J. Inst. Eng. India Ser. B
https://doi.org/10.1007/s40031-022-00815-6

ORIGINAL CONTRIBUTION

Identification of Biclusters in Huntington's Disease Dataset Using a New Variant of Grey Wolf Optimizer

Joy Adhikary[1] · Sriyankar Acharyya[1]

Received: 25 January 2022 / Accepted: 23 September 2022

Abstract Biclustering is a useful technique to identify subgroups of genes that have same type of expression characteristics with respect to some conditions in microarray gene expression data. This is a complex problem where meta-heuristic algorithms are more suitable to explore the large datasets for finding biclusters of optimal quality. In this paper, there is an attempt for the first time to choose biclusters with respect to shifting and scaling behaviors using Huntington's disease database applying Grey Wolf Optimizer (GWO) along with its proposed modified version namely, Enhanced Search Grey Wolf Optimizer (ES-GWO). ES-GWO incorporates strategies that make the search process more balanced with respect to exploration and exploitation compared to the state-of-the-art techniques (GWO, RM-GWO). The efficacy of ES-GWO is validated on several benchmark instances and compared with the existing meta-heuristic techniques (PSO, HS, Firefly, ABC and DE) based on convergence quality. Finally, from 100 biclusters produced by ES-GWO top 5 were separated. 7 genes common in those 5 biclusters have proved to be biologically significant.

Keywords Gene expression data · Meta-heuristics · Grey wolf optimization · Biclustering

Introduction

Microarray method is widely used to study the expression levels of several genes in organs or cells. It gives large-throughput expression matrices that can be used to compare the expression levels of genes in different clinical conditions [1, 2]. These high dimensional matrices contain gene expression levels that can reflect gene activities related to different physiological status [3]. Microarray datasets corresponding to different diseases are considered and analyzed to find some special subsets of genes having similarity in their expressional behavior. The dataset used here is taken from Huntington's disease which is a rare type genetic disorder. To develop this disease, a person needs only one copy of an inactive gene. Apart from the genes in the sex chromosomes, a person receives two copies of the entire gene set-one copy from each parent [4]. A parent with a genetic defect can transmit a copy with defect or a healthy copy. Therefore, each child in the family is 50% likely to have a genetic predisposition. It causes further deterioration of nerve cells and acts on different parts of the brain, consequently affecting the movement, behavior, and perception. It becomes difficult to walk, to think, to swallow and to speak. Eventually, that person will need full-time care. Signs and symptoms appear in people in their 30–50 s [4–6].

Biclustering [2] is a kind of data mining approach that identifies subgroups of genes in the high-throughput expression matrices. These subgroups of genes are identical in expression patterns under some selected conditions. A subgroup of genes represents some cellular processes that are only active under some subset of conditions [7]. The biclustering problem can be solved using Swarm Intelligence (SI) which is one of the most popular branches of meta-heuristic algorithms. The method considered here is Grey Wolf Optimizer (GWO) which belongs to the group of SI-based algorithms. GWO is based on social hierarchy mechanism of wolves and their hunting strategies.

This research has proposed a new variant of GWO, namely, Enhanced Search Grey Wolf Optimizer (ES-GWO)

✉ Sriyankar Acharyya
srikalpa8@gmail.com

[1] Department of Computer Science and Engineering, Maulana Abul Kalam Azad University of Technology, Kolkata, India

Published online: 07 November 2022

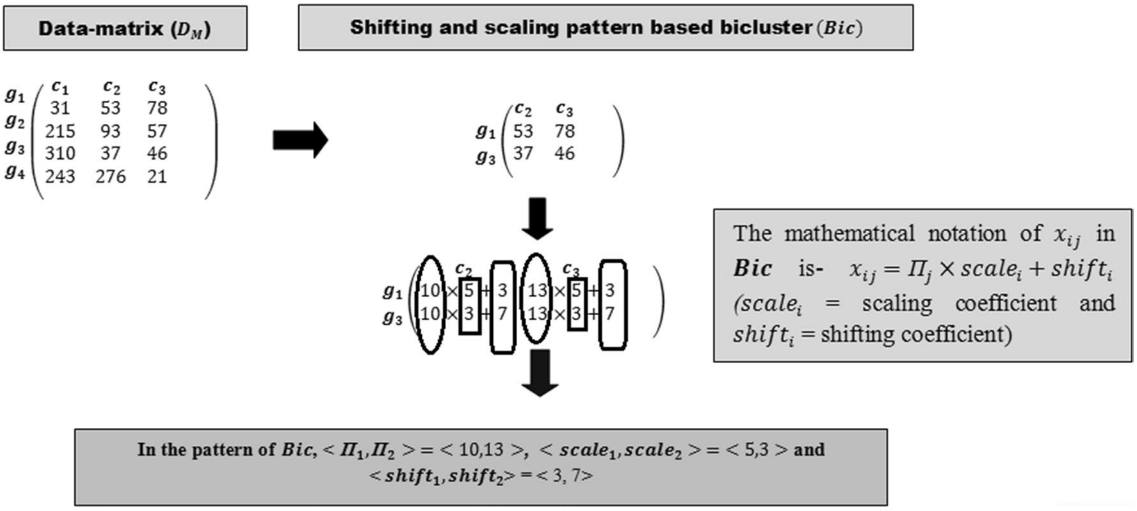
which has been implemented to identify biclusters based on shifting and scaling characteristics. The proposed variant used randomized movement that provides decent exploration in search process. This research has also incorporated an inertia weight strategy called, weight cosine control factor strategy [8]. It makes search process more balanced. The efficacy of proposed variant has tested on several unimodal and multimodal benchmark functions. Statistical test has performed to validate the result of ES-GWO [9]. The results yielded by the proposed variant have also been validated with the existing methods like Particle Swarm Optimization (PSO) [10], Harmony Search (HS) [11], Firefly [12], Artificial Bee Colony (ABC) [13] and Differential Evolution (DE) [14]. This research attempts for the first time to identify shifting and scaling behavior-based biclusters from Huntington's disease datasets. Results of ES-GWO on real life dataset have been validated with that of the state-of-the-art of GWO (GWO, RM-GWO). The efficacy of ES-GWO has been observed on the both data dataset. In each dataset, ES-GWO successfully finds biological relevant biclusters. Finally, from 100 optimal biclusters identified by ES-GWO top 5 were separated on the basis of cost function. 7 genes common in those top 5 biclusters have been selected and validated (by *p*-value) to be biologically significant.

Related Works

Cheng and Church (CC) [15] introduced the concept of bicluster, involving some selected genes and some selected conditions having a good similarity measure. The concept bicluster overcomes several problems associated with traditional clustering methods. Researchers used Mean Squared Residue (MSR) [15] to measure the coherence present in genes and conditions belonging to a bicluster [15]. The strength of Mean Squared Residue (MSR) is only capturing the constant and shifting biclusters [16]. It is not able to detect scaling biclusters. In the investigations [16], researchers used a new cost function called Scaling Mean Squared Residue (SMSR) to detect the scaling patterns effectively. Huang et al. [7] introduced an algorithm named Condition-Based Evolutionary Biclustering (CBEB). It is based on Evolutionary Algorithms (EA). It is used to detect biclusters by the parallelizing search strategy. This work incorporates MSR metric with predefined threshold to obtained better results [7]. Thangavel et al. [17] proposed a hybrid algorithm called, Hybrid PSO-SA-BIC that combines binary PSO and Simulated Annealing together. PSO-SA-BIC identified highly correlated biclusters having larger volume [17]. In the literature [2], researchers proposed an algorithm, named Evolutionary Biclustering based on Expression Patterns (Evo-Bexpa). Researchers used a cost function that measure quality, volume, overlapping amount and gene variance of biclusters based on shifting and scaling pattern. Adhikary et al. [18] reported Random Move Grey Wolf Optimizer (RM-GWO) to find biclusters on Parkinson's disease dataset. This research has considered RM-GWO [18] as state-of-the art algorithm to validate the efficacy of the proposed algorithm (ES-GWO).

