Performance Evaluation of Ensemble-Based Machine Learning Techniques for Prediction of Chronic Kidney Disease

K. M. Zubair Hasan and Md. Zahid Hasan

Abstract Chronic kidney disease (CKD) is widespread and related with enhanced risk of cardiovascular disease and end-stage renal disease, which are possibly escapable through early detection and treatment of individuals at risk. Machine learning algorithm helps medical experts to diagnose the disease correctly in the earlier stage. Therefore, machine-predicted analysis has become very popular in recent decades that can efficiently recognize whether a patient has certain kidney disease or not. In this regard, we propose an ensemble method based classifier to improve the decision of the classifiers for kidney disease diagnosis efficiently. Ensemble methods combine multiple learning algorithms to achieve better predictive performance than could be obtained from any of the constituent learning algorithms. In addition, Data is evaluated by using tenfold cross-validation and performance of the system is assessed on receiver operative characteristic curve. Extensive experiments on CKD datasets from the UCI machine learning repository show that our ensemble-based model achieves the state-of-the-art performance.

Keywords Machine learning · Classification · Chronic kidney disease (CKD) · Ensemble method · Data mining · Healthcare informatics

1 Introduction

According to the report of the Centers for Disease Control and Prevention (CDC), kidney disease causes millions to die each year. A good number of people in the world with kidney damage and slightly reduced kidney function are not conscious of having CKD. Therefore, strategies for early detection and cure of people with CKD are consequently required worldwide. A computer-aided diagnosis system based on sophisticated machine learning techniques is required for healthcare data to mine

K. M. Zubair Hasan · Md. Zahid Hasan (✉)
Daffodil International University, 102 Sukrabadh, Mirpur Road, Dhaka 1207, Bangladesh
e-mail: zahid.cse@diu.edu.bd

K. M. Zubair Hasan
e-mail: kmzubair.hasan@gmail.com

N. R. Shetty et al. (eds.), *Emerging Research in Computing, Information, Communication and Applications*, Advances in Intelligent Systems and Computing 882,
https://doi.org/10.1007/978-981-13-5953-8_34

hidden pattern from data for effective decision-making. The purpose of our current study is to evaluate the performance of several ensemble-based machine learning techniques correctly classifying CKD patients based on clinical datasets. We consider five machine learning classifiers, namely Adaptive Boosting, Bootstrap Aggregating, Extra Trees, Gradient Boosting, and Random Forest Classifier. Finally, a large set of standard performance metrics is used to design the computer-aided diagnosis system for estimating the performance of each machine learning and artificial intelligence classifier. The metrics we used include Classification Accuracy, Sensitivity, Precision, Specificity, Negative Predictive Value, False Positive Rate, False Negative Rate, F1-Score, and Error Rate of Classification.

The paper is organized as follows. Section 2 discusses the related works. Section 3 introduces the materials and methods. Section 4 describes the proposed ensemble-based machine learning techniques applied to the dataset. Section 5 presents the classification performance matrices. These comparisons' study is discussed in Sect. 6. Finally, we concluded our work in Sect. 7.

2 Related Work

Ho et al. [1] presented a computer-aid diagnosis tool for CKD classification analyzing ultrasonography images. This system used for detecting and classifying distinctive various stages of CKD. The K-means clustering was performed for detecting regions in an ultrasonic image as preprocessing step. Estudillo-Valderrama et al. [2] proposed the feasibility study of using a distributed system for the management of alarms from chronic kidney disease patients inside Enefro project. Charleonnan et al. [3], explored four machine learning approaches including K-nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR), and decision tree classifiers for predicting the chronic kidney disease. In Hsieh et al. [4] showed a real-time system to consider chronic kidney disease by using only ultrasound images. Support vector machine technique is used to predict and classify CKD stages form the ultrasound images. The authors in [5] recommended diverse methods to influence the hierarchical structure in ICD-9 codes to develop the performance of predictive models. A novel feature engineering approach is proposed in this study to influence this hierarchy, while immediately diminishing feature dimensionality. An intelligent system based on artificial neural networks in Chiu et al. [6] for detecting chronic kidney disease for evaluating the extremity of a patient. Three types of artificial neural networks have been used in this model including backpropagation network (BPN), modular neural network (MNN), and generalized feedforward neural networks (GRNN).

