Journal of Information Systems Engineering and Management
2025, 10(20s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Enhanced Hybrid Feature Selection with HRFBP Algorithm for Medical Data Classification

N. Kumar [1,2], Dr. T. Christopher[3]

[1]Assistant Professor, Dept. of Computer Science, Dr.N.G. P Arts and Science College, Coimbatore, Tamil Nadu, India
[2] Research Scholar, PG and Research Dept. of Computer Science, Government Arts College, Udumalpet, Tamil Nadu, India
&
[3]Associate Professor, PG and Research Dept. of Computer Science, Government Arts College, Coimbatore, Tamil Nadu, India

ARTICLE INFO

Received: 19 Dec 2024

Revised: 29 Jan 2025

Accepted: 10 Feb 2025

ABSTRACT

Medical industry produces a significant portion of data whereas by adopting various machine learning models which can make accurate predictions about public healthcare that can be generalised. Machine learning is ideal for examining complex high dimensional data because of its flexibility and automation compared to conventional approaches. Still, feature extraction is not performed effectively and hence overall classifier accuracy is reduced considerably. To overcome the above mentioned problems, in this work, Hybrid Random Forest with Back Propagation (HRFBP) neural network algorithm is proposed. Initially, the datasets are collected which is preprocessed using K-Means Clustering (KMC) algorithm. It is used to handle the missing values and error rates efficiently. Then, the feature extraction is done by using Modified Principal Component Analysis (MPCA) which is focused to extract the significant features from the given medical dataset. After that, the feature selection is done by using EFS algorithm which generates best fitness values via objective function. EFS is done dependent on integrating numerous FS rather than a single FS to handle the FS issue. The possibility of EFS method is that merging the outcomes of a various single FS methods like Entropy Elephant Herding Optimization (EEHO), Adaptive Firefly Optimization Algorithm (AFOA) and Entropy Butterfly Optimization Algorithm (EBFO) acquire improved outcomes rather than utilizing a single FS methodology. Finally, the medical dataset classification is performed using HRFBP algorithm. The HRFBP algorithm performs training and testing process which learns a set of weights for prediction for the class label of features. HRFBP increases the classifier accuracy and reduces the error rates prominently. From the experimental result, it concluded that the proposed HRFBP algorithm provides better performance in terms of the higher accuracy, sensitivity, specificity and lower execution time rather than the existing algorithms

Keywords: Medical dataset classification, Hybrid Random Forest with Back Propagation (HRFBP) neural network algorithm, Ensemble Feature Selection (EFS), Entropy Elephant Herding Optimization (EEHO), Adaptive Firefly Optimization Algorithm (AFOA) and Entropy Butterfly Optimization Algorithm (EBFO)

1. Introduction

Diagnosing the presence of disease is actually tedious process, as it requires depth knowledge and rich experience. In general, the prediction of disease lies upon the traditional way of examining medical report, for instance, heart disease report such as Electrocardiogram (ECG), Magnetic Resonance Imaging (MRI), Blood Pressure (BP), stress tests by a medical practitioner. Now days, a large volume of medical data is available in medical industry and acts as a great source of predicting useful and hidden facts in almost all medical problems. These facts would really in turn, help the practitioners to make accurate predictions [1]. The novel techniques and concepts have also been contributing themselves in yielding highest prediction accuracy over medical data.

The availability of clinical datasets and knowledge mining methodologies encourages the researchers to pursue research in extracting knowledge from clinical datasets. Different data mining techniques have been used for mining rules, and mathematical models have been developed to assist the clinician in decision making. In this information era the advancement of computerized database system facilitates the enhancement of decision making and diagnosis in medical science. Analysis of clinical datasets by using data mining methodology, techniques, and tools helps to develop a knowledge based system that can assist clinicians in decision making [2]. Clinical dataset consists of information about the current condition of the patients [3]. The information includes data of patient profile, physical check-up, and laboratory results. Mining knowledge from clinical dataset refers to the discovery of hidden valuable knowledge, to develop clinical expert system

In medical classification process, feature extraction is a method of dimension reduction. It is known to be an effective way in both reducing computational complexity and increasing accuracy. Besides feature selection, feature extraction is also an effective way to reduce the dimension of the data. Then, only the significant components are retained for classification [4]. Different types of dimensionality reduction techniques, including unsupervised approaches such as principal component analysis (PCA) and independent component analysis (ICA), as well as supervised approaches such as linear discriminant analysis, have been used. Among these methods, the PCA method can ensure that most information of the medical dataset can be preserved in a small amount of significant Principal Components (PCs) [5]

In current era, many feature selection approaches are utilized on medical database to extract most relevant data. These selection techniques are performed on medical data for various disease prediction. In order to reduce the computational complexity, most of the feature selection methods employ evolutionary and heuristic approaches. Such methods can handle high dimension optimization problems with acceptable results and time [6]. Swarm based algorithms are considered to be the most accepted nature inspired Meta heuristic techniques. SI (swarm intelligence) is one of the AI based method that provides collective behaviours for self-organized and decentralized systems. It contains a population of simple actors communicating locally and only within their surrounding

Data classification is one of the most important machine learning techniques. Classification plays an important role in organizing the data [7]. Classification assigns items in a collection to target categories or classes and the goal of classification is the accurate prediction of target class for each class of data. A prediction related to disease classification in earlier stages could help in treating disease patients

The objective of this research is to build a classifier that will predict the presence or absence of a disease by learning from the minimal set of attributes that has been extracted from the clinical dataset. The main aim of this research work is the medical data classification using HRFBP algorithm. There is numerous research and methodologies introduced but the medical data classifier accuracy is not ensured significantly. The existing approaches has drawback with error rates and inaccurate classification results. To overcome the abovementioned issues, in this research, HRFBP algorithm is proposed to improve the overall classification performance. The main contribution of this research is preprocessing, feature extraction, feature selection and disease classification. The proposed method is used to provide more accurate results using effective algorithms for the given benchmark datasets

The rest of the paper is organized as follows: a brief review of some of the literature works in preprocessing, feature extraction, feature selection, and classification methods on diseases are presented in Section 2. The proposed methodology for HRFBP scheme is detailed in Section 3. The experimental results and performance analysis discussion is provided in Section 4. Finally, the conclusions are summed up in Section 5

2. Related work

In [8], Chandra et al (2021) discussed the concept of BPNN which is used for transmitting entire error back to lessen the loss is termed as BPNN. It is considered BPNN for classification as it is flexible, less complex, and performs better with noise-free data. The experimental analysis has been carried out by gathering dataset from UCI storehouse. Popular datasets like cancer, diabetes, heart, and liver are chosen for study. The classifier efficiency has been shown by observing its lower RMSE value and better accuracy with other factors also. By developing a BPNN-based classifier system, it may ascertain physicians to deal with health-related problems

In [9], Leema et al (2016) uses Artificial Neural Network (ANN) trained by drawing in the relative advantages of Differential Evolution (DE), Particle Swarm Optimization (PSO) and gradient descent based backpropagation (BP) for classifying clinical datasets is used. The DE algorithm with a modified best mutation operation is used to enhance the search exploration of PSO. The ANN is trained using PSO and the global best value obtained is used as a seed by the BP. Local search is performed using BP, in which the weights of the Neural Network (NN) are adjusted to obtain an optimal set of NN weights. Three benchmark clinical datasets namely, Pima Indian Diabetes, Wisconsin Breast Cancer and Cleveland Heart Disease, obtained from the University of California Irvine (UCI) machine learning repository have been used. The performance of the trained neural network classifier used in this work is compared with the existing gradient descent backpropagation, differential evolution with backpropagation and particle swarm optimization with gradient descent backpropagation algorithms

In [10], Shah et al (2017) introduced Probabilistic Principal Component Analysis (PPCA) which has reputation to deal with the problem of missing values of attributes. This research presents a methodology which uses the results of medical tests as input, extracts a reduced dimensional feature subset and provides diagnosis of heart disease. The methodology extracts high impact features in new projection by using PPCA. It extracts projection vectors which contribute in highest covariance and these projection vectors are used to reduce feature dimension. The selection of projection vectors is done through Parallel Analysis (PA). The feature subset with the reduced dimension is provided to Radial Basis Function (RBF) kernel based Support Vector Machines (SVM). The RBF based SVM serves the purpose of classification into two categories i.e., Heart Patient (HP) and Normal Subject (NS). The methodology is evaluated through accuracy, specificity and sensitivity over the three datasets of UCI i.e., Cleveland, Switzerland and Hungarian. The statistical results achieved through the technique are presented in comparison to the existing research showing its impact.

In [11], Hassanien et al (2018) introduced the classification of ElectroCardioGram (ECG) heartbeats into normal or abnormal. The approach is based on the combination of swarm optimization algorithms with a modified PannTompkins algorithm (MPTA) and SVMs. The MPTA was implemented to remove ECG noise, followed by the application of the Extended Features Extraction Algorithm (EFEA) for ECG feature extraction. Then, elephant herding optimization (EHO) was used to find a subset of ECG features from a larger feature pool that provided better classification performance than that achieved using the whole set. Finally, SVMs were used for classification. The results show that the EHOSVM approach achieved good classification results in terms of five statistical indices: accuracy, 93.31%; sensitivity, 45.49%; precision, 46.45%; F-measure, 45.48%; and specificity, 45.48%. Furthermore, the results demonstrate a clear improvement in accuracy compared to that of other methods when applied to the MITBIH arrhythmia database

In [12], Dash et al (2019) suggested a novel hybrid swarm intelligence-based meta-search algorithm is used by combining a heuristic method called conditional mutual information maximization with chaos-based firefly algorithm. The combined algorithm is computed in an iterative manner to boost the sharing of information between fireflies, enhancing the search efficiency of chaos-based firefly algorithm and reduces the computational complexities of feature selection. The meta-search model is implemented using a well-established classifier, such as support vector machine as the modeller in a wrapper approach. The chaos-based firefly algorithm increases the global search mobility of fireflies. The efficiency of the model is studied over high-dimensional disease datasets and compared with standard firefly algorithm, particle swarm optimization, and genetic algorithm in the same experimental environment to establish its superiority of feature selection over selected counterparts

In [13], Alam et al (2019) discussed that medical data classification is considered to be a challenging task in the field of medical informatics. Although many works have been reported in the literature, there is still scope for improvement. In this paper, a feature ranking based approach is developed and implemented for medical data classification. The features of a dataset are ranked using some suitable ranker algorithms, and subsequently the Random Forest classifier is applied only on highly ranked features to construct the predictor. We have conducted extensive experiments on 10 benchmark datasets and the results are promising. We present highly accurate predictors for 10 different diseases, as well as suggest a methodology that is sufficiently general and is expected to perform well for other diseases with similar datasets

3. Proposed methodology

In this work, HRFBP algorithm is proposed to improve the medical dataset classification performance. The proposed system has four main steps such as preprocessing, feature extraction, feature selection and classification. The overall block diagram of the proposed HRFBP system is shown in Fig 1

3.1. Input dataset collection

In this work, data sets are taken from the UCI machine repository. The datasets are such as Pima Indians Diabetes, Heart-Statlog, Hepatitis, and Fertility data sets.

