DISEASE CATEGORIZATION WITH CLINICAL DATA USING OPTIMIZED BAT ALGORITHM AND FUZZY VALUE

Mr.Naresh Patel K M[1] **Mr.Shivkumarswamy G M[2]** **Dr.Ashoka K[3]**

[1]Research Scholar, CSE Research center, Assistant Professor, B.I.E.T., Davangere

[2]Assistant Professor, Dept. Of EE, B.I.E.T., Davangere

[3]Associate Professor, Dept. Of ISE, B.I.E.T., Davangere

Email address: nareshpatela.is@gmail.com, shivkumar_gm@rediff.com, ashoka_kkd1@yahoo.com

Abstract

In this paper, design a Bat-based Random Forest (BbRF) framework to enhance the performance of categorizing diseases with fuzzy values which also protect the privacy of the developed scheme. It involves pre-processing, attributes selection, fuzzy value generation, and classification. Additionally, the developed framework is implemented in Python tool and patient disease datasets are used for implementation. Moreover, pre-processing remove the error and noise, attributes are selected based on the duration of diseases. Finally, classify the patient disease based on the generated fuzzy value. To prove the efficiency of the developed framework, attained results are compared with other existing techniques in terms of accuracy, sensitivity, specificity, F-measure, and precision.

KEYWORDS: Bat-based Random Forest, Fuzzy value, Optimization.

DOINumber: 10.14704/NQ.2022.20.12.NQ77174 **NeuroQuantology2022;20(12): 2006-2015**

1. INTRODUCTION

Nowadays, the classification of medical diseases based on Machine Learning (ML) is the most challenging task because of nature data that contains imprecise, uncertain, and incomplete information [1]. Moreover, medical data analysis is the most sensitive problem which is essential for developing a classification system and accurately predicting and easily diagnosing disease [2, 3]. Generally, Fuzzy Logic (FL) is analogous to the human thinking framework which is used to control the vagueness of the data [4]. Also, FL proves as a valuable tool to classify the problems which are used to employ the rules, improve interpretability and provide more insight into classification structure [5].

By increasing the use of computational intensive techniques, ML has the efficiency to create revolution consequently, still has some clinical problems like a quick decision, less reliability, and less accuracy [6]. Additionally, many medical datasets contain redundant, noise, incomplete and irrelevant information that may degrade the performance of classification [7]. Furthermore, the performance of disease prediction is processed based on the excellence of medical data also used for the classification process [8]. Moreover, classification is the most significant part to categorize the disease in an accurate way for predicting and diagnosing disease [9]. Also, it is an important task in pattern recognition and ML which extracts the real-world problems also construct to predict the target class from the medical dataset [10]. Many classification algorithms are designed to predict the disease it attain less accuracy, low flexibility, and slow convergence because of less capability and more time to predict disease [11].

To overcome these issues FL framework is utilized for converting the numerical input types into corresponding linguistic terms such as low, medium, and high [12]. Consequently, FL is useful for handling uncertainty problems by defining the relationship value for various linguistic terms [13]. New techniques are developed to enhance the development of agriculture, industry, medicine, and commerce. Particularly, in the medical field healthcare sector is used for providing better service at a reasonable cost also enhancing the hospital information system by monitoring the health of the patient.

The arrangement of this article is structured as follows. The related works based on the classification of diseases are detailed in section 2 and the system model and problem statements are elaborated in section 3. Also, the process of the proposed methodology is described in section 4. Finally, the achieved outcomes are mentioned in section 5 and the conclusion about the developed model is detained in section 6.

2. RELATED WORKS

A few recent literature surveys based on the disease diagnosis are detailed below,

G.Thippa Reddy et al [11] has proposed a fuzzy logic system with an adaptive genetic framework for diagnosing the disease in an early stage. It can minimize the risk of death and treatment costs. The attained results show the efficiency of the developed technique but the error rate is high because of data complexity.

Farhad Pourpanah et al [14] developed Fuzzy based neural network scheme to enhance feature selection and overcome classification problems. However, misclassification issues have occurred during classification.

Zhennao et al [16] proposed a chaotic bacterial foraging algorithm by gauss mutation framework for resolving the issues of parameter tuning. Moreover, simulation results indicate the performance of Fuzzy k-Nearest Neighbour (FKNN). Furthermore, SVM and kernel learning is used for feature selection. The limitation of this approach is,it has the problem of less sensitivity.