3 Materials and Methods

Our research uses a CKD dataset [7], which is openly accessible at UCI machine learning laboratory. This dataset comprises 400 instances with 150 samples without kidney disease (not presence) and 250 samples with kidney disease (presence). Missing values in the dataset may effect in the accuracy of disease prediction. So to avoid this, we consider 158 without missing samples and then apply Extra Tress algorithm for evaluating feature importance. After getting feature importance, we compute impute (mean) for missing values of numerical type attributes for increasing the number of samples as preprocessing steps. Thenceforth, for predicting kidney

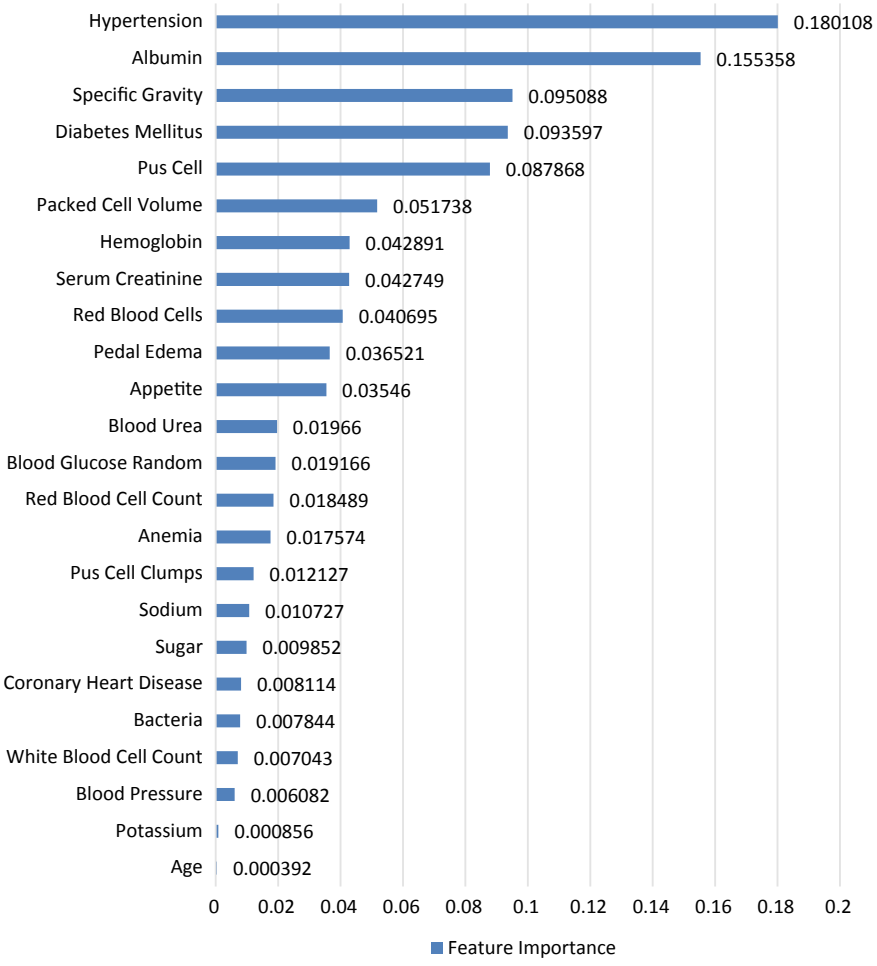


Fig. 1 Importance of attributes in CKD datasets

diseases, 13 distinct parameters (Hypertension, Albumin, Specific Gravity, Diabetes Mellitus, Pus Cell, Packed Cell Volume, Haemoglobin, Serum Creatinine, Red Blood Cells, Pedal Edema, Appetite, Blood Urea, and Blood Glucose Random) have been taken out of 24 attributes into account considering the feature importance. Figure 1 shows the importance of attributes in CKD dataset.

4 Machine Learning Techniques

4.1 Adaptive Boosting (AdaBoost)

Adaptive Boosting is an ensemble machine learning meta-algorithm technique that creates a robust classifier from a number of weak classifiers. For improving performance, it incorporated extra copies of the classifier on the same dataset, however where the weights of incorrectly classified samples are adjusted and adjusts them to represents the final output of the boosted classifier.

Given M training data $\{(x_1, y_1), \ldots, (x_M, y_M)\}$, x_i is a vector corresponding to an input sample data, associated with Q input attributes, and y_i is a target variable with a class label of either -1 or $+1$ and a set of weak learners $\{K_1, \ldots, K_L\}$ each of which outputs a classification $K_j(x_j) \in \{-1, 1\}$ for each classifier. An initial weight is set for $1/n$. Equal weight is assigned to all the instances in the kidney datasets. AdaBoost is explained using pseudocode [8].

