Optimization Based Feature Selection Algorithm with Twin-Bounded Support Vector Machine for Medical Dataset Classification

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

[1] Assistant Professor, PG and Research Department of Computer Science
Government Arts College, Coimbatore – 641018
chris.hodcs@gmail.com

[2] Research Scholars, PG and Research Department of Computer Science
Government Arts College, Udumalpet-642126
&
Assistant Professor, Department of Computer Science,
Dr.N.G.P. Arts and Science College, Coimbatore – 641048
pkn.kumaar@gmail.com

Abstract

The medical business demands innovative technical tools to objectively examine information.Currently, there is a lot less emphasis placed on the value of knowledge in medicine as a work needing computer assistance in everyday clinical scenarios.Now that the healthcare industry has expanded beyond the confines of the medical environment, clinical decision support systems are increasingly often used to refer to AI systems.The existing system has issue with optimized feature selection process and error rates lead poor classification performance. Also, the earlier stages of disease prediction are not achieved effectively. The Adaptive Firefly Optimization Algorithm (AFOA) and Twin Bounded Support Vector Machine (TBSVM) algorithm are therefore introduced in this study to address the aforementioned issues.The K-Means Clustering (KMC) technique is used to pre-process the datasets once they have been gathered. It is used to effectively manage the missing data and error rates.The AFOA method, which produces the best fitness values using an objective function, is then used to choose the features.The emphasis is on choosing the greatest characteristics through the best fitness values. The TBSVM technique is then utilized to classify the medical dataset.The inclusion of a regularization term to structural risk minimization, or TBSVM, aims to maximize the margin.Additionally, it helps to boost classification accuracy and save training time.It was found that the suggested AFOA-TBSVM algorithm delivers superior performance than the current algorithms in terms of the greater accuracy, sensitivity, and specificity as well as the shorter execution time. These findings were based on the results of the experiments.

Keywords: Medical dataset classification, feature selection, Adaptive Firefly Optimization Algorithm (AFOA) and Twin Bounded Support Vector Machine (TBSVM) algorithm.

1. Introduction

Machine Learning (ML) techniques are being increasingly used for detection and diagnosis of diseases for its accuracy and efficiency in medical data classification. On the basis of diverse medical test results, several ML algorithms are being used to diagnose various sorts of illnesses.With the use of these ML approaches, many ailments may be identified more quickly and accurately[1] [2]. In order to better categorize data into distinct categories, researchers are concentrating on creating superior machine learning (ML) models for classification. These models comprise a variety of input characteristics.The quantity of study being done in this area is enormous, and it is producing new findings that are having a significant influence on humanity.This field is producing vast amounts of data. These data must be analysed in order to uncover the underlying

patterns that are significant and pertinent. For clinical diagnosis, these patterns might be employed.

Computational or artificial intelligence has increasingly been used in medical diagnosis in recent years.In order to help a doctor, diagnose a patient's ailment, machine intelligence-aided decision systems are often being employed.Usually, a doctor builds up her expertise based on the signs and symptoms of their patients and the proven diagnosis.Thus, a doctor's experience has a significant impact on their ability to make a diagnosis. The use of computerized medical decision support systems has emerged as a realistic strategy to help doctors quickly and properly diagnose patients since it is now very simple to obtain and store a huge quantity of information digitally [3]. As the objective of such a system is to anticipate (i.e., diagnose) a new case based on the records and data that are already available, it may be thought of as a classification job.One of the most difficult problems in medical informatics is believed to be categorization tasks of this kind.

To increase the accuracy of medical data categorization, many academics have investigated uses of algorithms.Techniques for classifying data that are more accurate in classifying will provide data that is sufficient to identify possible patients, increasing the accuracy of the diagnosis[4]. In order to evaluate various illnesses and develop models that may either help with the diagnosis or identify the disease sooner, data mining and machine learning approaches have been applied more effectively.

Filter, wrapper, and embedding methods may all be used to select features.The highest scoring features are those that are most closely associated to the data class, and each feature is given a value of significance (score) as part of a filter procedure to create a ranking.The performance of a classifier in a given subset of features is determined by a search algorithm, and wrapper techniques pick features based on that performance.Several subsets of characteristics are utilized as input data to the classifier once it has been selected [5].Then, it selects the subset with the characteristics that provide the maximum level of accuracy.In conclusion, built-in approaches involve the selection of features as an integral element of the process of training a model, such as the Recursive Feature Elimination (RFE) of Variables and the Least Absolute Shrinkage and Selection Operator (LASSO).

In previous work, single feature selection algorithm is utilized. Hence, the classifier accuracy is not ensured prominently. Soft computing-based feature selection approach is utilized in this study to increase classifier accuracy in order to address the aforementioned issues.Complex medical data may be analysed using an area of computer science known as "soft computing" [6]. The diagnostic, treatment, and outcome prediction they are capable of making from the relevant relationship set in a data set may be employed in a variety of clinical situations.The term "soft computing" refers to the practice of combining many technologies, either as an integral component of an integrated approach to the resolution of a problem or in order to carry out a specific job that is then followed by a second method that carries out another activity.

In [7], the Binary classification is carried out using the Feature Selection method combined Twin Bounded Support Vector Machine (FSTBSVM) method.It made use of eight benchmark data sets that were widely utilized by scientists who created machine learning techniques for categorizing medical data.The results of the trial demonstrate how

well the new FSTBSVM performs. Utilizing classification accuracy, sensitivity analysis, and specificity, the FSTBSVM's robustness is evaluated.The findings demonstrate that the suggested FSTBSVM is a highly promising strategy for classification and that it can get excellent results with less characteristics than the original data sets.

The categorization of medical data using ML algorithms is the primary goal of this research project.Despite several studies and approaches, there is no assurance of the correctness of the medical data classifier.The current methods have limitations with regard to mistake rates and erroneous categorization outcomes.The Twin Bounded Support Vector Machine (TBSVM) and Adaptive Firefly Optimization Algorithm (AFOA) algorithms are introduced in this study to enhance the overall classification performance in order to address the aforementioned problems.Preprocessing, feature selection, and illness categorization are the primary contributions of this study.With the help of efficient algorithms, the suggested technique produces more reliable findings for the provided benchmark databases.

