*Intelligent Automation & Soft Computing*
DOI: 10.32604/iasc.2023.029127
***Article***

Tech Science Press

# An Intelligent Hybrid Ensemble Gene Selection Model for Autism Using DNN

**G. Anurekha[*] and P. Geetha**

Department of Information Science and Technology, College of Engineering, Anna University, Chennai, 600025, India
*Corresponding Author: G. Anurekha. Email: anu21rekha@gmail.com

Received: 25 February 2022; Accepted: 05 May 2022

**Abstract:** Autism Spectrum Disorder (ASD) is a complicated neurodevelopmental disorder that is often identified in toddlers. The microarray data is used as a diagnostic tool to identify the genetics of the disorder. However, microarray data is large and has a high volume. Consequently, it suffers from the problem of dimensionality. In microarray data, the sample size and variance of the gene expression will lead to overfitting and misclassification. Identifying the autism gene (feature) subset from microarray data is an important and challenging research area. It has to be efficiently addressed to improve gene feature selection and classification. To overcome the challenges, a novel Intelligent Hybrid Ensemble Gene Selection (IHEGS) model is proposed in this paper. The proposed model integrates the intelligence of different feature selection techniques over the data partitions. In this model, the initial gene selection is carried out by data perturbation, and the final autism gene subset is obtained by functional perturbation, which reduces the problem of dimensionality in microarray data. The functional perturbation module employs three meta-heuristic swarm intelligence-based techniques for gene selection. The obtained gene subset is validated by the Deep Neural Network (DNN) model. The proposed model is implemented using python with six National Center for Biotechnology Information (NCBI) gene expression datasets. From the comparative study with other existing state-of-the-art systems, the proposed model provides stable results in terms of feature selection and classification accuracy.

**Keywords:** Autism spectrum disorder; feature selection; ensemble gene selection; microarray; gene expression; deep neural network; meta-heuristic

## 1 Introduction

ASD is a complicated neurodevelopmental disorder that is often identified in toddlers. Children with ASD suffer from interminable difficulties such as social interaction, speech, non-verbal communication, and restricted or repetitive actions [1]. The prevalence of the disease is increasing worldwide, and the estimated probability of occurrence in the United States is one in 44 children [2]. The probability of behavioral improvement has been drastically increased when ASD is diagnosed in its earlier stages [3]. Microarrays are involved in analyzing the gene expression of the autism-affected patients' from its control, contributing to novel insights [4]. In the microarray dataset, genes are the features. The microarray data is voluminous and has a high dimension, which suffers from a problem of dimensionality. Generally, the number of samples of microarray data is in the range of hundreds, and the number of corresponding genes is ten times greater. Thus, the gene selection process is both a time-consuming and challenging issue because of the huge imbalance of data in terms of feature selection and classification. An efficient gene selection and classification model is required to select the appropriate gene features and classify them to effectively detect ASD. Although, from the literature, there are several feature selection and classification techniques [5,6] proposed in the past, the selecting and classification of relevant genes from high dimensional microarray data is an important and challenging issue for the early diagnosis of ASD efficiently.

In the view of an epidemiologist, the subset of genes with high classification accuracy contributes insights about the disease, but techniques that could produce stable and robust result are desired. Meanwhile, gene selection strategies that are used for cancer gene identification cannot be used for autism as the variance of gene expression is high in corresponding to the same class [7]. Hence, it is necessary to focus research on several ensemble techniques for stable and robust gene selection [8]. An ensemble-based feature selection model consists of three major types: Data perturbation, functional perturbation, and hybrid ensemble [9]. Applying a single feature selection model to all the sample data is termed "data perturbation". Employing various feature selection techniques on the same sample of data is termed "functional perturbation" whereas, applying different feature selection techniques to the different sampled versions is termed a hybrid ensemble. Motivated by all these observations, a hybrid ensemble-based gene selection model has been proposed for efficient gene feature selection and classification for a better and more effective diagnosis of ASD.

The remainder of the manuscript is organized as follows: Section 2 briefly outlines the related works. Section 3 describes the proposed model. Section 4 presents the experimental findings and discussion. Finally, Section 5 details the conclusion and the future scope.

## 2 Related Works

Discussion of the existing approaches from the literature can complement the proposed methodology. A hybrid feature selection method in [10] for ASD is designed to analyze Single Nucleoid Polymorphism (SNP) to classify diseased SNPs from their control. In this method, Conditional Mutual Information Maximization (CMIM) with Support Vector Machine–Recursive Feature Elimination (SVM-RFE) is used for SNP selection. A combination of statistical measures and wrapper approach was proposed in [11]. In this model, three statistical measures are employed for the initial selection of genes, namely two-sample t-test, feature correlation with class, and Wilcoxon rank-sum. Later, the wrapper approach incorporating binary Particle Swarm Optimization (PSO) with Support Vector Machine (SVM) is utilized for final gene optimization and classification. The accuracy obtained by this approach is 92.1%. A hybrid deep learning model [12] is proposed with information gain ratio for dimension reduction and Deep Belief Network (DBN) based on Gaussian Restricted Boltzmann Machine (GRBM) for classification of samples from disease to its control. K. Sekaran and M. Sudha proposed a gene selection strategy in [13] for recognizing the autistic gene in the genomic microarray data. The optimum features are selected by the signal-to-noise ratio holding a logistic sigmoid function, the Hilbert–Schmidt Independence Criterion Lasso, along with a regularized genetic algorithm employed for gene classification. The ASD's prediction accuracy rate is better using this method than compared with the other existing systems.

Various researchers from the literature have experimented with ensemble models for feature selection in recent years. In [14] Bakrutheen et al. -Pour proposed multiple filter such as relieff and Fast Correlation Based Filter (FCBF), are used to reduce the dimensionality of the initial set of genes from microarray data. An improved binary gravitational search algorithm is used in their approach to optimize the selected genes. In this ensemble model, the authors used a simple voting strategy for selecting top n-ranked genes. Classifiers like Decision Tree (DT), K-Nearest Neighbor (KNN), and SVM are used for the performance evaluation of selected optimal genes. A hybrid ensemble-filter wrapper feature selection approach is described in [15] for medical data classification. In this model, authors have adopted a subset-based filter approach such as Correlation-based Feature Selection (CFS) and Consistency (CONS) and a rank-based filter approach such as chi-square test, information gain, and relieff for initial gene selection. The proposed system uses maximum cardinality to select the top n-rank feature subset. A sequential feature selection model is employed to obtain the optimal features.

From the literature, it is clear that many researchers have focused on gene selection based on improving classification accuracy rather than obtaining stable and robust results. Furthermore, microarray data are a highly imbalanced dataset that may can lead to overfitting problems, and high variance in gene expression values leads to misclassification. Hence an intelligent model is necessary to mitigate the problems identified by the existing approaches. In this paper, a novel Intelligent Hybrid Ensemble Gene Selection (IHEGS) model is proposed. The proposed IHEGS model classifies the best performing genes with high rank from multiple feature selection techniques and they are aggregated moving towards the final gene subset thus excluding the poor performing and redundant genes. For that reason, it reduces the dimensionality problem considerably. Hence, towards the motivation of finding a stable and robust autism gene subset from high-dimensional microarray data, the following contributions are carried out in this work.