Liaqat Ali et al [15] developed SVM based expert system to predict heart failure. Furthermore, the designed model indicates better performance also attains 91.83% accuracy. However, classification and regression issues are happened because of vast data.

Hager Ahmed et al [12] proposed a real-time system to predict heart disease also label the current health status of the patient. Moreover, the main aim is to identify an optimal path for attaining high accuracy to predict heart disease. Finally, attain accuracy rate is 94.9% but the execution time is high when comparing other techniques.

The key contribution of our research is summarized as,

- ❖ Initially, the healthcare of various patient datasets is collected and they are trained in the system.
- ❖ Consequently, a novel BbRF is designed to classify the patient disease based on the fuzzy results.
- ❖ Hereafter, preprocessing is utilized to remove the error and noise present in the dataset.
- ❖ Then select the attributes based on the duration of disease, generate the fuzzy value using the fitness of bat optimization
- ❖ Moreover, categorize the disease with fuzzy values based on the duration of a week, month, and year
- ❖ Finally, the validated metrics are compared with other models in terms of accuracy, execution time, sensitivity, specificity, F-measure, and precision.

3. PROBLEM DEFINITION

As well, the prediction of disease using ML is the most trending topic in the early diagnosis of disease. But it contains some issues for predicting diseases such as less accuracy, error rate, misclassification, and long execution time. The main issues to categorize disease are diagnosisdisease problem, pattern recognition, and classification problems. These issues are motivated to develop a new optimized technique based on machine learning using fuzzy values. Furthermore, using the fuzzy value easily categorizes the disease also attains the best accuracy and less execution time.

4. PROPOSED METHODOLOGY

The categories of disease are more significant to enhance healthcare quality. So designing a Bat-based Random Forest (BbRF) to enhance the performance of categorizing disease with fuzzy value also protect the privacy of the developed scheme. Initially, the healthcare dataset of various patients is collected and trained in the system.

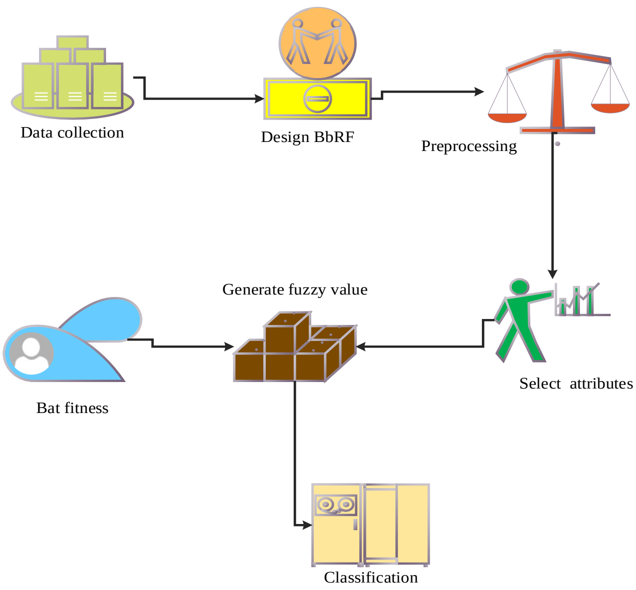

Fig.1 Proposed methodology

Then the collected dataset is updated to the designed model to categorize the disease with fuzzy values based on the duration of a week, month, and year. Moreover, pre-processing is processed to eliminate the errors and noise present in the dataset. The architecture of the developed technique is shown in fig.1. Additionally, select the attributes based on the relevant features from the dataset. After that, generate the weight based on the fuzzy rule to categorize disease in a numerical format. Finally, update the fitness of bats in the classification layer to enhance the categories of disease. Based on the duration of the dataset provide fuzzy and developed techniques attained results are compared with other states of the art techniques.

4.1 Data collection

The proposed approach utilized patient healthcare data to categorize the disease based on fuzzy values. The collected dataset contains patient ID, age, gender, Zip code, disease name, and duration. Additionally, the total number of diseases present in the dataset is 24 with various patients' IDs. Moreover, 70% of data are used for training and 30% of data are used for testing. Based on the duration of the disease, the developed framework generates the fuzzy value also categorizes the disease. Moreover, the collected dataset is given to the developed BbRF model for further processing.

