Optimization of Prediction Method of Chronic Kidney Disease Using Machine Learning Algorithm

Pronab Ghosh
Department of Computer Science & Engineering
Daffodil International University
Dhaka, Bangladesh
pronab1712@gmail.com

F. M. Javed Mehedi Shamrat
Department of Software Engineering
Daffodil International University
Dhaka, Bangladesh
javedmehedicom@gmail.com

Shahana Shultana
Department of Computer Science & Engineering
Daffodil International University
Dhaka, Bangladesh
shahanashomi246@gmail.com

Saima Afrin
Department of Computer Science & Engineering
Daffodil International University
Dhaka, Bangladesh
saima.cse@diu.edu.bd

Atqiya Abida Anjum
Department of Computer Science & Engineering
United International University
Dhaka, Bangladesh
aanjum151026@bscse.uiu.ac.bd

Aliza Ahmed Khan
Department of Computer Science & Engineering
Daffodil International University
Dhaka, Bangladesh
alizaahmedkhancse22ju@gmail.com

2020 15th International Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP) | 978-1-6654-1554-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/iSAI-NLP51646.2020.9376787

***Abstract*—Chronic Kidney disease (CKD), a slow and late-diagnosed disease, is one of the most important problems of mortality rate in the medical sector nowadays. Based on this critical issue, a significant number of men and women are now suffering due to the lack of early screening systems and appropriate care each year. However, patients' lives can be saved with the fast detection of disease in the earliest stage. In addition, the evaluation process of machine learning algorithm can detect the stage of this deadly disease much quicker with a reliable dataset. In this paper, the overall study has been implemented based on four reliable approaches, such as Support Vector Machine (henceforth SVM), AdaBoost (henceforth AB), Linear Discriminant Analysis (henceforth LDA), and Gradient Boosting (henceforth GB) to get highly accurate results of prediction. These algorithms are implemented on an online dataset of UCI machine learning repository. The highest predictable accuracy is obtained from Gradient Boosting (GB) Classifiers which is about to 99.80% accuracy. Later, different performance evaluation metrics have also been displayed to show appropriate outcomes. To end with, the most efficient and optimized algorithms for the proposed job can be selected depending on these benchmarks.**

Keywords—Support Vector Machine, AdaBoost, Linear Discriminant Analysis, Gradient Boosting.

I. INTRODUCTION

Kidney disease develops very slowly without revealing any symptoms. There are various forms of kidney disease around the world. So most doctors usually waste their precious time to detect whether a patient is affected by kidney disease or not. In this paper, we are basically working on "Chronic Kidney Disease"[1] based on different performance indices to figure out which algorithm is best to use in this type of problem. The fast prediction of the disease can help save thousands of lives worldwide before severe damage has been done to the patients. Besides, machine learning algorithms can be used [2] to detect this disease. To detect this rising disease, several machine learning algorithms can be trained [3] depending on medical patients' data. But the challenge is to get the most accurate prediction in the shortest time.

The crucial aim of the proposed research is to build a kidney disease system focused entirely on machine-learning. The research aims to solve various algorithms, such as SVM, AB, LDA, and GB to classify the people affected by kidney disease. To make it more accurate, this study is performed using different performance assessment metrics such as False Negative Rate (FNR), Accuracy (ACC), Precision (PRE), Negative predictive value (NPV), F1 Score (F1), False Discovery Rate (FDR), Standard Deviation (SD), Specificity (SPE), Mean Absolute Error (MAE), Mean Squared Error (MSE), Sensitivity (SEN), Root Mean Squared Error (RMSE), False Positive Rate (FPR), ROC, AUC, Error Rate, and Execution time to properly evaluate classifiers performance. The main aims of this research are:

- All missing value issues have been solved through the imputation method of K-Nearest Neighbors to obtain more reliable outcomes.
- With the aid of a standard scaler technique, all features are prepossessed to hold the values within the range of [0, 1].
- The assessment process of different models has been experimented with the 80:20 distinction.
- This study elucidates accuracy, error rate, execution time, AUC, and ROC figures to demonstrate the efficiency of different classifiers.

II. LITERATRE REVIEW

Over the past researchers have shown the use of machine learning algorithms to perform various calculations and evaluate data to come up with decisions to better human life. From the field of math and science to business, medicine to every day human life, experiments using machine learning algorithms bring fruitful results.