Hepatitis data symptoms and results in various people of both the genders of various age groups. Hepatitis symptoms include fatigue, anorexia, big liver, etc.

Fertility dataset collected from 100 volunteers provide a semen sample analyzed according to the WHO 2010 criteria. Sperm concentration are related to socio-demographic data, environmental factors, health status, and life habits in UC Irvine machine learning repository which consists of 100 instances and 10 attributes. The fertility attributes are such as season, age, childish disease, accident or serious trauma, surgical intervention, high fevers in last year, frequency of alcohol consumption, smoking habit, number of hours spent sitting per day and diagnosis.

The Pima datasets consist of several medical predictor (independent) variables and one target (dependent) variable. The variables include the number of pregnancies the patient, their BMI, insulin level, age, glucose, blood pressure, skin thickness, diabetes pedigree function and outcome.

Heart Statlog data contains attributes such as age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang and oldpeak. Age implies patient age in year, sex implies gender, cp is chest pain, trestbp is resting blood pressure, chol is serum cholesterol, fbs is fasting blood sugar, restecg is resting electrocardiographic result, mhrt is maximum heart rate, exang is exercise included angina and opk is old peak level, slope is slope of peak exercise, thalach is defect type, vessel is number of major vessels and class is heart disease.

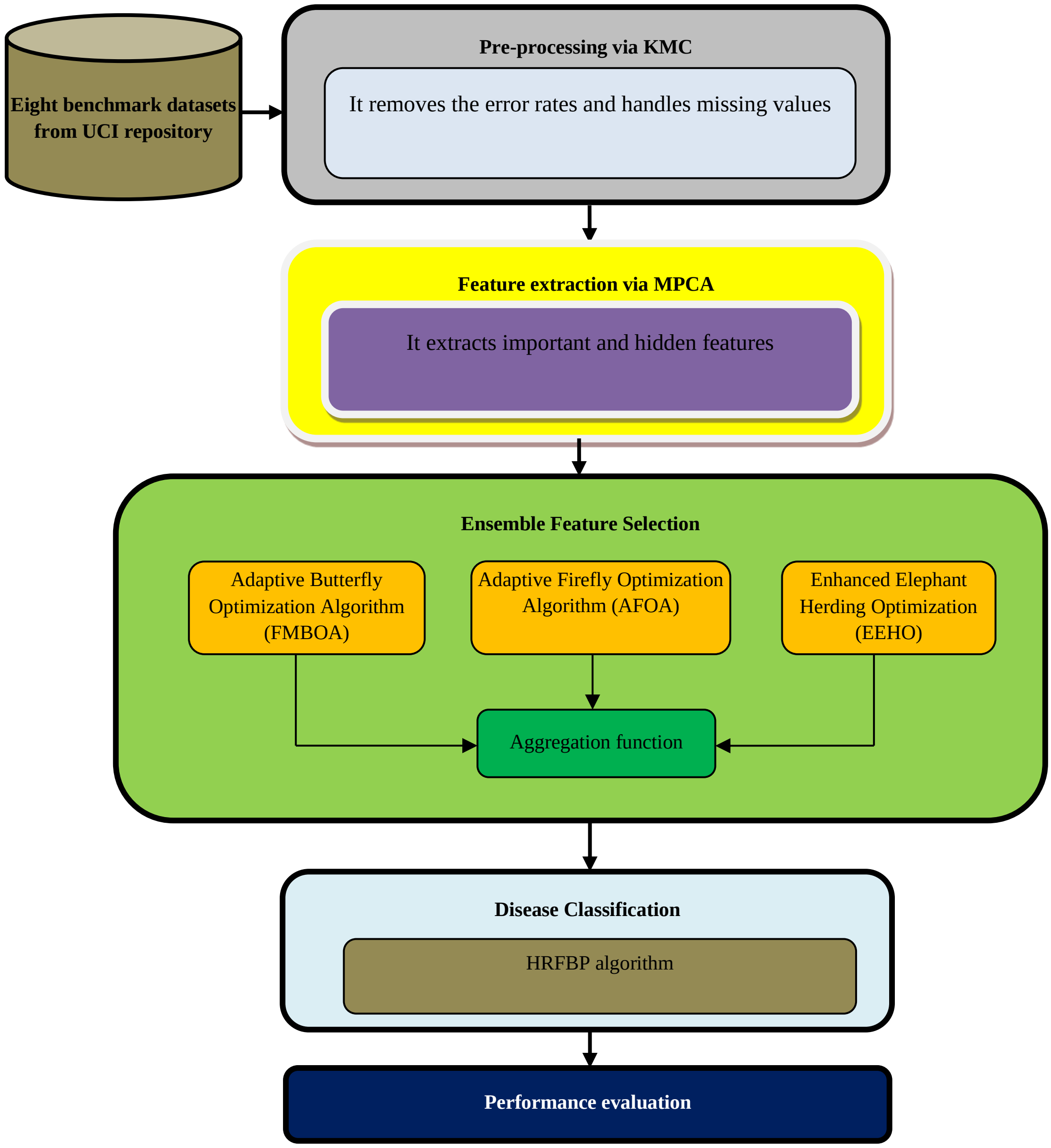
Eight benchmark datasets from UCI repository
Pre-processing via KMC
It removes the error rates and handles missing values
Feature extraction via MPCA
It extracts important and hidden features
Ensemble Feature Selection
Adaptive Butterfly Optimization Algorithm (FMBOA)
Adaptive Firefly Optimization Algorithm (AFOA)
Enhanced Elephant Herding Optimization (EEHO)
Aggregation function
Disease Classification
HRFBP algorithm
Performance evaluation

Fig 1 Overall block diagram of the proposed HRFBP algorithm

3.2 Pre-processing using K-Means Clustering (KMC) algorithm

In this work, pre-processing is done by using K-Means Clustering (KMC) algorithm which is used to increase the disease detection accuracy for the given dataset. The structured and unstructured datasets are utilized in this work. Unstructured datasets are Hepatitis data and fertility data. Structured datasets are Pima and Heart Statlog data.

KMC is an effective clustering technique used to separate similar data into groups based on initial centroids of clusters [14]. It uses the concept of Euclidean distance to calculate the centroids of the clusters. Starting from a random partitioning, the algorithm repeatedly (i) computes the current cluster centers (i.e. the average vector of each cluster in data space) and (ii) reassigns each data item to the cluster whose centre is closest to it. It terminates when no more reassignments take place. By this means, the intra-cluster variance, that is, the sum of squares of the differences between data features and their associated cluster centers is locally minimized. Fig 2 shows the example of KMC algorithm.

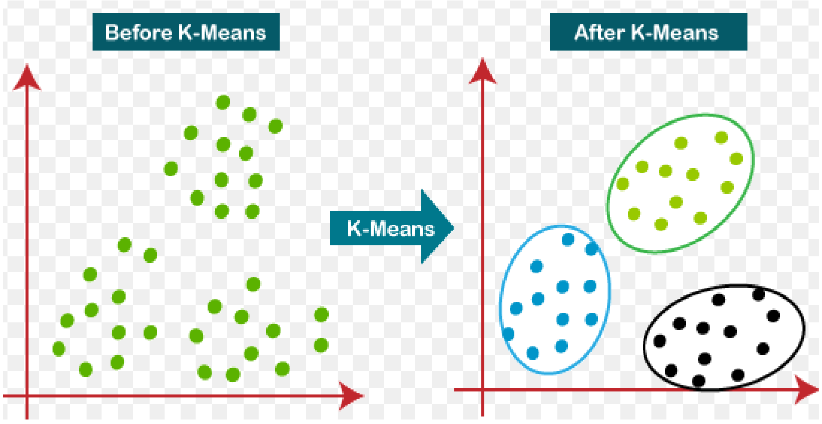

Fig 2 Example of KMC algorithm

K-means' strength is its runtime, which is linear in the number of data elements and its ease of implementation. In this work, the cluster number is kept equal to the number of classes. Calculate the Euclidean distance for finding the centroids of the clusters using the given below formula

$$d(i,j) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (1)$$

Where x_i and y_i are two points in Euclidean n-space

Algorithm 1: KMC algorithm

1. Choose a number of clusters k from structured and unstructured dataset (D)

2. Initialize cluster centers µ1,... µk

3. Pick k data points and set cluster centers to these points

4. Randomly assign points to clusters and take means of clusters

5. For each data point, compute the cluster center it is closest to and calculate the distance measure for finding the missing values using (1)

6. Assign the data point to this cluster

7. Re-compute cluster centers (mean of data points in cluster)

8. Find and remove error value and missing values

9. Stop when there are no new re-assignments

From the original dataset the instances having missing attributes are separated from the dataset. The dataset is divided into two sets where one set contains complete instances that do not contain any missing values and the other set contains incomplete instances which contains missing values. KMC is applied on complete instances set to obtain clusters of complete instances. Hence, the instances are taken one by one and the missing attributes are filled with their possible values. KMC is applied to the dataset from the resultant clusters, the newly added instance is validated that whether it is been clustered in the correct class or not. If it is in the correct cluster, then the assigned value is made as permanent then the procedure is continued with the next instance. If it is in the wrong cluster then the next possible value will be assigned and compared till it founds the value which put the instance in the correct cluster. Thus the preprocessing method is used to improve the disease classification accuracy effectively by using KMC algorithm

3.3 Feature extraction using Modified Principal Component Analysis (MPCA)

The PCA algorithm is used to obtain the hidden information and classify the data to identify medical diseases. The proposed system is molded to estimate the condition normal or not. It extracts based on the features such as problems with characteristics of Pima Indians Diabetes, Heart-Statlog, Hepatitis, and Fertility data sets.

The purpose of PCA is to reduce the large dimensionality of the data space (observed variables) to the smaller intrinsic dimensionality of feature space (independent variables), which are needed to describe the data economically. This is the case when there is a strong correlation between observed variables. By discarding minor components, the PCA effectively reduces the number of features and displays the data set in a low dimensional subspace [15]. PCA is a classical multivariate data analysis method that is useful in linear feature extraction. In this study the feature extraction algorithm based on PCA is chosen. The coefficients of these methods are used as feature vectors which efficiently represent medical dataset. The normal PCA algorithm is used for extracting features from small dataset and will miss important feature information. The PCA method cannot guarantee that the information related to the relevant classes is effectively compressed. To avoid the above mentioned issues, modified PCA is enhanced.