4.2 Bat-based Random Forest mechanism

Design BbRF technique to categorize the disease of patients based on the fuzzy value. Moreover, fuzzy value is generated based on the selected attributes which have been selected depending on the duration of the disease such as week month, and year. Additionally, bat optimization [17] is updated in the random forest [18] prediction layer for enhancing the prediction results of disease. The basic process of bat optimization is searched and identifies the differences between prey/food using sensing of echolocation. They selected the target with frequency and velocity also use loudness for searching the prey. Thus the loudness of the bat varies from large to minimum constant value. The purpose of using the bat in developed techniqueis easily categorize the disease of the various patient based on the searching behaviour. In the input layer, collected datasets are tested and trained to the system then the dataset is processed to the next stage

.

- **Preprocessing**

The most important step in machine learning algorithm is data preprocessing which is required to test and effectively train the data. Moreover, it contains the process of removing missing values, removes errors, and noises. The main task of preprocessing phase is removing the noises and errors present in the dataset which helps to enhance the quality of the dataset. Furthermore, preprocessing is performed using Eqn. (1),

$$P^{t+1} = T(s)(h(s) - k(s)) \tag{1}$$

Where $T(s)$ is denoted as collected input dataset, $h(s)$ is represented as original data, and $k(s)$ is denoted as error and noise.

- **Select attributes**

The selection of attributes is useful for determining minimal attributes subset also attain high accuracy in demonstrating original attributes. The main issues in attribute selection are noisy, misleading, and irrelevant attributes. Moreover, the collected dataset contains three attributes a week, a month, and a year. Mainly, the attribute selection is processed based on the duration of each patient. Thus the search for the good attribute contains highly correlated decision attributes and the selection of attributes is obtained using Eqn. (2)

$$A_s = R\sum_{r=1}^{m} J_s(t)h(s) + P^{t+1} \tag{2}$$

Where, $J_s(t)$ is denoted as the selected attributes and R is represented as a decision set of attributes and m is the subset of selected attributes.

- **Generate fuzzy value**

The fuzzy values are generated using the fuzzy membership function, in this phase update the bat fitness function for accurate generation of fuzzy values based on the selected attributes. Consequently, input data are transferred to the membership degree through the membership function. Thus the membership function is utilized to modify the input data into fuzzy values. Thus the analysed principle of membership value is performed using Eqn. (3)

$$F_v = \frac{\sum_{r=1}^{m} O_m B_a \left(1/\left\|a - a_j\right\|^{2/(n-1)}\right)}{\sum_{r=1}^{m}\left(1/\left\|a - a_j\right\|^{2/(n-1)}\right)} \tag{3}$$

Where, O_s is denoted as fuzzy membership score, n is represented as several training data, a is denoted as membership data variable, a_j is represented as fuzzy strength and B_a is denoted as the bat fitness function. Thus the calculation of the fuzzy membership score is obtained using Eqn. (4)

$$O_s = \begin{cases} 0.51 + \left(\dfrac{j}{n}\right) * 0.49, & j = 1 \\ \left(\dfrac{j}{n}\right) * 0.49, & j \neq 1 \end{cases} \tag{4}$$

Where, j is represented as a data class. Moreover, testing data by the selected attributes are transferred to the developed framework. In this stage, gained fuzzy value classifies the patient disease in the classification stage.

- **Classification**

The classification process is performed using RF with generated fuzzy value. Moreover, RF is one of the supervised machine learning algorithms which is used for regression and classification problems. It is processed by building a decision tree on different samples at last predicting the results based on the majority voting. Based on the fuzzy value, categorizes the patient disease which is obtained using Eqn. (5)

$$C(t)=\begin{Bmatrix} if\left(A_s=p\right) & 0 \\ if\left(A_s=q\right) & 0.5 \\ if\left(A_s=A\right) & 1 \end{Bmatrix} \tag{5}$$

Where, p is denoted as weekly disease, q is represented as monthly disease and A is denoted as a yearly disease. Finally, the developed framework accurately categorizes the disease with the help of fuzzy values and bat optimization. The process of the proposed BbRF is detailed in algorithm.1.

Algorithm:1 BbRF for categorizing patient disease with fuzzy value

Start
Input: patient disease dataset
Output: categorize disease
{
Initialization
For all $T(s)$ *//s=1,2,3,4…k*
{
Update the dataset *//input layer*
// patient name, patient ID, Zip code, etc.
}
End for
Pre-processing () *// improve the quality of the dataset*
For all $k(s)$
{
Remove error and noise
//remove patient name
}
End for
Select attributes () *// extract relevant attributes*
For all $J_s(t)$
{
Select duration *// week, month and year*
}
End for
Generate fuzzy value ()
Update fitness function of bat *// accurate classification of disease*
The fuzzy values are generated using Eqn. (3)
Generate $\mathbf{F}_v \rightarrow \mathbf{O}_s$ *//*F_v*- fuzzy value, //*O_s*- Fuzzy membership score*
Classification ()
// categorize disease with fuzzy value
If($A_s=0$ ***)*** *//*A_s *-attributes(duration)*

```
            {
            Categorize weekly disease                 // headache, stomach pain
            }
                else if( A_s = 0.5 )
                {
                Categorize monthly disease            // Jaundice
                }
                    Else if( A_s = 1 )
                    {
                    Categorize yearly disease         //cancer, heart attack
                    }
            End if
        Finest output solutions                       // output layer
        }
End
```

