International Journal of Hepatology Sciences
Online ISSN: 2664-9683, Print ISSN: 2664-9675
Received: 03-01-2019; Accepted: 05-02-2019; Published: 13-02-2019
www.hepatologyjournals.com
Volume 1; Issue 1; 2019; Page No. 09-13

Automatic detection of liver disease using voting ensemble method

Shawni Dutta[1], Samir Kumar Bandyopadhyay[2]
[1] Department of Computer Science, the Bhawanipur Education Society College, Kolkata, India
[2] Academic Advisor, the Bhawanipur Education Society College, Kolkata, India

Abstract
Live disease is one of the prominent causes of death which can be tackled by providing detection at an early stage along with possible countermeasures. Liver diseases caused by factors like genetic predisposition, infections and the environment. It requires diverse and targeted treatment options. The increasing of hepatic conditions worldwide is due to lifestyle actions such as intake of alcohol and drug with the consultation of physicians. The cause of numerous infections and disorders are not yet well understood. A voting ensemble method is proposed in this paper that considers influential factors responsible for liver disease. This predictive model aims to enhance forecasting reports with respect to other peer intelligent model. The enhanced efficiency reaches an accuracy of 77.2% which is quite promising towards early liver disease prediction.

Keywords: liver disease prediction, automated prediction, data mining, ensemble classifiers

1. Introduction
Liver cancer, hepatitis, non-alcoholic fatty liver disease and end-stage liver disease are studied extensively to improve the understanding of the mechanisms of disease progression. It guides decisions of therapy selection and timing of treatment. An alternative way to medical diagnosis is offered in this paper by implementing an intelligent model. Doctors often spend longer time to assess the enzyme values during normal diagnostic period while making a decision based on those enzymes. In this context, a contribution is made to medical diagnosis process by shortening time. The intelligent model proposed in this paper assists the physician to handle critical cases while proceeding through diagnostic period.

Data mining approaches are often found to be useful in examining and identifying hidden patterns from large amount of data for inferring conclusions. This analytical process is put into practice by executing machine learning algorithms for analysing the medical data. The machine learning model receives medical data as inputs and enables to learn by itself to cultivate the knowledge base. To provide informed decision by predicting the unknown data or label of given data is the objective of such learners.

A classification process has been carried out in this paper that identifies the patients according to whether they indeed suffer from a chronic Liver Disease or not. In this regard, two-phase classification based framework is approached by this paper. During the first phase, set of five classifiers such as K-Nearest Neighbors Classifier [1], Decision Tree Classifier [2], AdaBoost Classifier [3], Random Forest Classifier [4], and Gradient boosting algorithms [5] are implemented. Next, their performances are compared with respect to some selection criterions and top two learners are picked up. The predictions of these best two learners are assembled under a single platform with the intention of efficiency maximization. A voting strategy based ensemble method is constructed in this paper that aims to combine predictions of two best learners. The output obtained from this voting ensemble classifier finally predicts whether the medical patient needs medical assistance for liver disease treatment or not.

2. Related Work
Differential evolution for automatic rule extraction from medical databases is presented in [6] accompanying with tenfold cross-validation mechanism. The objective of this study is obtaining automatic classification of items in medical databases. Applying the method to liver disorder dataset reached accuracy of 64.74%, specificity of 45.08%, sensitivity of 79.84% and ROC curve area of 62.46 [6]. Another research ensured such as application of Fuzzy beans, Bocklisch membership function, and differential evolution algorithm was found in [7]. Experimental study concluded that liver disorders diagnosis obtained an accuracy of 73.9%.

Elizondo *et al* in [8] aimed in measuring the level of complexity of classification data sets. The proposed method reduces any two class classification problem to a sequence of linearly separable steps. The amount of reduction steps could be observed as measuring the degree of non-separability which in turn denotes the complexity of the problem.

