International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Medical Dataset Classification Using Ensemble Feature Selection and Back Propagation Neural Network Algorithm

Dr. T. Christopher [1] and N. Kumar [2]

Submitted: 02/05/2024 **Revised**: 15/06/2024 **Accepted**: 22/06/2024

Abstract: Due to the ongoing creation of digital data, the amount of medical data has significantly expanded in recent years. the many types of medical information, including reports, text, numbers, monitoring, and laboratory results. Because of a problem with a single optimisation technique in the current system, classification accuracy is not considerably guaranteed. Another significant issue is error rates, which prevents early illness prediction from being carried out effectively. This research work uses EFS (Ensemble Feature Selection) with BPNN (Back Propagation Neural Networks) to handle the afore mentioned issues. The input data is pre-processed using KMC (K-Means Clustering) algorithm, mainly for handling missing values and subsequently, EFS method is used to choose the features since it produces the best fitness values using an objective function. To solve the FS problem, EFS relies on integrating many FS rather than just one FS. Combining the results of multiple single FS approaches, such as EEHO (Entropy Elephant Herding Optimisation) and AFOA (Adaptive Firefly Optimisation Algorithm), is one alternative for the EFS method. And EBFO (Entropy Butterfly Optimization Algorithm) acquire improved outcomes rather than utilizing a single FS methodology. Finally, the medical dataset classification is performed using BPNN algorithm. With the help of the BPNN algorithm, a multilayer FFNN (feed forward neural networks) is trained. The class labels in tuples are predicted using weights that are learnt iteratively. The experimental findingsof the proposed EFS-BPNN algorithm demonstrates better values for accuracy, sensitivity, specificity, and execution time when compared with existing methods.

Key words: *Medical dataset classification, EFS, Entropy Elephant Herding Optimization (EEHO), Adaptive Firefly Optimization Algorithm (AFOA) and Entropy Butterfly Optimization Algorithm (EBFO) and Back Propagation Neural Network (BPNN)*

1. Introduction

The procedure of diagnosing the existence of a disease is truly laborious since it necessitates in-depth information and extensive expertise. In general, the conventional method of reviewing medical reports, such as heart illness reports such as ECG (Electrocardiogram), MRI (Magnetic Resonance Imaging), Blood Pressures, and Stress Tests by Medical Practitioners, depends upon the prognosis of disease. Today's medical business has access to a sizable amount of medical data, which can be used to forecast key details about practically all medical issues. The practitioners would then be better able to forecast the future thanks to these findings [1]. The most accurate predictions over medical data have been produced by using unique methodologies and concepts.

The researchers are encouraged to conduct research in information extraction from clinical datasets by using knowledge mining. DMT (Data mining techniques) have been employed to mine rules and construct mathematical models for assisting clinicians in their decisions. The development of computerised database systems in this information age aids the improvement of medical science decision-making and diagnosis. DMT tools are used to analyse clinical records in order to create a knowledge-based system that may help physicians make decisions [2]. Clinical databases provide data about individuals' current health problems, including profiles, physical examinations, and test findings [3]. The process of uncovering hidden knowledge from clinical records in order to create clinical experts

Earlier analyses are not consolidated Ensemble Feature Selection (EFS) for heart disease, PIMA, fertility and hepatitis diagnosis. FS turn into the fundamental process in numerous data mining purposes. Choosing suitable attributes in the information are significant, because unessential attributes could reduce the numerous classifiers accuracy [4]. FS techniques are generally separated into filter, wrapper, and embedded strategies. Utilizing a solitary attribute subset determination technique might create nearby optima. Other than these three notable FS draws near, another gathering of techniques are built over the previous FS strategies: ensemble FS [5]. EFS builds group of attribute subsets and afterward join these subsets to create accumulated outcomes. EFS techniques applied to join different FS strategies as opposed to utilizing a single FS method. Traditional soft computing algorithms have not proficiently working in the EFS of high dimensional

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

dataset issues [6]. Then, various meta-heuristic approaches are adjusted for FS issues

A significant issue in the quickly developing field of DMT is classification [7]. NN (Neural networks) are highly suited to handle issues in biomedical engineering because of their broad variety of applications and capacity to learn complicated and nonlinear connections, including noisy or less accurate input. According to the investigation, the Multilayer FFNN with BPNN algorithm employing 15 input characteristics provides the maximum accuracy. The networks were trained using BPNN algorithm with momentum and variable learning rate. Various test data were fed into the networks as input in order to analyse network performance. Various test data were supplied to the network as input in order to analyse network performance. Physicians may use this binary Heart Disease dataset classifier to help them categorise the dataset. It was clear from the neural network design that MLP NNs needed a smaller architecture in terms of hidden node counts than other NNs and assist in classifications of samples. As a result, the amount of parameters, such as weights and biases, needed to create an MLP NN is significantly less than those of other methods. The categorization structure for medical data is shown in Fig. 1.

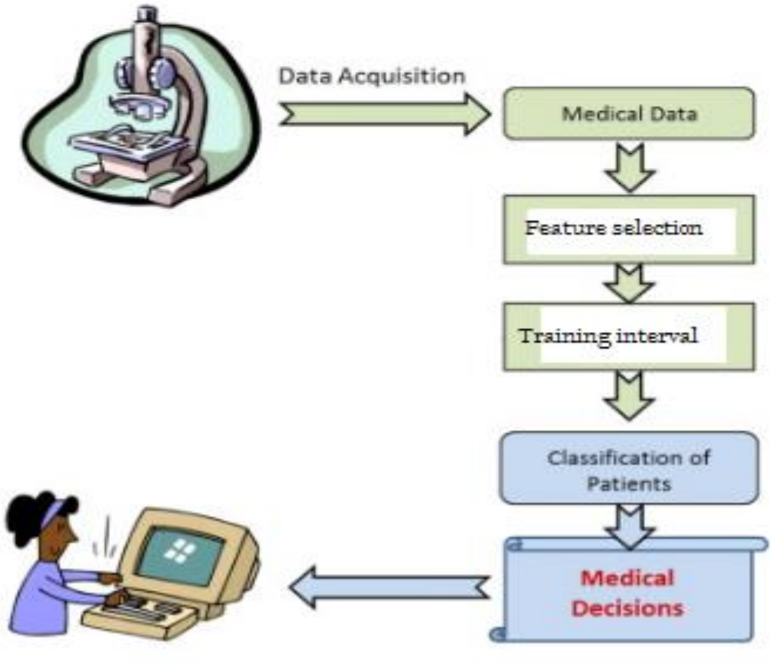

Fig 1 Structure of medical data classification

This work attempts to classify disease instances from even small sets of clinical datasets features. In order to classify medical data using the suggested EFS-BPNN method, this study is helpful. The accuracy of the medical data classifier is not entirely guaranteed, despite several research and methodologies that have been created. The present techniques have issues with error rates and inaccurate classification results. This work's suggested EFS-BPNN technique enhances overall classification performances while overcoming previously stated issues. This work executes classifications with pre-processes and feature selections as pre-phases Moreover, efficient algorithms are used for precise findings by algorithms for input datasets.

The remainder of the essay is structured as follows: In Section 2, there is a brief summary of some of the research on feature selection, pre-processing, and illness classification algorithms. In Section 3, the suggested technique for the EFS-BPNN system is described in depth. Section 4 provides the experimental findings and a discussion of the performance analysis. Finally, Section 5 summarises the findings.

2. Related work

Chandra et al. (2021) presented the idea of BPNN in [8], which is utilised for broadcasting the complete mistake back to reduce loss. It is regarded as a BPNN for classification since it is adaptable, simpler, and more effective with clean data. The experimental study was completed using data that was collected from the UCI repository. The study uses well-known datasets related to diabetes, cancer, the heart, and the liver. The classifier's effectiveness has been demonstrated by the decreased RMSE value and improved accuracy with additional parameters. Creating a classifier system based on BPNN could help identify doctors to handle medical issues.

For categorising clinical datasets, Leema et al. (2016) employed an ANN (Artificial Neural Network) trained using gradient descent-based BP (back propagation), PSO (Particle Swarm Optimisation), and DE (Differential Evolution). For better search explorations of PSO, a modified DE approach utilising mutations is used. PSO is used to train the ANN, and the e BP algorithm employs generated global best values as seeds. In order to produce the best sets of weights for NN, BP performs local searches where NN(Neural Network) weights are adjusted. Utilised were three standard clinical datasets from UCI's (University of California Irvine) ML (machine learning) repository namely: Prima Indian Diabetes; WDBC (Wisconsin Diagnostic Breast Cancer) and Cleveland Heart Disease. The suggested work and other classifier performances (gradient descent BP, DE with BP, and PSO with gradient descent BP algorithms) were compared with trained NN classifiers on these datasets.

Peng et al. (2016) in their study [10] proposed the usage of semi-supervised learning strategies to minimize tagged information using two breast cancer datasets of UCI ML repository. These two datasets are the subject of several experiments that are assessed. The testing outcomes show the algorithm's efficacy and efficiency, demonstrating that it is a potential automated detection technique for breast cancer.