For t in $1, \ldots, T$:

- Choose $h_t(x) \rightarrow [-1, 1]$:
  - Find weak learners that minimize $\in_t$, the summation of errors in weights for misclassified points

$$\in_t = \sum_{\substack{i=1 \\ h_t(x_i \neq y_i)}}^{n} w_{i,t}$$

  - Choose $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\in_t}{\in_t}\right)$
  - Add to ensemble $F_t(x) + \alpha_t h_t(x)$
  - Update weights for $i = 1, \ldots, m$ using

$$w_{t+1}(i) = \frac{w_t(i)\mathrm{e}^{-\alpha_t y_i h_t(x_i))}}{Z_t}$$

  Where Z_t is a normalization factor.

– Final hypothesis:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

H is determined as a weight majority vote of the weak hypothesis h_t, where each classifier is assigned weight α_t.

4.2 Bootstrap Aggregating (Bagging)

Bagging is another ensemble meta-algorithm used in statistical classification for reducing variance and supports to avoid overfitting. At first, an initial weight is set for kidney datasets [9].

For every classifier as T rounds:

- Normalize the weights
- Train the classifier and evaluate training error
- Choose: lowest error classifier
- Update the weights of the training data

Then, the final classifier is formed as the linear combination of T classifiers.

4.3 Extra Trees Classifier

Extra Trees is another modification of bagging classifier with ordinary tree-based techniques as far as accuracy and computational efficiency. The main differences with other tree-based algorithm are that it can split the node by choosing cut-points randomly and build the trees using the total learning samples [10].

Split_node(S)
Input: Spited node equivalent to local subset S equivalent
Output: a split where a_c is greater than a or not anything

- If **End_split(S)** is true then it will not return anything
- Or else, Choose k attributes $\{a_1, \ldots, a_k\}$ among all nonconstant (in S) candidate attributes.
- Draw k splits $\{s_1, \ldots, s_k\}$, where $s_i = \textbf{Choice\_random\_split}(s, a_i), \forall_i = 1, \ldots, k$;
- Return a split s_* such that s will be selected from the maximum score of S.

Choice_random_split(S,*a*)
Input: Local subset S and an attribute a.
Output: a split

- Draw a random cut-point a_C considering maximal attributes of subset S and minimal attributes of subset S
- Return the split based on $[a < a_C]$

End_split(S)
Input: Local subset S
Output: a Boolean value

- If the absolute value of S is less than the minimum sample size of the splitting node, then return TRUE;
- If every property is consistent at that point then return TRUE;
- If the outcome is steady in S, then return TRUE;
- Or else, return FALSE.

4.4 Gradient Boosting

Gradient boosting is an optimization machine learning algorithm on a suitable cost function for the prediction of kidney disease classification problem. In the process of gradient boosting, a series of predictor values are iteratively formed. The final predictor value is generated after iteratively calculating the weighted average values of the weak predictor. At every sequence, an extra classifier is invoked to boost the performance of the complete ensemble [11]. The algorithm for the predictive model is enlisted in Table 2.

Input: A training set of data points (x_i, y_i) from the given dataset
Output: A classification tree

- Initialization: Initialize model with a constant low value of T classifier.
- **For** i to number of iterations M **do**
- Calculate new weights of (x_0, y_0) to (x_i, y_i) with minimal prediction accuracy rate of T.
- Draw new weak classifier h_i on the pre-weighted prediction accuracy of T.
- Calculate weight β_i of new classifier.
- Update the model as $h_i + \beta_i$.
- **end for**
- **return** T

Gradient boosting adds weak learners to minimize the bias in addition to the variance to some degree, thus reducing the error [12].

4.5 Random Forest

Random Forest (RF) is a variant of ensemble classifier consisting of a collection of tree-structured classifiers $h(x, y_k)$, which is defined as multiple tree predictors y_k such that each tree relies upon the estimations of an arbitrary vector inspected independently and with a similar distribution for all trees in the forest. The randomization works in two ways: an arbitrary sampling of data for bootstrap samples as it

is done in Bootstrap aggregating (bagging) in Sect. 4.2 and random selection of input attributes for producing individual base decision trees [13]. Random forests become different in a way from other methods that a modified tree learning algorithm is utilized that chooses the differentiable candidate in the learning procedure, a random subset of the features. The cause for doing this is the relationship of the trees in a standard bootstrap sample. For example, if one or a couple of features are extreme indicators for the response variable (target output), these features will be chosen in a considerable lot of the decision trees, reasoning them to end up correlated.