In the MPCA approach, reduce the influence of the eigenvectors corresponding to the large eigenvalues by normalizing the jth element y_{ij}, of the ith feature vector y, with respect to its standard deviation,$\sqrt{\lambda_j}$. Hence, the new feature vector y_i' rewritten as

$$y_i' = [\frac{y_{i0}}{\lambda_0}, \frac{y_{i1}}{\lambda_1}, \dots . \frac{y_{i(r-1)}}{\lambda_{r-1}}] \qquad (2)$$

These normalized feature vectors are used to construct a new feature subspace. In this approach, it first normalizes the feature vectors by the square root of the corresponding eigenvalues, and then calculates the distance between the training and the testing features

In general, the linearly transform (PCA) can be expressed as following equation:

$$Y = TX \qquad (3)$$

where T is the transform matrix, X is the origional vectors and Y is the transformed vectors. In order to solve the transform matrix T, the following equation:

$$(\lambda I - S)U = 0 \qquad (4)$$

is used, where the matrices I , S, U and λ are the square matrix with unity along its diagonal, the covariance matrix of original images, the eigenvectors and the eigenvalues. U_j and $\lambda_j (j = 1,2, \dots m)$can be computed through the equation (2), with the eigenvalues ordered as $\lambda_1 \geq \lambda_2 \geq \cdots . \geq \lambda_m$. The eigenvectors U can be expressed as $U = [U_1, U_2, \dots \dots, U_m]$.

In the MPCA, training samples, which are relevant for a given application, are selected from a medical dataset, and the transformed matrix T' was obtained form these training samples. It can be expressed as the following equation:

$$Y = T'X \qquad (5)$$

$$V_N = b_1 u_1 + b_2 u_2 + \cdots . + b_N u_N \qquad (6)$$

$$S = \sum_{i=0}^{1} b_1 u_1 ; 1 < N \qquad (7)$$

Comparing the two equations (6) and (7), the difference lies in the transform matrix, and essentially lies in the samples for calculating the covariance matrix, one is from training samples, the other is from the whole hate speech dataset.

The main advantage of MPCA is dimension will be reduced by avoiding redundant information, without much loss. Better understanding of principal component analysis is through statistics and some of the mathematical techniques which are Eigen values, Eigen vectors. MPCA is a mathematical procedure that uses linear Transformations to map data from high dimensional space to low dimensional space. The low dimensional space can be determined by Eigen vectors of the covariance matrix. In this work, it is used to extract useful important acoustic features for medical dataset classification due to its error minimizing and de-correlating properties. The acquired input data of normal and disease have features like mean and standard deviation

(i) Mean= sum of no of data /total number of data (8)

(ii) Standard deviation: Standard Deviation is also known as root-mean square deviation as it is the square root of means of squared deviation from the arithmetic mean. $\sigma = \sqrt{}(\sum(x - x)/n)$ (9)

Algorithm 2: MPCA

```
1.  Start
2.  Find the mean value $S'$ of the given medical dataset S
3.  Subtract the mean value from S
4.  Obtain the new matrix A
5.  Covariance is obtained from the matrix i.e., $C = AA^T$ Eigen values are obtained from the covariance matrixes that are $V_1 V_2 V_3 V_4 \dots V_N$
6.  Finally Eigen vectors are calculated for covariance matrix $C$
7.  Any vector S can be written as linear combination of Eigen vectors using (6)
8.  Only Largest eigen values are kept to form lower dimension data set
9.  Match the combination of features in the given datasets (7)
10. Compute the feature using mean and Standard deviation (8) & (9)
11. Extract the more informative features (acoustic features)
12. End
```

The MPCA algorithm is used to extract the maximum and minimum occurring synchrostates and obtained the associated brain network parameters which are then fed into the discriminant

3.4 Ensemble Feature Selection (EFS)

Ensemble Feature Selection (EFS) techniques are defined to produce an optimum subset of features through combining multiple FS based on Enhanced Elephant Herding Optimization (EEHO), Entropy Butterfly Optimization Algorithm (EBFO) and Adaptive Firefly Optimization Algorithm (AFOA) is the intuition behind the ensemble learning. The common design of EFS is to aggregate the decisions of FS methods to develop the representation capability [16]. EFS techniques contain two main phases such as generation of diverse feature selectors and aggregation of the decisions.

3.4.1 Entropy Butterfly Optimization Algorithm (EBFO)

In this work, feature selection is done by using Entropy Butterfly Optimization Algorithm (EBFO) to select the optimal features from the medical dataset. EBFO is new nature-inspired algorithm that mimics food search (higher accuracy with selected features) and mating behavior of butterflies, to solve classification issues in medical disease diagnosis. The proposed EBFO Algorithm is mainly based on the foraging strategy of butterflies, which utilize their sense of smell for optimal selection of features in order to determine the location of nectar partner [16]. Based on scientific observations, it is found that butterflies have a very accurate sense of locating the source of fragrance (classification accuracy).

A butterfly will generate fragrance with some intensity which is correlated with its fitness (Classification accuracy), i.e., as a butterfly moves from one location to another, its fitness will vary accordingly. In EBFO Algorithm, whole concept of sensing and processing the modality is based on three important terms viz. sensory modality (c), stimulus intensity (I) and power exponent (a) for optimal selection of features [17]. In EBFO Algorithm, I is correlated with the fitness (accuracy) for the selection of features from medical dataset. Using these concepts, in EBFO Algorithm, the fragrance is formulated as a function of the physical intensity of stimulus as follows by equation (9),

$$f = cI^a \tag{10}$$

Where f is the perceived magnitude of the fragrance, i.e., how stronger the fragrance is perceived by other butterflies, c is the sensory modality which is generated via classification accuracy, In is the stimulus intensity and a is the power exponent dependent on modality. Thus a & c in the range $[0, 1]$. On the other hand, if $a = 0$, it means that the fragrance emitted by any butterfly cannot be sensed by the other butterflies at all. So, the parameter a

controls the behavior of the algorithm. Another important parameter is c which is also crucial parameter to find the speed of convergence and how the EBFO algorithm. To demonstrate above discussions in terms of a search algorithm, the above characteristics of butterflies are idealized as follows:

1. All butterflies are supposed to emit some fragrance which enables the butterflies (features) to attract each other (features).
2. Every butterfly will move randomly or toward the best butterfly emitting more fragrance.
3. The stimulus intensity of a butterfly is affected or determined by the landscape of the objective function.

There are three phases in EBFO such as (1) Initialization phase, (2) Iteration phase and (3) Final phase. In each run of EBFO, first the initialization phase is executed, then searching of optimal features are performed in an iterative manner and in the last phase, the algorithm is terminated finally when the best optimal selection solution is found. In the initialization phase, classification accuracy is computed in EBFO algorithm and its solution space. The values for the parameters used in EBFO are also assigned. The positions of butterflies (features) are randomly generated in the feature selection search space, with their fragrance and fitness values. After finishing initialization phase then algorithm starts the iteration phase. In each iteration, all butterflies in feature selection solution space move to new positions and then their classification accuracy values are evaluated. The algorithm, first fitness values are computed of all the butterflies on different positions in the solution space. Then these butterflies will generate fragrance at their positions using equation (10). In global search phase, the butterfly takes a step toward the fittest solution (g∗)(optimal features) which can be represented using equation (11),

$$x_i^{t+1} = x_i^t + (r^2 \times g^* - x_i^t) \times f_i * ECE_W \tag{11}$$

where x_i^t is the solution vector x_i for ith butterfly in iteration number t. Here, g^* represents the current best selected feature solution found among all the solutions in current iteration. Fragrance of ith butterfly is represented by f_i and $r \in [0,1]$ is a random number Local search phase can be represented by equation (12),

$$x_i^{t+1} = x_i^t + \left(r^2 \times x_j^t - x_k^t\right) \times f_i * ECE_W \tag{12}$$

where x_j^t and x_k^t are jth and kth butterflies from the feature selection solution space. If x_j^t and x_k^t belongs to the same swarm and $r \in [0,1]$ is a random number then equation (12) becomes a local random walk. Search for food and mating partner by butterflies can occur at both local and global scale for optimal selection of features from the dataset. Switch probability p is used in EBFO to switch between common global search to intensive local search. Till the stopping criteria are not matched, the iteration phase is continued. When the iteration phase is concluded, the algorithm outputs the best solution found with its best fitness. In the equation (11, 12), feature weight is also added to EBFO algorithm to select optimal number of features in the medical dataset. The EBFO algorithm focused to improve the accuracy of the classifier using optimal selection of features over the given medical dataset. Cross Entropy (CE) is used to measure the distance between two sampling distributions, solve an optimization problem by minimizing this distance, and obtain the optimal parameters of probability distribution. The CE method has good global search capability, excellent adaptability, and strong robustness.

$$CE = \frac{1}{N}\sum_{i=1}^{N} I_{s<r}\frac{f(x^i,v)}{g(x^i)} \tag{13}$$

where x^i represents a random sample from $f(x;v)$ with importance sampling density $g(x)$. The Kullback–Leibler divergence, i.e., the cross-entropy is introduced to measure the distance between two sampling distributions for obtaining the optimal importance sampling density

The overall steps involved in the proposed EBFO algorithm are shown in the algorithm 3. In the algorithm 3, initial population are generated via number of features in the medical dataset (Step 1), and then stimulus intensity I_i at x_i (Step 2) is computed based on the sensor modality c, power exponent a (Step 3). These factors are generated via the classification accuracy. Then it starts with stopping criteria (Step 4), for each butterfly in the dataset the fragrance value is computed (Step 6). After that find the best feature in the population(Step 8) and generated random number r (Step 10) . If $r < p$ then move towards the best butterfly by equation (11), else move randomly by equation (12). Then update a value (Step 17), and evaluate individuals according to their new position (Step 18). Finally end the process via end while (Step 19). The flowchart of the proposed Entropy Butterfly Optimization Algorithm (EBFO) algorithm is illustrated in Fig 3.