Using Bayesian Classifier, automatic diagnosis of Liver diseases is carried out in [9] by analysing the blood tests on liver functionality. In [10], early detection of liver diseases is utilised by implementing of Support Vector Machine (SVM), Boosted C5.0, and Naive Bayes (NB) classification algorithms for the early detection of liver diseases. Binish Khan et. al. in [11] have analysed various classification algorithms such as Random Forest, Logistic Regression and Separation Algorithm so as to find out the best classifier for determining the liver disease. Comparative study concluded that, Random Forest has shown highest accuracy and outperformed the other algorithms in the prediction of liver disease. Hoon Jin *et al*. in [12], implemented numerous classification techniques that assist the doctors to determine the disease quickly and efficiently. Classifiers include Naïve Bayes, Multi-Layer Perceptron, Decision Tree and k-NN for

implementation and evaluation based on several parameters like specificity, sensitivity and so on. The experimental results showed that in terms of precision, Naïve Bayes gave the better classification results whereas Logistic Regression and Random Forest provided better results in terms of recall and sensitivity.

3. Proposed Methodology

Machine learning approaches are utilized in this paper while constructing predictive models. Machine learning is useful when there is a large amount of example data and when the rules for prediction are unclear. In constructing the model, classification approach is employed rather than a regression approach. Given several influencing factors like Alkaline Phosphotase, Alamine Aminotransferase, and Aspartate Aminotransferase and many more, it can be predicted if the patient will have a liver disease suffering or not. Hence the prediction becomes a binary (yes/no) classification problem. The system flow diagram of the proposed methodology is shown in Figure 1.

