SIMPLEX BASED SOCIAL SPIDER OPTIMIZATION METHOD FOR IMPROVING MEDICAL DATA ANALYSIS

Monalisa Nayak[1], Soumya Das[2], Urmila Bhanja[3] and Manas Ranjan Senapati[4]

[1]Indira Gandhi Institute of Technology, Sarang, Dhenkanal, India
[2]Government College of Engineering, Kalahandi, India
[3]Indira Gandhi Institute of Technology, Sarang, Dhenkanal, India
[4]Veer Surendra Sai University of Technology, Burla, Sambalpur, India

ABSTRACT

Accurate and reliable prediction is the only way to prevent the disease transmission. Many machine learning models have been developed for prediction of large scale medical datasets. In this paper, Simplex based Social Spider Optimization method is used for classification of three types of medical datasets like heart disease, echocardiogram and hepatitis. The performance of the model is obtained by using Root Mean Square Error (RMSE) and time.

KEYWORDS

Social Spider Optimization (SSO), Simplex based Social Spider Optimization (SMSSO), Root Mean Square Error (RMSE), Support Vector Machine (SVM) & Random Forest Model (RFM).

1. INTRODUCTION

Public Health Organisation are working on the decreasing or preventing the transmission of disease which is possible only if the prediction of the disease is done in an accurate and reliable manner. Various models are developed for this work. Model performance varies for different datasets. Machine learning is an effective method to take decisions and prediction of a large amount of medical data provided from health care department. Autoregressive integrated moving average (ARIMA), support vector machine (SVM) and long short-term memory (LSTM) recurrent neural network were developed to predict the Hepatitis E which is a severe liver disease. These three methods are compared on the basis of RMSE, MAPE and MAE. Results obtained are 0.022, 0.0204 and 0.01 for ARIMA, SVM and LSTM respectively. ARIMA is solved using python, SVM is solved by MATLAB and LSTM by keras [1]. A model was developed using four algorithms like extreme gradient boosting (XGBoost), random forest (RF), decision tree (DCT) and logistic regression (LR). The optimal model is attained using the area under the receiver operating characteristics curve (AUC). The AUC value are 0.891, 0.829, 0.619 and 0.680 for XGBoost, RF, DCT and LR respectively which show that XGBoost is the optimal machine learning model for prediction of Hepatitis B surface antigen (HBsAg) [2]. A novel method is developed that finds important features with application of machine learning methods that results in increasing the accuracy of the prediction of cardiovascular disease. The results show an accuracy of 88.7% with help of hybrid random forest with a linear model (HRFLM) which is a prediction model [3]. Random forest model is a machine learning model that was developed to predict accurately the survival of the patient after echocardiography. This model uses clinical datasets and results in AUC>0.82. Here, there is comparison of non-linear models

(RFM) with linear models i.e. logic regression (LR). So, results show the non-linear model outperforming linear model [4].
This paper can be summarized as follows: Section 2 discusses about the classification methods. Implementation details are explained in Section 3 and in Section 4, results are obtained. Section 5 draws the Conclusion.

2. CLASSIFICATION METHODS

2.1. Social Spider Optimization (SSO)

In 2013, SSO technique was developed by Erik Cuevas [5, 6, and 7]. The SSO avoids premature convergence and getting trapped in local minima. SSO provides global solution in terms of accuracy and has a good balance between exploration and exploitation. In this method, there are two search agents that is female and male spiders. Here, the population of male is fewer than that of female population which is around 65-90% [5, 8]. The mathematical model is explained by Erik Cuevas in [5].The mathematical model can be described as:

2.1.1 Gender Assignment:

The In SSO, female population dominates over male population. The calculation of number of females M_f is shown below [5, 6, and 8]:

$$M_f = floor\ [(0.9\text{-}rand \times (0.25)) \times P] \quad (1)$$

where P= total population
rand= random value within [0,1]

Here, the female population is chosen randomly which is 65-90% of the whole population 'P'. Calculation of male population M_m is given as:

$$M_m = P\text{-}M_f \quad (2)$$

2.1.2. Fitness*:*

In the total population PP, each individual's fitness is calculated by receiving weight w_k which display the outcome value of every spider 'n' unbiased of gender.