The remaining portions of the essay are structured as follows: In Section 2, there is a short discussion of some of the studies in the field of illness classification, feature selection, and preprocessing.Section 3 provides specifics on the suggested technique for the AFOA-TBSVM scheme.Section 4 contains a description of the experimental performance analysis findings. Last but not least, Section 5 summarizes the results.

2. Related work

In [8], Wanget al (2021) based on deep feature fusion and improved routing, a proposed automated classification approach of breast cancer histopathology pictures is presented.In order to concurrently extract convolutional features and capsule characteristics, we first create a unique network with two parallel channels.After that, convolutional and capsule features are combined using a feature fusion technique to create more discriminating features.Additionally, upgraded routing was used to increase accuracy and speed by integrating the routing process into optimization. This was done in order to maximize efficiency.The findings of the experiments show that the combination of deep features and better routing achieves the greatest classification accuracy across the board.By combining convolution characteristics with capsule features, we show that this method has the capacity to identify malignant tumors from breast cancer histopathological pictures, which is helpful for clinical diagnosis.

In [9], Tomaret al (2014) proposed a machine learning technique for diagnosing cardiac problems called the Least Square Twin Support Vector Machine (LSTSVM), which is based on feature selection.This method weights each feature according to its F-score, and subsequently features are chosen based on their weight.A characteristic with a high F-score is given more weight. To improve the performance of the classifier, the grid search technique is also used to choose the optimal value for its parameters.In this research, the UCI repository's hearts tat log illness database was employed.A variety of training-test datasets have been used to assess the performance of models using various feature sets.LSTSVM model with 11 characteristics, according to the findings, has the maximum accuracy. In comparison to the preceding methods, the outcomes are highly encouraging.

In [10], Peng et al (2016) the requirement for labelled data, a semi-supervised learning technique was proposed.The UCI machine learning repository provided the research with two well-known benchmark breast cancer datasets.These two datasets are the subject of several experiments that are assessed.The experimental results show that the algorithm is a potential automated detection technique for breast cancer, demonstrating its efficacy and efficiency.

In [11], Sakriet al (2018) early recurrence prediction, when considered in light of recent technological breakthroughs, may assist patients in receiving therapy sooner.Accurate and efficient prediction is made feasible by the availability of vast amounts of data and cutting-edge techniques.This study compares how well various data mining methods already in use perform in predicting the recurrence of BC.In order to boost the prediction model's accuracy level, it incorporates a particle swarm optimization as a feature selection into three well-known classifiers, including naive Bayes, K-nearest neighbour, and rapid decision tree learner.

In [12], Carrera et al (2017) developed computerized processing of retinal images-based computer-assisted diagnostics to aid individuals in early detection of diabetic retinopathy.The primary objective is to develop an automated system that can categorize the severity of non-proliferative diabetic retinopathy based on any retinal picture.In order to do this, an early step of image processing separates blood vessels, micro-aneurysms, and hard exudates so that features can be extracted from them. These features may then be employed by a support vector machine to determine the retinopathy grade of each retinal image.A collection of 400 retinal pictures that have been graded on a 4-grade scale for non-proliferative diabetic retinopathy has been used to test this.

In [13], Ajam, N. (2015) Artificial neural network (ANN) have been tested for their potential to accurately diagnose diseases.An essential technique in medical diagnosis is classification, which is done using ANN for heart disease.To identify between the absence and presence of illness, a classifier using feed-forward back propagation neural network is utilized.There are 13 neurons in the input layer, 20 in the hidden layer, and only one neuron in the output layer.The test-and-error approach is used to choose an activation function and the number of neurons in the hidden layer.For the purpose of illness diagnosis, the data were taken from the UCI machine learning repository. As input and target data, respectively,the neural network's targets will be divided into categories with 0s designating absence of illness and 1s designating presence of disease. In the testing set, 88% of the instances could be classified, according to the outcome.

3. Proposed methodology

In this work, to enhance the performance of medical dataset categorization, the AFOA and TBSVM technique is suggested.Preprocessing, feature selection, and classification are the three primary processes of the suggested approach. Fig. 1 depicts the proposed system's overall block diagram.

3.1 Input dataset collection

The UCI machine repository is used to provide the data sets for this project. The data sets include those related to Pima Indians' hepatitis, diabetes, heart-stat log, and fertility.

Hepatitis affects individuals of both sexes and of all ages, and it has a variety of symptoms and effects. Fatigue, anorexia, a large liver, etc. are all hepatitis symptoms.

100 participants' fertility statistics were gathered, and a sample of semen was

examined in accordance with WHO 2010 standards.In the UC Irvine machine learning repository, which has 100 instances and 10 characteristics, sperm concentration is connected to sociodemographic information, environmental variables, health status, and lifestyle choices.Season, age, childhood illnesses, accidents or major trauma, surgeries, high fevers the previous year, alcohol use often, smoking habits, the amount of time spent sitting each day, and diagnoses are some of the factors that affect fertility.

The Pima datasets include a number of medical predictor factors that are considered independent, as well as one target variable that is considered dependent.The number of pregnancies that a patient has had, their body mass index, insulin level, age, glucose, blood pressure, skin thickness, diabetes pedigree function, and outcome are all characteristics that are taken into consideration.

Age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, and old peak are just a few of the characteristics that Heart Statlog data includes.Age and sex both imply the patient's year-old age, and Chest pain is indicated by the letters "cp," "trestbp," "chol," "fbs," "resecg," "resting electrocardiographic result," "maximum heart rate," "exang," "old peak level," "slope," "vessel," "number of main vessels," and "class" are all acronyms for coronary artery disease.