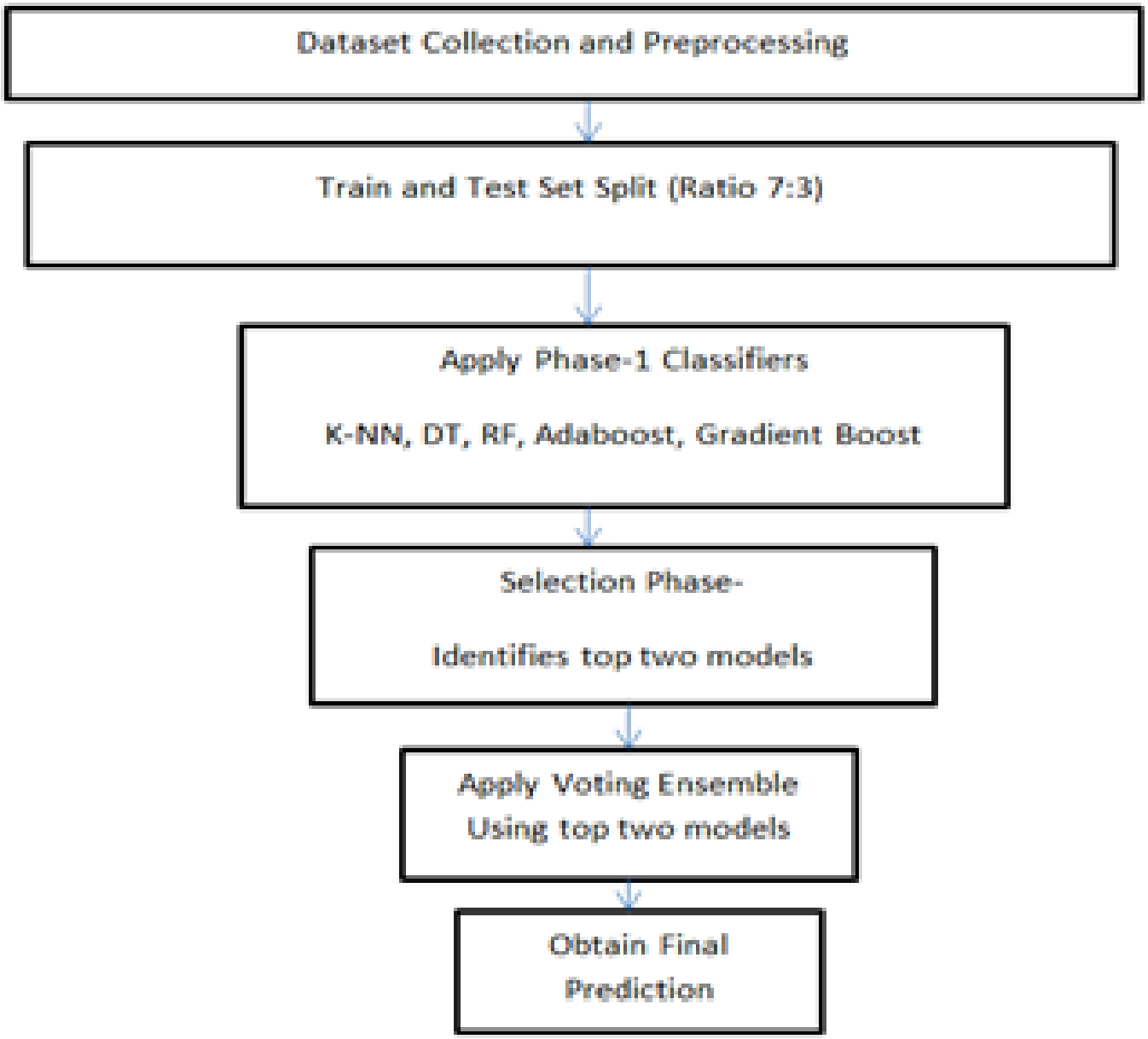

Fig 1: System Flow Diagram of Proposed Methodology

3.1 Dataset Collection and Pre-processing-

The data set used for the classifier training was obtained from the Machine Learning Repository University of California, Irvine [13]. The dataset can be formulated as collection of attributes as follows-Age of the patient, Gender: Gender of the Patients,Total Bilirubin, Direct Bilirubin, Alkaline Phosphotase, Alamine Aminotransferase, Asparatate Aminotransferase, Total Proteins, Albumin, Albumin and Globulin Ratio, and Selector field used to split the data into two sets labeled by the experts. All the features are turned out to be good predictor for diagnosis. This selector attribute is the target class that is to be predicted by classifier models. After collecting the dataset, data cleaning operation such as missing values replacement, irrelevant attribute elimination etc. are applied. This step is followed by attribute scaling operation. Relevant attributes are scaled into a range to be fitted into a classifier. All these applied pre-processing techniques will assist in obtaining transformed dataset. Before feeding these data into classifier model, it is partitioned into training set and testing set with a ratio of 7:3. Training set is fitted into the classifier model, and later

prediction is obtained for the testing set.

3.2 Methodology and Implementation-

Classifier model receives attributes of the dataset as input and inputs are mapped to target class by considering training data. The target class identifies whether patient need to proceed through diagnostic process or not. In this framework, numerous classifiers are implemented for early prediction of liver disease. The methodology proceeds by implementing two-phase classification. Each of these phases are explained as follows-

First Phase Classifiers

During the first phase, following classifiers such as K-Nearest Neighbors Classifier, Decision Tree Classifier, AdaBoost Classifier, Random Forest Classifier and Gradient boosting algorithms are implemented. Brief description of the classifiers are provided as follows-

a. K-Nearest Neighbors Classifier

K-Nearest Neighbour Classifiers [1] is often known as lazy learners. The classification procedure is a two stage process in which it identifies objects based on closest proximity of training examples in the feature space and then the classifier considers k number of objects as the nearest object while determining the class. The main challenge of this classification technique relies on choosing the appropriate value of k [1].

b. Decision Tree Classifier-

A Decision Tree (DT) [2] is a classifier that gains knowledge on classification by exemplifying the use of tree-like structure. Each target class is denoted as a leaf node of DT and non-leaf nodes of DT are used as a decision node that indicates certain test. The outcomes of those tests are identified by either of the branches of that decision node. Starting from the beginning at the root this tree are going through it until a leaf node is reached. It is the way of obtaining classification result from a decision tree [2].