Using AB ensemble classifiers, the authors in [4] performed human activity recognition. The data for the experiment data was gather from human body sensor. The high performance shows the feasibility of the algorithm. Developmental Dysplasia of Hip (DDH) in infants can be deadly. In [5] the authors use SVM technique to detect DDH in acoustic non-invasive data. Using an acoustic noise of 10-2500Hz the data is collected and the performance of the model is calculated. Here, SVM gives an accuracy rate of 79%. ECG signal frequently contains noises and speckles. After filtering through various image processing techniques, machine learning algorithm is used in [6]. Here, LDA is used to features from the input ECG signals and SMV to recognize pattern. Specificity, Sensitivity and mean square root is calculated to measure to efficiency of the algorithms. In [7] the authors studied the data of socio-demographic, clinical and magnetic resonance imaging to predict Alzheimer's disease using GB algorithm. The algorithm predicted the disease on the aspect of socio economical state, age, education, gender, and mini-mental state exam. The model achieved 91.3% accuracy in prediction.

From the research gap, SVM, AB, LDA, and GB classifiers have been addressed to detect CKD. The data are preprocessed before the algorithms are implemented on them. Here, the main object of the study is to find the best and most optimal algorithm for prediction of kidney disease.

III. Research Methodology

A. System Overview:

This deals with the theoretical ability of the research work. It will give a piece of clear information about the concept of work. For the study, the dataset is collected form an online data repository. This data is cleaned using several pre-processing techniques. The feature selection is done on the dataset obtained from the repository. Once the data is cleaned and processed, it is divided in training set and test set data. Four machine learning classification algorithms are trained using the training data. After the algorithms are train, they are implemented on the test data to obtain prediction. In this study, the accuracy and performance of the prediction among the four algorithms are compared to determine the most efficient algorithm to predict the chronic kidney disease among the patients. The following Fig. 1 is illustrated to show the proposed model of this research.

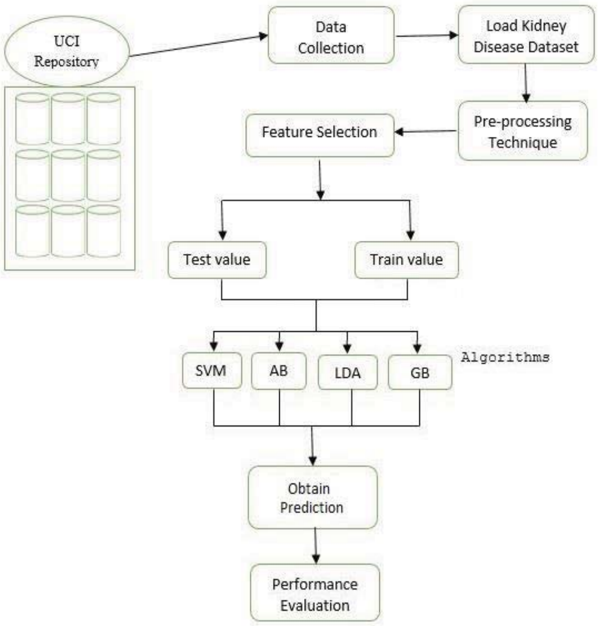

Fig. 1. The proposed model of kidney disease detection

B. Evaluation Criteria on Performance Measure Indices:

There are four different learning techniques used to represent the model together with some performance indices such as precision, recall, F1-score, specificity. The outcomes of performance measure indices are dependent on TP, TN, FP, and FN. [16-17]

True Positive = A list of reported cases exactly classified with CKD.

False Positive = A list of confirmed incidents incorrectly classified with CKD.

True Negative= A list of reported instances exactly classified with CKD.

False Negative= A list of confirmed instances exactly classified with CKD.

IV. Implementation

A. Different Machine Learning Libraries:

The proposed model implemented via Jupiter Notebook using basic python libraries are coded in the Python programming language including Panda - an open-source library that performs a superior function [8], Pyplot - generates the same view Matplotlib as similar to MatLab, Seaborn - helps to draw interesting and informative statistical graphics. A variety of machine learning techniques is used to solve real-world [9] problems used by Sklearn Python libraries [10].

