Grenze International Journal of Engineering and Technology, June Issue

Prediction and Diagnosis of Cardio Vascular Diseases using Machine Learning

Ravipati Naga Deepthi[1], Uddanti Sri Sai Ramya[2], Ramana Venkata Rami Reddy[3], Nakka Soma Sekhar[4] and VBKL Aruna[5]
[1-5]VRSEC(JNTUK)/ECE/Kanuru, Vijayawada, India
Email: 198w1a04n4@vrsec.ac.in, ramyauddanti1701@gmail.com , ramanaramireddy123@gmail.com, somasekhar94410@gmail.com, kameswari.aruna@gmail.com

Abstract*—Cardiovascular disease is one of the most hazardous conditions impacting people nowadays and significantly changes current lifestyle.One of the most lethal human diseases, cardiovascular disease has a detrimental impact on processing.Further health evaluations, treatments, and innovative characteristics are still required from the patient to the distant hospital. A precise and speedy verification of cardiovascular events is crucial for the early diagnosis and treatment of heart failure. It's frequently believed that the ritualistic medical history is spectacularly ineffective at diagnosing cardiovascular diseases. With non-invasive methods like machine learning, it may be possible to distinguish between healthy people and people with heart disease with precision and efficiency.There are numerous machine learning strategies for CVD prediction.The most complete and effective strategy, however, is determined by the data set being used.One of the most deadly human ailments is cardiovascular disease, which has a terrible impact on processing methods. The data is first pre-processed using data from before coronary artery disease.Then, we can anticipate CVDs early on by applying machine learning methods.Naive Baye's,SVM,Random Forest,Gradient Booster and Logistic Regression are the most often used algorithms. Large data sets are suited for Random forest. The publicly accessible Kaggle database, which was integrated with Google Collaboratory and a PYTHON programme, can be used as the input database.We shall categorise a person's heart ailment into phases after classifying heart diseases. Based on the likelihood of developing a heart block, there are four categories. Arterial blockage is the stage 1 and stage 2 condition, while chronic heart disease is the stage 3 and stage 4 condition.

***Index Terms*— CVDs , Machine Learning, Google collaboratory, Classification, Data preprocessing.**

I. INTRODUCTION

Cardiovascular disease (CVD), a chronic and complicated disease resulting from heart and vascular conditions, is currently the main cause of chronic disability and early mortality on a global scale.The usual methods of treatment are medical and surgical ones. Yet, these treatments are unable to reverse CVD. The quality of life for those with CVD is also considerably improved by these medicines.As a result, the primary focus of contemporary CVD care is prevention measures.According to recent studies, early intervention could save nearly 80% of CVD-related premature deaths. Moreover CVD has a slow onset and a long incubation period; as a

Grenze ID: 01.GIJET.9.2.81

result, it is frequently diagnosed at a more advanced stage.Thus, early recognition of high-risk groups for CVD is essential for its management and prevention.
The use of Csk prediction models to identify high-risk populations that could benefit from early intervention to reduce their likelihood of developing CVD has been supported by an increasing number of CVD prevention and control guidelines. The majority of the current risk prediction models for CVD were built using conventional statistical methods.A model is deemed to be established if it meets the independence and linearity requirements. This has an effect on both the use of external verification and the precision of the prediction model because it is unable to adequately represent the intricate interaction between variables.
The machine learning (ML) algorithm, a well-known statistical method, can successfully handle non-linearity, variable redundancy, and inter-variable interactions. It is frequently utilised in the field of CVD prevention and control because it may be used to explore potential risk factors for CVD in order to improve its prediction performance.Notwithstanding its advantages, there are still concerns about its ability to predict CVD. According to related studies, ML algorithms perform more accurately when making predictions than traditional statistical methods.In contrast, study found that the predictive performance of logistic regression (LR) was comparable to that of machine learning techniques.
According to this study, there is a direct link between high CVD incidence and high prevalence of the disease's risk factors, such as metabolic syndrome, hypertension, and obesity. The majority of these models are based on a feature screening strategy, despite the fact that Chinese researchers have lately produced prediction models based on Cox regression and ML algorithms.
According to this study, there is a direct link between high CVD incidence and high prevalence of the disease's risk factors, such as metabolic syndrome, hypertension, and obesity. The majority of these models are based on a feature screening strategy, despite the fact that Chinese researchers have lately produced prediction models based on Cox regression and ML algorithms.

