



## Predicting Alzheimer's Disease Using a Modified Grey Wolf Optimizer and Support Vector Machine

Haider A. Abbas Mohammed<sup>1</sup>, Ali A Kasim Jizany<sup>2</sup>, Iman Mohammed Mahmood<sup>1</sup>, Qusay Kanaan Kadhim<sup>3\*</sup>

<sup>1</sup> Department of Computer Science, College of Basic Education, University of Diyala, Diyala 32001, Iraq

<sup>2</sup> College of Administration and Economics, University of Baghdad, Baghdad 10071, Iraq

<sup>3</sup> Department of Computer Science, College of Science, University of Diyala, Baqubah 32001, Diyala, Iraq

Corresponding Author Email: [dr.qusay.kanaan@uodiyala.edu.iq](mailto:dr.qusay.kanaan@uodiyala.edu.iq)

Copyright: ©2024 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/isi.290228>

### ABSTRACT

**Received:** 11 August 2023

**Revised:** 24 November 2023

**Accepted:** 7 December 2023

**Available online:** 25 April 2024

#### Keywords:

*predictive modelling, grey wolf optimizer, Alzheimer's disease, support vector machine, gene selection, bioinformatics*

Alzheimer's Disease (AD), which is left untreated, is the primary cause of neurodegenerative dementia, which primarily affects those over 65. Memory and thought processes are gradually disrupted by the irreversible brain disruption and, eventually, the capacity to fulfil simple tasks. Detecting AD early prevents its progression and diminishes its signs. The study aimed to build a classification model that might predict the early stages of Alzheimer's disease using accurate early-stage gene expression data from blood obtained from a clinical Alzheimer's dataset. The datasets used in this work were gathered from the Gene Expression Omnibus (GEO) and are GSE63060 and GSE63061. It has the right rows for (569 samples) and columns for (16382 genes). The suggested GWO gene selection method aims to identify the ideal feature subset for medical data. A model for predicting early Alzheimer's disease is proposed based on the modified Grey Wolf Optimizer and Support Vector Machine (GWO-SVM). These methods will help reduce the number of trivial and redundant genes in the original datasets. We attained an accuracy of 82%-88% using only the Support Vector Machine method. When we utilized several evolutionary algorithms to implement gene selection, we observed an increase in accuracy of 6%-11%, with modified Grey Wolf optimization with Crossover (CGWO) called the proposed algorithm (CGWO-SVM) achieving the highest accuracy of (97.41%) with the Alzheimer's disease dataset in comparison to the other competitive schemes in the existing literature. Thus, the proposed system is suitable for picking more accurate classification and interesting genes to increase classification accuracy while decreasing gene dimensions.

## 1. INTRODUCTION

Alzheimer's disease is a degenerative brain sickness that causes memory loss and cognitive decline over time. This harms the brain's nerve cells involved in language and memory. 65 years after the symptoms' onset, the prevalence rises significantly with age. The AD is responsible for 60-90% of all dementia cases [1]. In this paper aims to early prediction for Alzheimer's disease at the highest possible accuracy and the least error by selecting the most informative genes. Due to the importance lack of a particular treatment for AD, a new technique called the microarray technique was used to discover the genes that reason the illness. Microarray technology is a valuable medical device that biologists use to track gene expression levels in different organisms. Through a variety of practical situations, microarray data analysis aims to assist in accurate medical diagnosis across several genes and determining the best treatment for a number of ailments [2]. The data of gene expression obtained by microarray technology, on the other hand, suffers from a high dimensionality curse. The deletion is longer critical if the genes lack important information or are redundant [3]. The gene selection approach is a methodology for solving this

problem that only extracts the best collection of characteristics (genes) to create classifier models. Gene selecting is the method for selecting a lesser portion from the genes from a wider gene collection that solely contains important genes. A machine learning algorithm is used in this study to predict AD using gene expression data [4]. Support vector machines (SVMs) were originally proposed. As a statistical learning theory-based machine learning algorithm. A sizable amount of data is used in medical applications and solar power generation prediction to understand the activity of genes and forecast the unknown class label using a training set [5]. For the purpose of projecting the short-term resource load for cloud computing, the Grey Wolf Optimization Algorithm model that has been enhanced by external optimization and support vector machines has been used [6]. The experiments shown that the suggested model could properly characterize the complex trends of cloud computing resource short-term load and effectively improve the accuracy of short-term resource load prediction.

Mohammed [7] presented a technique based on machine learning (SVM). To find AD genes that are prevalent across the entire genome, the proposed technique was tested by merging gene expression data using data from a specific gene