5 Classification Performance Measurements

Five ensemble-based supervised machine learning techniques were applied for the classification of chronic kidney disease attributes. Classification performance was estimated by tenfold cross-validation. The classification models were assessed on the nine quality measures [14]. These quality measures for the analysis of classification were examined closely. Absence attributes of chronic kidney disease were measured as non-ckd class, and attributes with the presence of chronic kidney disease were measured as ckd class. Here,

True positive (TP)—Correct positive prediction where samples with ckd predicted as ckd.
False positive (FP)—Incorrect positive prediction where samples with non-ckd predicted as ckd.
True negative (TN)—Correct negative prediction where samples with non-ckd predicted as non-ckd.
False negative (FN)—Incorrect negative prediction where samples with ckd predicted as non-ckd.

The quality measures are executed to evaluate the efficiency of each machine learning classifier in the distinction between presence and absence with samples of chronic kidney disease. They are expressed as follows:

$$\text{Classification Accuracy} = \frac{(TN+TP)}{(TN+FP+FN+TP)}$$
$$\text{Sensitivity} = \frac{TP}{(TP+FN)}$$
$$\text{Precision} = \frac{TP}{(TP+FP)}$$
$$\text{Specificity} = \frac{TN}{(TN+FP)}$$
$$\text{Negative Predictive Value} = \frac{TN}{(TN+FN)}$$
$$\text{False Positive Rate} = \frac{FP}{(FP+TN)}$$
$$\text{False Negative Rate} = \frac{FN}{(FN+TP)}$$
$$\text{Rate of Misclassification} = \frac{(FN+FP)}{(TN+FP+FN+TP)}$$
$$\text{F1 Score} = 2 * \frac{(\text{Precision*Recall})}{(\text{Precision+Recall})}$$

6 Result and Discussion

The efficiency of five ensemble-based machine learning approaches was accounted that all classifiers are trained with all 13 classes discussed in methods and materials section, following a tenfold cross-validation to illustrate statistically robust results for the prediction of kidney disease. We employed the Python software to execute all classification tasks. Predicted results based on the confusion matrix is arranged in Tables 1, 2, and 3 for Adaptive Boosting (AdaBoost), Bootstrap Aggregating (Bagging), Extra Trees, Gradient Boosting, and Random Forest, respectively.

Figure 2 shows the classification results of these ensemble machine learning algorithms. It is obvious from the outcomes that Bootstrap Aggregating (Bagging) predicts most noteworthy of true positives (presence of kidney disease classified as presence) (Table 2; Fig. 2) and combined Adaptive Boosting (AdaBoost) and Extra Trees algorithm predicts the most noticeable number of true negatives (absence of kidney disease classified as absence of kidney disease) (Tables 1 and 3; Fig. 2). Adaptive and gradient boosting confusion matrix (Tables 1 and 2) indicates that it has second-best true positives (Fig. 2). Tables 2 and 3 show the confusion matrix of Extra Trees and Random Forest classifier, which indicates that this classifier gives the third most elevated number of true positives (Fig. 2).

Table 1 Confusion matrix for tenfold cross-validation using adaptive boosting and bootstrap aggregating

Adaptive boosting				Bootstrap aggregating			
	Predicted class				Predicted class		
	ckd	Non-ckd	Actual total		ckd	Non-ckd	Actual total
Actual class				Actual class			
ckd	61	1	62	ckd	62	0	62
Non-ckd	0	38	38	Non-ckd	4	34	38
Total predicted	61	39	100	Total predicted	66	34	100

Table 2 Confusion matrix for tenfold cross-validation using extra trees and gradient boosting

Extra trees				Gradient boosting			
	Predicted class				Predicted class		
	ckd	Non-ckd	Actual total		ckd	Non-ckd	Actual total
Actual class				Actual class			
ckd	60	2	62	ckd	61	1	62
Non-ckd	0	38	38	Non-ckd	2	36	38
Total predicted	60	40	100	Total predicted	63	37	100

Table 3 Confusion matrix for tenfold cross-validation using random forest

Random forest			
		Predictive class	
	CKD	Non-CKD	Actual total
Actual class			
ckd	60	2	62
Non-ckd	3	35	38
Total predicted	63	37	100