There are some recent trends in feature extraction which may be relevant in this context. In the researches [19], Fonseca et al proposed a system that incorporated two different approaches (feature engineering approach and learning representations from data). Feature engineering approach uses Gradient Boosting Machine (GBM) and learning representation CNN learns features from the audio signal. Das et al. [20] reported a deep neural network model for sentiment evaluation on live-streamed tweets on Coronavirus. In the literature, [21] machine learning methods have been introduced on a novel iris recognition system. The main objective of this model [21] is to achieve virtually perfect classification accuracy. A fuzzy neural network-based model [22] has been applied on human face recognition system where the images are considered in spatial and frequency domain to identify different orientations and scales [22]. Jana et al proposed a technique of automatic epileptic seizure prediction by extracting features from EEG signal [23] using Dense Convolutional Network.

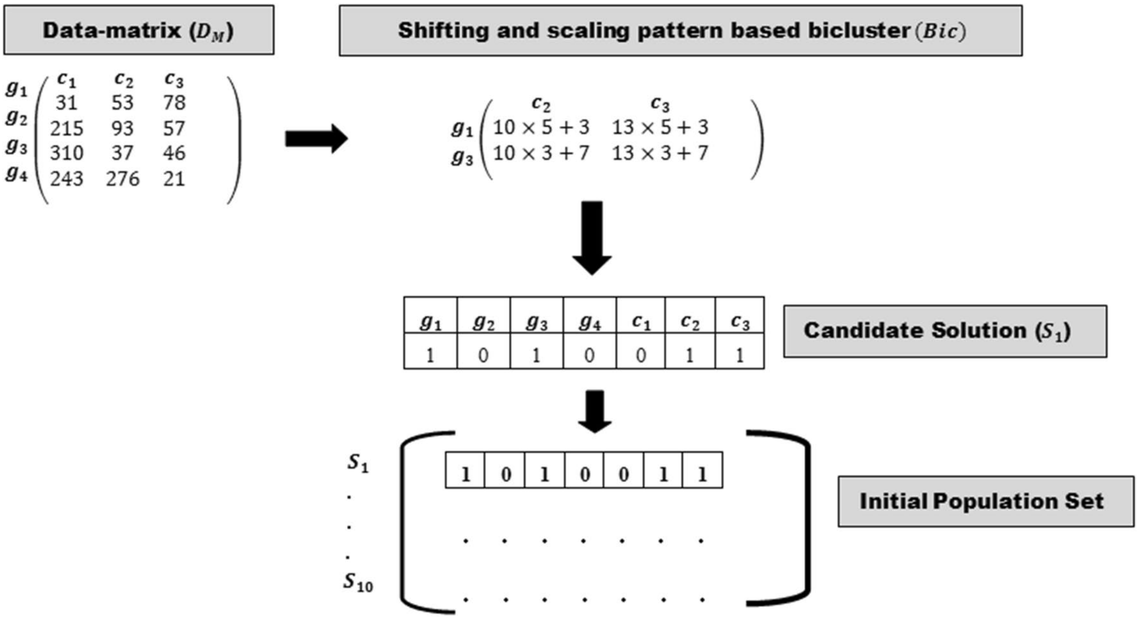
Feature extraction techniques are widely used in various research fields. In [24] researchers used a deep neural model to recognize the facial expression. In ECG signal analysis, feature extraction methods [25–27] have been widely used. Gupta et al. [28] used Fractional Fourier Transform and Independent Principal Component Analysis (IPCA) to extract features from noisy ECG signals. In the literature [27, 29], analysis of ECG signal is done to detect heart disease in proper time depending on Fractional Wavelet Transform (FrWT) and Principal Component Analysis (PCA). The advancement of these techniques helps to identify minor changes in ECG signal that plays a crucial role in health care system [30–34]. Feature extraction techniques are also used in monitoring of blood pressure (BP) [30] and detecting R-peak in electrocardiogram signal [26, 35–38]. Gupta et al. [31] used popular nonlinear technique, namely, chaos theory for R-peak detection on noisy ECG signal. In [36], researchers have used Probabilistic Principal Component Analysis (PPCA) to detect R-peaks for diagnosing heart abnormalities.



The mathematical notation of x_{ij} in Bic is- $x_{ij} = \Pi_j \times scale_i + shift_i$ ($scale_i$ = scaling coefficient and $shift_i$ = shifting coefficient)

In the pattern of Bic, $<\Pi_1,\Pi_2> = <10{,}13>$, $<scale_1,scale_2> = <5{,}3>$ and $<shift_1,shift_2> = <3, 7>$

Fig.1 Mathematical notation of bicluster (***Bic***)


Data-matrix (D_M)
Shifting and scaling pattern based bicluster (Bic)
g_1 g_2 g_3 g_4
c_1 c_2 c_3
31 53 78
215 93 57
310 37 46
243 276 21
c_2 c_3
g_1 $10 \times 5 + 3$ $13 \times 5 + 3$
g_3 $10 \times 3 + 7$ $13 \times 3 + 7$
Candidate Solution (S_1)
S_1
S_{10}
Initial Population Set

g_1	g_2	g_3	g_4	c_1	c_2	c_3
1	0	1	0	0	1	1

1	0	1	0	0	1	1

Fig.2 An example deriving bicluster (***Bic***) from Data-matrix (D_{M})

Biclustering

Biclusters

A bicluster [1, 2, 39] consists of a collection of genes where the expression patterns look similar with respect to a set of experimental conditions. Let ***Bic*** be a bicluster made of a collection of genes (g) and conditions (c). Each entry in ***Bic*** is represented by x_{ij} $\left(x_{ij} \in \boldsymbol{Bic}\right)$, $i \in g$ and $j \in c$. This research work only identified bicluster based on shifting and scaling pattern [2, 39]. Suppose $\boldsymbol{D_M}$ is a 4×3 matrix, where rows (genes) are $\langle g_1, g_2, g_3, g_4 \rangle$ and columns (conditions) $\langle c_1 c_2, c_3 \rangle$. From $\boldsymbol{D_M}$**,** a bicluster based on shifting and scaling pattern (***Bic***) is chosen, where genes are denoted by $\langle g_1, g_3 \rangle$ and the conditions are denoted by $\langle c_2, c_3 \rangle$ (Fig. 1).

Candidate Solution Generation

In this research, every sub-matrix (bicluster) is considered as a candidate or member solution. The length of a member solution is determined by the addition of the count of rows and columns of the Data-matrix (D_M). The complete candidate solution generation process is depicted in Fig. 2. Each candidate solution is a combination of bits 1 and 0. 1 and 0 bits denote the situation whether the respective gene or condition is taken into account or not.