B. Dataset Collection:

Data is thought of as the first and global parts of the research field. As a large number of patient records are collected from one of the most popular sites so as to get the most efficient outcomes. In this study, CKD disease dataset has been applied to predict deadly diseases from the UCI Machine Learning Repository [11]. We picked 25 individual features with 400 entries taken in a CSV format [12] from their database in where 250 are Kidney disease, KD class and 150 is a not-KD class. There are three types of data included such as float64(11), int64(1), and object(14). All features of categorical data such as objects have been converted into numbers by label encoding [13]. Take an example, we do label "not-KD" and "KD" with respect to 0 and 1.

C. Data Pre-processing:

After collecting a number of raw data from one of the most popular repositories, the dataset is to be applied in the preprocessing section. In this area, a number of missing values is detected to fulfill the data demand. Almost every column in the dataset has missing values that observe in the Table I. As the following dataset contains a number of missing values, it creates a number of complex situations in order to predict an accurate outcome. This major problem must be solved by two well-known methods in order to handle missing values. Among them, the median technique, the average of the two middle numbers, has been used to solve this vital problem. After applying this method, there have been observed no missing values depicted in Fig. 2.

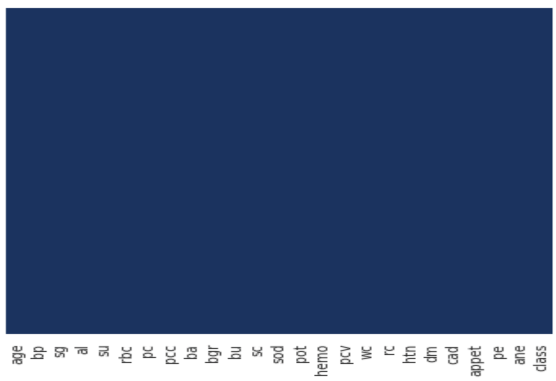

Fig. 2. No missing values in chronic kidney disease dataset

After the successful completion of the preprocessing technique, the proposed dataset attributes have been used for selection. To reduce the dimensionality, the process of feature selection has been employed. To get a better prediction rate [14-15], the narrow subsections of the appropriate features are extracted from its dataset.

D. Spliting Training and Testing set:

The dataset is partitioned into two parts: training and testing. In the training part, more than 70% of data is given to get predicted attributes value where the rest of the values is assigned for testing data. After finishing the training as well as the testing process, our machine is ready for doing classification [30]. To achieve an accurate outcome, four separate machine learning tools such as SVM, AB, LDA, and GB classifiers have been driven to predict disease stages.

TABLE I. VARIOUS ATTRIBUTES WITH MISSING VALUES

Features	Features Code	Number of missing values
Age	age	9
blood pressure	bp	12
specific gravity	sg	47
albumin	al	46
sugar	su	49
red blood cells	rbc	152
pus cell	pc	65
pus cell clumps	pcc	4
bacteria	ba	4
blood glucose random	bgr	44
blood urea	bu	19
serum creatinine	sc	17
sodium	sod	87
potassium	pot	88
hemoglobin	hemo	52
packed cell volume	pcv	70
white blood cell count	wc	105
red blood cell count	rc	130
hypertension	htn	2
diabetes mellitus	dm	2
coronary artery disease	cad	2
appetite	appet	1
pedal edema	pe	1
anemia	ane	1
class	class	0

E. Support Vector Machine:

Support vector machine [18] is also known as support vector networks and supervised learning models associated with machine learning algorithms. It also analyzes data using for regression and classification and the working process has been described by Bhagile et al. [19].

$$\mathrm{Y} = sign\ (\textstyle\sum_{i=1}^{N} y_i \alpha_i\ (\mathrm{x} * x_i) + b) \qquad (1)$$

From equation 10, $(\mathrm{x} * x_i)$ is known as labeled training and works as an input vector. Fig. 3 shows the working process of the implemented algorithm.

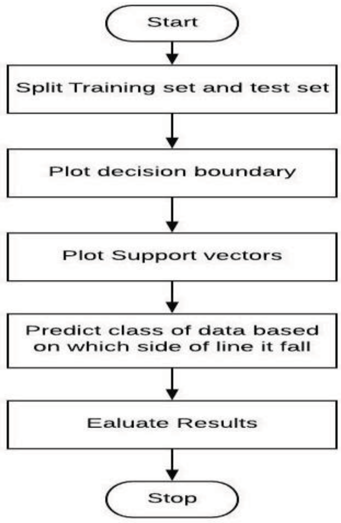

Fig. 3. The working process of SVM algorithm.

