Predicting Cardiovascular Disease with Machine Learning Algorithms: A Review

Maitri Bhavsar[1], Manish Patel[2]

Assistant professor, Computer Engineering, Sankalchand Patel University, Visnagar, India[1]
Professor, Information Technology, Sankalchand Patel University, Visnagar, India[2]

mhbhavsarce_spce@spu.ac.in, mmpatelit_spce@spu.ac.in

Abstract: Early detection of cardiovascular disease symptoms is one of the hardest things for professionals to do. Cardiovascular disease comes in many forms, including stroke, congenital heart disease (CHD), peripheral artery disease (PAD), and coronary artery disease (CAD). Comparing several feature selection methods to accurately predict cardiovascular disease is the main objective of this study. The renowned random forest, support vector classifier, k-nearest neighbors, Naive Bayes, and gradient boosting model have been taken into consideration in order to support the comparative accuracy and define the best predictive analytics. These algorithms use data analysis to forecast when heart failure will occur. This study processes the data to predict coronary illness. Finding more effective datasets, however, is essential to the effectiveness of the machine learning model. We have reviewed several machine learning algorithms that are currently in use, together with their benefits and drawbacks, in this work. We have also talked about a few outstanding research questions that will support future studies in this area.

Keywords: SVM, Random Forest, KNN, Naïve Bayes, Gradient Boosting

I. INTRODUCTION

It is impossible to overestimate the significance of the heart to the well-being of humans; nonetheless, issues arise when blood flow is insufficient, ultimately leading to cardiovascular failure. Genetics and changing lifestyles are two factors contributing to the global burden of diseases and rising healthcare costs [1]. Two of the most conventional signs and symptoms of the illness are discomfort in the left shoulder or elbow and pain in the chest.

However, if there are no symptoms, many diseases may lie undiagnosed for a very long time. The conventional approach to diagnosing this illness involves reviewing the patient's previous medical records, although this approach is rarely sufficient. An alternative would be to gather and examine the patient's data for potential future preventative actions. The World Health Organization lists heart disease as one of the leading causes of silent and not enforceable death worldwide. It is challenging to identify heart a medical condition early using clinical data. Early illness prediction is achieved by the application of machine learning techniques and technologies [2].

In this case, the study's application of machine learning techniques has proven essential. Because the cardiovascular system is so intricate, it must be managed to prevent potentially catastrophic situations. To identify cases of cardiac disease, numerous researchers have used methods for machine learning [4].

In an effort to identify patterns and forecast the onset of disease, they examined decision tree approaches utilizing information from the UCI repository's Cleveland cardiovascular disease database. The research's objective is to produce an application that uses the K-means algorithm for attribute grouping and neural network back propagation to forecast cardiac problems more accurately [5].

Many workable methods for employing machine learning to identify heart disease have been put out by researchers. As a result, research on creating medical applications with different machine learning techniques and algorithms has been released [9].

This is how the leftover of the sections is organized. In Section 2, the machine learning algorithms are covered. Section 3, lists the advantages and disadvantages of the various machine learning techniques now in use. Open research questions in this area are presented in Section 4, Our conclusion is drawn in Section 5.

II. MACHINE LEARNING ALGORITHMS

A. Support Vector Machine

SVMs are a kind as machine learning algorithm that is capable of both prediction and classification. The SVM supervised learning method is used to split the data into two categories. The margins between sorted data are output by an SVM as widely apart as feasible.

B. Random Forest

According to its name, "In order to increase the data set's forecast accuracy, Random Forest is a classifier that builds many decision trees on different subsets and averages them. Rather than depending just on a single decision tree, the random forest determines the outcome by computing the majority vote of each tree's forecasts and applying that information to the forecast.

C. Naïve Bayes

The Naïve Bayes Classifier is one of the easiest, fastest, and most efficient classification algorithms out there. The likelihood that a specific object will be encountered by the classifier is the basis for predictions.

D. K-Nearest Neighbor

It is a simple classification method that establishes the class of a new data point by comparing it to the classes of the k nearest neighbors in the training set. In our case, KNN was able to forecast the likelihood of heart illness based on whether the features of a patient and those of known cases were similar enough.

E. Gradient Boosting

It's an ensemble approach that builds up a number of weak learners over time by having each learner concentrate on the shortcomings of the one before it. It incorporates their conjectures to arrive at a trustworthy judgment. When used to forecast cardiac disease, gradient boosting, which iteratively improves predictions based on previous iterations, may be able to pick up on minute trends in the data.

