A Hybrid Grey Wolf-Whale Optimization Algorithm for Classification of Corona Virus Genome Sequences using Deep Learning

Muthulakshmi Murugaiah
Department of Computer Science and Engineering,
Kalasalingam Academy of Research and Education, India
m.muthulakshmi@klu.ac.in

Murugeswari Ganesan
Department of Computer Science and Engineering,
Manonmaniam Sundaranar University, India
gmurugeswari@msuniv.ac.in

Abstract: *Genome sequence data is widely accepted as complex data and is still growing in an exponential rate. Classification of genome sequences plays a crucial role as it finds its applications in the area of biology, medical and forensics etc. For classification, Genome sequences can be represented in terms of features. More number of less significant features leads to lower accuracy in classification task. Feature selection addresses this issue by selecting the most important features which aids to improve the accuracy and lessens the computational complexity. In this research, Hybrid Grey Wolf-Whale Optimization Algorithm (HGWWOA) is proposed for Genome sequence classification. The proposed algorithm is evaluated using 23 benchmark objective functions along with Convolutional Neural Network classifier and its efficiency is verified using a novel metric namely "Feature Reduction Rate". The proposed optimization algorithm can be applied for any optimization problems. In this research work, the proposed algorithm is used for classification of Corona Virus genome sequences. Performance comparison of the proposed and existing algorithms was carried out and it is evident that the performance of proposed algorithm exceeds the previous algorithms with an accuracy of 98.2%.*

Keywords: *Corona virus genome, deep learning, feature selection, GWO, hybrid optimization, WOA.*

Received July 2, 2020; accepted December 15, 2022
https://doi.org/10.34028/iajit/20/3/5

1. Introduction

Classification of genome sequences gains extensive attention as it finds its applications in significant areas such as Disease prediction, Genome analysis and Forensics etc. Feature extraction helps to represent the lengthy Genome sequences in terms of set of quantitative values which is called Feature vector and can be used for further processing like classification. The high complex genomic data may consist of large number of irrelevant features which may results in lower classification accuracy. So, there should be a trade off made between the number of features and accuracy. Feature selection helps to upgrade the classification accuracy by selecting the more relevant features for classification. It also prevents complex calculations by minimizing the number of features in a dataset [4]. Three types of feature selection methods are available namely Filter, Wrapper and Embedded methods. In the filter method, each feature is assigned with a relevance score and then the feature with high score is selected. In case of wrapper methods, the classifier's accuracy is used as a measure for selecting the features [3]. Embedded method embeds the feature selection process within the classifier construction [16]. Embedded method combines the qualities and advantages of filter and wrapper feature selection methods [28]. Optimization algorithms are the under the category of Wrapper feature selection method. Bio-inspired optimization approach is the evolving research area which is derived from the concepts and inspiration of the nature's biological evolution to design optimization techniques. There are numerous bio-inspired optimization algorithms that are developed and designed to imitate the group of insects or animals by defining and applying deterministic or stochastic rules in solving different optimization problems [24] and is also called as Swarm Intelligence optimization algorithms. Optimization algorithms can decrease the computational complexity while retaining or improving the classification accuracy [1]. The objective of the research is to develop an efficient optimization technique for feature reduction without the degradation of classification performance. The proposed optimization technique is tested for corona virus genome sequence classification. The main contribution of this research is that the benefits of two different optimization techniques are combined which can effectively minimizes the number of features for classification. In order to evaluate the proposed optimization technique, 23 benchmark objective functions and a deep learning algorithm are used. Additionally, the best objective function is also identified.

2. Related Works

A survey on important principles of feature selection along with the most significant and current applications was presented by Wang *et al*. [31]. Chuzhunova *et al*. [6] designed a feature selection technique using Gamma test. This method was adopted for feature selection and classification of large sub units of ribosomal Ribo Nucleic Acid (rRNA). A comprehensive survey of feature selection techniques especially applicable to genomic data was carried out by Tadist *et al*. [28]. Feature selection aids in reducing the complexity in analysing the genome data. Lo Bosco and Pinello [15] designed a feature selection technique for DNA sequences which depends on Motif Independent Measure (MIM). This technique utilize k-mers counting for feature selection and picked the more relevant k-mers correlated to the testing dataset by calculating Kullback-Leibler divergence between the training and testing dataset.

Aghdam *et al*. [2] presented a method to select the features for classification of biological dataset by combining Ant Colony Optimization and Bayesian classification algorithms. Garcia-Diaz *et al*. [9] presents a feature selection algorithm namely Grouping Genetic Algorithm (GGA) for classification of cancer using RNA sequence data. GGA combines an Extreme Learning Machine and the objective function to evaluate the performance of feature selection. A precise and quick feature selection method for omics data was proposed by Perez-Riverol *et al*. [22]. The author developed a dynamic framework and an R package named feseR to solve the feature selection problems. Leclercq *et al*. [13] developed BioDiscML which is a Biomarker discovery tool. This tool helps in automatic selection of features for the discovery of Biomarkers in omics data.