- An Intelligent Hybrid Ensemble Gene Selection (IHEGS) model is proposed.
- The data perturbation module employs the Standardized Cost pruning Extra Tree Classifier (SCP-ETC) for initial gene subset selection and feature cut-off process.
- Three swarm intelligence-based algorithms, Confidence Interval-Based Flamingo Search Algorithm (CI-FSA), Hinge Loss-Based Garra-Rufa Optimization (HL-GRO), Euclidean Distance-Based Seagull Optimization Algorithm (ED-SOA), are proposed in the functional perturbation module for selecting the final gene subset.

## 3 Proposed Model

The proposed IHEGS model is depicted in Fig. 1. The key components of the proposed model are data partitioning, data similarity, the IHEGS module which in turn contains two submodules: The data perturbation module, and functional perturbation module, and finally the validation module.

### 3.1 Data Partitioning

To begin with, the microarray data is shuffled and divided into n partitions as $Dp_1$, $Dp_2$, $Dp_n$ by employing stratified subsampling. Through this practice, the data in each partition is balanced. Each partition contains *m* rows of samples and *n* columns of genes which are denoted by Eq. (1).

$$Dp^T_{M\times N} = [Dp^1_{m\times n}, Dp^2_{m\times n}, Dp^3_{m\times n} \ldots\ldots..Dp^t_{m\times n}]. \tag{1}$$

### 3.2 Data Similarity

This module emphasizes redundant data partition removal. A one-way ANOVA test [16] was utilized to detect the similarity among the data groups. For every group, a null hypothesis and an alternative hypothesis are formulated as given in Eq. (2)

$$\begin{aligned} &H_0 : The\ \ two\ \ groups\ \ mean\ \ are\ \ significantly\ \ same \\ &H_1 : The\ \ two\ \ groups\ \ mean\ \ are\ \ not\ \ significantly\ \ same \end{aligned} \tag{2}$$

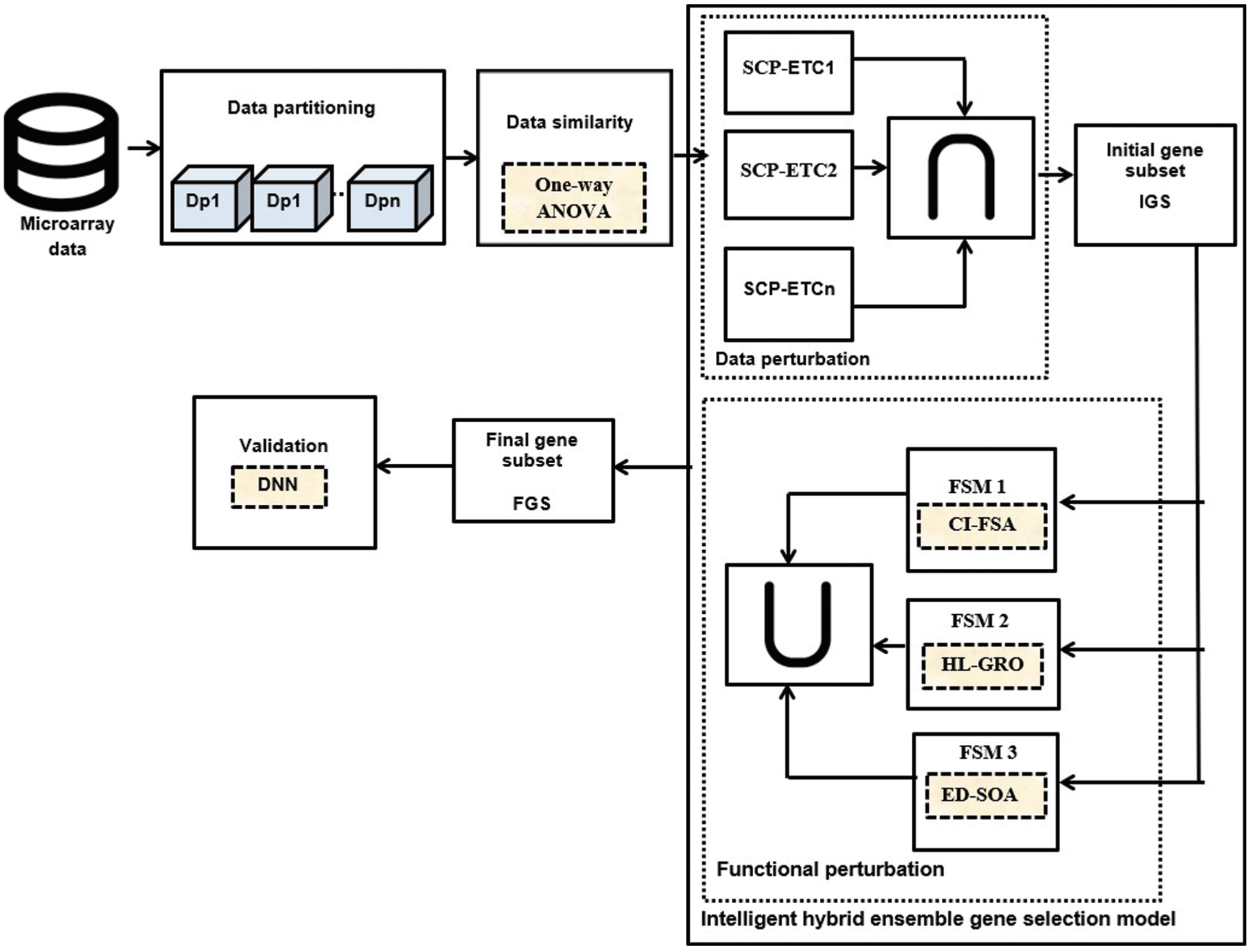


**Figure 1:** Proposed model

The one-way ANOVA test was carried out with a significance of 0.05. The significance of the hypothesis proves 95% confidence interval on the F-value. The groups with a similar spread of data are eliminated, and the obtained data frame is detailed in Eq. (3).

$$Dp^T_{M\times N} = \Omega''[Dp^1_{m\times n},\ Dp^2_{m\times n},\ Dp^3_{m\times n} \ldots\ldots.Dp^t_{m\times n}] \tag{3}$$

where $\Omega''$ signifies the similarity measures of the sampled data.

### *3.3 Intelligent Hybrid Ensemble Gene Selection*

The objective of the proposed model is to predict the autism gene subset from high-dimensional microarray data. The proposed IHEGS model is formed by merging data perturbation and functional perturbation modules. The sampled data partition $Dp^T_{M\times N} = \Omega''[Dp^1_{m\times n},\ Dp^2_{m\times n},\ Dp^3_{m\times n} \ldots\ldots.Dp^t_{m\times n}]$ is given as input. Unique features namely $[At_1,\ At_2,\ At_3,\ \ldots.At_n]$ are encompassed by IHEGS for each data sample. The highly relevant genes are identified by the Gene Selection (GS) score $[\gamma_1,\ \gamma_2,\ \gamma_3,\ \gamma_4,\ \gamma_5,$

$\ldots.\gamma_n$]. Therefore, maximizing the GS score to predict the autism gene subset is the overall objective function. The formulation is given in Eq. (4).

$$Obj = [\gamma_1 At_1 + \gamma_2 At_2 + \gamma_3 At_3 + \gamma_4 At_4 + \gamma_5 At_5 + \ldots. + \gamma_n At_n] \tag{4}$$