F. AdaBoost:

The algorithms of Boosting merge different weak classifiers to form strong classifiers to enhance the classification accuracy [20]. Another practical algorithm, Adaptive Boosting, had been suggested by Friedman et al. in 1997 by Fried and Schicher, although later in 2000. It was shown that LogitBoost overcame this situation through better generalizations. Boosting algorithms solves a variety of medical issues, namely the protein structure class detection in [21], cancer detection in [22] and breast cancer identification in [23]. The process of Adaptive Boost is shown in Fig. 4.

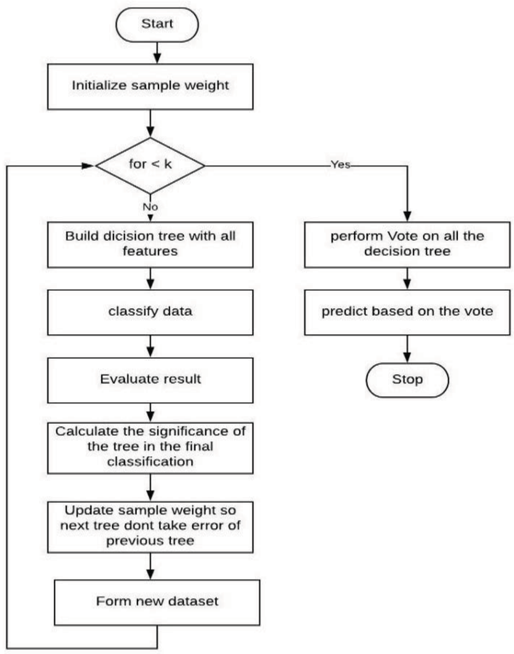

Fig. 4. The illustration process of AdaBoost algorithm.

G. Linear Discriminant Analysis

Linear Discriminant Analysis [24] has been a general form of Fisher's linear discriminant, a process included in statistics, pattern recognition, and machine learning to discover a linear combination of attributes that describes more than two groups of events. Each of C categories has a mean like μi and the same covariance like Σ. Then the scatter between variability of class might be described by the sample means of covariance class [25]. Fig. 5 and equation 2 show the steps of LDA.

$$\sum b = \frac{1}{C}\sum_{i-1}^{C}(\mu_i - \mu)(\mu_i - \mu)^T \quad (2)$$

H. Gradient Boosting:

Gradient boosting addresses the issues regarding a classification and regression [26]. That form a diagnostic model in the shape of an ensemble of weak forecasting analytics, usually trees of decisions. The model is structured in a phase-wise manner, congruous with other boosting method, and by allowing optimization of an arbitrary differentiable loss function. Our algorithm should be able to handle the addition of some new estimator hm (x) [27] in equation 3.

$$F_{m+1}(x) = F_m(x) + h_m(x) = y \quad (3)$$

The flow of the GB machine learning technique has been portrayed in Fig. 6.

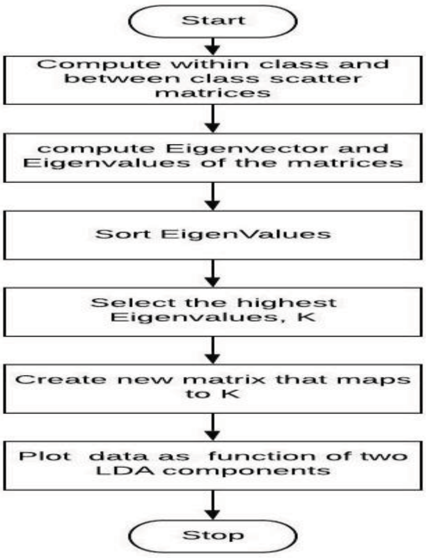

Fig. 5. The process of LDA algorithm

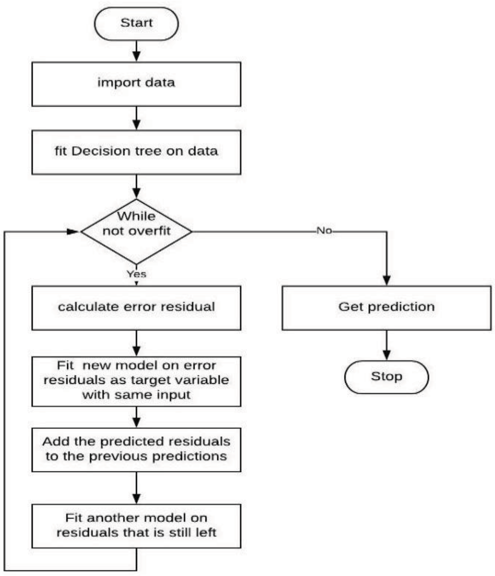

Fig. 6. The flow of the Gradient classifier.