network inside the human brain. The results of the procedure provide an examination of genes linked to Alzheimer's disease, as well as an assessment of their reliability. Limitation only a few genes were the subject of the investigation; other genes have not been investigated. Perera et al. [8] used of gene expression datasets namely: (GSE5281). The Principal Components Analysis (PCA) and (RF) The best subset is found using a feature selection algorithm. Models for classification: Naive Bayes (NB), Decision trees (DT), (SVM), and (RF), have proved predictive in distinguishing between Cognitively Normal, Mild Cognitive Impairment, SVM and RF classification models were successful for both healthy patients and those with AD, achieved accuracy were (93.76%), and (87.48%). That of the lesser sample sizes in addition to unstable nature of this algorithm being presented. Muhammed Niyas et al. [9] using the AutoML technology Add Data Bio (JADBIO). Classification algorithms are: Random Forests (RF), Support Vector Machines (SVM), and Ridge Logistic Regression (LR). The use datasets (GSE63060, GSE63061, GSE46579 and GSE120584) from AD samples. The system achieved was (85%) accuracy. The study's weakness is that the training set is possible to contain a large amount from noise, which could have an impact on model performance, and low accuracy. Abdulwahab et al. [10] employed Support Vector Machine (SVM), Naive Bayes (NB), and K-Nearest Neighbor (KNN) algorithms to analyze the gene expression of AD patients and healthy individuals. The system achieved an accuracy was (91.8%) with (500) genes. Unproven data with missing values. El-Gawady et al. [11] proposed a gene selection algorithm based on statistical Chi squared ( $\chi^2$ ), ANOVA, and MI methods for selecting the best gene subset. They used in this experiment a dataset namely (GEO: GSE44770, GSE44771, and GSE44768). The SVM, RF, LR, and Adaboost models were used to analyze the gene expression of AD patients and healthy individuals. The average accuracy was (95.5%) by SVM. Limitation this takes a lot of time to locate the best features for given budget range.

In this paper aims to early prediction for Alzheimer's disease at the highest possible accuracy and the least error by selecting the most informative genes, an efficient gene selection algorithm is proposed, which mainly depends on Information Gain (IG) technique. In addition, a training algorithm for tuning Support Vector Machine (SVM) is proposed based on an enhanced version of Grey Wolf Optimizer with Crossover Operator, called (CGWO). The SVM tuned via CGWO was tested based on two Alzheimer datasets (GSE63060 and GSE63061). The gene expression omnibus (GEO), a publicly accessible data source, provided the dataset used in this work. It was made available on August 5, 2015 by the National Center for Bioinformatics Information (NCBI). The AddNeuroMed Cohort contributed the accession numbers for the dataset, which are GSE63060 and GSE63061. These two datasets were combined to create the AD dataset, which has more samples overall. Microarray technology was used to track the level of gene expression in the AD dataset. Its rows correspond to (569 samples) and columns to (16382 genes). It is composed of 142 MCIs, 182 CTLs, and 245 AD patients.

## 2. METHODOLOGY

In this section of the suggested process, the microarray AD data set is loaded in its raw form, among other crucial steps.

Then, as shown in Figure 1, Gene selection techniques employ the Min-Max strategy to standardize data, optimization gray wolf optimizer with crossover, and classification with a Support Vector Machine. The Gene Expression Omnibus (GEO: GSE63060 and GSE63061) of the National Center for Bioinformatics (NCBI) was the source of the data, which is publicly accessible. Next, we created the AD dataset by merging two datasets into one. The Alzheimer's disease dataset including 16382 genes and 569 samples, comprising 245 Alzheimer's disease and 142 mild cognitive defect patients (MCI). The microarray usually is used for representing the genetic datasets. In other word, the contents of such datasets are thousands of genes, which not all of them are important or relevant to the classification process [12]. The size of the dataset and the nature of the genes as well effect on the performance of the classification models – i.e., the classification accuracy, and on the time of learning process. Meaning that, these two factors could cause an over-fitting to the training process, which confuses the learning procedures. Therefore, designing an efficient gene selection method is an important preprocessing step, which selects the best minimal number of genes.

The preprocessing for the data is a bit noisy generated by microarray technologies. The dataset needs to be uniform to reduce the variation in expression measurements. The process of normalizing the data collection can be normalized using the min-max method. Gene expression values are calculated so that each gene's lowest value is zero and its greatest value is one.

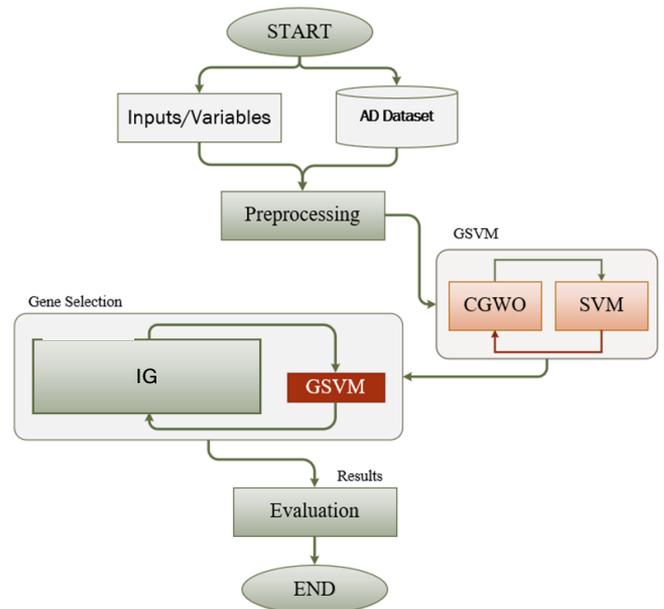


Figure 1. Proposed methods (CGWO-SVM)

### 2.1 Gene selection methods

A set of microarray tests indicates the varied expression from multiple genes in many contexts. A dimensionality of microarray data is an issue because most of the genes are unrelated to the categorization process. As a result, gene selection approaches are effective in the 182 Control Subjects (CTL). Data dimensionality can be reduced by removing redundant and unnecessary genes [13]. The goal of the gene selection approach is to identify the small number from genes that generates good results, which can AD's computational

expenses will be reduced, and its effectiveness will increase categorization. In this study, a variety of gene selection techniques, including Information Gain (IG), were used to find helpful genes that are directly associated with illness diagnosis [14].