### *3.3.1 Data Perturbation*

The selection of the initial gene subset is achieved by the data perturbation module. In this module, the same feature selection strategy is applied to all data samples and aggregated by intersection operation. Which in turn assists in the recognition of a predictive subset of appropriate genes with higher classification accuracy. Small perturbations are prioritized by the data perturbation module. It is functional for avoiding instability problems when selecting the initial gene subset. To select an initial gene subset in the data perturbation module, SCP-ETC is proposed. SCP-ETC is made up of decision tree stumps. Each decision tree in the Extra Trees Forest is created from the training data. Then, the trees are arbitrarily chosen as the gene feature subset. The best feature to separate the data is determined by each tree as per calculated entropy at every test node. A slew of an independent decision tree is produced by the random selection of features. The importance of a feature is measured by the entropy of the genes. It is presented in Eq. (5).

$$E(d) = \sum\nolimits_{i=1}^{n} -\hbar_i \log_2(\hbar_i)^{\cdot} \tag{5}$$

$$\Re(E,\ d) = IG - E. \tag{6}$$

$$\text{where}\, IG = -\frac{\hbar}{\hbar + \mathrm{N}}\log_2\left(\frac{\hbar}{\hbar + \mathrm{N}}\right) - \frac{\mathrm{N}}{\hbar + \mathrm{N}}\log_2\left(\frac{\mathrm{N}}{\hbar + \mathrm{N}}\right). \tag{7}$$

where $\Re(E,\ d)$ signifies the gain, based on the entropy of genes, $IG$ indicates the information gain, P and N signifies the data's binary class. To avoid overfitting decision tree, a standardized cost pruning is used to eliminate tree fragments such as branches and roots. It also promotes the healthy growth of the decision tree. The standardized cost pruning is calculated as shown in Eq. (8).

$$\aleph_{\vartheta}(T) = \aleph(T) + \vartheta|T|. \tag{8}$$

where $\aleph_{\vartheta}(T)$ signifies the tree's cost complexity measure, $|T|$ indicates the number of terminal nodes, and $\vartheta$ symbolizes the complexity parameters. And, the small subsets of genes are obtained by the data perturbation. After maximum iterations, those genes are saved in the form of data frames by grouping them in a ranking manner. The mean GS ranking score $\overline{\gamma}(At_i)$ is calculated as shown in Eq. (9)

$$\overline{\gamma}(At_i) = \frac{1}{z}\sum\nolimits_{z=1}^{Z} \gamma_{z,i}. \tag{9}$$

$$IGS = \bigcap\nolimits_{z=1}^{Z} F_{z,i}. \tag{10}$$

where $F_{z,i}$ signifies the primary gene subset and $IGS$ is the initial gene subset. The common features that are selected across data partitions by the data perturbation module are retained by the intersection operation, and the uncommon features are removed.

### *3.3.2 Functional Perturbation*

The selection of a final gene subset of the ASD is identified by the functional perturbation module. The three proposed swarm intelligence techniques (CI-FSA, HL-GRO, and ED-SOA) are applied to the initial gene subset *IGS* selected by the data perturbation module.

*CI-FSA (Feature Selection Model 1)*

Feature selection model 1 employed in the proposed IHEGS is the Confidence Interval-based Flamingo Search Algorithm (CI-FSA). The Flamingo Search Algorithm (FSA) is inspired by the migratory and foraging behavior of flamingos. The basic understanding and theory of FSA are detailed in [17]. A confidence interval parameter is added to the existing FSA to avoid the fluctuations caused by the randomness of the diffusion factor. The confidence interval-based diffusion factor provides an exact range to control the exploitation and exploration to find the best features. In the proposed CI-FSA flamingos are the features, and fitness value is the best score of the features. The foraging and migratory behavior of the proposed CI-FSA are detailed below.

Foraging Behavior: Initialize Flamingo population P, Migrating flamingos $MP_b$ and Maximum Iteration $I_{max}$. Assume the Flamingo ($Ab_j$ ) has the most food in the $j^{th}$ dimension. Consider $Ab_{ij}$ is the position of the $i^{th}$ and $j^{th}$ dimension of the flamingo population. The uncertainty of each flamingo disturbs the foraging behavior of an individual that encounters an error in the information transmitted. To simulate this, the maximum distance of the flamingo's beak scan in foraging behavior is quantified, as given in Eq. (11).

$$|\Re_1 \times Ab_j + \eta_2 \times At_{ij}| \tag{11}$$

where, $\eta_2$ is a random number between −1 or 1, where $\Re_1$ is a random number that follows a standard normal distribution. The scanning range during the beak behavior is retained within a range that is given by Eq. (12).

$$\Re_2 \times |\Re_1 \times Ab_j + \eta_2 \times At_{ij}|. \tag{12}$$

where, $\Re_2$ denotes a random number that follows the standard normal distribution. Assume food is abundant in the population $Ab_j$, the distance between the flamingos is changed and the traveling is computed as $\eta_1 \times Ab_j$, where $\eta_1$ is a random number of −1 or 1 represents the range of search. To sum up, the moving step of flamingos foraging in the $t^{th}$ iteration is the scanning range of the flamingo's beak and the moving distance of their feet given by Eq. (13).

$$b^t_{ij} = \eta_1 \times Ab^i_j + \Re_2 \times |\Re_1 \times Ab^i_j + \eta_2 \times At^t_{ij}|. \tag{13}$$

The equation for updating the location of flamingos' foraging behavior is given in Eq. (14).

$$At^{t+1}_{ij} = (At^t_{ij} + \eta_1 \times Ab^t_j + \Re_2 \times |\Re_1 \times Ab^t_j + \eta_2 \times At^t_{ij}|)/k. \tag{14}$$

$At^{t+1}_{ij}$ represents the position of the $i^{th}$ flamingos in the $j^{th}$ dimension of the population in the $(t+1)$ *th* iteration, $At^t_{ij}$ represents the position of the $i^{th}$ flamingos in the $j^{th}$ dimension in the $t$ iteration of the flamingo's population, namely, the position of the flamingo's feet, $Ab^t_j$ represents the $j^{th}$ dimension position of the flamingos with the best fitness in the population in the $t$ iteration, $k$ is the diffusion factor calculated as given in Eq. (15). That follows a confidence interval with chi-square distribution of $n$ degrees of freedom.

$$k = \left[At^t_{ij} - Z_{\beta/2}\frac{S.D}{\sqrt{n}},\ At^t_{ij} + Z_{\beta/2}\frac{S.D}{\sqrt{n}}\right] \tag{15}$$

$Z_{\beta/2}$ denotes the confidence interval value, and $S.D$ denotes the standard deviation of the sample. It is used to increase the size of the flamingos' foraging range to simulate the chance of individual selection in nature increase their global merit-seeking ability. $\Re_1 = N(0,\ 1)$ and $\Re_2 = N(0,\ 1)$ are random numbers that follow a standard normal distribution, $\eta_1$ and $\eta_2$ are randomized by −1 or 1.