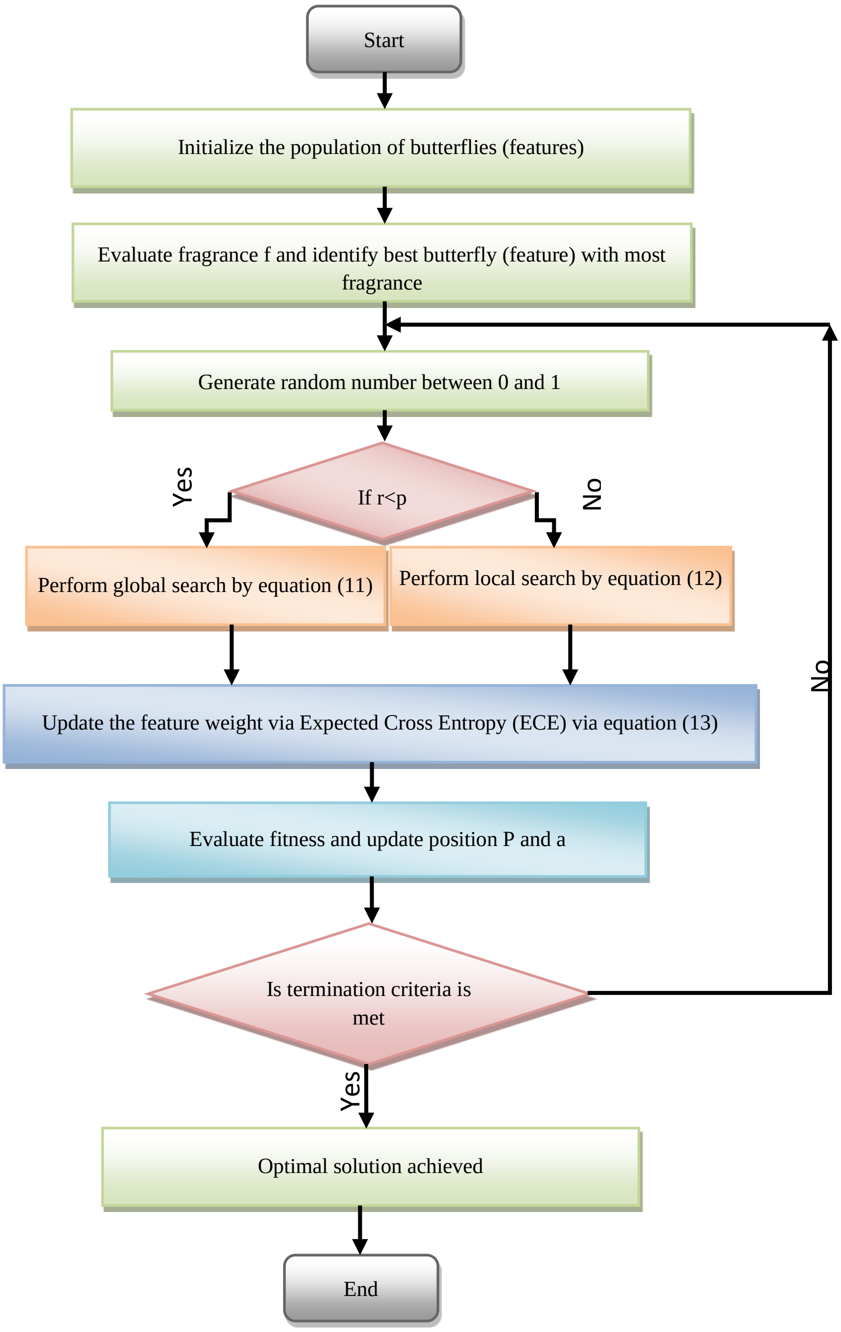
Start
Initialize the population of butterflies (features)
Evaluate fragrance f and identify best butterfly (feature) with most fragrance
Generate random number between 0 and 1
Yes
If r<p
No
Perform global search by equation (11)
Perform local search by equation (12)
No
Update the feature weight via Expected Cross Entropy (ECE) via equation (13)
Evaluate fitness and update position P and a
Is termination criteria is met
Yes
Optimal solution achieved
End

Fig 3 flowchart of Entropy Butterfly Optimization Algorithm (EBFO)

Algorithm 3: Entropy Butterfly Optimization Algorithm (EBFO)

Input: Medical datasets (Pima Indians Diabetes, Heart-Statlog, Hepatitis, and Fertility data sets)

Objective function: Classifier accuracy, $f(x), x = (x_1, x_2, \dots., x_{dim})\ dim = no. of\ dimesnions$

Output: Selection of optimal features

```
Generate initial population of $n$ butterflies $x_i = (i = 1,2,\dots,n)$ via number of features in the dataset
Stimulus Intensity $I_i$ at $x_i$ is found by classification accuracy $f(x_i)$
Define sensor modality $c$, power exponent $a$ and switch probability $p$
While stopping criteria not met do
    For each butterfly $f$ in population do
        Calculate fragrance for $f$ using equation (11) and generate weight via entropy by equation (13)
    End for
    Find the best butterfly
    For each butterfly $f$ in population do
        Generate random number r
        If $r < p$ then
            Move towards the best butterfly (optimal features) by equation (11) and generate weight via entropy by equation (13)
        Else
            Move randomly using the equation (12)
        End if
    End for
    Update the value of a
    Evaluate individuals(features) according to their new position
End while
Output the best solution found
```

3.4.2 Adaptive Firefly Optimization Algorithm (AFOA) algorithm

In this work, feature selection is performed via AFOA algorithm over given datasets. Firefly algorithm is inspired by biochemical and social aspects of real fireflies. Real fireflies produce a short and rhythmic flash that helps them in attracting (communicating) their mating partners and also serves as protective warning mechanism. Firefly Algorithm (FA) formulates this flashing behavior with the objective function of the problem to be optimized. FA works on the principle of the firefly's flashing lights. The intensity of the light supports a firefly group shift to intense and attractive positions which is plotted to produce best resolution over the seeking place.

This mechanism normalizes few of the firefly features and can be shown as given below [18]:

(i) Every firefly is attracted to a different irrespective of their sex.

(ii) The brightness formed via the firefly is openly comparative to its attractiveness and it is among two fireflies, the firefly along with higher brightness attracts the one which has lesser brightness. A firefly shifts arbitrarily if it is not capable to discover a brighter nearest firefly.

In the statistical form, firefly's brightness depends on the objective function. Fig 4 shows the basic mechanism of the firefly algorithm

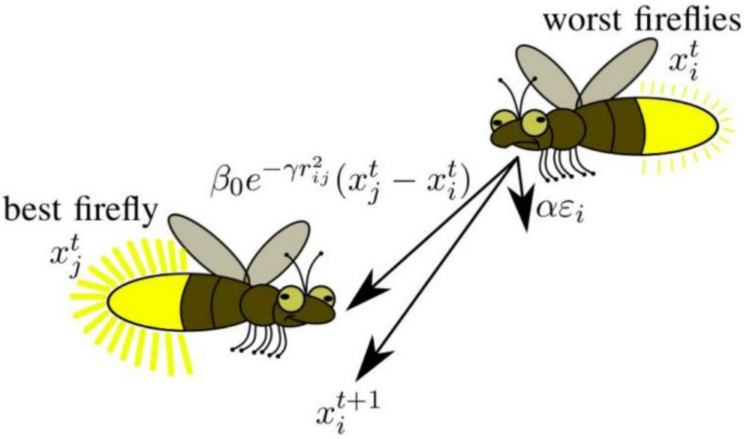

Fig 4 Basic mechanism of the firefly algorithm

Firefly algorithm is selected for its potential of giving best solutions for multi objective problems. For a maximization problem, the brightness can simply be proportional to the objective function. For simplicity, it is assumed that, the attractiveness of a firefly is determined by its brightness or light intensity which in turn is associated with the encoded objective function. These phases are repeated iteratively until the convergence criteria are satisfied

a) Attractiveness and Light Intensity at the source: The light intensity changes based on the inverse square law as follows

$$I(r) = \frac{I_0}{r^2} \quad (14)$$

Where I(r) is the light intensity at attractiveness r^2

Attractive generated by random assignment of features

b) While the intermediate is given, the light intensity is as follows:

$$I(r) = I_0 \exp(-\gamma r) \quad (15)$$

Where I_0 is the absorption coefficient of the medium

c) To avoid the singularity, the following Gaussian form of the approximation is considered

$$I(r) = I_0 \exp(-\gamma r^2) \quad (16)$$

The attractiveness of a firefly is proportional to the light intensity seen by the adjacent fireflies. A new solution is obtained by considering the possibility of variation and changing the pixels randomly. Hence the attractiveness β of a firefly is as follows

$$\beta = \beta_0 \exp(-\gamma r^m) \quad (17)$$

Where β_0 is the attractiveness at r=0.

Distance between any two fireflies (features) i and j are positioned and the distance is computed as follows

$$r_{i,j} = \sqrt{\sum_{k=1}^{d}(x_{i,k} - x_{j,k})^2} \quad (18)$$

Where $x_{i,k}$ is the kth factor of the spatial match x_i of the kth firefly and d is the amount of dimensions. Adaptation parameter is introduced for both absorption and random parameter, resulting in an AFOA. These changes improve the global search and local search capability by changing the parameter linearly during the iterations period [19]. AFOA is applied for producing the optimal features for the given datasets via selecting the best features in terms of higher fitness values.

Calculate the parameter α as follow:

$$\alpha(t+1) = \left(1 - \frac{t}{MaxG}\right)\alpha(t) \quad (19)$$

α adapts the value with the distance deviation degree of the optimization, so as to improve the solution precision and convergence speed. At the same time, in order to enhance the flexibility of population it is rewritten as follow

$$\alpha = \alpha_{min} + (\alpha_{max} - \alpha_{min}) \times ||x_i - x_{best}||/L_{max} \quad (20)$$

Where $L_{max} = (x_{worst} - x_{best})$ (21)

$\alpha_{max} and\ \alpha_{min}$are the maximum and minimum features respectively. In Eq. (21), x_{worst} represents the position of the worst individual at the generation t firefly, and L_{max} is the distance between the worst individual and the global optimal individual x_{best}. In the early stage of the algorithm, the firefly individuals are scattered in the whole space, and most of them have a large distance from the globally optimal individuals. At this point, the value of $\|x_i - x_{best}\|$ is larger, $L_{max}$and $(\alpha_{max} - \alpha_{min})$ are fixed values. Therefore, Eq. (20) showed that the value of α is larger in the early stage, which has better global optimization effect. As the algorithm implementation, fireflies individual i is attracted by brighter fireflies than itself, and close to global optimal features. In the later period, fireflies individuals i will gather around the global optimal individuals, the value of $\|x_i - x_{best}\ \|$ is smaller at this point, which is propitious to improve the searching optimal features over the given datasets. In each iteration, the value of α is changed with the position of the optimum, which improves the convergence speed of the algorithm. According to the above analysis, the step size factor α changes adaptively and dynamically according to the distance between the individuals of fireflies, which balances the ability of algorithm development and search.

In this research, a new fitness function assumes accuracy and execution time which is given by

$$f(x) = \frac{\left({}^{I_d}/_{I_t}\right) \times \left({}^{I_f}/_{P_{init}{}^{i}}\right)}{exp^{-{}^{e_E}/_{e/_M} + H_{accuracy}}} \quad (22)$$

where I_d is the number of dropped features. m_t is the total number of features sent with higher accuracy

I_f is the features in the dataset i.