Sakri et al. (2018) addressed in [11] how early recurrence prediction might assist patients in receiving therapy sooner. Accurate and quick prediction is made feasible by the availability of vast amounts of data and cutting-edge methodologies. This study examined the efficacies of DMTs in selecting right features and improving forecasts of breast cancer occurrences in datasets using classifiers including NB (Naive Bayes), KNN (K-nearest neighbour), and fast decision tree learner combined with PSO.

Gradient descendent BPNN were used for classifying clinical instances by Elgin Christo et al. (2019)'s in [12]. The study used bio-inspired techniques for feature selections. The study processes clinical data where Hot deck imputations handled missing values, and min-max normalizations modifies data. AdaBoostSVM classifier was used as fitness functions in wrappers that select features using bio-inspired algorithms like DE (Differential Evolution), Lion Optimisations, and GWO (Glowworm Swarm Optimisation). Each of these algorithms selected subsets of traits, resulting in three feature subset outputs. Correlation-based ensemble feature selections then identified important feature subsets. Ensembles selected apt features based on correlations. Gradient descendent BPNN trained on these features. Ten-fold cross-validations were used to train and assess classifier performances and accuracies were assessed on Hepatitis and WDBC datasets from UCI's ML repository.

Devikanniga et al. (2018) focused on developing a hybrid classifier model in [13] that uses bone mineral density measurements to distinguish between osteoporotic patients and healthy individuals. Then, an ANN classifier based on the monarch butterfly aided in quick diagnostics and thus prevention of osteoporosis. The study's experints were done on two datasets namely lumbar spine and femoral neck utilising Ten-fold cross-validations. The experimental outcomes demonstrated the classifier's effectiveness and showed that it consistently beat the alternative methods.

EHO (Elephant Herding Optimisation algorithm), a nature-based ML approach Nayak et al. (2020) proposed, was verified for identifications of lung, breast, and cervical cancers in multiple cancer datasets. EHO's results of classifications with/without feature selection were validated using RMSE (Root Mean Square Error), in comparisons with LLWNN, PSO, and EHO RMSE values were executed.

3. Proposed methodology

EFS-BPNN technique is suggested in this study to enhance performances of medical dataset classifications using three processes namely pre-processing, feature selections, and classifications. Fig. 2 displays the proposed EFS-BPNN system's overall flow diagram.

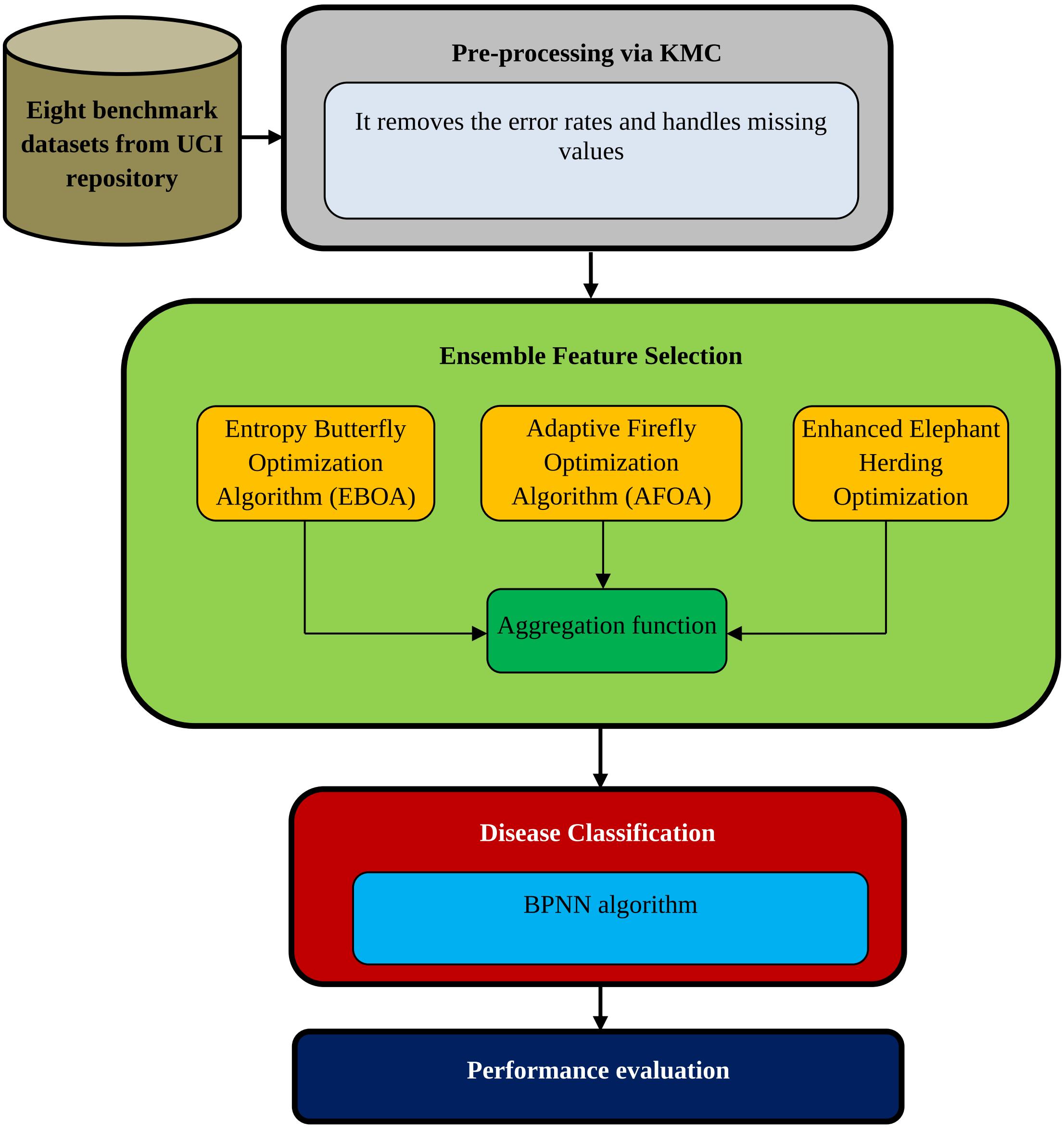
Eight benchmark datasets from UCI repository
Pre-processing via KMC
It removes the error rates and handles missing values
Ensemble Feature Selection
Entropy Butterfly Optimization Algorithm (EBOA)
Adaptive Firefly Optimization Algorithm (AFOA)
Enhanced Elephant Herding Optimization
Aggregation function
Disease Classification
BPNN algorithm
Performance evaluation

Fig 2 Overall block diagram of the proposed EFS-BPNN algorithm

3.1. Input dataset collection

This work’s datasets obtained from UCI ML repository included the datasets: Pima Indians with diabetes; Heart diseases and hepatitis, and fertility.

Hepatitis affects people of both sexes and age groups and causes a variety of symptoms. Fatigue, anorexia, a large liver, and other symptoms are hepatitis symptoms.

Semen samples from fertility datasets collected from 100 people from UCI ML repository was evaluated based on WHO 2010 guidelines. Sperm concentrations connected to socio demographics, environmental variables, health conditions, and lifestyle choices encompassing 10 characteristics of 100 cases were used . Season, age, childhood diseases, mishaps or severe trauma, surgeries, high fevers the year before, alcohol use frequency, and smoking habits were a few of the variables that impact fertility, daily sitting time estimates, and diagnosis.

The Pima databases contain outputs of patients with pregnancies, BMI, insulin levels, ages, glucose levels, blood pressures, skin thicknesses, and family histories of diabetes, together with counts of medical predictors (independent) factors.

Heart Statlog data included information about chest pains, levels of serum cholesterol and fasting blood sugars, electrocardiographic results, max. heart rates, old peaks/slopes on faults and exercise based angina. The ages/genders of the patients were also used.

3.2 Pre-processing using KMC algorithm

The KMC method is employed in this work's pre-processing in order to improve the dataset's illness identification accuracy. In this work, both structured and unstructured datasets are used. Hepatitis data and fertility data are examples of unstructured datasets. Pima and Heart Statlog data are examples of structured datasets.

KMC is a practical clustering technique that uses cluster beginning centroids to classify similar data [15]. The Euclidean distance concept is used to identify the cluster centroids. Starting with a random partitioning, (i) data point are re-assigned to clusters whose centres are closest to them while (ii) continuous re-computations of cluster centres based on averages of. These procedure were stooped when there were no more assignments. Intra-cluster variances computed as sums of squares of differences between features and cluster centres were reduced. Figure 3 depicts algorithmic example of KMC.

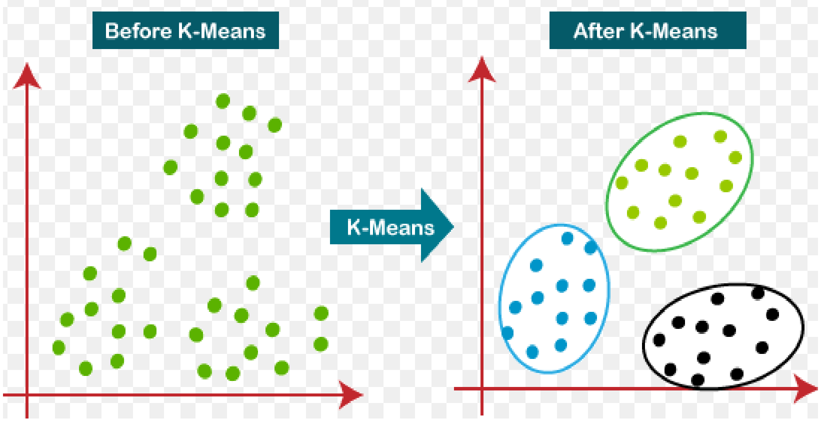

Fig 3 KMC algorithm

KMC implementation's efficacy and their linear runtimes are two important advantages. The counts of clusters in this work is maintained at one per class. Utilise the formula below to determine the Euclidean distance in order to determine the cluster centroids.

