cancer: colon and lymphoma. The tests demonstrated the suggested method for identifying subsets of relevant genes and improving classification accuracy by reducing the number of inputs.

Three different feature selection methods were used in the study by [11]: IG, random forest (RF), and a wrapper for genetic algorithm and SVM (GA/SVM) are all examples of machine learning algorithms. C4.5 (decision tree), naive Bayes (NB), RF, k-nearest neighbor (KNN), SVM with linear kernel, and SVM with Gaussian kernel were among the six classification methods employed. The created a machine learning approach called SVM (2018). On the AD dataset, the proposed method was tested (GEO: GSE5281). We may use the suggested approach to choose groupings of minor genes that can be efficiently taught together.

This paper has faced many challenges such as the curse of dimensionality: The number of genes in the gene expression data is very huge (containing thousands of genes). Commonly, not all genes are useful; some genes are irrelevant and redundant information in the dataset. Therefore, working on this huge number of genes is difficult. The application of the prediction model to this dataset: This presents another difficulty as it should be implemented with the least possible error, least execution time, and highest achievable accuracy.

The proposed method has been evaluated based on combining gene expression data with a particular network of genes to discover full-spectrum AD genes throughout a whole genome in the human brain. This study of genes that are linked to Alzheimer’s disease and their reliability was assessed [12] by the use of this approach. Jameel and Abdullah [12] applied principle component analysis (PCA) to gene expression data with high dimensionality to extract efficient genes. Gene selection helps in reducing computation requirements, understanding data, improving the predictor performance, and reducing the effect of the curse of dimensionality. Following that, an artificial neural network (ANN) was used to examine the retrieved representation of the performance. The ANN is a supervised classification approach for determining which genes are required for illness diagnosis. A finding of the suggested method indicated that interactions between genes might aid in the detection of diseases [13].

3. MICROARRAY TECHNOLOGY

Deoxyribonucleic Acid (DNA) technology is used in the micro arraying of the samples. DNA chips, or microarrays, are tiny pieces of DNA that have been attached to a solid support [14]. The term “microarray” refers to a group of devices that use this technique for identifying and determining properties of components present in complex matrices (such as cell cultures) based on many copies of the same DNA sequence that can be found in a single place, that is an organism's unique representation of a gene. This spot is arranged in rows and columns in Figure 1 [15].

Each gene's expression level is recorded as an image (CEL File), see Figure 1. Subsequently, data may be extracted from such a picture using particular software [16]. There are several software packages available for analyzing microarray data. The most commonly used is the Limma package, which is a module of analysis tools for raw CEL files [14].

The DNA microarray is a surface used by biologists to monitor the expression levels of many genes in an organism. This is done by comparing the gene expression level in an unhealthy cell to that of a healthy cell. This can help identify genes responsible for different diseases [15]. The technology of microarrays is a great tool for studying gene expression. This method enables scientists to simultaneously assess the expression levels of thousands of genes. This data collected from microarray experiments generally takes the shape of a matrix. This matrix contains information about the experimental conditions and the expression levels of the genes. The data is classified into one of two types: rows and columns (gene sequence or genes). Any study relies on the collection of this information shown in Figure 2 [16].

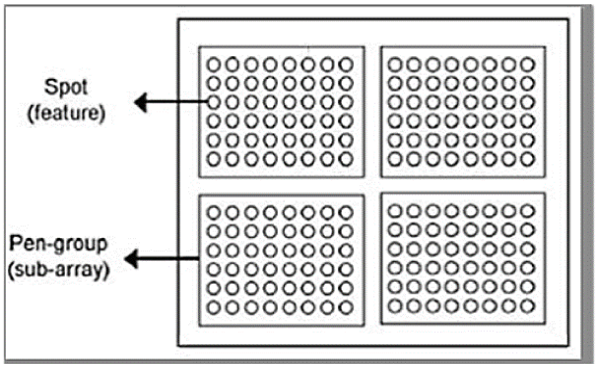

Figure 1. The surface of a DNA microarray [15]

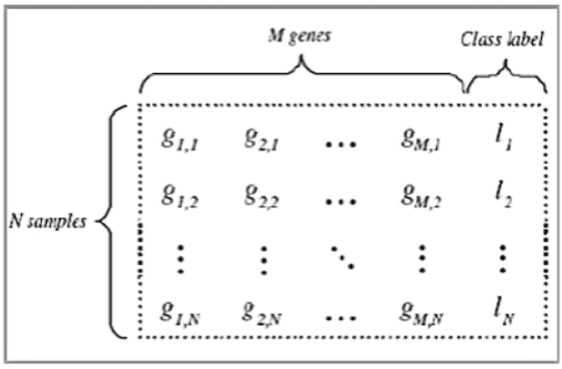

Figure 2. Indications the gene expression data matrix [17]

4. DATASET

The NCBI GEO provided the microarray AD datasets, a public functional genomics data repository of high-throughput gene expression data. Microarray data from postmortem central nervous system (CNS) tissues of 245 AD patients, 142 mild cognitive impairment (MCI) patients, and 182 healthy controls are included in the GSE63060 and GSE63061 dataset (released on Jan. 29, 2011) [17]. The GSE1297 (GPL29311, nano string counter human myeloid innate immunity panel v2) dataset (published January 8, 2021) contains gene expression data from the frontal cortex (FC) and cerebellum (CB) of 22 AD patients and 9 healthy controls [16]. GSE5281 comprises 161 samples, with 74 non-demented controls and 87 Alzheimer's disease patients [18]. GSE36980 (released Apr. 17, 2013) contains microarray data from the grey matter of the frontal and temporal cortices, as well as hippocampi, derived from 80 postmortem brains of 38 Alzheimer's patients and 42 controls [16]. GSE132903 (published on January 20, 2015) contains microarray data from 97 Alzheimer's disease patients and 98 healthy controls [18]. The GSE33000 dataset (released Jan. 8, 2021) contains gene expression data from the FC and CB of 310 AD patients, 157 MCI, and 157 control subject [19]. The current study did not need the approval of an ethical committee or patient consent. Table 1 summarizes the transcription datasets that were eventually used in this study.

4.1. Data preprocessing

Data processing is the process of modifying data before using a data mining technique. Because data will almost certainly be faulty and rife with anomalies, it is not suitable for beginning a data mining procedure. Data preprocessing may transform the data to meet the demands of each data mining technique, allowing it to be processed that might not be possible otherwise. It is a powerful instrument that allows you to handle and analyze complex data. Techniques for data preparation and reduction are used. The first, which contains information normalization, by selecting features, seeks to decrease the data's complexity, while the second, known as feature elimination, identifies outliers or anomalies in a set of data [19]. The use of an effective data preprocessing method procedure may transform an ultimate dataset got into a reliable and appropriate data mining algorithm's source used subsequently.

