A Performance Analysis of Cardiovascular Disease using Machine Learning Approaches

Maria Sultana Keya[1], Minhaz Uddin Emon[2], Faridul Islam Suny[3], Himu Akter[4], Sabiha Jannat Anni[5], Raihana Zannat[6] and Ohidujjaman[7]

[1,2,3,4,5,7] *Department of Computer Science and Engineering,*
[6]*Department of Software Engineering*
Daffodil International University
Dhaka 1207, Bangladesh
Corresponding Author: Ohidujjaman

ABSTRACT : *Cardiovascular disorder (CVD) is a generic term for complications that influence the heart and the blood vessels. Cardiovascular disorder is more prevalent in people over the age of 50, and the chance of having it grows as you reach adulthood. Timely identification of various potential problems of injury can decrease the disease. Therefore, in this article developing a model that can forecast the risk of cardiovascular disorder (CVD) in patients with the greatest analysis and for prediction. Therefore different types of machine learning algorithms such as Random Forest, Logistic Regression, Decision Tree, Bagging, AdaBoost, Naive Bayes. The best analysis is found in AdaBoost classifier. This algorithm shows the best confusion matrix in the outcome.*

KEYWORDS: *AdaBoost classifier, cardiovascular, correlation metrics, feature extraction, machine learning.*

Date of Submission: xx-xx-xxxx Date of acceptance: xx-xx-xxxx

---

I. INTRODUCTION

The value of 140/90 millimeters of mercury, blood pressure, the continuous elevation of arterial pressure is defined as high blood pressure. A broad range of long-term complications are available: heart disease, stroke, kidney failure, and so on. It is common in the world for coronary heart disease, heart failure, coronary artery disease, ischemic heart disease, cardiovascular disease, left heart hypo-plastic syndrome, and other forms of heart disease. A study by the World Health Organization (WHO) reports that the main cause of death is now cardiovascular disease (CVD) and an estimated 17.9 million people are killed annually [4].Due to the plaque on the artery ruptures and a clot then forms, blocking circulation, most heart attacks occur. The heart disease diagnosis was based on the medical experience of the patients. Diabetes is regarded as one of the seven main controllable risk factors for CVD by the American Heart Association. In order to improve their diet and other lives, all people at high risk are encouraged. Factors of style and potentially take drugs to control elevated characteristics [5]. Large datasets are obtained by the healthcare industry, hence efficient techniques such as machine learning are crucial to manage them adequately. Machine learning is an application of artificial intelligence and is a type of algorithm that predicts results accurately [18]. The use of machine learning in the study of cardiovascular disease is the best strategy to diagnosis, prediction, management, and other aspects of clinical administration. Machine learning can help people make a tentative decision about CVD according to their daily physical examination findings. In this research the risk factor of cardiovascular diseases are analyzed by using some machine learning model.

