IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 35, No. 6, December, 2021, pp. 447-456
Journal homepage: http://iieta.org/journals/ria

Firefly Optimization Based Noise Additive Privacy-Preserving Data Classification Technique to Predict Chronic Kidney Disease

Preet Kamal Kaur*, Kanwal Preet Singh Attwal, Harmandeep Singh

Department of Computer Science and Engineering, Punjabi University, Patiala, Punjab 147002, India

Corresponding Author Email: phd_preet@csepup.ac.in

https://doi.org/10.18280/ria.350602

Received: 1 June 2021
Accepted: 20 October 2021

Keywords:
chronic kidney disease, data perturbation, firefly optimization algorithm, privacy-preserving data classification, random forest

ABSTRACT

With the continuous advancements in Information and Communication Technology, healthcare data is stored in the electronic forms and accessed remotely according to the requirements. However, there is a negative impact like unauthorized access, misuse, stealing of the data, which violates the privacy concern of patients. Sensitive information, if not protected, can become the basis for linkage attacks. Paper proposes an improved Privacy-Preserving Data Classification System for *Chronic Kidney Disease* dataset. Focus of the work is to predict the disease of patients' while preventing the privacy breach of their sensitive information. To accomplish this goal, a metaheuristic Firefly Optimization Algorithm (FOA) is deployed for random noise generation (instead of fixed noise) and this noise is added to the least significant bits of sensitive data. Then, random forest classifier is applied on both original and perturbed dataset to predict the disease. Even after perturbation, technique preserves required significance of prediction results by maintaining the balance between utility and security of data. In order to validate the results, proposed method is compared with the existing technology on the basis of various evaluation parameters. Results show that proposed technique is suitable for healthcare applications where both privacy protection and accurate prediction are necessary conditions.

1. INTRODUCTION

Data mining has played very important role in the development of healthcare sector. Mining helps to take out the essential outcomes which further support decision making process and also, allow creation of necessary plans or policies needed for trouble-free functioning of healthcare functions and operations. Mining enables gaining insights into various diseases, predicting the diseases of patients at earlier stage and also assisting doctors in providing timely treatment to patients. According to this, it can be perceived that health related data poses a high value for researchers as well as for healthcare workers. Through access to this information, very good relation can be maintained between patient and clinician (doctor) enabling effective and correct conduction of healthcare practices. But at the same time, patient's data may contain sensitive attributes which should not be disclosed such as Aadhar number, age, address, hospital visits, lab results, some sensitive disease or its cause etc. So, issue of privacy in data mining needs to be addressed and confidentiality of data should be maintained while accessing or analyzing it. To fetch required information from data, discover unknown patterns from it and also, to prevent sensitive information from disclosure, mining techniques are combined with privacy preservation approaches giving rise to the field of Privacy Preserving Data Mining (PPDM). Original data of patients is transformed in such a way that adversaries can't access it and still, it is useful enough to take out the required outcomes.

In the current scenario, healthcare organization uses information and communication technology to monitor the patients [1]. Patient's data is collected from the Electronic Health Records (EHR), sensors, Radio Frequency Identification Tag (RFID) that helps in decision making and improves the quality of service [2]. Records contain the patient's personal information (such as identity data) along with medical details. Thus, illegal access to the records affects the privacy of patients [3]. To safeguard the sensitive data from leakage, various data disturbance approaches can be utilized such as anonymization, cryptography or perturbation [4]. Anonymization method makes the individual data indistinguishable using suppression and generalization techniques. In the cryptographic technique, sensitive information is encrypted using a secret key and encryption algorithm. Either these methods don't provide the required privacy or they bring in the overhead of encryption-decryption and are less efficient for the larger data sets as data utility poses great concern [5]. Data perturbation method is more popular due to simplicity and its advantage to treat different attributes independently. Data perturbation techniques transform sensitive information with an advantage of maintaining the utility of data [6]. Thus, perturbed data can be employed for research and analysis without breaking patient privacy [7]. On the other side, patient's data is used for disease prediction but due to large amount of data processing, it manually consumes a lot of time. Thus, various machine learning classifiers are applied to predict the disease in an automated and intelligent manner.

To overcome these challenges, in this paper, focus is towards developing a Privacy Preserving Data Classification Technique for prediction of Chronic Kidney Disease. In the proposed technique, meta-heuristic Firefly Optimization Algorithm has been applied to generate random noise. Noise

is added in the least significant bits of the sensitive data to provide privacy. On the other side, random forest technique is used to predict the disease and results are analyzed for both original and perturbed datasets.

Rest of the paper is as follows. Section 2 outlines existing work done in the field. Section 3 illustrates systematic flow and methodology of the proposed technique. Section 4 represents implementation results of the technique. Different performance metrics used for its evaluation as well as for its validation are discussed. Section 5 presents the conclusion and future scope of this research.

2. RELATED WORK

In section 2.1 and 2.2, various data perturbation techniques and different classifiers are studied that have commonly been used in the domain of healthcare. Inferences and challenges drawn from the literature are elaborated in section 2.3.

2.1 Data perturbation techniques

Data perturbation method transforms the sensitive data before publishing in such a way that privacy of an individual is preserved while maintaining the important data properties. Different ways studied to perform perturbation of data are shown below:

2.1.1 Noise addition

In this technique, random noise is added in the sensitive data using Eq. (1).

$$P = S + N \tag{1}$$

where, P denotes the perturbed data, S and N define sensitive data and noise respectively.

2.1.2 Noise multiplication

In this technique, random noise is multiplied with the sensitive data using Eq. (2).

$$P = SXN \tag{2}$$

where, P denotes the perturbed data, S and N define the sensitive data and noise respectively.

2.1.3 Min-max normalization

In this technique, sensitive data is linearly transformed and normalized using Eq. (3).

$$v' = \frac{v - \min(A)}{\max(A) - \min(A)}\left((new_{\max(A)} - new_{\min(A)}) + new_{\min(A)}\right) \tag{3}$$

where, A denotes the sensitive attribute, max and min represent the maximum value and minimum value in the attribute, new_{max} and new_{min} denote the new boundary value range for A.

2.1.4 Micro-aggregation

It is a technique used to satisfy the k-anonymity constraint in dataset. In k-anonymization, whole dataset is converted into different groups where each group holds at least k records such that any record in the group can't be identified from k-1 other records in it. Micro-aggregation helps to replace all the values in group with the centroid value (arithmetic mean in case of numerical attributes or it can be some other value based on the value range of an attribute). Aim should be to achieve the k-anonymous dataset while lessening the degradation of quality of data [8].

