Journal of Scientific & Industrial Research
Vol. 82, August 2023, pp. 841-850
DOI: 10.56042/jsir.v82i08.3398

Autism Gene Subset Selection from Microarray data – A Wrapper Approach

Anurekha G[1]* & Geetha P

[1]Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai 600 062, Tamil Nadu, India

[2]Department of Information Science and Technology, College of Engineering, Anna University, Guindy, Chennai 600 025, Tamil Nadu, India

Received 14 October 2021; revised 20 June 2023; accepted 29 June 2023

Autism spectrum disorder is a complex neurodevelopment disorder that affects an individual's social behavior. Microarray analysis is an extensively used technique to detect autism. Microarray data can provide additional insight into the etiology of the disorder. Identifying the specific set of genes associated with autism from complex microarray data poses a significant research challenge due to its high dimensionality. However, Gene subset selection is classified as an np-hard problem that can be handled by the meta-heuristic algorithm. In this paper, a novel meta-heuristic Game Theory Based Whale Optimization Algorithm is proposed. The proposed algorithm uses a two-person zero-sum game theory and convergence parameter to increase convergence rate and avoid local optima. The performance of the proposed algorithm is tested with 23 mathematical benchmark functions and compared with other state-of-the-art algorithms. Further, the proposed algorithm is employed as a wrapper-based gene subset selection model with a support vector machine. Furthermore, the outcomes demonstrate that the gene selection model utilizing a wrapper-based approach is capable of effectively identifying a subset of autism-related genes with desirable accuracy.

Keywords: Autism spectrum disorder, Dimensionality reduction, Feature selection, Meta-heuristic, Whale optimization algorithm

Introduction

Autism is a complex neuro-developmental disorder that affects an individual's social communication and behavior. The precise cause for autism is still unknown. Some risk factors that cause autism are siblings with autism, elderly parents, low birth weight, chemical imbalance, genetic factors, prenatal environmental, and stress. Microarray is a futuristic technology capable of analyzing thousands of genes simultaneously. These microarrays can be used to study gene expression, genome mapping, transcription factor, Single Nucleotide Polymorphism (SNP), pathogen identification, and toxicity.[1] Autism microarray data represents the variability of expression values to the same class.[2] Consequently, the existing gene selection methods[3–5]for cancer microarray data cannot be used for autism. Identifying the minimum subset of genes with high classification accuracy is the process of gene (feature) subset selection. Computationally this process is classified as an np-hard problem.[6] Meta-heuristic approaches may provide a desirable solution.[7] The primary aim of meta-heuristic approaches is to identify the global optimum solution from the vast solution space. This paper proposes a novel meta-heuristic, "Game Theory Based Whale Optimization Algorithm (GTBWOA)". The proposed GTBWOA algorithm is used as a wrapper-based feature selection model to select the autism gene subset.

A Whale Optimization Algorithm (WOA)[8] has been proposed by Mirjalili & Lewis for solving a global optimization problem. It mimics the hunting behavior of a humpback whale by simulating operations such as search of prey, encircling prey, and bubble-net attacking. The variants of WOA algorithms proposed for solving global optimization problems are Levy Flight trajectory based WOA[9], Improved chaotic map based WOA[10], modified WOA[11], Chaotic WOA[12] and Hybrid WOA[13], in which to mitigate the stagnation problem of the existing WOA, intelligent techniques are employed. An ensemble data mining technique[2] for autism gene selection has been proposed by Latkowski & Osowski. In their work, data mining techniques such as Fisher discriminant analysis, ReliefF algorithm, Two-sample t-test, Kolmogorov–Smirnov test, Kruskal–Wallis test, Stepwise regression method, Feature correlation with class, and SVM-RFE (Support Vector

*Author for Correspondence
E-mail: anu21rekha@gmail.com

Machine-Recursive Feature Elimination) are applied to the microarray data. The subset of best features obtained is given as input to the classifiers. A two-stage classifier of Support Vector Machine (SVM) and Random Forest (RF) is used to analyze the performance of the selected gene subset. A hybrid model[14] by Alzubi *et al*. comprises CMIM (Conditional Mutual Information Maximization) and SVM-RFE to select and classify the informative SNP from the normal ones for autism detection. A data-driven autism gene selection[15] using signal processing and machine learning techniques has been proposed by Antovski *et al*. In their work, signal processing techniques are used for gene selection and machine learning algorithms like SVM, K-Nearest Neighbor (KNN), and RF techniques are used for classification. Detecting methylomic biomarker of autism has been demonstrated.[16] In their work, Analysis of Variance (ANOVA) test, Chi-square, Mutual Information, Pearson's correlation coefficient, and t-test are used for initial gene ranking. Later, RFE is used to refine the gene subset. The performance of the resultant gene subset is analyzed using logistic regression, SVM, KNN, RF, and Naïve Bayes. From the literature study, it has been identified that the variants of whale optimization algorithm suffer from local optima and premature convergence; hence, it is desirable to enhance the performance of existing WOA in both the intensification and diversification phase of the algorithm. Meanwhile, microarray data needs an algorithm to handle high dimensional data and identify autism genes subsets with high classification accuracy. Therefore, the primary contributions of this paper are:

1. A Novel meta-heuristic Game Theory Based Whale Optimization Algorithm.
2. Two-person zero-sum game theory is employed to update the whale position.
3. A Convergence parameter (α) is introduced to avoid local optima.
4. The proposed GTBWOA is utilized as a wrapper gene selection approach for selecting autism genes subset from high dimensional microarray data.