Table 1. Details of AD dataset

Datasets	Alzheimer Disease AD	Normal	MCI	Total No.	Age	Number of Attributes (genes)
GSE63060, GSE63061	245	182	142	569	Range [55 to 91]	16382
GSE5281	87	74	NA	161	Range [60 to 90]	54675
GSE1297	22	9	NA	31	Range [55 to 88]	22283
GSE132903	97	98	NA	195	Range [57 to 92]	42179
GSE33000	310	157	157	624	Range [50 to 90]	39280

4.1.1. Data Normalization

Raw data is rarely in a form that data mining techniques can utilize. Normalizing raw data values to another form with characteristics more suitable for modeling and analysis is known as data normalization [18], [20]. All genes should be measured in the same unit of measurement. As a result, it eliminates any variance between significant small and large values that might skew the results. As shown in (1) is one approach for data normalization in which the minimum and maximum values of the original variables are replaced with a mean value. In contrast, z-score and min-max normalization are two ways to accomplish this. In min-max normalization, (1) is utilized to compute the value [21].

$$\hat{V} = \frac{V - min_a}{max_a - min_a} * (new\_max_a - new\_min_a) + new\_min_a \quad (1)$$

where V is representing the gene value, min_a is the minimum original value for any gene, max_a is the maximum original value for any gene, and New-max_a and new-min_a are the maximum and minimum intervals of values.

4.2. Gene selection in high dimensional datasets

This set of genes (sometimes thousands) vastly outnumbers the sample used in microarray data. Then, many typical approaches are ineffective or computationally demanding when dealing with such data. The truth is that not every single one of the thousands of genes has to be classified [7], [22]. The majorities of genes are irrelevant and negatively influence classification performance. The value of the features is diminished, and more genes are discovered to be linked with the disease than those that were previously identified. This increases the complexity of the problem, causes computational strain, and generates useless noise in classification methods [23], [24]. As a result, identifying a minimal number of genes is crucial, known as informative genes that may be sufficient for acceptable classification. The most significant subset of genes, on the other hand, is frequently unknown [25], [26]. Filter and wrapper methods are two common gene selection techniques. Filter procedures evaluate all characteristics in terms of their quality based on a univariate scoring statistic and relate every gene to the class label based on that relation. Before classifying algorithms are used, the top-rated genes are chosen. In contrast, wrapper approaches necessitate that a gene selection method be combined and uses a classifier to evaluate each gene subset's categorization performance. The Feature selection techniques are illustrated in Figure 3.

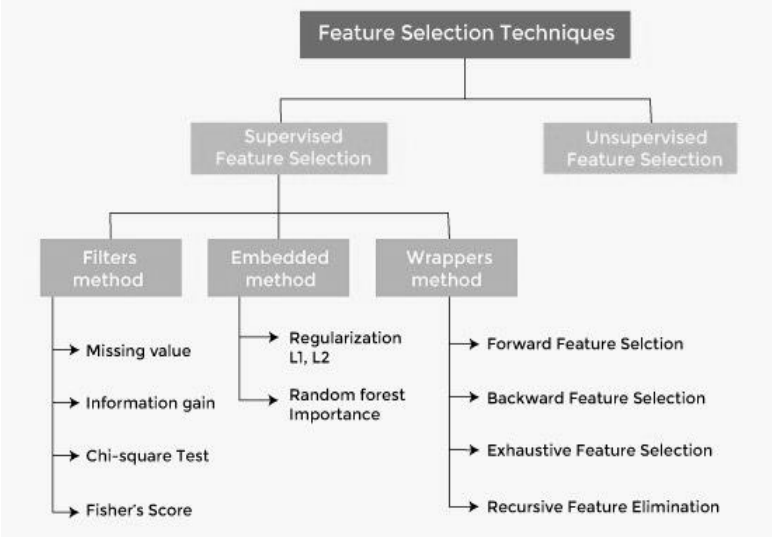

Figure 3. Feature selection techniques [27]

This top group of genes is chosen according to the performance of the classifier in each subset. Because the filter procedure cannot measure gene relationships, and the wrapper method requires a significant computational investment, it is not viable [28], [29]. Such individualizing comparative genomic hybridization (CGH) applications and gene selection techniques may reduce computational costs and increase classification accuracy. This method can also optimize the efficacy of AD classification by finding a small subset of genes that produce high results [27], [30]. Several gene selection strategies, such as IG and NPO, have been utilized in this research to isolate relevant genes that are linked with a disease's diagnosis immediately.

5. METHOD

In this part, the suggested technique entails several key procedures, such as importing the microarray AD data set in its raw form. Then, utilizing the Min–Max approach, normalize the data. We approached feature selection by combining a filter (IG), wrapper (wrapper-based NPO), and supervised method (SVM). The filter wrapper removes genes having the most redundancy (i.e. they have a high correlation among themselves) and identifies the best sub genes based on NPO. Chosen genes pass through the SVM to be further compressed. The resulting, compressed genes enable the classifying of AD and non-AD individuals by using SVM. The process flow of the proposed structure may be seen in Figure 4.

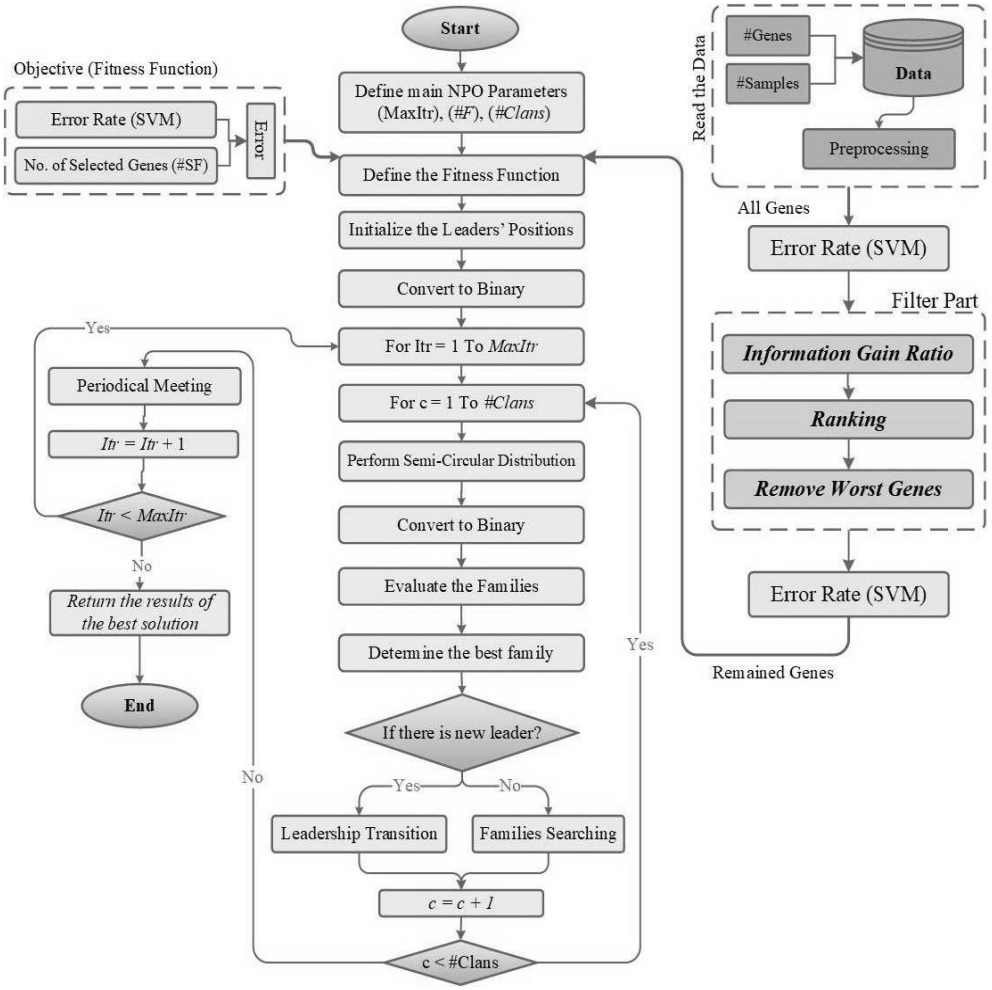

Figure 4. Algorithm of NPO-SVM classifier model

5.1. Gene selection using entropy and information gain

Entropy is the basic rule of information theory used as an equation for computing the similarity of the characteristics. If the samples are completely homogenous, for example, entropy is 0 for them, the