II. LITERATURE REVIEW

Dominic et. al. proposed to realize the alignment of adventure lines with heart diseases[6]. Naive Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), Neural Network (NN), Association Rules are being used where DT shows the highest accuracy 98%.Wosiak et. al. performed to the rhythm unsupervised special item and collecting to promote T of statistical analysis[7]. Statistical inferences are evaluated by result and that is performing on clustering. Brown et. al. identified to count the present instance and check the risk of cardiovascular diseases[8]. The 244 incidental things were recognizing several meta-analyses that observed three outcomes. No evidence is found that the risks of CVD increment. Restricted in methodology. Kim et. al. used as compatible and dependable vaticinatined method obligate in diagnosis of CVD [9]. Restricted to collect sample data in a particular association that is not plucking adequate data from acceptable hospitals. Alic et. al. performed to distribute diabetes and cardiovascular affliction taking Bayesian Networks (BNs) and Artificial Neural Networks (ANNs). Best accuracy stands for alignment CVD 97.92% and diabetes 99.51% parallelism of ANNs and BNs to detect excellent choice for acquiring the highest accuracy [10]. Angayarkanni et. al. utilized several machine learning algorithms to forecast CVD within non diabetes and diabetes patients [11]. Low, medium and high risk of CVD are performed. Tadesse et. al. identified data driven access that may automate that system[12]. For image input CNN skeleton is planned. This assessment is affirmed on diversity ECG dataset and competitive execution is gained. Restricted in features that are not generalized. Harvay et. al. proposed several diagnosis and treatment of four dominant CVDs[13]. CVDs risk creator appraisement for women that performed to object intervention patterns to restriction advancement. Gayathri et. al. performed diagnosis patterned can conduct to development [14]. More than 90% accuracy is gained by some strategy. An actual dataset bearing medical list by using association rules. Few experiment remittance is searched to lead the more experiment in equal approach. Data cleaning is a serious problem for the heart disease database. Mathew et. al. found an indicative upliftment in heart rate [15]. Halting thickness pest is a significant barrier to defeat in shots to alleviate the fardel of CVD in the people. Miller et. al. [16] identified bio markers of cardiovascular diseases [16]. Many echo cardiography labs do not feature diastolic operation by investigation. BNP (Brain natriuretic peptide) is suitable bio markers for evaluating usefulness of treatment of CVD. Restrictions the process handy for use. Moses et al. proposed numerous patterns for CVD diagnosis applying data mining by an ECG signal [17]. LZW (Lempet Ziv Welch) and Huffman coding are used for ECG compression. EM clusters used to remove in the spatial domain ensure more proper medical diagnosis.

III. PROPOSED MODEL

According to figure 1 it is revealed a separate concise overview of each portion of the proposed model.

Input Data: This research work used 69964 data where 12 attributes are observable and then all of them are floating data. Moreover there is a decision class and class variable. The criteria are 0 and 1 where the number of 0 is 36834, the number of 1 is 34150.

Missing Value Handling: For missing value handling there have used mean value. Usually there are two steps to overcoming missing values, whether abandoning the data or replacing the missing data. It has been dealing with the missing value here by substituting the mean value with the outcomes.

Split Data: In this research 70% data are used for train model and 30% data are used for testing purposes.

Feature Extraction: Feature extraction is a term generally for approaches to develop significant variables to solve these problems while still classifying the functionality with adequate accuracy. The step to a successful model construction is properly optimized feature extraction.

Classifier: This study worked with numerous classifiers for machine learning. The classifier is used by Random Forest, Decision Tree, Logistics, Bagging, AdaBoost, and Naive Bayes where the random forest is accessible. The algorithm of the random forest consists of various decision trees, each with the same nodes, but using different data that leads to different leaves.In order to find a response, it merges the decisions of several decision trees, indicating the aggregate of all these decision trees.The bagging classifier is also used here for the meta-estimator ensemble that applies base classifiers on random subsets of the original dataset and then compiles individually. Then it is to allow exponentially increasing precision, there have used the AdaBoost classifier to incorporate it with other classifiers. To get a wider spectrum of classification tasks, naive bayes is used here.

Classification Report: In the classification report it was able to find out the confusion matrix for all algorithms.

Best Algorithms: In this study the result depends on some part of this research. However the algorithm which gives the best true positive, false positive, true negative, and false negative is the best algorithm in this analysis.