II. Related Works

Clinicians should focus on tracking changes in NT-proBNP, Ddimer, Killip, cTnI, and LDH in order to enhance the clinical results of ACS patients.This suggestion is supported by research on models for patients with acute coronary syndrome that use machine learning to predict in-hospital mortality. The methodologies employed in this paper are the study's design and the patient variables and definition, building a model to predict hospital mortality risk for ACS patients, statistical investigation.
Muktevi Sri Venkatesh proposed a method for anticipating cardiovascular illness using machine learning techniques including K-Nearest Neighbour (KNN), Support Vector Machines (SVM), Logistic Regression (LR), Naive Bayes (NB), and Random Forest (RF). Patients with chronic CVD infection and healthy individuals can be predicted using a few learning calculations employing KsssNN, SVM, LR, NB, and RF[1]. The Logistic Regression classifier demonstrated its prowess in forecasting with the greatest outcomes in terms of precision and quickest execution time, according to the reenactment's results.
For the purpose of predicting the risk of developing heart disease, Ibomoiye Domor Mienye and Yanxia Sun[8] proposed a method based on an enhanced ensemble learning approach that uses techniques such as logistic regression (LR), l k nearest neighbour (KNN), support vector machine (SVM), linear discriminant analysis (LDA), classification and regression tree (CART), gradient boosting (GB), and random forest (RF). The proposed technique performed best with precision of 96%, sensitivity of 91%, accuracy of 93%, and F1 score of 93%.
In January 2022, A. Reyana, Sandeep Kautish, and Rajkumar Gangappa Nadakinamani[5] developed an approach for analysing clinical data in order to forecast cardiovascular illness using machine learning methods. This approach makes use of M5P Tree, Random Tree, REP Tree, Linear Regression, J48, JRIP, and Naive Bayes algorithms. Two datasets—Statlog (heart) and Hungarian—are made available in this study for machine learning classification techniques such Naive Bayes, Linear Regression, Random Tree, REP Tree, J48, and JRIP. To select the best machine learning model, the performance of the proposed CDPS was evaluated using a number of variables. When it comes to the prediction of patients with cardiovascular illness, the Random Tree model performed amazingly well, with the best RMSE of 0.0231, accuracy of 100%, the lowest MAE of 0.0011, and the quickest prediction time of 0.01 (secs).
In December 2021[12], M. D. Amzad Hossen, Tahia Tazin, and Sumiaya Khan published a method for supervised machine learning-based cardiovascular disease analysis and prediction that makes use of methods like Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT). The goal of this research was to determine which machine learning classifier would be most effective in predicting heart disease given the dataset

provided. Three distinct classifiers' results were compared. The logistic regression technique performed better than the other two classifiers, with an accuracy of 92%.

A method on Early and Accurate Prediction of Heart Disease Using Machine Learning Model was proposed in 2021[3] by B Padmaja, Chintala Srinidhi, Kotha Sindhu, and Kalali Vanaja. It uses methods like Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting, and Naive Bayes. A variety of data mining techniques are used for classification.The random forest data mining classification algorithm was used with a minimum set of 14 required characteristics, and it generated good results.

Apurb Rajdhan, Milan Sai, and Avi Agarwal[7] proposed a method for applying machine learning to detect heart illness in April 2020. Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), and Naive Bayes are employed in this method (NB).The goal of the study was to identify the most efficient ML algorithm for heart disease diagnosis.

The accuracy score of various machine learning algorithms for heart disease prediction is evaluated in this work using data from the UCI machine learning repository[10].The Random Forest algorithm, with an accuracy score of 90.16%, is the best algorithm for foretelling heart disease, according to the study's findings.Weicheng Sun and Ping Zhang proposed a method for the prediction of cardiovascular disorders using machine learning in May 2021, which uses methods such as Support Vector Machine (SVM), Logical Regression (LR), and Random Forest (RF).The results show that the SVM-based method can provide an average area under the ROC curve of 78.84%, proving that SVM outperforms LR & RF for our data.

III. METHODS

To finish our work, we used a heart disease dataset that we acquired from Kaggle.As seen in table 1, this dataset contains 917 instances of 9 different variables, including age, sex, cp, trestbps, chol, fbs, restecg, thalach, and exang.The dataset is initially cleaned up and processed using pre-processing techniques.The total number of patient records showed was 917.The data scientist can have a better knowledge of the dataset's validity with the use of data visualisation tools.Five machine learning classifiers are exposed to the dataset: Logistic Regression (LR, SVM, GB, RF), Gradient Booster, Random Forest, Naive Bayes, and SVM (GB).The most accurate classifier can be used to determine which one is the best. The trends connected to diseases are separated by datasets.From the data, a training dataset and a testing dataset are produced.920 records in total, together with 76 medical-related features, were obtained. The list of 9 qualities that the system is based on is displayed in the table below.