entropy value of the evenly divided samples is one [24], [31]. Because of the dataset's higher dimension and small sample size, classifying these data is difficult. From thousands of attributes that are usually investigated, very few of these attributes display relevance to disease. So, must keep only the relevant features [25]. Studying the accurate selection of genes for the categorization process is aided by a gene profile. $E(X) = - R^{+}\log_2 (R^{+}) -R^{-}\log_2 (R^{-})$ for a selection of bad and good characteristics. The entropy formula as (2) [20]:

$$\text{Entropy } (X)= \sum_{i=1}^{m} -(R_i \log_2 R_{i)} \quad (2)$$

where R_i is prior probability X are categorical variables, and I is the classification system's category number. When there are two classification issues, it is a specific case (where M number of classes). Let J be a gene with n different potential values ($j1, j2... jn$). The following is the entropy:

$$Entropy(i/j) = - \sum_{j=1}^{N} p(j) \sum_{i=1}^{m} p(\frac{i}{j}) \log_2 (p(\frac{i}{j})) \quad (3)$$

where $p(\frac{i}{j})$ is the conditional probability of variables I after attribute J is constant and all numbers of attribute and M field of the classes. The entropy plays a big role in calculating information gain 27[26]. Following the distribution of the dataset's characteristics. The entropy of all the characteristics in the dataset is then computed. It then separates the data into the form of groups of features. Each group's entropy is determined independently, its total entropy is calculated by adding the entropies of all groupings. The entropy is then subtracted after the distribution from the entropy before distribution [27]. Where gene J and category I are not relevant.

$$IG\ (J) = Entropy(S) - Entropy\ (i/j) = Zero. \quad (4)$$

While if relevant $Entropy(S) > Entropy(i/j)$ this leads to information gene (IG) $(J) > 0$. The higher the connection between J and I, the greater the difference between J and I. The larger value of the feature information gain increases its importance for the classification. So, it is chosen genes with greater IG to indicate the original high-dimensional gene first then use as a basic in gene selection [29]. The IG flowchart describes IG algorithm steps as shown in Figure 5.

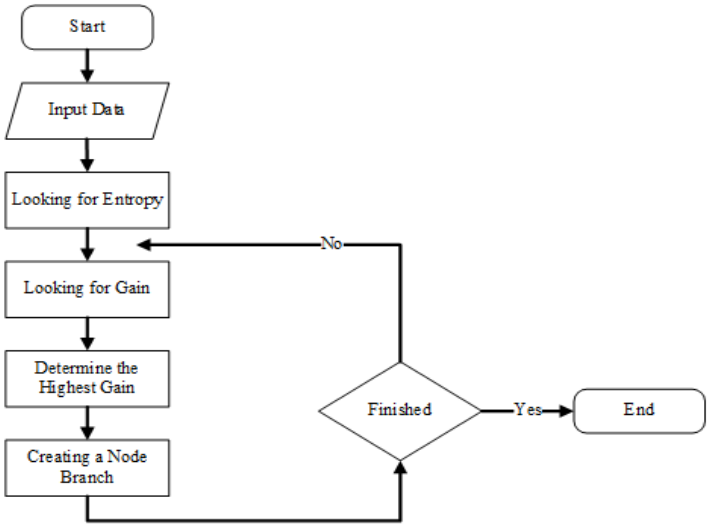

Figure 5. Flowchart for IG

Where the input data set has an attribute, and the outputs required is select E attributes subset of attributes. First, establish attributes for classification. Second compute the entropy for each class of all

samples, depending on the probability using the (2). Then, using (3), for each value of a single property, compute the conditional probability, which is then applied to all characteristics to generate conditional entropy. Compute information gain by (4) for all attributes. Results values of information gain are arranged in ascending order and select the highest values to depend on the threshold value.

5.2. Nomadic people optimizer

Starting with a concept of nomadic people and their mode of existence, this section discussed the fundamentals of NPO; this is used as the basis for the proposed method. NPO is the algorithm associated with social aspects and is developed around considering the movement of nomads in search of food [32]. The NPO algorithm can easily deal with large scale problems.

The NPO is a new metaheuristic algorithm based on the swarm-like behavior of nomadic people [32]. It mimics the movements and hunts for the sustenance of nomadic tribes, involves grazing grass or water, and how they've managed to live for hundreds of years by relocating to the best and most suited environments [32]. The multi-swarm technique was used to create the algorithm, containing a large number of clans, each of which is vying for the greatest spot for the best answer regarding a leader's location [32]. The validation of the process is done based on 36 unrestricted benchmark functions.

The NPO mimics how nomads live and their movements as they look for grass and water for their animals and other sources of life. The NPO is also designed to capture the way of life of the nomads and their existence for many years and how they keep moving from one place to another, constantly looking for comfort [32]. The NPO design follows the multi-swarm approach because it is made up of various clans, with every clan having a leader who also happens to be the best member in the respective clan. Nomads, being herders, use much of their time migrating from place to place as they search for natural life sources [32]. The animals benefit from water and food from the nearby sources, and they provide food and other needs to those who own them.

5.2.1. Main steps

The NPO method has five main operators: the first meeting, the semi-circular distribution, and the search for families, leadership is being transitioned, and finally, a regular meeting is being held [32].

a. Initial gathering (initialization)

A group of leaders (σ), where $\sigma_i = \{\sigma 1, \sigma 2, \dots, \#Clans\}$ is initialized in a random manner by use of this (5).

$$\overrightarrow{\sigma_c} = (\mathrm{UB} - \mathrm{LB}) \times Rand + \mathrm{LB} \quad (5)$$

b. A circular distribution (Local search—exploitation)

A set of families (*x*), in which $Xi = \{X1; X2; \dots; \#Families\}$ is dispersed about the relevant community head *Q*. mathematically the distribution of points is given as:

$$X = (Rd \times \sqrt{R_1}) \times \cos(\theta) + X_0 \quad (6)$$

$$Y = (Rd \times \sqrt{R_2}) \times \sin(\theta) + Y_0 \quad (7)$$

where X_0 and Y_0 stand for the coordinates of the point of origin, and R_1 and R_2 indicates the random coordinates in the perimeter of the circle [32]. In the meantime, θ is the angle value of the point which is a random value lying between $[0.2\pi]$. In this manner, the tent distribution in a random manner about the tent of the leader needs an *X* coordinate but the non-needed *Y* coordinate is excluded [32]. Using that the equation is devised to suit this scenario, as shown in (8).

$$\overrightarrow{X_c} = \overrightarrow{\sigma_c} \times \sqrt{R} \times \cos(\theta) \quad (8)$$

where $\overrightarrow{x_c}$ is representative of the position of the family, $\overrightarrow{\sigma_c}$ stands for the leader's position for the same clan © with R representing a random number in the range [0,1].

c. Families searching (global search-exploration)