Migratory Behavior: The flamingos migrate to the next area due to the scarcity of food in the present area. Assume that the location of the food-rich area in the $j^{th}$ dimension is $Ab_j$, the formula for the migration of the flamingo's population is as given in Eq. (16).

$$At_{ij}^{t+1} = At_{ij}^{t} + \omega \times \left(Ab_j^t - At_{ij}^t\right) \tag{16}$$

where $\omega = N(0, n)$ represents the random number based on a normal distribution with degrees of freedom $n$ that increases the search space and simulates the randomness of individual behaviors of flamingos in the specific migration process. Finally, when the maximum iteration is reached an optimal solution is obtained. The obtained optimal solution is substituted to the objective function and sorted to find the most relevant features. The most relevant features are framed inside the data frame using $FGS1$ as given in Eq. (17).

$$FGS1 = \mathrm{A}[At_1^d, At_2^d, At_3^d, At_4^d, At_5^d, At_6^d, At_7^d \ldots At_n^d] \tag{17}$$

*HL-GRO (Feature Selection Model 2)*

HL-GRO is employed as the feature selection model 2 in the functional perturbation module. It is a modified version of the existing GRO algorithm. The GRO algorithm is inspired by the behavior of Garra-Rufa fish, commonly used in massage centers for the commercial purposes detailed in [18]. To avoid a high exploration rate and to identify the worst leaders among mobile and follower fish, the proposed work has combined learning rate and hinge loss. Here, the fish is the gene profile. The detailed steps of HL-GRO are explained as follows

Step 1: Initially, for each leg fishes are equally distributed, and one leader is assigned to each group.

Step 2: Based on the optimal convergence speed of the fish, it has to be decided whether it is a leader or follower. Then, the number of leaders is chosen as per the problem complexity. The respective optimal points are selected for the objective functions.

Step 3: A few fish that travel into other group during each iteration offer the best optimal value among all groups. The traveling fish is deemed to be random and is indicated as mobile fish. However, the number of fish is kept constant.

Step 4: For every leader, sub-global fitness is assessed and the global fitness for each iteration is determined.

Step 5: After a random number of iterations, all the fish must follow one leader. First, the number of Followers can be calculated as given in Eq. (18).

$$\Psi = \frac{n - Ab_i^L}{Ab_i^L} \tag{18}$$

where $n$ implies the total particles and $Ab_i^L$ denotes the number of best leaders. Now, for selecting the leaders along with the best group, the objective function for every fish is found and grouped.

Step 6: Using Eqs. (19) and (20), the change in the numbers of mobile and follower fish is calculated for the worst leaders.

$$At_i = \mathrm{int}\left(\wp \times \bar{\lambda} \sum\nolimits_{i=1}^{n} \zeta_i\right) \tag{19}$$

$$At_{ij}^t = Max\left(At_{ij}^{t-1} - At_i\right),\ (0) \tag{20}$$

where $\bar{\lambda}\sum_{i=1}^{n}\zeta_i$ is the hinge loss multiplied with learning rate, $\wp$ implies a maximum number of mobile fish for one group, and $At_{ij}^t$ denotes the number of followers and mobile fish for the ith leader, correspondingly. For the best leader, the changes in the mobile and follower numbers can be calculated as given in Eqs. (21) and (22).

$$Ab_i = \sum\nolimits_2^n At_i \tag{21}$$

$$At_{ij}^t = At_{ij}^{t-1} - Ab_i \tag{22}$$

To conclude, an optimal solution is offered by the maximum iteration attained to substitute in the objective function along with a group for finding the most relevant features. Utilizing HL-GRO (B), the most appropriate features are framed inside the data frame as shown in Eq. (23).

$$FGS2 = \mathrm{B}[At_1^d,\ At_2^d,\ At_3^d,\ At_4^d,\ At_5^d,\ At_6^d,\ At_7^d \ldots At_n^d] \tag{23}$$

*ED-SOA (Feature Selection Model 3)*

The third gene selection strategy used in the functional perturbation module is the ED-SOA. It is based on the former SOA. SOA is a bio-inspired algorithm inspired by the seagull behavior in nature described in [19]. The proposed ED-SOA is an advanced version of the existing SOA. To avoid low convergence problems with the existing SOA algorithm, it should be adapted to find the balanced value that generates an accurate position of the search agents. Hence, a Euclidean distance-based SOA (ED-SOA) is proposed. The ED-SOA controls the slow convergence of the dimensions by handling dimension rates while updating the new position of the search agents. The selection of the features is based on the migration and attacking behaviors of a seagull. The exploration and exploitation approach of ED-SOA are illustrated below in detail.

Step1: Exploration

During the optimization process, three rules are followed to move seagulls from one group to another. The three rules are

Rule 1: To avoid the congestion between two seagulls, a variable is introduced for the evaluation of the new search agent position as shown in Eq. (24).

$$\vec{At_s} = \chi \times (Ab_s^t) \tag{24}$$

wherein $\vec{At_s}$ denotes the search agent doesn't collide, $Ab_s^t$ indicates the search agent's current position, $t$ signifies the current iteration, and $\chi$ symbolizes the movement behavior of the search agent in a provided search space as given in Eq. (25).

$$\chi = \eta_s - (t \times (\eta_s/max_{iter})). \tag{25}$$

where $\eta_s$ is presented for controlling the frequency of employing variable $\chi$ which is linearly reduced as of $\eta_s$ to 0. The value $\eta_s$ is fixed to 2 and $t = 0, 1, 2\ldots..max_{iter}$.

Rule 2: After avoiding the collision, the search agent moves towards the best neighbor that is modeled as in Eq. (26).

$$AT_s = \beta \times (AB_s^t - Ab_s) \tag{26}$$

where $AT_s$ indicates the search agent's position $Ab_s$ towards the best-fit search agent $AB_s^t$ (that is, the fittest seagull). For suitable balancing between exploration and exploitation behavior, $\beta$ is responsible for randomizing and computing as $\beta = 2 \times \eta^2 \times rd$. Wherein, $rd$ implies a random number that lies within the range of [0, 1].

Rule 3: Lastly, utilizing Euclidean distance, the position of the search agent is updated. It is formulated as shown in Eq. (27).

$$D_s = \sqrt{\sum\nolimits_{i=1}^n (\vec{At_{si}} - AT_{si})^2}. \tag{27}$$

wherein, $D_s$ denotes the distance between the search agent and the best-fit ones (i.e., the best seagull whose fitness value is less).