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

Where x_i, y_i Euclidean points in n space

Algorithm 1: Algorithmic KMC

1. Identify cluster counts (k) from structured/unstructured datasets,.

2. Create the cluster centres µ1,… µk.

3. Decide cluster centres of k data points) and use their locations.

4. Select cluster means and randomly allocate points to clusters.

5. Determine cluster centres nearest to data points, use distance measures and locate missing values using (1).

6. Add clusters to data points.

7. Recalculate cluster centres (cluster means of points).

8. Identify and remove incorrect or missing values.

9. Stop when all assignments are over.

Instances from original datasets that lack certain attributes were eliminated and dataset is divided into two sets, one including cases that are totally complete and have no missing values and the other containing instances that are partially complete and have missing values. Full instances were clustered by KMC to yield independent instances, and attributes without values were normalized. Newly entered instances were examined to determine their groupings in appropriate classes after KMC on datasets from generated clusters. When assigned points are in proper clusters, they are made permanent, and procedures repeated for all occurrences. Otherwise, next values are allocated and compared until the correct clusters are determined. As a consequence, the preprocessing method successfully increases the accuracy of sickness categorization by utilising the KMC algorithm.

3.3 EFS

The idea behind ensemble learning is that EFS strategies are specified to provide an optimal subset of features by integrating various FS based on EEHO, EBFO, and AFOA. The typical EFS design involves combining the outcomes of FS approaches to provide a representational capacity [16]. The development of several feature pickers and the aggregate of the decisions are the two fundamental aspects of EFS approaches.

3.3.1 EBFO

In this study, the best characteristics from the medical dataset are chosen by the use of EBFO. EBFO is a novel algorithm that draws inspiration from nature and imitates the food-seeking (more accurate with certain attributes) and butterfly mating behaviour to solve categorization issues in the medical illness diagnosis. The suggested EBFO Algorithm mimic butterfly's sense of smell to select most advantageous traits and locate nectar partners [17]. Based on research, it has been shown that butterflies can identify the source of smell with a high degree of classification accuracy.

Butterfly's fitness values correlate to (classification accuracy) their intensities which change as they travel from one location to another. EBFO Algorithm makes extensive usages of sensory modalities (c), intensities of stimuli (I), and power exponents (a) to deliver optimal feature selections [18]. I is related to fitness (accuracy) for EBFO algorithm's choice of features from medical datasets. EBFO algorithm formulates smell as functions of physical stimuli intensities using these concepts and equation (9).

$$f = cI^{a} \tag{2}$$

Where f represents perceived intensities of scents, c represents sensory modalities created by classification accuracies, In represents stimulus intensities, and a represents power exponents that are dependent on modalities. a and c in the interval [0, 1]. However, if a = 0, it indicates that none of the other butterflies are able to smell. Thus, the behaviours of algorithm depends on parameters a. C is yet another crucial variable that affects the EBFO algorithm's operation and the rate of convergence. In terms of searches, the following idealised descriptions of butterfly characteristics are given as examples:

1. Butterflies exude scents that attract other butterflies.

2. Butterflies migrate at random or in specific directions and best of them emits maximum smell.

3. Goal functions determine stimulus intensities of butterflies.

EBFO operates in the phases of initializations (start of runs), iterations and finalizations (on finding best optimum selections). Classification accuracy is calculated using the EBFO method and its solution space during the startup phase. Additionally, the values for the EBFO's parameters are assigned. In the feature selection search space, butterfly locations (features), together with fitness and smell values, are generated at random. After finishing the starting phase, the algorithm starts the iteration phase. During each iteration, each butterfly in the feature selection solution space is moved to a new location, and the classification accuracy values are then computed. All butterfly fitness values are initially computed throughout the solution space in the procedure. These butterflies will then release scent where they are by applying equation (3). During the period of global search, denoted by equation (3), the butterfly advances towards the fittest solution (g)(optimal features).

$$x_i^{t+1} = x_i^t + (r^2 \times g^* - x_i^t) \times f_i * ECE_W \tag{3}$$

where x_i^t stands for i^th butterfly's solution vectors x_i in iterations counts $t.$, g^* stands for selected feature solutions (current best) in current iterations. i^th butterfly's fragrance is f_i and $r \in [0,1]$ are randomized values in local searches depicted by Equation (4),

$$x_i^{t+1} = x_i^t + \left(r^2 \times x_j^t - x_k^t\right) \times f_i * ECE_W \tag{4}$$

where x_j^t feature selection space's j^th butterfly and x_k^t feature selection space's k^th butterfly. When these butterflies belong to the same swarm and $r \in [0,1]$ implies randomized numbers then equation (11) turns into local random walks. Butterflies have the ability to look for food and a partner at both local and global sizes in order to get the greatest selection of features from the dataset. Switch probabilities p are used in EBFO to move from general global searches to focused local searches. Up until the halting requirements are not fulfilled, the iteration process is continued. At the end of iterations, best possible solutions with highest fitness are obtained. To choose the ideal amount of features in the medical dataset, the EBFO algorithm additionally has feature weight applied in equation (4). The goal of the EBFO method was to increase the classifier's accuracy by making the best possible feature choices from the available medical dataset. By minimising the distance between two sample distributions, an optimisation issue is solved, and the best probability distribution parameters are then obtained. This technique is known as cross entropy (CE). The CE approach provides high resilience, outstanding flexibility, and good global search capabilities.

$$CE = \frac{1}{N}\sum_{i=1}^{N} I_{s<r} \frac{f(x^i, v)}{g(x^i)} \tag{5}$$

where x^i implies random samples from $f(x;v)$ with sampling densities $g(x)$. To determine ideal significance sampling densities, Kullback-Leibler divergences, or cross-entropies are used to calculate distances between sampling distributions.

The suggested EBFO algorithm's general phases are displayed in algorithm 2's flowchart. In step 1 of algorithm 2, the starting population is created by counting the features in the medical dataset, and step 2 of the algorithm computes the stimulus intensity I_i at x_i using sensor modalities c and power exponents a. Following are applied for stopping criteria (Step 4), a butterfly's fragrance value is determined (Step 6). Finding the population's greatest feature came next (Step 8), and then r was developed (Step 10). If r>p, move using equation (4), else move at random using equation (12). Then, in steps 17 and 18, you update a value and evaluate people in light of their new standing. Finally, use the end while command (Step 19) to finish the procedure. Fig. 4 depicts EBFO algorithm.

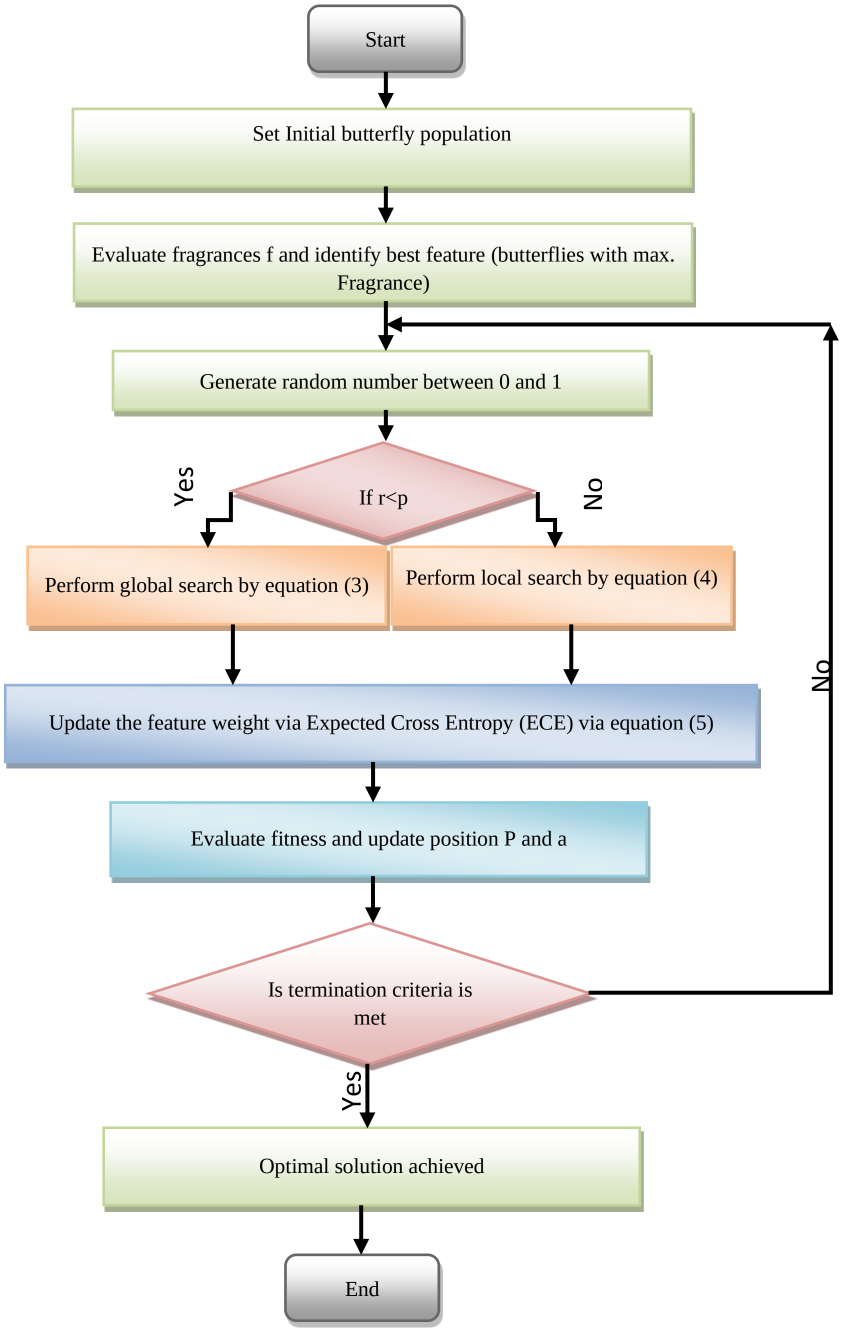
Start
Set Initial butterfly population
Evaluate fragrances f and identify best feature (butterflies with max. Fragrance)
Generate random number between 0 and 1
If r<p
Yes
No
Perform global search by equation (3)
Perform local search by equation (4)
Update the feature weight via Expected Cross Entropy (ECE) via equation (5)
Evaluate fitness and update position P and a
Is termination criteria is met
No
Yes
Optimal solution achieved
End

