Original Article

CARDIAC DISEASE PREDICTION USING SMOTE AND MACHINE LEARNING CLASSIFIERS

[1]Sudipta Priyadarshinee and [2]*Madhumita Panda
[1]Research Scholar and [2]Associate Professor, Department of Computer Science, G.M. University, Sambalpur, Odisha, India
[1]sudiptapatel88@gmail.com and [2]mpanda.gmu@gmail.com

DOI: 10.47750/pnr.2022.13.S08.108

Abstract

Cardiovascular Disease (CVD) is presently the biggest reason of death globally. Clinical data analytics face a huge hurdle when attempting to predict cardiac disease. Massive amounts of raw data generated by the healthcare business are transformed into meaningful insights using machine learning techniques. The goal is to use of machine learning models that can enhance the predictability of cardiac patient survival. This paper employs eight machine learning classifiers: Decision Tree (DT), Extra Tree (ET), Random Forest (RF), Adaptive Boosting (AdaBoost), Ridge Classifier (RC), Linear Discriminant Analysis (LDA) and Light Gradient Boosting Machine (Light GBM) for prediction of cardiac disease. Synthetic Minority Oversampling Technique (SMOTE) is used to resolve the issue of unbalance dataset. Experiment outcomes demonstrate that SMOTE technique improves the accuracy of the selected classifier's output and Random Forest achieves highest accuracy with 95.12% applying SMOTE in predicting the survival of cardiac illness.

Keywords-Cardiovascular Disease (CVD), Machine Learning, Classifier, SMOTE

1. INTRODUCTION

According to World Health Organization (WHO), the most common reason of mortality in the world is cardiac disorder [1]. It is difficult to diagnose cardiovascular disease (CVD) because of a wide a number of risk elements, such as excessive cholesterol, high blood pressure, an inappropriate pulse rate, diabetes, and a number of other problems [2] and also the diagnosing is a difficult job which requires knowledge and ability. A wrong diagnosis could cause the patients' death or a person may become disabled. Medical experts and practitioners can forecast cardiac disease with the aid of the disease prediction model. The enormous volume of information that can be acquired with the use of digital gadgets (either by the patient or in the hospital) can be combined with machine learning techniques to diagnose and predict diseases.

Machine learning is one of the most well-liked and fastest developing sectors of artificial intelligence. By lowering the error in prediction and actual results, a variety of machine learning approaches are utilised to better grasp interaction between multiple components that is complicated and non-linear [3]. In order to predict cardiac illness many classification approaches are used in medical data mining [4]. Due to the ever-growing amount of medical information, machine learning algorithms must be employed to assist medical professionals in analysing data and producing diagnostic decisions that are exact and accurate.

The remainder part of the paper is organized as follows: A review of various heart-related works are presented in the Section 2. The proposed framework, as well as the various algorithms used to classify the given dataset are explained in the Section 3. Section 4 presents the outcomes of implementing the proposed methodology. The conclusion and future work are finally summarised in Section 5.

2. RELATED WORK

This section presents a discussion of various algorithms used in the prediction of cardiovascular disease.

The authors [5] have used K-nearest Neighbour, Naive Bayes, Decision Tree and Random Forest methods for accurate cardiac disease prediction. According to the findings of the experiments, KNN has achieved the highest accuracy (90.78%).

In this study [6], researchers have used two machine learning methods namely, Naive Bayes and Decision Tree to predict cardiovascular disorder. Decision Tree was determined to be most efficient model with a prediction accuracy of 91 percent.

In this paper [7] the clinical support system (CDSS) was investigated for cardiac failure analysis. In their study, they compared the performance of five machine learning algorithms such as Neural Network (NN), Support Vector Machine(SVM), Fuzzy Genetic Expert System, Classification and Regression Tree(CART) and Random Forest . With an accuracy of 87.6 percent, CART performed well among all other classifiers.

The research [8] is done taking real life data of healthy and ill patients of a local clinic using Stochastic Gradient Descent (SGD) Classifier, K-Nearest Neighbor Classifier, Random Forest Classifier, Logistic Regression Classifier and finally the mentioned models are combined using ensemble method where the classification is done using majority voting .The ensemble method resulted in a 90 percent accuracy rate as compared to others.

This paper [9] works on Navies Bayesian classification techniques and compares it with different other techniques like SMO (Sequential Minimal Optimization), Bayes Net and MLP (Multi-Layer Perception). Effective Performance was seen with Naïve Bayes giving an accuracy of 89.77%. The results are finally encrypted using AES security algorithms and then presented to user.