The designed model enhances the performance of the developed technique by attaining better experimental results. Additionally designed model overcomes the issues of misclassification and less prediction by generating fuzzy values.

Table.1 categorize patient disease with fuzzy value

Patient ID	Zip code	Duration	Fuzzy value	Disease
198336	577003	Week	0.0	Cough
198339	577005	Year	1.0	HIV
198346	577218	Month	0.5	Jaundice
198354	577213	Year	1.0	Kidney failure
198352	577601	Week	0.0	Sinusitis
198342	577221	Month	0.5	Tuberculosis

Subsequently, generate the fuzzy value based on the selected attributes of the developed framework. If the fuzzy value is 0 means, classified as weekly diseases, fuzzy value is 0.5 means, classified as monthly diseases and the fuzzy value is 1 means, classified as yearly diseases. Using the score of fuzzy membership score, generate the fuzzy score which is used to categorize the diseases efficiently. Hereafter, update the fitness function of the bat in the classification layer to categorize the disease. Finally, classify the patient disease based on the fuzzy value.

5. RESULTS AND DISCUSSION

In this approach, various patient healthcare dataset is collected and are trained to the system. Subsequently, developed technique training and validation graphs of accuracy vs. loss are shown in fig.2.

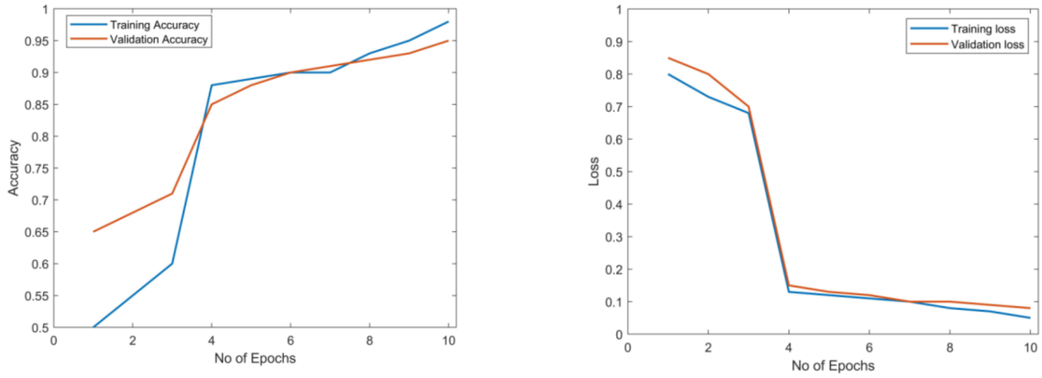

Fig.2Accuracy Vs. Loss

Furthermore, the introduced method classifies the patient disease based on the fuzzy value. Finally, a successive score of the proposed model is compared with other models to verify the efficiency of the designed model.

5.2 PERFORMANCE METRICS

Moreover, the developed approach is validated using existing methods in terms of accuracy, sensitivity, specificity, precision, execution time, and F-measure like Hybrid Genetic System (HGS) [11], Hybrid technique for Feature Selection and Data Classification (HFS-DC) [19], Enhanced Fuzzy KNN (EF-KNN) [26], Heart Failure Prediction (HFP) [20], and Stacking and Fuzzy Colour (SFC) [13] Model.The achieved performance metrics of the developed technique is compared with other existing techniques such as HGS, HFS-DC, HFP, EF-KNN, and SFC.