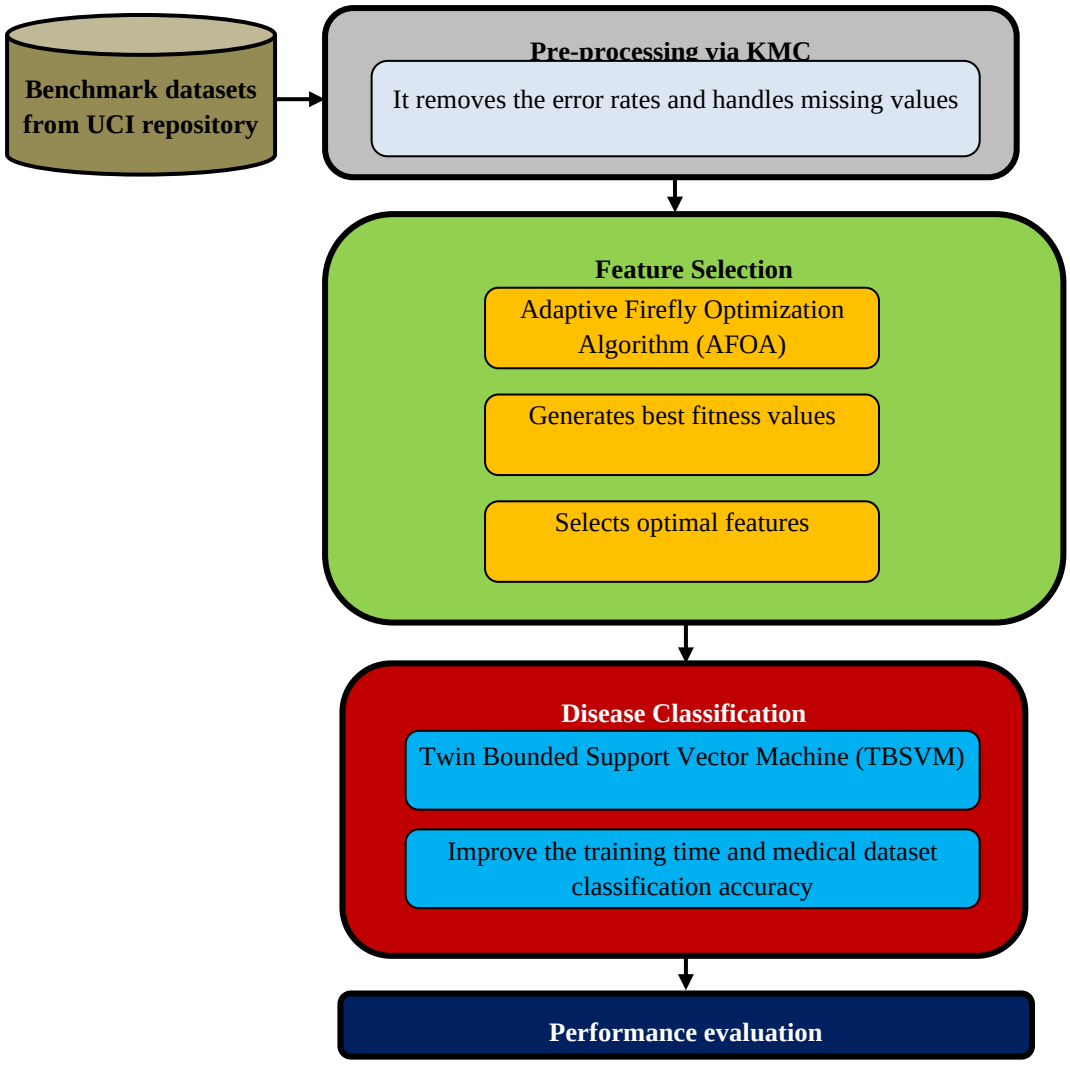

Fig 1AFOA-TBSVM algorithm's overall block diagram

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

To improve the accuracy of illness diagnosis for the provided database, pre-processing is carried out in this study utilizing the KMC method.The structured and unstructured datasets are utilized in this work. Unstructured datasets are Hepatitis data and fertility data. Structured datasets are Pima and Heart Stat log data.

Based on the initial centroids of clusters, KMC is a powerful clustering approach for classifying comparable data into groups[14].The centroids of the clusters are computed using the principle of Euclidean distance.The procedure continually computes the current cluster centres and I reassign each piece of data to the cluster with the closest center to it after randomly splitting the data. When there are no more transfers, it comes to an end.By doing so, the intra-cluster variance, also known as the sum of squares of the differences between data features and their associated cluster centres, is locally minimized. Specifically, this implies that the intra-cluster variance may be reduced to a greater extent. Fig 2 shows the example of KMC algorithm.

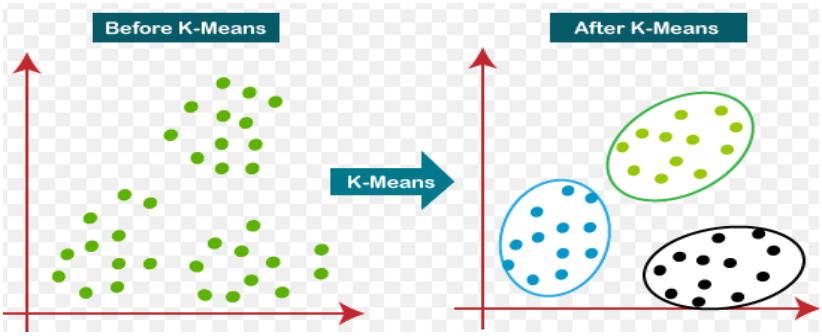

Fig 2 Example of KMC algorithm

The speed and simplicity of K-means implementation make them particularly advantageous.The number of clusters is maintained in this work at one per class.Utilize the following formula to calculate the Euclidean distance and identify the cluster centroids.

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

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

Algorithm 1: KMC algorithm

1. From the structured and unstructured datasets, choose k clusters (D)
2. Cluster centres in the beginning µ1,…µk
3. Set cluster centres to these locations using k data points.
4. Give clusters points at random, then calculate the cluster means.
5. Calculate the distance measure for each data point to determine which cluster center it is closest to in order to identify the missing values by (1)
6. Identify this cluster as the data point's home
7. Recalculating cluster centres
8. Find and remove error value and missing values
9. When no fresh re-assignments are received, stop

The instances with missing characteristics are removed from the database from the original instances.The database is partitioned into two sets, one of

which includes full examples that are devoid of any missing values, and the other of which contains incomplete instances that do have some gaps in their data.When KMC is applied to a collection of full instances, complete instance clusters are produced.In order to fill in the missing characteristics with potential values, the instances are examined one at a time.The newly inserted instance is checked to see whether it has been clustered in the right class after KMC is applied on the database from the generated clusters.The given value is made permanent and the method is then repeated with the next instance if it is in the right cluster.The value that places the instance in the proper cluster will be discovered by assigning and comparing the next feasible value if it is in the incorrect cluster.As a result, the KMC algorithm is utilized in the preprocessing approach to successfully increase the illness classification accuracy.