Background Study

Whale Optimization Algorithm

The existing WOA[8] involves three operations namely the encircling prey mechanism, the bubble-net attacking method, and the search for prey mechanism.

Encircling Prey Mechanism

As the optimal solution in the search space is unknown, this mechanism assumes that the whale with the maximum fitness is the optimal solution or near the optimal solution. Hence the entire search agent travels towards the identified optimal solution at the initial iteration. This mechanism is modeled as in Eq. (1) and Eq. (2).

$$\vec{D} = |\vec{C}.\vec{X}^*(t) - \vec{X}(t)| \quad \ldots (1)$$

$$\vec{X}(t+1) = \vec{X}^*(t) - \vec{A}.\vec{D} \quad \ldots (2)$$

where, t represents the current iteration, $\vec{A}$ and $\vec{C}$ denotes the coefficient vectors, X^* represents the position vector of the best solution, $\vec{X}$ represents the position vector. The search agents will follow the current best whale to identify the prey. The coefficient vectors are calculated by Eq. (3) and Eq. (4).

$$\vec{A} = 2\vec{a}.\vec{r} - \vec{a} \quad \ldots (3)$$

$$\vec{C} = 2.\vec{r} \quad \ldots (4)$$

In this context, $\vec{r}$ refers to the random vector, and $\vec{a}$ gradually decreases in a linear manner from 2 to 0.

Bubble-Net Attacking Mechanism

The bubble net attacking method further involves two more operations namely the "shrinking encircling mechanism" and the "spiral updating mechanism".

Shrinking Encircling Mechanism

Increasing the value of 'a' in Eq. (3) achieves the shrinking encircling mechanism. As 'a' value decreases, the corresponding reduction in 'A' can be observed

Spiral Updating Position

It initially calculates the whale located position i.e., (x, y) and prey location at (x*,y*). To replicate the spiral movement resembling that of a whale, a cosine function is employed, as described in Eq. (5)

$$\vec{X}(t+1) = \overrightarrow{D'}.e^{bl}.\cos(2\pi l) + \overrightarrow{X^*}(t) \quad \ldots (5)$$

where, $\overrightarrow{D'} = |\overrightarrow{X^*}(t) - \vec{X}(t)|$ indicates the distance between the i^{th} whale and prey, b-constant, l- random no in [−1,1]. The factor that decides the switch between the "shrinking encircling mechanism" or the "spiral updating position" is given by the random variable p. Variable p has a 50% probability of choosing the spiral or circular movement of the whale which is modeled in Eq. (6).

$$\vec{X}(t+1) = \begin{cases} \overrightarrow{X^*}(t) - \vec{A} \,.\, \vec{D} if p < 0.5 \\ \overrightarrow{D'} \,.\, e^{bl} . \cos(2\pi l) + \overrightarrow{X^*}(t) if p \geq 0.5 \end{cases} \quad \ldots (6)$$

Search for Prey

Here a random search agent X_{rand} from the current population pool is chosen instead of the best agent with the highest fitness in the diversification phase. The search agent randomly searches for the best solution, mimicking the hunting behavior of whales. To control this random search, parameter A is engaged. When the value of 'A' exceeds 1, a random search agent is chosen, and the position of the whale is updated using Eq. (7) and Eq. (8).

$$\vec{D} = |\vec{C} . \overrightarrow{X_{rand}} - \vec{X}| \quad \ldots (7)$$

$$\vec{X}(t+1) = \overrightarrow{X_{rand}} - \vec{A} . \vec{D} \quad \ldots (8)$$

If the A value is less than 1, then the current search agent position is updated in Eq. (1)

Methodology

Data Pre-processing

This module aims to simplify the gene selection process. Microarray data are high-dimensional voluminous datasets. In this paper, data pre-processing is carried out to reduce the number of genes. Subsequently, a t-test is conducted to pre-select the genes prior to the actual process of gene subset selection. From the literature, the t-test[17] outperforms in filtering the genes with high classification accuracy compared with other statistical methods. Let MA(Pid, V) be the microarray data. Where, Pid is the probe id and V is the normalized signal value. Initially, the microarray data is divided into the autistic group and the non-autistic group. Let MA(Pid(A), V(A)) and MA(Pid(NA), V(NA)) be the autistic and non-autistic data respectively. A t-test is applied to the groups for identifying the p-values. Later, sort the p-values in ascending order. Select the genes with low p-values thus the filtered gene subset F[MA(Pid, V)] of size m is obtained. As a result of the preprocessing, 3000 genes from genomics data, 3000 genes from transcriptomics data, and 8000 CpG sites from epigenomics data are selected for further gene subset selection process.