V. Results And Discussion

A. Experimental Outputs Among Different Methods:

After the fruitful evaluation technique on the dataset, a large amount of data has been divided into training and testing. To detect a patient has a KD or not, four approaches of classification with regression were used. Performance metrics are used to justify diverse algorithm methods. Positive classification occurs if and only if a person has symptoms of kidney disease, a person does not have a kidney disease (KD), negative classification occurs. GB Classifier shows the highest outcome among all algorithms. The predictive outcomes have been depicted in Table II.

TABLE II. Performance Measurement Criteria

Dimension	**Support Vector Machine**	**AdaBoost**	**Linear Discriminant Analysis**	**Gradient Boosting**
ACC	99.56%	97.91%	97.91%	99.80%
SEN	99%	98%	98%	99%
SPE	99%	98%	98%	98%
PRE	99%	99%	99%	98%
NPV	97%	92.30%	92.30%	99%
FPR	0%	0%	0%	0%
FDR	0%	0%	0%	0%
FNR	0%	2.77%	2.77%	0%
F1	99%	98%	98%	99%
SD	0%	17.05%	17.05%	0%
MAE	0%	2.08%	2.08%	0%
MSE	0%	2.08%	2.08%	0%
RMSE	0%	14.43%	14.43%	0%

From the tables II, the accuracy rate among the four algorithms are compared. Here it is seen that the GB give the accuracy of 99.80% with 99% recall score. , and AB and LDA gives accuracy of 97.91%. The visualization of comparison is shown in Fig. 7.

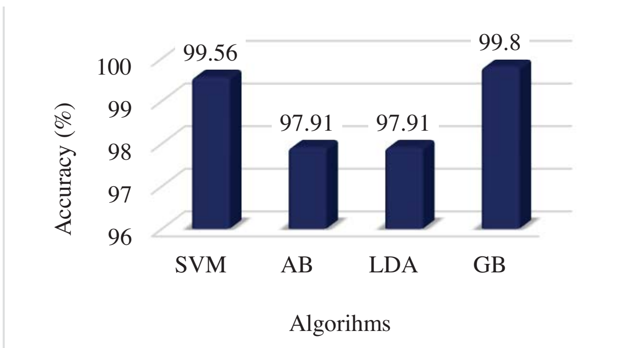

Fig. 7. Accuracy comparisons among different algorithms

B. Execuion Time Measurement over the Models:

The prediction rate is totally dependent on the dataset along with the model preprocessing technique. Besides, time takes a prominent role compared to others. As we know from Fig. 8 that the lowest predictable ratio of run time comes from SVM, On the other hand, GB takes the highest time to achieve a predictable score. Other popular algorithms generate lower time periods to find out the represented outcomes.

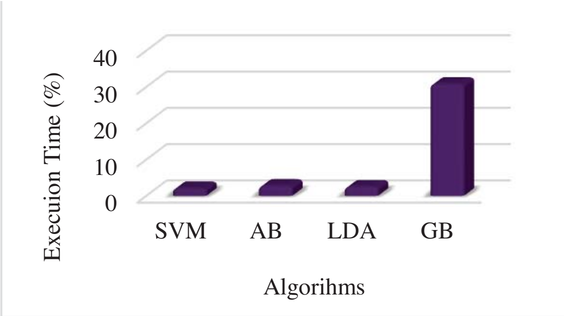

Fig. 8. A Prediction Time Comparison Occurred on Performed Algorithms.

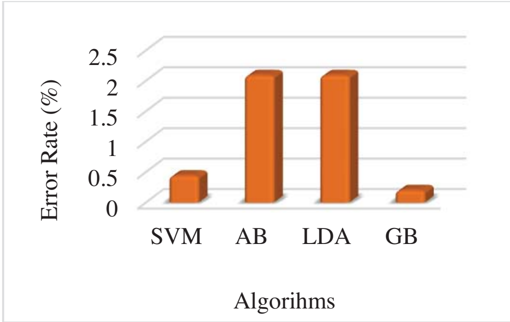

Fig. 9. The error rates of the introduced models.