Cost Function

In this experiment researchers has used a cost function (Cost(***Bic***)) to calculate the cost of each bicluster. Cost(***Bic***) is calculated by the components: transposed virtual error, volume, overlapping amount and gene variance [2] of each bicluster. The expression of cost function is shown by equation (1). W_{v}, W_{ov} and W_{var} are three most important parameters used in cost function (Eq. 1). W_{v} is used to control the volume of a bicluster. W_{ov} and W_{var} are used to control the overlapping amount and gene variance of a bicluster [2]. The tuning of parameter values is done in sect. 5.1. The standard of bicluster based on shifting and scaling pattern [1, 2] is determined by the minimization of this cost function.

$$\mathrm{Cost}(\boldsymbol{Bic}) = \frac{VE^t(\boldsymbol{Bic})}{VE^t\left(D_M\right)} + W_{\mathrm{v}} \times \mathrm{Volume}(\boldsymbol{Bic}) + W_{\mathrm{ov}} \times \mathrm{Overlap}(\boldsymbol{Bic}) + W_{\mathrm{var}} \times \frac{1}{1+\mathrm{Var}(\boldsymbol{Bic})} \tag{1}$$

Transposed Virtual Error (VE^t) is the most important component of the cost function [1, 2]. VE^t (Eq. 5) measures the quality of each bicluster.

$$\rho_i = \frac{1}{|J|}\sum\nolimits_{j=1}^{|J|} x_{ij} \tag{2}$$

The virtual condition (ρ_i) of i-th gene is defined in Eq. 2, where, x_{ij} is the value of gene expression and $|J|$ is the total count of conditions.

$$\widehat{\rho_i} = \frac{\rho_i - \mu_p}{\sigma_p} \tag{3}$$

In the above equation (3), $\widehat{\rho_i}$ denotes the standardized value of virtual condition, μ_p denotes the mean and σ_p denotes the standard deviation.

$$\widehat{x_{ij}} = \frac{x_{ij} - \mu_{c_j}}{\sigma_{c_j}} \tag{4}$$

Equation (4) shows $\widehat{x_{ij}}$ is standardized bicluster, σ_{c_j} and μ_{c_j} denote the standard deviation and average of expression measures respectively for condition j.

$$\mathrm{VE}^t(\boldsymbol{Bic}) = \frac{1}{|I|\cdot|J|}\sum_{i=1}^{I}\sum_{j=1}^{J}\mathrm{abs}\left(\widehat{x_{ij}} - \widehat{\rho_i}\right) \tag{5}$$

Table 1 Specification of Huntington's disease datasets for experiments

Sl. no.	Dataset accession no.	Collected from	Total number of genes	Total number of samples
1	GSE3790 [4]	GEO, NCBI	22283	201
2	GSE26927 [41]	GEO, NCBI	20589	118

Bicluster volume (Volume(Bic)) is determined from Eq. 6. The smaller the volume, the larger will be the chance of getting a perfect bicluster pattern [1, 2].

$$\mathrm{Volume}(\boldsymbol{Bic}) = \left(\frac{-\ln(|I|)}{\ln(|I|)+w_g}\right) + \left(\frac{-\ln(|J|)}{\ln(|J|)+w_c}\right) \tag{6}$$

In Equation (6), $|I|$ and $|J|$ refer to the count of genes and conditions. w_g and w_c are used to control the volume of a bicluster [2]. The values of w_g and w_c are tuned in sect. 5.1.

The overlapping term (Overlap(***Bic***)) is the count of occurrences of a member of a bicluster in other biclusters [2]. The larger the value of Overlap(***Bic***) the more will be the scope for getting a faulty pattern.

$$\mathrm{Overlap}(\boldsymbol{Bic}) = \frac{\sum_{i\in I, j\in J} W\left(x_{ij}\right)}{|I|\cdot|J|\cdot\left(\mathrm{num_b}-1\right)} \tag{7}$$

where, the weight matrix is denoted by W and $\mathrm{num_b}$ denotes the count of biclusters.

Gene variance (Var(***Bic***)) (Eq. 8) denotes the average of variances of expressions of genes present in a bicluster [2]. The more the gene variance the less is its significance.

$$\mathrm{Var}(\boldsymbol{Bic}) = \frac{1}{|I|\cdot|J|}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(x_{ij} - \mu_{g_i}\right)^2 \tag{8}$$

In Equation (8), μ_{g_i} is the arithmetic mean of expression values of gene i.

Materials and Methods

In this section, description of datasets and methods are defined.

Materials

In this experiment, researchers used Huntington's disease [4, 5] datasets. The specification and source of the datasets are described in Table 1. Proposed variant ES-GWO and state-of-the-art of GWO (GWO, RM-GWO) [8, 40] are applied to obtained 100 shifting and scaling pattern-based biclusters from disease datasets.

Methods

Meta-heuristic algorithms are becoming an important part in the optimization field for their characteristics like robustness, simplicity and problem independence. This experiment applied a meta-heuristic technique (GWO) and its modified version (ES-GWO) to identify optimal quality biclusters based on shifting and scaling pattern.

Grey Wolf Optimizer (GWO)

Swarm intelligence (SI)-based meta-heuristic algorithms are inspired by the collective behavior of agents (birds, animals, insects, etc.) in a community. Grey wolf optimization is SI-based algorithm which is inspired by hunting strategies and social hierarchy mechanisms of wolves. This meta-heuristic algorithm mainly focuses on social hierarchy, fixing the prey and then attacking.

Social hierarchy: In this algorithm, wolves are categorized into four groups α (top-level or leader wolf), β (second best level), δ (third level) and ω (lowest level) [8, 40]. The alpha wolf (α) takes the decision on searching and hunting a prey.

Encircling prey: Prior to the hunting phase, wolves are locating the prey. In this step, positions of the wolves are altered by the equation (9). Difference vector $\left(\overrightarrow{\mathrm{Diff}}\right)$ is measured by Equation 10. A and C are calculated by equations (11) and (12) [8, 40].

$$\vec{Y}(t+1) = \overrightarrow{Y_\mathrm{p}}(t) - A.\overrightarrow{\mathrm{Diff}} \tag{9}$$

$$\overrightarrow{\mathrm{Diff}} = \left| C \cdot \overrightarrow{Y_\mathrm{p}}(t) - \vec{Y}(t) \right| \tag{10}$$

where, t denotes the current step of iteration , $\vec{Y}_\mathrm{p}$ is the location of prey,$\vec{Y}$ is the location of a grey wolf. A and C are the coefficients.

$$A = 2 \times a \times r_1 - a \tag{11}$$

$$C = 2 \times r_2 \tag{12}$$

$$a = 2 - 2 \times (t/\max \mathrm{Iter}) \tag{13}$$

where, t and $\max \mathrm{Iter}$ are iterations no. and the maximum count of iterations.r_1 and r_2 are the random variables, r_1 ϵ [0,1] , r_2 ϵ [0,1]. a and C control the exploitative and explorative behavior of search, respectively.