In the NPO, the searchable section is implemented whenever a local best solution is lacking within the swarm 32[31]. In this situation, families look for improved positions. Every family goes in a different direction, searching for space using random steps and directions created by the Levy flight formula as (9):

$$\overrightarrow{X_{i^{new}}} = X_i^{\vec{d}} + (a_c * (\sigma_c - X_i) \oplus Levy) \quad (9)$$

where $\overrightarrow{X_{inew}}$ and $X_i^{\vec{d}}$ stands for the families' new and old positions, a_c stands for the clan's area and σ_c. a_c may be computed by use (10).

$$a_c = \frac{\sum_{i=1}^{\phi} \sqrt{I\sigma c_{\_}}\, xi_1^{61d}\, J^2}{\phi} \tag{10}$$

The movement of the families disburses in various directions using random-sized steps, which are produced by the Levy flight (λ_c) equation as shown in (11):

$$\text{Levy} \sim u = t^{-\lambda} \; (1 < \lambda \leq 3) \tag{11}$$

d. Leadership transition (exploitation)

Checks are made for each of the clans to find any new family that is more suitable than the leader of that clan that family takes the leadership and vice versa [32].

e. Periodic meetings (exploitation–exploration)

Periodic meetings vary but leader redeployment in the wilderness is consistent [32]. Leaders go from the meetings to update their locations. The update happens by including variance between the most powerful leader's place and that of the ordinary leader as shown in (12).

$$\Delta\text{Pos} = \Psi \left(\frac{\sqrt{\sum_i^D (E - n)}}{\#D} \right) \tag{12}$$

The direction variable Ψ leads the ordinary Leaders to good locations based on the best sheikh's fitness value, as shown:

$$\Psi = \begin{cases} -1^1 & \text{if } f(\sigma) \geq 0 \end{cases} \tag{13}$$

Otherwise, the ordinary leaders appraise the places through (12). It is an equation the representation from a part of survey steps in NPO.

$$\overrightarrow{\sigma_{c^{new}}} = \overrightarrow{\sigma_c N} + \Delta\text{Pos}(\sigma^E - \sigma_C^N) * \frac{IT}{\#T} \tag{14}$$

The meeting room approach (MRA) and the main phases of NPO are outlined below and illustrated in Figures 6 and 7 respectively.

Algorithm: nomadic people optimizer (NPO)

Input: Number of Clans (#Clans), Number of Families (Φ), Number of Iterations (#T)
Output: The Most Excellent Sheikh
Procedure:
Create the objective function f (x),
Initialize the Leaders σ_{c^O}, c= {1, 2, 3,… # Clans}
Calculate each leader's fitness value using the formula below f(x)
Repeat (Itr):
For = 1 to # Clan
Divide the solutions/families in a semi-circular form around the leader through (4)
Compute each solution's fitness value $X^c{}_i$ via f(x)
Set the best $X^c{}_i$ in the clan c as σ^B
If σ^B is better than the original σ^O Then, Swap them $\sigma^O = \sigma^B$
Else: Explore the search space using the following steps:
Compute the average distance between all of the families using the formula (6)
Transfer the family to the new location by using eq. 5
Compute each solution's fitness value $X^c{}_i$ via f(x)
Set the best $X^c{}_i$ in the clan C as σ^B
If σ^B is better than the original σ^O Then, Swap them $\sigma^O = \sigma^B$
End if
End For
Implement the Periodical Meeting
Loop Until (Itr > #T)
Return σ^E

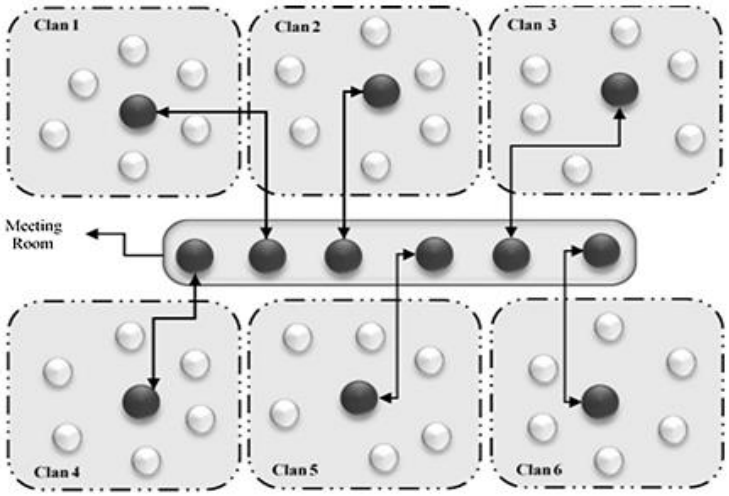

Figure 6. Meeting room approach [28]

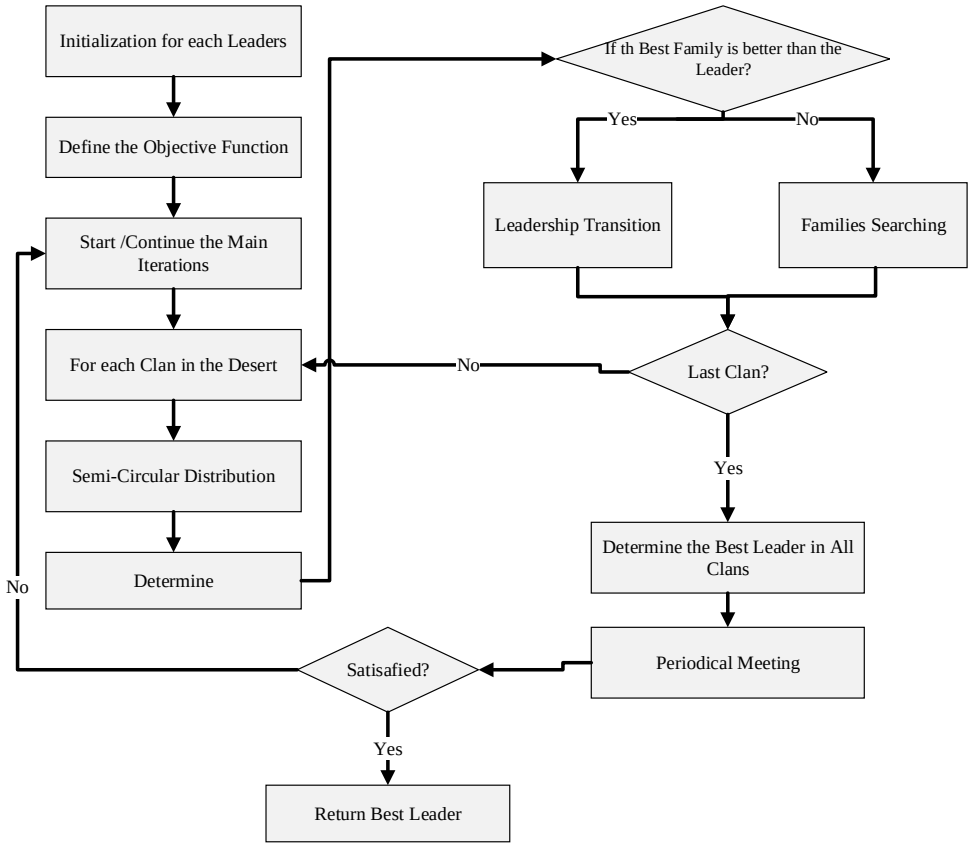

Figure 7. Flowchart of NPO [28]

5.2.2. Source of inspiration

The NPO gets inspiration from the lifestyle of nomadic people [32]. The suggested algorithm is a simulation of how the nomads behave whenever they search for life sources, including grazing fields and water [32]. The fundamental component of the algorithm comprises many clans, and every clan looks for the best place or most appropriate solution based on the position of the leader. The interaction between the clans gets inspiration based from the notion of groups of human beings being guided by their leaders [32]. The leaders of the clans occasionally meet in a given room to select a general best leader who controls all the

other leaders [32]. The meeting room approach ensures a balance between the examination and exploitation capabilities of the suggested NPO. NPO has two phases for the part of exploitation, but the exploration is undertaken using another step [32]. The local NPO search is executed using the peculiar formula of distribution. However, the global search ability has a levy flight equation that creates a step to move the families to reach the new positions.