In this study [10], researchers have used four machine learning methods Logistic Regression, Naive Bayes, Decision Trees and Random Forest to predict cardiovascular disorder taking data set from UCI machine learning repository. According to the experimental findings, the Random Forest technique has achieved the maximum accuracy (90.16%) as compared to other algorithms.

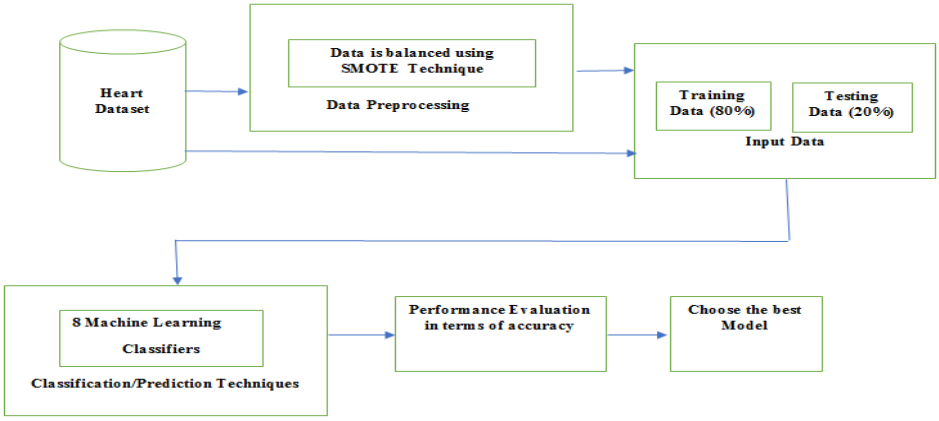
The researchers [11] have employed bagging and boosting in improving the prediction of heart disease for weak classifiers. Accuracy was further seen to be improved using feature selection techniques.

This study [12] focuses on the risk factor-based prediction of cardiac disease. The effectiveness of prediction methods like Binary Logistic Classification, K-Nearest Neighbors and Naive Bayes was assessed. These fundamental classifiers are contrasted with ensemble modelling methods like stacking, boosting, and bagging. With a final accuracy of 75.1%, the Random Forest, KNN and SVM stacked model was found to be the most successful.

3. Proposed Methodology

A cardiac disease dataset from Kaggle [13] which includes 299 occurrences of 13 characteristics are used. Based on the split criteria, we divide the dataset into a training set (comprised of 80% of the data) and a test set (consisting of 20% of the data). We then run eight distinct machine learning methods on the imbalance data set to compare their performance. The SMOTE technique is then employed on the same dataset to address the issue of data imbalance. At, the end, the algorithms of machine learning are then employed on the balanced dataset and the rate of accuracy is calculated. The primary goal was to find the method that could best classify the given dataset.

Fig.1. Proposed Frame Work



3.1. Description of Dataset

The Kaggle [13] heart dataset has been used for the experiment. The dataset includes 299 patients' medical records with heart problems who were accumulated during the time of follow-up, with each profile of the patient containing 13 clinical attributes. There are 194 men and 105 women among the 299 records. All of the patients are beyond the age of 40. In the target class, 1 denotes the deceased and 0 denotes the alive. Table 1 provides an overview of the data set.

Table. 1 Specification of Dataset

Sr No	Attributes	Description	Measured In	Range
1	Age	The patient's age	Years	40 - 95
2	Anaemia	Red blood cell or haemoglobin deficiency	Boolean	0, 1
3	Creatinine phosphokinase (CPK)	CPK enzyme levels in the blood	mcg/L	23 -7861
4	Diabetes	Whether the patient has diabetes	Boolean	0, 1
5	ejection fraction	percent of blood leaving	Percentage	14 - 80
6	high_blood_pressure	If someone has hypertension	Boolean	0, 1
7	Platelets	The number of platelets in the blood	kilo platelets/Ml	25.01-850.00
8	serum creatinine	Creatinine concentration in the blood	mg/dL	0.50 -9.40
9	serum sodium	Sodium levels in the blood	mEq/L	114 -148
10	Sex	Woman or man	Binary	0, 1

3. 2 Imbalance Nature of Dataset

There are 203 instances of the negative class (0) and 96 instances of the positive class (1) in the given dataset. Thus, there is an unbalanced distribution of classes. One of the key causes of decreased classification model accuracy is unequal distribution. As a result of their imbalance problem in dataset, many machine learning algorithms fail to discover good patterns for both positive and negative classes. Furthermore, as the positive classes are few in number, the outcomes provided by this class are typically unsuccessful. The imbalanced characteristics of the presented dataset is well tackled by SMOTE approach [14], which is one of the proposed work's significant contributions.