A basic purpose from gene selection approaches is to reduce the dataset's dimensionality in computing space. Because genes are frequently unimportant, it is a method for choosing a subset of genes from a larger dataset. Whatever the machine learning techniques, these procedures are normally carried out prior to the use of gene selection based on performance measures and machine learning methods. Gene selection method IG was a technique for identifying a selection of genes used in this study's categorization right away.

## 2.2 Information Gain (IG)

Genes that are less helpful in creating a solid predictive model could be eliminated through effective gene selection. Also, genes unrelated to the target variable must be eliminated because they may increase computational costs and hinder the model from performing at its best [15]. The most crucial gene using the Information Gain (IG) method. Differences between entropy and conditional entropy can be used to assess how important a gene is in a certain category [16]. The ability of the predictor variable to categorize the dependent variable is determined by the IG, a kind of filter-based gene selection [17]. The information theory-based IG technique determines the statistical dependence between two variables [18]. The IG important in calculates the reduction in entropy from the transformation of a dataset. It can be used for gene selection by evaluating the information gain of each variable in the context of the target variable.

The IG between the two variables X and Z is formulated mathematically, as Eq. (1):

$$IG(X|Z) = H(X) - H(X|Z) \quad (1)$$

$H(X|Z)$  is the conditional entropy for variable X given Z, while  $H(X)$  is the entropy for variable X. When determining the IG value for an attribute, the conditional entropies for all possible values of that attribute must be subtracted from the target variable's overall dataset entropy [19]. Along with computing the entropy  $H(X)$  and conditional entropy  $H(X|Z)$  for the probability distribution P, as Eq. (2), and Eq. (3):

$$H(X) = - \sum_{x \in X} P(x) \log_2(x) \quad (2)$$

$$H(X|Z) = - \sum_{x \in X} P(x) \sum_{z \in Z} P(x|z) \log_2(P(x|z)) \quad (3)$$

## 2.3 Grey Wolf Optimizer with Crossover (CGWO)

The four sorts of wolves represent the hierarchy of wolves: alpha, beta, delta, and omega. The top, second, the remaining wolves are referred to as omega, while the third, fourth, and fifth wolves are termed alpha, beta, and delta, respectively [20]. When it comes to hunting or optimization in the GWO, alpha, beta, and delta lead. They point out the best places or areas for searching to the other wolves. The alpha, beta, and delta wolves determine the prey's likely position during the iterative search process. The GWO takes into account the hunting, social hierarchy, and searching habits of grey wolves [21]. Because of its decreased unpredictability and different numbers of participants assigned to the global and local search

processes, the GWO method can be easily applied and converges quickly. Data indicates that it performs more efficiently than other bionic algorithms including the PSO method. Because of its improved performance, apps have received more attention [22]. The selections of features and bands, parameter approximation, automatic control, power dispatching, multi-objective optimization, and shop scheduling have all been worked on. However, the typical GWO method was created using the grey wolves' placements, which have equal relevance but lack rigorous conformity with their social order [23]. The binary and multi-objective GWO algorithms are examples of recent advancements in GWO algorithms that tend to blend together and maintain their original form when merged with other algorithms. The GWO algorithm can likely be improved if the grey wolves' search and hunt positions also follow the social hierarchy [24], as shown in Figure 2. Because the proposed GWO gene selection technique seeks to determine the optimal feature subset for medical data, we have chosen to use the GWO in this research.

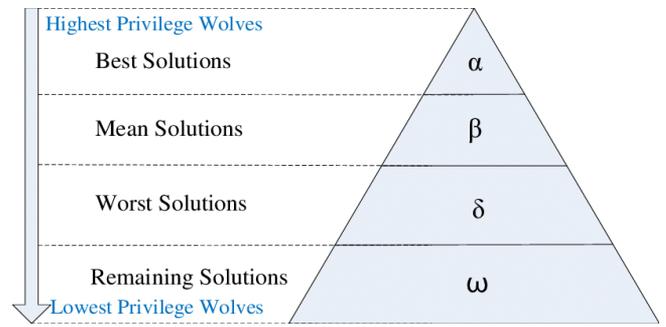


Figure 2. Hierarchy of grey wolf [22]

The updated version of the Grey Wolf Optimizer (GWO) algorithm in this study contains a new step, it can be found at the conclusion of every loop. Stated otherwise, the initial iteration of the GWO is implemented, followed by the crossover operator's enhancement of the current optimal solution, or alpha. An enhanced version of Grey Wolf Optimizer (GWO) based on crossover operators for enhancing the local search process. The suggested method is known as "Grey Wolf algorithm with Crossover (CGWO)", which is given as follows:

Inputs: Obtain every parameter needed to solve a certain optimization issue.