Gene Selection

The proposed GTBWOA with SVM classifier is built together as a wrapper-based gene subset selection model. The reason behind the use of an SVM classifier is that it can perform well for a two-class problem. The GTBWOA is engaged as it can find the optimized solution in high-dimensional solution space. The GTBWOA is explained in detail in forthcoming section. The filtered gene subset of size m is given as input to the gene selection module as mentioned in Algorithm 1. The proposed GTBWOA algorithm identifies the genes subset. The performance of the selected genes subset is estimated in terms of classification accuracy of the SVM classifier, as mentioned in steps 2 and 3 of Algorithm 1. This process is repeated iteratively for N times until the gene subset achieves desirable classification accuracy. The frequently occurred genes are ranked based on their cumulative frequencies for all subsets of genes obtained as given in steps 4 and 5 of Algorithm 1. The selected genes with high rank are given as the output of this module.

Algorithm 1: Gene Selection
Input: F[MA(P_{id}, V)]
Output: Autism Gene subset $Gs(i)$
1: while F[MA(P_{id}, V)] != null
2:$Gs(i)$=GTBWOA(F[MA(P_{id}, V)])
3: CA= SVM ($Gs(i)$)
4: if CA is desirable
5: Gene Ranking($Gs(i)$)
6: else
7: Go to step2
8: end if
9: end while
10: return $Gs(i)$

Game Theory Based Whale Optimization Algorithm

The proposed GTBWOA is the enhanced version of the existing WOA. A convergence parameter (α) and a two-person zero-sum coin matching game are introduced to enhance the existing WOA, as discussed in this section.

Convergence Parameter

In the existing WOA, search agents will follow the current best whale to identify the prey; hence there is a chance of local optima convergence. To avoid the trap of local optima, a convergence parameter (α) is introduced. The convergence parameter aims to reduce the magnitude of change of the search agents[11] following the current best in encircling prey mechanism or the random agent chosen in the search of prey mechanism. The convergence parameter (α) is the sum of coefficient vectors $\vec{A}$ and $\vec{C}$ as mentioned in Eq. (12). Hence the equation of encircling the prey, spiral updating position, and search of prey can be written as given in Eq. (9), Eq. (10), and Eq. (11), respectively

$$\vec{X}(t+1) = \frac{\vec{X}^*(t) - \vec{A}.\vec{D}}{\alpha} \quad \ldots (9)$$

where, $\vec{D} = \frac{|\vec{C}.\vec{X}^*(t) - \vec{X}(t)|}{\alpha}$

$$\vec{X}(t+1) = \frac{\overrightarrow{D'}.e^{bl}.\cos(2\pi l) + \overrightarrow{X^*}(t)}{\alpha} \quad \ldots (10)$$

where, $\overrightarrow{D'} = \frac{|\overrightarrow{X^*}(t) - \vec{X}(t)|}{\alpha}$

$$\vec{X}(t+1) = \frac{\overrightarrow{X_{rand}} - \vec{A}.\vec{D}}{\alpha} \quad \ldots (11)$$

where, $\vec{D} = \frac{|\vec{C}.\overrightarrow{X_{rand}} - \vec{X}|}{\alpha}$

$$\alpha = \vec{A} + \vec{C} \quad \ldots (12)$$

Random Variable p

The factor that decides the switch between the "shrinking encircling mechanism" or the "spiral updating position" is given by the random variable p. In the existing WOA, it is assumed that p has a 50% probability of choosing the spiral or circular movement of the whale. To avoid premature convergence for the global optimization problem, modification is necessary to the diversification phase of the algorithm. Henceforth, a two-person zero-sum game theory[18] is employed to decide the whale's position. The purpose of preferring game theory[19] is to make appropriate decisions from choices of decisions. Let P1 and P2 be the players with strategies X, Y respectively. Players P1 and P2 toss a fair coin simultaneously. The possible outcome is either head (H) or tail (T). This scenario is mathematically modeled as given in Eq. (13) and Eq. (14).

$$P(X=x) = \begin{cases} 1 \; if x is H \\ -1 \; otherwise \end{cases} \quad \ldots (13)$$

$$P(Y=y) = \begin{cases} 1 \; if y is T \\ -1 \; otherwise \end{cases} \quad \ldots (14)$$

The winning condition for player P1 is when toss_coin(P1) = H and toss_coin(P2) = H. The winning condition for player P2 is when toss_coin (P2) =T and toss_coin(P2) =T. The coin is tossed again if toss_coin(P1) = H and toss_coin(P2) = T or toss_coin (P1) = T and toss_coin(P2) = H. When player P1 wins the game, the whale position is updated by encircling prey mechanism, i.e., Eq. (9), and when player P2 wins, the whale position is updated by spiral updating position, i.e., Eq. (10). The Bubble-Net Attacking strategy with two-person zero-sum game theory is mathematically modeled as given in Eq. (15).

$$\vec{X}(t+1) = \begin{cases} \frac{\overrightarrow{X^*}(t) - \vec{A}.\vec{D}}{\alpha} if p1 \; wins \\ \frac{\overrightarrow{D'}.e^{bl}.\cos(2\pi l) + \overrightarrow{X^*}(t)}{\alpha} if p2 \; wins \end{cases} \quad \ldots (15)$$