c. Random Forest Classifier

Ensemble approach facilitates several machine learning algorithms to perform together to obtain higher accuracy of the entire system. Random forest (RF) [4] exploits the concept of ensemble learning approach and regression technique applicable for classification based problems. This classifier assimilates several tree-like classifiers which are applied on various sub-samples of the dataset and each tree casts its vote to the most appropriate class for the input.

d. AdaBoost Classifier

Boosting is an efficient technique where several unstable learners are assimilated into a single learner in order to improve accuracy of classification [3]. Boosting technique applies classification algorithm to the reweighted versions of the training data and chooses the weighted majority vote of the sequence of classifiers. AdaBoost [3] is a good example of boosting technique that produces improved output even when the performance of the weak learners is inadequate.

e. Gradient Boosting Classifier

Gradient boosting [10] algorithm is another boosting technique based classifier that exploits the concept of decision tree. It finds models which decrease the loss function obtained from trained samples. From these calculations the errors are measured and analysed for optimal prediction of results. Loss function calculates the range of detected rate which compares with desired target. Onward stepwise process is most popular method for updating different with various attributes. The accuracy is optimized by reducing loss function and adding base learners at all stages.

Implementation of Phase 1 Classifiers-

While implementing aforementioned classifiers, it is necessary to put concentration on parameter-tuning because this will enhance performance of the models. This framework utilised the K-NN classifier for the value k=5 considering all the evaluating metric for obtaining maximised results. On the other hand, ensemble classifiers, such as, Random Forest, AdaBoost and Gradient Boost classifiers are built based on 500 numbers of estimators on which the boosting is terminated. After constructing these classification models, training data are fitted into it. Later the testing dataset are used for prediction purpose. After the prediction is done, performance of the classifiers are evaluated based on the predicted value and the actual value.

3.3 Selection Procedure-

During this procedure, aforementioned classifiers are evaluated as well as compared with respect to pre-defined metrics. Use of this metrics will assist in justifying the performance of best problem-solving approach. These metrics are discussed as follows-

1. Accuracy

Accuracy [14] is a metric that ascertains the ratio of true predictions over the total number of instances considered. However, the accuracy may not be enough metric for evaluating model's performance since it does not consider wrong predicted cases with different weights.

2. F1-Score-

For compensating the above addressed problem, we consider two more metrics known as, Recall and Precision. *Precision* [14] identifies the ratio of correct positive results over the number of positive results predicted by the classifier. *Recall* [14] denotes the number of correct positive results divided by the number of all relevant samples. *F1-Score* or *F-measure* [14] is a parameter that is concerned for both recall and precision and it is calculated as the harmonic mean of precision and recall.

3. MSE-

Mean Squared Error (MSE) [14] is another evaluating metric that measures absolute differences between the prediction and actual observation of the test samples.

Mathematically, the aforementioned metrics can be defined as follows with given True Positive, True Negative, False Positive, False Negative as TP, TN, FP, FN respectively-

Accuracy = TP+TN/ (TP+FP+TN+TP)

Recall = TP/ (TP+FN) Precision= TP/(TP+FP)

F1- Measure or F1-Score = 2* Recall * Precision / (Recall + Precision)

MSE= $(\sum_{i=1}^{N}(Xi - Xi')^2$ / **N)** where Xi is the actual value and Xi' is the predicted value.

Lower value of MSE and higher values of accuracy and F1-Score signifies a better performing model.

After computing the performance of above specified classifiers, two best learners are identified and then fed into the next phase of classification which ensembles these two weak learners in order to attain maximized efficiency.

3.4. Second Phase Classification-

In this phase, an ensemble method is proposed that assembles prediction of two best learners in order to maximize the performance of predictive tool. For this

purpose, a voting strategy based ensemble method [15-16]. is proposed in this paper. Using voting strategy, it is potential to make a good choice out of multiple possible solutions. Hence, multiple classifiers may cast their preference for one or more solutions. Considering majority preferences, final decision is drawn for problem-solving approach. It is possible to obtain a better solution when several potential algorithms work in harmony to solve the same problem domain. Using ensembles of different classifiers has the advantage that not all of them will make the same mistake. Selection procedure picks up top two learners and this Voting ensemble method assimilates those learners in order to draw final problem-solving inference.