Initialization: Using the search space, randomly generate each search agent using an uniform distribution method, which depends on the upper and lower bounds for the objective function. The generating method is as follows, in Eq. (4):

$$W_i.Position_j = (U - L) * Rand + L \quad (4)$$

where,  $W_i$  represents the  $i^{th}$  search agent,  $Position_j$  represents the  $j^{th}$  dimension in each solution, U and L represent the upper and lower boundaries for the search, and  $Rand$  represents a random number in range 0 and 1. It is worth to mention that that  $i$  is in range  $[1, Size]$ , while  $j$  is in range  $[1, Dim]$ . The values of U and L depend mainly on the nature of the optimization problem.

Determine  $W_{Alpha}$ ,  $W_{Beta}$ , and  $W_{Delta}$ : In this step, the solution should be sorted ascending to identify the best wolves. If there is a new solution with fitness better than the previous, then keep it as  $W_{Alpha}$ , otherwise, keep the previous  $W_{Alpha}$ .

Crossover Operator: In the event that the optimal solution from this iteration is the same as the prior one, then, the alpha is updated via crossover it with the beta as these two solutions are the best two solutions found so far.

The type of encoding used in this study is the real encoding, which means there are specific types of crossover operator to be used. In this study, the Blended Crossover (*BLX - a*). Is utilized for this matter, as follows:

(1) Set P1 and P2 as two parents, where the first parent (P1) denotes the WAlpha, and the second parent (P2) denotes the Wbeta.

(2) Determine Xmax and Xmin, Where Xmax denotes the maximum value between P1 and P2, while Xmin denotes the minimum value between them.

(3) Generate two random solutions as offsprings (C1, C2), in range [R1, R2] where

$$R1=Xmax+Ia \tag{5}$$

$$R2=Xmin-Ia \tag{6}$$

where,

$$I=R1-R2 \tag{7}$$

and a randume value  $\in [0, 1]$ .

(4) Evaluate the new generated offsprings (C1, C2)

(5) Compare the fitness of C1 and C2 with the parents P1, and P2.

(6) Keep the original W Alpha and W Beta and replace the best offspring with the poorest parent, if any, as the new alpha and beta.

## 2.4 Classifier using Support Vector Machine (SVM)

The SVM is utilized to classify a feature after they have been selected, which allows for a method for extracting characteristics that is both effective and efficient as well as a set of criteria for classifier [25]. A unique hyper-plane depicts the discriminative classification technique SVM [26]. Indicate

Figure 3. Due to its excellent precision and ability to interpret data in multiple dimensions [27], many different applications, including bioinformatics, signal processing, and computer vision, make extensive use of the SVM classifier [28].

The SVM excels at addressing the two-class issue. The linear discriminant function's generic formula is represented by as  $w \cdot x + b = 0$ . To distinguish the samples between the two groups without using noise, an optimal hyper-plane is used, which is mathematically described in the Eq. (8).

$$pi[w \cdot x + b] - 1 \geq 0, = 1, 2, N \tag{8}$$

Then, reduce  $lwl \ 2$  in the Eq. (5). As a result, a Lagrange function's saddle point with Lagrange multipliers  $\alpha$  solves an optimization problem.

We denote the ideal discriminant function, as in the Eq. (9).

$$(x) = sign \{ (w * x) + b * \} = sign \{ \sum ai N * i = 1. p (xi * - x) + b * \} \tag{9}$$

Data can be categorized even when they cannot be divided linearly thanks to the Support Vector Machine (SVM), which maps data to a high-dimensional feature space. The data is changed such that the separator can be drawn as a hyperplane when a separator between the various categories is selected. The extreme vectors or points that help create the hyperplane are chosen by the SVM. The approach is referred to as the support vector machine because these extreme situations are known as support vectors. In Figure 3, a decision boundary hyperplane is used to group two distinct categories together.

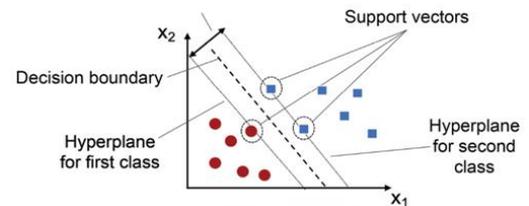


Figure 3. Support Vector Machine (SVM) [29]

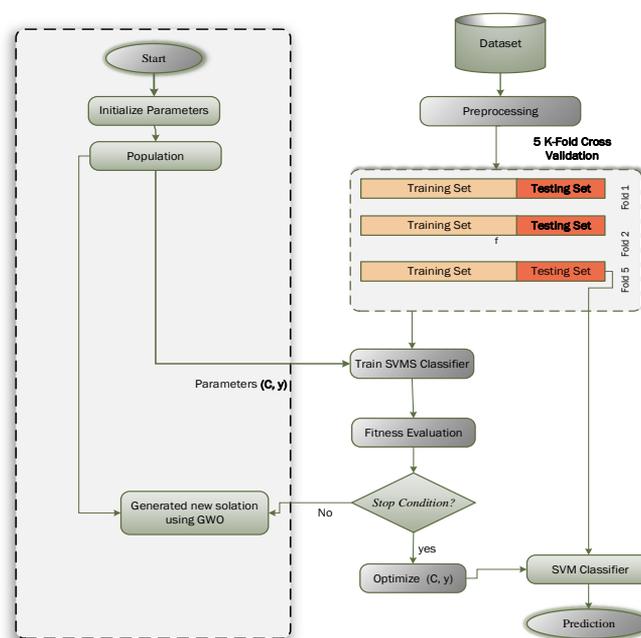


Figure 4. Flowchart GWO for tuning SVM (GSVM)