A novel framework for feature selection was developed by Chatzilygeroudis *et al*. [5] which is based on Genetic Algorithm. The method was applied to Single Cell RNA sequence to reduce the complexity in analysis. A hybrid feature selection method which combines Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) was introduced by Elsadekl *et al*. [7]. The hybrid PSO-GA approach helps in dimensionality reduction of DNA Copy Number Variation (CNV) data. Qin *et al*. [23] developed a framework for a feature selection by combining the traditional feature selection categories such as filter, wrapper and embedded. In the first phase, Gene Co-expression Network (WGCNA), random forest and minimal redundancy maximal relevance (mRMR) are utilized whereas the second phase uses improved binary Salp Swarm Algorithm. OmicSelector, a web application was developed by Stawiski *et al*. [27]. This application has the capability to perform both feature selection and classification for genomic data.

3. Outline of the Proposed Work

The input Corona virus genome sequences are collected from the National Centre for Biotechnology Information (NCBI) database [11]. A total of 120 features are extracted from the genome sequences using Frequency based Feature Extraction Technique (FFET) [20]. Afterwards, the extracted features are given as input to the proposed A Hybrid Grey Wolf-Whale Optimization Algorithm (HGWWOA) for feature selection. Subsequently, the selected features are fed into Convolutional Neural Network (CNN) [21] for classification. Finally, the performance of the feature selection with classifier is estimated using various performance metrics. The outline or architecture of the proposed work is shown in Figure 1.

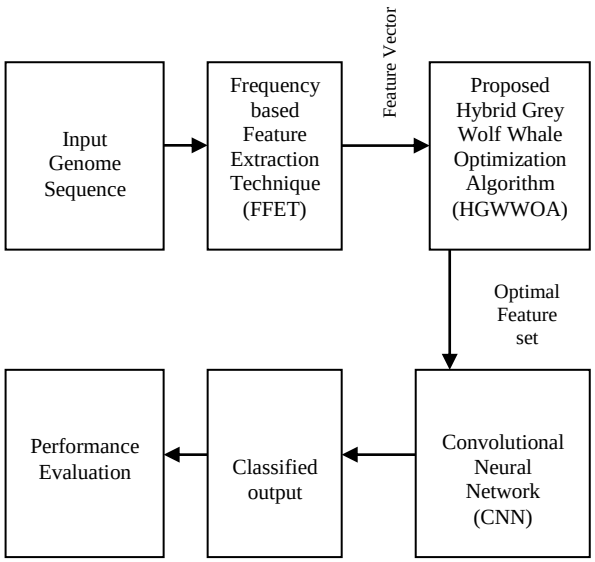

Figure 1. Architecture of the proposed work.

4. Proposed Hybrid Grey Wolf-Whale Optimization Algorithm (HGWWOA)

The basic intention of this work is to propose a hybrid model of two swarm intelligence optimization algorithms. In general, there are two phases for swarm intelligence optimization algorithms namely exploration phase and exploitation phase. In HGWWOA, the exploration phase is similar to Grey Wolf Optimization Algorithm (GWO) [18] and is followed by the exploitation phase which is similar to Whale Optimization Algorithm (WOA) [17, 19]. The process flow of HGWWOA is shown in Figure 2.

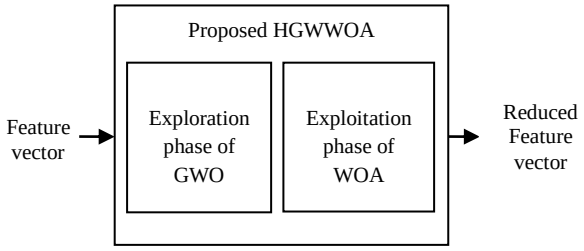

Figure 2. Process flow of proposed HGWWOA.

The exploration phase denotes the process of identifying the optimistic area of the search space as widely as possible. However, the exploitation indicates the capability of local search around the promising regions acquired in the exploration phase. At first, the population is initialized to investigate the search space to identify the global solution area as per Grey Wolf optimization concept and hence the best solution is obtained. The first or best fittest solution is agreed as alpha (α), next best fittest solution is taken as beta (β) and the following next best fittest solution is accepted as delta. The remaining candidate solutions are assumed to be omega (δ). This is known as leadership quality of wolves. The exploitation phase in the proposed HGWWOA which is similar to Whale optimization algorithm is supervised by alpha, beta and delta. On behalf of this modification, the hybrid approach outperforms the individual algorithms. The mathematical representation of Grey Wolf Optimization algorithm for encircling the prey is expressed below

$$\begin{gathered}\vec{D} = |\vec{Z} \cdot \vec{X}_f(t) - \vec{X}(t)| \\ \vec{X}(t+1) = \vec{X}_f(t) - \vec{Y} \cdot \vec{D} \\ \vec{Y} = 2 \cdot \vec{v} \cdot \vec{r_1} - \vec{v} \\ \vec{Z} = 2 \cdot \vec{r_2}\end{gathered} \tag{1}$$