C. Error Rate Among Various approaches:

Evaluating error rate [28] , accuracy of any algorithm can be measured. This helps gauging the performance of the algorithms. The four algorithms gave low errors rate altogether but the GB gave only 0.20% error after implementation on the dataset. However, AB and LDA generate the mid-lowest accuracy of 97.91% with 1.0 specificity score. The error comparison of the models are shown in Fig. 9.

D. Detection Using ROC and AUC Curves:

The diagnostic ability of the classifiers can be confused by using the confusion matrix and the receiver working property (ROC) curve [29]. The matrix of confusion is also referred to as the contingency matrix in the machine learning studies area. True Positive (TP) happens when the classifiers identified the data successfully while in False Positive (FP) the classifiers cannot identify the data. To measure the performance, AUC and ROC metrics were generated in Fig. 10.

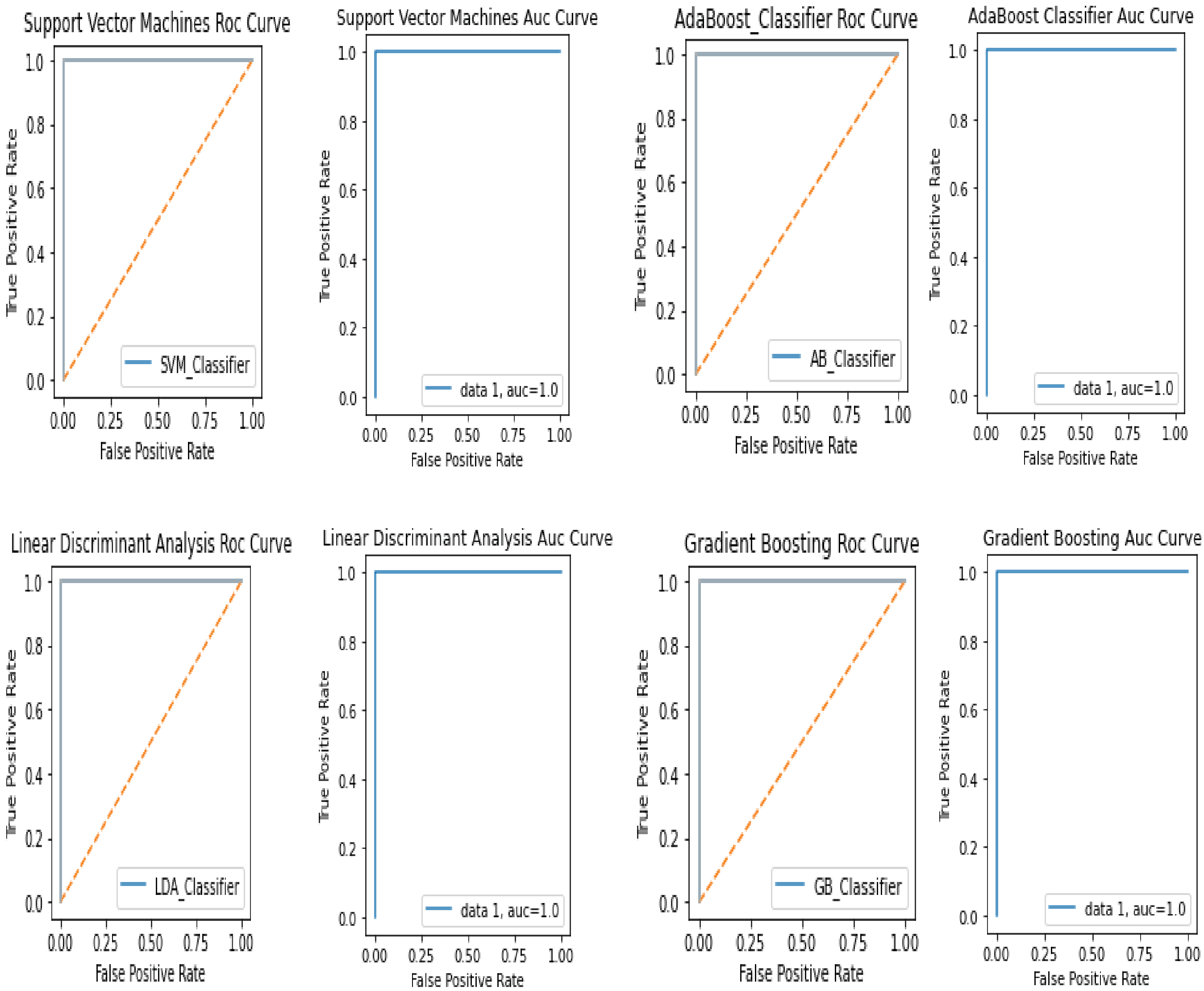

Fig. 10. ROC and AUC curves of all inroduced models.