III. RELATED WORK

A. (CRISP-DM) methodology [3]

Heart and blood vessel problems are known as cardiovascular diseases (CVDs), and they are a leading global cause of disability and early mortality. Early detection of individuals at increased risk of CVD is necessary to avoid premature deaths. Computational intelligence advancements and the massive volume of data generated every day have enabled the development of recognition systems that can recognize important information and hidden patterns in healthcare contexts. The aim of this work is to determine if a patient has a cardiovascular disease (CVD) by using data mining techniques (DMTs) to clinical information gleaned from the diagnosis.

Merits: In summary, this study demonstrated that it is feasible to employ DMMs to predict the presence or absence of CVDs using actual data from Electronic Health Records (EHR). It was feasible to obtain scores over 70% for some of the created DMMs, which is deemed satisfactory.

Demerits: Prior to putting into practice, a decision-support system based on one of the models developed during this study in a medical setting, for example, by using stronger algorithms, such as neural networks, to improve the model's performance.

B. Feature selection (filter, wrapper, embedded) [6]

The main objective of this study is to compare several feature selection algorithms in order to properly predict cardiovascular disease. In order to accomplish this, this study used a two-stage feature subset retrieving technique: first, we took into consideration Three popular algorithms for feature selection (filter, wrapper, and embedded); second, we extracted a feature subset obtained by utilizing these three techniques a common "True" condition based on a Boolean process. In order to establish the optimal predictive analytics and validate the comparative accuracy, several popular models have been examined, including ANN, XGBoost, Naive Bayes, k-nearest neighbors, random forests, and support vector classifiers.

Merits: Developing feature selection methods are an important yet difficult task for improving the performance of prediction models. This work, however, offered an improved scenario for cardiovascular disease feature selection by proposing a multi-stage learning algorithm based on feature selection for reproducing the dataset.

Demerits: This study has certain limitations that should be taken in to account. First off, there were only 13 features in the low-dimensional attribute used in this investigation. Further studies should incorporate other high-dimensional feature data sets to evaluate the suggested methodology.

C. Ensemble learning [7]

The World Health Organization lists heart disease as one of the leading causes of silent and non- communicable death worldwide. It is challenging to identify heart illness early using clinical data. In the past few years, machine learning has drawn more and more interest from researchers in the medical sciences. Early disease prediction is achieved through the application of machine learning techniques and methodologies.

Merits: Using a novel approach in this study to identify heart illness early on using machine learning techniques. The results showed that the ML Ensemble model had the highest predictive accuracy for cardiovascular illness.

Demerits: There is still opportunity for improvement even if diverse clinical data and several machine learning techniques have produced state-of-the-art results.

D. Machine Learning Classification [8]

Daily blockages of heart caused by plaque reduce the blood supply to the heart. This phase of coronary artery disease is known as the early stage. Plaque has substance-filled walls made of cholesterol. Gradually, less oxygen is needed for heart function. The wall and valve will contract during this period, giving the reference image an odd aspect. With a better degree of accuracy, their suggested approach is utilized to anticipate coronary artery disease early.

Merits: The most important characteristic of the suggested conception has been employed and evaluated for the prediction; additionally, the noise-free dataset makes the filter essential.

Demerits: The SVM algorithm will be tested on a large database in order to improve efficiency. The suggested method needs to be expanded to include a Graphical User Interface (GUI) for the cardiovascular sector's medical center.

E. Machine Learning Algorithms [11]

After obtaining the data set, pre-processing steps such as label encoding, normalization, and outlier identification were completed. The processed data was classified using techniques such as SVM, K- Nearest Neighbor, Decision Tree, Random Forest, Naive Bayes, Gradient Boosting, and Logistic Regression.

Merits: The investigation found that several methods performed differently and that each had unique benefits and drawbacks for the diagnosis of cardiovascular disease. The adoption of clinical settings could be facilitated and usability enhanced by a web application.

Demerits: Appropriate classification algorithms combined with suitable data-cleaning techniques can lead to more accurate analysis and prediction.

F. Extreme Learning Machine technique [12]

A primary application of machine learning systems is the identification of heart disease, which strikes millions of individuals. Patients with heart disease have many traits in common that can be utilized to accurately diagnose the condition, a host of variables, such as blood pressure, cholesterol, age, gender, and many others.

Merits: The Extreme Learning Machine (ELM) algorithm is suggested as the representation for these variables. For people who may potentially with risk of heart disease, the proposed system offers a warning system in place of costly medical exams.

Demerits: The Cleveland Clinic Foundation has collected information on about 300 patients so that the method could be tested. With this method heart problems could be anticipated with about 80% accuracy.

G. Ensemble learning [13]

This research suggests an ensemble-based method to estimate an individual's risk of cardiovascular disease using machine learning (ML) and deep learning (DL) models. In order to forecast cardiovascular disease, there are six classification methods. A dataset of cardiovascular disease cases that is made available to the public is used to train models. To identify significant features related to cardiovascular illness, they employ random forest (RF).