Step 2: Exploitation

As the Seagulls change their speed and position of attack by maintaining their altitude using their wings and weight. While attacking the prey, a spiral movement behavior occurs in the air. This behavior in the *x*, *y*, *and z* planes is explained as given in Eq. (28).

$$\begin{aligned} x' &= \Im \times \cos(\kappa) \\ y' &= \Im \times \sin(\kappa) \\ z' &= \Im \times \kappa \\ r &= u \times e^{\kappa v} \end{aligned} \tag{28}$$

whereas $r$ denotes the radius of every turn of the spiral, $\kappa$ implies random number, $u$ and $v$ are the constants for defining the spiral shape, and $e$ denotes the natural logarithm's base. Thus, the position of the search agent is updated as shown in Eq. (29).

$$\vec{At_s} = (D_s \times x' \times y' \times z') + \overrightarrow{AB}_s \tag{29}$$

where, $\vec{At_s}$ stores the best solutions and updates the other search agent position. The proposed ED-SOA begins with an arbitrarily generated population. Based on the best search agent, the position of the other search agent could be updated throughout the iteration process and $\chi$ is linearly reduced from $\eta_s$ to 0. The variable $\beta$ is tuned for a smooth transition between exploration and exploitation. Finally, an optimum solution is obtained when maximum iteration is reached. Substituting the obtained global optimum in the objective function and group to identify the most appropriate genes. Utilizing ED-SOA ($C$), the most relevant genes are framed in the data frame as shown in Eq. (30).

$$FGS3 = C[At_1^d,\ At_2^d,\ At_3^d,\ At_4^d,\ At_5^d,\ At_6^d,\ At_7^d \ldots At_n^d] \tag{30}$$

Overall, the functional perturbation ensemble component aggregates its inputs via a union operation, as shown in Eq. (31), to obtain the best feature subset, i.e., the final gene subset. The intelligence of various feature selection can be leveraged through the function perturbation cycle, and the obtained final gene set is less vulnerable to overfitting.

$$FGS = \bigcup_{k=1}^{n} FGS_k \tag{31}$$

Lastly, the final gene subset $FGS = [At_1^d,\ At_2^d,\ At_3^d,\ At_4^d,\ At_5^d,\ At_6^d,\ At_7^d \ldots At_n^d]$ is obtained. The Algorithm 1 in Fig. 2 exhibits the proposed IHEGS overall outline.

### *3.4 Validation*

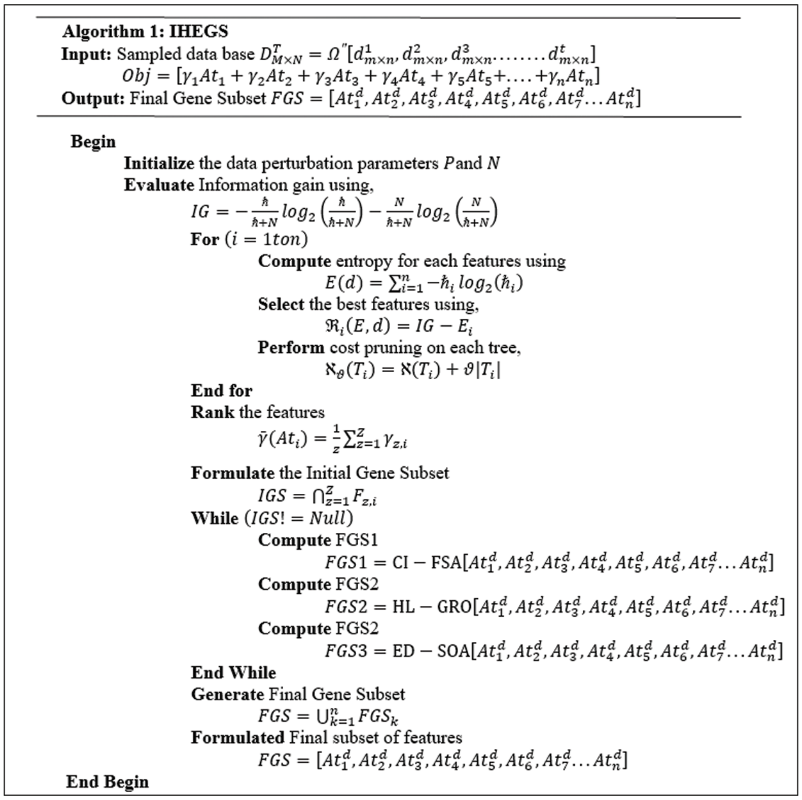
Validation is a process where selected genes are trained and evaluated with testing data to predict ASD genes. Validation is performed using a DNN model. The Fully Connected Neural Network (FCNN) is initiated by the selected features that are given by Eq. (32).

$$\mathrm{I}_1 = [At_1^d,\ At_2^d,\ At_3^d,\ At_4^d,\ At_5^d,\ At_6^d,\ At_7^d \ldots At_n^d] \tag{32}$$

Thereafter, the input features get aggregated with their corresponding weights to get trained. The hidden layer output is given in Eq. (33).

$$\mathrm{I}_2 = Act \sum_{i=1}^{n} w_i At_i + b \tag{33}$$

where $b$ denotes the bias, which is initialized randomly, $w_i$ is the weight of the respective input node, and $Act$ denotes the activation function. The activation function for hidden neurons is leakyRelu and sigmoid for output neurons. Finally**,** the output layer provides the prediction of ASD. Based on the result, the error rate $(\xi_r = \mathrm{I}_i - \hat{\mathrm{I}}_i)$ has been calculated. Now, to minimize the loss, backpropagation has been done, backpropagation updates the new weights to match the targeted result.



```
Algorithm 1: IHEGS
Input: Sampled data base $D^T_{M\times N} = \Omega''[d^1_{m\times n}, d^2_{m\times n}, d^3_{m\times n} \ldots\ldots\ldots d^t_{m\times n}]$
        $Obj = [\gamma_1 At_1 + \gamma_2 At_2 + \gamma_3 At_3 + \gamma_4 At_4 + \gamma_5 At_5 + \ldots. + \gamma_n At_n]$
Output: Final Gene Subset $FGS = [At^d_1, At^d_2, At^d_3, At^d_4, At^d_5, At^d_6, At^d_7 \ldots At^d_n]$

Begin
    Initialize the data perturbation parameters $P$and $N$
    Evaluate Information gain using,
        $IG = -\frac{\hbar}{\hbar+N} log_2\left(\frac{\hbar}{\hbar+N}\right) - \frac{N}{\hbar+N} log_2\left(\frac{N}{\hbar+N}\right)$
        For $(i = 1ton)$
            Compute entropy for each features using
                $E(d) = \sum_{i=1}^{n} -\hbar_i\, log_2(\hbar_i)$
            Select the best features using,
                $\Re_i(E,d) = IG - E_i$
            Perform cost pruning on each tree,
                $\aleph_\vartheta(T_i) = \aleph(T_i) + \vartheta|T_i|$
        End for
        Rank the features
            $\bar{\gamma}(At_i) = \frac{1}{z}\sum_{z=1}^{Z} \gamma_{z,i}$
        Formulate the Initial Gene Subset
            $IGS = \bigcap_{z=1}^{Z} F_{z,i}$
        While $(IGS! = Null)$
            Compute FGS1
                $FGS1 = \mathrm{CI} - \mathrm{FSA}[At^d_1, At^d_2, At^d_3, At^d_4, At^d_5, At^d_6, At^d_7 \ldots At^d_n]$
            Compute FGS2
                $FGS2 = \mathrm{HL} - \mathrm{GRO}[At^d_1, At^d_2, At^d_3, At^d_4, At^d_5, At^d_6, At^d_7 \ldots At^d_n]$
            Compute FGS2
                $FGS3 = \mathrm{ED} - \mathrm{SOA}[At^d_1, At^d_2, At^d_3, At^d_4, At^d_5, At^d_6, At^d_7 \ldots At^d_n]$
        End While
        Generate Final Gene Subset
            $FGS = \bigcup_{k=1}^{n} FGS_k$
        Formulated Final subset of features
            $FGS = [At^d_1, At^d_2, At^d_3, At^d_4, At^d_5, At^d_6, At^d_7 \ldots At^d_n]$
End Begin
```

**Figure 2:** The proposed IHEGS algorithm

## 4 Results and Discussion

The results obtained from the proposed IHEGS model for selecting ASD genes were analyzed and compared with existing techniques.