Fig 4 flowchart of EBFO

Algorithm 2: EBFO

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

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

Outputs: Optimal feature selections

1. Generate initial n butterfly populations $x_i = (i = 1,2, \ldots, n)$ based on dataset features
2. Stimuli intensities I_i at x_i obtained using classification accuracies $f(x_i)$
3. Define sensor modalities c, power exponents a and switch probabilities p

```
Do until criteria for stopping is not met
Do for butterflies $f$ in population
Compute fragrances using Equation (3) and compute entropy dependent weights based on equation (5)
End for
    Find best butterflies
    For butterflies $f$ in populations do
        Generate randomized values r
        If $r < p$ then
            Shift towards best butterflies (optimal features) using equation (3) and compute entropy dependent weights based on equation (5)
        Else
            Move in a randomized manner based on equation (4)
        End if
    End for
    Update a's values
    Feature evaluations based on their changed positions
End while
Display best solutions outcomes
```

3.3.2 AFOA algorithm

AFOA method is used in this study to choose features from provided datasets. The physiological and social The Firefly algorithm is based on characteristics of real fireflies which generate quick flashes that are signals of impending dangers as well as in luring (communicating with) their spouses. The Firefly Algorithm (FA) uses the problem's objective function to produce this flashing characteristic. The same idea underlies FA and firefly flashing lights. A firefly group is prompted by the brightness of the light to go to alluring and intense locations that are designed to produce the best resolution over the desired place.

A couple of the firefly traits are normalised by this technique, which may be demonstrated as follows [19]:

(i) All fireflies, regardless of their sexual orientation, are drawn to different people.

(ii) Because the brightness of a firefly is directly proportional to its attractiveness, when two fireflies are present, the one with the greater brightness attracts the one with the lower brightness. If it can't discover a brighter firefly in the region, a firefly will randomly change direction.

The brightness of the firefly is statistically influenced by the objective function. The basic operation of the firefly algorithm is shown in Fig. 5.

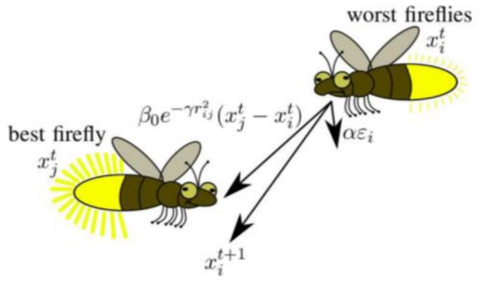

Fig 5 Basic mechanism of the firefly algorithm

FA was selected as it has the ability to provide best answers to multi-objective problems. For a maximisation problem, the brightness can simply be proportional to the objective function. To keep things simple, it is assumed that the attraction of a firefly is determined by its brightness or light intensity, which is linked to the encoded goal function.

Until the convergence criteria are met, these phases are iteratively repeated. a) Attractiveness and Light Intensity at the Source: According to the inverse square law, the amount of light varies as follows.

$$I(r) = \frac{I_0}{r^2} \tag{6}$$

Where I(r) stands for light intensities at attractiveness r^2

Attractions created in random assignments of features

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

$$I(r) = I_0 \exp\,(-\gamma r) \tag{7}$$

Where I_0 stands for medium's absorption coefficients

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

$$I(r) = I_0 \exp\,(-\gamma r^2) \tag{8}$$

Attractions of fireflies are proportional to light intensities as seen by adjacent ones. New solutions are achieved in variations and randomly changing pixels. Firefly's attractiveness β is computed as:

$$\beta = \beta_0 \exp(-\gamma r^m) \tag{9}$$

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

$$r_{i,j} = \sqrt{\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2} \tag{10}$$

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. AFOA is produced by introducing adaptations for both random and absorption parameters. These improvements improve the capacity of both local and global search by altering the parameter linearly during the duration of iterations [20]. AFOA is used to give the best features for the datasets that are provided by selecting the best features in terms of higher fitness values.

Calculate the parameter α as follow:

$$\alpha(t+1) = \left(1 - \frac{t}{MaxG}\right) \alpha(t) \tag{11}$$

α adjusts the value to the degree of the optimization's distance deviation in order to improve the solution's correctness and speed of convergence. It is also altered as follows to improve the population's adaptability.

$$\alpha = \alpha_{min} + (\alpha_{max} - \alpha_{min}) \times ||x_i - x_{best}||/L_{max} \tag{12}$$

Where $L_{max} = (x_{worst} - x_{best})$ (13)

$\alpha_{max} and\ \alpha_{min}$are the maximum and minimum features respectively. In Eq. (13), the location of the worst individual at generation t fireflies are represented by x_{worst}, and distances between worst and global optimal individuals are represented by L_{max}. In early stages, the firefly individuals are dispersed across space, with majority of them being at long distances from globally ideal individuals. At this stage, the value of $\|x_i - x_{best}\|$ is greater, and $L_{max}$and $(\alpha_{max} - \alpha_{min})$ are fixed. Eq. (12) revealed that higher values of early on have greater total optimisation impacts. As a result of algorithm's implementation, individual fireflies i are drawn to other fireflies that are brighter. As time goes on, firefly individuals i will cluster around the global ideal people. At this point, the value of x_i-x_best is lower, which is advantageous for enhancing the search for optimal features throughout the provided datasets. The value of is altered in accordance with the location of the optimum at each iteration, speeding up the algorithm's convergence [16]. The data above shows that the step size factor adjusts dynamically and adaptively depending on the distance between firefly individuals, balancing the capacity to move quickly and easily

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

$$f(x) = \frac{\left({}^{I_d}/_{I_t}\right) \times \left({}^{I_f}/_{{P_{init}}^{i}}\right)}{exp^{-{}^{e_E}/_{e/_M} + H_{accuracy}}} \tag{14}$$

where I_d is the counts of dropped features. m_t is the total counts 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.

$$x_i = x_i + \beta_0 e^{-\gamma r^2}(x_j - x_i) + \alpha(rand - \tfrac{1}{2}) \tag{15}$$

Where x_i and $x_j\ is$ distance between two firefly features

Each feature's fitness value is computed inside the population. The counts of features in a batch is chosen at random in the first generation. Each firefly's fitness value is computed. The selection process is then utilised to choose two particular fireflies. Along with having a higher brightness, Firefly also has the greatest fitness value, making it the feature of choice for the following generation.

Algorithm 3: AFOA for feature selection

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

Output: Optimal features

```
1.    Objective function $(x)$, $x$ = $(x1,...,)T$ take into account high accuracies of classifiers
2.    Set initial firefly populations $xi$ ($i$ = 1, 2, . . . , $n$)
```

```
3.  Light intensities $Ii$ at $xi$ obtained using $f(xi)$
4.  Specify coefficients of light absorptions $\gamma$
5.  while ($t$ < MaxGeneration)
6.  for $i$=1:$n$ all $n$ fireflies (features)
7.  for $j$=1:$i$ all $n$ fireflies (features)
8.  if ($Ij > Ii$), Move firefly $i$ towards $j$ in $d$-dimension;
9.  end if
10. Changes in attractiveness based on distances $r$ using exp[$-\gamma r$]
11. Use (14) and (15) to calculate the fitness function.
12. Use (10) to compute an objective model.
13. Compute fresh options and update light outputs
14. Eliminate unnecessary elements
15. Use (12) to update ideal characteristicss.
16. end for $j$
17. end for $i$
18. Sort the fireflies and identify their top qualities right now.
19. End while
20. Fireflies $i$ shift to more attractive ones
21. Return best features
```

According to method 3, the AFOA method is utilised to provide the best results based on fitness, which has a better measure for classifier accuracy. The fireflies are rated in the AFOA algorithm, and the best fitness values are used to choose the best firefly. The firefly iteration is continued after the innovative best solutions are added to the firefly pool. As a result, the AFOA algorithm is used in this study to choose the characteristics with the highest accuracy. When test datasets' extracted features are put to AFOA, they provide correlation matching with data features. If maximal brightness is achieved, the input test datasets will contain characteristics related to illnesses; otherwise, the input test datasets will contain features related to health.