5.3. Classification using support vector machine

Following feature selection, the classifier is performed using SVM to extract features effectively and a set of rules to do classification. A distinct hyper-plane represents SVM, which is a discriminative classification algorithm [33], [34]. Figure 8 illustrates it.

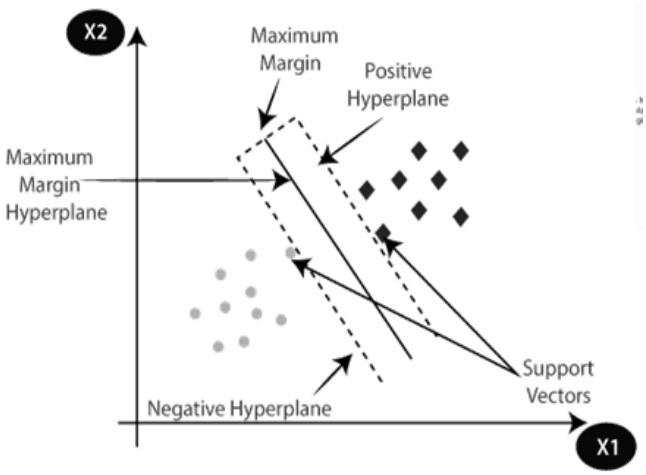

Figure 8. Support vector machine [33]

Because of its high accuracy and capacity to analyze data with high dimensionality, the SVM classifier is widely utilized in a variety of ways like bioinformatics, signal processing, and computer vision [35]. The two-class problem connected with Vapnik–Chervonenkis theories and structural principles are successfully solved by SVM. $W.x + b = 0$ is the general formula for the linear discriminant function. To differentiate the samples without noise, an optimal hyper plane is used between the two groups, which are mathematically provided in (15).

$$pi[w.\ x + b] - 1 \geq 0,\ i = 1, 2,..\ N \quad (15)$$

Then, reduce $\|w\|$ 2 in the (15), so the optimization issue is solved by the saddle point of a Lagrange function with Lagrange multipliers αi. The ideal discriminant function is denoted in (16).

$$(x) = sign\ \{(w * x) + b *\} = sign\ \{\sum \alpha i\ N * i{=}1.\ pi\ (xi * - x) + b *\} \quad (16)$$

Finally, using greater dimensional data, to minimize computational complexity, replace the interior product ($xi\ x$) with a linear kernel function ($x,\ x'$) in (16). As a result, the linear separability of calculated samples has improved, as well as the rewriting of the discriminant function as indicated in (17).

$$(x) = \{\sum \alpha i\ N * i{=}1.\ pi.\ (x,\ xi) + b *\} \quad (17)$$

Expression of genes microarrays are gaining traction as a tool for medical decisions in terms of diagnostics in Alzheimer's disease and other complicated disorders. Researchers are always working to create and deploy the to maximize the technology's benefits, and the most accurate decision support algorithms for the generation of gene expression patient profiles must be used. According to previous research, support vector machines (SVMs) perform the best among the well-known and often used methods for categorizing microarray gene expression data in K-nearest neighbors, the performance of back propagation neural networks, and probabilistic neural networks, weighted voting techniques, and decision trees are all significantly worse. The reasons behind this are as follows: i) SVMs have shown the capacity to not only

accurately classify items into suitable categories but also to identify instances when the evidence does not support the established categorization; and ii) the ability to choose a similarity function with ease, as well as the sparseness of the solution while working with large data sets, the ability to deal with large feature areas, and the capability to recognize outliers are all mathematical aspects that make SVM appealing for gene expression research.

5.4. Cross validation (CV)

To avoid over-fitting and improve the robustness of the results obtained, this work applies the 5-fold (CV) technique for divider the original datasets into 5 folds evenly of equal size. Of 5-folds, 4 folds are used for training our framework, and one-fold is held for testing.

6. THE PROPOSED METHOD

This section contains an assessment of the suggested approach as well as the study's experimental outcomes. Before using the data, it must be preprocessed and normalized; the methods are outlined in section 3. The evaluation metric is determined in terms of results that generate trustworthy information regarding the suggested methodology's efficacy and efficiency. Performance measures such as specificity, sensitivity, and F1-score are used to understand the link between the recommended approach's input and output values. The F1-score, specificity, and sensitivity may all be calculated using this formula. Furthermore, accuracy is one of the useful assessment metrics utilized to determine the efficacy of the suggested Alzheimer's disease categorization approach. The simplest intuitive performance metric is accuracy, which is just the ratio of all observations to accurately anticipated observations.

7. RESULTS AND DISCUSSION

Filtering strategies based on IG are one of the most prevalent ways to select genes. These strategies get a low computational cost and may be employed in large scale applications, microarrays, and other high-dimensional datasets. They estimate each gene’s discriminatory power separately, ignoring interactions between genes. Because these approaches do not consider the relationship between genes, they may be limited in terms of their usefulness. Because it considers gene interactions, one of most effective gene selection techniques is NPO. However, recursive feature elimination cannot eliminate superfluous features since the weight of one of them will be increased proportionally. This paper describes a combined approach that contains an NPO methodology and a strategy for decreasing redundancy in the chosen genes. The following are the procedures of the proposed procedure: i) the IG is used to rank genes, ii) considering expression reduces redundancy, iii) the NPO technique was used to choose the final gene set, and iv) the compressed genes are used to classify AD and non-AD individuals by using SVM.

This part looks at the results of NPO and asks, “Why is NPO effective?” In this situation, there are two key reasons: i) NPO is well-equipped for exploration and exploitation and ii) NPO has a sophisticated system for balancing exploration and exploitation capability. The exploration is used twice: initially when the leaders are introduced at the first meeting, and secondarily when they gather for the periodic meeting. The latter is when families search for anything in the foraging space, or more precisely, the third function. NPO's investigation varies from that of other metaheuristics in that it employs many members of swarms to explore the state space.