Algorithm 2 explains the steps involved in the proposed GTBWOA. Initially, the population is initialized as given in step 2. The fitness is calculated for each search agent. The search agent with high fitness (X*) is the best search agent. For each search agent, update the parameters a, A, c, l, and compute α. Later, a coin toss game for players P1 and P2 was initiated. If P1 wins and |A| is less than 1, update whale position by encircling prey mechanism (Eq. 9) else update whale position by the search for prey mechanism, Eq. (11). If player P2 wins, then update whale position by spiral updating mechanism, Eq. (10). If player P1 and P2 end up in a draw, then again play the game. This process is carried out until maximum iteration. Finally, the best solution is returned as output.

Algorithm 2: The Game Theory Based Whale Optimization Algorithm

Input: F[MA(Pid, V)]

Output: X*

1. Initialization{
2. whale population Xi (i=1, 2, ..n)
3. computefitness
4. X* = best search agent}
5. while (t<max_iteration)
6. for (each whale) do
7. update a, A, c and l
8. α = A+C
9. toss_coin (P1)
10. toss_coin (P2)
11. if(P1==H && P2==H)
12. if(|A| < 1) then
13. update whale position by Eq. (9)
14. else
15. select random whale Xrand
16. update whale position by Eq. (11)
17. end if
18. else if(P1==T && P2==T)
19. update whale position by Eq. (10)
20. else
21. go to step 10
22. end if
23. end for
24. t=t+1
25. end while}
26. return X*

GTBWOA Wrapper Gene Selection

To employ the proposed GTBWOA as a gene subset selector, it is necessary to formulate the following vital parameters. A random population P is generated to initialize the gene selection process. Fitness function[20,21] as given in Eq. (16) is applied over the randomly generated population P. To each population, fitness is calculated and the population with the highest fitness is marked as prey X^* and the search agents move towards the identified prey.

$$Fitness = \alpha\varepsilon(Gs) + \beta\frac{|Gs|}{|G|} \quad \ldots (16)$$

where, α and β represent the relative importance of error and feature respectively, β = α-1 and $\varepsilon(Gs)$ represent the error of the gene subset Gs by the SVM classifier. In a meta-heuristic algorithm formulating the solution, space is as important as designing the fitness function. Here the solution space contains a one-dimensional array named S with s binary elements. The value "1" represents the gene selected, where "0" represents the corresponding gene that is not selected. Since feature selection is a binary problem, the solution space is denoted by 0 or 1 and it is necessary to transfer the continuous numerical values into the discrete binary form. From the literature, the V-shaped transfer function[20] is used. The transfer functions V2, V3, V4 provide the denotable result on comparing with function V1. The proposed GTBWOA works as a gene selection algorithm for the given microarray dataset with these discussed parameters.

Results and Discussion

This section presents the description of the results and insights derived from the simulation.

Performance of GTBWOA

The proposed GTBWOA algorithm is simulated using MATLAB R2017B and tested over 23 mathematical benchmark functions which in turn contain 7 monomodal functions, 6 polymodal functions, and 10 fixed dimension polymodal functions.[8] The objective is to obtain the f_{min} value for the benchmark function from the given range of solution space. Here 30 search agents were employed with 500 iterations. The mean and standard deviation of the 23 mathematical functions is contrasted with those of other cutting-edge algorithms such as WOA[8], Particle Swarm Optimization (PSO)[22], Gravitational Search Algorithm (GSA)[23], Differential Evolution (DE)[24] given in Supplementary File 1. The search agent position is decided by the two-person zero-sum game theory. The utilization of the convergence parameter (α) in the proposed GTBWOA results in a lower standard deviation compared to other algorithms, thereby enhancing its effectiveness. This, in turn, implies that the algorithm has avoided premature convergence by a diverse population. The proposed GTBWOA has outperformed in the monomodal and polymodal function compared with other cutting-edge algorithms. The existing WOA algorithm has performed well in the fixed dimension polymodal functions f15and f18. The average and standard deviation of the proposed GTBWOA is contrasted with modified WOA[11] and Levy Flight Trajectory-based WOA[9] as given in Supplementary File 2. The proposed GTBWOA algorithm achieves the f_{min} value with less standard deviation for the given 23 mathematical benchmark functions.