3.3.3 EEHO

The clever heuristic EEHO algorithm is based on the nomadic lifestyle of elephants. The elephant herd mostly has the following two criteria for the selection of aspects from medical condition diagnosis, according to observation and research of the elephants. The first trait is that an elephant herd is divided into several clans, each of which has a patriarch and followers who adhere to his directives for the best selection of characteristics from illness diagnostics. Another defining characteristic of the herd is the lack of one adult male elephant. Young elephants will live away from the other elephants when they are adults. Clan updating and splitting, the two concepts that make up the core of EEHO, are both influenced by these two characteristics [21]. Equation (16) describes the update process for the clan updating operator, which is the initial characteristic of the elephant herd.

$$x_{n,i,j} = x_{i,j} + r * a * \left(x_{b,i} - x_{i,j}\right) * ECE_W \tag{16}$$

where old ($x_{i,j}$) and new ($x_{n,i,j},$) feature positions of clan i's elephants in medical datasets while α∈[0, 1] is the factoring scale; $x_{b,i}$ implies best fitness valued feature positions (classification accuracies) in clans i. r stands for randomized numbers with normal distributions in the interval [0, 1]. Equation (2) depicts the majority of individuals' update processes (features), but each clan's matriarch has not been changed [22]. Assuming specific feature values are observed for implying target features data in EEHO, the weights of medical dataset's features are determined. This weight value has now been changed for the EEHO algorithm. The weight from Expected Cross Entropy (ECE) is used to determine the relevance of the attributes. ECE is based on the Kullback-Leiber (KL) distance and measures the difference between the topic class's probability and its probability when a certain characteristic is present. You may explain computer equation (17)

$$\text{Cross Entropy (CE)(f)} = P(f)\sum_{i=1}^{|C|} P(f|c_i)\log\frac{P(f|c_i)}{P(c_i)} \tag{17}$$

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 (18),

$$\text{Information Entropy (IE)(f)} = -\sum_{i=1}^{|C|} P(f|c_i)\log(P(f|c_i)) \tag{18}$$

In a summary, combine equation (17) and equation (18) together; the ECE equation is as follows by equation (19),

$$\text{ECE (f)} = \frac{P(f)}{IE(f)+\varepsilon}\sum_{i=1}^{|C|}(P(f|c_i) + \varepsilon)\log\frac{P(f|c_i)+\varepsilon}{P(c_i)} \tag{19}$$

Information entropy equals zero (IE(f)=0) if feature f is unique to a single class. Therefore, it is necessary to include a small parameter as a regulator in the denominator. Equation (20)–(21) illustrates the matriarch's updating process for the feature selection process for illness detection.

$$x_{n,i,j} = \beta * x_{c,i} \tag{20}$$

$$x_{c,i} = \frac{1}{n_i} \times \sum_{j=1}^{n_i} x_{i,j} \tag{21}$$

where β stands for scales in the interval [0, 1]. Centre positions (features) in clans i is $x_{c,i}$ computed using equation (20), clan counts of elephants i is n_i and updates of matriarch positions (feature related to clans). Second elephant herd characteristic can be abstracted to provide separating operators and separation procedures are depicted in (22),

$$x_{w,i} = x_{min} + r * (x_{max} - x_{min}) \tag{22}$$

where $x_{w,i}$ x_{max} and x_{min} represent top and lower bounds of elephant's locations (feature positions), respectively. r implies values in the interval [0, 1], and represents positions (feature position) with poorest fitness values (classification accuracies) in clans i. Algorithm 4 displays the workings of the proposed EEHO algorithm. It begins with the population being initialised with the counts of features from the medical dataset, after which the fitness value (classification accuracy) is evaluated. Based on this, the poorest features (medical illness characteristics) are then eliminated from the clan, and the method is then started with t iterations up to T_max. Steps 6 through 8 are used to carry out two operations for each feature, such as clan update and the other being separating. Once these procedures have been carried out, step 11 to step 13 should be used to expel the worst elephant from the clan. Finally, in step 14, locate the best features. Similar to that, Fig. 6 shows the suggested system's flowchart.

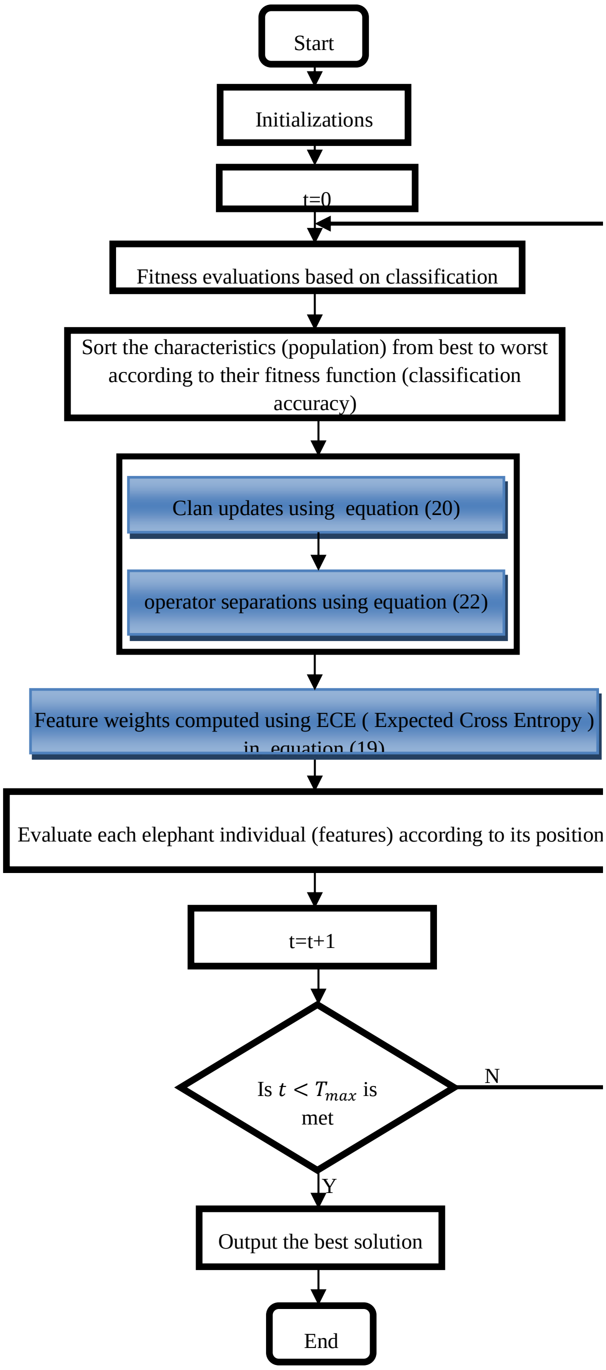
Start
Initializations
t=0
Fitness evaluations based on classification
Sort the characteristics (population) from best to worst according to their fitness function (classification accuracy)
Clan updates using equation (20)
operator separations using equation (22)
Feature weights computed using ECE (Expected Cross Entropy) in equation (19)
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 6 flowchart of EEHO algorithm

Algorithm 4: Enhanced Elephant Herding Optimization (EEHO) algorithm

```
1. Initialize parameter values and set population and feature counts
2. Evaluate fitness using positions of features and their classification accuracies
3. While $t < T_{max}$ do
4.   For $i = 1\ to\ n_c$ do
5.     For j=1 to $n_j$ ((Feature counts (elephants)) in clans) do
```

```
6.          Update $x_{i,j}$ and generate $x_{n,i,j}$ based on equations (16), produce feature weight by ECE (f) in equation (19)
7.          If $x_{i,j} = x_{b,i}$ then
8.              Update $x_{i,j}$ and generate $x_{n,i,j}$ based on the equations (20-21)
9.          End if
10.     End of for loop
11.     For $i = 1\ to\ n_c$ do
12.         Substitute worst features by modified network weights in clans i using equation(22)
13.     End of for loop
14.     Individual feature evaluations based on their new positions
15.  End of while  loop
```

3.3.4 Aggregation function

A suitable aggregation function is used to merge the many ranked results into a single ensemble list, assigning each characteristic a "overall score" based on its position (rank) in the initial records. Consider $L_k$be, the result of applying certain feature selection technique to kth bootstrap test (k = 1,..., B), as the ranking list. A final score is then calculated for each of the initial characteristics $f_i(i = 1, \dots N)$ using the formula $score_i = score\ (f_i) = aggr(r_{i1}, r_{i2}, \dots r_{iB})$. Here, aggr is a suitable aggregation function, and r_ik is the rank of the ith feature in the kth ranked list. The traits are ranked in the final ensemble list from most significant to least significant based on their total scores [38]. The optimal feature are selected which increases the medical dataset classification accuracy prominently.