2.1.5 Data swapping

In this process, values of sensitive attributes are swapped among different records so as to increase the uncertainty in data which further provides privacy. It maintains the statistical properties of the database i.e. different knowledge discovery tasks can be performed on this transformed data.

Existing methods studied in the field of data perturbation are briefly discussed below:

Kiran and Vasumathi [9] applied min-max normalization technique on original dataset (Matrix M) to distort it and take all the values into the specific range of 0 and 1. Obtained distorted data matrix M' is not same as real matrix. Then, it is multiplied with a negative number (shifting vector) for more security [10]. Technique is applied on 4 real-world datasets and evaluation is done by calculating values of privacy and accuracy parameters for NBTree classfier.

Kalaivani and Chidambaram [11] proposed multilevel trust based PPDM technique in which original data is changed before it gets published. Data is perturbed and made available to various data miners according to trust level maintained by them i.e. different data miners have differently perturbed copies of similar data. Still, if all the data placed with them can be joined and reconstructed in some way, then, it can give rise to diversity attack which reveals extra information about users. To protect the privacy, Gaussian noise is added randomly to the original data [12].

Jahan et al. [13] developed multiplicative perturbation method to change the data before publishing it to the data analyst for performing data mining functions. In multiplicative perturbation, combination of fuzzy logic and random rotation works. Different multivariate datasets downloaded from UCI machine learning repository are used as input. For confidential information, fuzzy logic is computed and then, original dataset is multiplied with random rotation matrix. Random rotation preserves distance measures which further helps to maintain the utility of data. K-means clustering is used to analyze the results obtained with both original and perturbed datasets.

Balasubramaniam and Kavitha [14] used geometrical data perturbation (GDP) technique to preserve the privacy of personal health records. Different type of information about patients is stored in different tables. Any of the table and from it, any number of columns can be chosen for perturbation. While applying geometric perturbation, firstly, original dataset is converted into matrix form. As geometric perturbation can only be applied to numerical data values, ASCII values are calculated for other type of data and stored in the matrix. Random matrix is created with values in the range of 1 to 9 and then, it is rotated in clockwise direction vertically. This random rotation matrix is multiplied with original data. Gaussian noise is also generated (range from 0 to 1). Finally, addition operation is performed between initially calculated product, transpose of rotation matrix and Gaussian noise [15]. Perturbed data is now outsourced to cloud where data retrieval and query processing takes place. This technique is compared with existing Advanced Encryption Standard (AES) method on the basis of time.

2.2 Data classification

Classification is a process of categorizing data into specific classes. Classifier is trained with the help of training data and various classification rules are discovered. These rules are further used to classify unknown data/tuples for obtaining important results in different domains. In this section, various classifiers used for disease prediction are explained.

2.2.1 Decision Tree (DT) Algorithm

Decision tree is a structure similar to flow chart which is trained with the help of labelled tuples. Root (R) of a tree contains whole population which has to be further divided. Internal nodes (I) represent test that has to be conducted on a particular attribute of data, each branch (B) comes out as outcome of the test and leaves (L) give different possible solutions (class labels). While traversing from root node to leaf node of a tree, set of classification rules are obtained which can now be used to predict the class label of an unknown tuple. The algorithm belongs to supervised learning field and applied in a loop to each child node until all samples at the node are of one same class [16, 17]. To select an attribute which has to be considered for splitting the node while constructing tree, information gain is used as a measure. Selected attribute needs to minimize the information required for classification of tuples and tree should be as simple as it can be. Chaurasia et al. [18] applied decision tree algorithm on Chronic Kidney Disease dataset and achieved 93% correct predictions. Author focused on identifying the key attributes that play major role to decide whether person is suffering from kidney disease or not.

2.2.2 K-Nearest Neighbor (KNN) Algorithm

K-Nearest Neighbor is also a supervised method of classification which works using the concept of analogy i.e. comparison is done between the new test tuple and similar/alike training tuples. Each training tuple contains value for every attribute in dataset which means to define a particular tuple, n attributes are needed. Training tuples collectively form pattern space of n-dimensions. Whenever any new data tuple comes, classifier searches this pattern space to find k tuples from the training set that are nearest to the new tuple. Mostly, Euclidean distance is used as a measure to find the closeness between tuples [16, 19]. From the 'k' identified neighbors (training tuples), majority class label is selected and assigned to new (test) tuple. If value of 'k' is 1 i.e. there is only one neighbor, then, label of that neighbor (tuple) is given to test tuple. Tikariha and Richhariya [20] used KNN classifier for prediction of Chronic Kidney disease and compared it with the results obtained by Support Vector Machine for same dataset.

2.2.3 Support vector machine

Support Vector Machine is a classification algorithm used to classify the points of data in n-dimensional space. It can be useful to classify both linear and non-linear data by finding the required hyper-plane. Hyper-plane is used to separate the classes like a decision boundary. It is found with the help of support vectors and number of planes depends upon number of features in data. When there is more than one hyper-plane, aim is to find the plane with maximum margin i.e. distance from nearest element of either class to the hyper-plane should be largest. This property makes SVM robust and reduces misclassification errors. SVM also has a capability to ignore outliers [16, 20].

Linear SVM: If straight lines (hyper-plane) can be drawn to classify the data, then dataset is said to be linearly separable. Suppose, if there are 2 classes, goal is to select the one having largest margin. Data belonging to either side of line represents different classes/categories present in dataset.

Non Linear SVM: If no straight line can be drawn to classify the dataset, then, data is said to be linearly inseparable. Kernel function is used to change low dimensional space of input to higher dimensional space i.e. inseparable problems are converted into separable problems. Non-linear decision boundary (such as concentric circles) is used to separate the classes.