So, weight w_k of each individual is given by:

$$w_n = \frac{J(PP_n) - worst_{PP}}{best_{PP} - worst_{PP}} \quad (3)$$

where $worst_{PP}$= worst individual

$best_{PP}$= best individual

$J(PP_n)$= the value of fitness which is given by evaluation of spider position PP_n as per the objective function $J(.)$

2.1.3 Vibration in their webs*:*

During optimization, usually the spiders interact by the vibration of their string in their given webs. The spider which gets the vibration dependent on the distance within them and the spider size that sends the vibration.

$$V_{na_{n,t}} = w_t . e^{-d_{n,t}^2} \quad (4)$$

$d_{k,l}$= Euclidian distance within n[th] and t[th] spiders

$V_{na_{n,t}}$ = Vibration formed by spider 'n' that is given by spider 't'.

In this method, the spiders feels only three vibrations from other spiders' like-(i) nearby spider with higher fitness ($V_n a_{n,c}$), (ii) the spider that is better in the swarm ($V_n a_{n,best}$) and (iii) the female spider that is nearby and accessible only to male spiders ($V_n a_{n,f}$).

2.1.4 Cooperative behaviour of female spiders*:*

There is cooperative behaviour between the social spiders. Female spider's cooperative behaviour is matched by a new operator shown in equation (15). In between the spiders an act of attraction or repulsion is developed due to the generation of vibration by the spiders which radiates all over the communal web that is calculated as:

$$A_n(l+1) = \{A_n(l) + \propto * V_n a_{n,c}(f_c - A_n(l)) + \beta * V_n b_{n,best}(f_a - f_n(l)) + \gamma(rand - 0.5) , \quad r < TH\} \quad (5)$$

$$A_n(l+1) = \{A_n(l) - \propto * V_n b_{n,c}(f_c - A_n(l)) - \beta * V_n a_{n,best}(f_a - f_n(l)) + \gamma(rand - 0.5) , \ r \geq TH\} \quad (6)$$

where $\propto, \beta, \gamma, r$and *rand* are random numbers between [0, 1]

i= iteration number

TH= threshold value

f_c = close member to spider *n* with a large weight

f_b =close member to spider *n* that is the best individual in the total population.

2.1.5 Cooperative Behaviour of Male Spiders:

Male spiders divides into dominant and non-dominant classes. Dominant classes are better fit than the non-dominant classes and are mostly attracted to the female spiders in the public web.

$$B_n(l+1) = \{B_n(l) + \propto * V_n b_{n,f}(f_s - B_n(l) + \gamma(rand - 0.5) \ if \ w_{M_f+n} > w_{M_f+B}\} \quad (7)$$

$$B_n(l+1) = \{B_n(l) + \propto * \{\frac{\sum_{h=1}^{M_m} B_h(l) * w_{M_f+h}}{\sum_{h=1}^{M_m} w_{M_f+h}} - B_n(l)\}\} \ if \ w_{M_f+k} \leq w_{M_f+B} \quad (8)$$

where f_s= female individual i.e. nearest to *n* male member

$\frac{\sum_{h=1}^{M_m} B_h(l)*w_{M_f+h}}{\sum_{h=1}^{M_m} w_{M_f+h}} - B_n(l)$= weighted mean of the male population M

2.1.6 Mating Operator

This operator is used to modify search agents. Equation 2 calculates the mating radius where the dominant class males mate with the females. When more number of males and females are present then a roulette wheel is employed to arbitrarily search the parents according to their fitness. Thus, a new spider is produced by mating of male and female spiders. After new spider generation, the male and female spider's fitness are designed and it is related with the worst spider of the whole population. Hence, if the new generated spider is fit then it enters the population whereas the spider with less fitness is removed.