3.3 Feature selection using Adaptive Firefly Optimization Algorithm (AFOA) algorithm

In this work, feature selection is performed via AFOA algorithm over given datasets. The biological and sociological features of actual fireflies serve as inspiration for the Firefly algorithm.Authentic fireflies emit a flash of light that is both brief and cyclical; this serves two purposes: first, it helps them attract potential mates, and second, it acts as a warning signal.The objective function of the issue that has to be improved is how the Firefly Algorithm (FA) formulates this flashing characteristic. FA operates in a manner similar to how fireflies flash their lights.A firefly group is supported by the light's intensity as they move to appealing and intense spots that are planned to create the greatest resolution over the sought-after location.

A number of the firefly traits are normalized by this technique, which may be shown as follows [15]:

(i) Regardless of their gender, each firefly is drawn to a distinct person.

(ii) When two fireflies are present, the attractiveness of one firefly is attracted by the other's greater brightness, which is a clear comparison between the brightness the firefly produces and its attraction.If a firefly cannot find a nearby firefly that is brighter, it will change direction at random.

The objective function determines how bright a firefly appears in the statistical form. The fundamental workings of the firefly algorithm are shown in Fig. 3.

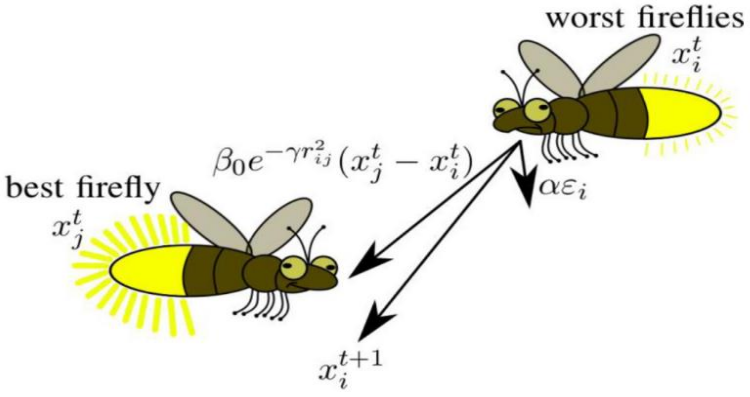

Fig 3 The firefly algorithm's fundamental workings

For its ability to provide the best solutions for multi-objective problems, the firefly algorithm is chosen.The brightness may easily be made proportional to the objective function if the issue is one of maximizing.For the purpose of simplicity, it is believed that the brightness or light intensity of a firefly, which is connected to the encoded goal function, determines its appeal.Until the convergence conditions are met, these steps are repeated recursively.

a) Light intensity varies according to the inverse square rule as follows: Attractiveness and Light Intensity at the Source

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

where the attractiveness's light intensity is denoted by I(r)r^2

By assigning traits at random, attractive

b)The light intensity is as follows while the intermediate is given:

$$I(r) = I_0 \exp(-\gamma r) \quad (3)$$

Where I_0 is a measure of how well a medium absorbs energy.

c) The following Gaussian form of the approximation is taken into consideration to prevent the singularity:

$$I(r) = I_0 \exp(-\gamma r^2) \quad (4)$$

The amount of light that the nearby fireflies can perceive directly affects how appealing a firefly is.By taking into account the potential for variation and randomly modifying the pixels, a new solution is created.Consequently, β a firefly's appeal consists of the following.

$$\beta = \beta_0 \exp(-\gamma r^m) \quad (5)$$

Where $\beta_0$is the attractiveness at r=0.

The following formula is used to calculate the distance between any two firefly (features) i and j.

$$r_{i,j} = \sqrt{\sum_{k=1}^{d}(x_{i,k} - x_{j,k})^2} \quad (6)$$

Where $x_{i,k}$ is the spatial match′s k[th] factor x_i of the k[th]d is the number of dimensions and firefly.An AFOA is produced by introducing an adaptation parameter for both the random and absorption parameters.By adjusting the parameter linearly over the iterations period, these adjustments enhance the capacity of both local and global search [20]. 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:

$$\alpha(t+1) = \left(1 - \frac{t}{MaxG}\right)\alpha(t) \quad (7)$$

α improves the solution accuracy and convergence speed by adjusting the value with the optimization's distance deviation degree.Additionally, it is revised as follows to increase the population's adaptability.

$$\alpha = \alpha_{min} + (\alpha_{max} - \alpha_{min}) \times ||x_i - x_{best}||/L_{max} \quad (8)$$

Where $L_{max} = (x_{worst} - x_{best})$ (9)

α_{max} and α_{min} are the maximum and minimum features respectively. In Eq. (9),

x_{worst} represents where the generation's worst person stands$t$ firefly, and $L_{max}$is a measure of how far the worst person is from the ideal person overall x_{best}. The firefly people are dispersed over the whole space in the early stages of the algorithm, and the majority of them are located at a great distance from the individuals who are globally optimum.Currently, the worth of$\|x_i - x_{best}\|$ is larger, $L_{max}$and $(\alpha_{max} - \alpha_{min})$ are constants. The value of α is thus bigger in the early stage, which has a stronger overall optimization impact, as shown by Eq. (8).Individual fireflies in the algorithm are drawn to fireflies that are brighter than themselves and nearer to the global ideal characteristics.When the time comes, I'll collect the firefly people near the most ideal people on the planet$\|x_i - x_{best}\|$is now smaller, making it easier to choose the best characteristics among the available datasets for searches. The location of the optimum is taken into account while changing the value of throughout each iteration, which accelerates the algorithm's convergence [16]. The step size factor, which balances the capacity of algorithm development and search, α varies adaptively and dynamically in accordance with the distance between the individuals of firefly, according to the study presented above.