2.3 Inferences drawn from the literature

- As healthcare data attributes such as blood pressure, sugar, age etc. play major role in deciding the disease and its outcomes, it is very necessary to maintain the utility of data. If data is changed to very large extent, then it will not be useful enough for the required purpose. Therefore, balance has to be maintained between usefulness of data and security of the patient.
- It is observed from the study that most popular techniques for data perturbation are geometric transformation, Gaussian noise, K-Means Clustering and Min-Max normalization [9-15].
- In geometric perturbation, sensitive data is transformed by performing various geometric functions such as rotation, translation, scaling or projection. They can't process very high volumes of data efficiently, e.g. random rotation consumes a considerable amount of time to provide better results while enforcing privacy.
- Gaussian noise addition depends on original data i.e. mean and standard deviation is drawn from the given input data for noise generation. In K- Means Clustering technique, sensitive data is changed into different clusters. Internal to each cluster, there is similar type of information regarding an attribute. For example, in the healthcare data, patients are grouped into different clusters based on age attribute. Main limitation of K-Means Clustering technique is that it forms many clusters if there is high variability in the attributes.
- Min-max normalization technique modifies the sensitive data and takes it between the specified ranges. However, this transformation is constant and breaks the relationship between sensitive and non-sensitive data. Besides, it also provides lesser accuracy.
- On the other side, various machine learning techniques are being used to classify the healthcare data. Support vector machine (SVM), K-nearest neighbor, and decision tree are mostly used for classification [16, 20].
- Out of these, the decision tree algorithm is the most preferred classifier to predict the disease but it also has a large number of limitations such as over-fitting and no global solution [18, 21, 22]. Small change in the input data can lead to significant changes in the decision tree.
- In K-NN algorithm, if chosen value of 'k' is incorrect, it leads to over-fitting or under-fitting of data to the model. Moreover, classifier doesn't perform well on unbalanced data (e.g. if instances of some class is more than the other, biased results may be produced). K-NN can't assume anything about distribution or discriminative functions from the training data and completely relies on memorizing all of the training instances. Generalization of data is done only when prediction query is submitted to the algorithm [16, 20, 23].

- In SVM, while dealing with non-linear data, kernel function needs to be selected. Matrix of kernel grows in quadratic manner with increase in size of training dataset. If data is high dimensional, many support vectors are generated. Due to this, high memory is required and training time is also increased. Also, interpretation of final SVM model is difficult for humans [16, 20, 24].

To deal with above said limitations, firefly algorithm is deployed for data perturbation in this research and instead of fixed noise, random noise is generated. To sustain the utility of healthcare data, noise is added to the least significant bit of sensitive attribute. Random Forest is going to be applied as a classification method in the work. It is an ensemble learning technique which has a potential to deal with large amounts of data. When combination of decision trees works together, prediction accuracy improves. Subset of features (not all) is considered for splitting the node at each level of tree and significance of each feature can also be explained by this classifier. Detailed functioning of Firefly algorithm and Random Forest for Chronic Kidney Disease prediction is elaborated in the subsequent section.

3. PROPOSED TECHNIQUE

Proposed technique adds random noise to the confidential information (which can later become medium for linkage attacks when combined with other databases) and precisely predicts the disease of patients. Disease prediction is done by applying random forest classifier initially on original dataset and then, on perturbed dataset. Performance is examined on the basis of evaluation parameters discussed in next section. Also, proposed technique is compared with the existing techniques related to this domain. Flowchart of the proposed technique is shown in Figure 1.

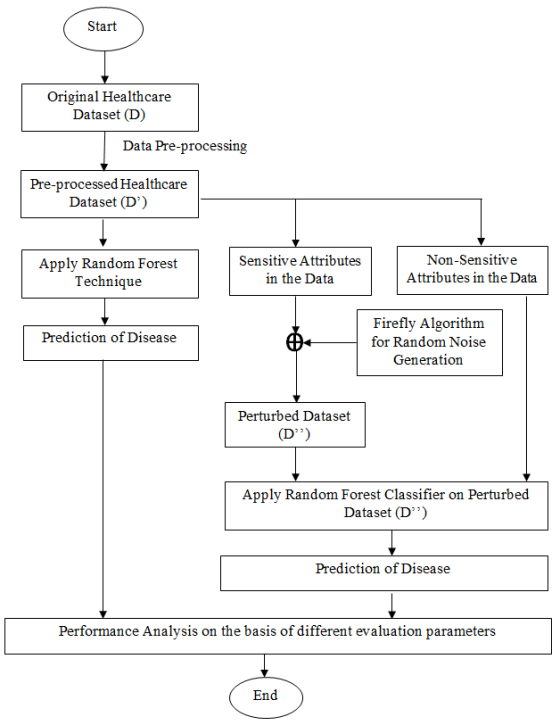

Figure 1. Flow chart for the proposed technique

- **Original Healthcare Dataset:** Chronic_kidney_disease healthcare dataset is downloaded from the UC Irvine Machine Learning Repository. The dataset contains total 25 attributes (24 other attributes + 1 class attribute) [25].
- **Data Pre-processing:** Chronic kidney dataset is pre-processed i.e. cleaned and formatted so that it can be utilised for prediction or analysis purpose.
- **Identification of sensitive and normal attributes:** Now, in the dataset, sensitive attributes violating privacy of patients are identified. Noise is to be added in these sensitive attributes and other normal attributes remain as it is.
- **Data Perturbation Technique:** In data perturbation, data is transformed with an addition of noise in it. In the proposed technique, random noise is generated based on Firefly Algorithm and added to the sensitive attributes of dataset so that privacy of patients is not breached.
- **Classifier:** Random Forest Data Classifier is used to predict the disease of patient by randomly picking samples for tree construction, using its sensitive as well as non-sensitive attributes. This classifier is applied on both modified as well as the original dataset.
- **Performance Analysis:** In this section, change in dataset after noise addition and variability in prediction results by applying classifier on modified dataset and original dataset is measured in terms of Accuracy, Precision, Recall, F-score, Mean Square error, Peak Signal-to-Noise Ratio, Structural Similarity Index, Execution time.

3.1 Firefly Optimization Algorithm and how it works

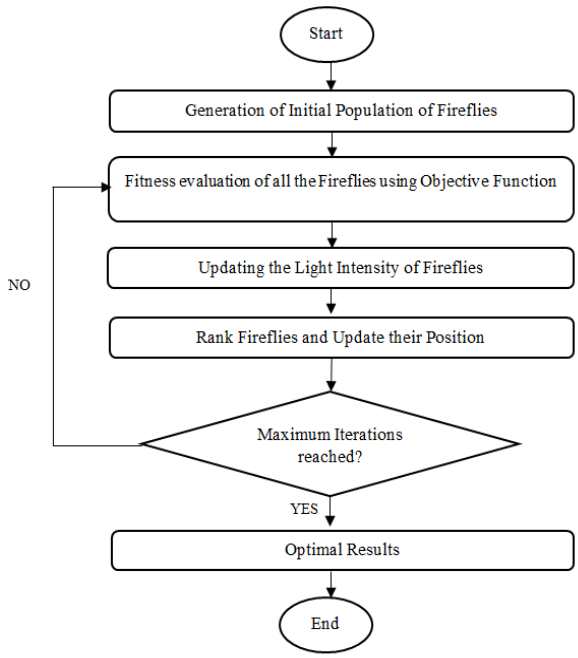

Figure 2. Flow chart for firefly optimization algorithm

Firefly algorithm is based on the fireflies that produce light while flying in the night [26]. The light is produced from the lower stomach of the firefly known as bioluminescence. This light is used by firefly to move or attract their mates or prey. This mechanism was formulated by author Yang and defined the following postulates:

- Fireflies are unisexual. Therefore, they can attract fireflies irrespective of their sex.
- The attractiveness between the fireflies is directly proportional to the light intensity. Therefore, low-intensity

firefly is attracted by high-intensity firefly.