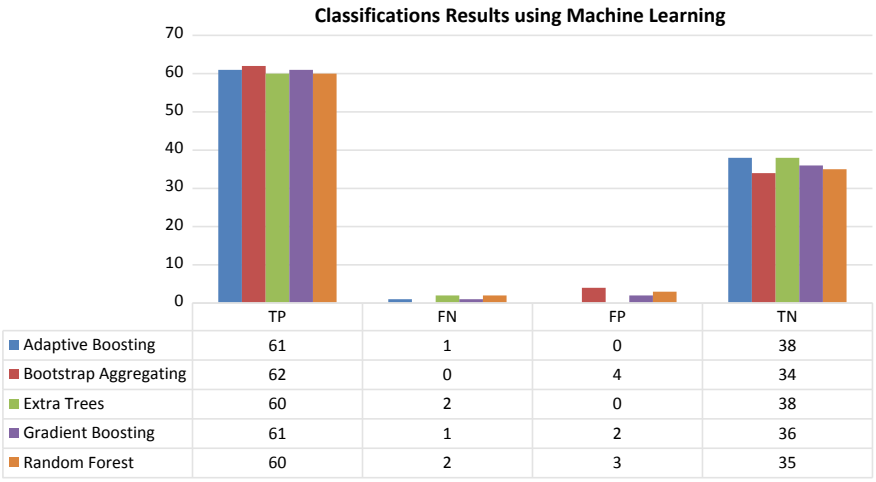
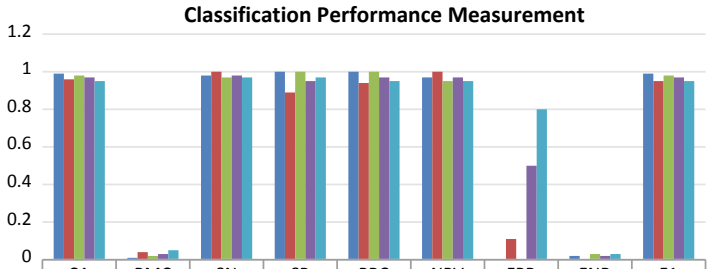


Fig. 2 Classification results of ensemble machine learning techniques

Figure 3 plots nine quality measures for classification performance measurements. Figure 3 specifies that Adaptive Boosting beats over all other machine learning algorithms with the most extreme classification accuracy of 99% while the second most noteworthy characterization exactness is accomplished by Extra Trees 98%. Furthermore, these two strategies have indicated the most extreme specificity and precision of 100% that this classifier is most appropriate for distinguishing proof of patients with chronic kidney disease (absence class). Bootstrap Aggregating achieves the highest sensitivity, the negative predictive value of 100% and the lowest false negative rate of zero percent, which indicates that this classifier is most appropriate for identification of people who are sick with chronic kidney disease (presence class). But, bagging also has the lowest specificity of 89% and false negative rate of 11% that is not fit for prediction of sick people with chronic kidney disease (absence class). Figure 3 associates F1 Score and Error Rate of Classification Adaptive Boosting shows the highest 99% and the lowest 1%, respectively.



	CA	RMC	SN	SP	PRC	NPV	FPR	FNR	F1
Adaptive Boosting	0.99	0.01	0.98	1	1	0.97	0	0.02	0.99
Bootstrap Aggregating	0.96	0.04	1	0.89	0.94	1	0.11	0	0.95
Extra Trees	0.98	0.02	0.97	1	1	0.95	0	0.03	0.98
Gradient Boosting	0.97	0.03	0.98	0.95	0.97	0.97	0.5	0.02	0.97
Random Forest	0.95	0.05	0.97	0.97	0.95	0.95	0.8	0.03	0.95