A novel fitness function that takes into account accuracy and execution time has been developed as part of this study.

$$f(x) = \frac{\left({I_d}/{I_t}\right)\times\left({I_f}/{P_{init}{}^{i}}\right)}{exp^{-{e_E}/{e_M}+Haccuracy}} \quad (10)$$

where I_d is the quantity of characteristics removed. $m_t$delivered with more accuracy overall in terms of the quantity of features.

I_f is the features in the dataseti.

$P_{init}{}^{i}$is the first characteristic.

e_E is the time of execution, e_M is the longest delay permitted.

$$x_i = x_i + \beta_0 e^{-\gamma r^2}\left(x_j - x_i\right) + \alpha(rand - \tfrac{1}{2}) \quad (11)$$

Where x_i and x_j is the separation of two firefly characteristics

Every characteristic in the population has its fitness value computed.The number of characteristics in each batch is chosen at random in the first generation.Each firefly's fitness score is determined. The decision is made to choose two fireflies later using the selection process.The next generation is chosen from Firefly because it has the best fitness value in addition to its increased brightness.

Algorithm 2: 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,...,)Tas an objective function, take greater classifier accuracy
2. create the basic firefly populationxi (i = 1, 2, . . . ,n)
3. Ii at $xi$is a measure of light intensity $f(xi)$
4. Describe the coefficient of light absorption γ
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. The distance from an object affects how attractive it is$r$ via exp[$-\gamma r$]
11. Determine fitness level using (10) and (11)
12. Create a model that is objective using (6)
13. Use updated light intensity calculations and novel solutions to (3)
14. Utilizing, reduce the unnecessary features (19)
15. Improve the best features by employing (8)
16. end for j
17. end for i
18. The current top characteristics of the firefly are ranked
19. end while
20. An attractive firefly changes its appearance.
21. The finest features back

Algorithm 2 explains thatAFOAbased on fitness, which has a better classifier accuracy measure, an algorithm is utilized to create the best results.The fireflies are sorted in the AFOA algorithm, and the best fitness values determine which fireflies are optimum. The firefly is iterated further using the innovative best solutions, which are added to the firefly pool. 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 Disease classification using Twin Bounded Support Vector Machine (TBSVM)

TBSVM is based on TWSVM,used [17] which, by resolving two smaller QPPs, creates two non-parallel hyperplanes.The training method speed is increased using the Successive Over Relaxation (SOR) algorithm [18]. By including the regularization term in the TBSVM's fundamental issue, the structural risk reduction concept is put into practice. The following are the fundamental issues with TBSVM.

$$\min \_w1, b1, \varepsilon_1, \varepsilon_2 1/2c_3(||w_1||^2 + b_1^2) + 1/2\varepsilon_2^T\varepsilon_2 + c_1e_2^T\varepsilon_1$$

$$\text{such that} \begin{cases} Aw_1 + e_1b_1 = \varepsilon_2 \\ -(Bw_1 + e_2b_1) + \varepsilon_1 \geq e_2 \\ \varepsilon_1 \geq 0 \end{cases} \quad (12)$$

$$\min \_w2, b1, \varepsilon_1, \varepsilon_2 1/2c_4(||w_2||^2 + b_2^2) + 1/2\varepsilon_{\eta 2}^T\eta_2 + c_2e_1^T\eta_1$$

$$\begin{cases} Aw_2 + e_2b_2 = \eta_2 \\ (Aw_2 + e_1b_2) + \eta_1 \geq e_1 \\ \eta_1 \geq 0 \end{cases} \quad (13)$$

Where $c_i, i = 1,2,3,4\ are\ penalty\ parameters$

The additional regularization term is the difference between the primal issues (1)–(2) of TBSVM and the primordial problems of TWSVM [16] $\frac{1}{2}c_3(||w_1||^2 + b_1^2)$ and $\frac{1}{2}c_4(||w_2||^2 + b_2^2)$ utilized to reduce the structural risk in issues (12) and (13)[39]. This adjustment causes two converging issues.

$$\max_{\alpha} e_2^{T-} \alpha^T G(H^T H + c_3 I)^{-1} G^T \alpha \quad (14)$$

$$0 \le \alpha \le c_1 e_2$$

$$\max_{\beta} e_1^T - 1/2 \beta^T H(G^T G + c_4 I)^{-1} G^T \alpha \quad (15)$$

$$s.t\ 0 \le \alpha \le c_1 e_1$$

Where $G = [Ae_1]$ and $H = [Be_2]$ Lagrange multipliers α and β and matrices are present($H^T H + c_3 I)^{-1}$ and $(G^T G + c_4 I)^{-1}$ not singular. after finding answers (w_1, b_1) and (w_2, b_2) decision to use the new design is derived from issues (3) and (4), respectively $x \in \mathbb{R}^n$ comparable to TWSVM, is allocated to a class

Assume that the denotation for all inputs in classes +1 and 1 is $A \in R^{n_+ \times m}$ and $B \in R^{n_- \times m} respectively$.The F-score

$$R_{positive} = \frac{(x_i^{(+)} - x_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \hat{x_i})^2} \quad (16)$$

$$R_{negative} = \frac{(x_i^{(-)} - x_i)^2}{\frac{1}{n_- - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(-)} - \hat{x_i})^2} \quad (17)$$

where x_i represents the whole whole's average of the i^{th} characteristic, and $x_i^{(+)}$are, respectively, the averages of the contrasting samples. $x_{k,i}^{(+)}$ and $x_{k,i}^{(-)}$are, respectively, the i^{th} and k^{th} characteristics of the positive and negative samples. Note that the squared distance between is the numerator of equation (16) $x_i^{(+)}$ and x_iEq. (17) numerator is the squared distance between the i^{th} feature and the xið−Þ for the i^{th} characteristic, and x_i. The sample variances for the positive and negative examples are matched by the denominators of equations (16) and (17), respectively.