$$S = \sum_{l=1}^{i} \frac{(P_n^{high} - P_j^{low})}{2-i} \quad (9)$$

where q= dimension of the problem
P_n = upper bound
P_j = lower bound

Pseudocode of SSO
SSO (Input: population size; Output: fittest spider)
Initialize the population size
While i<max iteration
Task of the gender by the use of equation (1) and (2)
Task of the fitness by the use of equation (3)
Vibration calculation by the equation (4)
Males and Females cooperative behaviour calculation by the use of equation (5-8)
Performance of the mating process by the use of equation (9)
End while

2.2. Simplex based Social Spider Optimization (SMSSO) algorithm

During complex problems the SSO leads to higher computational cost because of which there are local optima entrapment and poor convergence rate. Hence, simplex method is added to SSO so that it can enhance the local and global search abilities, increase the rate of convergence and help in escaping from getting stuck in local optima [5, 9].

Figure 1 show the graphical view of the simplex method in which the whole process is defined [5, 10, 11, and 12]. The simplex method process can be defined as given below:

(i) First, evaluation of each and every spider in the population is performed. After that the global test P_a & second best P_c are chosen by assuming T_b spider to be replaced. $f(P_a)$,$f(P_b)$&$f(P_c)$ are the corresponding values of fitness.

(ii) The middle position (P_f) of P_a & P_c is given as:

$$P_f = (P_a+P_c)/2 \quad (10)$$

(iii) The reflection point (P_h) is discussed as:

$$P_h = P_f+\alpha(P_f-P_b) \quad (11)$$

α= reflection coefficient which set to 1.

(iv) Comparison of fitness values between P_h & P_a.
If $f(P_h)<f(P_a)$ then the extension process is done as:

$$P_d = P_f + \gamma(P_h - P_f) \quad (12)$$

where γ = extension coefficient which is set to 2

So, there is comparison between fitness standards of global best P_a is with the extension point P_d
If $f(P_d) < f(P_a)$ then P_b is used as substitute of P_d else P_h is replaced with P_b

(v) There is comparison of fitness values of P_h with P_b

If $f(P_h) > f(P_b)$

The compression operation is activated as follows:

$$P_g = P_f + \beta(P_b - P_f) \quad (13)$$

Where β= condense coefficient is set to 0.

Hence, there is comparison of fitness standards of condense point P_g with P_b

If $f(P_g) < f(P_b)$

Then, P_b is used as substitute for P_g else P_h is replaced by P_b

(vi) If $If(P_a) <f(P_h) <f(P_b)$

The shrink operations are performed for the documentation of condex point P_e where β is considered as a shrink coefficient.

$$P_e = P_f - \beta(P_b - P_f) \quad (14)$$

If $f(P_e) < f(P_b)$ then P_b is a substitute with P_e else P_h is replaced by P_b.