Hunting: The three most efficient solutions are already stored as they are more aware of the location of the prey. In GWO, the positions of the searching wolves are modified based on the location of the α, β and δ according to equations (14), (15) and (16) [40].

$$\overrightarrow{\mathrm{Diff}_\alpha} = \left| C_1 \cdot \overrightarrow{Y_\alpha} - \vec{Y} \right|, \overrightarrow{\mathrm{Diff}_\beta} = \left| C_2 \cdot \overrightarrow{Y_\beta} - \vec{Y} \right|, \overrightarrow{D_{\mathrm{iff}\delta}} = \left| C_3 \cdot \overrightarrow{Y_\delta} - \vec{Y} \right| \tag{14}$$

$$\begin{aligned} &\overrightarrow{Y_1} = \overrightarrow{Y_\alpha} - A_1 \cdot \left(\overrightarrow{\mathrm{Diff}_\alpha}\right), \overrightarrow{Y_2} = \overrightarrow{Y_\beta} - A_2 \cdot \left(\overrightarrow{\mathrm{Diff}_\beta}\right), \\ &\overrightarrow{Y_3} = \overrightarrow{Y_\delta} - A_3 \cdot \left(\overrightarrow{\mathrm{Diff}_\delta}\right) \end{aligned} \tag{15}$$

$$\vec{Y}(t+1) = \frac{\overrightarrow{Y_1} + \overrightarrow{Y_2} + \overrightarrow{Y_3}}{3} \tag{16}$$

Enhanced Search Grey Wolf Optimizer (ES-GWO)

Premature convergence is one of the problems in meta-heuristic search. It occurs due to weak exploration in searching strategies. A way to escape from this premature convergence (local optimum) is to explore and exploit the entire search region in a balanced manner. The proposed algorithm (ES-GWO) improves explorative ability by incorporating random moves to the wolf's journey using random numbers generated by student's t distribution. The function $\mathrm{trnd}(df, 1, \mathrm{Dim})$ generates $1 \times \mathrm{Dim}$ random numbers by student's t distribution with degrees of freedom (df), Dim is the dimension [8]. The flowchart of the proposed algorithm (ES-GWO) is depicted by Figure 3.

$$\mathrm{UP}_i = Z_i(t+1) + \mathrm{trnd}(df, 1, \mathrm{Dim}) \tag{17}$$

Table 2 List of parameters and their tuned values

Values of the parameters used in cost function (Eq. 1)

Parameters	Value	Range
W_v	5.0	W_v ∈[0, 10]
W_OV	5.0	W_{OV} ∈ [0,10]
W_Var	0.1	W_{Var} ∈ [0 , 0.2]
w_g	0.25	w_g ∈ [0 , 0.5]
w_c	0.5	w_c ∈ [0 , 1]

Values of the Parameters used in GWO, RM-GWO, ES-GWO

maxIter=100, Population size=10

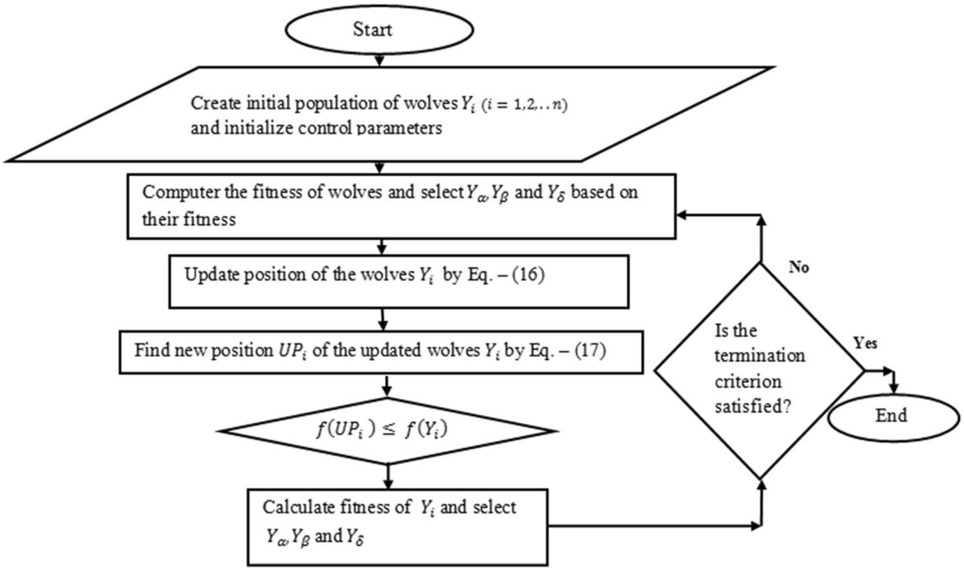
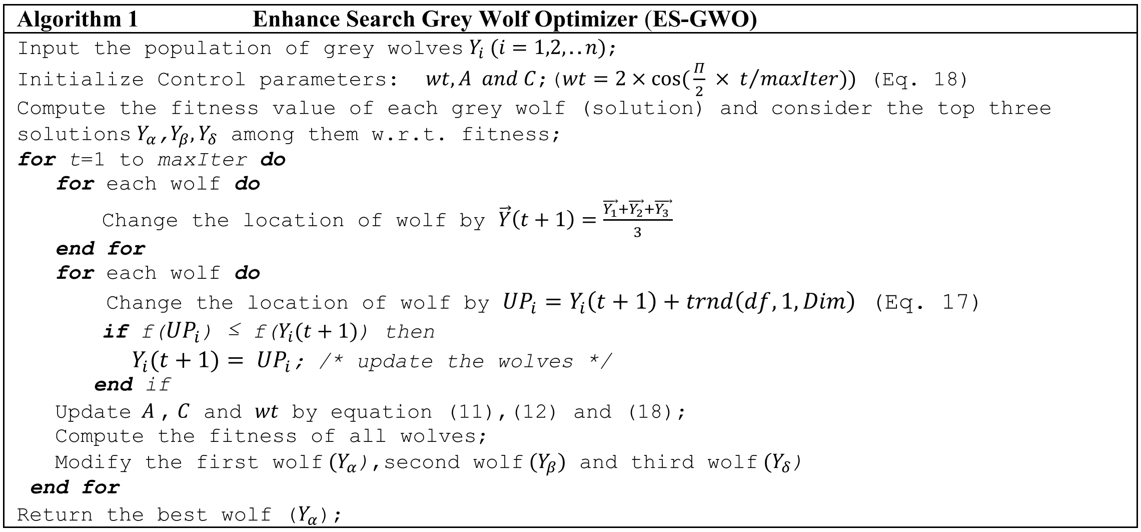

Start
Create initial population of wolves Y_i ($i = 1,2,..n$) and initialize control parameters
Computer the fitness of wolves and select Y_α, Y_β and Y_δ based on their fitness
Update position of the wolves Y_i by Eq. – (16)
Find new position UP_i of the updated wolves Y_i by Eq. – (17)
$f(UP_i) \leq f(Y_i)$
Calculate fitness of Y_i and select Y_α, Y_β and Y_δ
Is the termination criterion satisfied?
No
Yes
End

Fig. 3 Flowchart of Enhanced Search Grey Wolf Optimizer (ES-GWO)

In equation (17), UP_i is the modified solution and if the cost of UP_i is less than that of $Y_i(t+1)$, then $Y_i(t+1) = \mathrm{UP}_i$, otherwise old $Y_i(t+1)$ will be retained.

To make the search process more balanced, a weight cosine control factor strategy (Eq. 18) [8] is used in the proposed variant. In the original GWO [8] and RM-GWO [18], a was employed to control the degree of exploitation of the search. The value of a was linearly decreased, but here the proposed variant uses the weight cosine control factor strategy (wt) instead of linearly decreasing a. It helps the search process to avoid local optima in search space.