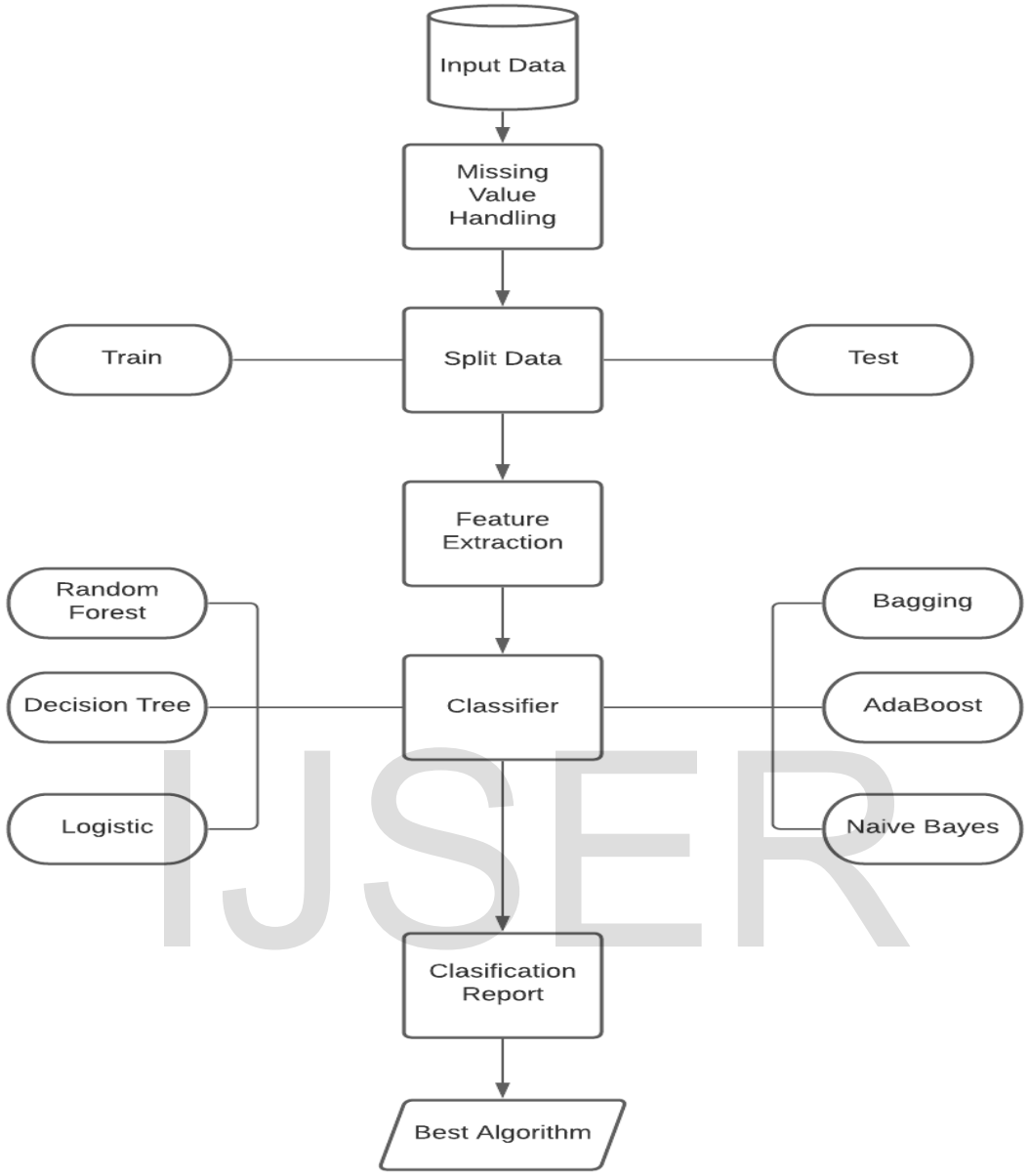

Fig.1. Proposed Model

IV. RESULT DISCUSSION AND ANALYSIS

Feature Extraction: This learning although has been working on machine learning with a vast quantity of data, and has chosen feature extraction. Extraction of features is a plan that outlines the data key features and decreases the initial dataset. However there have envisioned the findings in Fig. 2 by incorporating 2D PCA (two dimensional principal component analysis) through feature extraction where there is no cardio disease in the portion outlined in blue, the red part shows that there is a cardiac disease.