- If the light brightness between two fireflies is equal, then movement of fireflies is random.

The generations of novel solutions are performed on the basis of two components, namely random walk and glow of the lightning bugs. Light of the fireflies becomes the fundamental factor that should be linked up with the objective determination of the related problem. Initially, a random population of fireflies is generated. After that, fitness evaluation is done for all fireflies based on the objective function. Objective function helps in finding the optimal solution. Then, updating of fireflies' light intensity is there. Again, they are ranked and their positions are updated. Whole procedure is iterated a fixed number of times and optimal results are determined [27].

Updated position of firefly is determined using Eq. (4).

$$X' = X + \beta e^{-\gamma r^2}(Y - X) + \alpha \in \quad (4)$$

whereas, X' denotes the new position of a firefly, X, Y denotes the two fireflies in which attractiveness is to be calculated. r denotes the distance between the fireflies. β represents the attractiveness factor, γ defines the light absorption factor, $\alpha \in$ defines the random deviation factor. Flowchart of the firefly algorithm is shown in Figure 2.

3.2 Deployment of Firefly Algorithm for noise generation

In the proposed work, to secure the sensitive attribute, noise is added in it. Given below is the explanation of generating random noise using Firefly Algorithm:

Step 1: First of all, initial population of fireflies is defined along with number of iterations, light absorption factor and random deviation factor.

Step 2: Next, initial random positions of fireflies are determined.

Step 3: After that, the Euclidean distance between the fireflies is calculated.

Step 4: Two arguments are passed for fitness function to work: positions of fireflies and sensitive attribute of dataset. Sensitive attribute is read and converted into 8-bit binary number. 2 Least significant bits of Firefly's position are XORed with the LSB 2-bits of the sensitive attribute and fitness function is evaluated using the Mean Square Error (MSE). The reason behind taking MSE as an objective function is to generate the maximum difference between original and noisy attributes.

Step 5: Again, position of fireflies are updated based on the Eq. (4) and this whole process is repeated for fixed number of times.

Step 6: Out of all fireflies, firefly which gives maximum MSE is selected for providing optimal solution and its position is taken as a key to perturb the sensitive data.

3.3 Random Forest Classifier

Random Forest Classifier is based on ensemble learning. In this learning, many algorithms or classifiers that may be of same type or of different type are combined together to give the required decision. Random Forest technique works using the collection of decision trees where trees are built by randomly selected subsets from the training data. All the decision trees provide their prediction. After this, voting is performed and prediction result with highest votes is selected. Basic concept of Random Forest Classifier is shown in Figure 3:

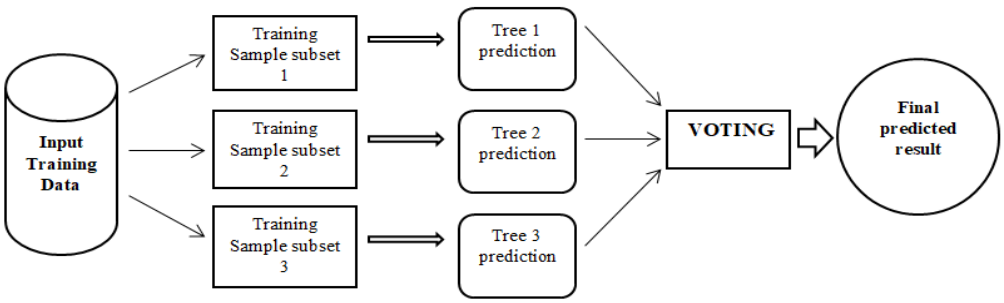

Figure 3. Basic working scheme of Random Forest Classifier

Random Forest Classifier provides higher prediction accuracy because rather than single model, union of models work in combination. It is based on the concept of bagging but actually is an extension of it. Bagging is basically bootstrap aggregation in which all features have to be taken under consideration for node splitting while making a tree whereas in random forest, subgroup of all the features is selected randomly and from this subset, best feature is used to split the node. It helps to improve performance and reduce variance from the results. For example, if there are 30 features, the random forest will be using only sure number of these features in each model, let's say five. 25 features are missed that would be effective. However, as per statement, decision trees are a part of random forest. So, in each tree, only five random features are used. But, if the number of trees in the forest is increased, then all or maximum features can be used. Therefore, error because of bias and error rate due to variance can be reduced by the usage of maximum features [28-30]. Therefore, it can be concluded that random forest based on ensemble learning is more powerful in comparison to a single decision tree so as to reduce bias error and limit over-fitting.

3.4 Random Forest Classifier for Chronic Kidney Disease

Globally, Chronic Kidney Disease (CKD) is becoming a common health issue and 10% of the world's population is generally affected. How to analyze CKD by systematic and automatic ways has few direct testimonials. So, how the machine learning (ML) method is used to diagnose CKD is highlighted in this paper. Abnormalities in various physiological data can be detected by ML algorithms with immense achievement. Research presents that random forest (RF) classifier attains the near-optimal performances for the identification of CKD subjects. Outcomes are perceptible and

collectively discussed. Thus, ML algorithms play an important function in CKD diagnosis with adequate durability and RF can be used for diagnosis of this kind of disease as proposed by our research [29].

4. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, various results are shown that have been obtained through experimentation performed on the proposed technique to validate it against the existing techniques. Standard healthcare *chronic kidney disease* dataset is taken from the UCI Machine Learning Repository [18]. This dataset contains 25 attributes and is collected from the hospital for nearly two months of the period to predict chronic kidney disease. Attributes in the dataset are: age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anemia and class. Simulations are performed in MATLAB 2017. System configurations are i5 processor, 8GB RAM and 1.80GHz operating frequency.