$P_{init}{}^{i}$ is the initial feature.

e_E is the execution time and e_M is the maximum allowable delay.

$$x_i = x_i + \beta_0 e^{-\gamma r^2}(x_j - x_i) + \alpha(rand - \tfrac{1}{2}) \quad (23)$$

Where x_i and x_j is distance between two firefly features

The fitness value of each feature is calculated in the population. In the first generation, the number of features in a batch is randomly selected. The fitness value of each firefly is calculated. Then, the selection procedure is used to decide on two selected fireflies. Firefly along with higher brightness has the highest fitness value which is selected for the next generation (feature selection)

Algorithm 4: AFOA for feature selection

Input data: Pima Indians Diabetes, Heart-Statlog, Hepatitis, and Fertility data sets

Output: Optimal features

```
Objective function $(x)$, $x = (x1,...,)T$ consider higher classifier accuracy as objective function
Produce initial population of fireflies $xi$ $(i = 1, 2, . . . , n)$
Light intensity $Ii$ at $xi$ is found via $f(xi)$
Describe light absorption coefficient $\gamma$
while ($t$ < MaxGeneration)
    for $i$=1:$n$ all $n$ fireflies (features)
        for $j$=1:$i$ all $n$ fireflies (features)
            if ($Ij$ > $Ii$), Move firefly $i$ towards $j$ in $d$-dimension;
            end if
            Attractiveness changes along with distance $r$ via exp$[-\gamma r]$
            Compute fitness function using (22) and (23)
            Compute objective model using (18)
            Estimate new solutions and update light intensity
            Reduce the redundant features
            Update the optimal features using (20)
        end for $j$
    end for $i$
    Rank the fireflies and find the current best features
end while
A firefly $i$ shifts to a more attractive
Return best features
```

Algorithm 4 explains that AFOA algorithm is used to produce optimal solutions based on the fitness which has higher classifier accuracy metric. In the AFOA algorithm, the fireflies are ranked and the optimal fireflies are chosen by best fitness values. The novel best solutions are appended to the firefly pool and the subsequent iteration of the firefly is continued. Hence, in this research, optimal feature selection is performed by using AFOA algorithm based on the higher accuracy features. Extracted features of test datasets are applied AFOA and these extracted features forms correlation matching with data features. If maximum brightness obtained then input test datasets diseases features otherwise for minimum brightness test input dataset is healthy feature

3.4.3 Enhanced Elephant Herding Optimization (EEHO)

Enhanced Elephant Herding Optimization (EEHO) algorithm is a heuristic intelligent algorithm based on the nomadic habits of elephants. Through the observation and study of the elephants, the elephant herd mainly has the following two characteristics for selection of features from medical disease diagnosis. The first characteristic is that there are multiple clans in an elephant herd, each of which has its own patriarch and members who follow the instructions of the patriarch for optimal selection of features from medical disease diagnosis. Another characteristic of the herd is that there is no adult male elephant. Young elephants will live alone from the elephants when they grow up. Inspired by these two characteristics, core idea of EEHO has two parts, one is clan updating, and the other is separating [20]. The first feature of the elephant herd can be abstracted as a clan updating operator, and its update process is as follows by equation (24),

$$x_{n,i,j} = x_{i,j} + r * a * \left(x_{b,i} - x_{i,j}\right) * ECE_W \tag{24}$$

where the old and new feature position from medical dataset of elephant j in clan i are $x_{i,j}$ and $x_{n,i,j}$, respectively; α∈[0, 1] is a scale factor; $x_{b,i}$ is the feature position with the best fitness value(classification accuracy) in clan i. r is a random number with a normal distribution in the range [0, 1]. Equation (2) represents the update process of most individuals (features), but the matriarch in each clan has not been updated [21]. In order to calculate the weight value of each feature for medical disease dataset in the EEHO, let us assume that when a certain feature value is observed, it gives a certain amount of data to the target feature. Finally this weight value is updated to EEHO algorithm. Based on the weight from Expected Cross Entropy (ECE), importance of features is selected. ECE is performed based on Kullback Leiber (KL) distance, and it reflects the distance between the probability of the topic class and the probability of the topic class under the condition of a specific feature. The computing equation (25) can be described as follows [22],

$$\text{Cross Entropy (CE)(f)} = P(f)\sum_{i=1}^{|C|} P(f|c_i)\log\frac{P(f|c_i)}{P(c_i)} \tag{25}$$

where f is the feature, $P(f)$ is the probability of data that contains appearing in the training set, $P(c_i)$ is the probability of class $c_i$in the training set, $P(f|c_i)$ is the probability of data that contains feature in class c_i, and $|C|$is the total amount of classes in training set. The formula of information entropy is given by equation (26),

$$\text{Information Entropy (IE)(f)} = -\textstyle\sum_{i=1}^{|C|} P(f|c_i)\log(P(f|c_i)) \tag{26}$$

In a summary, combine equation (25) and equation (18) together; the ECE equation is as follows by equation (27),

$$\text{ECE (f)} = \frac{P(f)}{IE(f)+\varepsilon}\sum_{i=1}^{|C|}(P(f|c_i)+\varepsilon)\log\frac{P(f|c_i)+\varepsilon}{P(c_i)} \tag{27}$$

If feature f exists only in one class, the value of information entropy is zero; that is $\text{IE(f)} = 0$. Hence should introduce a small parameter in the denominator as a regulator. Therefore, the update process of the matriarch for feature selection process for disease detection is shown in equation (28)–(29).

$$x_{n,i,j} = \beta * x_{c,i} \tag{28}$$

$$x_{c,i} = \frac{1}{n_i}\times\sum_{j=1}^{n_i} x_{i,j} \tag{29}$$

where β is a scale factor in the range [0, 1]. The centre position (feature position) in clan i is $x_{c,i}$ and it can be calculated by equation (20). The elephant number in clan i is n_i . In equation (30), the update of the matriarch position (feature position) is related to the information of all members (features) in the clan. The separating operator can be abstracted from the second feature of the elephant herd. The separation process is as follows in equation (30),

$$x_{w,i} = x_{min} + r * (x_{max} - x_{min}) \tag{30}$$

where $x_{w,i}$ is the position (feature position) with the worst fitness value (classification accuracy) in clan i; x_{max} and x_{min} are the upper and lower bound of the elephant's position(feature position), respectively; r is a random number with a normal distribution in the range [0, 1]. Algorithm 5 shows the working procedure of proposed EEHO algorithm. It starts with initialization of population via the number of features in the medical dataset and then evaluate the fitness value (classification accuracy), based on that eliminate worst features (medical disease features) in the clan, then start the procedure with t iteration to T_{max}. For each features two operations such as clan updating, and the other is separating is performed by step 6 to step 8. Once these operations are performed then remove the worst elephant from the clan via the step 11 to step 13. Finally find the best features in the step 14. Similarly the flowchart of the proposed system is illustrated in Fig 5

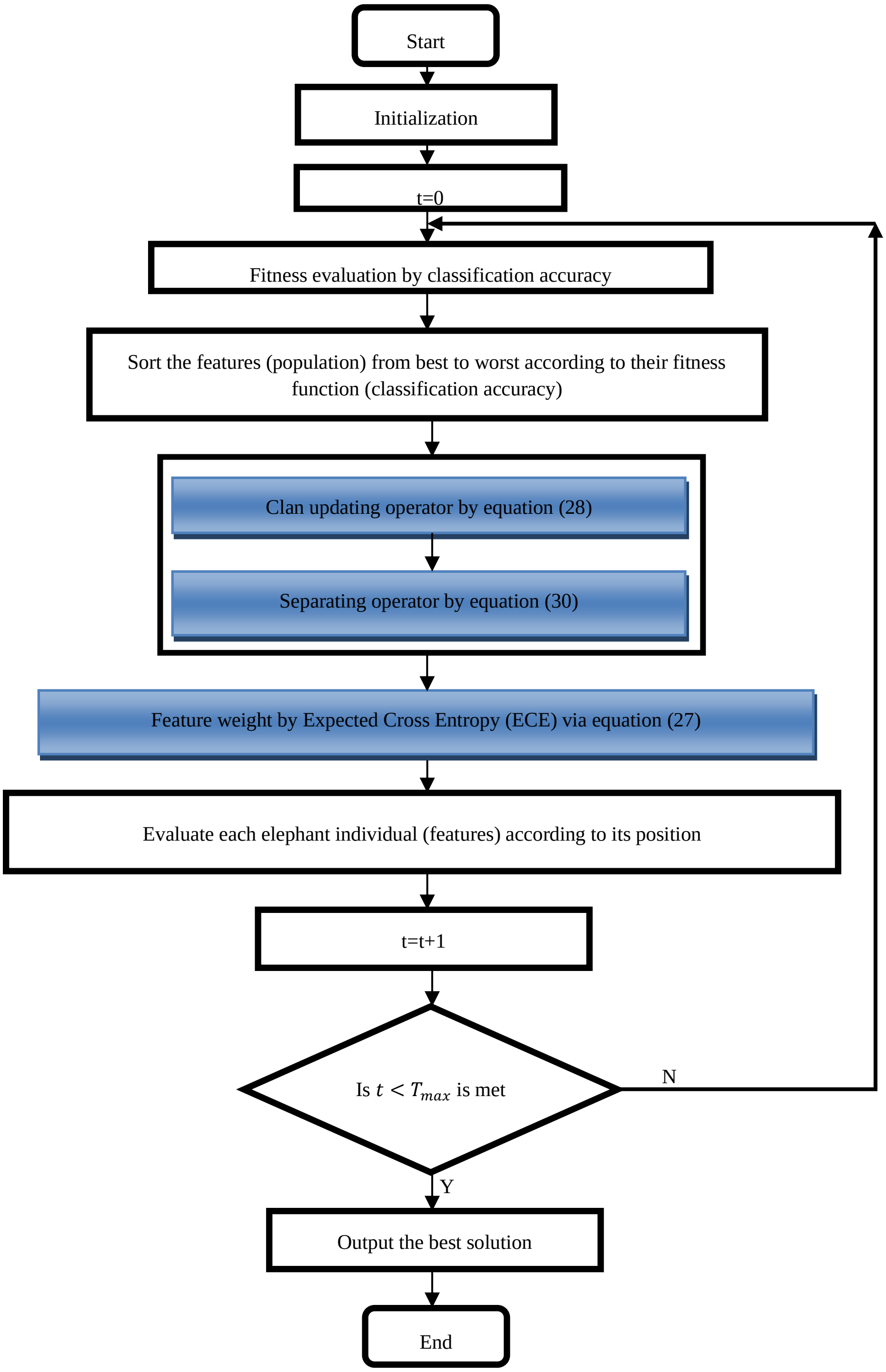
Start
Initialization
t=0
Fitness evaluation by classification accuracy
Sort the features (population) from best to worst according to their fitness function (classification accuracy)
Clan updating operator by equation (28)
Separating operator by equation (30)
Feature weight by Expected Cross Entropy (ECE) via equation (27)
Evaluate each elephant individual (features) according to its position
t=t+1
Is $t < T_{max}$ is met
N
Y
Output the best solution
End

Fig 5 flowchart of Enhanced Elephant Herding Optimization (EEHO) algorithm

Algorithm 5: Enhanced Elephant Herding Optimization (EEHO) algorithm

```
1. Initialization of number of populations via the number of features and parameters
2. Fitness evaluation via classification accuracy and their feature position
3. While $t < T_{max}$ do
```

```
For $i = 1\ to\ n_c$ do
    For j=1 to $n_j$ ( the number of elephants (Features) in one clan) do
        Update $x_{i,j}$ and generate $x_{n,i,j}$ based on the equation (24), generate feature weight by ECE (f) in equation (27)
        If $x_{i,j} = x_{b,i}$ then
            Update $x_{i,j}$ and generate $x_{n,i,j}$ based on the equations (28-29)
        End if
    End for
For $i = 1\ to\ n_c$ do
    Replace the worst elephant(features) in clan i by equation(30)
End for
Evaluate individuals(features) according to their new position
End while
```

3.4.4 Aggregation function

The diverse ranked results are combined into single ensemble list utilizing an appropriate aggregation function which allocates every feature a "overall score" depends on the attribute's place (rank) in the first records. Generally, consider $L_k$be the ranked list outcome from the function of a specified feature selection algorithm to the kth bootstrap test (k = 1,... ,B). For every one of the original features $f_i(i = 1, \dots N)$ a overall score is subsequently determined via $score_i = score\ (f_i) = aggr(r_{i1}, r_{i2}, \dots r_{iB})$. Here r_{ik} is the rank of the ith feature in the kth ranked list and $aggr$ is a reasonable aggregation function. Depends on their overall scores, the attributes are arranged, from the most significant to the most un-significant, in the final ensemble list [38]. The optimal feature are selected which increases the medical dataset classification accuracy prominently

3.5 Disease classification via Hybrid Random Forest with Back Propagation (HRFBP) algorithm

In this work, disease classification is done by using HRFBP algorithm. The neural network used in this work is a gradient descent backpropagation neural network with variable learning rates. BPNN consists of three layers: input layer, hidden layer, and output layer. The backpropagation algorithm only needs to update the weights of the features along this dataset. It can efficiently perform the updates for every single feature. Therefore, the training phase for a single neural network is expected to work extremely efficiently using hybrid RF classifier.