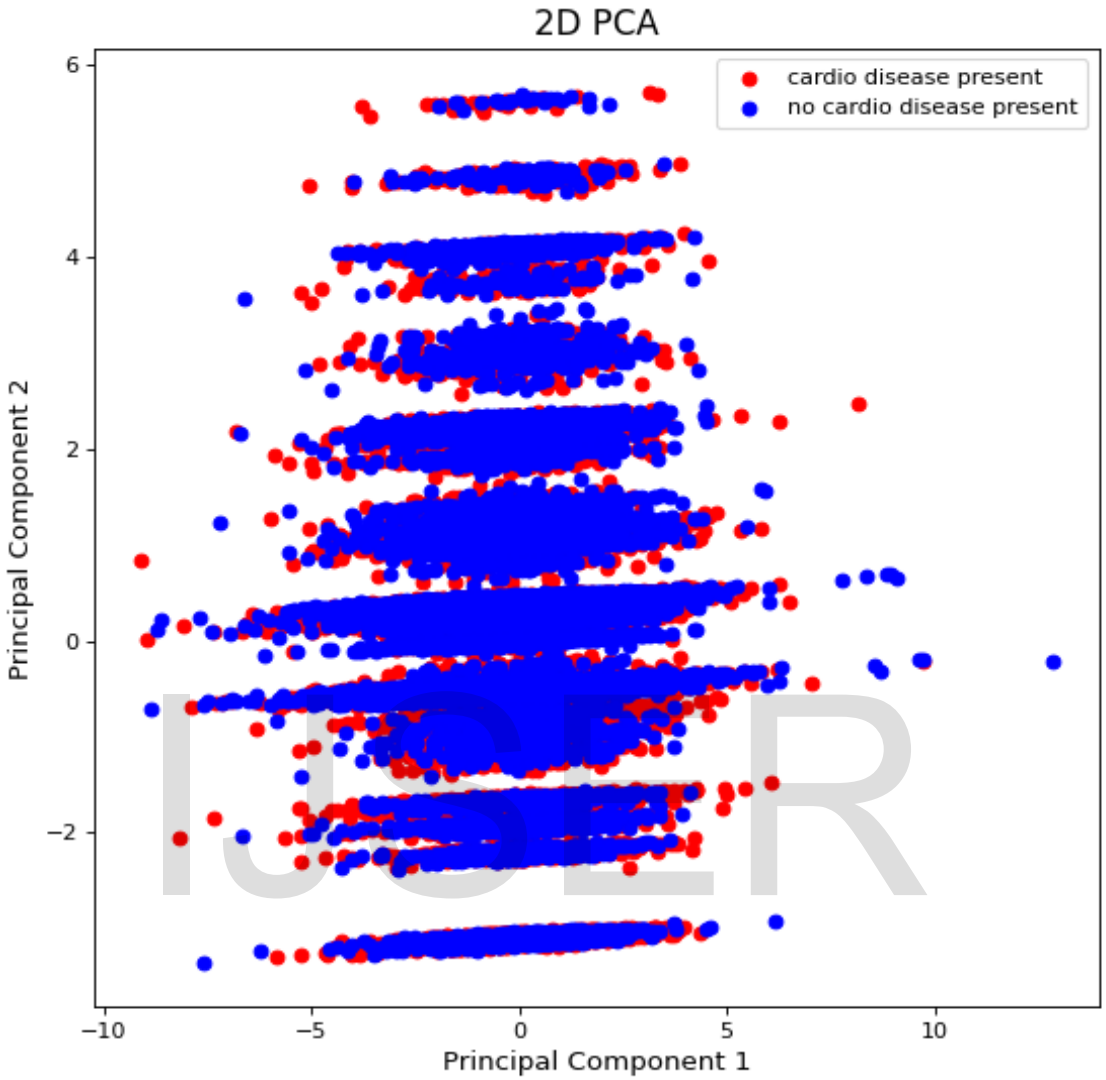

Fig. 2. Feature extraction using principal component analysis

Correlation Metrics: The Fig. 3 demonstrates the research's correlation matrix. Correlation metrics examine whether a connection persists among two or more variables or not. It emphasizes how they can impair cardiovascular disease in the diagram of the following correlation matrix, using several other attributes, including edge, height, and weight