4.1 Performance metrics

- **Mean Square Error:** Mean Square Error is computed by averaging the squared intensity of the original values and perturbed values (after noise addition) for an attribute [31]. It is calculated using Eq. (5) given below:

$$MSE = \frac{\sum_{i=1}^{length(attribute)} (Original_{attribute_{value}} - Perturbed_{attribute\_value})^2}{length(attribute)} \quad (5)$$

- **Peak Signal-to-Noise Ratio (PSNR):** PSNR measures how much noise is added in the attribute and it is calculated using Eq. (6) [32].

$$PSNR = 10 \log_{10} \frac{Peak^2}{MSE} \quad (6)$$

whereas, peak denotes the maximum value that can be represented for the particular attribute. If the class of an attribute is 8-bit long, peak value will be 255.

- **Structural Similarity Index (SSIM):** Structural Similarity Index tells about the similarity between original dataset and perturbed dataset. In the proposed technique, this measure is used to find the similarity between initial dataset and dataset obtained after addition of noise in the sensitive attribute. Inbuilt command 'ssim' of MATLAB is used for the calculation of this parameter.
- **Accuracy:** Accuracy is one of the most important metrics for evaluating the classification models. It is the fraction of predictions in which our model got right i.e. when classifier gave correct results. It is calculated using Eq. (7) [16].

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positve + True\ Negative + False\ Negative} \quad (7)$$

whereas,

True Positive: positive class correctly predicted
False Positive: positive class incorrectly predicted
True Negative: Negative class correctly predicted
False Negative: Negative class incorrectly predicted

- **Precision:** Precision is defined as the ratio of number of observations which are correctly predicted as positive to the total number of observations that are predicted positive [16, 33].

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positve} \quad (8)$$

- **Recall:** Recall is defined as percentage of observations of a positive class that have actually been predicted as belonging to that class [33].

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (9)$$

- **F-score:** F-score is defined as the harmonic mean of precision and recall [33].

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (10)$$

- **Execution Time:** Total time taken by the proposed technique to obtain the required results. In MATLAB, tic and toc commands are used to determine the total execution time.

Results obtained after each step of experimentation are shown below:

Step 1: Firstly, pre-processing of the dataset is done in order to deal with missing and inconsistent values. Irrelevant information is removed from it.

Step 2: Then, dataset is divided in the proportion of 70:30 where 70% denotes the training data and remaining 30% is for test data.

Step 3: After that, random forest technique is applied on the training data for disease prediction. Accuracy, precision, recall and F-score of classifier are measured.

Step 4: Next, the dataset is divided into two parts, sensitive and non-sensitive attributes. In the *chronic kidney disease* dataset, according to the proposed work, age attribute comes under sensitive and remaining attributes come under non-sensitive attributes. Age of patient when leaked, can give rise to linkage attacks i.e. combining this dataset with some other dataset having common attribute can lead to revelation of individual's identity which is violation of research limits and privacy requirements.

Step 5: Integer value of age (sensitive attribute) is transformed into 8-bit binary value and its least significant (LSB) 2-bits are extracted.

Step 6: Firefly algorithm is used for noise generation. Fitness evaluation is done using the Mean Square Error and position of firefly which leads to maximum MSE, is taken as a key to add noise in the age attribute. 2 LSBs of the key are XORed with 2 LSBs of age (converted into binary) to preserve the privacy of the patient without negatively impacting the prediction of the disease.

Initialization of various parameters for Firefly algorithm is shown in Table 1.

Table 1. Firefly Algorithm Parameters

Parameter	Value
Population Size	5
Iterations	100
Alpha (Randomization Parameter)	0.5
Gamma (Light Absorption Coefficient)	1

Step 7: Sensitive attribute (after addition of noise in it) is concatenated with the non-sensitive attributes and random forest classifier is applied to the perturbed dataset.

Step 8: Parameters such as accuracy, precision, recall and F-score are again calculated. Also, PSNR (Peak Signal to Noise Ratio), SSIM (Structural Similarity Index) and MSE (Mean Square Error) are measured.

Step 9: Performance analysis of the firefly algorithm for noise generation and random forest technique for disease prediction is done using various parameters. Comparative analysis is done for validation of the proposed technique against existing techniques.

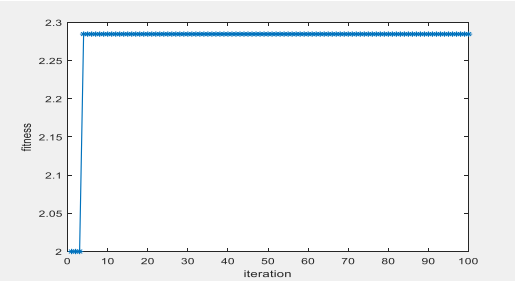

Figure 4. Fitness function to achieve high MSE

Fitness function for the Firefly algorithm to achieve high MSE is shown in Figure 4.

Experimental values of the evaluation parameters calculated for the proposed technique are shown in Tables 2 and 3. The results show that the proposed technique achieves good classification results while preserving privacy of patients.

Table 2. Calculation of various performance parameters for proposed technique

	MSE	PSNR (in dB)	SSIM	Execution Time (in seconds)
Proposed Technique	0.2175	54.756	0.99984	5.102601

Table 3. Comparison of classification results for original dataset and perturbed dataset

Parameter \ Dataset	Original Dataset	Perturbed Dataset
Accuracy Percentage (%)	97	95
Recall	0.97015	0.95522
Precision	0.98485	0.9697
F-score	0.97744	0.96241

Random Forest classifier is applied firstly on Original Dataset (without addition of noise in sensitive attribute) and then, on perturbed dataset (with addition of noise in sensitive attribute using Firefly Algorithm). Comparison of classification results on the basis of Recall, Precision and F-Score is represented in Figure 5. Accuracy comparison by applying classifier on Original and Perturbed dataset is depicted in Figure 6.