3.4 Disease classifications using BPNN

In this study, the BPNN algorithm is used to classify diseases. Gradient descent BPNN with variable learning rates is the type of neural network employed in this study. The input, hidden, and output layers are parts of BPNN whose architecture is depicted in Fig. 7.

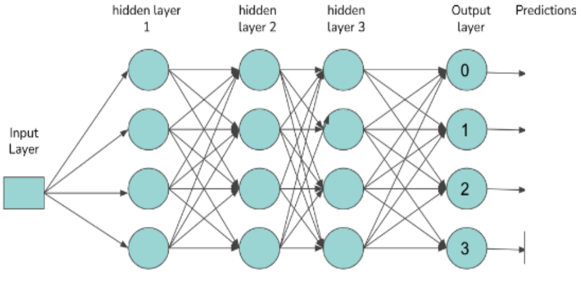

Fig 7 BPNN architecture

The most popular artificial neural network method is the basic FFNN BPNN algorithms. Algorithm modifies network weights using error propagations between outputs to inputs. By calculating the weights, the network reduces error during training. The gradient-descent method, in which the set of weight vectors made up of weights is modified by the learning parameters, can be used to carry out the minimization technique. It is possible to take into account network characteristics like learning rate, constant, training error, and counts of epochs. The weights are initialised to random values before training. Weights should be initialised with modest values to prevent saturation. The entire sample is randomly divided into training and test groups in order to assess the network's efficacy. The 80/20 rule, which states that 80% of the data are utilised for training and 20% are used for testing, is used to test the model. The training method is deemed effective in accordance with this classification strategy when the MSE reduces. The definition of an algorithm's training period is the number of epoch counts required to meet the halting condition.

The hidden layer's activation function is a sigmoidal function, whereas the output layer's activation function is a linear function. The following equation is used to compute the total counts of concealed nodes:

$$H = 2n + 1 \tag{23}$$

where H is the counts of hidden nodes and n is the counts of input nodes. 'e steps involved in this process are given below

Table 1: Parameter setting for BPNN

Parameter	Value	Meaning
N	Feature selected by EFS	Counts of input nodes (features)
H	2n+1	Counts of hidden nodes
H_{layer}	1	Hidden layer
O	-	Output

Algorithm 5: BPNN

Step 1: The disease features selected by the ensemble feature selector are given as the input of the BPNN. Initial parameters are initialized as shown in Table 1.

Step 2: The input of the hidden layer and the output of the hidden layer are calculated using equations (24) and (24):

$$a_{net} = \sum w_{i,j}\ O_i + \emptyset_j \tag{24}$$

Where $w_{i,j}$ are weights of input nodes where $\emptyset_j$ implies bias

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

(25)

In order to determine the output on the jth node in this study, the log-sigmoid activation function, which has a range of [0,1], will be used.

Step 3: Gradient descents computer error rates and learning rates rise on low error rates and reduce in the case of larger errors.

Step 4: Utilizing BP and gradient descents error and learning rate's new weights and bias are updated until error rates converge, Steps 2 and 3 are repeated.

Step 5: Use (26) to lower the mistake rate.

Step 6: Accurately classify the features for the provided datasets.The output error function at the output neuron is defined as;

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

(26)

n : output node counts of output layers.

t_k: kth unit's desired outputs.

o_k: kth unit's network outputs.

O_j : jth unit outputs.

O_i : ith unit outputs.

W_{ij} : weights of links from units i to j.

a_{net},j jth unit's net input activation functions

It helps with careful initial weights and bias selections, learning rates, momentums, network architectures, activation functions, and their gain values. Additionally, its utilization to increase training effectiveness and hasten network's convergence and effective modification of weights and enhance medical dataset classification performances. The weights are calculated using the gradient descent method, and the network is modified to reduce output error. The gradient descent BPNN method is seen in Fig. 8.

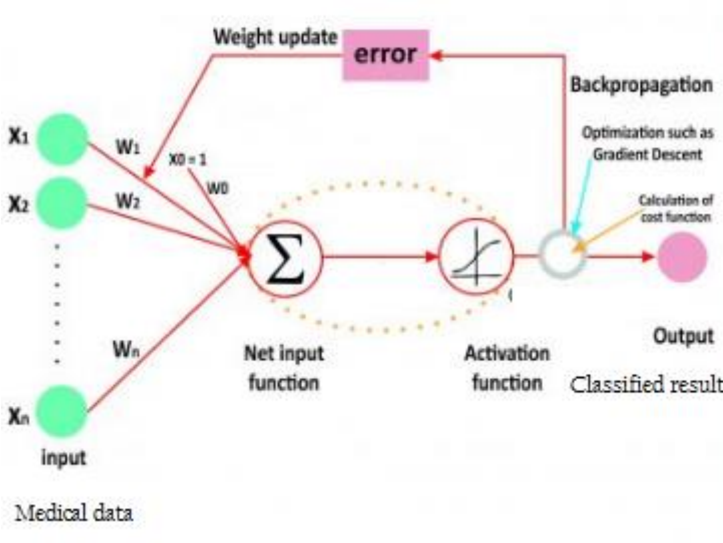

Fig 8 Gradient descent BPNN algorithm

1. Experimental result

To evaluate how well the EFS-BPNN performed, four data sets from the UCI ML repository were used. Input data for ML were normalised between [0, 1]. The present techniques that are taken into account include the TWin Support Vector Machine (TWSVM) [24], the Feature Selection Method Combined with Twin-Bounded Support Vector Machine (FSTBSVM) [25], and the AFOA-TBSVM that is evaluated with the proposed EFS-BPNN algorithm. The proposed method is compared to the existing algorithm in terms of performance measures including accuracy, sensitivity, specificity, and execution time. Hepatitis data was gathered from the following link: https://www.kaggle.com/datasets/codebreaker619/hepatitis-data.

Hepatitis information is gathered from the website https://www.kaggle.com/datasets/gabbygab/fertility-data-set.

The Pima data is gathered from the following source. https://www.kaggle.com/code/vincentlugat/pima-indians-diabetes-eda-prediction-0-906/data

The data for the Heart Statlog is gathered from the following URL. https://www.kaggle.com/datasets/shubamsumbria/statlog-heart-data-set?select=statlog.csv

Accuracy

Accuracies are determined as overall correctness of models and computed as total actual classifications $(\mathrm{T_p} + \mathrm{T_n})$ which are segregated by sums of classification parameters $(\mathrm{T_p} + \mathrm{T_n} + \mathrm{F_p} + \mathrm{F_n})$. Accuracy is computed as: :

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

(27)

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

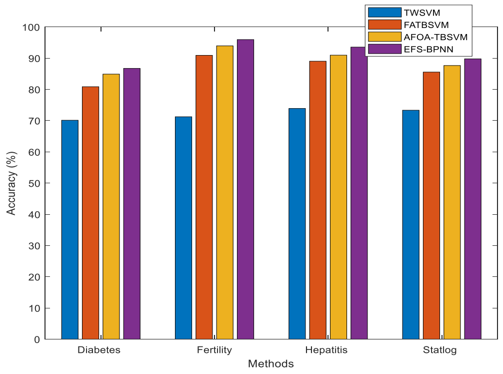

Fig 9 Accuracy

As can be seen in Fig. 9, both the current and new approaches are used to evaluate the comparability metric's accuracy. The x-axis uses the datasets and approaches, while the y-axis displays the accuracy value. The present processes make use of these algorithms. TWSVM, FATBSVM, and AFOA-TBSVM offer lesser accuracy for the supplied Pima, Heart-Statlog, Hepatitis, and Fertility datasets, but the suggested EFS-BPNN method offers superior accuracy. To increase classification accuracy, pre-processing techniques are employed to eliminate noise and fill in missing information. The best features are combined using the EFS process, and accuracy is increased using the gradient descent BPNN technique.

Sensitivity

Sensitivities are true positive rates, recalls, or probabilities of detection with certain measures and they are proportions of actual positives that are correctly identified, For Example, the percentage of sick people who were correctly identified as sick.

$$\text{Sensitivity} = \frac{\mathrm{T_p}}{\mathrm{T_p}+\mathrm{F_n}} \qquad (28)$$

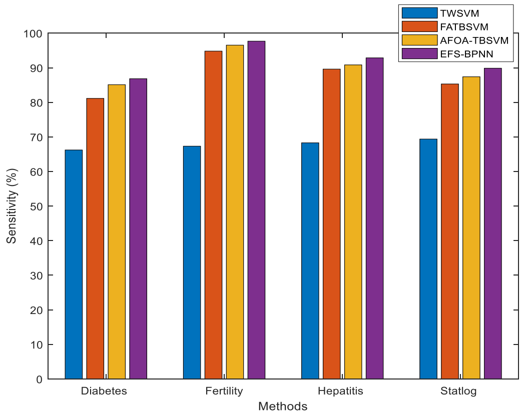

Fig 10 Sensitivity

The aforementioned Fig. 10 makes it evident that both the current and newly suggested approaches are used to evaluate the comparison metric's sensitivity. X-axis shows benchmarked methods while y-axis displays their corresponding values for sensitivity. The proposed EFS-BPNN method has more sensitivity for the supplied Pima, Heart-Statlog, Hepatitis, and Fertility datasets than the existing algorithms, such as TWSVM, FATBSVM, and AFOA-TBSVM. The recommended strategy improves sensitivity by choosing more relevant data. The findings indicate that by using the best features, the proposed EFS-BPNN algorithm enhances the performance of medical dataset categorization.