### *4.1 Dataset*

The six-benchmark microarray dataset from the NCBI with accession numbers GSE25507 (dataset 1) [7], GSE26415 (dataset 2) [20], GSE27044 (dataset 3) [21], GSE109905 (Dataset 4) [22], GSE42133 (Dataset 5) [23], and GSE123302 (Dataset 6) [24] were used for analysis. GSE25507 dataset contains gene expression of 82 autism and 64 controls with 54,614 probes. GSE26415 dataset comprises gene profiles of 21 adults with autism, 21 healthy mothers with autistic children, and 42 controls with 19,194 probes. GSE27044 microarray data holds the DNA methylation profile of autistic children and non-autistic siblings as control with 27,578 CpG sites. GSE109905 dataset comprises genome-wide DNA methylation of 38 ASD patients and 31Controls with 410620 CpG sites. GSE42133 dataset comprises a blood-based gene expression signature relevant to the brain pathophysiology of 87 ASD and 55 control with 47323 genes. GSE123302 dataset comprises 36459 genes with 59 ASD and 92 non-Typically

Developing (non-TD) and 120 Typically Developing (TD) samples. Here gene expression of umbilical cord blood between ASD and TD, and between Non-TD and TD, was analyzed.

### *4.2 Experimental Setup*

The proposed IHEGS model was implemented on the Windows 10 operating system with Intel i7 CPU @ 3.60 GHz processor and 8GB RAM. Python 3.7 with Numpy, Pandas, Seaborn, Matplotlib, Sklearn, Keras libraries are used for implementation.

### *4.3 Evaluation Metrics*

The proposed IHEGS model is evaluated using the metrics classification accuracy, precision, recall, f-measure, error rate, False Positive Rate (FPR), False Negative Rate (FNR), and False Discovery Rate (FDR). The corresponding formulae are given from Eqs. (34) to (41) respectively.

$$\text{Classification} \quad \text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{34}$$

$$\text{Precision} = \quad \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{35}$$

$$\text{Recall} = \quad \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{36}$$

$$\text{F1} \quad \text{score} = 2 * \frac{\text{Precision} \quad * \quad \text{Recall}}{\text{Precision} \quad + \text{Recall}}. \tag{37}$$

$$\text{Error} \quad \text{rate} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{38}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \tag{39}$$

$$\text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}}. \tag{40}$$

$$\text{FDR} = \frac{\text{FP}}{\text{FP} + \text{TP}}. \tag{41}$$

where TP is true positive, TN is true negative, FP is false positive, FN is false negative.

### *4.4 Evaluation of Proposed Ensemble Feature Selection Technique*

Here, for selecting ASD genes, the proposed IHEGS technique is compared based on the Jaccard Index (JI) and Spearman Correlation Coefficient (SCC) on the benchmark datasets. It is analyzed with the existing methods, namely: Feature similarity-based Feature Selection (FSFS) [25], Multi Cluster-based Feature Selection (MCFS) [26], Unsupervised Discriminative Feature Selection (UDFS) [27], PSO [28], and Genetic Algorithm (GA) [29]. Figures Figs. 3 and 4 depict the JI and SCC on the subset of 5% best features for the proposed IHEGS and the existing feature selection algorithm respectively.

The proposed IHEGS technique is efficient since it attains an average JI of 0.858 and SCC of 0.886. But the average JI and SCC of existing algorithms range from 0.51–0.65 and 0.53–0.76 respectively. Therefore, from the results, the proposed IHEGS is signified by the high value of JI and SCC. Hence, it is clear that the proposed IHEGS can select gene features with robustness when compared with the other state-of-the-art techniques. In general, it is observed that ensemble feature selection provides stable and robust results since it depends on the multiple feature selection technique and the utilization of data samples. The proposed IHEGS has handled the data efficiently by adopting data partitioning and the removal of redundant data by the data similarity module. Moreover, the ensemble of data perturbation and functional perturbation has resulted in selecting relevant gene features resulting in highly stable and robust results.

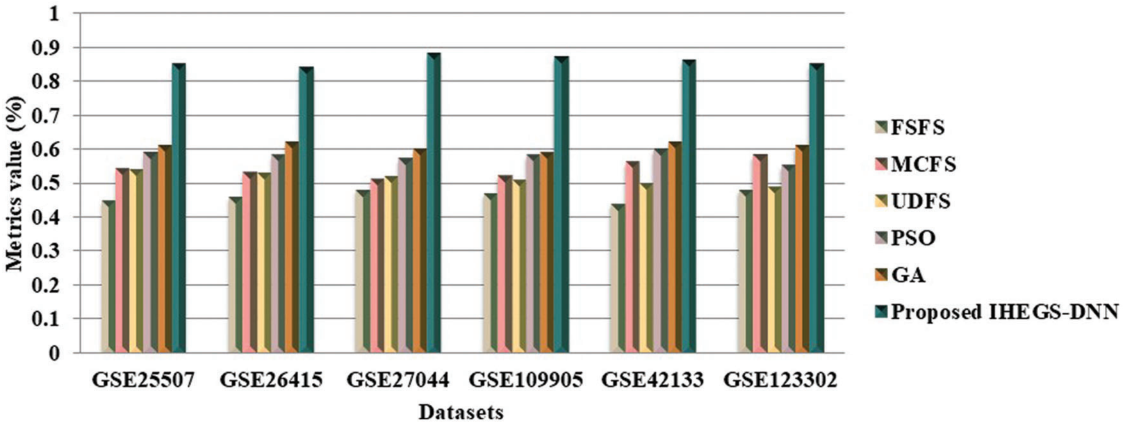


**Figure 3:** Analysis of the robustness of proposed IHEGS based on JI

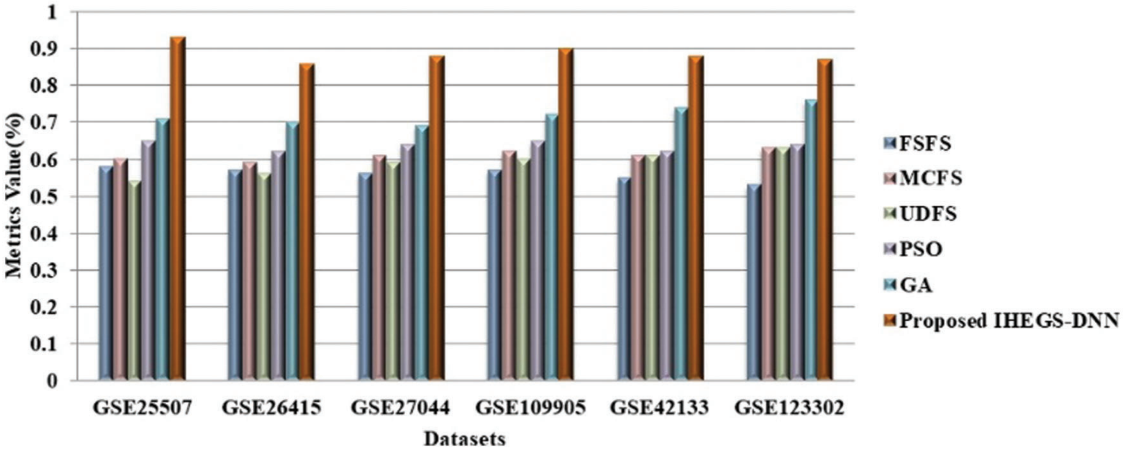


**Figure 4:** Analysis of the robustness of proposed IHEGS based on SCC

### *4.5 Evaluation of DNN-IHEGS for Predicting ASD*

Here, the proposed DNN- IHEGS model for selecting ASD gene is compared based on classification accuracy with the benchmark datasets and contrasted with the existing methodologies such as Naive Bayes-Feature Similarity-based Feature Selection (NB-FSFS) [25], Support Vector Machine-Multi Cluster-based Feature Selection (SVM-MCFS) [26], K-Nearest Neighbors-based Unsupervised Discriminative Feature Selection (KNN-UDFS) [27], Boosting based Particle Swarm Optimization (BOOST-PSO) [28], and Multi-Layer Perceptron based Genetic Algorithm (MLP-GA) [29]. The evaluation of the proposed IHEGS with the DNN model is illustrated in Tab. 1.