Algorithm 1 Enhanced Search Grey Wolf Optimizer



Algorithm 1 **Enhance Search Grey Wolf Optimizer (ES-GWO)**

```
Input the population of grey wolves $Y_i$ ($i = 1,2,..n$);
Initialize Control parameters:  $wt, A$ and $C$; ($wt = 2 \times \cos(\frac{\pi}{2} \times t/maxIter)$) (Eq. 18)
Compute the fitness value of each grey wolf (solution) and consider the top three
solutions $Y_\alpha, Y_\beta, Y_\delta$ among them w.r.t. fitness;
for t=1 to maxIter do
   for each wolf do
      Change the location of wolf by $\vec{Y}(t+1) = \frac{\vec{Y_1}+\vec{Y_2}+\vec{Y_3}}{3}$
   end for
   for each wolf do
      Change the location of wolf by $UP_i = Y_i(t+1) + trnd(df, 1, Dim)$ (Eq. 17)
      if $f(UP_i) \leq f(Y_i(t+1))$ then
         $Y_i(t+1) = UP_i$; /* update the wolves */
      end if
   Update $A$, $C$ and $wt$ by equation (11),(12) and (18);
   Compute the fitness of all wolves;
   Modify the first wolf ($Y_\alpha$),second wolf ($Y_\beta$) and third wolf ($Y_\delta$)
 end for
Return the best wolf ($Y_\alpha$);
```

Table 3 Result obtained by GWO and ES-GWO on unimodal functions (dimension 5)

Functions		Mean	Std dev	Median	Min	Max
F1	GWO	1.476E-11	2.91E-11	3.076E-12	1.83E-14	1.870E-10
	ES-GWO	**2.48E-14**	**7.25E-14**	**9.07E-16**	**2.19E-18`**	**4.56E-13**
F2	GWO	3.72E-04	0.0023	3.25E-06	2.846E-09	0.0164
	ES-GWO	**2.99E-07**	**8.74E-07**	**8.05E-09**	**8.42E-12**	**5.74E-06**
F3	GWO	**0.5990**	0.1864	0.6668	**5.72E-04**	0.6684
	ES-GWO	0.6493	**0.0577**	**0.6667**	0.3292	**0.6672**
F4	GWO	21.22	18.714	21.6472	0.1420	67.17
	ES-GWO	**2.2419**	**2.664**	**1.3880**	**0.1025**	**13.83**

Best results (minimum cost) are bolded

Table 4 Result obtained by GWO and ES-GWO on unimodal functions (dimension 10)

Functions		Mean	Std dev	Median	Min	Max
F1	GWO [18]	5.30E-07	7.27E-07	2.23E-07	2.35E-09	3.86E-06
	ES-GWO	**7.16E-09**	**1.15E-08**	**1.83E-09**	**4.83E-11**	**6.30E-08**
F2	GWO [18]	**2.795**	5.692	**0.440**	**0.0089**	25.38
	ES-GWO	4.686	**4.318**	2.914	0.3579	**17.0045**
F3	GWO [18]	0.7104	0.1110	0.6696	0.6671	1.030
	ES-GWO	**0.6686**	**0.0013**	**0.6683**	**0.6668**	**0.6727**
F4	GWO [18]	961.02	1.41E+03	587.68	**2.923**	7.68E+07
	ES-GWO	**29.845**	**37.495**	**19.662**	4.8749	**190.155**

Best results (minimum cost) are bolded

$$\mathrm{wt} = 2 \times \cos\left(\frac{\pi}{2} \times t/\max \mathrm{Iter}\right) \tag{18}$$

Results and Discussion

Experiments were built into the machine with Pentium Dual-Core CPU, 4 GB memory, Microsoft Windows 7 environment and MATLAB R2012b platform. Values of parameters are mentioned. To analyze the performances of algorithms on several benchmark functions. Convergence analysis of GWO and ES-GWO represent the convergence graphs compared with other meta-heuristic techniques. This gives the statistical description and shows the results of complexity of algorithms based on CEC 14. The results on real-life datasets included and has listed some genes identified in obtained biclusters based on biological significance.

Parameters Tuning

This section represents the used parameters. The tuned values of parameters are presented in Table 2.

Benchmark Instances: Results

This section measures the efficiency of GWO [40] and its proposed variant (ES-GWO) in 8 benchmark markings on different scales (sizes 5 and 10). Functions F1 to F4 are unimodal (Sphere (F1), Zakharov (F2), Dixon-price (F3) and

Table 5 Result obtained by GWO and ES-GWO on multimodal functions (dimension 5)

Functions		Mean	Std dev	Median	Min	Max
F5	GWO	**1.343**	**0.9895**	1.1982	4.84E-08	**3.699**
	ES-GWO	1.972	2.1427	**1.4087**	**2.70E-12**	12.109
F6	GWO	0.0931	0.0656	0.0906	1.529E-05	0.2709
	ES-GWO	**0.0467**	**0.0470**	**0.0414**	**6.74E-06**	**0.161**
F7	GWO	5.42E-05	1.13E-04	3.25E-05	1.01E-06	7.94E-04
	ES-GWO	**1.06E-06**	**1.88E-06**	**5.07E-07**	**4.96E-08**	**1.29E-05**
F8	GWO	**9.315**	**4.681**	**8.571**	**3.0670**	**24.229**
	ES-GWO	30.271	11.99	28.75	9.826	61.259

Best results (minimum cost) are bolded

Table 6 Result obtained by GWO and ES-GWO on multimodal functions (dimension 10)

		Mean	Std dev	Median	Min	Max
F5	GWO [18]	**4.929**	**2.397**	**4.418**	1.208	**12.76**
	ES-GWO	9.631	6.952	7.6321	**0.0015**	32.80
F6	GWO [18]	0.4096	0.2194	**0.364**	**0.0805**	1.0286
	ES-GWO	**0.3466**	**0.0869**	0.3635	0.1736	**0.5453**
F7	GWO [18]	0.0056	0.0062	0.0041	0.0011	0.0348
	ES-GWO	**4.26E-04**	**3.65E-04**	**3.04E-04**	**3.60E-05**	**0.0018**
F8	GWO [18]	37.673	11.72	37.212	16.673	62.294
	ES-GWO	**27.297**	**10.665**	**24.441**	**11.560**	**50.635**

Best results (minimum cost) are bolded

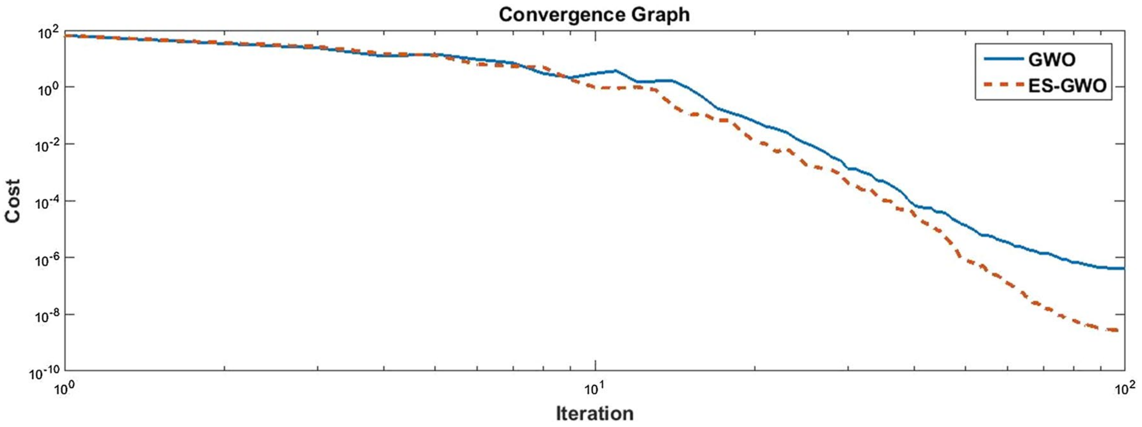


Fig.4 Convergence plot for sphere (F1) function

Shifted Sphere (F4)) and some from F5 to F8 is multimodal (Rastrigin (F5), Levy (F6), Ackly (F7) and Shifted Rastrigin's (F8)) [9]. For each benchmark activity, 50 independent runs were taken.