Where *t* refers to ongoing iteration, $\vec{X}_f(t)$ refers to the position vector of prey or best solution at iteration t. $\vec{X}(t+1)$ is the position vector at the current iteration. $\vec{X}$ refers to the search agent's position vector. $\vec{D}$ is the Distance vector which represents the distance between prey and the search agent. $\vec{Y}$ and $\vec{Z}$ are the co-efficient vectors. | | is used to obtain the absolute value. The value of $\vec{v}$ is decreased from two to zero through iterations, $\vec{r_1}$ and $\vec{r_2}$ are the random vectors lies in the range zero to one. The position of the search agent is calculated using the following formula

$$\begin{aligned}&\textit{Position } (X_i) = \textit{rand } (n, \textit{Dim}) \times (\textit{Upper Bound} - \\ &\textit{Lower Bound}) + \textit{Lower Bound}\end{aligned} \tag{2}$$

Where n refers to the population size (total number of instances) and Dim refers to the attributes which ranges from 1 to 120 in this research work. Upper bound and lower bound are the ranges or boundary of the function's search space. The first three best solutions obtained so far are taken as alpha, beta and delta. The hunting is usually supervised by the alpha which has more wisdom about the prey. The beta and delta also takes part in hunting and it has better awareness about the prey. The positions of the Omega can be updated as reported by the best search agent. The hunting behavior of Grey Wolf is mathematically devised as shown below

$$\begin{gathered}\vec{D_\alpha} = |\vec{Z_1} \cdot \vec{X_\alpha} - \vec{X}|, \ \vec{D_\beta} = |\vec{Z_2} \cdot \vec{X_\beta} - \vec{X}|, \\ \vec{D_\delta} = |\vec{Z_3} \cdot \vec{X_\delta} - \vec{X}| \\ \vec{X}_1 = \vec{X_\alpha} - \vec{Y_1} \cdot (\vec{D_\alpha}), \ \vec{X}_2 = \vec{X_\beta} - \vec{Y_2} \cdot (\vec{D_\beta}) \\ \vec{X}_3 = \vec{X_\delta} - \vec{Y_3} \cdot (\vec{D_\delta}) \\ \vec{X}(t+1) = \frac{\vec{X_1} + \vec{X_2} + \vec{X_3}}{3}\end{gathered} \tag{3}$$

In case of Search for Prey, the search agent make a move towards the prey to attack when the value of $|Y| < 1$. The search agent does a random search for the prey when the value of $|Y| \geq 1$ and this action is mathematically designed as

$$\begin{gathered}\vec{D} = |\vec{Z} \cdot \vec{X}_{rand} - \vec{X}| \\ \vec{X}(t+1) = \vec{X}_{rand} - \vec{Y} \cdot \vec{D}\end{gathered} \tag{4}$$

In the exploitation phase of HGWWOA, the spiral updating position in Whale Optimization Algorithm is done using the leadership hierarchy of GWO. The mathematical version of this approach is given below

$$\begin{gathered}\vec{D'_\alpha} = |\vec{X}_\alpha(t) - \vec{X}(t) \qquad \vec{D'_\beta} = |\vec{X}_\beta(t) - \vec{X}(t) \\ \vec{D'_\delta} = |\vec{X}_\delta(t) - \vec{X}(t) \\ \vec{D'}_{avg} = \frac{\vec{D'}_\alpha + \vec{D'}_\beta + \vec{D'}_\delta}{3} \\ \vec{X}_{avg}(t) = \frac{\vec{X}_\alpha(t) + \vec{X}_\beta(t) + \vec{X}_\delta(t)}{3} \\ \vec{X}(t+1) = e^{bh} \cdot \cos(2\pi h) \cdot \vec{D'}_{avg} + \vec{X}_{avg}(t)\end{gathered} \tag{5}$$

Where $\vec{D'}$ refers to the distance between the search agent and prey, b is the constant value and h denotes the random number in [-1,1]. The mathematical approach for updating the positions in proposed HGWWOA is summarized as follows

$$\vec{X}(t+1) = \begin{cases} \frac{\vec{X_1} + \vec{X_2} + \vec{X_3}}{3} & \text{if } p < 0.5 \\ e^{bh} \cdot \cos(2\pi h) \cdot \vec{D'}_{avg} + \vec{X}_{avg}(t) & \text{if } p \geq 0.5 \end{cases} \tag{6}$$

Where *p* signifies a random number which ranges from zero to one. The pseudo code of HGWWOA is given in Algorithm (1).