Convergence of GTBWOA

The convergence ability of the proposed GTBWOA for monomodal, polymodal, and fixed dimension polymodal benchmark functions is depicted in Figs 1–3. From the graph, the proposed GTBWOA algorithm performs well for both monomodal and polymodal

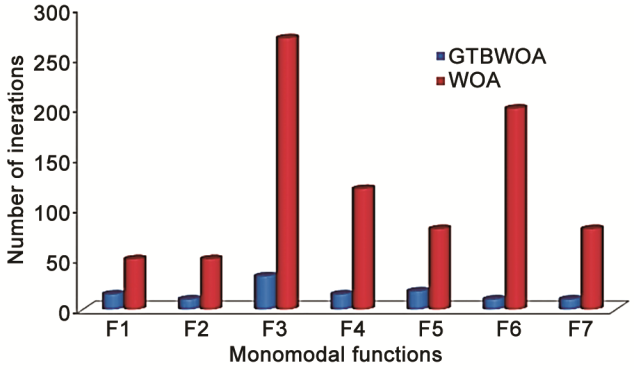

Fig. 1 — Convergence rate of GTBWOA over monomodal functions

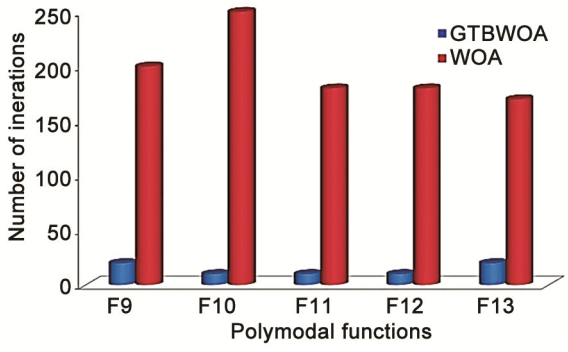

Fig. 2 — Convergence rate of GTBWOA over polymodal functions

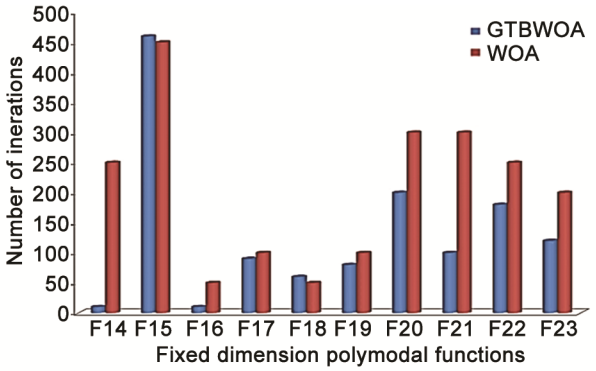

Fig. 3 — Convergence rate of GTBWOA over fixed dimension polymodal functions

functions. Whereas the convergence of the proposed GTBWOA is high in fixed dimension polymodal functions f15 and f18.

Performance of GTBWOA for Gene Selection

Dataset

The microarray dataset comprises of Genomics, Epigenomics, and Transcriptomics microarray dataset. For implementation purposes, pre-processed microarray data from Gene Expression Omnibus (GEO) with accession number GSE25507(25) (Genomics), GSE26415(26) (Transcriptomics), and GSE27044[27] (Epigenomics) is used. The GSE25507 microarray dataset comprises gene expression of 82 autistic children and 64 matched controls. The dataset GSE26415 comprises gene profiling data from 21 individuals diagnosed with autism and 21 healthy mothers having children with autism. The gene profile of 42 individuals with autism was compared to a control group consisting of individuals of the same age, sex, and IQ. The GSE27044 data comprises 27,578 CpG site DNA Methylation profile studies of autistic children with their non-autistic sibling as control. All three microarray datasets contain two classes.

Experimental Setup

The proposed gene selection methodology is implemented with MATLAB R2017B. The obtained gene subset is classified using SVM, KNN, ANN (Artificial Neural Network), and Random forest with default hyper-parameters. The performance evaluation of the proposed gene selection methodology is conducted in this study using k-fold cross-validation. To address the inherent randomness of the proposed GTBWOA algorithm, ten consecutive runs are averaged and presented in the table for thorough analysis. To validate the efficiency of the proposed GTBWOA, a comparison with filter-based techniques such as Information gain, ReliefF, FCBF (Fast Correlation Based Filter) and wrapper-based approaches like GA, PSO, ACO (Ant Colony Optimization) have been carried out.

Performance Analysis

The performance of the proposed GTBWOA as a wrapper gene selection algorithm is analyzed using metrics such as classification accuracy, precision, recall, F1 score, and error rate as mentioned in Eq. (17), Eq. (18), Eq. (19), Eq. (20), and Eq. (21), respectively. The metrics precision, recall, accuracy, and F1 score should increase, and the error rate should decrease so that the proposed GTBWOA performs well compared with other approaches.

$$Classification\ Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \ldots (17)$$

$$Precision = \frac{TP}{TP+FP} \quad \ldots (18)$$

$$Recall = \frac{TP}{TP+FN} \quad \ldots (19)$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \ldots (20)$$

$$Errorrate = \frac{FP+FN}{TP+TN+FP+FN} \quad \ldots (21)$$

where, the terms TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) represent components of the evaluation metrics. The performance of the proposed GTBWOA for top-ranked 200 genes of genomics and transcriptomics data is tabulated in Table 1 and Table 2, respectively. The performance of top-ranked 500 CpG sites of epigenomics data is depicted in Table 3.

From the comparative study, it is inferred that the proposed GTBWOA could find the global best solution from high dimensional data by avoiding local optima and premature convergence. GTBWOA outperforms other comparative gene selection models with high precision, recall, accuracy, F1 score, and low error rate. The obtained result of the proposed gene selection model is benchmarked with existing methods as given in Table 4.