The proposed DNN-IHEGS predicts the ASD genes accurately by obtaining a range of accuracy values between 93.12%–99%. The existing methodologies tend to achieve an overall classification accuracy rate ranging from 60.25% to 77.89%, which is comparatively lower than the proposed DNN-IHEGS model. Thus, the DNN-IHEGS model achieves a better classification accuracy rate and tends to be more robust compared with the existing methods. An ensemble of proposed three intelligent swarm-based metaheuristic algorithms such as CI-FSA, HL-GRO, and ED-SOA in the functional perturbation module has resulted in selecting the most appropriate gene features that can distinctly classify diseased autistic gene from their control. Thus, a highly relevant feature leads to a better accuracy score with a lower error rate. The proposed technique also provides a low false prediction that is, a misclassification of the data by obtaining low FPR, FNR, and FDR values as illustrated in Fig. 5.

**Table 1:** Evaluation of proposed DNN- IHEGS based on classification accuracy

| Datasets | NB-FSFS | SVM-MCFS | KNN-UDFS | BOOST-PSO | MLP-GA | Proposed DNN-IHEGS |
|---|---|---|---|---|---|---|
| GSE25507 | 61.23 | 60.25 | 62.54 | 68.98 | 75.86 | 98.78 |
| GSE26415 | 63.54 | 62.54 | 64.58 | 70.12 | 76.89 | 96.25 |
| GSE27044 | 62.12 | 61.58 | 65.47 | 71.24 | 77.89 | 99 |
| GSE109905 | 64.56 | 63.54 | 66.54 | 73.65 | 78.84 | 93.12 |
| GSE42133 | 65.41 | 64.12 | 68.89 | 74.58 | 77.21 | 94.65 |
| GSE123302 | 60.87 | 59.36 | 61.53 | 65.92 | 68.51 | 95.82 |

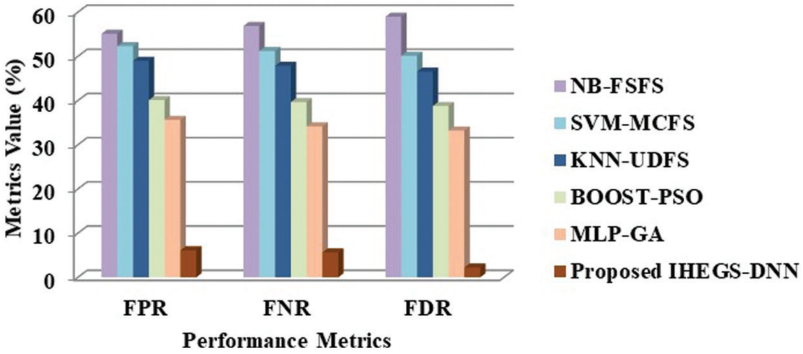


**Figure 5:** Analysis of proposed IHEGS-DNN based on FPR, FNR, and FDR

The metrics FPR, FNR, and FDR illustrate the possibility of misclassification. The lower value of the above metrics represents a better model. Accordingly, the proposed method achieves an average of 6.12% of FPR, 5.64% of FNR, and 2.23% of FDR, whereas the existing techniques tend to achieve an overall value ranging between 33.21% to 58.98% which leads to high misclassification and high error rate model. As a result of low FPR, FNR, and FDR the proposed IHEGS mode has obtained better precision, recall, f-measure, and low error rate as detailed in Tab. 2.

The stated metrics in Tab. 2 give a detailed analysis of reliability, sensitivity, and handling of complex data. The robust model has a higher value when the metrics are measured. Accordingly, the proposed technique tends to obtain an average precision of 95.95%, recall of 96.05%, F-Measure of 96.01%, and an error rate of 3.73% for the six benchmark datasets. Hence, the proposed IHEGS method achieved a better metrics value for detecting ASD and proved to be efficient. Thus, the proposed IHEGS method achieves better reliability and avoids misclassification of ASD genes as compared to the existing methods. The frequently selected genes and CpG sites by the proposed IHEGS with its confidence interval are given in Supplementary file 1. The selected gene subset is validated by the domain expert.

**Table 2:** Evaluation of proposed DNN-IHEGS based on various metrics for predicting ASD

| Dataset | Precision (↑) | Recall (↑) | f-measure (↑) | Error (↓) |
|---|---|---|---|---|
| GSE25507 | 98.6 | 98.4 | 98.6 | 1.22 |
| GSE26415 | 95.8 | 96 | 96.2 | 3.75 |
| GSE27044 | 99 | 99.2 | 99.1 | 1 |
| GSE109905 | 93.1 | 93 | 92.6 | 6.88 |
| GSE42133 | 93.9 | 94.3 | 94 | 5.35 |
| GSE123302 | 95.3 | 95.4 | 95.6 | 4.18 |

## 5 Conclusion

ASD is a neurodevelopment disorder that affects children between the ages of 2 and 5. Genetic and environmental factors influence the cause of ASD. Analyzing gene expression profiles from microarray data can provide meaningful etiologic insights into the disorder. But the complex patterns of genes due to irregular molecular interactions during the transcription process, and the sample size of microarray data make the data difficult for analysis. Hence an Intelligent Hybrid Ensemble Gene selection model was proposed to identify stable and robust gene subsets from high dimensional microarray data. The proposed model comprises both data perturbation and function perturbation modules. SCP-ETC was employed by the data perturbation module to identify the initial gene subset. The functional perturbation module was built with three intelligent swarm-based metaheuristic algorithms such as CI-FSA, HL-GRO, and ED-SOA to identify the final gene subset. The predicted autism genes subset was validated using the DNN model. 6 benchmark microarray datasets were used for the implementation. From the experimental results, it was proved that the proposed IHEGS model was stable and robust enough to identify the autism gene subset from the microarray dataset. The proposed IHEGS obtains an average JI of 0.858 and SCC of 0.886 and predicts the ASD with an average accuracy of 96.25%, precision of 95.95%, recall of 96.06, and an f-measure of 96.01% for the six NCBI benchmark datasets. The prediction model avoids misclassification and overall remains to be stable and robust as compared to existing state-of-the-art methods. In the future, the proposed IHEGS model could be used as a feature selection technique for various other engineering problems and high dimensional datasets that need optimization.