Algorithm 1. HGWWOA

Initialize the population X_i where i ranges from 1 to n
Initialize the parameters v,Y,Z,h and f
Compute the fitness of all search agents
Let X_α be the first best search agent
Let X_β be the second best search agent
Let X_δ be the third best search agent
while (t<100)
  for each search agent
  Update control parameter (v,Y,Z,h and f)
  If1 (f < 0. 5)
    If2 (|Y| < 1)
      Update the search agent's position using (2)
    else if2 |Y| ≥ 1)

Select (X_{rand})
Update the search agent's position using (3)
end if2
else if1 ($f \geq 0.5$)
Update the search agent's position using (4)
end if1
end for
Check & amend the search agents far away from the search space
Compute the fitness of amended search agents
Update the positions of $X_\alpha, X_\beta, X_\delta$
$t = t+1$
end while
Return X_α

5. Classification Using Proposed HGWWOA and Deep Learning

Machine learning and deep learning algorithms have accomplished high performance in the area of genome biology [14]. CNN, a deep learning model is utilized to evaluate the efficiency of proposed HGWWOA. CNN consists of three layers which is Convolutional layer, Pooling layer and fully connected layer. Convolutional layer is the basic building block of CNN which contains kernel which helps to extract features from the input. Max pooling and Softmax activation function is used in the pooling layer and fully connected layer respectively for the purpose of classification. The number of hidden layers and epochs in CNN is set to 100. The dataset is split into training and testing dataset using K fold Cross Validation for the purpose of classification where the value of K is set 10.

6. Experimental Result Analysis and Performance Evaluation

The dataset used for the experiment, results of the proposed and existing optimization algorithms and their performance with CNN classifier are described in this section.

6.1. Dataset

COVID-19 is a pandemic disease caused by a virus called Corona virus. It is found that there are 7 human infecting Corona virus strains are available. They are HCoV-229E, HCoV-OC43, HCoV-HKU1, HCoV-NL63, MERS-CoV, SARS-CoV and SARS2-CoV2 [10]. SARS2-CoV2 is the novel human infecting Corona virus which was named as COVID-19. For experimental analysis, samples of 1000 Corona virus Genome sequences which includes all the seven strains are downloaded from NCBI database and it includes the Genome sequences of COVID-19 patients of various countries across the world.

6.2. Performance Evaluation

The proposed optimization algorithm is evaluated by implementing it with various objective (or) fitness functions along with CNN classifier and by calculating the performance metrics for feature reduction and classification. The performance of HGWWOA is juxtaposed against the existing optimization algorithms.

6.3. Performance Metrics

The performance of optimization is evaluated using ametric called Feature Reduction Rate (FRR). Feature Reduction Rate is the novel measure introduced in this work to measure the features reduction capability of optimization technique. It is defined using the number of features selected by the algorithm and the total number of features. The formula for calculating FRR is as follows.

$$FRR = \frac{Total\ number\ of\ Features - Number\ of\ features\ selected}{Total\ Number\ of\ Features} \times 100 \quad (7)$$

The metrics used to measure the performance of classification algorithm are shown in Table 1.

Table 1. Performance metrics.

Sl. No.	Performance Metrics	Formula
1	Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
2	Specificity	$\frac{TN}{TN+FP}$
3	Precision	$\frac{TP}{TP+FP}$
4	Recall	$\frac{TP}{TP+FN}$
5	F1 Score	$2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$
6	Error Rate	$\frac{FP+FN}{TP+TN+FN+FP}$

Where TP, TN, FP, and FN refers to True Positive, True Negative, False Positive and False Negative respectively.

6.4. Performance Analysis of HGWWOA using Benchmark Objective Functions

An objective function is used as a single figure of merit to summarise how close a proposed design solution is to achieving the set goals. Objective functions are used to guide simulations towards optimal design solutions. To evaluate the performance of the proposed HGWWOA in detail, 23 standard benchmark functions [17, 18] are utilized. The position of the search agent (x_i) was calculated using Equation (1). Upper bound and lower bound are the ranges, boundary of the function's search space associated with each objective function which is given in Table 2. Values of x_i with respect to ranges can be applied to each fitness or objective function which provides different results for

optimization. The performance of the 23 benchmark objective functions (or) fitness functions with the proposed HGWWOA is evaluated using CNN classifier and the accuracy is shown in Table 2. In Table 2, functions f_1 to f_7 are called as Unimodal benchmark functions. Functions f_8 to f_{13} are called multimodal benchmark functions f_{14} and f_{23} to are called as fixed dimensional multimodal benchmark functions. Experimental results show that the HGWWOA works efficient with the objective function f_2 which selects 22 features from 120 features and shows an accuracy of 98.2%. The functions f_1, f_7, f_{10}, f_{14} and f_{18} gives an average accuracy of 90% and above. The objective function f_2 is used in this research work for optimization.

$$f_2(x)=\sum_{i=1}^{n}|x_i|+\prod_{i=1}^{n}|x_i| \tag{8}$$

Where (x_i) represents the search agent's position and n refers to the population size.