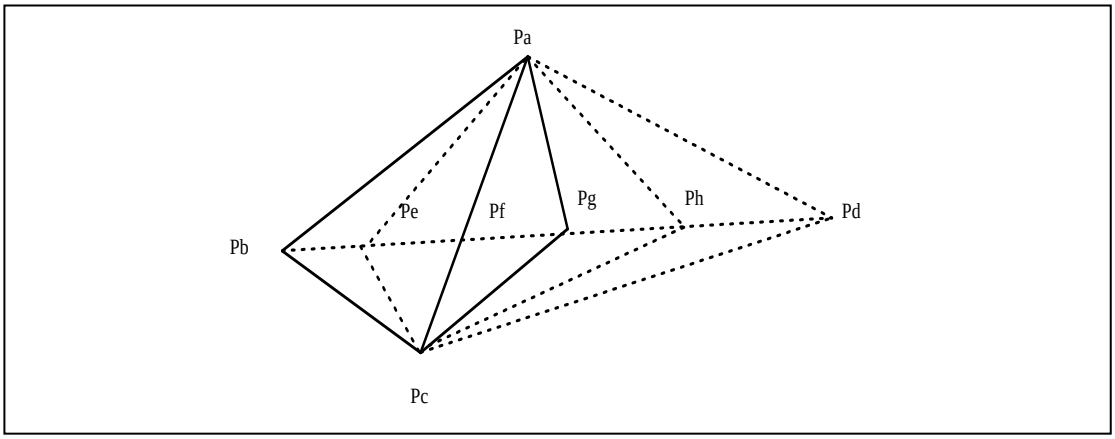

Fig.1. Schematic structure of the simplex method

SMSSO (Input: Initialization of spider population,: $\propto, \beta, \gamma, r, i$; Output: fitness value f(P_{new}))
The description of the SMSSO algorithm is discussed as follows:

Step-1: Initialization of the spider population is done where every spider is considered as one clustering vector with the equation 1 and 2.
Step-2 setting of some parameters like maximum size of iterations, $\propto, r, \beta, i, \gamma$ *and the threshold value (TH). Step-3: For all* P_k*, the fitness f(*P_k*) is given as [4]:*

$$f(Y, X) = \sum_{i=1}^{n} min(\|x_i - z_v\| | v = 1,2 \ldots . V) \quad (15)$$

The Y dataset with n attributes and V clusters with X as a center vector of a cluster. Here the in between each datum y_i *and the center of cluster* z_v *is minimized. Then, weights assignment is done as per equation 3.*
Step-4: the best solution is found and with the use of equation 4 the three spiders are calculated from other spiders.
Step-5: the position of the female spider is reorganised by equation 5.
Step-6: the position of the male spider is reorganised by equation 6.
Step-7: the mating radius is designed by equation 7.
Step-8: when the spiders are present in the mating region, then new spiders are created
Step-9: otherwise the worst spider position reorganised with the simplex method by the help of equations 8-12.
Step-10: if f(P_{new}*) > f(*P_{worst}*), then the new spiders are replaced by worst spiders, otherwise step 10.*
Step-11: the fitness rate of the spider is designed and the best solution is found.
Step-12: i=i+1
Step-13: Stop otherwise move to Step-5.

3. FINDING AND ANALYSIS

3.1. Dataset Details: The heart disease dataset can be found at [13]. The dataset contains total 270 rows and 13 columns. The Echocardiogram dataset is available at [14]. The dataset contains 132 rows and 12 columns. The hepatitis dataset is available at [15]. The dataset contains 150 rows and 13 columns.

3.2. Implementation Details: The performance validation of the classifier is done using three datasets. 70% of the dataset is considered as training and 30% of the dataset is taken as

testing. Pre-processing of data is attained by omitting the missing values. Then, data normalization is performed using min-max normalization which is discussed below:

$$X_{normalised}=\frac{Xoriginal\ -Xminimum}{Xmaximum\ -Xminimum} \quad (16)$$

The parameters used to assess the classifier's performance are:

1. **Root Mean Square***:*

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(Y_n - T_n)^2}{N}} \quad (17)$$

where Yn = predicted output

T_n = target
N = data sample size

2. Computational Cost: time for execution of the classifier.

First dataset is trained with SSO-NN and then there is comparison of results with SMSSO-NN. The output is in binary format where for malignant it is set as 1 and for benign it is set as 0.The classifier takes 100 iterations with 50 runs. Computation takes place using MATAB 14 with Intel (R) Core (TM) I3-40300 PC with clock speed of 1.9 GHz.

4. RESULTS AND DISCUSSIONS

The results are shortened with tables and graphs. Performance increment in the tables are shown in bold format.

Table 1 describes the RMSE and Time in seconds of SMSSO and SSO of different datasets like Hepatitis, Electrocardiogram and Heart disease. Table 2 show the classification accuracy comparison between SMSSO-NN and SSO-NN with Hepatitis, Electrocardiogram and Heart disease datasets. Table 3 obtains the confusion matrix of SMSSO-NN in Heart Disease, Echocardiogram and Hepatitis.