Eq. (18) yields the feature ranking (FR), with a higher score indicating a feature's importance.

$$FR(i) = \frac{R_{positive}(i) + R_{negative}(i)}{2} \quad (18)$$

technique is comparable to our method [19].In addition, it entails first determining the mean between the scores for each class and then individually calculating the scores for each class.The ranking of positive and negative samples for the ith feature is determined by Eqs. (16) and (17), respectively, given a training vector xk with k = 1,..., m and the number of positive and negative occurrences being n+ and n, respectively.

Algorithm 1 provides a brief summary of the process to calculate feature ranking for FSTBSVM.

The suggested feature rank varies from F-score in that it divides the sample variance of the relevant instance by the sum of the squared distances of positive and negative samples, while F-score divides the same by the total of the sample variance of the instances.

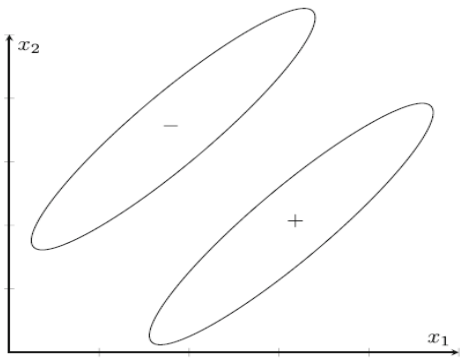

Fig 4 Bi-dimensional dataset

A bi-dimensional data collection is shown in Fig 4. There are two groups in this data set that can't be linearly separated by x1 and x2 alone.Given by (6) and (7), the F-rank sum has both denominators that are much greater than the respective numerators, as specified in (8); as a result, both characteristics have a small F-rank.An aslant line may be used to divide these groups, but the F-rank does not disclose any information about the qualities that are shared between them.However, it is just as easy to use and efficient as the F-score substitute despite this drawback.The algorithm's ability to examine positive and negative samples independently, with each grouping having a direct impact on the creation of the hyperplanes that TBSVM utilizes, as opposed to the F-score, which creates the hyperplanes for TBSVM using all samples, is one of its benefits.

4. Experimental result

In this part, we looked at four data sets from the UCI machine repository to assess the efficacy of feature selection paired with TBSVM.In order for the features to be situated in the [0, 1] range, the samples are normalized before machine learning.The suggested AFOA-TBSVM algorithm is assessed alongside the known approaches, including TWin Support Vector Machine (TWSVM) [17] and Feature Selection Method Combined with Twin-Bounded Support Vector Machine (FSTBSVM) [7].The performance parameters that are evaluated between proposed and current algorithms include things like accuracy, sensitivity, specificity, and execution time.

From the following link the hepatitis data is collected:

https://www.kaggle.com/datasets/codebreaker619/hepatitis-data.

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

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

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

The results of the performance comparisons are shown in Table 1.

Table 1. Results of performance comparison

Metrics and dataset	Methods		
	TWSVM	FATBSVM	AFOA-TBSVM
Accuracy - Diabetes	70.12	80.85	84.89
Accuracy-Fertility	71.24	90.90	93.93
Accuracy-Hepatitis	73.89	89.03	90.96
Accuracy-Statlog	73.33	85.54	87.64
Sensitivity -	66.23	81.13	85.20

Diabetes			
Sensitivity - Fertility	67.33	94.38	96.94
Sensitivity - Hepatitis	68.3	89,62	90.84
Sensitivity -Stat log	69.23	85.67	87.74
Specificity- Diabetes	65.30	80.62	84.84
Specificity- Fertility	65.33	81.82	84.55
Specificity - Hepatitis	62.30	82.62	85.84
Specificity - Stat log	64.39	70.31	85.40
F-measure -- Diabetes	65.89	80.12	84.14
F -measure - Fertility	65.23	85.93	89.45
F- measure - Hepatitis	62.56	85.84	87.79
F- measure -Stat log	64.67	85.49	87.57
Execution time- Diabetes (sec)	45	34	20
Execution time- Fertility(sec)	47	30	18
Execution time- Hepatitis(sec)	41	31	16
Execution time- Stat log(sec)	48	41	27

Accuracy

The entire actual classification parameters are calculated to determine accuracy, which is defined as the model's overall correctness ($T_p + T_n$) which may be separated based on the total number of criteria for categorization ($T_p + T_n + F_p + F_n$). This is how the accuracy is calculated:

$$\text{Accuracy} = \frac{T_p+T_n}{(T_p+T_n+F_p+F_n)} \quad (19)$$

In this case Tp would be a true positive, Tn would be a true negative, Fp would be a false positive, and Fn would be a false negative.

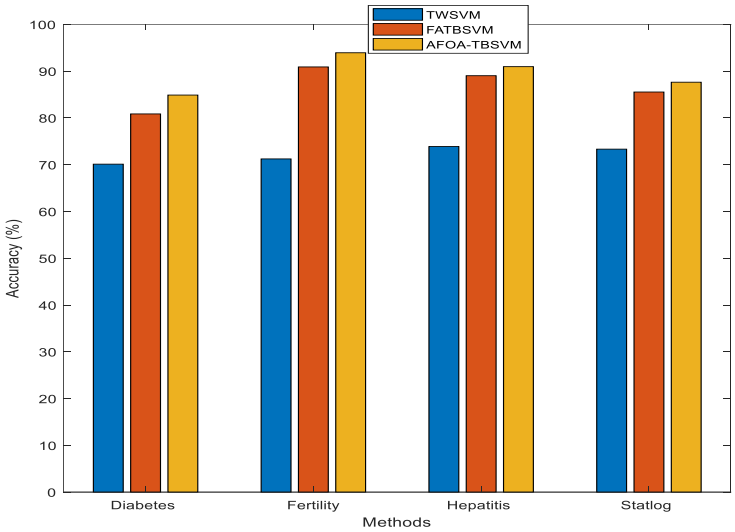

Fig 5 Accuracy

The comparison metric's correctness is assessed using both the current and new methods, as can be seen in the aforementioned Fig. 5.The information and techniques used for the x-axis are taken, and the accuracy value is shown on the y-axis.The suggested AFOA-TBSVM algorithm gives more accuracy for the supplied data than the current approaches, such as algorithms TWSVM and FATBSVM. Heart-Statlog, Pima, Hepatitis and Fertility datasets. The employment of a pre-processing technique results in greater classification accuracy via filling missing values and removal of noise. The conclusion drawn from the finding is that the AFOA-TBSVMalgorithm produces high accuracy results of 84.89% while the other existing TWSVM and FATBSVM methods produces 70.12% and 80.85% for the diabetes datasets.