3.3. Classification Algorithms

The following eight machine learning algorithms have been used in this work**.**

References	Model	Description
15	**Random Forest (RF)**	It's a form of machine learning called supervised learning, and it's used to do things like classifying data or making predictions. For classification problems, RF gives the most popular tree class as its result.
16	**Decision Tree (DT)**	It is useful for both classification and regression. A decision tree is applied to construct a structure that looks like a tree. It has mainly three node (1) Root node (2) interior node (3) Leaf node.
17	**Extra Tree (ET)**	An ensemble machine learning algorithm is Extra Tree. It is a particular bundle of decision trees and is attached with another ensembles of decision tree techniques, including bootstrap aggregation and random forest.

18	**Adaptive Boosting (AdaBoost)**	Adaptive Boosting is referred to as AdaBoost. It is applied in conjunction with additional methods to enhance their performance. It works on boosting to train weak algorithms into strong algorithms.
19	**Logistic regression (LR)**	Logistic regression is primarily concerned with classification problems. It is a statistical model and a predictive analysis algorithm. It uses the idea of probability to analyze binary data, where the values of one or more factors determine the result.
20	**Linear Discriminant Analysis (LDA)**	To execute a classification task, Linear Discriminant Analysis is used. Because the fundamental goal of Linear Discriminant Analysis is used to distinguish examples of classes by moving them linearly to a different feature space.
21	**Light Gradient Boosting Machine (LightGBM)**	It is a memory-efficient, high-performing gradient-boosting framework that makes use of decision tree branches to boost model quality. While most boosting algorithms split a tree along its depth, LightGBM divides it along its leaves to get the optimum solutions.
22	**Ridge Classifier**	It uses a regression methodology derived from the Ridge regression method to address the problem by altering the label data to lie inside the range [-1, 1]. Using the prediction value, multi-output regression determines which class to focus on while working with multiclass data.

4. RESULTS AND DISCUSSION

This section displays the experimental outcomes of all heart patients with the dataset taken from Kaggle [13]. All experiments were carried out in a simulation environment using python. First, on the given dataset, the experiment is carried out using eight machine learning algorithms. Then SMOTE technique is employed to balance the dataset. Afterwards, on the balanced dataset, machine learning approaches are again applied and accuracy is estimated.

4.1. Performance Metrics

Our proposed model is estimated in terms of accuracy [23] using the following equation.

$$Accuracy = \frac{TP(True\ Positive)+TN(True\ Negative)}{TP+TN+FN(False\ Negative)+FP(False\ Positive)}$$

On the suggested model employing the heart dataset [13], two various simulation methodologies have been implemented (a) without oversampling (b)with oversampling. Table 2 displays the hyperparameter values utilised in this investigation of algorithms.

Table 2. Hyperparameter Values of Machine Learning Algorithms

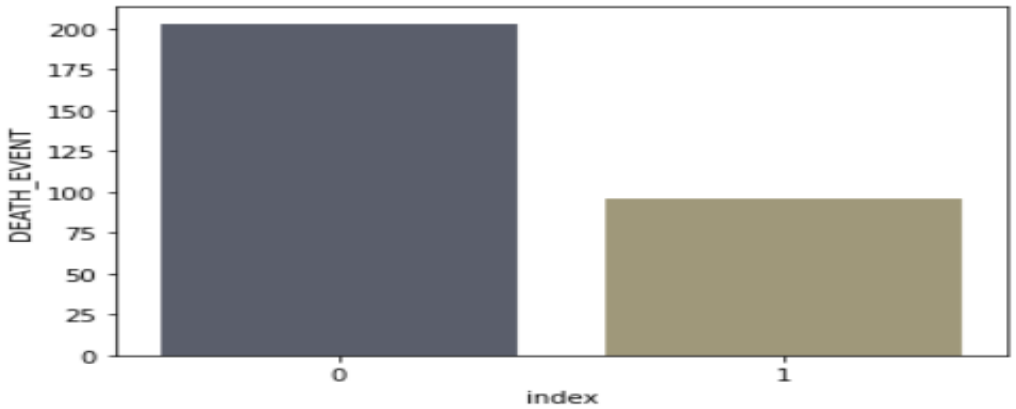
Algorithms	**Hyperparameter Values**
Random Forest	n_estimator=100, random_state=0
Decision Tree	max_depth=2, Random_state=1
Extra Trees	max_iter=500, Random_state=42
Adaptive Boosting	n_estimators=500,learning_rate=0.05, Random_state=42
Logistic regression	max_iter=140, random_state=1
Linear Discriminant Analysis	solver='lsqr', shrinkage='auto'
Light GBM	n_estimator=200, random_state=123