Table 1. Details of Dataset

Details of dataset	RMSE		TIME(in seconds)	
	SMSSO-NN	SSO-NN	SMSSO-NN	SSO-NN
Heart Disease	0.0023	0.0345	5.2314	6.2316
Echocardiogram	0.0153	0.0231	4.3187	5.9834
Hepatitis	0.0126	0.0121	4.6721	7.5641

Table 2. Comparison of Classification Accuracy

Dataset	Class	Affected	Not affected	% Classification
Heart Disease	Affected	127	3	97.7
	Not Affected	10	140	93.33
Dataset	Class	Alive	Dead	% Classification
Echocardiagram	Alive	50	05	90.9
	Dead	4	73	94.8
Hepatitis	Alive	115	3	97.45
	Dead	2	30	93.75

Table 3. The confusion matrix obtained from SMSSO in Heart Disease, Echocardiogram and Hepatitis

Types of Datasets	Techniques	Classification accuracy in percentage (%)
Heart Disease	"SSO-NN"	93.4
	"SMSSO-NN"	**95.52**
Echodiagram	"SSO-NN"	90.5
	"SMSSO-NN"	**92.9**
Hepatitis	"SSO-NN"	90.3
	"SMSSO-NN"	**95.6**

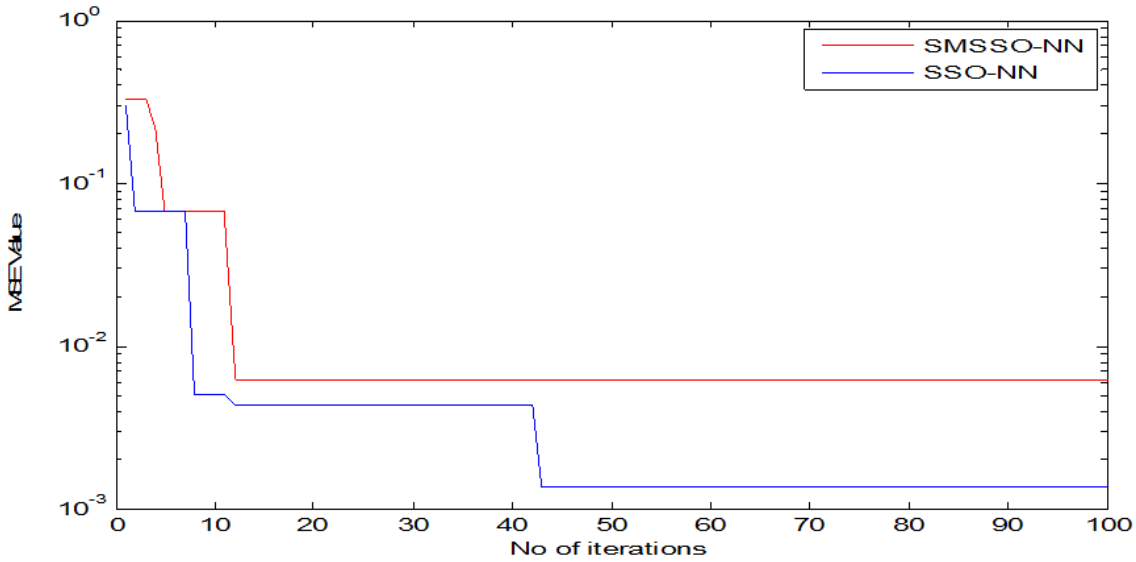

Fig 2. MSE convergence curve of Electrocardiogram

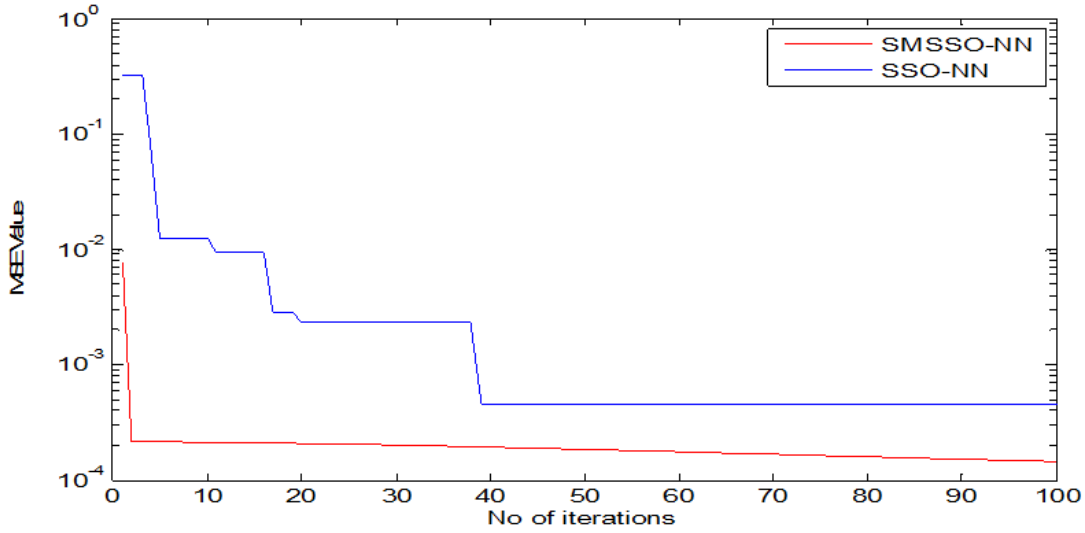