Since the proposed wrapper-based gene subset selection model uses GTBWOA for gene subset selection it outperforms the existing gene selection methods with high accuracy and a minimum number of genes as depicted in Table 4.

Table 1 — Performance analysis of genomics data

Feature selection technique	Classifier	Precision(↑)	Recall(↑)	Accuracy(↑)	Error(↓)	F1score(↑)
Information gain	SVM	66.12%	65.15%	64.84%	35.15%	65.64%
	KNN	69.23%	57.69%	58.59%	41.40%	62.93%
	ANN	75.38%	65.33%	67.18%	32.81%	70%
	Random forest	70.76%	63.01%	64.06%	35.93%	66.66%
ReliefF	SVM	66.15%	65.15%	64.84%	35.15%	65.64%
	KNN	69.23%	57.69%	58.59%	41.40%	62.93%
	ANN	55.38%	62.06%	60.15%	39.84%	58.53%
	Random forest	56.92%	50%	49.21%	50.78%	53.23%
FCBF	SVM	66.15%	65.15%	64.84%	35.16%	65.65%
	KNN	69.23%	57.69%	58.59%	41.41%	62.93%
	ANN	78.46%	68%	70.31%	29.69%	72.86%
	Random forest	69.23%	58.44%	59.38%	40.63%	63.38%
GA	SVM	90%	93.10%	91.67%	8.33%	91.53%
	KNN	86.67%	94.20%	90.67%	9.33%	90.28%
	ANN	87.33%	91.61%	89.67%	10.33%	89.42%
	Random forest	88.67%	90.48%	89.67%	10.33%	89.56%
PSO	SVM	88.67%	89.26%	89.00%	11%	88.96%
	KNN	87.33%	88.51%	88%	12%	87.92%
	ANN	88%	89.19%	88.67%	11.33%	88.59%
	Random forest	88.67%	89.86%	89.33%	10.67%	89.26%
ACO	SVM	88%	88.59%	88.33%	11.67%	88.29%
	KNN	86.67%	87.25%	87%	13%	86.96%
	ANN	84.67%	86.9%	85.67%	14.33%	85.52%
	Random forest	86%	87.76%	87.00%	13%	86.87%
GTBWOA (proposed)	SVM	98%	98.66%	98.33%	1.67%	98.33%
	KNN	95.33%	94.70%	95%	5%	95.02%
	ANN	96.67%	97.32%	94.44%	5.56%	96.99%
	Random forest	95.33%	96.62%	92.50%	7.50%	95.97%

Table 2 — Performance analysis of transcriptomics data

Feature selection technique	Classifier	Precision(↑)	Recall(↑)	Accuracy(↑)	Error(↓)	F1score(↑)
Information gain	SVM	58.54%	61.54%	61.90%	38.10%	60%
	KNN	41.46%	43.59%	45.24%	54.76%	42.50%
	ANN	56.10%	53.49%	54.76%	45.24%	54.76%
	Random forest	36.59%	37.50%	39.29%	60.71%	37.04%
ReliefF	SVM	58.54%	61.54%	61.90%	38.10%	60%
	KNN	41.46%	43.59%	45.24%	54.76%	42.50%
	ANN	60.98%	52.08%	53.57%	46.43%	56.18%
	Random forest	53.66%	52.38%	53.57%	46.43%	53.01%
FCBF	SVM	58.54%	61.54%	61.90%	38.10%	60%
	KNN	41.46%	43.59%	45.24%	54.76%	42.50%
	ANN	46.34%	50%	51.19%	48.81%	48.10%
	Random forest	56.10%	47.92%	48.81%	51.19%	51.69%
GA	SVM	94.67%	93.42%	94%	6%	94.04%
	KNN	93.33%	90.32%	91.67%	8.33%	91.80%
	ANN	94%	92.76%	93.33%	6.67%	93.8%
	Random forest	95.33%	91.67%	93.33%	6.67%	9.46%

(*Contd*)

Table 2 — Performance analysis of transcriptomics data

Feature selection technique	Classifier	Precision(↑)	Recall(↑)	Accuracy(↑)	Error(↓)	F1score(↑)
PSO	SVM	92.67%	90.26%	91.33%	8.67%	91.45%
	KNN	92%	93.88%	93%	7%	92.93%
	ANN	93.33%	94.59%	94%	6%	93%
	Random forest	92.67%	92.05%	92.33%	7.67%	92.36%
ACO	SVM	96%	95.36%	95.67%	4.33%	95.68%
	KNN	94.67%	94.04%	94.33%	5.67%	94.35%
	ANN	94.67%	95.95%	95.33%	4.67%	95.30%
	Random forest	94%	95.92%	95%	5%	94.95%
GTBWOA (proposed)	SVM	99.33%	98.68%	99%	1%	99%
	KNN	94.67%	95.95%	95.33%	4.67%	95.30%
	ANN	94.67%	93.42%	94%	6%	94.04%
	Random forest	94%	93.38%	93.67%	6.33%	93.69%