Ridge Classifier	max_iter=200, random_state=0

4.2 Performance of Classifiers Accuracy without SMOTE

The experiments were first carried out on imbalance dataset as shown in Fig.2 to find the accuracy using 8 machine learning methods as listed in section 3.3.

Fig. 2. Imbalance Heart Dataset



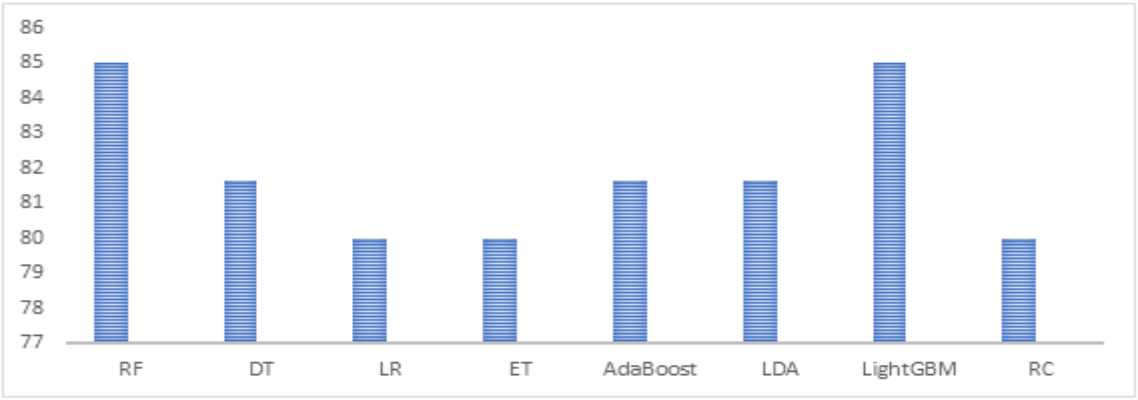
The execution of the classifier's accuracy on imbalanced data set is shown below in Table 3.

Table 3. Classification Accuracy Without SMOTE

Classifier	Accuracy (%)	Inaccuracy (%)
Random Forest	85	15
Decision Tree	81.66	18.34
Logistic Regression	80	20
Extra Tree	80	20
Adaptive Boosting (Adaboost)	81.66	18.34
Linear Discriminant Analysis	81.66	18.34
Light GBM	85	15
Ridge Classifier	80	20

Table 3 and graphical results of Fig.3 clearly shows that Random Forest and LightGBM gave the highest accuracy of 85 percent.

Fig. 3. Classification Accuracy Without SMOTE



4.3 Performance of Classifiers Accuracy with SMOTE

The SMOTE pre-processing technique is considered as one of the most dependable and effective pre-processing strategies in the machine learning and information mining industry [14]. To expand the amount of data instances, using Euclidean distance, SMOTE generates random false minority data from its closest neighbours. Because new instances are formed based on original features, they become identical to the original data. In this study, after SMOTE technique is used on the dataset, it raises the number of data samples from 299 to 406 and the dataset is now balanced as seen in Fig. 4.

Fig. 4. Balanced Heart Dataset

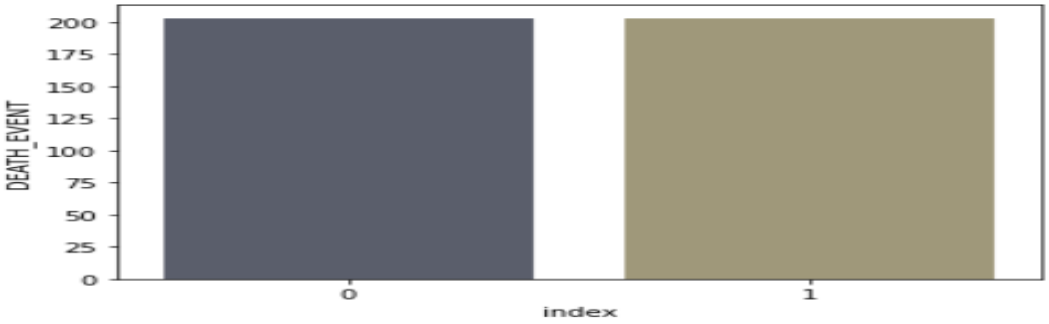


Table 4 displays the accuracy of the classifier's after using SMOTE technique.