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)

$$\text{Sensitivity} = \frac{T_p}{T_p + F_n} \quad (20)$$

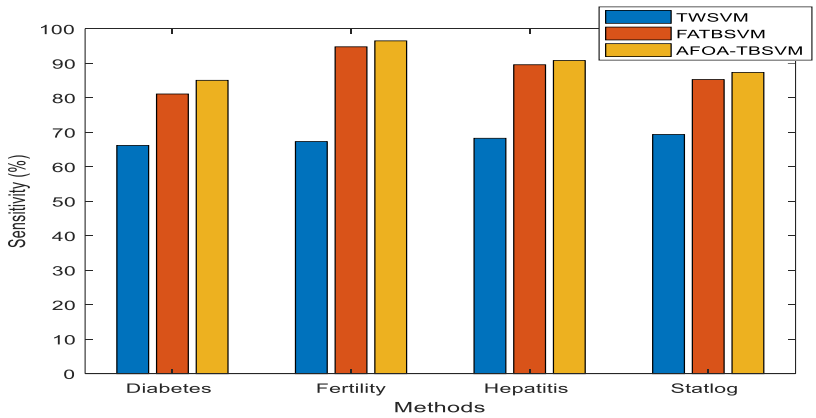

Fig 6 Sensitivity

The comparison metric's sensitivity is assessed using both the current and new methods, as can be seen in the aforementioned Fig 6.The approaches are used for the x-axis, and the sensitivity value is represented on the y-axis. The suggested AFOA-TBSVM algorithm gives more sensitivity for the supplied Pima, Heart-Statlog, Hepatitis, and Fertility datasets than the current approaches, such as TWSVM and FATBSVM algorithm, which provide lesser sensitivity.The proposed method improves the sensitivity through the selection of more relevant information. The conclusion drawn from the finding is that theAFOA-TBSVMalgorithm produces high recall results of 85.11% while the other existing TWSVM and FATBSVM produces 66.23% and 81.14% for the diabetes datasets.

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).

$$\text{Specificity} = \frac{T_n}{T_n + F_p} \quad (21)$$

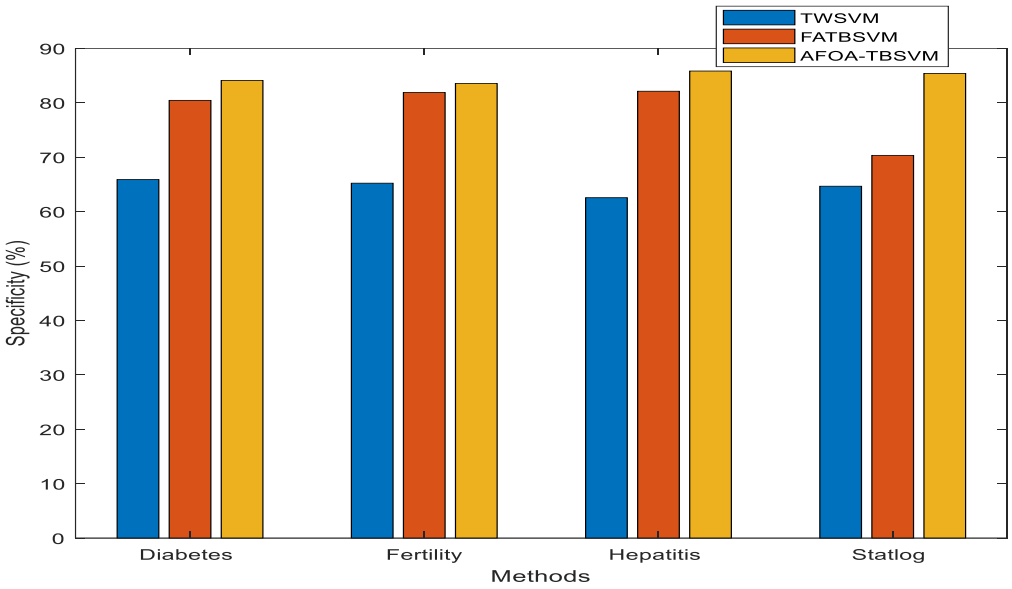

Fig 7 Specificity

From the above Fig 7, 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 and FATBSVM algorithm provides lower specificity whereas the proposed AFOA-TBSVM 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. AFOA-TBSVM increases the performance in the training stability. Thus the training process of datasets is much more stable. Thus the result concludes that the proposed AFOA-TBSVM algorithm increase the

classification performance through the optimal features.

Eq. (18) gives the feature ranking, with a higher score indicating a feature's importance.

F-measure

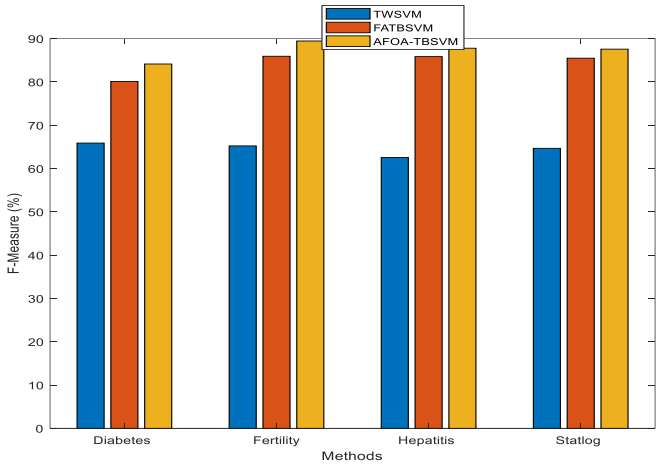

Fig 8 F-measure

From the Fig 8, the comparison values for F-measure metric using existing and proposed algorithm is evaluated. The existing TWSVM and FATBSVM for the provided medical datasets, approaches produce lower F-measure, whereas the suggested AFOA-TBSVM algorithm delivers greater F-measure.AFOA algorithm is used to provide optimal features. Hence the proposed algorithm provides higher accuracy of classification and better performance for given medical datasets.The conclusion drawn from the finding is that the AFOA-TBSVM algorithm produces high f measure results of 84.14% while the other existing TWSVM and FATBSVM produces 65.89% and 80.12% for the diabetes datasets.