6.5. Performance Evaluation of Existing Optimization Algorithms and Proposed HGWWOA

The performance of optimization is also evaluated by FRR. FRR of HGWWOA is compared with existing individual algorithms such as Grey Wolf Optimization Algorithm (GWO), WOA, Salp Swarm optimization Algorithm (SSA) [8], PSO algorithm [30] and few hybrid optimization algorithms. Performance comparison of FRR is shown in Figures 3 and 4.

In Table 2, in the function $f_{12}(x)$, $y_i = 1+\frac{x_i+1}{4}$

$$u(x_i,a,k,m)=\begin{cases} k(x_i-a)^m & x_i > a \\ 0 & -a < x_i < a \\ k(-x_i-a)^m & x_i <- a \end{cases} \tag{9}$$

Where x_i refers to the position of the search agent.

From Figures 3 and 4, it is obvious that the FRR of proposed HGWWOA is higher when compared to existing individual as well as hybrid optimization algorithms. The performance of the proposed and existing optimization algorithms such as GWO, WOA, SSA and PSO with CNN classifier is shown in Table 3. The performance of the proposed HGWWOA is also juxtaposed with the performance of the other existing hybrid algorithms which was shown in Table 4. From Table 4, it is observed that the exploitation phase of Whale Optimization algorithm performs better as the HPSO-WOA and the proposed HGWWOA gives better accuracy than other existing hybrid algorithms. From Tables 3 and 4, it is clear that the proposed hybrid algorithm outperforms the former individual as well as hybrid algorithms both in terms of FRR and classification accuracy.

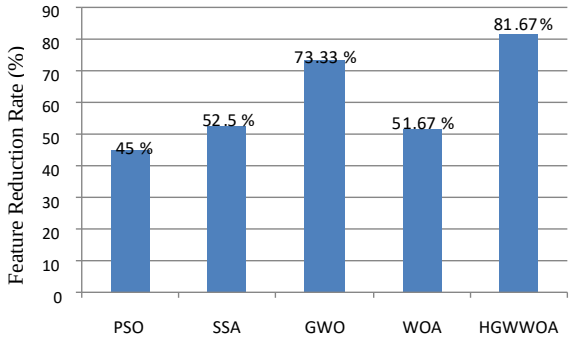

Figure 3. FRR comparison of HGWWOA with existing algorithms.

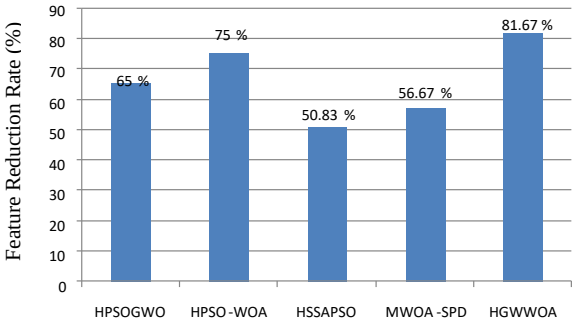

Figure 4. FRR comparison of HGWWOA with existing hybrid algorithms.

Table 2. Performance of proposed HGWWOA with various objective functions (Total Number of features-120).