Table 3 shows the results of unimodal benchmark functions (F1 to F4) for dimension = 5. Unimodal function measures the strength of exploitation of an algorithm. The proposed variant ES-GWO has outperformed GWO in function F1, F2 and F4. In case of function F3, it has performed well in respect of most of the metrics. Results on unimodal benchmark functions with dimension = 10 has been shown in Table 4. In this observation, ES-GWO has outperformed

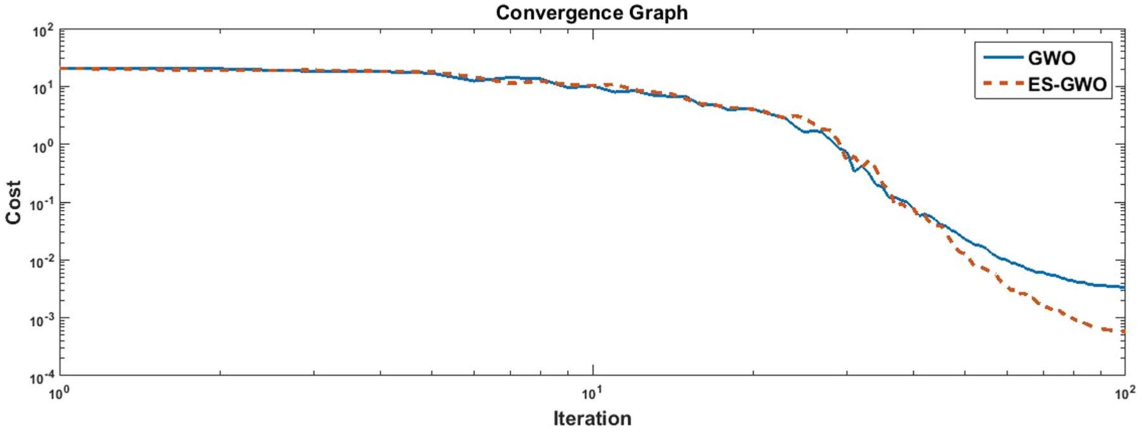


Fig.5 Convergence plot for akly (F7) function

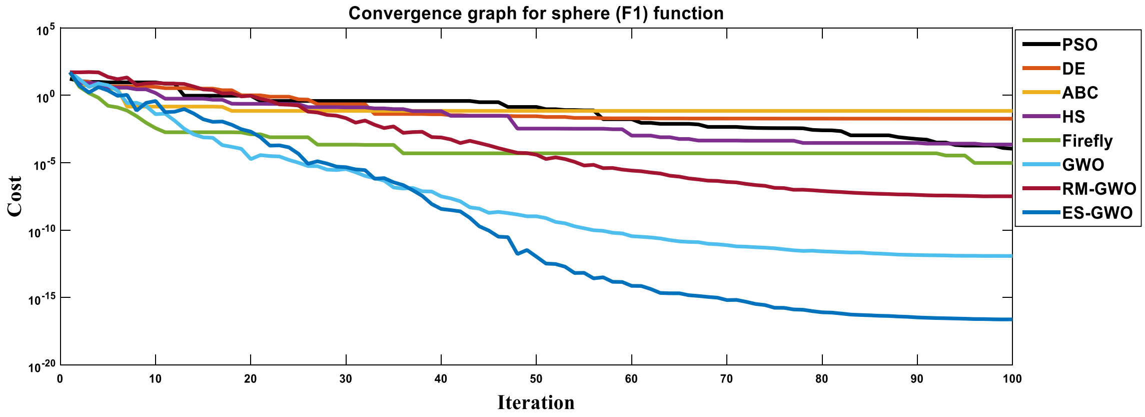


Fig.6 Convergence plot for sphere (F1)

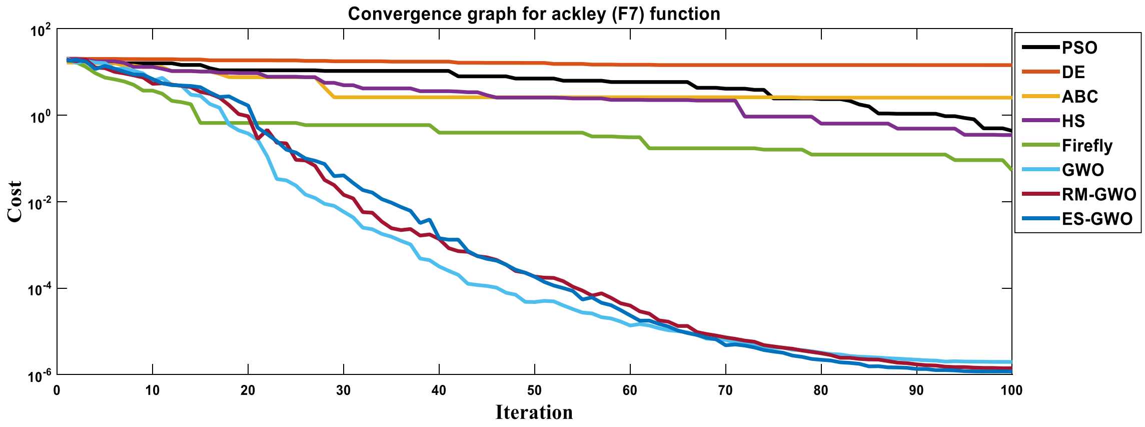


Fig.7 Convergence plot for ackly (F7)

GWO on functions F1, F3 and F4. Overall performance of ES-GWO is well in case of unimodal functions.

Tables 5 and 6 show the results of multimodal benchmark functions (F5–F8) on dimension 5 and 10, respectively. Multimodal function evaluates the strength of exploration of an algorithm. It also measures the ability of avoiding local optima. ES-GWO outperformed in function F6 and F7 (Table 5). Table 6 also shows the superiority of ES-GWO. Here, proposed variant outperformed on functions F7 and F8. In case of function F6, ES-GWO performed better in most of the metrices. The improvement in overall performance of ES-GWO justifies enhancement in exploration capability to avoid local optima.

Convergence Analysis

The convergence nature of GWO and ES-GWO has been studied here on benchmark functions F7 and F1. In Figures 4 and 5, *X*-axis presents the iterations and *Y*-axis presents the corresponding cost of solutions. The convergence of the proposed variant ES-GWO is clearly shown in Figures 4 and 5, respectively. In both cases, the convergence accuracy of the proposed variant is better than GWO. In both cases, ES-GWO shows strong exploration ability that helps to explore search space more effectively. Compared to the other algorithm, it reaches fast to the effective part of the search space.