## 2.5 Grey Wolf Optimizer with Crossover for Support Vector Machine (GSVM)

The main issue with almost all of machine learning models is the over-fitting problem, especially when dealing with high-dimensional datasets. In general, there are two main solutions for such issues, either by reducing the total of genes and keep the most important and relevant subset of features, or, by enhancing the training algorithm, which means tuning the hyper-parameters values if any. The success of any model depends mainly on proper parameterization, and in case of SVM, proper kernel choice [27].

The original version of the GWO has been implemented successfully in developing and solving different optimization problems. However, there is a need to enhance the exploitation ability of the algorithm. Choosing the top three answers is what essentially determines how the GWO is exploited, which are called “ $W_{Alpha}$ ”, “ $W_{Beta}$ ”, and “ $W_{Delta}$ ”.

However, SVM suffers from a simple problem, choosing the kernel require tuning a set of two parameters in order to decrease the chances of over or under-fitting. In order to solve such a problem, an efficient tuning algorithm needs to be developed. An enhancement on GWO's search functionality has been created. A crossover operator decreases the chances of the algorithm of falling in the local optima, because it discovers more regions far from the current best solution once there no new improvements [29].

The CGWO is used for tuning the hyper-plane parameters of support vector machine, to enhance the performance of SVM, this lessens the possibility of over-fitting when a right values to the parameters are found. The main parameters need to be tuned are two only,  $\gamma$  and  $c$ , therefore the solution representation is simple [30]. The steps of the proposed algorithm are exactly same as Algorithm (CGWO); however, it is utilized for a specific problem, represented with two dimensions only. The final model is called GSVM, which is illustrated in the following flowchart Figure 4.

## 2.6 Evaluation measures

Common performance indicators, such as classification accuracy and loss, were employed. True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are four parameters that are used to evaluate the overall prediction ability of a model Eq. (10). This works by looking at the sample numbers and correctly classifying them based on a ratio for the total number of samples tested [31].

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \quad (10)$$

Precision: Is determined by subtracting the false positives from the total of the true positives in the following Eq. (11) [31].

$$\text{Precision} = \frac{TP}{TP+FP} \quad (11)$$

Recall: The total number of true positives multiplied by the sum of the false negatives and true positives in the following Eq. (12).

$$\text{Recall} = \frac{TP}{TP+FN} \quad (12)$$

F measure: Recall and precision are used in a formula to determine accuracy, as demonstrated. in the following Eq. (13):

$$F - \text{measure} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (13)$$

## 3. RESULTS AND DISCUSSION

A proposed method was used IG gene selection methods to find relevant genes. The data on gene expression is read first. After that, the data is normalized using the min-max method. Techniques for selecting genes were used to reduce the number of genes to a level that was near to the sample size. A description of the AD datasets based on the number of unique genes, as well as the genes themselves chosen using IG, for that reason, it is the main algorithm in this study for reducing the sizes of the high-dimensional datasets and keeping the most important genes used Information Gain (IG), as shown in Table 1.

**Table 1.** The selected data's summary

Methods	Sample	Original genes	Selected genes
IG	569	16382	500

When comparing to unprocessed datasets and alternative gene selection methods, the suggested methodology utilizing SVM without IG on the AD dataset, the model can get accuracy (92.34%). The compares the results of the average classification accuracy and loss, as shown in Table 2. The alternative gene selection method using IG outperforms previous gene selection methods, with the SVM model achieved was (94.89%).

**Table 2.** The average accuracy and loss are compared using the AD dataset

Methods	SVM	
	Accuracy	Loss
RawData	84.92%	0.7952
SVM	92.34%	0.5614
IG	94.89%	0.3588

The suggested improvement for the GWO algorithm has been tested in solving a variety of challenging numerical optimization issues. When compared against a number of well-known optimization techniques, including the original GWO algorithm, it performed better.

The following Figures 5(a-d) show the convergence curves for the employed test functions for each proposed technique. Each image is divided into two sections: the Figure 5(a-b) shown the test function's 3D in the first and the convergence of the algorithms in the second (CGWO and the standard GWO), as shown in the Figure 5(c-d). According to the Figure 5(a-d), it can be noticed that CGWO has better performance, and faster convergence towards the global minima solution, because it has better local search ability when hybridized with an evolutionary operator (i.e., crossover).

The primary goal of improving GWO's performance is to create and implement a training strategy for adjusting the key SVM hyper-parameters. Based on the two datasets, the results of CGWO using SVM are shown and discussed in this subsection. Table 3 demonstrates how well the suggested CGWO-SVM algorithm performs in adjusting the SVM hyper-parameters and maintaining the stability of the search process. This indicates that running CGWO-SVM ten times yields results that are nearly identical with minimal variation.