ROC curve is diagnosed because the receiver working characteristic curve in which AUC is the vicinity under the ROC curve. If the rating of AUC is excessive, the performance of the version must be excessive, and vice versa. The ratings of Support Vector system, Linear Discriminant Analysis, and Gradient Boosting Classifier provide the best rating all of them in each ROC and AUC curves. Support Vector Machine that represents fourth model in the carve and provides 1.0 score in both curves. The score of Linear Discriminant Analysis (LDA) gives the mid-lowest score in both ROC and AUC curves. The score of Gradient Boosting Classifier gives the highest predictable score in both ROC and AUC curves.

VI. Conclusion

In this paper, four distinct algorithms were selected to get a precise expectation rate over the introduced dataset. Contrasting all presented approaches, the fruitful results have been gotten from GB classifier. These models effectively generate a 99.80% accuracy rate while AB, and LDA (97.91%) provides a low score. Besides the GB classifier requires more time compared to others to give a prediction and highest predictable score in both ROC and AUC curves. Since an exact pace of expectation is without a doubt reliant on the pre-processing strategy, the methods of the pre-processing must deal with cautiously to accomplish recognized outcomes precisely.

Acknowledgment

The authors are grateful and pleased to all the researchers in this research study.

References

[1] “Chronic kidney disease symptoms, treatment, causes & prevention-american kidney fund, ” http://www.kidneyfund. org/kidney disease/ chronic-kidney-disease-ckd/, (Accessed on 31/07/2020).

[2] F.M. Javed Mehedi Shamrat, Md. Asaduzzaman, A.K.M. Sazzadur Rahman, Raja Tariqul Hasan Tusher, Zarrin Tasnim “A Comparative Analysis Of Parkinson Disease Prediction Using Machine Learning Approaches” International Journal of Scientific & Technology Research, Volume 8, Issue 11, November 2019, ISSN: 2277-8616, pp: 2576-2580.

[3] F. M. Javed Mehedi Shamrat, Md. Abu Raihan, A.K.M. Sazzadur Rahman, Imran Mahmud, Rozina Akter, “An Analysis on Breast Disease Prediction Using Machine Learning Approaches” International Journal of Scientific & Technology Research, Volume 9, Issue 02, February 2020, ISSN: 2277-8616, pp: 2450-2455.

[4] Alghamdi, Raghad A. Makawi, Eman A. Albiety, Tayeb Brahimi, Akila Sarirete: Sensor Based Human Activity Recognition Using Adaboost Ensemble Classifier, In: Procedia Computer Science, Volume 140, (2018), Pages 104-111, ISSN 1877-0509, https://doi.org/10.1016/j.procs. 2018.10.298.

[5] M. E. Alam, N. Smith, D. Watson, T. Hassan and K. Neupane, "Early Screening of DDH using SVM Classification," 2019 SoutheastCon, Huntsville, AL, USA, 2019, pp. 1-4, doi: 10.1109/SoutheastCon42311.2019.9020565.

[6] Varatharajan, R., Manogaran, G. & Priyan, M.K. A big data classification approach using LDA with an enhanced SVM method for ECG signals in cloud computing. Multimed Tools Appl 77, 10195–10215 (2018). https://doi.org/10.1007/s11042-017-5318-1

[7] L. V. Fulton, D. Dolezel, J. Harrop, Y. Yan, C. P. Fulton, Classification of Alzheimer’s Disease with and without Imagery Using Gradient Boosted Machines and ResNet-50. Brain Sci. 2019, 9, 212.

[8] H. Zhang, C. Hung, W. C. Chu, P. Chiu and C. Y. Tang, "Chronic Kidney Disease Survival Prediction with Artificial Neural Networks," 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, pp. 1351-1356, 2018.