Merits: The study suggested using deep learning and ensemble-based machine learning techniques to forecast cardiovascular illness. The Machine Learning Ensemble model proved to be the most precise forecaster cardiovascular disease, based on the results of the study.

Demerits: May use a variety of techniques to identify the ideal qualities for our upcoming projects. To improve and increase the accuracy of the assessment, new datasets could be used. Ultimately, the prediction problem for more effective cardiovascular disease detection may be addressed using deep learning and reinforcement learning techniques.

H. Heartbeat Monitoring and Arrhythmia Detection Based on ECG Systems [10]

Predicting heart illnesses is aided by the latest advancements in data mining and machine learning- based medical support systems. They used a Cleveland Clinic dataset for this study, which includes nine attributes: age, gender, serum blood sugar during fasting, cholesterol in mg/dl, resting blood pressure (in mm Hg), and maximum heart rate reached during the ECG, chest pain during exercise, and type of chest pain. 25% of the dataset is employed in experiments, while the remaining 75% is accustomed to train the model using the classifier.

Merits: The aim is to create a classifier that can distinguish between patients with heart disease and healthy individuals. A model that functions well when data is entered in real time.

Demerits: It requires a lot of cardiovascular equipment, which raises the cost and puts it beyond the means of the average person.

IV. OPEN RESEARCH QUESTION

Appropriate classification algorithms combined with suitable data-cleaning techniques can lead to more accurate analysis and prediction [11], Rather than studying various methods for data cleaning and pruning to get a suitable dataset, the majority of this work has concentrated on using classification models to predict heart disease. But it's important to remember that, in terms of accuracy, a well- trimmed and cleaned data collection usually performs better than a disorganized data set with missing values. [13] The main aim of this work is to evaluate the effectiveness of suggested deep learning and machine learning classifiers for cardiovascular illness identification. The dataset for cardiovascular disease is utilized in the experiments conducted for this study. This study makes use of both deep learning and machine learning. There are two methods used to conduct these experiments. The first step toward improving outcomes is the application of machine learning models. Ensemble model, which was ML-based, did well on the provided dataset. The goal is to obtain more trustworthy outcomes.

V. CONCLUSION

The rise in mortality from cardiovascular disease has raised the demand for a quick and accurate prediction system. The purpose of this research was to determine the most effective machine-learning method for spotting cardiovascular problems. The KNN, Naive Bayes, Decision Tree, Gradient Boosting, Random Forest, Support Vector Machine, and Logistic Regression methods were evaluated using data set. An investigation found that several methods performed differently and that each had unique benefits and drawbacks for the diagnosis of cardiovascular disease. Subsequent studies could look into larger datasets, more machine learning methods, and the integration of Explainable AI and Federated Learning. An adoption of clinical settings could be facilitated and usability enhanced by a web application.