Fig 3. MSE convergence curve of Hepatitis

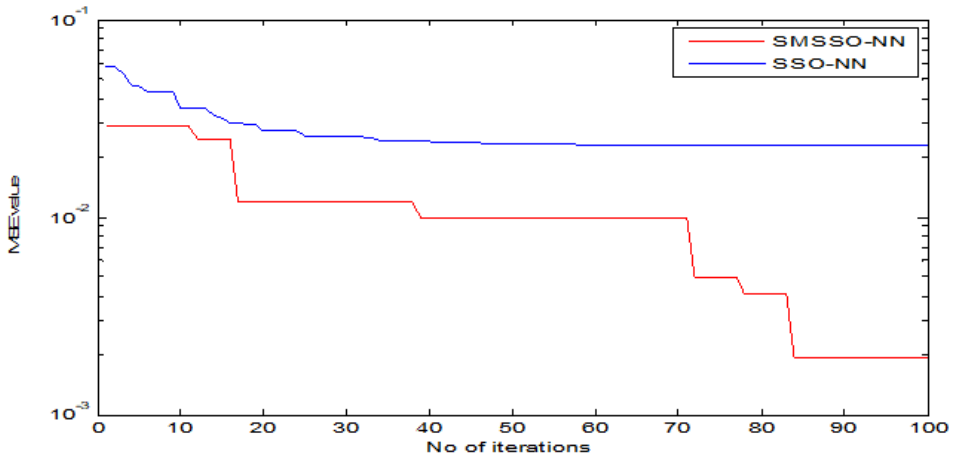

Fig 4. MSE convergence curve of heart disease

Figure 2, 3 and 4 describes the MSE convergence curve of Electrocardiogram, Hepatitis and primary tumor

5. CONCLUSION

Medical data keeps varying from time to time. The data essentially contains variables and constraints but it functions as a structural model for the organization. Early prediction is the key in case of medical data. In that context many methods have been developed to analyze medical data.

In this paper a comparison is presented between SSO and SMSSO for medical data. Three types of medical data i.e. Heart disease, Echocardiagram and Hepatitis are taken for analysis. The methods are compared and results are stated in form of graphs and figures. From the results the superiority of the method can be proved.

In future more no of datasets may be taken in accordance with complex networks for more promising results.