Execution time

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

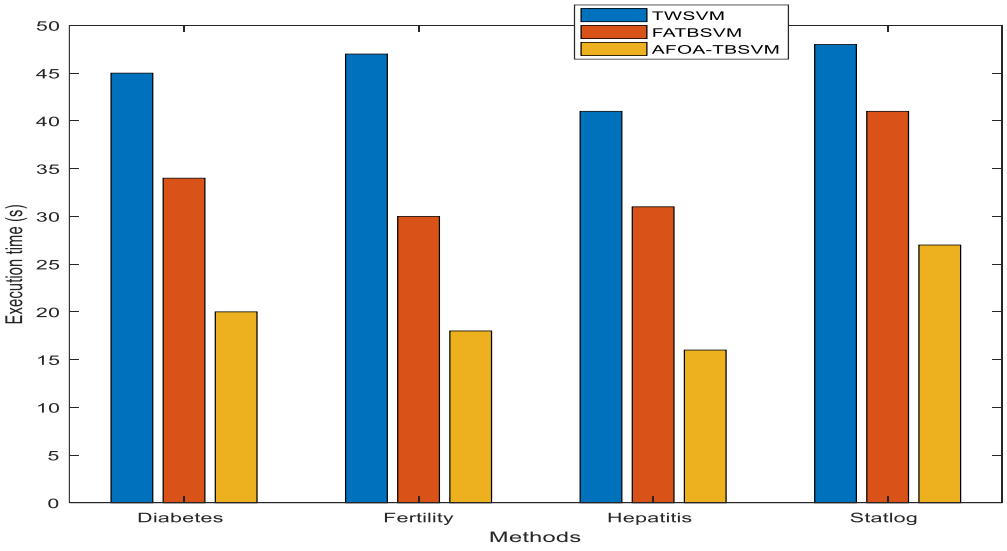

Fig 9 Execution time

From the above Fig 9, 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 and FATBSVM algorithms provide higher execution time whereas the proposed AFOA-TBSVM algorithm provides lower execution time for the given datasets. Thus the result concludes that the proposed AFOA-TBSVM algorithm increase the medical dataset performance through the optimized features

5. Conclusion

The AFOA-TBSVM technique is introduced in this study to enhance the performance of medical dataset categorization. This work contains four main modules such as pre-processing, feature selection and classification. By adding missing data and eliminating noise, the KMC algorithm improves classification performance.Then, feature selection is performed via AFOA which is used to select the most relevant and useful features.Finally, classification is done by using TBSVM algorithm which provides more accurate medical dataset classification performance. Based on the findings of the experiments, it was determined that the suggested AFOA-TBSVM algorithm offers superior accuracy, sensitivity, and specificity while simultaneously reducing the amount of time required for its execution in comparison to the algorithms that are currently in use. In future work, ensemble algorithms can be developed for the given datasets.

References

1. Izonin, Ivan, et al. "Towards Data Normalization Task for the Efficient Mining of Medical Data." *2022 12th International Conference on Advanced Computer Information Technologies (ACIT)*. IEEE, 2022.
2. Liu, Na, et al. "A novel ensemble learning paradigm for medical diagnosis with imbalanced data." *IEEE Access* 8 (2020): 171263-171280.
3. Yu, Yuhai, et al. "Deep transfer learning for modality classification of medical images." *Information* 8.3 (2017): 91.
4. Milovic, Boris, and Milan Milovic. "Prediction and decision making in health care using data mining." *Kuwait Chapter of the Arabian Journal of Business and Management Review* 1.12 (2012): 126.
5. Begum, S., Chakraborty, D., & Sarkar, R. (2015, December). Data classification using feature selection and kNN machine learning approach. In *2015 International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 811-814). IEEE.
6. Shen, Liming, et al. "Evolving support vector machines using fruit fly optimization for medical data classification." *Knowledge-Based Systems* 96 (2016): 61-75.
7. 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.

8. Wang, Pin, et al. "Automatic classification of breast cancer histopathological images based on deep feature fusion and enhanced routing" *Biomedical Signal Processing and Control* 65 (2021): 102341
9. Tomar, Divya, and Sonali Agarwal. "Feature selection based least square twin support vector machine for diagnosis of heart disease." *International Journal of Bio-Science and Bio-Technology* 6.2 (2014): 69-82.
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, SapiahBinti, NurainiBinti Abdul Rashid, and Zuhaira Muhammad Zain. "Particle swarm optimization feature selection for breast cancer recurrence prediction." *IEEE Access* 6 (2018): 29637-29647.
12. Carrera, Enrique V., Andrés González, and Ricardo Carrera. "Automated detection of diabetic retinopathy using SVM." *2017 IEEE XXIV international conference on electronics, electrical engineering and computing (INTERCON)*. IEEE, 2017.
13. Ajam, N. (2015). Heart diseases diagnoses using artificial neural network. *IISTE Network and Complex Systems*, *5*(4).
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. Liu, Jingsen, et al. "A dynamic adaptive firefly algorithm with globally orientation." *Mathematics and Computers in Simulation* 174 (2020): 76-101.
16. 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
17. Chandra, Suresh, and R. Khemchandani. "Twin support vector machines for pattern classification." *IEEE Trans. Pattern Anal. Mach. Intell* 29.5 (2007): 905-910.
18. Wen, Zaiwen, Wotao Yin, and Yin Zhang. "Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm." *Mathematical Programming Computation* 4.4 (2012): 333-361.
19. Chen, Yi-Wei, and Chih-Jen Lin. "Combining SVMs with various feature selection strategies." *Feature extraction*. Springer, Berlin, Heidelberg, 2006. 315-324.