TABLE I. DATASET ATTRIBUTES

S.No	Attributes	Valuetype
1.	Age	Numerical
2.	Sex	Numerical
3.	Cp	Numerical
4.	Trestbps	Numerical
5.	Cho	Numerical
6.	Fbs	Numerical
7.	Restecg	Numerical
8.	Thalach	Numerical
9.	Exang	Numerical

IV. DESIGN PROCESS

A. Data Collection

- The quality and amount of our data have an impact on the precision of our model.
- Typically, the result of this stage is a data representation—Guo condenses it as a table—that we will utilise for training.
- Previously gathered data can still be used at this level by using datasets like Kaggle, UCI, etc.

B. Data Preparation

- Acquire data and make it training-ready. reasonable train/evaluation split? Depending on the field, the accessibility of the data, the specifics of the dataset, etc., 80/20, 70/30, or anything comparable may be appropriate.
- Clean anything that may need to be cleaned (remove duplicates, correct errors, deal with missing values, normalisation data type conversions, etc.)
- The impacts of the specific order in which we gathered and/or otherwise prepared our data are eliminated by randomising the data.
- Use data visualisation to uncover pertinent associations between variables or gaps between classes (bias alert!), or carry out more exploratory investigations.
- Group the sets according to training and assessment purposes.

C. Choose a Model

Chosen the most effective technique for the job at hand. Many algorithms exist.

D. Train the Model

- To respond or forecast accurately as often as is practical is the goal of training.
- Linear regression example: The computer software would need to learn the values of m(orW) and b. (x is input,y is output)
- Each step of the procedure involves training.

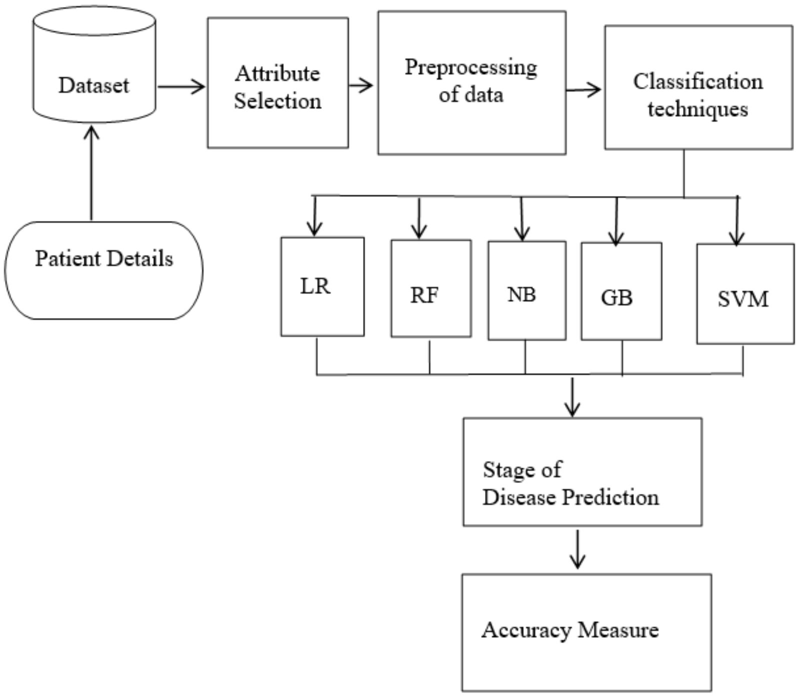

Fig: Flowchart Model

Algorithms used:

Random Forest
Random Forest techniques are used in both regression and classification.Predictions are made based on a tree-like organisation of the data.When used on large data sets, the Random Forest approach may still produce the same findings even when a significant portion of record values are missing.

Logistic Regression
Typically, the logistic regression approach is used for binary classification problems. The logistic regression technique uses the logistic function to restrict the output of a linear equation to the range between 0 and 1, as opposed to fitting a straight line or hyperplane.

Naive Bayes

The Nave Bayes algorithm is a supervised learning method for classification issues that is based on the Bayes theorem. It is mostly employed in text categorization with a large training set.Being a probabilistic classifier, it makes predictions based on the likelihood that an object will occur.

Support Vector Machine (Svm):

The vector machine, which can be used to address both classification and regression problems, is one of the most well-liked supervised learning techniques. According to the SVM approach, n- dimensional space can be classified by drawing the best line or decision border.

Gradient Booster (GB)

Gradient booster is a method that stands out for its prediction speed and accuracy, especially with large and complex datasets.Bias errors and variance errors are the two types of errors. Using the gradient boost method, we may lessen the bias error of the model.

V. RESULTS

Information on 917 patients was included in the dataset that was used for this inquiry.By using algorithms like Logistic Regression, Gradient Booster, SVM, Naive Bayes, and Random Forest, we can determine whether a person has CVD or not.The evaluation metric used in this inquiry was accuracy.