4.1 Experimental Results

In this section, performance of each of the aforementioned classifiers is shown with respect to performance evaluation metrics

Table 1: Performance Summary of all phase-1 Classifiers.

Performance Measure Metrics	K-NN Classifier	Decision Tree Classifier	AdaBoost Classifier	Gradient Boost Classifier	Random Forest Classifier
Accuracy	69.95%	68.39%	72.54%	74.09%	71.5%
F1-Score	0.7	0.68	0.73	0.74	0.72
MSE	0.3	0.32	0.27	0.26	0.28

Table 2: Performance of Proposed Voting Ensemble Method

Performance Measure Metrics	Accuracy	F1-Score	MSE
Voting Ensemble Method	77.2%	0.77	0.23

Analysis

During phase-1, numerous classifiers are implemented on the liver disease dataset and as result predictions are obtained. Prediction outcomes are evaluated against some metrics which assist in identifying two best learners. In this case as shown in Table 1, the Adaboost and Gradient Boost classifier turned Out to be the best classifiers. After selecting these models, voting ensemble method proposed in this paper assembles them in order to maximize the efficiency of the predictive model. These models are assembled using 'hard' voting strategy during implementation. Table 2 concludes that this method is quite prominent in terms of prediction over other specified models. An accuracy of 77.2%, F1-Score of 0.77 and MSE of 0.23 is offered by this proposed predictive model. Figure 2 and 3 describes prediction performance of each classifier with respect to Accuracy, F1-Score and MSE.