Specificity

Specificities are true negative rates that measure proportions of actual negatives that were correctly identified. For Example, percentage of healthy people who were correctly identified as healthy.

$$\text{Specificity} = \frac{\mathrm{T_n}}{\mathrm{T_n}+F_p} \qquad (29)$$

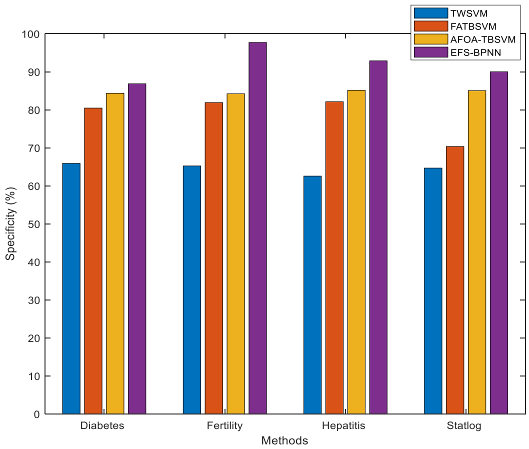

Fig 11 Specificity

Specificity values of existing and proposed methodologies were evaluated and depicted in Fig. 11 shown above. X-axis shows benchmarked methods while y-axis displays their corresponding values for Specificity. The suggested EFS-BPNN algorithm gives more specificity than the current approaches, such as TWSVM, FATBSVM, and AFOA-TBSVM, for the supplied Pima, Heart-Statlog, Hepatitis, and Fertility datasets. The recommended strategy improves sensitivity by choosing more relevant data. The training stability performance is enhanced by EFS-BPNN. Dataset training is significantly more stable as a consequence. The findings show that by using the best features, the proposed EFS-BPNN algorithm enhances classification performance.

F-measure

The outcomes of feature ranks obtained from Equation (18) are used for selecting important features based on their higher scores.

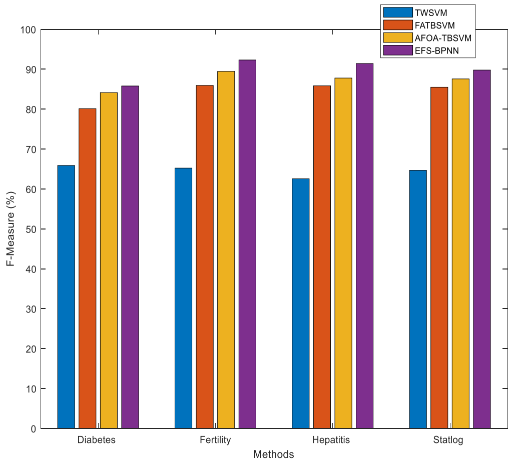

Fig 12 F-measure

The comparative values for the F-measure metric utilising the existing and suggested algorithms are assessed from Fig. 12. For the provided medical datasets, the proposed EFS-BPNN algorithm offers a higher F-measure than the current TWSVM, FATBSVM, and AFOA-TBSVM methods. The suggested classifier successfully predicted with an F1 score of 84% and no characteristics that were misidentified. To offer the best characteristics, the EFS algorithm is applied. Consequently, the suggested algorithm offers better performance and higher classification accuracy for the provided medical datasets.

Execution time

The proposed algorithm exhibits lesser execution times

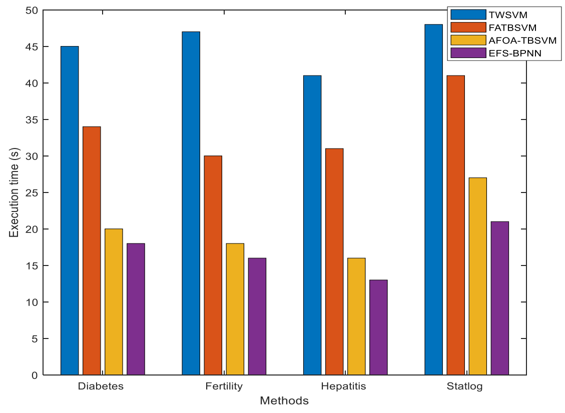

Fig 13 Execution time

The comparisons for evaluations are done in terms of execution times of existing and recommended approaches, as shown in Fig. 13 above. The x-axis represents the techniques, while the y-axis displays the execution time value. For the provided datasets, the suggested EFS-BPNN algorithm executes more quickly than current techniques like TWSVM, FATBSVM, and AFOA-TBSVM. The findings indicate that the recommended EFS-BPNN method makes use of optimised characteristics to enhance the performance of the medical dataset.

2. Conclusion

The EFS-BPNN technique is suggested in this study to enhance the performance of classifying medical datasets. Pre-processing, feature selection, and classification are some of the four primary components of this study. By adding missing values and eliminating noise, the KMC algorithm improves classification performance. Subsequently features are selected using EFS for identifying most pertinent and practical characteristics. EFS are used to identify the more important characteristics for patients with different illness classifications. To get better results than using a single FS approach, EFS is implemented by combining three FS methods, such as EEHO, EBFO, and AFOA. Both the EBFO and EEHO approaches derive attribute weight from ECE. It is done by utilising a distance metric that measures the disparity between the topic class's probability and possibility under a certain attribute condition. The feature subsets that might be employed in the suggested classification strategy were produced using the aggregation function. The best attribute subset to select will rely on how well and efficiently a classification can be made based on the results of a diagnostic. Finally, classification is carried out using the BPNN algorithm, which performs better in terms of accuracy for medical dataset classification. A collection of weights are learned iteratively to predict the class label of tuples. According to the experimental findings, the suggested EFS-BPNN algorithm offers superior accuracy, sensitivity, and specificity over current algorithms, as well as shorter execution times. In future work, feature extraction and hybrid classification algorithm can be developed for the given datasets

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 BP neural network." *Computational and mathematical methods in medicine* 2015 (2015).

[4] Park, H.W., Li, D., Piao, Y. and Ryu, K.H., 2017, A hybrid feature selection method to classification and its application in hypertension diagnosis. In International Conference on Information

Technology in Bio-and Medical Informatics (pp. 11-19). Springer, Cham

[5] Seijo-Pardo B., I. Porto-Díaz, V. Bolón-Canedo, A. Alonso-Betanzos, Ensemble feature selection: homogeneous and heterogeneous approaches, Knowl. Based Syst. 118 (2017) 124–139

[6] Arora, J., Agrawal, U., Tiwari, P., Gupta, D. and Khanna, A., 2020, Ensemble Feature Selection Method based on recently developed Nature-inspired algorithms. In International Conference on Innovative Computing and Communications (pp. 457-470). Springer, Singapore

[7] Amato, Filippo, et al. "Artificial neural networks in medical diagnosis." *Journal of applied biomedicine* 11.2 (2013): 47-58

[8] Chandra Sekhar, Ch, et al. "Effectiveness of BP 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] Peng, Lingxi, et al. "An immune-inspired semi-supervised algorithm for breast cancer diagnosis." *Computer methods and programs in biomedicine* 134 (2016): 259-265.

[11] Sakri, Sapiah Binti, Nuraini Binti Abdul Rashid, and Zuhaira Muhammad Zain. "Particle swarm optimization feature selection for breast cancer recurrence prediction." *IEEE Access* 6 (2018): 29637-29647

[12] Elgin Christo, V. R., et al. "Correlation-based ensemble feature selection using bioinspired algorithms and classification using BP neural network." *Computational and mathematical methods in medicine* 2019 (2019)

[13] Devikanniga, D., and R. Joshua Samuel Raj. "Classification of osteoporosis by artificial neural network based on monarch butterfly optimisation algorithm." *Healthcare technology letters* 5.2 (2018): 70-75.

[14] Nayak, Monalisa, et al. "Elephant herding optimization technique based neural network for cancer prediction." *Informatics in Medicine Unlocked* 21 (2020): 100445.

[15] 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

[16] Ng, W.W., Tuo, Y., Zhang, J. and Kwong, S., 2020. Training error and sensitivity-based ensemble feature selection. International Journal of Machine Learning and Cybernetics, 11(10), pp.2313-2326

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

[18] 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

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

[20] 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

[21] 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.

[22] 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

[23] 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

[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

[26] Menaga, D., Ambati, L.S. & Bojja, G.R. Optimal trained long short-term memory for opinion mining: a hybrid semantic knowledgebase approach. Int J Intell Robot Appl 7, 119–133 (2023). https://doi.org/10.1007/s41315-022-00248-w

[27] D. Menaga and A. Lakshminarayanan, "A Method for Predicting Movie Box-Office using Machine Learning," 2023 4th International Conference on Electronics and Sustainable

Communication Systems (ICESC), Coimbatore, India, 2023, pp. 1228-1232, doi: 10.1109/ICESC57686.2023.10192928.

[28] D. Kamalakkannan, D. Menaga, S. Shobana, K. V. Daya Sagar, R. Rajagopal & Mohit Tiwari (2023) A Detection of Intrusions Based on Deep Learning, Cybernetics and Systems, DOI: 10.1080/01969722.2023.2175134

[29] Manivannan R, Venkateshwaran G, Sivakumar S, Kumar MH, Jacob MS. Privacy-Preserving Image Storage on Cloud Using An Unified Cryptographic Authentication Scheme. Salud, Ciencia y Tecnología - Serie de Conferencias 2024; 3:609 . https://doi.org/10.56294/sctconf2024609.