When compared to standard SVM using default parameter settings, for each of the ten run runs, the mean accuracy shows

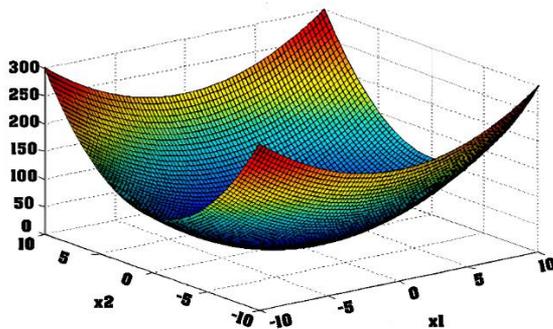
that CGWO-SVM generally obtained extremely good accuracy.

**Table 3.** Performance of CGWO-SVM on two different datasets

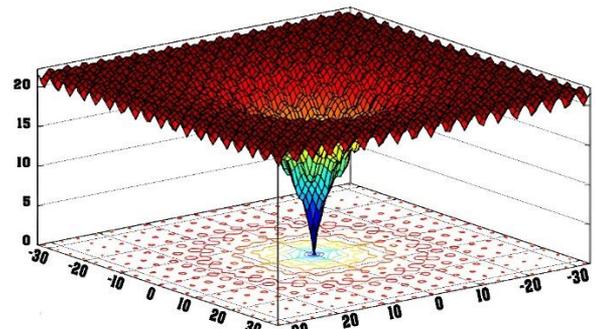
Dataset	SS	MaxItr	Best Acc (%)	Mean Acc (%)	Recall (%)	Precision (%)	F1 (%)
GSE63060	10	100	84.09	81.12	80.92	80.01	79.29
		200	86.17	83.29	82.69	81.00	80.00
		300	88.10	86.23	84.93	82.10	81.05
	20	100	90.54	87.12	89.57	84.30	82.00
		200	93.00	90.23	91.13	85.20	80.61
		300	87.00	85.34	83.97	81.19	81.02
	30	100	89.10	87.23	85.09	83.60	83.51
		200	93.09	91.21	90.73	86.70	84.00
		300	97.41	94.03	93.09	88.00	80.00
GSE63061	10	100	83.01	80.02	82.10	81.00	72.10
		200	85.3	81.20	83.02	81.25	73.10
		300	89.12	83.11	84.25	82.50	76.11
	20	100	90.21	84.01	89.02	83.37	79.10
		200	92.21	86.03	90.00	85.89	82.12
		300	86.07	83.01	84.20	81.21	74.00
	30	100	90.10	84.01	86.36	83.02	80.00
		200	91.08	87.02	89.18	85.10	83.00
		300	96.18	89.11	90.21	87.00	85.00

**Table 4.** The SVM based on the proposed gene selection algorithms

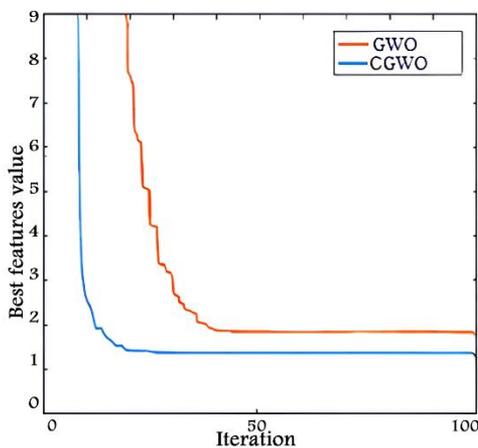
Dataset	Model	Acc (%)	Recall (%)	Precision (%)	F1 (%)
GSE63060	SVM	81.12	80.92	80.01	79.29
	GWO-SVM	83.29	82.69	81.00	80.00
	IG-SVM	86.23	84.93	82.10	81.05
	Proposed	97.41	89.57	84.30	82.00
GSE63061	SVM	80.02	82.10	81.00	72.10
	GWO-SVM	81.20	83.02	81.25	73.10
	IG-SVM	84.01	89.02	83.37	79.10
	Proposed	96.18	90.21	87.00	85.00



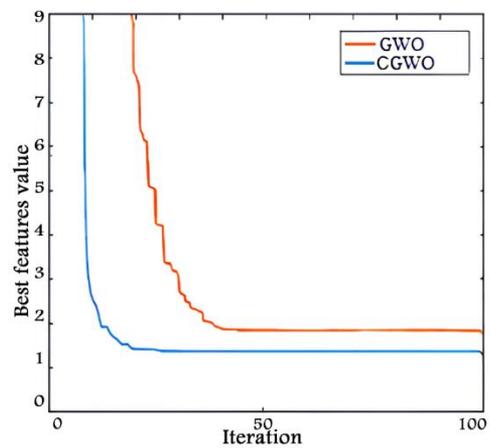
(a)



(b)



(c)



(d)

**Figure 5.** The convergence for proposed method

The results in the previously presented tables are summarized in the following Table 4, which displays a comparison between the best-obtained results with/without CGWO, and SVM based default parameters' values.