Table 4. Classification Accuracy with SMOTE

Classifier	Accuracy (%)	Inaccuracy (%)
Random Forest	95.12	4.88
Decision Tree	91.4	8.6
Logistic Regression	82.92	17.02
Extra Tree	90.24	9.76
Adaptive Boosting (Adaboost)	92.68	7.32
Linear Discriminant Analysis	82.92	17.08
Light GBM	90.24	9.76
Ridge Classifier	84.14	15.86

After balancing the heart dataset, it is found that all the above classifier's performance is increased. Random Forest gives the best accuracy of 95.12 percent. Ada Boost had the second highest accuracy of 92.68 percent.

Fig. 5. Comparison of Classifier Accuracy with and without SMOTE

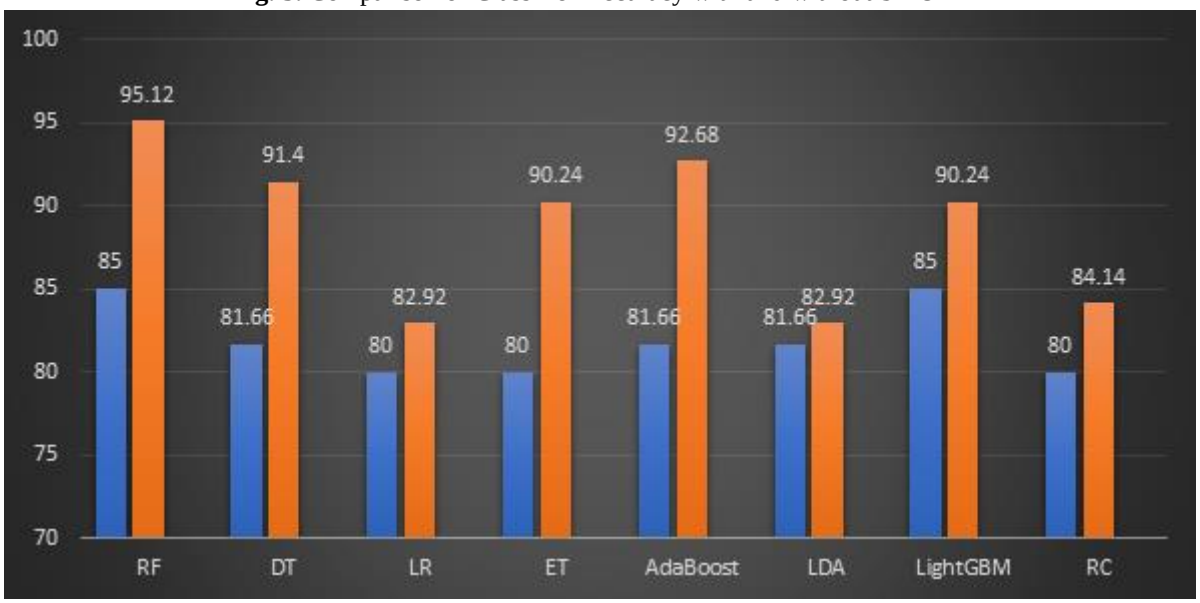


Fig.5 gives a graphical overview of accuracy of all algorithms with and without SMOTE. It is seen that the accuracy of algorithms has been increased after applying SMOTE.