[30] Menaga, D., Saravanan, S. GA-PPARM: constraint-based objective function and genetic algorithm for privacy preserved association rule mining. Evol. Intel. 15, 1487–1498 (2022). https://doi.org/10.1007/s12065-021-00576-z

[31] Journal of Autonomous Intelligence (2024) Volume 7 Issue 1

doi: 10.32629/jai.v7i1.734,Deep learning-based cancer disease classification using Gene Expression Data, J. Dafni Rose*,K. Vijayakumar, D. Menaga

[32] Karn, A. L., Bagale, G. S., Kondamudi, B. R., Srivastava, D. K., Gupta, R. K., & Sengan, S. (2022). Measuring the determining factors of financial development of commercial banks in selected SAARC countries. Journal of Database Management (JDM), 33(1), 1-21.

[33] Raj, K. B., Seth, J. K., Gulati, K., Choubey, S., & Patni, I. (2022, July). Automated Cyberstalking Classification using Social Media. In 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES) (pp. 1-6). IEEE.

[34] Chaubey, P. K., Arora, T. K., Raj, K. B., Asha, G. R., Mishra, G., Guptav, S. C., ... & Alhassan, M. (2022). Sentiment Analysis of Image with Text Caption using Deep Learning Techniques. Computational Intelligence and Neuroscience, 2022.

[35] Rajkumar, N., Sriram, V. P., Badrinarayanan, M. K., Raj, K. B., & Patel, R. (2022). A novel framework for the automated healthcare disaster based on intellectual machine learning. World Journal of Engineering, 20(5), 801-807.

[36] Bhavana Raj, K. (2022). Industry 4.0: Smart Manufacturing in Industries-The Future. Machine Learning and Data Science: Fundamentals and Applications, 67-74.

[37] Reddy, S., & Raj, K. B. (2021). Security and Well-being in Tech-Savvy Urban Communities. In Interdisciplinary Research in Technology and Management (pp. 306-309). CRC Press.

[38] Sanil, H. S., Singh, D., Raj, K. B., Choubey, S., Bhasin, N. K. K., Yadav, R., & Gulati, K. (2021). Role of machine learning in changing social and business eco-system–a qualitative study to explore the factors contributing to competitive advantage during COVID pandemic. World Journal of Engineering, 19(2), 238-243.

[39] Yadav, S., Sudman, M. S. I., Dubey, P. K., Srinivas, R. V., Srisainath, R., & Devi, V. C. (2023, October). Development of an GA-RBF based Model for Penetration of Electric Vehicles and its Projections. In 2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS) (pp. 1-6). IEEE.

[40] Jana, S., Ghosh, A., & Guha, B. (2021). IPL 2019: Evaluating the performance of teams by DEA & SEM. Malaya Journal of Matematik Vol. S, 1, 41-45.

[41] Jana, S., Ghosh, A., & Guha, B. (2021). IPL 2019: Evaluating the performance of teams by DEA & SEM. Malaya Journal of Matematik Vol. S, 1, 41-45.

[42] Ghosh, A., Dey, M., Guha, B., Jana, S., & Sarkar, A. (2021). Performance evaluation & rankings of players in IPL 2019 by DEA & SEM. Malaya Journal of Matematik, Vol. S, 1, 46-56.

[43] Nandi, B., Jana, S., & Das, K. P. (2023). Machine learning-based approaches for financial market prediction: A comprehensive review. Journal of AppliedMath, 1(2).

[44] Baid, Y., Ghosh, A., Jana, S., & Giri, A. Evaluation of the conspicuous EPL matches for sponsorships using data envelopment analysis.

[45] Jana, S., Sharma, E. S., Khan, A., Maji, A. K., & Pal, R. K. (2022, December). Generating a Suitable Hash Function Using Sudoku for Blockchain Network. In International Conference on Frontiers in Computing and Systems (pp. 161-171). Singapore: Springer Nature Singapore.

[46] Ghosh, A., Dey, M., Guha, B., Jana, S., & Sarkar, A. (2021). Performance evaluation & rankings of players in IPL 2019 by DEA & SEM. Malaya Journal of Matematik, Vol. S, 1, 46-56.

[47] Subrata, J., Ghosh, A., Ghorui, N., & Serdeira Azevedo, P. Ranking of Financial Apps Using Fuzzy Ahp and Fuzzy Marcos: An Application of Multi-Criteria Decision-Making (Mcdm) Techniques.

[48] Swain, S., Gupta, R. K., Ratnayake, K., Priyanka, P. D., Singh, R., Jana, S., ... & Giri, L. (2018). Confocal imaging and k-means clustering of GABAB and mGluR mediated modulation of Ca2+ spiking in hippocampal neurons. ACS chemical neuroscience, 9(12), 3094-3107.

[49] Swain, S., Gupta, R. K., Ratnayake, K., Priyanka, P. D., Singh, R., Jana, S., ... & Giri, L. (2018). Confocal imaging and k-means clustering of GABAB and mGluR mediated modulation of Ca2+ spiking in hippocampal neurons. ACS chemical neuroscience, 9(12), 3094-3107.

[50] Tătăranu, E., Diaconescu, S., Ivănescu, C. G., Sârbu, I., & Stamatin, M. (2016). Clinical, immunological and pathological profile of infants suffering from cow's milk protein allergy. Romanian journal of morphology and embryology = Revue roumaine de morphologie et embryologie, 57(3), 1031–1035.

[51] Ciongradi, C. I., Sârbu, I., Iliescu Halițchi, C. O., Benchia, D., & Sârbu, K. (2021). Fertility of Cryptorchid Testis-An Unsolved Mistery. Genes, 12(12), 1894. https://doi.org/10.3390/genes12121894

[52] Ciongradi, C. I., Filip, F., Sârbu, I., Iliescu Halițchi, C. O., Munteanu, V., & Candussi, I. L. (2022). The Impact of Water and Other Fluids on Pediatric Nephrolithiasis. Nutrients, 14(19), 4161. https://doi.org/10.3390/nu14194161

[53] Ciongradi, C. I., Benchia, D., Stupu, C. A., Iliescu Halițchi, C. O., & Sârbu, I. (2022). Quality of Life in Pediatric Patients with Continent Urinary Diversion-A Single Center Experience. International journal of environmental research and public health, 19(15), 9628. https://doi.org/10.3390/ijerph19159628

[54] Popa, Ș., Apostol, D., Bîcă, O., Benchia, D., Sârbu, I., & Ciongradi, C. I. (2021). Prenatally Diagnosed Infantile Myofibroma of Sartorius Muscle-A Differential for Soft Tissue Masses in Early Infancy. Diagnostics (Basel, Switzerland), 11(12), 2389. https://doi.org/10.3390/diagnostics11122389

[55] Mohammed, A. H. (2021). Fish Schooling And Sorensen Trust Based Wireless Sensor Network Optimization. International Journal, 9, 6.

[56] Mohammed, A. H. DDoS Malicious Node Detection by Jaccard and Page Rank Algorithm in Cloud Environment.

[57] Mohammed, A. H. (2021). Invasive Weed Optimization Based Ransom-Ware Detection in Cloud Environment.

[58] Purohit, S. (2023). California Geographical Society, 96162, California, United States. Journal of Environmental Science and Public Health, 7, 176-184.

[59] Purohit, S. Role of Industrialization and Urbanization in Regional Sustainable Development–Reflections from Tier-II Cities in India.

[60] Purohit, M. S. (2012). Resource management in the desert ecosystem of Nagaur district_ An ecological study of land _agriculture_ water and human resources (Doctoral dissertation, Maharaja Ganga Singh University).

[61] Faisal, L., Rama, V. S. B., Roy, S., & Nath, S. (2022). Modelling of Electric Vehicle Using Modified SEPIC Converter Configuration to Enhance DC–DC Converter Performance Using MATLAB. In Smart Energy and Advancement in Power Technologies: Select Proceedings of ICSEAPT 2021, Volume 2 (pp. 643-653). Singapore: Springer Nature Singapore.

[62] Faisal, L., Rama, V. S. B., Yang, J. M., Wajid, A., & Ghorui, S. K. (2022, May). Performance and Simulation Analysis of IPMSyncRM (Internal Permanent Magnet Synchronous Reluctance Motor) for Advanced Electric Vehicle Design. In 2022 3rd International Conference for Emerging Technology (INCET) (pp. 1-6). IEEE.

[63] Faisal, L., Rama, V. S. B., Yang, J. M., Wajid, A., & Ghorui, S. K. (2022, May). Performance and Simulation Analysis of IPMSyncRM (Internal Permanent Magnet Synchronous Reluctance Motor) for Advanced Electric Vehicle Design. In 2022 3rd International Conference for Emerging Technology (INCET) (pp. 1-6). IEEE.

[64] Mohd, R., & Faisal, L. (2022). Smart Agricultural Practices using Machine Learning techniques For Rainfall Prediction: A case Study of Valkenburg station, Netherlands. Mathematical Statistician and Engineering Applications, 71(4), 8451-8462.