#### 4. CONCLUSION

This paper presents the proposed early prediction model for Alzheimer's disease with the highest possible accuracy and the slightest error by selecting the most informative genes. The proposed algorithm, consisting of several sub-algorithms, collaborates to achieve optimal classification accuracy with a minimal subset of selected genes.

The min-max approach is used to standardize the data. For addressing a dimensionality curse and other challenges connected to data nature, However, as most genes are thought to be uninformative about examined classes, and expression data is highly noisy and redundant, only a small subset of genes may have distinct profiles for various classes of samples. Tools to address these difficulties are, therefore, of utmost importance. These tools are trained to strongly identify subsets of informative genes from a large set of highly dimensional data. As this strategy can reduce the data dimensionality and frequently lead to better analyses, the gene selection process is commonly seen in this context as an essential pre-processing step to evaluate these data; the IG is utilized as a gene selection approach. Using the Grey Wolf Optimizer(GWO) and improvement by the crossover operator decrease the probability of local optima trapping the GWO algorithm by updating alpha and beta positions once they get stuck. The CGWO enhanced the classification performance of SVM by tuning the hyper-plane parameters, the final algorithm called CGWO-SVM, and the dataset findings by assembling a subset with valuable information to improve the classification precision. In future work, we can apply the proposed system to other datasets such as diabetes, prostate, breast cancer, etc. The deletion is no longer critical if the deleted genes lack important information or are redundant.

#### REFERENCES

- [1] Park, C., Ha, J., Park, S. (2020). Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset. *Expert Systems with Applications*, 140: 112873. <https://doi.org/10.1016/j.eswa.2019.112873>
- [2] Jameel, N., Abdullah, H.S. (2021). Intelligent feature selection methods: A survey. *Engineering and Technology Journal*, 39(1): 175-183. <https://doi.org/10.30684/etj.v39i1B.1623>
- [3] Madeeh, O.D., Abdullah, H.S. (2021). An efficient prediction model based on machine learning techniques for prediction of the stock market. *Journal of Physics: Conference Series*, 1804(1): 012008. <https://doi.org/10.1088/1742-6596/1804/1/012008>
- [4] Karim, A., Hassan, A., Alawi, M. (2017). Proposed handwriting arabic words classification based on discrete wavelet transform and support vector machine. *Iraqi Journal of Science*, 58(2C): 5511-1168. <https://doi.org/10.24996/ijcs.2017.58.2c.19>
- [5] Sadiq, A.T., Chawishly, S.A., Sulaka, N.J. (2018). Intelligent methods to solve null values problem in databases evolutionary DNA computing algorithm view project cipher applications view project intelligent methods to solve null values problem in databases article info. *Journal of Advanced Computer Science and Technology Research*, 2(2): 91-103.
- [6] Ahmed, H., Soliman, H., Elmogy, M. (2020). Early detection of Alzheimer's disease based on Single Nucleotide Polymorphisms (SNPs) analysis and machine learning techniques. *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy, ICDABI 2020, Sakheer, Bahrain*, pp. 15-20. <https://doi.org/10.1109/ICDABI51230.2020.9325640>
- [7] Mohammed, S.J. (2018). A proposed Alzheimer's disease diagnosing system based on clustering and segmentation techniques. *Engineering and Technology*, 36(2): 160-165. <https://doi.org/10.30684/etj.36.2B.12>
- [8] Perera, S., Hewage, K., Gunarathne, C., Navarathna, R., Herath, D., Ragel, R.G. (2020). Detection of novel biomarker genes of Alzheimer's disease using gene expression data. *MERCon 2020 - 6th International Multidisciplinary Moratuwa Engineering Research Conference, Proceedings, Moratuwa, Sri Lanka*, pp. 1-6. <https://doi.org/10.1109/MERCon50084.2020.9185336>
- [9] Muhammed Niyas, K.P., Thiyagarajan, P. (2021). Feature selection using efficient fusion of Fisher Score and greedy searching for Alzheimer's classification. *Journal of King Saud University-Computer and Information Sciences*, 34(8): 4993-5006. <https://doi.org/10.1016/j.jksuci.2020.12.009>
- [10] Abdulwahab, A., Attya, H., Ali, Y.H. (2020). Documents classification based on deep learning. *International Journal of Scientific & Technology Research*, 9(2): 62-66.
- [11] El-Gawady, A., Makhlof, M.A., Tawfik, B.S., Nassar, H. (2022). Machine learning framework for the prediction of Alzheimer's disease using gene expression data based on efficient gene selection. *Symmetry*, 14(3): 491. <https://doi.org/10.3390/sym14030491>
- [12] Khadhim, B.J., Kadhim, Q.K., Shams, W.K., Ahmed, S.T., Wahab Alsiadi, W.A. (2023). Diagnose COVID-19 by using hybrid CNN-RNN for chest X-ray. *Indonesian Journal of Electrical Engineering and Computer Science*, 29(2): 852-860. <https://doi.org/10.11591/ijeecs.v29.i2.pp852-860>
- [13] Mahajan, S., Bangar, G., Kulkarni, N. (2020). Machine learning algorithms for classification of various stages of Alzheimer's disease: A review. *International Research Journal of Engineering and Technology (IRJET)*, 7(8): 817-824.
- [14] Nazeeh, I., Hadi, T.H., Mohammed, Z.Q., Ahmed, S.T., Kadhim, Q.K. (2023). Optimizing blockchain technology using a data sharing model. *Indonesian Journal of Electrical Engineering and Computer Science*, 29(1): 431-440. <https://doi.org/10.11591/ijeecs.v29.i1.pp431-440>
- [15] Andono, P.N., Shidik, C.F., Prabowo, D.P., Pergiawati, D., Pramunendar, R.A. (2022). Bird voice classification based on combination feature extraction and reduction dimension with the K-nearest neighbor. *International Journal of Intelligent Engineering and Systems*, 15(1): 262-272. <https://doi.org/10.22266/ijies2022.0228.24>
- [16] Kavitha, C., Mani, V., Srividhya, S.R., Khalaf, O.I. (2022). Early-stage Alzheimer's disease prediction using