The RF algorithms form a family of classification methods that rely on the combination of several decision trees. The particularity of such ensembles of classifiers is that their tree based components are grown from a certain amount of randomness. Based on this idea, RF is defined as a generic principle of randomized ensembles of decision trees. The basic unit of RF (the so-called base learner) is a binary tree constructed using recursive partitioning.

The RF tree base learner is typically grown method in which binary splits recursively partition the tree into homogeneous or near homogeneous terminal nodes (the ends of the tree). A good binary split pushes data from a parent tree-node to its two daughter nodes so that the ensuing homogeneity in the daughter nodes is improved from the parent node. RF is often a collection of hundreds to thousands of trees, where each tree is grown using a bootstrap sample of the original data. In addition to the randomization introduced by growing the tree using a bootstrap sample of the original data, a second layer of randomization is introduced at the node level when growing the tree.

Rather than splitting a tree node using all variables, RF selects at each node of each tree, a random subset of variables, and only those variables are used as candidates to find the best split for the node. The purpose of this two-step randomization is to de-correlate trees so that the forest ensemble will have low variance, a bagging phenomenon. RF trees are typically grown deeply. Although it has been shown that large sample consistency requires terminal nodes with large sample sizes, empirically, it has been observed that purity or near purity is often more effective when the feature space is large or the sample size is small. This is because in such settings, deep trees grown without pruning generally yield lower bias. In such cases, deep trees promote low bias, while aggregation reduces variance. The construction of RF is hybrid with BPNN algorithm which is described in the following main steps

RF separates the chosen arbitrary subset from the root node to a child node constantly until every tree arrives at a leaf node without pruning [23]. Every tree builds the detection of the attributes and the target variable separately and decisions in favour of the last tree class. RF chooses the overall classification dependent on the majority voting method. The development of RF could described in the subsequent steps,

Algorithm 6: HRFBP

Step 1: The disease features selected by the ensemble feature selector are given as the input of the BPNN.

Step2: Generates N amount of bootstrap features from the specified dataset

Step3. Every node takes a random sample of attributes of size m where $m < M$. (M refers to the total number of attributes).

Step4: Builds a split via the m features chosen in Step 2 and computes the k node through the best split point. ("k" refer to next node). Train basic RF classifier

Step5: A repeat splitting tree until only one leaf node is attained and the tree is terminated

Step6: The algorithm is trained on every bootstrapped independently

Step 7: The input of the hidden layer and the output of the hidden layer are calculated using equations (31) and (32):

$$a_{net} = \sum w_{i,j} O_i + \emptyset_j \tag{31}$$

Where $w_{i,j}$ are the weights of each input nodes and $\emptyset_j$ is the bias

$$O_j = \frac{1}{1+e^{-a_{net,j}}} \tag{32}$$

This research will use log-sigmoid activation function which has a range of [0,1] in finding the output on the jth node

Step7: Utilizes the trees classification voting to assemble the prediction data from the (n) trained trees

Step8: Utilizes the highest voted features to construct the final RF model

Step 10: Train BPNN classifier and the error rate is computed using gradient descent algorithm. When error rate is low, the learning rate increases, whereas when the error rate is high and the learning rate is decreased

Step 11: The new weights and bias are updated based on the error rate and learning rate using gradient descent backpropagation algorithm. The Step 2 and Step 3 are repeated till the error rate converges

The output error function at the output neuron is defined as;

$$E = \frac{1}{2}\sum_{k=1}^{n}(t_k - o_k(\alpha_k))^2 \tag{33}$$

n : number of output nodes in the output layer.

t_k: desired output of the kth output unit.

o_k: network output of the kth output unit.

O_j : Output of the jth unit.

O_i : Output of the ith unit.

W_{ij} : weight of the link from unit i to unit j.

a_{net}, j net input activation function for the jth unit

Step 12: Reduce the error rate using (33)

Step 13: Classify the accurate features for the given datasets

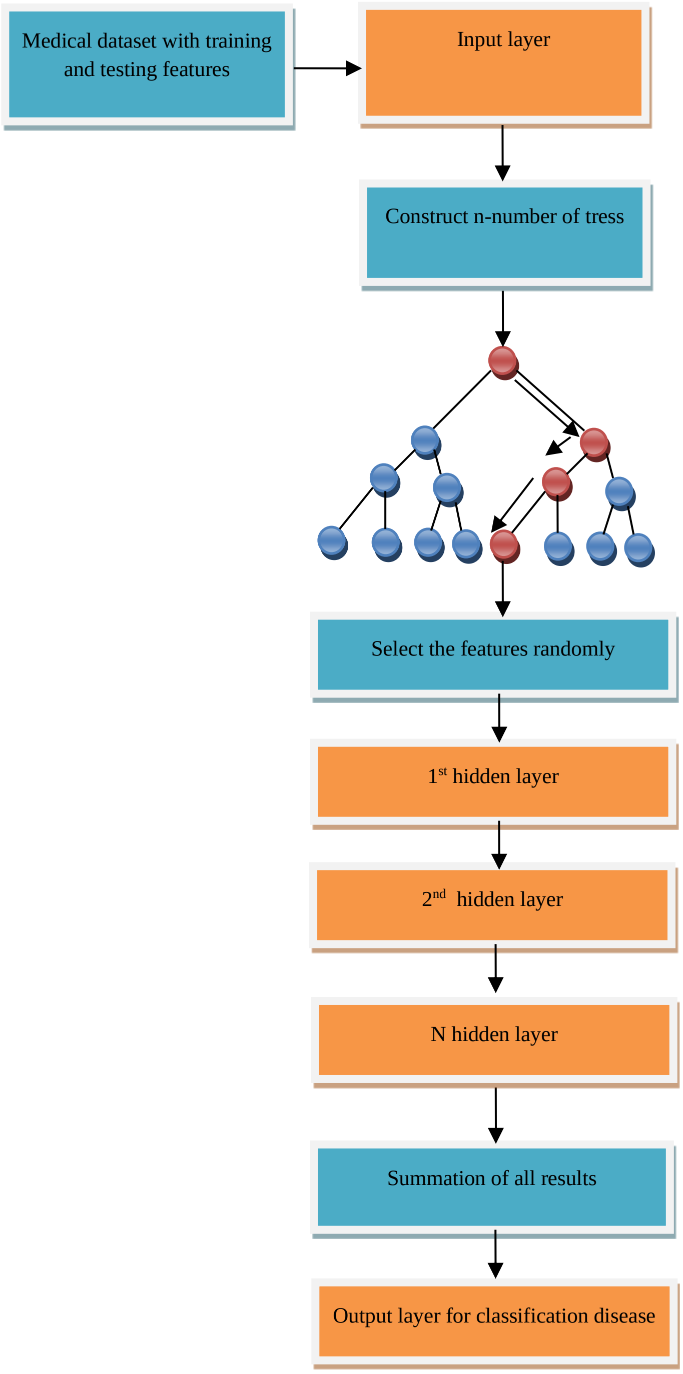
Medical dataset with training and testing features
Input layer
Construct n-number of tress
Select the features randomly
1st hidden layer
2nd hidden layer
N hidden layer
Summation of all results
Output layer for classification disease

Fig 6 Structure of HRFBP classifier algorithm

Fig 6 refers the structure of RF classifier algorithm for given four datasets. RF helps to define the tree with many decisions at training base and mode of class's output by tree. RF is basically mixture of tree predictor where every tree relies and the value of random vector sampled separately alongwith the same level of distribution. The basic rule is with the purpose of a group of “weak learners” able to come together to form a “strong learner”. RF is a wonderful tool for making classification considering they do not overfit since of the law of large numbers. The appropriate randomness is introduced to build them more precise classifiers and regressors in terms of quality. Single decision tree are advisable due to high variance and bias. But RF supports to alleviate these issues and finding the acceptable balance between both extremes. Samples selected based on aggregation function and its features are feed into RF classifier. Finally HRFBB is effectively classifying the hepatitis, Pima, heart disease and fertility output over the respective datasets

4. Experimental result

To evaluate the performance of HRFBP, the four data sets are taken from the UCI machine repository. The samples are normalized before machine learning such that the features are located in the [0, 1] range. The existing methods are considered as TWin Support Vector Machine (TWSVM) [24], Feature Selection method combined with Twin-Bounded Support Vector Machine (FSTBSVM) [25], AFOA-TBSVM and EFS-BPNN which is evaluated alongwith the proposed HRFBP algorithm. The performance metrics are such as accuracy, sensitivity, specificity and execution time which is compared between existing and proposed algorithms.

From the following link the hepatitis data is collected: https://www.kaggle.com/datasets/codebreaker619/hepatitis-data.

From the following link the fertility data is collected: https://www.kaggle.com/datasets/gabbygab/fertility-data-set.