REFERENCES

[1] Organization, W.H. (2023) Cardio vascular diseases, World Health Organization. Available at: https://www.who.int/healthtopics/cardiovascular-diseases (Accessed: 23 May 2023).
[2] M. A. Alim, S. Habib, Y. Farooq and A. Rafay 2020 Robust Heart Disease Prediction: A Novel Approach based on Significant Feature and Ensemble learning Model 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), pp 1–5
[3] Barbara Martins, Diana Ferreira, Cristiana Neto, Antonio Abelha,Jose Machado , “Data Mining for Cardiovascular Disease Prediction,” © Springer, 5 January 2021 .
[4] S. Patel, J. Patel and S. Tejalupadhyay,"Cardiovascular Disease prediction using Machine learning and Data Mining Technique", Journal - IJCSC, vol. 7, 2022. Available: 10.090592/IJCSC.2016.018.
[5] Malav, A., Kadam, K. and Kamat, P. prediction of heart disease using K-means and artificial neural network Available at: https://www.researchgate.net/publication/319486202_Prediction_Of_Heart_Disease_Using_KMeans_and_Artificial_Neural_Network_as_Hybrid_Approach_to_Improve_Accuracy (Accessed: 22 May 2023).
[6] Najmul Hasan and Yukun Bao, “Comparing different feature selection algorithms for cardiovascular disease prediction,” © Springer, October 2020.
[7] M. A. Alim, S. Habib, Y. Farooq and A. Rafay 2020 Robust Heart Disease Prediction: A Novel Approach based on Significant Feature and Ensemble learning Model 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), pp 1–5
[8] Dr. Joy Iong Zong Chen, P. Hengjinda, “Early Prediction of Coronary Artery Disease (CAD) by Machine Learning Method-A Comparative Study” Journal of Artificial Intelligence and Capsule Networks (2021) Vol.03/ No.01 http://irojournals.com/aicn/DOI: https://doi.org/10.36548/jaicn.2021.1.002
[9] R. Kumar, P. Kumar, R. Tripathi, G. P. Gupta, A. K. M. N. Islam, and M. Shorfuzzaman, “Permissioned blockchainanddeep-learningforsecureandefficientdatasharinginindustrialhealthcaresystems,” IEEE Transactions on Industrial Informatics, p. 1, 2022.
[10] Divya Annepu, Gowtham G, “Cardio vascular Disease Prediction Using Machine Learning Techniques” International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072
[11] Kazi Mostaq hridoy, Saiful Islam Akash, Fahmida Afrose Dipti, Mahamudul Hasan, Jesan Ahammed ovi, Md Al-Imran, Dr. Sarwar Jahan " Heart Disease Algorithm using Machine Learning Algorithms" 2023 4th International Conference on Big Data Analytics and Practices (IBDAP) | 979-8-3503-0019- 2/23/$31.00 ©2023 IEEE | DOI: 10.1109/IBDAP58581.2023.10271997
[12] S. Ismaeel, A. Miri and D. Chourishi 2015 Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis 2015 IEEE Canada International Humanitarian Technology Conference (IHTC2015), pp 1–3
[13] Abdullah Alqahtani, Shtwai Alsubai, Mohammed Sha, Lucia Vilcekova, and Talha Javed, “Cardiovascular Disease Detection using Ensemble Learning,” Hindawi, August 2022.
[14] Dangare, C. and Apte, S. (2012) A data mining approach for prediction of heart disease using Neural Networks, SSRN. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstractid=2175569 (Accessed: 22 May 2023).
[15] Welcome to the UC Irvine Machine Learning Repository (no date) UCI Machine Learning Repository. Available at: https://archive.ics.uci.edu/ (Accessed: 19 June 2023).
[16] https://www.kaggle.com/sulianova/cardiovascular-disease-dataset
[17] C. J. Harrison and C. J. S. Gibbons, “Machine learning in medicine: a practical introduction to natural language processing,” BMC Medical Research Methodology, vol. 21, no. 1, p. 158, 2021.
[18] R. Kumar, P. Kumar, R. Tripathi, G. P. Gupta, A. K. M. N. Islam, and M. Shorfuzzaman, “Permissioned blockchainanddeep-learningforsecureandefficientdatasharinginindustrialhealthcaresystems,” IEEE Transactions on Industrial Informatics, p. 1, 2022.
[19] Karl Patrik Kresoja, Matthias Unterhuber, Rolf Wachter,Holger Thiele,Philipp Lurz, “A cardiologist’s guide to machine learning in cardiovascular disease prognosis prediction” Basic Research in Cardiology (2023) 118:10 https://doi.org/10.1007/s00395-023-00982-7 © Springer, 20 March 2023.
[20] Sibo Prasad Patro, Neelamadhab Padhy, Dukuru Chiranjevi, “Ambient assisted living predictivemodel for cardiovascular disease prediction using supervised learning” Evolutionary Intelligence https://doi.org/10.1007/s12065-020-00484-8, © Springer-Verlag GmbH Germany, part of Springer Nature 2020.
[21] Saravanan Srinivasan, Subathra Gunasekaran, SandeepKumar Mathivanan, Benjula Anbu Malar M. B, Prabhu Jayagopal& Gemmachis Teshite Dalu, “An active learning machine technique-based prediction of cardiovascular heart disease from UCI repository database” Scientific Reports | (2023) 13:13588 | https://doi.org/10.1038/s41598-023-40717-1.
[22] Ahmad Shaker Abdalrada, Jemal Abawajy, Tahsien Al Quraishi, Sheikh Mohammed Shariful Islam, “Machine learning models for

prediction of co-occurrence of diabetes and cardiovascular diseases: a retrospective cohort study" Journal of Diabetes & Metabolic Disorders (2022) 21:251–261, https://doi.org/10.1007/s40200-021-00968-z, © Springer, 12 January 2022.

[23] Niloy Biswas, Md Mamun Ali, Md Abdur Rahaman, Minhajul Islam, Md. Rajib Mia, Sami Azam, Kawsar Ahmed, Francis M. Bui, Fahad Ahmed Al-Zahrani, and Mohammad Ali Moni, "Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques" Hindawi BioMed Research International Volume 2023, Article ID 6864343, https://doi.org/10.1155/2023/6864343.

[24] Madhukar Chavan, Shashank Kumar Singh, Saurabh Bansod and Prashant Pal, "Design and Implementation of Heart Disease Prediction Using Artificial Neural Network" 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM) | 979-8-3503-4779- 1/23/$31.00 ©2023 IEEE| DOI: 10.1109/ICONSTEM56934.2023.10142