- machine learning models. *Front. Public Health*, 10: 1-13. <https://doi.org/10.3389/fpubh.2022.853294>
- [17] Kadhim, Q.K. (2017). Classification of human skin diseases using data mining. *International Journal of Advanced Engineering Research and Science*, 4(1): 159-163. <https://doi.org/10.22161/ijaers.4.1.25>
- [18] Dhandapani, P. (2022). Multi-channel convolutional neural network for prediction of leaf disease and soil properties. *International Journal of Intelligent Engineering and Systems*, 15(1): 318-328. <https://doi.org/10.22266/ijies2022.0228.29>
- [19] Ebiaredoh-mienye, S.A., Swart, T.G., Esenogho, E. (2022). A machine learning method with filter-based feature selection for improved prediction of chronic Kidney disease. *Bioengineering*, 9(8): 350. <https://doi.org/10.3390/bioengineering9080350>
- [20] Long, W., Cai, S., Jiao, J., Tang, M. (2020). An efficient and robust grey wolf optimizer algorithm for large-scale numerical optimization. *Soft Computing*, 24(2): 997-1026. <https://doi.org/10.1007/s00500-019-03939-y>
- [21] Mohammed, H.A.A., Nazeeh, I., Alisawi, W.C., Kadhim, Q.K., Ahmed, S.T. (2023). Anomaly detection in human disease: A hybrid approach using GWO-SVM for gene selection. *Revue d'Intelligence Artificielle*, 37(4): 913-919. <https://doi.org/10.18280/ria.370411>
- [22] Zangenehjad, F., Jiang, Y., Gao, Y. (2023). GNSS observation generation from smartphone android location API: Performance of existing apps, issues and improvement. *Sensors*, 23(2): 777. <https://doi.org/10.3390/s23020777>
- [23] Gao, Z.M., Zhao, J. (2019). An improved grey Wolf optimization algorithm with variable weights. *Computational Intelligence and Neuroscience*, 2019: 2981282. <https://doi.org/10.1155/2019/2981282>
- [24] Wang, J.S., Li, S.X. (2019). An improved grey wolf optimizer based on differential evolution and elimination mechanism. *Scientific Reports*, 9(1): 1-21. <https://doi.org/10.1038/s41598-019-43546-3>
- [25] Taha Ahmed, S., Malallah Kadhem, S. (2021). 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)*, 15(16): 95. <https://doi.org/10.3991/ijim.v15i16.24191>
- [26] Ahmed, S.T., Kadhem, S.M. (2022). Early Alzheimer's disease detection using different techniques based on microarray data: A review. *International Journal of Online and Biomedical Engineering (IJOE)*, 18(4): 106-126. <https://doi.org/10.3991/ijoe.v18i04.27133>
- [27] Tahseen Ali, A., Abdullah, H.S., Fadhil, M.N. (2021). Voice recognition system using machine learning techniques. *Materials Today: Proceedings*, 47-60. <https://doi.org/10.1016/j.matpr.2021.04.075>
- [28] Ahmed, S.T., Kadhem, S.M. (2023). Optimizing Alzheimer's disease prediction using the nomadic people algorithm. *International Journal of Electrical and Computer Engineering (IJECE)*, 13(2): 2052-2067. <https://doi.org/10.11591/ijece.v13i2.pp2052-2067>
- [29] Dhahi, S.H., Dhahi, E.H., Khadhim, B.J., Ahmed, S.T. (2023). Using support vector machine regression to reduce cloud security risks in developing countries. *Indonesian Journal of Electrical Engineering and Computer Science*, 30(2): 1-8. <https://doi.org/10.11591/ijeecs.v30.i2.pp1-1x>
- [30] Hameed, E.M., Hussein, I.S., Altameemi, H.G.A., Kadhim, Q.K. (2022). Liver disease detection and prediction using SVM techniques. *2022 3rd Information Technology to Enhance E-Learning and Other Application (IT-ELA)*, Baghdad, Iraq, pp. 61-66. <https://doi.org/10.1109/IT-ELA57378.2022.10107961>
- [31] Ahmed, S.T., Kadhim, Q.K., Alsultani, H.S.M., Abd almahdy, W.S. (2021). Applying the MCMSI for online educational systems using the two-factor authentication. *International Journal of Interactive Mobile Technologies (IJIM)*, 15(13): 162. <https://doi.org/10.3991/ijim.v15i13.23227>