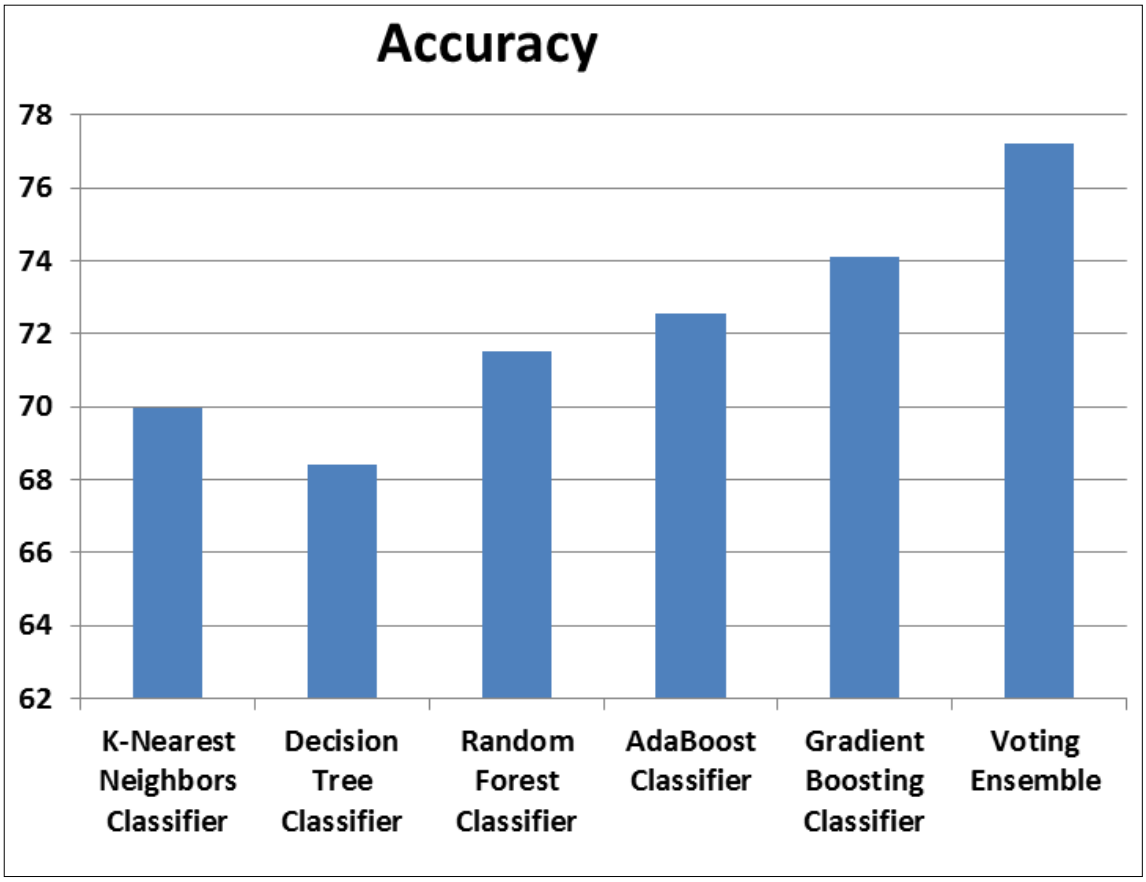

Fig 2: Prediction Performance of all classifiers in terms of Accuracy.

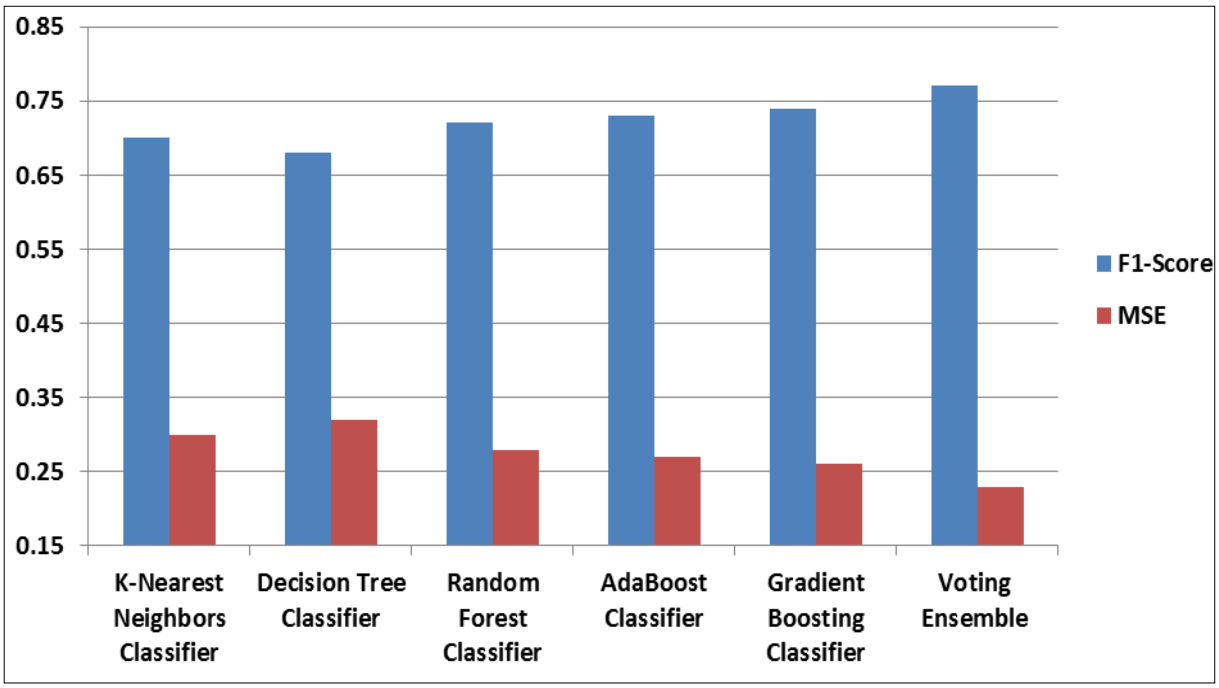

Fig 3: Prediction Performance of all classifiers in terms of F1-Score and MSE

Conclusion

Accordingly to [Add Ref], liver disease is one of the top ten leading causes of death in India and it is responsible for over 2.4% of Indian deaths per annum. Hence, attention to this disease can heal the medical care system by predicting at an early stage. An intelligent automated module is proposed in this paper that assembles two boosting classifiers such as Adaboost and Gradient Boosting algorithm to obtain predictive results. In this context, voting strategy based approach is employed in order to obtain superior prediction results. Experimental results have shown significantly better prediction results of the proposed method in comparison with other specified classifier models.