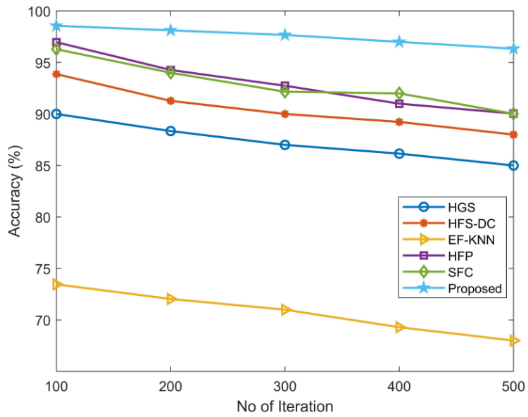

Fig.3 Comparison of accuracy

Moreover, the HGS and SFC replicas attained 90% and 96.29%, HFS-DC technique gained 93.87%. Also, EF-KNN and HFP techniques gained accuracy rates are 73.46% and 96.97%. The developed BbRF attained high accuracy while comparing other techniques to categorize the diseases as 98.56%. Also, the comparison of accuracy is graphically represented in fig.3.

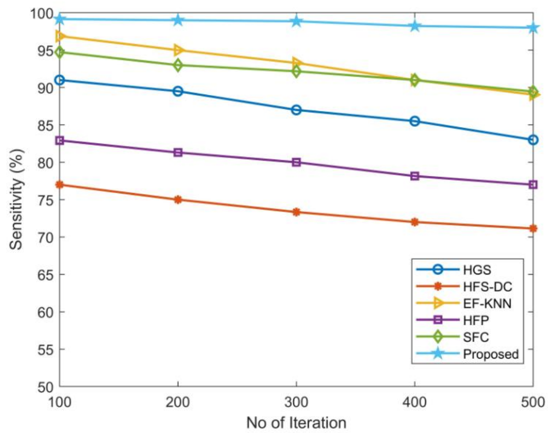

Fig.4 Comparison of sensitivity

Moreover, the HGS replica attained 91%, and the HFS-DC technique gained 77%. Also, EF-KNN and HFP techniques gained sensitivity rates are 96.87% and 82.92%.The SFC technique gained a 94.73% in sensitivity rate and the existing approaches have achieved lower sensitivity of almost 96% only. Also, the developed BbRF method has achieved a 99.15% high sensitivity value than other methods. Also, the comparison of sensitivity is graphically represented in fig.4.

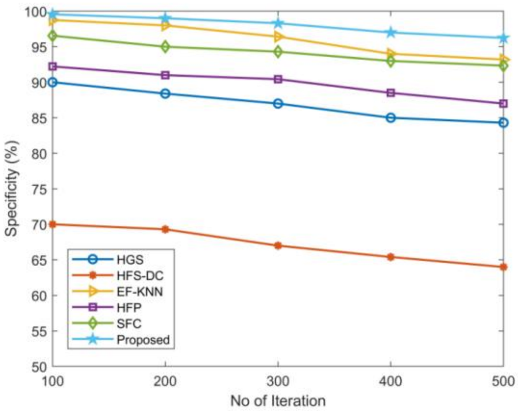

Fig.5 Comparison of Specificity

Additionally, the HGS replica attained 90% for 100 iterations. Also, HFS-DC and EF-KNN techniques gained specificity rates are 70% and 98.75%. Additionally, HFP and SFC methods attain 92.22% and 96.55%. The developed technique attained high specificity while comparing other techniques to categorize disease as 99.56%. The comparison of specificity is illustrated in fig.5.

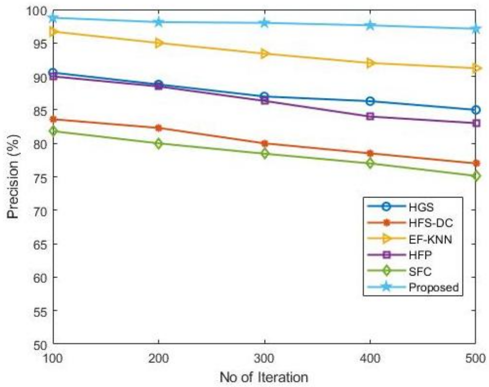

Fig.6 Comparison of Precision

Initially, the HGS technique gained a 90.56% in precision rate and the HFS-DC replica achieved 83.6% precision in 100 iterations. Moreover, the EF-KNN and HFP techniques gained 96.7% and 90%% precision rates. Finally, the developed BbRFtechnique gained a 98.76% in precision rate. Thus the comparison of precision rates is detailed in fig.6.