From the following link the Pima data is collected https://www.kaggle.com/code/vincentlugat/pima-indians-diabetes-eda-prediction-0-906/data

From the following link the Heart Statlog data is collected https://www.kaggle.com/datasets/shubamsumbria/statlog-heart-data-set?select=statlog.csv

Accuracy

Accuracy is determined as the overall correctness of the model and is computed as the total actual classification parameters $(T_p + T_n)$ which is segregated by the sum of the classification parameters $(T_p + T_n + F_p + F_n)$. The accuracy is computed as like :

$$\text{Accuracy} = \frac{T_p+T_n}{(T_p+T_n+F_p+F_n)} \tag{34}$$

Where Tp is true positive, Tn is true negative, Fp is false positive and Fn is false negative

Table 1: Results of performance comparison for Accuracy

Metrics and dataset	Methods				
	FATBSVM	AFOA-TBSVM	TWSVM	EFS-BPNN	HRFBP
Accuracy -Diabetes	80.85	84.89	70.12	86.1	89.12
Accuracy-Fertility	90.90	93.93	71.24	98.2	97.15
Accuracy-Hepatitis	89.03	90.96	73.89	92.3	98.31
Accuracy-Statlog	85.54	87.64	73.33	90.02	93.45

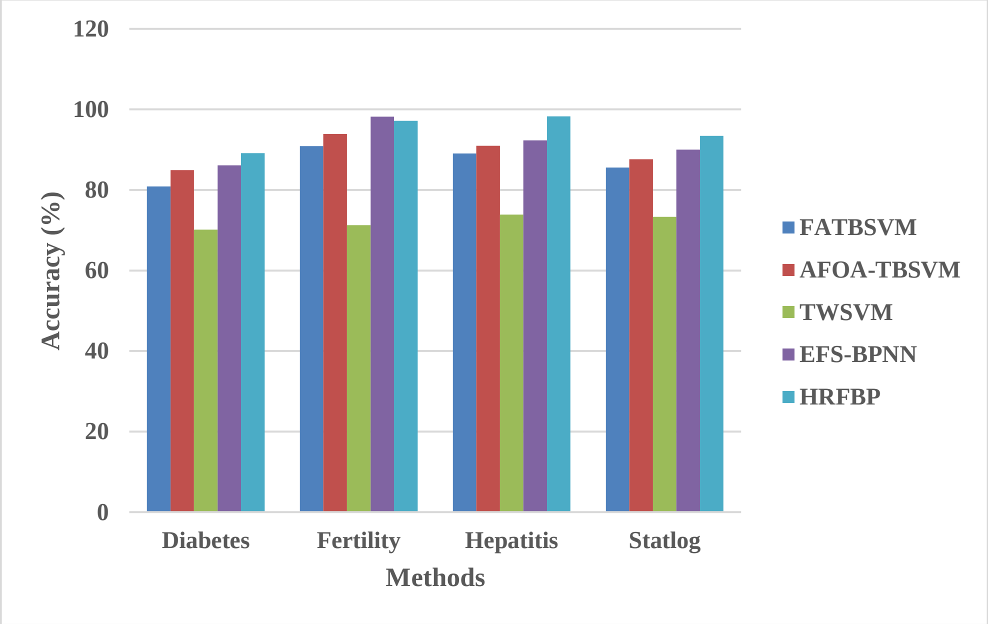

Fig 7 Accuracy

From the above Fig 7, it can be observed that the comparison metric is evaluated using existing and proposed method in terms of accuracy. For x-axis the datasets along with methods are taken and in y-axis the accuracy value is plotted. The existing methods are such algorithms TWSVM, FATBSVM, AFOA-TBSVM and EFS-BPNN provide lower accuracy whereas proposed HRFBP algorithm provides higher accuracy for the given Pima, Heart-Statlog, Hepatitis and Fertility datasets. Pre-processing method is used to increase the classification accuracy in higher via filling missing values and removal of noise. The important features are extracted via MPCA algorithm effectively. The optimal features are aggregated using EFS process and accuracy is increase via HRFBP algorithm

Sensitivity

Sensitivity is also called the true positive rate, the recall, or probability of detection in some fields measures and it the proportion of actual positives that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition)

$$\text{Sensitivity} = \frac{T_p}{T_p + F_n} \qquad (35)$$

Table 2: Results of performance comparison for Recall

Metrics and dataset	Methods				
	FATBSVM	AFOA-TBSVM	TWSVM	EFS-BPNN	HRFBP
Recall-Diabetes	81.14	85.11	66.23	86.15	89.21
Recall-Fertility	94.82	96.55	67.33	99.01	96.15
Recall -Hepatitis	89.62	90.84	68.30	92.04	99.25
Recall- Stat log	85.31	87.40	69.39	90.05	92.41

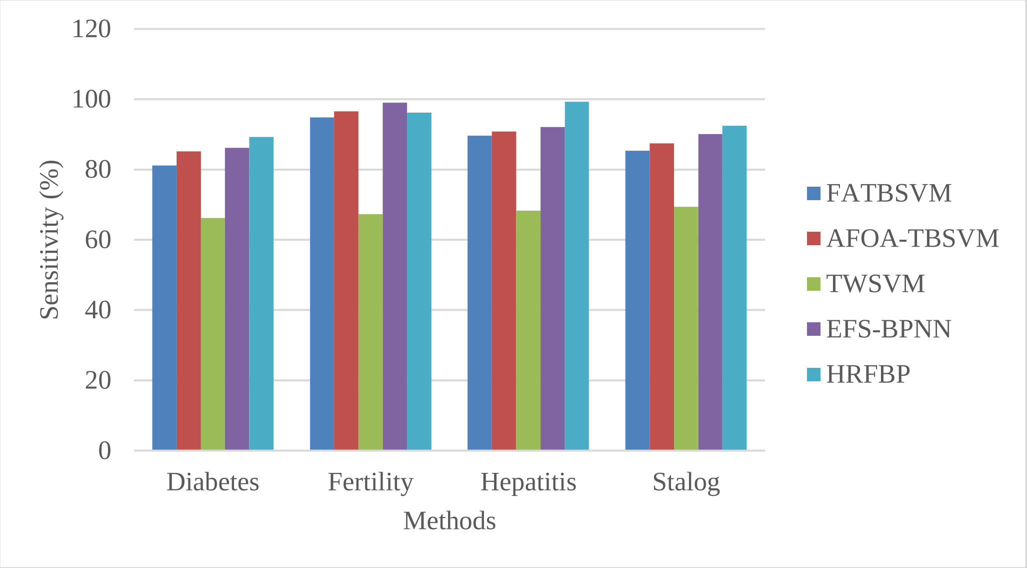

Fig 8 Sensitivity

From the above Fig 8, it can be observed that the comparison metric is evaluated using existing and proposed method in terms of sensitivity. For x-axis the methods are taken and in y-axis the sensitivity value is plotted. The existing methods are such as TWSVM, FATBSVM, AFOA-TBSVM and EFS-BPNN algorithm provides lower sensitivity whereas the proposed HRFBP algorithm provides higher sensitivity for the given Pima, Heart-Statlog, Hepatitis and Fertility datasets. MPCA is focused to extract more informative features. Thus, the result concludes that the proposed HRFBP algorithm increase the medical dataset classification performance through the optimal features

Specificity

Specificity (also called the true negative rate) measures the proportion of actual negatives that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition).

$$\text{Specificity} = \frac{\mathrm{T_n}}{\mathrm{T_n}+F_p} \quad (36)$$

Table 4. Results of performance comparison for Specificity

Metrics and dataset	Methods				
	FATBSVM	AFOA-TBSVM	TWSVM	EFS-BPNN	HRFBP
Precision -Diabetes	79.13	83.20	64	86.12	89.25
Precision -Fertility	78.13	83.20	63.90	99.01	99.15
Precision -Hepatitis	82.38	84.94	61.27	92.08	98.35
Precision -Stat log	85.67	87.74	63.23	90.05	92.15

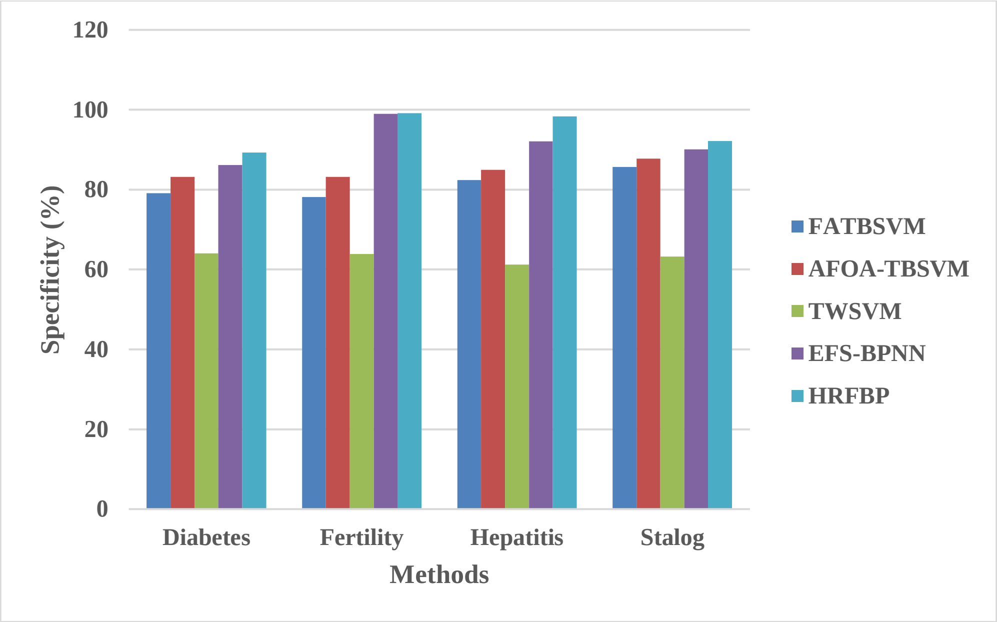

Fig 9 Specificity

From the above Fig 9, it can be observed that the comparison metric is evaluated using existing and proposed method in terms of specificity. For x-axis the methods are taken and in y-axis the specificity value is plotted. The existing methods are such as TWSVM, FATBSVM, AFOA-TBSVM and EFS-BPNN algorithms provide lower specificity whereas the proposed HRFBP algorithm provides higher specificity for the given Pima, Heart-Statlog, Hepatitis and Fertility datasets. The proposed method improves the sensitivity through the selection of more relevant information. EFS-HRFBP increases the performance in the training and testing stability. Thus the training and testing process of datasets is much more stable. Hence the result concludes that the proposed HRFBP algorithm increase the classification performance through the optimal features

F-measure

The feature ranking is obtained, where the higher the score, the more important the feature.