Sl. No	Function	Range	No. of Features selected	Accuracy (CNN) (%)
1	$f_1(x)=\sum_{i=1}^{n} x_i^{\,2}$	[-100,100]	20	95.40
2	$f_2(x)=\sum_{i=1}^{n}\lvert x_i\rvert+\prod_{i=1}^{n}\lvert x_i\rvert$	[-10,10]	**22**	**98.20**
3	$f_3(x)=\sum_{i=1}^{n}\left(\sum_{j-1}^{i} x_j^{\,2}\right)$	[-100,100]	36	88.00
4	$f_4(x)=\max_i\left\{\lvert x_i\rvert, 1\le i\le n\right\}$	[-100,100]	23	89.60
5	$f_5(x)=\sum_{i=1}^{n-1}[100(x_{i+1}-x_i^{\,2})^2+(x_i-1)^2$	[-30,30]	37	75.00
6	$f_6(x)=\sum_{i=1}^{n}([x_i+0.5])^2$	[-100,100]	36	83.33
7	$f_7(x)=\sum_{i=1}^{n} i x_i^{\,4}+random[0,1]$	[-1.28,1.28]	25	93.33
8	$f_8(x)=\sum_{i=1}^{n} x_i\sin(\sqrt{\lvert x_i\rvert})$	[-500,500]	44	80.00
9	$f_9(x)=\sum_{i=1}^{n}[x_i^{\,2}-10\cos(2\pi x_i)+10]$	[-5.12,5.12]	55	78.00
10	$f_{10}(x)=20\exp(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2}})-\exp(\frac{1}{n}\sum_{i=1}^{n}\cos(2\pi x_i))+20+e$	[-32,32]	33	93.20
11	$f_{11}(x)=\frac{1}{4000}\sum_{i=1}^{n} x_i^{\,2}-\prod_{i=1}^{n}\cos(\frac{x}{\sqrt{i}})+1$	[-600,600]	42	86.40
12	$f_{12}(x)=\frac{\pi}{n}\left\{10\sin(\pi y_1)+\sum_{i=1}^{n-1}(y_i-1)^2[1+\sin^2(\pi y_{i+1})]+(y_n-1)^2\right\}+\sum_{i=1}^{n} u(x_i,10,100,4)$	[-50,50]	63	84.40
13	$f_{13}(x)=0.1\{\sin^2(3\pi x_1)+\sum_{i=1}^{n}(x_i-1)^2[1+\sin^2(3\pi x_i+1)]+(x_n-1)^2[1+\sin^2(2\pi x_n)]\}+\sum_{i=1}^{n} u(x_i,5,100,4)$	[-50,50]	43	89.60
14	$f_{14}(x)=-\sum_{i=1}^{n}\sin(x_i)\cdot\left(\left(\sin(\frac{i x_i^{\,2}}{\pi})\right)\right)^{2m}, m=10$	$[0,\pi]$	35	92.00
15	$f_{15}(x)=\left[e^{-\sum_{i=1}^{n}\left(x_i/\beta\right)^{2m}}-2e^{-\sum_{i=1}^{n} x_i^{\,2}}\right]\cdot\prod_{i=1}^{n}\cos^2 x_i, m=5$	[-20,20]	43	84.00
16	$f_{16}(x)=\left\{\left\lvert\sum_{i=1}^{n}\sin^2(x_i)\right\rvert-\exp(-\sum_{i=1}^{n} x_i^{\,2})\right\}\cdot\exp\left\lvert-\sum_{i=1}^{n}\sin^2\sqrt{\lvert x_i\rvert}\right\rvert$	[-10,10]	31	80.00
17	$f_{17}(x)=\left(x_2-\frac{5.1}{4\pi^2}x_1^{\,2}+\frac{5}{\pi}x_1-6\right)^2+10\left(1-\frac{1}{8\pi}\right)\cos x_1+10$	[-5,5]	59	88.13
18	$f_{18}(x)=\left[1+(x_1+x_2+1)^2(19-14x_1+3x_1^{\,2}-14x_2+6x_1x_2+3x_2^{\,2})\right]\times\left[30+(2x_1-3x_2)^2\times(18-32x_1+12x_1^{\,2}+48x_2-36x_1x_2+27x_2^{\,2})\right]$	[-2,2]	36	90.63
19	$f_{19}(x)=-\sum_{i=1}^{4} c_i\exp\left(-\sum_{j=1}^{3} a_{ij}\left(x_j-p_{ij}\right)^2\right)$	[1,3]	41	78.33
20	$f_{20}(x)=-\sum_{i=1}^{4} c_i\exp\left(-\sum_{j=1}^{6} a_{ij}\left(x_j-p_{ij}\right)^2\right)$	[0,1]	48	86.20
21	$f_{21}(x)=-\sum_{i=1}^{5}\left[(X-a_i)(X-a_i)^T+c_i\right]^{-1}$	[0,10]	39	81.19
22	$f_{22}(x)=-\sum_{i=1}^{7}\left[(X-a_i)(X-a_i)^T+c_i\right]^{-1}$	[0,10]	43	75.00
23	$f_{23}(x)=-\sum_{i=1}^{10}\left[(X-a_i)(X-a_i)^T+c_i\right]^{-1}$	[0,10]	42	74.40

Table 3. Performance of proposed HGWWOA and existing optimization algorithms.

Sl. No.	Optimization Techniques	Total Number of features	No. of Features selected	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1 Score (%)	Error Rate (%)
1	PSO	120	66	92.93	66.67	100.00	91.76	80.00	7.07
2	SSA	120	57	94.60	95.48	71.43	75.00	78.13	5.40
3	GWO	120	32	96.40	79.45	80.56	95.14	88.55	3.60
4	WOA	120	58	95.92	77.78	100.00	92.24	87.50	4.08
5	**HGWWOA**	120	**22**	**98.20**	**87.50**	**100.00**	**97.65**	**93.33**	**1.80**

Table 4. Performance of proposed HGWWOA and existing hybrid optimization algorithms.

Sl. No.	Optimization Algorithms	Algorithms for Hybridization	Total Number of features	No. of Features selected	Accuracy (%) (CNN)	Error Rate (%)
1	HPSOGWO [25]	PSO & GWO	120	42	89.20	10.80
2	HPSO-WOA [29]	PSO & WOA	120	30	96.80	3.20
3	HSSAPSO [26]	SSA & PSO	120	59	90.32	9.68
4	MWOA-SPD [12]	WOA & SSA	120	52	93.94	6.06
5	**Proposed HGWWOA**	**GWO & WOA**	**120**	**22**	**98.20**	**1.80**

The performance of HGWWOA with CNN is also evaluated for different number of epochs and the result is shown in Table 5. When the epoch is set to 200 and 400, there is no much difference in accuracy while for 500 epochs, the classifier provides an accuracy of 99.10, but it is a time consuming process. So, by considering the running time, the proposed HGWWOA utilized CNN with 100 epoch.