References

1. Cunningham P, Delany SJ. "K -Nearest Neighbour Classifiers," *Mult. Classif. Syst*, 2007, 1-17, doi: 10.1016/S0031-3203(00)00099-6.
2. Sharma H, Kumar S. "A Survey on Decision Tree Algorithms of Classification in Data Mining," *Int. J. Sci. Res.* 2016; 5(4):2094-2097. doi: 10.21275 /v5i4.nov162954.
3. Friedman J, Hastie T, Tibshirani R. "Additive logistic regression: a statistical view of boosting," *Ann. Stat.* 2000; 28(2):337-407. doi: 10.1214/aos/1016218223.
4. Breiman L. "Random Forests," *Mach. Learn.*, vol. 2001; 45(1):5-32, doi: 10.1017/CBO 9781107 415324.004.
5. Natekin A, Knoll A. "Gradient boosting machines, a tutorial," *Front. Neurorobot*, 2013, 7. DEC, doi: 10.3389/fnbot.2013.00021.
6. De Falco I. "Differential Evolution for automatic rule extraction from medical databases," *Appl. Soft Comput. J.* 2013; 13(2):1265-1283. doi: 10.1016/j.asoc.2012.10.022.
7. Luukka P. "Fuzzy beans in classification," *Expert Syst. Appl.* 2011; 38(5):4798-4801. doi: 10.1016/j.eswa.2010.09.167.
8. Elizondo DA, Birkenhead R, Gamez M, Garcia N, Alfaro E. "Linear separability and classification complexity," *Expert Syst. Appl.* 2012; 39(9):7796-7807. doi: 10.1016/j.eswa.2012.01.090.
9. Venkata Ramana B, Babu SPM, Venkateswarlu N. "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis," *Int. J. Database Manag. Syst.* 2011; 3(2):101-114. doi: 10.5121/ijdms.2011.3207.
10. ME S, Engy A, Ali I El-Desouky. "Prediction of Liver Diseases Based on Machine Learning," no. January. 2018; 723:229-239, doi: 10.1007/978-3-319-74690-6.
11. Khan B, Shukla PK, Ahirwar MK. "Strategic Analysis in Prediction of Liver Disease Using Different Classification Algorithms," *Int. J. Comput. Sci. Eng.* 2019; 7(7):71-76, doi: 10.26438/ijcse/v7i7.7176.
12. Jin H, Kim S, Kim J. "Decision factors on effective liver patient data prediction," *Int. J. Bio-Science Bio-Technology*. 2014; 6(4):167-178. doi: 10.14257/ijbsbt.2014.6.4.16.
13. Michael Kahn, St Louis. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
14. M H, MN S. "A Review on Evaluation Metrics for Data Classification Evaluations," *Int. J. Data Min. Knowl. Manag. Process*. 2015; 5(2):01-11. doi: 10.5121/ijdkp.2015.5201.
15. Leon F, Floria SA, Badica C, "Evaluating the effect of voting methods on ensemble-based classification," *Proc. - 2017 IEEE Int. Conf. Innov. Intell. Syst. Appl. INISTA 2017*, no, 2018, 1-6. doi: 10.1109/INISTA.2017.8001122.
16. Sontakke S, Lohokare J, Dani R, "Diagnosis of liver diseases using machine learning," 2017 Int. Conf. Emerg. Trends Innov. ICT, ICEI, 2017, 129-133. doi: 10.1109/ETIICT.2017.7977023.