Table 5. Results of performance comparison for F- Measure

Metrics and dataset	Methods				
	FATBSVM	AFOA-TBSVM	TWSVM	EFS-BPNN	HRFBP
F-measure --Diabetes	80.12	84.14	65.89	86.21	87.45
F -measure -Fertility	85.93	89.45	65.23	93.15	94.51
F- measure -Hepatitis	85.84	87.79	62.56	92.21	95.61
F- measure -Stat log	85.49	87.57	64.67	90.31	92.15

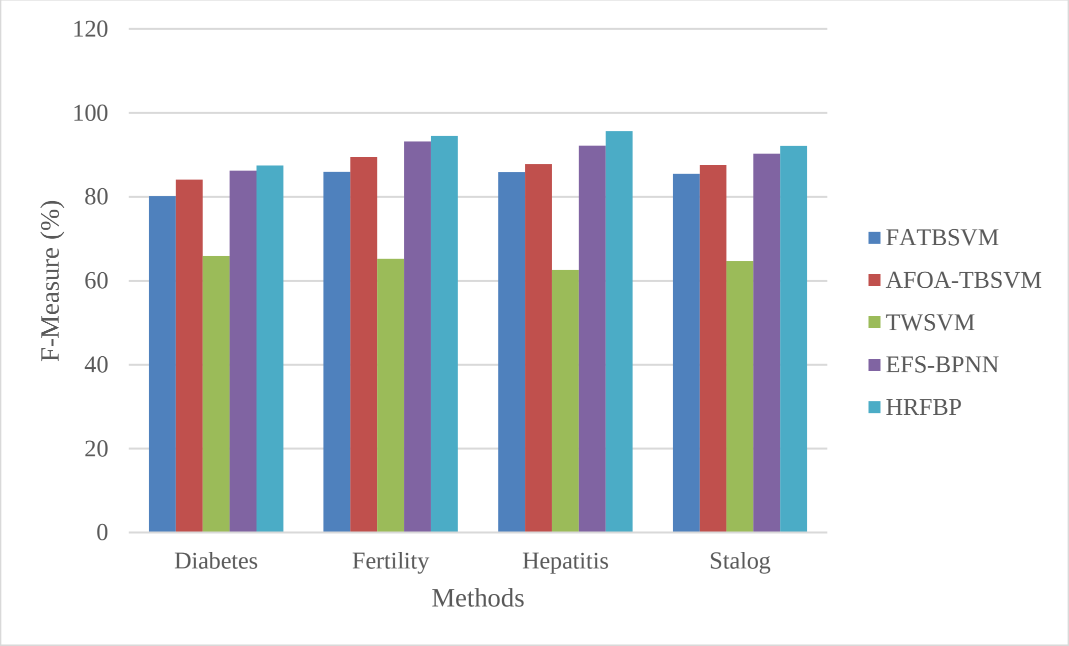

Fig 10 F-measure

From the Fig 10, the comparison values for F-measure metric using existing and proposed algorithm is evaluated. The existing TWSVM, FATBSVM, AFOA-TBSVM and EFS-BPNN methods provide lower F-measure whereas proposed HRFBP algorithm provides higher F-measure for the given medical datasets. The proposed classifier demonstrated that F1 score of 87.5% in the prediction, without any incorrectly identified features. EFS algorithm is used to provide optimal features. Hence the proposed algorithm provides higher accuracy of classification and better performance for given medical datasets

Execution time

The system is better when the proposed algorithm executes in less time consumption

Table 6. Results of performance comparison for Execution Time

Metrics and dataset	Methods				
	FATBSVM	AFOA-TBSVM	TWSVM	EFS-BPNN	HRFBP
Diabetes	34	45	18	20	15
Fertility	30	47	16	18	14
Hepatitis	32	42	13	16	11
Stalog	42	48	22	31	18

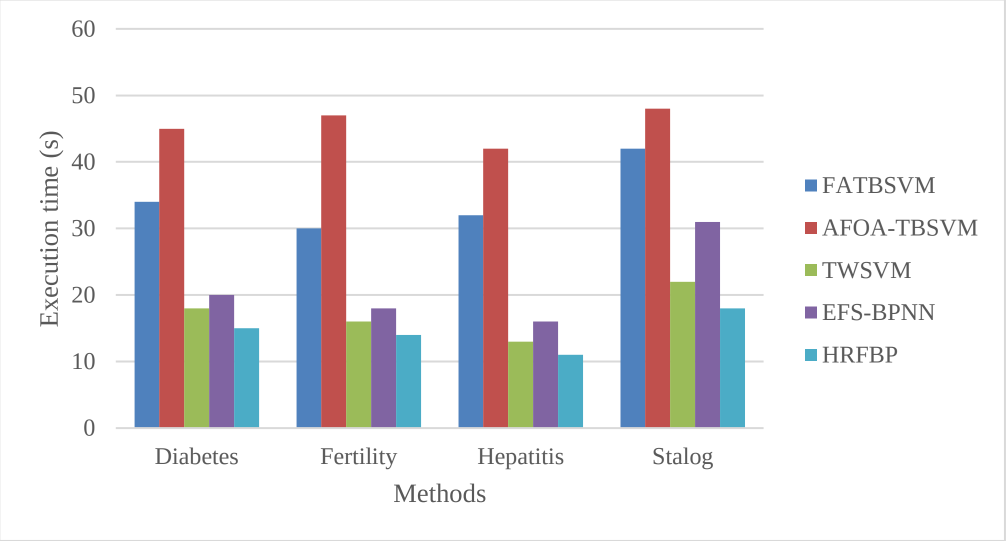

Fig 11 Execution time

From the above Fig 11, it can be observed that the comparison metric is evaluated using existing and proposed method in terms of execution time. For x-axis the methods are taken and in y-axis the execution time value is plotted. The existing methods are such as TWSVM, FATBSVM, AFOA-TBSVM and EFS-BPNN algorithms provide higher execution time whereas the proposed HRFBP algorithm provides lower execution time for the given datasets. Thus the result concludes that the proposed HRFBP algorithm increase the medical dataset performance through the optimized features

5. Conclusion

In this work, HRFBP algorithm is proposed to improve the medical dataset classification performance. This work contains five main modules such as pre-processing, feature extraction, feature selection and classification. KMC algorithm is used to increase the classification performance by filling the missing values and noise removal. Then, feature extraction is done by using MPCA algorithm which is focused to extract more informative features. After that, feature selection is performed via EFS which is used to select the most relevant and useful features. EFS is introduced to recognize the more significant attributes for disease classification patients. EFS is executed by EEHO, EBFO and AFOA which is attained superior outcomes rather than utilizing single FS method. The optimal feature subsets obtained using aggregation function which could be used in the proposed classification method. The optimal attribute subset could choose depends on the classification performance and efficiency over the diagnosis outcomes. Finally, classification is done by using HRF with BPNN algorithm which provides more accurate medical dataset classification performance. Training and testing data is improved via speed and accuracy metrics. By using weight values of neural network and voting process RF classifier, the error rates are reduced significantly. From the experimental result, it concluded that the proposed EFS-HRFBP algorithm provides higher accuracy, sensitivity, specificity and lower execution time rather than the existing algorithms. In future work, the proposed algorithms can be applied for larger image databases via various modality images.

References

[1] Pendyala, Vishnu S., and Silvia Figueira. "Automated medical diagnosis from clinical data." *2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE, 2017.

[2] El-Sappagh, Shaker H., and Samir El-Masri. "A distributed clinical decision support system architecture." *Journal of King Saud University-Computer and Information Sciences* 26.1 (2014): 69-78.

[3] Nahato, Kindie Biredagn, Khanna Nehemiah Harichandran, and Kannan Arputharaj. "Knowledge mining from clinical datasets using rough sets and backpropagation neural network." *Computational and mathematical methods in medicine* 2015 (2015)

[4] Qureshi, Naeem Ahmed, et al. "Application of principal component analysis (pca) to medical data." *Indian Journal of Science and Technology* 10.20 (2017): 1-9.

[5] Yang F, Mao K. Robust feature selection for microarray data based on multicriterion fusion. IEEE/ACM Trans ComputBiolBioinform 2011;8

[6] Bhat, Mahima, and Maya V. Karki. "Feature selection based on PCA and PSO for multimodal medical image fusion using DTCWT." *arXiv preprint arXiv:1701.08918* (2017)

[7] Aljawarneh, Shadi, Vangipuram Radhakrishna, and Gunupudi Rajesh Kumar. "An imputation measure for data imputation and disease classification of medical datasets." *AIP Conference Proceedings*. Vol. 2146. No. 1. AIP Publishing LLC, 2019.

[8] Chandra Sekhar, Ch, et al. "Effectiveness of Backpropagation Algorithm in Healthcare Data Classification." *Green Technology for Smart City and Society*. Springer, Singapore, 2021. 289-298.

[9] Leema, N., H. Khanna Nehemiah, and Arputharaj Kannan. "Neural network classifier optimization using differential evolution with global information and back propagation algorithm for clinical datasets." *Applied Soft Computing* 49 (2016): 834-844

[10] Shah, Syed Muhammad Saqlain, et al. "Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis." *Physica A: Statistical Mechanics and its Applications* 482 (2017): 796-807.

[11] Hassanien, Aboul Ella, Moataz Kilany, and Essam H. Houssein. "Combining support vector machine and elephant herding optimization for cardiac arrhythmias." *arXiv preprint arXiv:1806.08242* (2018).

[12] Dash, Sujata, Ruppa Thulasiram, and Parimala Thulasiraman. "Modified firefly algorithm with chaos theory for feature selection: A predictive model for medical data." *International Journal of Swarm Intelligence Research (IJSIR)* 10.2 (2019): 1-20.

[13] Alam, Md Zahangir, M. Saifur Rahman, and M. Sohel Rahman. "A Random Forest based predictor for medical data classification using feature ranking." *Informatics in Medicine Unlocked* 15 (2019): 100180.

[14] Mohamad, Ismail Bin, and Dauda Usman. "Standardization and its effects on K-means clustering algorithm." *Research Journal of Applied Sciences, Engineering and Technology*6.17 (2013): 3299-3303

[15] Taguchi, Y. H., and Yoshiki Murakami. "Principal component analysis based feature extraction approach to identify circulating microRNA biomarkers." *PloS one* 8.6 (2013): e66714

[16] Arora, S. and Singh, S., 2019. Butterfly optimization algorithm: a novel approach for global optimization. Soft Computing, 23(3), pp.715-734.

[17] Tubishat, M., Alswaitti, M., Mirjalili, S., Al-Garadi, M.A. and Rana, T.A., 2020. Dynamic butterfly optimization algorithm for feature selection. IEEE Access, 8, pp.194303-194314

[18] Liu, Jingsen, et al. "A dynamic adaptive firefly algorithm with globally orientation." *Mathematics and Computers in Simulation* 174 (2020): 76-101.

[19] Liu, Changnian, et al. "Adaptive firefly optimization algorithm based on stochastic inertia weight." *2013 Sixth International Symposium on Computational Intelligence and Design*. Vol. 1. IEEE, 2013

[20] Wang, G.G., Deb, S. and Coelho, L.D.S., 2015, Elephant herding optimization. In 2015 3rd International Symposium on Computational and Business Intelligence (ISCBI) , pp. 1-5.

[21] Li, J., Lei, H., Alavi, A.H. and Wang, G.G., 2020. Elephant herding optimization: variants, hybrids, and applications. Mathematics, 8(9), pp.1-25

[22] Shang, S., Shi, M., Shang, W. and Hong, Z., 2016. Improved feature weight algorithm and its application to text classification. Mathematical Problems in Engineering, vol.2016,no. 7819626, pp.1-12

[23] Zhu, Min, et al. "Class weights random forest algorithm for processing class imbalanced medical data." *IEEE Access* 6 (2018): 4641-4652.

[24] Chandra, Suresh, and R. Khemchandani. "Twin support vector machines for pattern classification." *IEEE Trans. Pattern Anal. Mach. Intell* 29.5 (2007): 905-910

[25] de Lima, Márcio Dias, Juliana de Oliveira Roque e Lima, and Rommel M. Barbosa. "Medical data set classification using a new feature selection algorithm combined with twin-bounded support vector machine." *Medical & Biological Engineering & Computing* 58.3 (2020): 519-528