Table 5. Performance of HGWWOA and CNN for different epochs.

No. of epochs	Accuracy
100	98.20
300	98.60
500	99.10

The improvement in classification accuracy is proved by comparing the performance of the proposed HGWWOA before and after applying optimization. The classification accuracy using CNN classifier for the Corona virus genome dataset before and after optimization was shown in Figure 4.

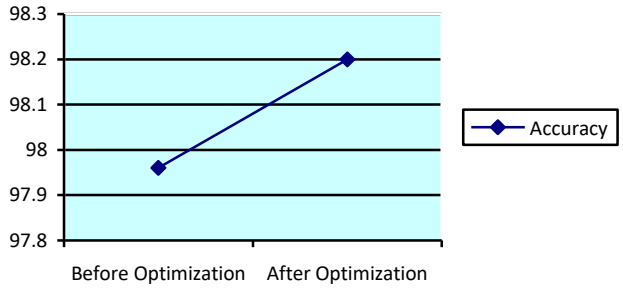

Figure 4. Performance of HGWWOA.

From Figure 4, it is obvious that the classification accuracy improves after performing optimization using the proposed HGWWOA.

7. Conclusions

Classification of genome sequences plays a vital role in predicting the genetic diseases. HGWWOA is proposed for classification of Corona virus genome sequences which helps in feature reduction for classification process without compromising the accuracy. This is because the proposed hybrid method combines the merits of both GWO and WOA optimization algorithms. The proposed optimization algorithm is examined with 23 benchmark objective functions and is evaluated with CNN classifier. Among 23 objective functions, the function f2 is declared to be the best performing objective function. From the experimental results, it is evident that the proposed HGWWOA with the objective function f_2 and CNN classifier performs better when compared against the state of art optimization algorithms by giving an accuracy of 98.2%. Using the proposed HGWWOA, the better accuracy with optimal feature set is achieved for classification of genome sequences. The proposed optimization algorithms can be applied to solve any optimization problems irrespective of the research area.

References

[1] Afshar M. and Usefi H., "High-Dimensional Feature Selection for Genomic Datasets," *Knowledge-Based Systems*, vol. 206, pp. 106370, 2020.

[2] Aghdam M., Tanha J., Naghsh-Nilchi A., and Basiri M., "Combination of Ant Colony Optimization and Bayesian Classification for Feature Selection in a Bioinformatics Dataset," *Journal of Computer Science and Systems Biology*, vol. 2, no. 3, pp. 186-199, 2009.

[3] Ahuja J. and Ratnoo S., "Feature Selection

Using Multi-Objective Genetic Algorith M: a Hybrid Approach," *INFOCOMP Journal of Computer Science*, vol. 14, no. 1, pp. 26-37, 2015.

[4] Al-Janabi M. and Ismail M., "Improved Intrusion Detection Algorithm Based on TLBO and GA Algorithms" *The International Arab Journal of Information Technology*, vol. 18, no. 2, pp. 170-179, 2021.

[5] Chatzilygeroudis K., Vrahatis A., Tasoulis S., and Vrahatis, M., "Feature Selection in Single-Cell RNA-seq Data via a Genetic Algorithm," *in Proceedings of the International Conference on Learning and Intelligent Optimization*, Athens, pp. 66-79, 2021.

[6] Chuzhanova N., Jones A., and Margetts S., "Feature Selection for Genetic Sequence Classification," *Bioinformatics (Oxford, England*), vol. 14, no. 2, pp. 139-143, 1998.

[7] Elsadekl S., Makhlouf M., El-Sayed B., and Mohamed H.., "Hybrid Feature Selection using Swarm and Genetic Optimization for DNA Copy Number Variation," *International Journal of Engineering Research and Technology*, Vol. 12, no. 7, pp. 1110-1116, 2019.

[8] Faris H., Mirjalili S., Aljarah I., Mafarja M., and Heidari A., *Nature-Inspired Optimizers*, Springer, 2020.

[9] Garcia-Díaz P., Sánchez-Berriel I., Martínez-Rojas J., and Diez-Pascual A., "Unsupervised Feature Selection Algorithm for Multiclass Cancer Classification of Gene Expression RNA-Seq Data," *Genomics*, vol. 112, no. 2, pp. 1916-1925, 2020.

[10] https://www.cdc.gov/coronavirus/types.html, Last Visited, 2021.

[11] https://www.ncbi.nlm.nih.gov/, Last Visited, 2020.

[12] Krithiga R. and Ilavarasan E., "A Novel Hybrid Algorithm to Classify Spam Profiles in Twitter," *Webology*, vol. 17, no. 1, pp. 260-279, 2020.