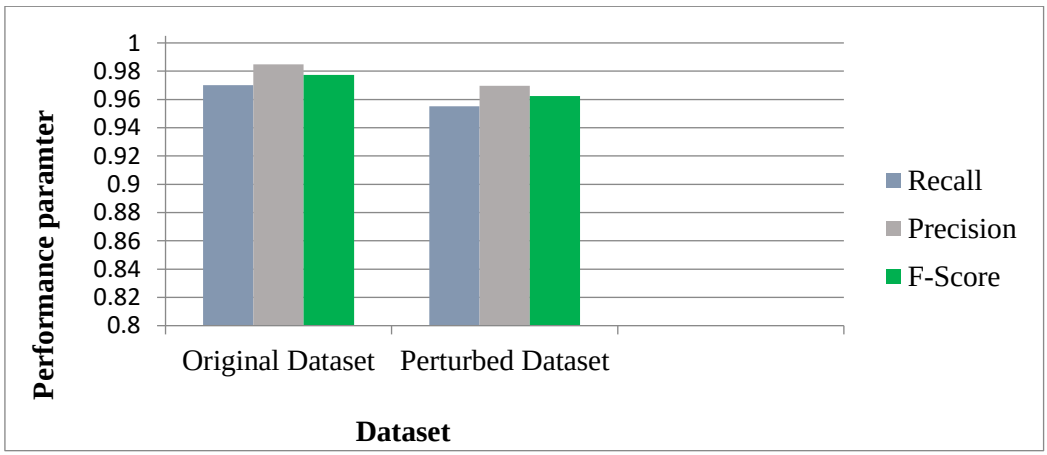

Figure 5. Comparison of classification results for original and perturbed dataset on the basis of recall, precision and F-score

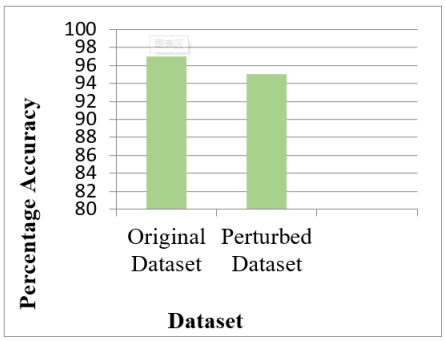

Figure 6. Comparison of classification accuracy for original and perturbed dataset

4.2 Comparative analysis

Proposed technique adds noise in the sensitive attribute using the firefly algorithm and predicts the accuracy using the random forest technique. Comparison of results for original and perturbed dataset is discussed above in detail. In this section, the experimental results of the proposed technique are compared with the existing techniques. Firefly algorithm is compared with the existing noise addition techniques on the basis of MSE, PSNR and SSIM and accuracy of random forest technique (when combined to privacy preservation) is compared with the existing classification techniques for disease prediction. Table 4 represents the comparison of noise addition methods.

Mean Square Error of Firefly algorithm is higher while Peak Signal-Noise Ratio and Structural Similarity Index are lower than Gaussian noise addition and min-max normalization noise addition methods. In the proposed technique, MSE is used as an objective function and its value is maximized. Highest value of MSE is chosen so that difference between original age and perturbed age becomes maximum which is required for the work. If there is more variation in data, more security can be provided for sensitive information of patients. Parallel to data perturbation, it has been assured that data is changed only up to the extent where its utility can be maintained for providing accurate results i.e. proposed work also takes care of the intrinsic balance management between privacy and usefulness of data.

Various classification techniques such as Support Vector Machine, K-Nearest neighbor and Decision Tree algorithm [18, 20] have mostly been used to predict the chronic kidney disease. In these existing methods, privacy concern has not been taken into consideration. In the proposed work, after noise addition in data for privacy preservation, ensemble learning (Random Forest) is applied. Classification accuracy of these above mentioned existing techniques are compared with that of proposed technique. Results are discussed in tabular form (Table 5).

Comparison of the proposed technique (Random Forest with Privacy Preservation) with different classification methods is shown below in graphical manner (Figure 7).

As presented above, proposed technique (using Firefly Algorithm and Random Forest Classifier) gives better accuracy than SVM, KNN and Decision tree classifiers [18, 20] even after perturbation of dataset by noise addition. It can be concluded that this technique provides required privacy to the patients while exhibiting the significant results for disease prediction.

Table 4. Comparison of Firefly Algorithm with the existing noise addition techniques

Noise Addition Technique	MSE	PSNR (in dB)	SSIM
Gaussian Noise	0.038541	62.272	0.99997
Min-Max Normalization	0.014806	66.426	0.99999
Firefly Algorithm	0.2175	54.756	0.99984

Table 5. Accuracy comparison of the proposed technique with the existing classification techniques

Classification Technique	Accuracy
Support Vector Machine	73.75%
K-Nearest Neighbor	78.75%
Decision Tree Algorithm	93%
Random Forest Technique with Privacy Preservation (Proposed)	95%

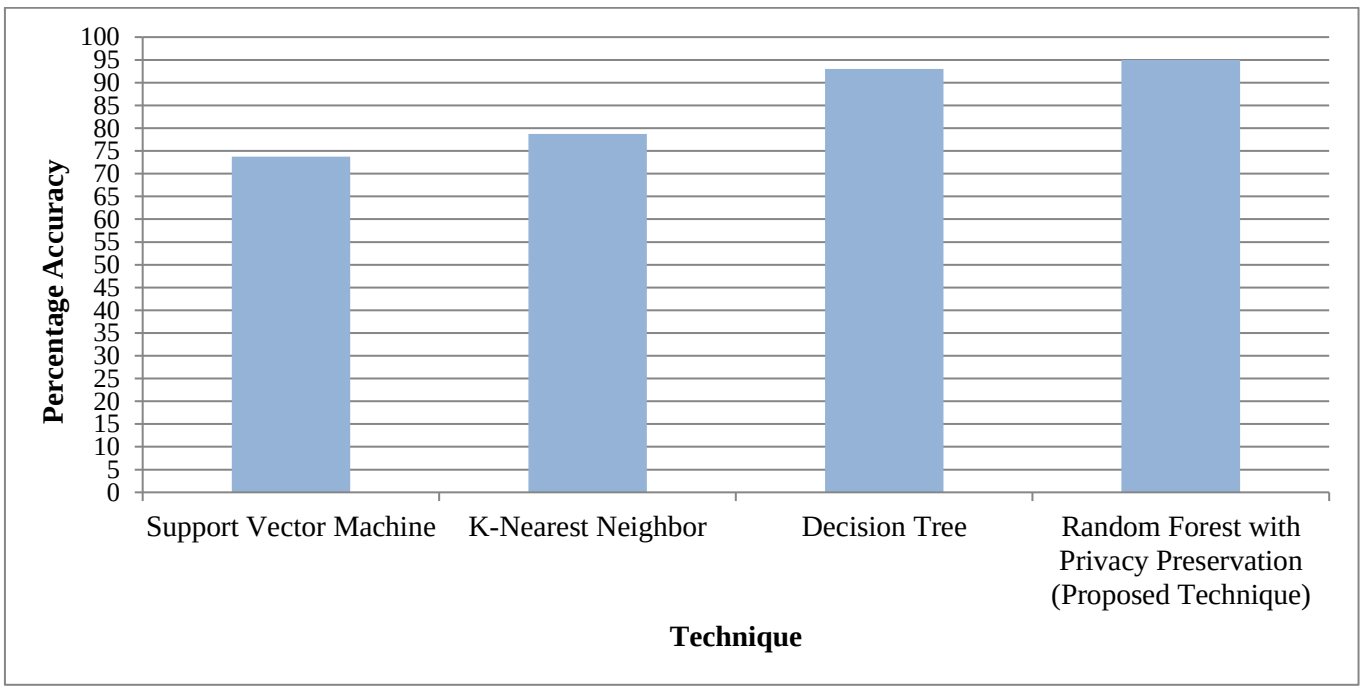

Figure 7. Accuracy comparison of proposed technique with various classification methods