Table 3 — Performance analysis of epigenomics data

Feature selection technique	Classifier	Precision(↑)	Recall(↑)	Accuracy(↑)	Error(↓)	F1score(↑)
Information gain	SVM	51.67%	48.44%	48.33%	51.67%	50%
	KNN	60.83%	50.69%	50.83%	49.17%	55.30%
	ANN	57.50%	46.62%	45.83%	54.17%	51.49%
	Random forest	65%	53.06%	53.75%	46.25%	58.43%
ReliefF	SVM	51.67%	48.44%	48.33%	51.67%	50%
	KNN	60.83%	50.69%	50.83%	49.17%	55.30%
	ANN	52.50%	50.81%	50.83%	49.17%	51.64%
	Random forest	57.50%	50.74%	50.83%	49.17%	53.91%
FCBF	SVM	51.67%	48.44%	48.33%	51.67%	50%
	KNN	60.83%	50.69%	50.83%	49.17%	55.30%
	ANN	43.33%	49.06%	49.17%	50.83%	46.02%
	Random forest	62.50%	55.15%	55.83%	44.17%	58.59%
GA	SVM	96%	95.36%	95.67%	4.33%	95.68%
	KNN	94.67%	94.04%	94.33%	5.67%	94.35%
	ANN	94.67%	95.95%	95.33%	4.67%	95.30%
	Random forest	94%	95.92%	95%	5%	94.95%
PSO	SVM	89.33%	91.16%	90.33%	9.67%	90.24%
	KNN	90%	91.84%	91%	9%	90.91%
	ANN	92%	90.79%	91.33%	8.67%	91.39%
	Random forest	93.33%	9.96%	96.67%	6.33%	93.65%
ACO	SVM	92.67%	90.26%	91.33%	8.67%	91.45%
	KNN	92%	93.88%	93%	7%	92.93%
	ANN	93.33%	94.59%	94%	6%	93.96%
	Random forest	92.67%	92.05%	92.33%	7.76%	92.36%
GTBWOA (proposed)	SVM	96%	96.64%	96.33%	3.67%	96.2%
	KNN	94.67%	94.04%	94.33%	5.67%	94.35%
	ANN	98.67%	99.33%	99%	1%	**99%**
	Random forest	98%	97.5%	97.67%	2.33%	97.67%

Table 4 — Comparison with existing work

Dataset	Existing work	Accuracy	Proposed GTBWOA accuracy
GSE25507	Antovski *et al.*[15]	99%(300 genes)	98.33% (200 genes)
GSE26415	Kim *et al.*[28]	93.80%	99%
	Sekaran *et al.*[29]	97.6%	
GSE27044	Feng *et al.*[16]	99.7(678 CpG site)	99% (500 CpG site)

Conclusions

Autism gene subset selection from high dimensional microarray dataset is a significant research challenge. This paper aims to contribute towards etiology by identifying the autism genes subset. Therefore, a novel meta-heuristic GTBWOA algorithm has been proposed. The idea of introducing a convergence parameter and two-person zero-sum game theory to the existing WOA algorithm has obtained efficient results. The proposed GTBWOA outperforms existing WOA and other cutting-edge algorithms. The proposed GTBWOA has overcome local optima and increased the convergence rate with a diverse population. The proposed GTBWOA is employed as a wrapper-based gene selection model with an SVM classifier. The proposed GTBWOA algorithm is comparatively the best in identifying autism gene subsets with high accuracy from high dimensional, non-linear, and scattered data. As a potential future direction, the proposed approach could be extended to encompass real-time clinical trials involving complex diseases, as well as diverse engineering domains where the attainment of global optima is crucial.

References

1 Selvaraj S & Natarajan J, Microarray data analysis and mining tools microarray data analysis and mining tools, *Bioinformation*, **6(3)** (2011) 95.

2 Latkowski T & Osowski S, Data mining for feature selection in gene expression autism data, *Expert Syst Appl*, **42(2)** (2015) 864–872, DOI:10.1016/j.eswa.2014.08.043.

3 Ke L, Li M, Wang L, Deng S, Ye J, & Yu X, Improved swarm-optimization-based filter-wrapper gene selection from microarray data for gene expression tumor classification, *Pattern Anal Appl*, **26(2)** (2023) 455–472, DOI: 10.1007/s10044-022-01117-9.

4 Azadifar S, Rostami M, Berahmand K, Moradi P & Oussalah M, Graph-based relevancy-redundancy gene selection method for cancer diagnosis, *Comput Biol Med*, **147(1)** (2022) 105766, DOI: 10.1016/ j.compbiomed.2022.105766.

5 Aziz R M, Application of nature inspired soft computing techniques for gene selection: a novel frame work for classification of cancer, *Soft Comput*, **26(22)** (2022) 12179–96, DOI: 10.1007/s00500-022-07032-9 6.

6 Chen B, Hong J & Wang Y, The minimum feature subset selection problem, *J Comput Sci Technol*, **12(2)** (1997) 145–153.

7 García-Torres M, García-López F C, Melián Batista B, Moreno Pérez J A & Moreno Vega J M, Solving feature subset selection problem by a hybrid metaheuristic, *Proc First Int Work hybrid Metaheuristics as part 16th Eur Conf Artif Intell*, 2004, 59–68.