Comparison of results with other meta-heuristics

In this section the performance of the proposed GWO variant is compared with other meta-heuristics. The convergence rate characteristics of an algorithm determine the optimality of performance. Convergence curve of the proposed variant (ES-GWO) can directly be compared with that of the other algorithms (PSO [10], HS [11], Firefly [12], ABC [13] and DE [14]) and state-of-the-art algorithms (GWO, RM-GWO). The comparative analysis has been represented through Figures 6–7. ES-GWO has shown better efficacy in each case. In Figures 6 and 7,

Table 7 Mann-Whitney U-test results between GWO and ES-GWO

Function	F1	F2	F3	F4	F5	F6	F7	F8
p-value	5.58E-09	7.92E-05	1.34E-05	3.00E-06	0.1338	0.3060	3.92E-06	3.42E-13

Table 8 Test program given by CEC 2014

Test program [7]

```
for i=1 to 1000000
  y=0.5+(double)i
  y=y+y; y-y/2;y=y×y;
  y=sqrt(y); y=log(y);
  y=exp(y);y=y/(y+2);
end for
```

GWO and RM-GWO have become competitive with ES-GWO. Most of the algorithms are supposed to be stuck to local optima.

Mann-Whitney U-test to analyze the performance of ES-GWO

This section performed statistical test on 8 benchmark functions. This experiment used Mann-Whitney U-test statistical test [9, 39]. This is a nonparametric test to check the equality of medians of populations taken from two independent samples. It is used to analyze the performance of GWO and its proposed variant ES-GWO. The results are tested with respect to *p*-value. In case of function F1, *p*-value is 5.58E-09. It denotes that the null hypothesis of equal medians is rejected at the level of 5% significance. Results are described in Table 7.

Complexity Analysis

This section measured the complexity of the algorithm based on the CEC 2014 benchmark suits [8]. For the complexity analysis (in secs), the parameters T_0, T_1 and T_2 were used in the same way as defined in CEC 14. T_0 is the computation time provided by the specified test program (Table 8) given in CEC 2014 [8]. T_1 is the time required to directly calculate 2×10^5 evaluation of the F18 function without using an algorithm. T_2 is the computation time needed by the algorithm to compute 2×10^5 evaluations of the F18 function. The complexity of the algorithm (Table 9) is estimated by $(T_2 - T_1)/T_0$ [8].

Table 9 Algorithm complexities by GWO, ES-GWO

Algorithm		Dimension=10
T_0	–	0.19
T_1	–	0.85
GWO	$(T_2-T_1)/T_0$	8.26
ES-GWO	$(T_2-T_1)/T_0$	8.19

Table 10 Experimental results on Huntington's disease dataset (best result bolded)

Dataset	Algorithm	Mean of Cost(*Bic*)
GSE3790 [4]	GWO	5.523E+03
	RM-GWO	5.028E+03
	ES-GWO	**2.765E+03**
GSE26927 [41]	GWO	5.561E+03
	RM-GWO	4.981E+03
	ES-GWO	**2.876E+03**

Best results (minimum cost) are bolded

Huntington's Disease Dataset

The cost of each bicluster is measured by cost function. In this experiment, GWO, RM-GWO and proposed variant ES-GWO applied on Huntington's disease dataset. Details of Huntington's disease dataset (Table 1) is defined. In Table 10, Mean of Cost(*Bic*) is defined the mean of cost of 100 biclusters. Investigations on the Huntington's disease dataset indicate that the proposed ES-GWO surpasses state-of-the art algorithms in terms of performance. ES-GWO successfully finds least cost biclusters from each dataset. It clearly shows the strong efficacy of ES-GWO. The minimum cost is given (bolded) in Table 10. In each dataset, RM-GWO has performed well, but ES-GWO has done exceptionally well because of its judicious mixture of exploitation and exploration.

Biological Significance of the Biclusters

This section introduces a detailed study on biclusters obtained from which some significant results have been extracted here from biological point of view. ES-GWO has identified 100 biclusters from the dataset GSE26927 [41]. Out of them top 5 biclusters have been selected considering the cost value (least cost) and from those biclusters (top 5), 10 common genes have been identified (Table 11). The dataset GSE26927 [41] has 118 samples (normal and disease) and among all the samples the *p*-values of those 10 common genes are obtained. It is observed (Table 11) that the *p*-values of 7 genes are less than 0.05 which proves the

Table 11 Biological relevance of top 5 biclusters of GSE26927 dataset

Analysis on top 5 (based on cost) bicluster GSE26927 [41]		
ID of obtained common genes among top 5 biclusters	Gene symbol of obtained common genes	Significant genes (based on p value < 0.05) from obtained common genes
ILMN_10272	USH2A	USH2A
ILMN_12129	PYHIN1	–
ILMN_12579	GTF3A	GTF3A
ILMN_13146	DNPEP	DNPEP
ILMN_21359	RPP21	RPP21
ILMN_22298	KCNE3	–
ILMN_25568	GKAP1	GKAP1
ILMN_29431	LRAT	LRAT
ILMN_3520	PLA2G2D	–
ILMN_8126	SYK	SYK

biological significance of those genes [41] in the context of Huntington's disease.

Conclusion and Future Work

In gene expression analysis, biclustering algorithms, in contrast to clustering algorithms, identify groups of genes that exhibit similar patterns of activity under specific subset of the experimental conditions. Identification of biclusters is always important in this analysis, as they consist of a small group of genes involved in cellular processes. Interesting cellular processes are active only under a subset of conditions. In medicine, it can be useful to prioritize precision medicine based on changes in cell function. Biclusters are considered as important biological features that explain certain cellular functions leading to disease. Apart from that, biclustering analysis is of great help to researchers as it saves a lot of time in analyzing huge microarray data sets. This research analyzes the dataset of Huntington's disease by capturing different subsets of genes that show similarity with respect to a subset of conditions. The proposed variant ES-GWO yields better results compared to other variants on different benchmark functions involving different scales. The performance of the proposed variant is also statistically verified with respect to other meta-heuristic algorithms. To identify the optimal quality biclusters with respect to shifting and scaling pattern, ES-GWO is efficiently applied on two Huntington's dataset and the results show the strong efficacy of method. Moreover, ES-GWO has produced better results compared to other variants in identifying optimal biclusters containing biologically significant genes.

Funding The authors declare that no funding information is available for this research.

Declarations

Conflict of interests The authors declare that they have no competing financial interests or personal relationships that could affect the work presented in this article.

References

1. B. Pontes, R. Giráldez, J.S. Aguilar-Ruiz, Biclustering on expression data: a review. J. Biomed. Inf. **57**, 163–180 (2015)
2. B. Pontes, R. Giráldez, J.S. Aguilar-Ruiz, Configurable pattern-based evolutionary biclustering of gene expression data. Algorithms Mol. Biol **8**(1), 1–22 (2013)
3. S. Liu, C. Xu, Y. Zhang, J. Liu, B. Yu, X. Liu, M. Dehmer, Feature selection of gene expression data for cancer classification using double RBF-kernels. BMC Bioinf. **19**(1), 1–14 (2018)
4. A. Hodges, D.S. Andrew, K.A. Aaron, A. Kuhn, T. Sengstag, G. Hughes, L.A. Elliston, C. Hartog, D.R. Goldstein, D. Thu, Z.R. Hollingsworth, Regional and cellular gene expression changes in human Huntington's disease brain. Human Mol. Genet. **15**(6), 965–977 (2006)
5. A. Neueder, G.P. Bates, A common gene expression signature in Huntington's disease patient brain regions. BMC Med. Genom. **7**(1), 1–23 (2014)
6. T. Gasser, Genetics of Huntington's disease. Curr. Opin. Neurol. **18**(4), 363–369 (2005)
7. Q. Huang, D. Tao, X. Li, A. Liew, Parallelized evolutionary learning for detection of biclusters in gene expression data. IEEE/ACM Trans. Comput. Biol. Bioinf. **9**(2), 560–570 (2011)
8. J. Adhikary, S. Acharyya, Randomized balanced grey wolf optimizer (RBGWO) for solving real life optimization problems. Appl. Soft Comput. **117**, 108429 (2022)
9. A.R. Jordehi, Enhanced leader PSO (ELPSO): a new PSO variant for solving global optimisation problems. Appl. Soft Comput. **26**, 401–417 (2015)
10. F. Cecconi, M. Campenni, PSO (particle swarm optimization) one method many possible application, in *Agent-Based Evolutionary Search*. (Springer, Berlin, 2010), pp.229–254
11. Z.W. Geem, J.H. Kim, G.V. Loganathan, A new heuristic optimization algorithm: harmony search. Simulation **76**(2), 60–68 (2001)