4.3 Other facets of proposed research

Nature inspired algorithms are emerging as an advancement to solve various complex problems in engineering. They tend to find the solution of problem through optimization of objective function by imitating the behavior of natural beings existing together. Bio-inspired algorithms need be explored more for the branch of Privacy Preserving data mining in healthcare applications. In this research, field of genetic algorithm has been merged with data classification and machine learning to predict the disease of patients' in secured manner. Concept of meta-heuristic optimization has been utilized for random noise addition. Firefly Algorithm uses MSE as a parameter for fitness function to select the key through which the values of sensitive attributes are perturbed. Perturbation has been done only in least significant bits so that original information (private) is concealed but classification results are not adversely affected. Prediction using Random Forest has high training speed because it deals with subset of features at a time rather than all of them. Variance gets averaged (ensemble learning) and total error rate is minimized. On the other side, Firefly algorithm divides the whole

population into different subgroups because of strong local attraction. It can deal with multi-modal problems with high efficiency as well as success rate and it doesn't suffer from pre-mature convergence. Although, initial random positions of fireflies and random deviation parameter improves the exploration of search space to move from local best to global best solution but still, firefly, in some instances, can fall into local optima. Further, aim is to apply and test different variants of Firefly Algorithm for healthcare field. It can be accomplished with the help of parameter tuning or further modification of location updating formula. Instead of straight walk, path of fireflies in the algorithm can be guided by different distributions like Brownian, Logarithmic etc. [26, 27]. This research can become the basis for real time applications in which data is continuously accessed for various data mining and knowledge discovery tasks but security is a bigger challenge.

5. CONCLUSION AND FUTURE SCOPE

Healthcare dataset contains a large amount of heterogeneous data. Data mining is applied to extract useful information from it. This data contains both sensitive and non-sensitive attributes. This research is focused on applying classification technique on healthcare data and reaching on certain decision about the patient's health. In addition to decision making, it is taken under consideration that patient's sensitive information is protected and probability of linkage attacks is reduced. So, before discovering knowledge from raw data, it is changed in certain ways to disguise the sensitive information while preserving the particular data property that is critical for building meaningful data-mining model. Perturbation techniques have to handle the intrinsic trade-off between preserving data privacy and data utility, as perturbing data usually reduces data utility.

Experimentation results by applying classifier on both the datasets are compared on the basis of 4 performance parameters: Accuracy, Recall, Precision and F-score. For original dataset, 97% accuracy, 0.97015 recall value, 0.98485 precision value and 0.97744 F-score value has been achieved. On the other side, perturbed dataset gives 95% accuracy, 0.95522 value for recall, 0.96970 and 0.96241 values for precision and F-score respectively. From these results, it is observed that, even after perturbation of sensitive data, it is providing good accuracy for classification/disease prediction results as discussed above. To validate the results of proposed technique, it is compared with existing data perturbation and classification techniques. Firefly algorithm is compared with Min-Max normalization and Gaussian noise addition methods on the basis of Mean Square Error, Structural Similarity Index and Peak Signal-Noise Ratio. From the values of SSIM and PSNR, it is found that by using both of these methods to add the noise in sensitive attribute, there is very much similarity between original and perturbed dataset. Due to this, necessary and sufficient privacy couldn't be provided to the patients. In firefly algorithm, noise addition is up-to the required extent and also, calculated MSE between original and perturbed age (sensitive attribute) is maximum when compared to above discussed noise addition techniques. For prediction of disease, using 70:30 ratio for training and testing data, it is observed that the proposed Privacy Preserving Data Classification technique (by using Firefly Algorithm and Random Forest Classifier) gives better accuracy i.e. 95% as compared to 73.75%, 78.75% and 93% of SVM, K-NN and decision tree algorithms respectively.

Key benefits of the proposed technique are random noise generation, better accuracy and secure prediction. In this research, MSE is taken as a parameter for calculation of fitness function which is further used to generate the optimal key for addition of noise in the sensitive data. In future, more than one parameter can be used in the objective/fitness function to select the key for noise addition. Moreover, Convolutional neural network (CNN) can be explored to further improve the accuracy of proposed technique.

REFERENCES

[1] Vitabile, S., Marks, M., Stojanovic, D., Pllana, S., Molina, J.M., Krzyszton, M., Salomie, I. (2019). Medical data processing and analysis for remote health and activities monitoring. In High-Performance Modelling and Simulation for Big Data Applications, pp. 186-220. https://doi.org/10.1007/978-3-030-16272-6_7

[2] Sharan Vinothraj, A., Hariraj, L.K., Selvarajah, V. (2020). Implementation of RFID Technology in Managing Health Information in a Hospital. Int J Cur Res Rev, 12(20): 177-182. http://dx.doi.org/10.31782/IJCRR.2020.122029

[3] Sohail, M.N., Jiadong, R., Uba, M.M., Irshad, M. (2019). A comprehensive looks at data mining techniques contributing to medical data growth: A survey of researcher reviews. In Recent Developments in Intelligent Computing, Communication and Devices, 752: 21-26. https://doi.org/10.1007/978-981-10-8944-2_3

[4] Jiang, L., Chen, L., Giannetsos, T., Luo, B., Liang, K., Han, J. (2019). Toward practical privacy-preserving processing over encrypted data in IoT: An assistive healthcare use case. IEEE Internet of Things Journal, 6(6): 10177-10190. https://doi.org/10.1109/JIOT.2019.2936532

[5] Jin, H., Luo, Y., Li, P., Mathew, J. (2019). A review of secure and privacy-preserving medical data sharing. IEEE Access, 7: 61656-61669. https://doi.org/10.1109/ACCESS.2019.2916503

[6] Kumar, A., Kumar, R. (2020). Privacy preservation of electronic health record: Current status and future direction. In Handbook of Computer Networks and Cyber Security, 715-739. https://doi.org/10.1007/978-3-030-22277-2_28