Fig. 3 Classification performance of ensemble machine learning techniques

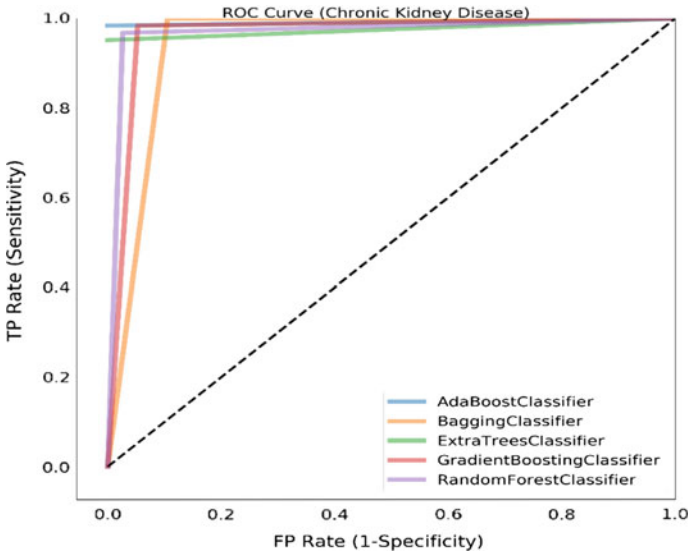


Fig. 4 ROC for ensemble machine learning techniques

In agreement to the previously mentioned assessment criteria, a receiver operating characteristic (ROC) curve [15] is used and the area under the curve (AUC) to evaluate the advantage and disadvantage of the classifier. ROC is unbiased of the two classes and significant when the number instances of the two classes change through training [16]. The region under ROC curve must be close to 1 for the best classifier. Figure 4 shows that AdaBoost classifier beats every single other procedure in expectation of

quality of kidney ailment and other ensemble-based learning techniques perform the approximate similar result in the prediction of the essence of kidney disease.

7 Conclusion

This prediction of kidney disease may spare the life of people and can have a real effect on its cure. In this paper, we propose an ensemble method based machine learning algorithm to improve the performance of the classifier for kidney disease. In contrast with quite a few classic prediction algorithms, the classification accuracy of our proposed ensemble learning algorithm achieves 99% classification accuracy.

References

1. Ho, C.-Y., Pai, T.-W., & Peng, Y.-C. Ultrasonography image analysis for detection and classification of chronic kidney disease. In *Sixth International Conference on Complex, Intelligent, and Software Intensive Systems* (pp. 624–629), Palermo.
2. Estudillo-Valderrama, M. A., et al. (2014, November). A distributed approach to alarm management in chronic kidney disease. *IEEE Journal of Biomedical and Health Informatics, 18*(6), 1796–1803.
3. Charleonnan, A., Fufaung, T., Niyomwong, T., Chokchueypattanakit, W., Suwanna-wach, S., & Ninchawee, N. (2016). Predictive analytics for chronic kidney disease using machine learning techniques. In *Management and Innovation Technology International Conference (MITicon)* (pp. 80–83), Bang-San.
4. Hsieh, J.-W., Lee, C.-H., Chen, Y.-C., Lee, W.-S., & Chiang, H.-F. (2014). Stage classification in chronic kidney disease by ultrasound image. In *International Conference on Image and Vision Computing*, New Zealand, ACM, 2014 (pp. 271–276).
5. Singh, A., Nadkarni, G., Guttag, J., & Bottinger, E. (2014). Leveraging hierarchy in medical codes for predictive modeling. *In Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '14)* (pp. 96–103). ACM, New York, NY, USA, 2014.
6. Chiu, R. K., Chen, R. Y., Wang, S.-A., & Jian, S.-J. (2012). Intelligent systems on the cloud for the early detection of chronic kidney disease. In *International Conference on Machine Learning and Cybernetics* (pp. 1737–1742), Xian, China.
7. Asuncion, A., & Newman, D. J. UCI Machine Learning Repository [Online]. Available http://www.ics.uci.edu/~mlearn/MLRepository.html. (2007).
8. Rojas, R. (2009). AdaBoost and the super bowl of classifiers a tutorial introduction to adaptive boosting. Freie University, Berlin, Technical Report.
9. Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.
10. Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63*(1), 3–42.
11. Manna, S., Biswas, S., Kundu, R., Rakshit, S., Gupta, P., & Barman, S. (2017). A statistical approach to predict flight delay using gradient boosted decision tree. In *International Conference on Computational Intelligence in Data Science (ICCIDS)* (pp. 1–5), Chennai, India.
12. Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*(4), 367–378.

13. Anbarasi, M. S., & Janani, V. (2017). Ensemble classifier with Random Forest algorithm to deal with imbalanced healthcare data. In *International Conference on Information Communication and Embedded Systems (ICICES)* (pp. 1–7), Chennai.
14. Kanmani, S., Uthariaraj, V. R., Sankaranarayanan, V., & Thambidurai, P. (2007). Object-oriented software fault prediction using neural networks. *Information and Software Technology, 49*(5), 483–492.
15. Metz, C. E. (1978). Basic principles of ROC analysis. In L. M. Freeman (Ed.), *Seminars in nuclear medicine* (Vol. 4, pp. 283–298). Amsterdam: Elsevier.
16. Dwivedi, A. K. (2016). Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Computing and Applications, pp. 1–9.