5. CONCLUSION AND FUTURE WORK

The use of algorithms of machine learning to process raw health data from the heart will help save the lives of cardiac patients. The mortality rate can be managed by evaluating variables that contribute to heart failure and taking preventive actions. This work proposes a machine learning-based strategy that is both effective and efficient for predicting the survival of heart patients. This paper employs eight classification models such as Decision Tree (DT),Extra Tree (ET),RandomForest(RF),AdaptiveBoosting(AdaBoost),RidgeClassifiers(RC),LinearDiscriminantAnalysis (LDA),Logistic Regression(LR) and Light Gradient Boosting Machine (LightGBM**)** to predict cardiac illness. SMOTE is used to address the issue of class imbalance. It's also been discovered that using the SMOTE technique improves the accuracy of the selected classifier's output and Random Forest achieves highest accuracy with 95.12% with SMOTE in prediction of cardiac disease. In future we'll apply a variety of other classification techniques, particularly ensemble methods to enhance the performance of the models and will also experiment with several other types of balancing approaches to address the unbalanced situation.

REFERENCES

[1] WHO. The Top 10 Causes of Death. Accessed: Dec. 30, 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10- causes-of-death

[2] C. Fryar, T.-C. Chen, and X. Li, ‘‘Prevalence of uncontrolled risk factors for cardiovascular disease: United states, 1999-2010,’’ in NCHS Data Brief, vol. 103. Aug. 2012, pp. 1–8.

[3] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE. 2017;12(4):e0174944.

[4] Ramalingam VV, Dandapath A, Raja MK. Heart disease prediction using machine learning techniques: a survey. Int J Eng Technol. 2018;7(2.8):684–7.

[5] Shah, Devansh, Samir Patel, and Santosh Kumar Bharti. "Heart disease prediction using machine learning techniques." SN Computer Science 1.6 (2020): 1-6.

[6] Krishnan, Santhana, and S. Geetha. "Prediction of Heart Disease Using Machine Learning Algorithms." 2019 1st international conference on innovations in information and communication technology (ICIICT). IEEE, 2019.

[7] G. Guidi, M. C. Pettenati, P. Melillo, and E. Iadanza, ‘‘A machine learning system to improve heart failure patient assistance,’’ IEEE J. Biomed. Health Informat., vol. 18, no. 6, pp. 1750–1756, Nov. 2014.

[8] Atallah, Rahma, and Amjed Al-Mousa. "Heart disease detection using machine learning majority voting ensemble method." 2019 2nd international conference on new trends in computing sciences (ictcs). IEEE, 2019.

[9] Repaka, Anjan Nikhil, Sai Deepak Ravikanti, and Ramya G. Franklin. "Design and implementing heart disease prediction using naives Bayesian." 2019 3rd International conference on trends in electronics and informatics (ICOEI). IEEE, 2019.

[10] Rajdhan, Apurb, et al. "Heart disease prediction using machine learning." International Journal of Research and Technology 9.04 (2020): 659-662.

[11] Latha, C. Beulah Christalin, and S. Carolin Jeeva. "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques." Informatics in Medicine Unlocked 16 (2019).

[12] Shorewala, Vardhan. "Early detection of coronary heart disease using ensemble techniques." Informatics in Medicine Unlocked 26 (2021): 100655.

[13] https://www.kaggle.com/andrewmvd/heart-failure-clinical- data/discussion/193109

[14] R. Blagus and L. Lusa, ‘‘Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models,’’ BMC Bioinf., vol. 16, no. 1, pp. 1–10, Dec. 2015.

[15] L. Breiman, ‘‘Random forests,’’ Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[16] Shouman, Mai, Tim Turner, and Rob Stocker. "Using decision tree for diagnosing heart disease patients." Proceedings of the Ninth Australasian Data Mining Conference-Volume 121. 2011.

[17] A. Sharaff and H. Gupta, ‘‘Extra-tree classifier with metaheuristics approach for email classification,’’ in Proc. Adv. Comput. Commun. Comput. Sci. Singapore: Springer, 20

[18] El Hamdaoui, Halima, et al. "Improving Heart Disease Prediction Using Random Forest and AdaBoost Algorithms." International Journal of Online & Biomedical Engineering 17.11 (2021).

[19] C. R. Boyd, M. A. Tolson, and W. S. Copes, ‘‘Evaluating trauma care: The TRISS method,’’ J. Trauma, Injury, Infection, Crit. Care, vol. 27, no. 4, pp. 370–378, Apr. 1987.

[20] https://machinelearningmastery.com/linear-discriminant-analysis-for-machine-learning/

[21] https://machinelearningmastery.com/light-gradient-boosted-machine-lightgbm-ensemble/

[22] https://www.datatechnotes.com/2020/07/classification-example-with-ridge-classifier-in-python.html

[23] Saboor, Abdul, et al. "A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms." Mobile Information Systems 2022 (2022).