**Acknowledgement:** We thank Dr. Amala for the validation of the results obtained by the proposed system.

**Funding Statement:** The authors received no specific funding for this study.

**Conflicts of Interest:** The authors declares that they have no conflicts of interest to report regarding the present study.

## References

[1] American Psychiatric Association, "What is autism spectrum disorder?," American psychiatric pub, 2018. [online]. Available: https://www.psychiatry.org/patients-families/autism/what-is-autism-spectrum-disorder.

[2] M. J. Maenner, K. A. Shaw, A. V. Bakian, D. A. Durkin, M. S. Esler *et al.,* "Prevalence and characteristics of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2018," *MMWR. Surveillance Summaries*, vol. 70, no. 11, pp. 1–16, 2021.

[3] G. Anurekha and P. Geetha, "Performance analysis of supervised approaches for autism spectrum disorder detection," *International Journal of Trend in Research and Development (IJTRD), Special Issue RIET*, pp. 16–18, 2017.

[4] O. Pain, A. J. Pocklington, P. A. Holmans, N. J. Bray, H. E. O'Brien *et al.,* "Novel insight into the etiology of autism spectrum disorder gained by integrating expression data with genome-wide association statistics," *Biological Psychiatry*, vol. 86, no. 4, pp. 265–273, 2019.

[5] X. Zhang, X. Sun, W. Sun, T. Xu, P. Wang *et al.,* "Deformation expression of soft tissue based on bp neural network," *Intelligent Automation & Soft Computing*, vol. 32, no. 2, pp. 1041–1053, 2022.

[6] X. Zhang, J. Zhou, W. Sun and S. K. Jha, "A lightweight CNN based on transfer learning for COVID-19 diagnosis," *Computers, Materials & Continua*, vol. 72, no. 1, pp. 1123–1137, 2022.

[7] M. D. Alter, R. Kharkar, K. E. Ramsey, D. W. Craig, R. D. Melmed *et al.,* "Autism and increased paternal age related changes in global levels of gene expression regulation," *PLoS ONE*, vol. 6, no. 2, pp. e16715, 2011.

[8] W. Awada, T. M. Khoshgoftaar, D. Dittman, R. Wald and A. Napolitano, "A review of the stability of feature selection techniques for bioinformatics data," in *Proc. of the IEEE Conf. on Information Reuse and Integration (IRI 2012)*, Las Vegas, NV, USA, pp. 356–363, 2012.

[9] R. Alzubi, N. Ramzan and H. Alzoubi, "Hybrid feature selection method for autism spectrum disorder SNPs," in *Proc. of the IEEE Conf. on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2017)*, Manchester, UK, pp. 1–7, 2017.

[10] K. L. Chiew, C. L. Tan, K. S. Wong, K. S. C. Yong and W. K. Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," *Information Sciences*, vol. 484, pp. 153–166, 2019.

[11] S. S. Hameed, R. Hassan and F. F. Muhammad, "Selection and classification of gene expression in autism disorder: Use of a combination of statistical filters and a GBPSO-SVM algorithm," *PLoS ONE*, vol. 12, no. 11, pp. 1–25, 2017.

[12] N. Samy, R. Fathalla, N. A. Belal and O. Badawy, "Classification of autism gene expression data using deep learning," in *Proc. of the Int. Conf. on Intelligent Data Communication Technologies and Internet of Things (ICICI 2019)*, Coimbatore, India, vol. 38, pp. 583–596, 2019.

[13] K. Sekaran and M. Sudha, "Predicting autism spectrum disorder from associative genetic markers of phenotypic groups using machine learning," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 3, pp. 3257–3270, 2021.

[14] A. Rouhi and H. Nezamabadi-Pour, "A hybrid-ensemble based framework for microarray data gene selection," *International Journal of Data Mining and Bioinformatics*, vol. 19, no. 3, pp. 221–242, 2017.

[15] N. Singh and P. Singh, "A hybrid ensemble-filter wrapper feature selection approach for medical data classification," *Chemometrics and Intelligent Laboratory Systems*, vol. 217, pp. 104396, 2021.

[16] G. A. Churchill, "Using ANOVA to analyze microarray data," *Biotechniques*, vol. 37, no. 2, pp. 173–177, 2004.

[17] W. Zhiheng and L. Jianhua, "Flamingo search algorithm: A new swarm intelligence optimization algorithm," *IEEE Access*, vol. 9, pp. 88564–88582, 2021.

[18] A. S. Jaber, H. A. Abdulbari, N. A. Shalash and A. N. Abdalla, "Garra rufa-inspired optimization technique," *International Journal of Intelligent Systems*, vol. 35, no. 11, pp. 1831–1856, 2020.

[19] G. Dhiman and V. Kumar, "Seagull optimization algorithm: Theory and its applications for large-scale industrial engineering problems," *Knowledge-Based Systems*, vol. 165, pp. 169–196, 2019.

[20] Y. Kuwano, Y. Kamio, T. Kawai, S. Katsuura, N. Inada *et al.,* "Autism-associated gene expression in peripheral leucocytes commonly observed between subjects with autism and healthy women having autistic children," *PLoS ONE*, vol. 6, no. 9, pp. e24723, 2011.

[21] R. S. Alisch, B. G. Barwick, P. Chopra, L. K. Myrick, G. A. Satten *et al.,* "Age-associated DNA methylation in pediatric populations," *Genome Research*, vol. 22, no. 4, pp. 623–632, 2012.

[22] R. Kimura, M. Nakata, Y. Funabiki, S. Suzuki, T. Awaya *et al.,* "An epigenetic biomarker for adult high-functioning autism spectrum disorder," *Scientific Reports*, vol. 9, no. 1, pp. 1–7, 2019.

[23] T. Pramparo, K. Pierce, M. V. Lombardo, C. C. Barnes, S. Marinero *et al.,* "Prediction of autism by translation and immune/inflammation coexpressed genes in toddlers from pediatric community practices," *JAMA Psychiatry*, vol. 72, no. 4, pp. 386–394, 2015.

[24] C. E. Mordaunt, B. Y. Park, K. M. Bakulski, J. I. Feinberg, L. A. Croen *et al.,* "A meta-analysis of two high-risk prospective cohort studies reveals autism-specific transcriptional changes to chromatin, autoimmune, and environmental response genes in umbilical cord blood," *Molecular Autism*, vol. 10, no. 1, pp. 1–21, 2019.

[25] Y. Li, S. J. Hu, W. J. Yang, G. Z. Sun, F. W. Yao *et al.,* "Similarity-based feature selection for learning from examples with continuous values," in *Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD 2009)*, Berlin, Heidelberg, pp. 957–964, 2009.

[26] D. Cai, C. Zhang and X. He, "Unsupervised feature selection for multi-cluster data," in *Proc. KDD*, New York, NY, United States, pp. 333–342, 2010.

[27] Y. Yang, H. T. Shen, Z. Ma, Z. Huang and X. Zhou, "L2,1-norm regularized discriminative feature selection for unsupervised learning." in *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI 2010)*, Barcelona, Spain, pp. 333–342, 2010.

[28] B. Xue, M. Zhang and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1656–1671, 2013.

[29] F. Tan, X. Fu, Y. Zhang and A. G. Bourgeois, "A genetic algorithm-based method for feature subset selection," *Soft Computing*, vol. 12, no. 2, pp. 111–120, 2008.