Other swarm-based algorithms, on the other hand, frequently employ a special between the global solution and the whole swarming mechanism. Additionally, utilizing a direction variable W, the MRA described in this research push the usual leaders are expected to follow the best leader. This parameter directs them in the direction of better locales, specifically, for their clans to be in a better position. Does the direction variable have two values were used in this study? 1 or -1, depending on whether the fitness values were positive or negative, if these numbers do not fit them in future investigations, depending on their case studies, the researcher may utilize different values. When NPO was used to perform multimodal test functions, its exploration ability was at its peak. NPO was effective in locating the best solutions. Two operators make up the exploitation stage: leadership transition and semicircular distribution. The NPO's local search mechanism is represented by the first operation, while the second operator takes use of the solutions provided by the first two operators. Furthermore, the clans with the second, third, and fourth operators each indicate a different search method, indicating that NPO #Clans have search methods (no. of clans). Families in every clan strive for better sites to relocate to in each iteration (generation) of NPO, resulting in the discovery of leaders and the enhancement of clans within. Even if the leaders do not outperform the finest leader on the planet, they're still there indicating the improvement in NPO, which improves the search process when MRA is used. Figure 6 depicts a generic structure of NPO as well as an overview of the procedures in both stages. Except

for structural factors, such as the number of clans and iterations, NPO appears to be devoid of any regulating parameters. However, these variables do not affect on NPO's search behavior, it can influence the chances of getting the best responses, or, to be more accurate, at the cost of execution time, more families or clans will discover the finest options faster.

7.1. Performance measures

We employed various conventional performance metrics to validate the suggested approach's performance, including classification accuracy, precision, recall, F-score, and the receiver operating characteristic (ROC) curve, which is summarized by the area under the curve (AUC). True positive (TP), true negative (TN), false positive (FP), and false negative (FN) are the four variables that are used to evaluate the model's overall predictive performance, as indicated by (18). It operates by dividing the total number of test samples by the number of successfully identified samples.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{18}$$

The ratio of TP to the total of TP and FN is known as recall, also referred to as sensitivity, as indicated by (19).

$$Recall = \frac{TP}{TP+FN} \tag{19}$$

Precision, also known as positive predictive value is another performance indicator employed in this study, as seen in (20).

$$Precision = \frac{TP}{TP+FP} \tag{20}$$

F-Measure is used to neutralize the bias in precision and recall, as demonstrated in (21) because it takes the harmonic mean of both accuracy and recall.

$$Fscore = 2\left(\frac{Precision \times Recall}{Precision + Recall}\right) \tag{21}$$

7.2. Experimental results

We applied the technique to five open data sets to look for the most discriminatory genes, and they compare the strategy to other well-known methods. To make a comparison between the results, support vector machines are used in all gene selection strategies for categorization. The accuracy, sensitivity, and specificity of a classifier are the measures used to evaluate different approaches. The proportion of correctly identified samples as a whole is known as accuracy. Table 2 summarizes the AD data's description depending on the number of original genes and those selected by IG and NPO. The majority of genes in the original dataset are not relevant at all to class label prediction because they are not used subsequently. As a result, the gene selection strategies described above yield less informative genes, allowing for better classification performance by disregarding irrelevant genes.

Table 2. The conclusion of the data selected

Techniques	Samples	Genes of Origin	Genes Selected
IG	569	16382	90
NPO			30
IG	31	22283	300
NPO			250
IG	161	54675	1,215
NPO			1,115
IG	195	42179	1,300
NPO			1,100
IG	624	39280	285
NPO			322

Five cross-validations are used to evaluate performance because there are only a few samples. The data is randomly split into five non-overlapping partitions of equal size. The test set is split into four parts, with each part used as a testing tool in turn. The remaining 1/5 of the data is used for training, while 4/5 is utilized to test. The process is repeated five times. The arithmetic mean from partitioning mistakes is the total error.

In conclusion, the results of utilizing the suggested technique are presented and other conventional methods based on classifier accuracy, sensitivity, and specificity. The accuracy, sensitivity, and specificity of different strategies are displayed in Tables 3 to 5. These tables list each technique's average precision, sensitivity, and accuracy, respectively. These tables also indicate the number of genes picked in each strategy. When using a gene selection technique before classifying, the classifier performance is enhanced in all methods except the IG and NPO approaches to choosing informative genes, see Table 1. When conducting NPO gene selection in the third step, the sensitivity and specificity of the classifier are reduced, as demonstrated by Tables 3 and 4. The proportion of genes that were highly expressed in the different approaches is shown in Table 5.

Table 3. Summarizes different gene selection approaches in terms of accuracy

Datasets	W1	W2	W3	W4	W5
GSE63060 and GSE63061	80.5	82.1	83.3	85.6	87.0005
GSE5281	83	84	90	92	96.8944
GSE1297	80	82	86	90	93.0031
GSE132903	81	83	87	88	92.4124
GSE33000	81	82	83	87	92.0961

W1 = No-sel: classifying without the use of genes.
W2 = IG: categorization following the use of IG for gene selection.
W3 = IG-SVM: classification after using SVM and IG for gene selection.
W4 = SVM-NPO: the SVM algorithm and the NPO approach were used to select and classify genes.
W5 = IG-NPO-SVM: classified following the use of the suggested framework for gene selection.

Comparing the results of IG-NPO (using the IG and NPO for gene selection) and SVM-NPO (via SVM algorithm and NPO for gene selection) leads to better performance. Our proposed method (IG-NPO-SVM) exceeds all other techniques in terms of accuracy, sensitivity, and specificity.

Table 4. Based on sensitivity, several gene selection strategies are compared

Dataset	W1	W2	W3	W4	W5
GSE63060 and GSE63061	80.6	81.1	84.3	85.6	87.001
GSE5281	83	84	90	92	95.8944
GSE1297	80	82	86	90	92.0031
GSE132903	81	83	87	88	92.4124
GSE33000	81	82	83	87	91.0961

Table 5. Different gene selection approaches are compared based on specificity

Dataset	W1	W2	W3	W4	W5
GSE63060 and GSE63061	81.16	82.01	83.03	84.06	86.001
GSE5281	82.03	83.05	88.01	90.01	94.0944
GSE1297	80.01	82.93	85.20	86.25	91.1031
GSE132903	81.91	83.81	87.70	88.62	91.4124
GSE33000	80.95	82.82	83.91	87.01	90.0961

Table 6 summarizes the findings from the AD dataset that has been published. When we compare the outcomes of our suggested framework to the findings in Table 6, our technique is more accurate than others. Table 7. Shows the average categorization accuracy and a particular gene's findings. When compared to other gene selection methods, the other gene selection approach employing NPO works well with the SVM model.

Table 6. Some articles have published their findings for the AD dataset

Reference	[7]	[5]	[17]	[2]	[13]	[35]	[31]
Accuracy	92.4 %	90.6 %	91.0 %	73.0 %	87.0 %	93.0 %	85.0 %
Number of Select Genes	565	15	500	150	100	200	3734

Table 7. On the AD dataset, the results of average accuracy and a specified gene

Proposed Models	Dataset	Accuracy (%)	Selected Gene
SVM+NPO	GSE1297	93.0031	250
	GSE5281	96.8944	1,115
	GSE63060 and GSE63061	87.0005	30
	GSE132903	92.4124	1,100
	GSE33000	92.0961	322

Table 8 illustrates the implementation of several algorithms such (Naïve Bayes, SMO (SVM), k-NN, and RF using Weka, and the suggested system showed the accuracy better than the previous algorithms. Table 9 shows the experimental results of the suggested technique, including the sample size of each dataset, as well as the training and test sample sizes. Also included are classification accuracy, precision, recall, F-score, and AUC. On five of the datasets, the suggested technique achieves a flawless classification performance of 1.00. The suggested technique achieved classification accuracy of 93.0031, 96.8944, 87.0005, 92.4124, and 92.0961 on the remaining five datasets, respectively.