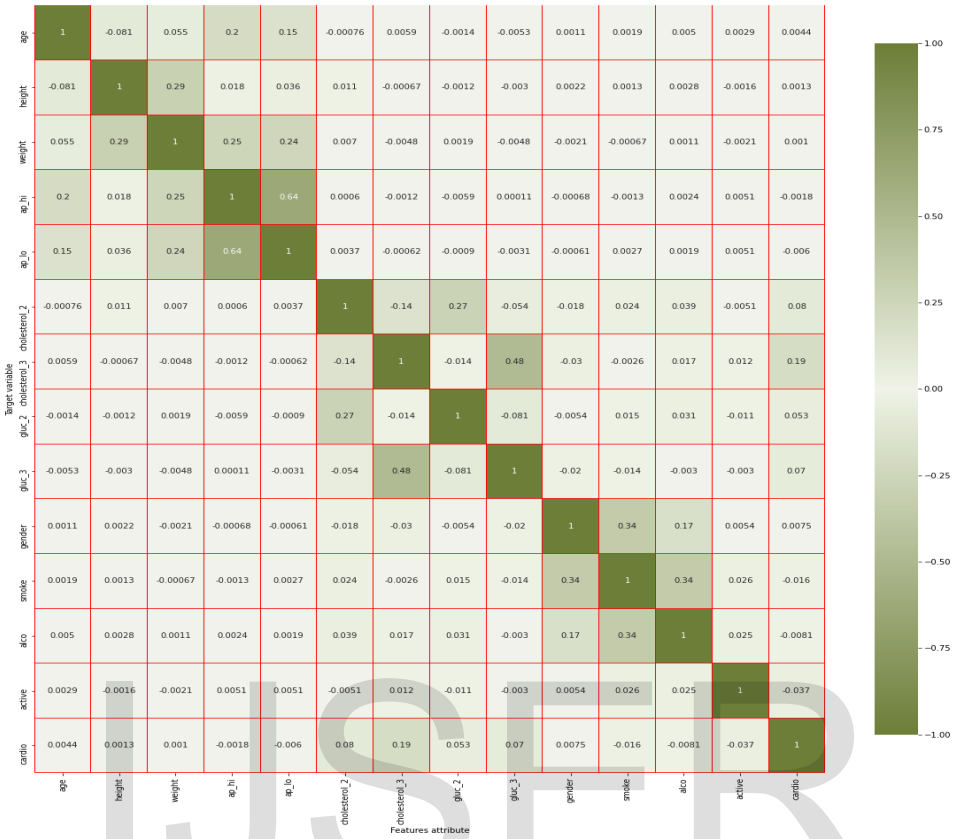

Fig. 3. Correlation metrics among feature attribute to target variable

Results: In this research table 1 represents the confusion matrix of applying classifier. The best result is shown in AdaBoost classifier. In this analysis the predictive negative, predictive positive, actual negative, and actual positive highest values are found from AdaBoost classifier. By using adaboost classifier the true negative value is 8864, and true positive value is 3626.

Table 1: Confusion metrics analysis for applying classifier

Model Name	Label	Predictive Negative	Predictive Positive
Random Forest	Actual Negative	6467	4170
	Actual Positive	5097	5262
Logistic Regression	Actual Negative	6467	4170
	Actual Positive	5097	5262
Decision Tree	Actual Negative	5694	4943
	Actual Positive	4922	5437
Bagging	Actual Negative	6858	3779
	Actual Positive	5754	4605
AdaBoost	Actual Negative	**8864**	1773
	Actual Positive	6733	**3626**
Naive Bayes	Actual Negative	7408	3229
	Actual Positive	5768	4591

V. CONCLUSION

The empirical analysis indicates that diverse approaches to the prediction of CVD in machine learning have been tested. The outcome shows that a more specialized approach to confusion metrics is to be implemented. However

more related traits are used to enhance the early prediction investigation and to define the risk thresholds for effective treatment improvements. Moreover the details are found in this study that the changes in the number of attributes is varied. The investigation relies on an open dataset in the implementation of the machine learning algorithm where it develops CVD prediction. This research finds the correctness of the algorithm applied and then define the effective technique. For the better prediction performance for the diagnosis of CVD, the future hybrid algorithms or improved computer algorithms will be implemented.