8 Mirjalili S & Lewis A, The whale optimization algorithm, *Adv Eng Softw*, **95** (2016) 51–67.

9 Ling Y, Zhou Y & Luo Q, Lévy flight trajectory-based whale optimization algorithm for global optimization, *IEEE Access*, **5** (2017) 6168–6186.

10 Oliva D, Abd El A M & Ella H A, Parameter estimation of photovoltaic cells using an improved chaotic whale optimization algorithm, *Appl Energy*, **200** (2017) 141–54, DOI: 10.1016/j.apenergy.2017.05.029.

11 Sivalingam R, Chinnamuthu S & Dash S S, A modified whale optimization algorithm-based adaptive fuzzy logic PID controller for load frequency control of autonomous power generation systems, *Automatika*, **58(4)** (2017) 410–421, DOI: 10.1080/00051144.2018.1465688.

12 Kaur G & Arora S, Chaotic whale optimization algorithm, *J Comput Des Eng*, **5(3)** (2018) 275–284, DOI: 10.1016/j.jcde.2017.12.006

13 Luo J & Shi B, A hybrid whale optimization algorithm based on modified differential evolution for global optimization problems, *Appl Intell*, **49(5)** (2019) 1982–2000.

14 Alzubi R, Ramzan N & Alzoubi H, Hybrid feature selection method for autism spectrum disorder SNPs, *2017 I EEE Conf Comput Intell Bioinforma Comput Biol* (IEEE) 2017, 1–7.

15 Antovski A, Kostadinovska S, Simjanoska M, Eftimov T, Ackovska N & Bogdanova A M, Data-driven autism biomarkers selection by using signal processing and machine learning techniques, *Bioinformat 2019 - 10th Int Conf Bioinformat Model Methods Algorithms, Proceedings*; *Part 12th Int Jt Conf Biomed Eng Syst Technol* (BIOSTEC)2019, 201–208.

16 Feng X, Hao X, Xin R, Gao X, Liu M, Li F, Wang Y, Shi R, Zhao S & Zhou F, Detecting methylomic biomarkers of pediatric autism in the peripheral blood leukocytes, *Interdiscip Sci Comput Life Sci*, **11(2)** (2019) 237–246, DOI: 10.1007/s12539-019-00328-9

17 Bonilla Huerta E, Duval B & Hao J K, A hybrid LDA and genetic algorithm for gene selection and classification of microarray data, *Neurocomputing*, **73(13–15)** (2010) 2375–2383.

18 Ferguson T S, A course in game theory, World Scientific 2020.

19 Bauso D, *Game theory: Models, numerical methods and applications. Vol. 1, Foundations and Trends in Systems and Control* (Now Publishers) 2014, 379–522.

20 Mafarja M, Jaber I, Ahmed S & Thaher T, Whale optimisation algorithm for high-dimensional small-instance feature selection, *Int J Parallel Emergent Distrib Syst*, **36(2)** (2019) 1–17, DOI: 17445760.2019.1617866

21 Xue B, Zhang M & Browne W N, New fitness functions in binary particle swarm optimisation for feature selection, *2012 IEEE Congr Evol Comput* (IEEE) 2012, 1–8.

22 He Q & Wang L, An effective co-evolutionary particle swarm optimization for constrained engineering design problems, *Eng Appl Artif Intell*, **20(1)** (2007) 89–99.

23 Rashedi E, Nezamabadi-pour H & Saryazdi S, GSA: A gravitational search algorithm, *Inf Sci (Ny)*, **179(13)** (2009) 2232–2248, DOI: 10.1016/j.ins.2009.03.004.

24 Storn R & Price K, Differential Evolution - A simple evolution strategy for fast optimization, *Dr Dobb's J*, **22(4)** (1997) 18–24.

25 Alter M D, Kharkar R, Ramsey K E, Craig D W, Melmed R D, Grebe T A, Bay R C, Ober-Reynolds S, Kirwan J, Jones J J & Turner J B, Autism and increased paternal age related changes in global levels of gene expression regulation, *PLoS One*, **6(2)** (2011) e16715.

26 Kuwano Y, Kamio Y, Kawai T, Katsuura S, Inada N, Takaki A & Rokutan K, Autism-associated gene expression in peripheral leucocytes commonly observed between subjects with autism and healthy women having autistic children, *PLoS One*, **6(9)** (2011) 24732.

27 Alisch R S, Barwick B G, Chopra P, Myrick L K, Satten G A, Conneely K N & Warren S T, Age-associated DNA methylation in pediatric populations, *Genome Res*, **22(4)** (2012) 623–632.

28 Kim S H, Kim I B, Oh D H & Ahn D H, Predicting autism spectrum disorder using blood-based gene expression signatures and machine learning, *Eur Neuropsychopharmacol*, **27(1)** (2017) S1090.

29 Sekaran K and Sudha M, Predicting autism spectrum disorder from associative genetic markers of phenotypic groups using machine learning, *J Ambient Intell Humaniz Comput*, **12(3)** (2021) 3257–3270.