We calculated the classification accuracy to assess the suggested method's performance. On AD datasets, Figure 9 depicts the suggested method's classification accuracy. In a classification problem, the ROC curve is commonly used to evaluate the model's performance. It operates by calculating the AUC at different thresholds. It is one of the measures used to assess a classifier's performance, and it has a high level of tolerance when it comes to categorizing data with lower class imbalance.

Table 8. Accuracy comparison with other methods on AD datasets

Methods	Dataset	Accuracy
Naive Bayes	GSE1297	67.7419%
SMO	GSE1297	61.2903%
Ibk (KNN)	GSE1297	61.2903%
RF	GSE1297	74.1935%
Proposed Models SVM+NPO	GSE1297	93.0031
Naive bays	GSE5281	77.6398%
SMO	GSE5281	95.6522%
Ibk (KNN)	GSE5281	84.472$
RF	GSE5281	87.5776%
Proposed Models SVM+NPO	GSE5281	96.8944
Naive bays	SE63060 and GSE63061	70%
SMO	GSE63060 and GSE63061	87.75%
Ibk (KNN)	GSE63060 and GSE63061	60%
RF	GSE63060 and GSE63061	81.25%
Proposed Models SVM+NPO	GSE63060 and GSE63061	87.0005
Naive bays	GSE132903	4.1026%
SMO	GSE132903	91.7494%
Ibk (KNN)	GSE132903	82.0513%
RF	GSE132903	81.0256%
Proposed Models SVM+NPO	GSE132903	92.4124
Naive bays	GSE33000	4.8077%
SMO	GSE33000	91.6667%
Ibk (KNN)	GSE33000	87.5256%
RF	GSE33000	82.5321%
Proposed Models SVM+NPO	GSE33000	92.0961

Table 9. Experimental results on all datasets

Datasets	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)	AUC (%)
GSE1297	93.0031	93.0031	93.0031	93.0031	93.0031
GSE5281	96.8944	96.8944	96.8944	96.8944	96.8944
GSE63060 and GSE63061	87.0005	87.0005	87.0005	87.0005	87.0005
GSE132903	92.4124	92.4124	92.4124	92.4124	92.4124
GSE33000	92.0961	92.0961	92.0961	92.0961	92.0961

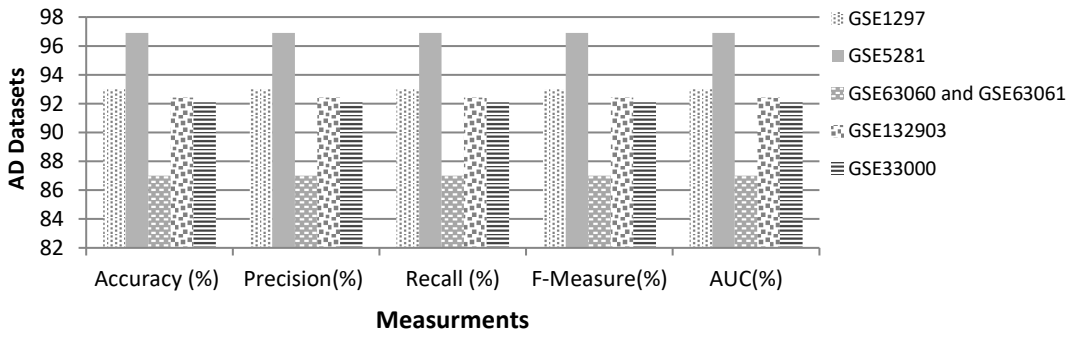

Figure 9. Classification accuracy on five microarray AD datasets

8. CONCLUSION

Discriminating between AD sufferers and healthy people subjects using gene expression from blood and brain is an important task. The proposed SVM model based on the NPO algorithm is a promising tool for diagnosing AD. The ability of the model to predict patients with early stages of AD could be of great clinical value in the future, helping to diagnose and monitor the disease progression. The proposed SVM model is a promising tool for the diagnosis of AD, and more studies are needed to validate its efficacy. In this study, 5-fold cross-validation was used, and a minor variance in prediction error may be obtained, demonstrating the method's robustness. The performance of the model constructed with IG, NPO, and SVM was evaluated by using numerous statistical performance assessors and compared. The best model with excellent accuracy, sensitivity, and specificity, discriminate between Alzheimer's patients and healthy controls was chosen based on the performance of the IG-based classification model.

The study's most significant contribution includes development and use of a unique metaheuristic based on the nature of nomadic populations' swarm-like behavior for unconstrained (normal and large-scale) optimization problems. Another contribution of this research is the use of a new multi-swarm technique. NPO is based on nomads' movement searching for food sources. This MRA is proposed as a technique of clan communication and balance exploration and exploitation.

This MRA is a cooperative multi-swarm approach based on how individuals communicate in groups. A MRA employs radical information exchange among the leaders of the swarms, who are represented by the best local solutions. Furthermore, NPO requires only two structural parameters: the number of clan leaders and the number of families. In the future, the suggested NPO might be improved to address concerns relating to constrained optimization. The includes programming language features that can be used to build programs for various domains, including artificial intelligence and machine learning. It can also be used to solve machine learning optimization issues like finding the most significant characteristics in classifying/grouping issues and training an artificial neural network (i.e. modifying the neural networks' weights) (i.e. the feature selection problem). Added variants will be produced in the future, including multiple objective NPOs.

REFERENCES

[1] C. M. Karch, C. Cruchaga, and A. M. Goate, "Alzheimer's disease genetics: From the bench to the clinic," *Neuron*, vol. 83, no. 1, pp. 11–26, Jul. 2014, doi: 10.1016/j.neuron.2014.05.041.

[2] S. J. Mohammed, "A proposed Alzheimer's disease diagnosing system based on clustering and segmentation techniques," *Engineering and Technology*, vol. 36, no. 2, pp. 160–165, 2018, doi: 10.30684/etj.36.2B.16.

[3] S. Gauthier *et al.*, "Management of behavioral problems in Alzheimer's disease," *International Psychogeriatrics*, vol. 22, no. 3, pp. 346–372, May 2010, doi: 10.1017/S1041610209991505.

[4] F. I. Yasser, B. H. Abd, and S. M. Abbas, "Detection of confusion behavior using a facial expression based on different classification algorithms," *Engineering and Technology Journal*, vol. 39, no. 02, pp. 316–325, 2021.

[5] S. Cahill, M. Pierce, P. Werner, A. Darley, and A. Bobersky, "A systematic review of the public's knowledge and understanding of Alzheimer's disease and dementia," *Alzheimer Disease & Associated Disorders*, vol. 29, no. 3, pp. 255–275, Jul. 2015, doi: 10.1097/WAD.0000000000000102.

[6] H. M. Ahmed and M. A. Taha, "A brief survey on modern Iris feature extraction methods," *Engineering and Technology Journal*, vol. 39, no. 01, pp. 123–129, 2021.

[7] A. Bir-Jmel, S. M. Douiri, and S. Elbernoussi, "Gene selection via a new hybrid ant colony optimization algorithm for cancer classification in high-dimensional data," *Computational and Mathematical Methods in Medicine*, vol. 2019, pp. 1–20, Oct. 2019, doi: 10.1155/2019/7828590.

[8] C. Park, J. Ha, and S. Park, "Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset," *Expert Systems with Applications*, vol. 140, Feb. 2020, doi: 10.1016/j.eswa.2019.112873.