The classification of heart disease is done using machine learning algorithms

- ➢ The one who has no heart disease,the model will predict it as '0',which implies the person has no chance of getting heart disease.
- ➢ The one who has heart disease,the model will divide the stages of disease into 4 types based on chance of getting heart blockage.
- ➢ Based on the severity of arterial blockage, heart disease can be divided into 4 phases (stages 1 through 4).
- ➢ Artery blockage>50% indicates presence of heart disease. The stage of cardiac disease increases with the size of the obstruction.
- ➢ Heart attacks can occur at any time in people with stage 3 and stage 4 chronic heart disease.
- ➢ The person who are suffering from cvd can be classified as stage 1,2,3 and 4 namely coronary heart disease,stroke heart disease, peripheral arterial heart disease, and aortic disease respectively.

Accuracy

It is defined as the ratio of accurate predictions to all predictions made by the classification algorithm.

$Accuracy = TP + TN/(TP + TN + FP + FN)$.

- Collected dataset from kaggle.

TABLE II. ACCURACY VALUES

Algorithms	Accuracy
Logistic Regression	0.804878
SVM	0.707317
Random Forest	0.985366
Naïve Bayes	0.780488
Gradient Booster	0.964321

- Logistic Regression, SVM, Naive Bayes, Gradient Booster, and Random Forest are the algorithms.
- Nave Bayes, SVM, Gradient Booster, and Logistic Regression all have accuracy values of 0.780488, 0.804878, 0.707317, and 0.964321 respectively. The Random Forest algorithm, which has an accuracy of 0.985366, is the most accurate of these five algorithms.

REFERENCES

[1] Ke J., et al. machine learning-based approaches for predicting in-hospital mortality for patients with acute coronary syndrome. 2022 Mar 1;53:127-34 the Emergency Medicine American Journal.

[2] Stevens BR and CJ Pepine emerging importance of machine learning in research and application of cardiovascular disease. Cardiology Research and Practice, American Heart Journal Plus, 2021 November 1, 11:100050.Early

and accurate cardiac disease prediction using machine learning model.
[3] Padmaja, B., et al. TURCOMAT 12.6 (2021): 4516-4528 Turkish Journal of Computer and Mathematics Education.
[4] Fernandez-Cabo, E. Lara-Pezzi, X. Rossello, V. Fuster, F. Benito, J. P. Manzano, J. C. Silla, & (2020). The definition of cardiovascular risk in young, asymptomatic people is improved by machine learning. Journal of the American College of Cardiology, 76(14), 1674-1685.
[5] Zhou, Y., et al. (2019). cardiovascular event prediction for percutaneous coronary intervention using machine learning. 73(9S1), 127–127. Journal of the American College of Cardiology.C Tangiographydata. *International Journal of Cardiology*, *335*, 130-136.
[6] "Detection of Cardiovascular Disease Using Machine Learning Classification Models," Hana H. Allawi and S. Alsuwat Manal. The International Journal of Engineering Research & Technology's ISSN (2021) is 2278-0181.
[7] Manasa Gummalla, Yiye Zhang, Kelly Michalak, Evan Sholle, Zhuoran Xu, Ashley N. Beecy, and others. Predicting heart failure all-cause mortality or 30-day unplanned readmission using machine learning and electronic health data 1, no. 2 (2020): 71–79 in the Journal of Cardiovascular Digital Health.
[8] Wang, Z., Sun, Y., and Mienye, I. D. (2020). An improved ensemble learning approach for predicting the risk of heart disease. Unlocked Medical Informatics, 20, 100402.
[9] W. Benjamins, M. W. Yeung, T. Maaniitty, A. Saraste, R. Klén, P. van der Harst, et al (2021). utilising machine learning to enhance clinical and coronary data integration to better identify patients for advanced cardiac imaging data froma CT angiogram. 335, 130–136, International Journal of Cardiology.
[10] Kim, Y., et al. cardiovascular risk assessment based on deep learning and employing retinalimage-based coronary artery calcium scores. e306-e316 in The Lancet Digital Health, 3(5).
[11] R.G. Nadakinamani et al., 2022. Analysis of Clinical Data for Prediction of Cardiovascular Disease Using Machine Learning. Cognitive neuroscience and machine learning, 2022.
[12] "Supervised analysis and prediction of cardiovascular illness using machine learning." M. D. Hossen et al. Issues in Engineering Mathematics 2021 (2021)
[13] 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.
[14] Shah, Devansh, Samir Patel, and Santosh Kumar Bharti. "Heart disease prediction using machine learning techniques." *SN Computer Science* 1 (2020): 1-6.