[13] Leclercq M., Vittrant B., Martin-Magniette M., Scott Boyer M., Perin O., Bergeron A., Fradet Y., and Droit A., "Large-Scale Automatic Feature Selection for Biomarker Discovery in High-Dimensional Omics Data," *Frontiers in Genetics*, vol. 10, no. 452, 2019.

[14] Leung M., Delong A., Alipanahi B., and Frey B., "Machine Learning in Genomic Medicine: a Review of Computational Problems and Data Sets," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 176-197, 2015.

[15] Lo Bosco G. and Pinello L., "A New Feature Selection Methodology for K-Mers Representation of DNA Sequences," *in proceedings of the International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*, Naples pp. 99-108, 2014.

[16] Ma S. and Huang J., "Penalized Feature Selection and Classification in Bioinformatics," *Briefings in Bioinformatics*, vol. 9, no. 5, pp. 392-403, 2008.

[17] Mirjalili S. and Lewis A., "The Whale Optimization Algorithm," *Advances in Engineering Software*, vol. 95, pp. 51-67, 2016.

[18] Mirjalili S., Mirjalili S., and Lewis A., "Grey Wolf Optimizer," *Advances in Engineering Software*, vol. 69, pp. 46-61, 2014.

[19] Mohammed H., Umar S., and Rashid T., "A Systematic and Meta-Analysis Survey of Whale Optimization Algorithm," *Computational Intelligence and Neuroscience*, vol. 2019, 2019.

[20] Muthulakshmi M. and Murugeswari G., "A Novel Feature Extraction from Genome Sequences for Taxonomic Classification of Living Organisms," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 2, pp. 1436-1451, 2021.

[21] Nguyen N., Tran V., Ngo D., Phan D., Lumbanraja F., Faisal M., Abapihi B., Kubo M., and Satou K., "DNA Sequence Classification by Convolutional Neural Network," *Journal Biomedical Science and Engineering*, vol. 9, no. 5, pp. 280-286, 2016.

[22] Perez-Riverol Y., Kuhn M., Vizcaíno J., Hitz M., and Audain, E., "Accurate and Fast Feature Selection Workflow for High-Dimensional Omics Data," *PloS one*, vol. 12, no. 12, 2017.

[23] Qin X., Zhang S., Yin D., Chen D., and Dong X., "Two-Stage Feature Selection For Classification Of Gene Expression Data Based on An Improved Salp Swarm Algorithm," *Mathematical Biosciences and Engineering*, vol. 19, no. 12, pp. 13747-13781, 2022.

[24] Qin Y., Yalamanchili H., Qin J., Yan B., and Wang J., "The Current Status and Challenges in Computational Analysis of Genomic Big Data," *Big Data Research*, vol. 2, no.1, pp. 12-18, 2015.

[25] Singh N. and Singh S., "Hybrid Algorithm of Particle Swarm Optimization and Grey Wolf Optimizer for Improving Convergence Performance," *Journal of Applied Mathematics*, vol. 2017, 2017.

[26] Singh N., Singh S., and Houssein, E., "Hybridizing Salp Swarm Algorithm with Particle Swarm Optimization Algorithm for Recent Optimization Functions," *Evolutionary Intelligence*, pp. 1-34, 2022.

[27] Stawiski K., Kaszkowiak M., Mikulski D., Hogendorf P., Durczynski A., Strzelczyk J., Chowdhury D., and Fendler W., "Omicselector: Automatic Feature Selection and Deep Learning

Modeling for Omic Experiments," *BioRxiv*, 2022.
[28] Tadist K., Najah S., Nikolov N., Mrabti F., and Zahi A, "Feature Selection Methods and Genomic Big Data: a Systematic Review," *Journal of Big Data*, vol. 6, no. 79, pp. 1-24, 2019.
[29] Trivedi I., Jangir P., Kumar A., Jangir N., and Totlani R., "A Novel Hybrid PSO-WOA Algorithm for Global Numerical Functions Optimization," *Advances in Computer and Computational Sciences*, pp. 53-60, 2018.
[30] Wang D., Tan D., and Liu L., "Particle Swarm Optimization Algorithm: An Overview," *Soft Computing*, vol. 22, no. 2, pp. 387-408, 2018.
[31] Wang L., Wang Y., and Chang Q., "Feature Selection Methods for Big Data Bioinformatics: A Survey from the Search Perspective," *Methods*, vol. 111, pp. 21-31, 2016.

Muthulakshmi Murugaiah received her Doctorate in Computer Science and Engineering from Manonmaniam Sundaranar University, Tirunelveli in 2016. She is currently working as Assistant Professor in Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Tamil Nadu, India. Her research work focuses on Bioinformatics and Computational Biology.

Murugeswari Ganesan received her Ph.D. in Computer Science and Information Technology in the year 2017 from Manonmaniam Sundaranar University. She is currently working as Associate Professor in Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India. Her area of interest includes Digital Image Processing, Bio informatics and Computational Biology.