REFERENCES

[1]. Elizabeth G. Nabel, M.D., "Cardiovascular disease", The New England Journal of Medicine, 349(1), 60-72, 2003.
[2]. Jesmin Nahar and Tasadduq Imam et al, "Association rule mining to detect factors which contribute to heart disease in males and females", Journal of Expert Systems with Applications Vol.40, PP.1086–1093, 2013.
[3]. Swati Shilaskar et al, "Feature selection for medical diagnosis: Evaluation for cardiovascular diseases", Journal of Expert System with Application, Vol.40, PP.4146-4153, 2013.
[4]. Petra A. Karsdorp and Merel Kindt et al, "False Heart Rate Feedback and the Perception of Heart Symptoms in Patients with Congenital Heart Disease and Anxiety", International Journal of behavioral Medicine, Vol.16, PP.81-88, 2009.
[5]. Dominic, V., Gupta, D., Khare, S., "An effective performance analysis of machine learning techniques for cardiovascular disease" Applied Medical Informatics., 36(1), 23-32, 2015.
[6]. Wosiak, A., Zakrzewska, D., "Integrating correlation-based feature selection and clustering for improved cardiovascular disease diagnosis", Complexity, Volume 2018, Article ID 2520706, https://doi.org/10.1155/2018/2520706
[7]. Brown, M. C., Best, K. E., Pearce, M. S., Waugh, J., Robson, S. C., Bell, R., "Cardiovascular disease risk in women with preeclampsia: systematic review and meta-analysis", European journal of epidemiology, 28(1), 1-19, 2013.
[8]. Kim, H., Ishag, M. I. M., Piao, M., Kwon, T., Ryu, K. H., "A data mining approach for cardiovascular disease diagnosis using heart rate variability and images of carotid arteries", Symmetry in Systems Design and Analysis, 13 June 2016.
[9]. Ali c, B., Gurbeta, L., Badnjevic´, A., " Machine learning techniques for classification of diabetes and cardiovascular diseases" 6th Mediterranean Conference on Embedded Computing (MECO) (pp. 1-4), June, 2017.
[10]. Angayarkanni, G., Hemalatha, S., "Towards Analyzing the Prediction of Developing Cardiovascular Disease using Implementation of Machine Learning Techniques", International Journal for Research in Applied Science & Engineering Technology (IJRASET) Vol. 8 Issue VIII August, 2020.
[11]. Tadesse, G. A., Zhu, T., Liu, Y., Zhou, Y., Chen, J., Tian, M., Clifton, D., "Cardiovascular disease diagnosis using cross-domain transfer learning", 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 4262- 4265)., July, 2019.
[12]. Harvey, R. E., Coffman, K. E., Miller, V. M., "Women-specific factors to consider in risk, diagnosis and treatment of cardiovascular disease", Women's Health, 11(2), 239-257, 2015.
[13]. Gayathri, P., Jaisankar, N., "Comprehensive study of heart disease diagnosis using data mining and soft computing techniques", International Journal of Engineering and Technology 5(3):2947-2958, June, 2013.
[14]. Mathew, B., Francis, L., Kayalar, A., Cone, J., "Obesity: effects on cardiovascular disease and its diagnosis", The Journal of the American Board of Family Medicine, 21(6), 562-568, 2008.
[15]. Miller, V. M., Redfield, M. M., McConnell, J. P., "Use of BNP and CRP as biomarkers in assessing cardiovascular disease: diagnosis versus risk", Current Vascular Pharmacology, 5(1), 15-25, 2007.
[16]. Moses, D., "A survey of data mining algorithms used in cardiovascular disease diagnosis from multi-lead ECG data", Kuwait Journal of Science, 2015.
[17]. Emon, M. U., et al. & Kaiser, M. S., "Performance Analysis of Machine Learning Approaches in Stroke Prediction" 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), (pp. 1464-1469), 2020, November