[9] K. Sekaran and M. Sudha, "Diagnostic gene biomarker selection for Alzheimer's classification using machine learning," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 12, pp. 2348–2352, 2019, doi: 10.35940/ijitee.L3372.1081219.

[10] C. D. A. Vanitha, D. Devaraj, and M. Venkatesulu, "Gene expression data classification using support vector machine and mutual information-based gene selection," *Procedia Computer Science*, vol. 47, no. C, pp. 13–21, 2015, doi: 10.1016/j.procs.2015.03.178.

[11] L. Scheubert, M. Luštrek, R. Schmidt, D. Repsilber, and G. Fuellen, "Tissue-based Alzheimer gene expression markers-comparison of multiple machine learning approaches and investigation of redundancy in small biomarker sets," *BMC Bioinformatics*, vol. 13, no. 1, 2012, doi: 10.1186/1471-2105-13-266.

[12] N. Jameel and H. S. Abdullah, "Intelligent feature selection methods: A survey," *Engineering and Technology Journal*, vol. 39, no. 1B, pp. 175–183, Mar. 2021, doi: 10.30684/etj.v39i1B.1623.

[13] S. T. Ahmed and S. M. Kadhem, "Early Alzheimer's disease detection using different techniques based on microarray data: A review," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 18, no. 04, pp. 106–126, Mar. 2022, doi: 10.3991/ijoe.v18i04.27133.

[14] M. Babu and K. Sarkar, "A comparative study of gene selection methods for cancer classification using microarray data," in *2016 Second International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, Sep. 2016, pp. 204–211, doi: 10.1109/ICRCICN.2016.7813657.

[15] G. B. Whitworth, "An introduction to microarray data analysis and visualization," in *Methods in Enzymology*, 2nd ed., vol. 470, no. C, Elsevier Inc., 2010, pp. 19–50.

[16] S. M. M. Osowski, "Data mining methods for gene selection on the basis of gene expression arrays," *International Journal of Applied Mathematics and Computer Science*, vol. 24, no. 3, pp. 657–668, Sep. 2014, doi: 10.2478/amcs-2014-0048.
[17] L. Dey and A. Mukhopadhyay, "Microarray gene expression data clustering using PSO based K-means algorithm," *International Journal of Computer Science and its Applications*, vol. 1, no. 1, pp. 232–236, 2005.
[18] R. Sehrawat, *Data preparation*. springer, 2017.
[19] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," *Big Data Analytics*, vol. 1, no. 1, Dec. 2016, doi: 10.1186/s41044-016-0014-0.
[20] A. K. Hassan and S. N. Mohammed, "A novel facial emotion recognition scheme based on graph mining," *Defence Technology*, vol. 16, no. 5, pp. 1062–1072, Oct. 2020, doi: 10.1016/j.dt.2019.12.006.
[21] R. Lin, C. Wang, and C.-W. Tung, "A machine learning classifier for predicting stable MCI patients using gene biomarkers," *International Journal of Environmental Research and Public Health*, vol. 19, no. 8, Apr. 2022, doi: 10.3390/ijerph19084839.
[22] R. Díaz-Uriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, no. 1, Dec. 2006, doi: 10.1186/1471-2105-7-3.
[23] J. Zahoor and K. Zafar, "Classification of microarray gene expression data using an infiltration tactics optimization (ITO) algorithm," *Genes*, vol. 11, no. 7, Jul. 2020, doi: 10.3390/genes11070819.
[24] B. Sahu, S. Dehuri, and A. Jagadev, "A study on the relevance of feature selection methods in microarray data," *The Open Bioinformatics Journal*, vol. 11, no. 1, pp. 117–139, Jul. 2018, doi: 10.2174/1875036201811010117.
[25] C. A. Finney, F. Delerue, and A. Shvetcov, "Artificial intelligence-driven meta-analysis of brain gene expression data identifies novel gene candidates in Alzheimer's Disease," *medRxiv*, 2022, doi: 10.1101/2022.02.02.22270347.
[26] A. A. Soofi and A. Awan, "Classification techniques in machine learning: Applications and issues," *Journal of Basic & Applied Sciences*, vol. 13, pp. 459–465, Jan. 2017, doi: 10.6000/1927-5129.2017.13.76.
[27] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *Journal of Biomedical Informatics*, vol. 85, pp. 189–203, Sep. 2018, doi: 10.1016/j.jbi.2018.07.014.
[28] H. Ali Abdulmohsin, H. B. A. Wahab, and A. M. J. A. Hossen, "A new hybrid feature selection method using T-test and fitness function," *Computers, Materials & Continua*, vol. 68, no. 3, pp. 3997–4016, 2021, doi: 10.32604/cmc.2021.014840.
[29] W. M. S. Abedi, A. T. Sadiq, and I. N. Ibraheem, "Modified deep learning method for body postures recognition," *International Journal of Advanced Science and Technology*, vol. 29, no. 2, pp. 3830–3841, 2020.
[30] M. N. K. P. and T. Thiyagarajan, "Feature selection using efficient fusion of fisher score and greedy searching for Alzheimer's classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 4993–5006, Sep. 2022, doi: 10.1016/j.jksuci.2020.12.009.
[31] S. Q. Salih and A. A. Alsewari, "A new algorithm for normal and large-scale optimization problems: Nomadic people optimizer," *Neural Computing and Applications*, vol. 32, no. 14, pp. 10359–10386, Jul. 2020, doi: 10.1007/s00521-019-04575-1.
[32] Y. Tian and Z. Qi, "Review on: Twin support vector machines," *Annals of Data Science*, vol. 1, no. 2, pp. 253–277, Jun. 2014, doi: 10.1007/s40745-014-0018-4.
[33] H. A. Jeiad, "Indian number handwriting features extraction and classification using multi-class SVM," *Engineering and Technology Journal*, vol. 36, no. 1, pp. 33–40, 2018.
[34] S. T. Ahmed and S. M. Kadhem, "Using machine learning via deep learning algorithms to diagnose the lung disease based on chest imaging: A survey," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 15, no. 16, pp. 95–112, Aug. 2021, doi: 10.3991/ijim.v15i16.24191.
[35] N. Mahendran, P. M. D. R. Vincent, K. Srinivasan, and C.-Y. Chang, "Improving the classification of Alzheimer's disease using hybrid gene selection pipeline and deep learning," *Frontiers in Genetics*, vol. 12, pp. 1–13, Nov. 2021, doi: 10.3389/fgene.2021.784814.

BIOGRAPHIES OF AUTHORS

Shaymaa Taha Ahmed M.Sc. (2015) in (India), currently a postgraduate student studying Ph.D. in the Department of Computer Sciences, The University of Technology, Baghdad, Iraq. Affiliation in University of Diyala Department Computer Science, College: Basic of Education Specialization: Computer Science\Information System. Research interests in cloud computing, deep learning, machine learning, AI, and data mining. She can be contacted at: Shaymaa.taha.ahmed@basicedu.uodiyala.edu.iq and mrs.sh.ta.ah@gmail.com.

Suhad Malallah Kadhem Ph.D. in Computer Sciences/2003/Computer Science Department/University of Technology. Scientific Specialization: Natural Language Processing. Scientific Title: Assistant Professor/20012. Scientific research interest: natural language processing, machine translation, artificial intelligence, information security, information hiding, machine learning, deep learning and she can be contacted at email: 110102@uotechnology.edu.iq.