12. X. Yang, Firefly algorithm, levy flights and global optimization. In. research and development intelligent systems XXVI, pp. 209–218 (2010)
13. D. Karaboga, B. Basturk, A powerful and efficient algorithm for numerical function optimization: Artificial Bee Colony (ABC) algorithm. J. Global Optim. **39**(3), 459–471 (2007)
14. S.M. Guo, C.C. Yang, P.H. Hsu, J.S.H. Tsai, Improving differential evolution with a successful-parent- selecting framework. IEEE Trans. Evolut. Comput. **19**(5), 717–730 (2015)
15. Y. Cheng, G.M. Church, Biclustering of expression data. Intell. Syst. Mol. Biol. **8**(2000), 93–103 (2000)
16. A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, A novel coherence measure for discovering scaling biclusters from gene expression data. J. Bioinf. Comput. Biol. **7**(05), 853–868 (2009)
17. K. Thangavel, J. Bagyamani, R. Rathipriya, Novel hybrid PSO-SA model for biclustering of expression data. Proc. Eng. **30**, 1048–1055 (2012)
18. J. Adhikary, S. Acharyya, Identification of biologically relevant biclusters of gene expression dataset of Parkinson's disease using grey wolf optimizer. In proceedings of international conference on industrial instrumentation and control, pp. 119-128 (2022)
19. E. Fonseca, R. Gong, D. Bogdanov, O. Slizovskaia, E. Gomez, X. Serra,: Acoustic scene classification by ensembling gradient boosting machine and convolutional neural networks, In: Detection and classification of acoustic scenes and events (2017)
20. S. Das, A.K. Kolya, Predicting the pandemic: sentiment evaluation and predictive analysis from large-scale tweets on Covid-19 by deep convolutional neural network. Evolut. Intell. **15**(3), 1913–1934 (2022)
21. S. Adamović, V. Miškovic, N. Maček, M. Milosavljević, M. Šarac, M. Saračević, M. Gnjatović, An efficient novel approach for iris recognition based on stylometric features and machine learning techniques. Future Generat. Comput. Syste. **107**, 144–157 (2020)
22. D. Bhattacharjee, D.K. Basu, M. Nasipuri, M. Kundu, Human face recognition using fuzzy multilayer perceptron. Soft Comput. **14**(6), 559–570 (2010)
23. R. Jana, S. Bhattacharyya, S. Das, Epileptic seizure prediction from EEG signals using DenseNet. In: 2019 IEEE symposium series on computational intelligence (SSCI), pp. 604-609 (2019)
24. J.D. Bodapati, U. Srilakshmi, N. Veeranjaneyulu, FERNet a deep CNN architecture for facial expression recognition in the wild. J. Inst. Eng. (India) Ser. B **103**(2), 439–448 (2022)
25. V. Gupta, M. Mittal, V. Mittal, N.K. Saxena, A critical review of feature extraction techniques for ECG signal analysis. J. Inst. Eng. (India) Ser. B **102**(5), 1049–1060 (2021)
26. V. Gupta, N. K. Saxena, A. Kanungo, P. Kumar, S. Diwania, PCA as an effective tool for the detection of R-peaks in an ECG signal processing. Int. J. Syst. Assur. Eng. Manag. 1-13 (2022)
27. V. Gupta, M. Mittal, V. Mittal, A novel FrWT Based arrhythmia detection in ECG signal using YWARA and PCA. Wirel. Personal Commun. **124**(2), 1229–1246 (2022)
28. V. Gupta, M. Mittal, V. Mittal, An efficient low computational cost method of R-peak detection. Wirel. Personal Commun. **118**(1), 359–381 (2021)
29. V. Gupta, M. Mittal, Arrhythmia detection in ECG signal using fractional wavelet transform with principal component analysis. J. Inst. Eng. (India) Ser. B **101**(5), 451–461 (2020)
30. V. Gupta, M. Mittal, V. Mittal, N.K. Saxena, BP signal analysis using emerging techniques and its validation using ECG signal. Sens. Imag. **22**(1), 1–19 (2021)
31. V. Gupta, M. Mittal, V. Mittal, Chaos theory and ARTFA: emerging tools for interpreting ECG signals to diagnose cardiac arrhythmias. Wirel. Personal Commun. **118**(4), 3615–3646 (2021)
32. V. Gupta, M. Mittal, V. Mittal, Chaos theory: an emerging tool for arrhythmia detection. Sens. Imag. **21**(1), 1–22 (2020)
33. V. Gupta, M. Mittal, QRS complex detection using STFT, chaos analysis, and PCA in standard and real-time ECG databases. J. Inst. Eng. (India) Ser. B **100**(5), 489–497 (2019)
34. V. Gupta, M. Mittal, V. Mittal, A. Gupta, An efficient AR modelling-based electrocardiogram signal analysis for health informatics. Int. J. Med. Eng. Inf. **14**(1), 74–89 (2022)
35. V. Gupta, M. Mittal, R-peak detection for improved analysis in health informatics. Int. J. Med. Eng. Inf. **13**(3), 213–223 (2021)
36. V. Gupta, M. Mittal, V. Mittal, FrWT-PPCA-based R-peak detection for improved management of healthcare system. IETE J. Res., 1–15 (2021)
37. V. Gupta, M. Mittal, V. Mittal, Y. Chaturvedi, Detection of R-peaks using fractional Fourier transform and principal component analysis. J. Ambient Intell. Human. Comput. **13**(2), 961–972 (2022)
38. V. Gupta, M. Mittal, Efficient R-peak detection in electrocardiogram signal based on features extracted using Hilbert transform and Burg method. J. Inst. Eng. (India) Ser. B **101**(1), 23–34 (2020)
39. J. Adhikary, S. Acharyya, Identification of biologically relevant biclusters from gene expression dataset of duchenne muscular dystrophy (DMD) disease using elephant swarm water search algorithm. In: emerging technologies in data mining and information security, pp. 147–157 (2021)
40. S. Mirjalili, S.M. Mirjalili, A. Lewis, Grey wolf optimizer. Adv. Eng. Softw. **69**, 46–61 (2014)
41. S. Mukherjee, Immune gene network of neurological diseases: Multiple sclerosis (MS), Alzheimer's disease (AD), Parkinson's disease (PD) and Huntington's disease (HD). Heliyon **7**(12), e08518 (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.