REFERENCES

[1] Yanhui Guo et.al. "Prediction of hepatitis E using machine learning models", September 17, 2020.https://doi.org/10.1371/journal.pone.0237750

[2] Xiaolu Tian et al. "Using Machine Learning Algorithms to Predict Hepatitis B Surface Antigen Seroclearance", Vol 2019 https://doi.org/10.1155/2019/6915850

[3] Senthilkumar Mohan; Chandrasegar Thirumalai; Gautam Srivastava, Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques 10.1109/ACCESS.2019.2923707, IEEE Access (Volume: 7), Page(s): 81542 – 81554

[4] Manar D. Samad et al. "Predicting Survival From Large Echocardiography and Electronic Health Record Datasets: Optimization With Machine Learning" J Am Coll Cardiol Img. 2019 Apr, 12 (4) 681–689

[5] Zhou, Y et al.: "A Simplex method-based Social Spider Optimization algorithms for clustering analysis", Engineering Applications of Artificial Intelligence, 64, 67-82 (2017).

[6] Cuevas E; Cienfuegos M. "A new algorithm inspired in the behavior of the social spider for constrained optimization". Expert Syst. Appl. Elsevier, vol-41, issue-2, pp 412–425, 2014, 10.1016/j.eswa.2013.07.067

[7] Cuevas, E.; Cienfuegos, M.; Zaldívar, D.; Pérez-Cisneros, M.: A swarm optimization algorithm inspired in the behaviour of the social-spider, Expert Syst. Appl. 40(16), 6374–6384(2013). 10.1016/j.eswa.2013.05.041

[8] Aviles, L.: Sex-ratio bias and possible group selection in the social spider anelosimus eximius, University of Chicago, Press J, 128(1), 1–12(1986). https://www.jstor.org/stable/2461281

[9] Spendley, W.G.R.F.R.; Hext, G.R.; Himsworth, F.R.: Sequential application of simplex designs in optimisation and evolutionary operation, Taylor & Francis, 4(4), 441–461(1962), doi: 10.2307/1266283

[10] Nelder, J.A.; Mead, R.: A simplex method for function minimization, Comput. J. 30, 8–313(1965). 10.1093/comjnl/7.4.308

[11] Yen, J.; Lee, B.: A simplex genetic algorithm hybrid. Evolutionary Computation, IEEE Conf.175–180(1997). 10.1109/ICEC.1997.592291

[12] Storn, R.; Price, K.: Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. J. Global Optim. 11, 341–359(1997). 10.1023/A:100820282132

[13] https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29

[14] https://archive.ics.uci.edu/ml/datasets/Echocardiogram

[15] https://archive.ics.uci.edu/ml/datasets/Hepatitis

AUTHORS

Soumya Das completed her master's degree from IIIT Bhubaneswar. Currently, she is working as an Assistant Professor in the Department of Computer science Engineering, Govt. College of Engineering, Kalahandi, Her area of research include soft computing, optimization techniques, and big data. Data Mining

Monalisa Nayak completed her master's degree from S'O'A University. Currently, she is working as an Assistant Professor in the Department of Electronics & Communication Engineering, Indira Gandhi Institute of Technology, Sarang. Her area of research include soft computing, optimization techniques, and big data. Data Mining.

Dr. Urmila Bhanja received her PhD. from the Dept. of E & ECE, Indian Institute of Technology, Kharagpur, West Bengal, India in the year 2011. Currently, she is working as an Associate Professor in the Department of Electronics & Communication Engineering, Indira Gandhi Institute of Technology, Sarang, India. She has published many papers in reputed international journals and conferences. She is also a reviewer for many reputed international journals and conferences. Her area of research includes optical communication, optical network, wireless communication, wireless network, soft computing, optimization techniques, and big data.

Dr. Manas Ranjan Senapati received his PhD. from the Dept. of CSE from BPUT. Currently, he is working as an Associate Professor in the Department of Computer science Engineering, Veer Surendra Sai University of Technology, Burla. He has published many papers in reputed international journals and conferences. He is also a reviewer for many reputed international journals and conferences. His area of research includes, soft computing, optimization techniques, and big data. Data Mining, Big data Analysis, Pattern Analysis, Clustering.