[9] J. Sivapriya, V. K. Aravind , S. S. Siddarth and S. Sriram, ” Breast Cancer Prediction Using Machine Learning,” International Journal of Recent Technology and Engineering (IJRTE), Vol. 8, Issue. 4, Nov 2019.

[10] P. Ghosh, M.Z. Hasan, O.A. Dhore, A.A. Mohammad and M. I. Jabiullah, “On the Application of Machine Learning to Predicting Cancer Outcome”, Proceedings of the International Conference on Electronics and ICT – 2018, organized by Bangladesh Electronics Society (BES), Dhaka, Bangladesh on 25-26 November, 2018, pp: 60.

[11] “UCI machine learning repository: Early stage of Chronic kidney disease dataset, https://archive.ics.uci.edu/ml/datasets/Chronic KidneyDisease, (Accessed on 05/06/2020).

[12] P. Yildirim, "Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction," 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), Turin, pp. 193-198, 2017.

[13] P. Ghosh, M. Z. Hasan and M. I. Jabiullah, “A Comparative Study of Machine Learning Approaches on Dateset to Predicting Cancer Outcome", Journal of the Bangladesh Electronic Society, Vol. 18, Num. 1-2, June, December 2018, ISSN: 1816-1510, pp:81-86.

[14] Zheng, Lijuan, H. Wang, and S. Gao, "Sentimental feature selection for sentiment analysis of Chinese online reviews", International Journal of Machine Learning and Cybernetics, Vol. 6, March, 2015.

[15] U. N. Dulhare and M. Ayesha, "Extraction of action rules for chronic kidney disease using Naïve bayes classifier," IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, pp. 1-5, 2016.

[16] F. M. J. M. Shamrat, P. Ghosh, M. H. Sadek, M. A. Kazi, S. Shultana “Implementation of Machine Learning Algorithms to Detect the Prognosis Rate of Kidney Disease” 2020 IEEE International conferenvce for Innovation in Technology, 2020.

[17] S. Manikandan, "Heart attack prediction system," International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, pp. 817-820, 2017.

[18] Y. Amirgaliyev, S. Shamiluulu and A. Serek, "Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods," 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), Almaty, Kazakhstan, pp. 1-4, 2018.

[19] S. A. Hannan, V. D. Bhagile, R. R. Manza, R. J. Ramteke,” Diagnosis and Medical Prescription of Heart Disease Using Support Vector Machine and Feed forward Backpropagation Technique, ” ,International Journal on Computer Science and Engineering, Vol. 02, No. 06, pp. 2150-2159, 2010.

[20] A. U. Islam and S. H. Ripon, "Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, pp. 1-6, 2019.

[21] Y. Cai, K. Feng, W. Lu and K. Chou, "Using LogitBoost classifier to predict protein structural classes", Journal of Theoretical Biology, vol. 238, no. 1, pp. 172-176, 2006.

[22] M. Dettling, and P. Buhlmann, "Boosting for tumor classification with gene expression data", Bioinformatics, vol. 19, no. 9, pp. 1061-1069, 2003.

[23] W. Zhang, F. Zeng, X. Wu, X. Zhang and R. Jiang, "A Comparative Study of Ensemble Learning Approaches in the Classification of Breast Cancer Metastasis," 2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, Shanghai, pp. 242-245, 2009.

[24] A. Nishanth and T. Thiruvaran, "Identifying Important Attributes for Early Detection of Chronic Kidney Disease," in IEEE Reviews in Biomedical Engineering, vol. 11, pp. 208-216, 2018.

[25] “A simple overview on Linear Discriminant Analysis,” https://en.wikipedia.org/wiki/Linear_discriminant_analysis, (Accessed on 08/06/2020).

[26] F. M. Javed Mehedi Shamrat, Zarrin Tasnim, Pronab Ghosh, Anup Majumder, Md. Zahid Hasan “Personalization of Job Circular Advertisement to Candidates Using Decision Tree Classification Algorithm” 2020 IEEE International conferenvce for Innovation in Technology, 2020.

[27] F. M. J. M. Shamrat, M. Asaduzzaman, P. Ghosh, M. D. Sultan, and Z. Tasnim, “A Web Based Application for Agriculture: “Smart Farming System," International Journal of Emerging Trends in Engineering Research, July 2020, ISSN: 2347-3983.