[7] Kundalwal, M.K., Chatterjee, K., Singh, A. (2019). An improved privacy preservation technique in health-cloud. ICT Express, 5(3): 167-172. https://doi.org/10.1016/j.icte.2018.10.002

[8] Rodríguez-Hoyos, A., Estrada-Jiménez, J., Rebollo-Monedero, D., Parra-Arnau, J., Forné, J. (2018). Does k-anonymous microaggregation affect machine-learned macrotrends? IEEE Access, 6: 28258-28277. https://doi.org/10.1109/ACCESS.2018.2834858

[9] Kiran, A., Vasumathi, D. (2020). Data mining: min–max normalization based data perturbation technique for privacy preservation. In Proceedings of the Third International Conference on Computational Intelligence and Informatics. Singapore: Springer, pp. 723-34. https://doi.org/10.1007/978-981-15-1480-7_66

[10] Jain, Y.K., Bhandare, S.K. (2011). Min max

normalization based data perturbation method for privacy protection. International Journal of Computer & Communication Technology, 2(8): 45-50. https://doi.org/10.47893/ijcct.2013.1201

[11] Kalaivani, R., Chidambaram, S. (2014). Additive Gaussian noise based data perturbation in multi-level trust privacy preserving data mining. International Journal of Data Mining & Knowledge Management Process, 4(3): 21-29. https://doi.org/10.5121/IJDKP.2014.4303

[12] Hu, Z., Luo, Y., Zheng, X., Zhao, Y. (2020). A novel privacy-preserving matrix factorization recommendation system based on random perturbation. Journal of Intelligent & Fuzzy Systems, 38(4): 4525-4535. https://doi.org/10.3233/JIFS-191287

[13] Jahan, T., Narasimha, G., Rao, V.G. (2016). A multiplicative data perturbation method to prevent attacks in privacy preserving data mining. International Journal of Computer Science and Innovation, 1(1): 45-51.

[14] Balasubramaniam, S., Kavitha, V. (2015). Geometric data perturbation-based personal health record transactions in cloud computing. The Scientific World Journal, 2015: 927867. https://doi.org/10.1155/2015/927867

[15] Krishnan, C., Lalitha, T. (2020). Attribute-Based Encryption for Securing Healthcare Data in Cloud Environment. PalArch's Journal of Archaeology of Egypt/Egyptology, 17(9): 10134-10143.

[16] Han, J., Kamber, M., Pei, J. (2011). Data Mining Concepts and Techniques (3rd ed.). Elsevier. https://doi.org/10.1016/C2009-0-61819-5

[17] Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. International Journal of Advanced Computer Science and Applications, 11(2): 612-619. https://doi.org/10.14569/IJACSA.2020.0110277

[18] Chaurasia, V., Pal, S., Tiwari, B.B. (2018). Chronic kidney disease: A predictive model using decision tree. International Journal of Engineering Research and Technology, 11(11): 1781-1794. https://ssrn.com/abstract=3298343

[19] Mittal, K., Aggarwal, G., Mahajan, P. (2019). Performance study of K-nearest neighbor classifier and K-means clustering for predicting the diagnostic accuracy. International Journal of Information Technology, 11(3): 535-540. https://doi.org/10.1007/s41870-018-0233-x

[20] Tikariha, P., Richhariya, P. (2018). Comparative study of chronic kidney disease prediction using different classification techniques. In Proceedings of International Conference on Recent Advancement on Computer and Communication, 34: 195-203. https://doi.org/10.1007/978-981-10-8198-9_20

[21] Zhou, X., Lu, P., Zheng, Z., Tolliver, D., Keramati, A. (2020). Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree. Reliability Engineering & System Safety, 200: 106931. https://doi.org/10.1016/j.ress.2020.106931

[22] Biplob, M.B., Sheraji, G.A., Khan, S.I. (2018). Comparison of different extraction transformation and loading tools for data warehousing. In 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET), pp. 262-267. https://doi.org/10.1109/ICISET.2018.8745574

[23] Cunningham, P., Delany, S.J. (2021). k-nearest neighbour classifiers-A tutorial. ACM Computing Surveys (CSUR), 54(6): 1-25. https://doi.org/10.1145/3459665

[24] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing, 408: 189-215. https://doi.org/10.1016/j.neucom.2019.10.118

[25] Asuncion, A., Newman, D. (2007). UCI machine learning repository. Chronic_Kidney_Disease Data Set. https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease, accessed on Jul. 03, 2015.

[26] Yang, X.S. (2010). Firefly algorithm, Levy flights and global optimization. In Research and development in intelligent systems XXVI: 209-218. https://doi.org/10.1007/978-1-84882-983-1_15

[27] Kumar, V., Kumar, D. (2021). A systematic review on firefly algorithm: past, present, and future. Archives of Computational Methods in Engineering, 28(4): 3269-3291. https://doi.org/10.1007/s11831-020-09498-y

[28] Jabbar, M.A., Deekshatulu, B.L., Chandra, P. (2016). Intelligent heart disease prediction system using random forest and evolutionary approach. Journal of Network and Innovative Computing, 4(2016): 175-184.

[29] Subasi, A., Alickovic, E., Kevric, J. (2017). Diagnosis of chronic kidney disease by using random forest. In CMBEBIH 2017, 589-594. https://doi.org/10.1007/978-981-10-4166-2_89

[30] Yadav, D.C., Pal, S. (2020). Prediction of heart disease using feature selection and random forest ensemble method. International Journal of Pharmaceutical Research, 12(4): 56-66. https://doi.org/10.31838/ijpr/2020.12.04.013

[31] Abdulkareem, N.M., Abdulazeez, A.M., Zeebaree, D.Q., Hasan, D.A. (2021). COVID-19 world vaccination progress using machine learning classification algorithms. Qubahan Academic Journal, 1(2): 100-105. https://doi.org/10.48161/qaj.v1n2a53

[32] Singh, S.S., Sachdeva, R., Singh, A. (2020). An optimized approach for underwater image dehazing and colour correction. In Proceedings of the International Conference on Innovative Computing & Communications (ICICC). http://dx.doi.org/10.2139/ssrn.3565932

[33] Attwal, K.P.S., Dhiman, A.S. (2020). Investigation and comparative analysis of data mining techniques for the prediction of crop yield. International Journal of Sustainable Agricultural Management and Informatics, 6(1): 43-74. https://dx.doi.org/10.1504/IJSAMI.2020.106540