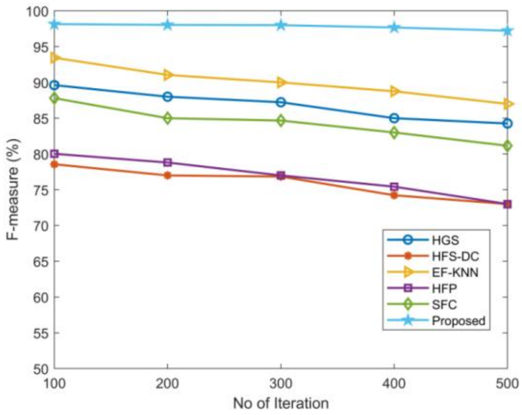

Fig.7 Comparison of F-measure

The comparison of the F-measure is detailed in fig.7.Initially, the HGS technique gained 89.62% in F-measure, the HFS-DC replica achieved 78.56% in precision in 100 iterations and EF-KNN methods attain 93.45% in F-measure. Moreover, HFP and SFC models attain 80.03% and 87.80% F-measure. Finally, the developed replica attain a high f-measure of 98.15%.

6. CONCLUSION

In this paper, design BbRF for the accurate classification of patient diseases based on the fuzzy value. The developed technique attains better performance to categorize the diseases using generated fuzzy values. The errors and noises present in the dataset are removed in preprocessing stage and then select the relevant attributes to enhance prediction results. Moreover, fuzzy values are generated based on the duration of diseases. Finally, the proposed model attained results are compared with other state-of-the-art techniques and the developed design attain better results in an accuracy of 98.56%, sensitivity of 99.15%, specificity of 99.56%, the precision of 98.76%, and F-measure as98.15%. It proves the efficiency of the developed framework also designed model can efficiently categorize disease.

REFERENCES

1. Ji, Miaomiao, and Zhibin Wu. "Automatic detection and severity analysis of grape black measles disease based on deep learning and fuzzy logic." Computers and Electronics in Agriculture 193 (2022): 106718.
2. Thakkar, Harshil, et al. "Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis." Clinical eHealth 4 (2021): 12-23.
3. Aamir, Khalid Mahmood, et al. "A Fuzzy Rule-Based System for Classification of Diabetes." Sensors 21.23 (2021): 8095.
4. Adhikari, Mainak, and AmbigavathiMunusamy. "ICovidCare: Intelligent health monitoring framework for COVID-19 using ensemble random forest in edge networks." Internet of Things 14 (2021): 100385.
5. Halim, Zahid. "An ensemble filter-based heuristic approach for cancerous gene expression classification." Knowledge-Based Systems 234 (2021): 107560.
6. Sarker, Iqbal H. "Machine learning: Algorithms, real-world applications and research directions." SN Computer Science 2.3 (2021): 1-21.
7. Cui, Sunan, et al. "Introduction to machine and deep learning for medical physicists." Medical physics 47.5 (2020): e127-e147.
8. Thabtah, Fadi, et al. "Data imbalance in classification: Experimental evaluation." Information Sciences 513 (2020): 429-441.

9. Das, Himansu, BighnarajNaik, and H. S. Behera. "Medical disease analysis using neuro-fuzzy with feature extraction model for classification." Informatics in Medicine Unlocked 18 (2020): 100288.
10. Zhang, Huaizheng, et al. "DeepQoE: A multimodal learning framework for video quality of experience (QoE) prediction." IEEE Transactions on Multimedia 22.12 (2020): 3210-3223.
11. Reddy, G. Thippa, et al. "Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis." Evolutionary Intelligence 13.2 (2020): 185-196.
12. Ahmed, Hager, et al. "Heart disease identification from patients' social posts, machine learning solution on Spark." Future Generation Computer Systems 111 (2020): 714-722.
13. Toğaçar, Mesut, BurhanErgen, and ZaferCömert. "COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches." Computers in biology and medicine 121 (2020): 103805.
14. Pourpanah, Farhad, et al. "A hybrid model of fuzzy min–max and brain storm optimization for feature selection and data classification." Neurocomputing 333 (2019): 440-451.
15. Ali, Liaqat, et al. "An optimized stacked support vector machines based expert system for the effective prediction of heart failure." IEEE Access 7 (2019): 54007-54014.
16. Cai, Zhennao, et al. "An intelligent Parkinson's disease diagnostic system based on a chaotic bacterial foraging optimization enhanced fuzzy KNN approach." Computational and mathematical methods in medicine 2018 (2018).
17. Paul, Angshuman, et al. "Improved random forest for classification." IEEE Transactions on Image Processing 27.8 (2018): 4012-4024.
18. Yılmaz, Selim, and Ecir U. Küçüksille. "A new modification approach on bat algorithm for solving optimization problems." Applied Soft Computing 28 (2015): 259-275.