Grenze International Journal of Engineering and Technology, Jan Issue

Technical Review on Prediction and Diagnosis of Cardiovascular Diseases using Machine Learning

R.NagaDeepthi[1], U.SriSaiRamya[2], R.VenkataRamiReddy[3], N.SomaSekhar[4], K.Gnandeep[5] and VBKL Aruna[6]
[1-6]Department of Electronics and Communication Engineering,Velagapudi Ramakrishna Siddhartha Engineering College, Kanuru, Vijayawada, AndhraPradesh
Email: {deepthiravipati7, ramyaauddanti1701, ramanaramireddy123, somasekhar94410, gnandeepkoneru1262, kameswari.aruna}@gmail.com

***Abstract*—The most serious sickness is cardiovascular disease that people nowadays has a severe impact on their daily lives. Therefore, additional medical diagnosis and treatment are needed, as well as information transformation from the patient to the distant hospital. An accurate and prompt identification of heart disease is essential.Cardiovascular disease diagnosis by standard medical history has been deemed unreliable in several ways. Non-invasive techniques, including machine learning, are effective and trustworthy for classifying healthy and cardiac disease patients. Nevertheless, there are numerous machine learning approaches for CVD prediction. However, the data set used determines which method is the most accurate and effective. First, data pre-processing techniques are used to prepare the data. Then, we may predict CVDs at an early stage by applying machine learning methods. LR, SVM, Random Forest, Gradient Booster, and Decision Tree are the most often used algorithms. Large datasets are appropriate for Random Forest.**

***Index Terms— Terms*—Cardio VascularDiseases, Machine Learning, Data Pre-Processing.**

I. INTRODUCTION

The most well-known dangerous infection in the world is cardio vascular disease(CVD), which kills more people each year than any other illness combined. 17.9 million people world wide died from cardio vascular disease(CVD) in 2016, accounting for 31% of all fatalities. It is difficult to diagnose heart disease because of diabetes, high blood pressure, excessive cholesterol, irregular heartbeat, and many other problems. Several data mining techniques have been used to assess the severity of heart illness in humans. Several methods, such as the Kth Nearest Neighbor (KNN).and decision trees, are used to categorise the disease's severity(DT). Since cardiac illness has a complex character, it requires cautious management. If you don't, it could harm your heart or lead to an early death. For the purpose of identifying different types of metabolic syndromes, the perspectives of medical science and data mining are utilised. In order to accurately forecast cardiac disease and analyse data, datamining with categorization is crucial. For the purpose of predicting cardiac disease, a number of techniques have been utilized to abstract knowledge from known data mining techniques. In the review article, it is discussed that a lot of reading has been done to create a prediction model that combines not just one or two separate methodologies, but also several. Utilizing heart rate time series, however, introduce neural networks. To ascertain the patient's precise condition in relation to heart disease, clinical records should be reviewed for Left Bundle Branch Block, Right Bundle Branch Block, Atrial Fibrillation , Normal Sinus Rhythm, Sinus

Grenze ID: 01.GIJET.9.1.515

Bradycardia , Atrial Flutter , Premature Ventricular Contraction , and Second-degree Block (BII). The radial basis function network (RBFN) uses 70% of the dataset for training and the remaining 30% for classification. This essay also provides an introduction to the subject ofresearch and medicine's use ofcomputer-aided decision support systems (CADSS). Datamining techniques have been used in the healthcare sectorand have been found toproduce faster and more accurate illness predictions. TheGA is suggested for heart disease diagnostics. This method makes use of effective association rules deduced from the GA for tournament selection, crossover, and mutation, which results in the new proposed fitness function. We do experimental validation utilising the well-known Cleveland dataset and data from a UCI machine learning repository. Later on, we'll evaluate how our results compare to some of the more popular supervised learning techniques. The most effective evolutionary method, particle swarm optimization (PSO), is described, and some rules are generated for heart disease. With the use of encoding techniques, the rules have been applied at random, improving accuracy as a whole. Pulse rate, sex, age, anda number of other symptoms are used to predict heart disease. As we've seen, the ML method using neural networks produces results that are more precise and dependable.

II. LITERATURE SURVEY

S.No	Year & Title	Author	Methodology	Journal	Remarks
1.	**31December2021** Machine learning-based in-hospitalmortality prediction models for patientswith acute coronarysyndrome.	JunKe, MD, YiweiChen, MPH, Xiaoping Wang, BS, Zhiyong Wu, MSc, QiongyaoZhang, BS, Yangpeng Lian, MSc, Feng Chen, BS	Study design, factors, and patient information Statistical analysis is used to build a prediction model of cszin-hospital mortality risk for ACS patients.	Journal of Big Data.	Based on the research, we advise doctors to pay close attention to NT pro BNP, D-dimer, Killip, cTnI, and LDH alterations since doing so may enhance the clinical results of ACS patients.
2.	**1September2021** Emerging role of machine learning in cardio vascular disease investigation and translations.	R.Cuocolo,T.Perillo, E.DeRosa, L.Ugga, M.Petretta	Cardio vascular disease, Clinical diagnosis, Machine learning Prediction models, Bio ethical considerations	Cardiolog y Researchand Practice	These were speedup during theCOVID-19 epidemic, and theensuing CVD translational implications of ML are guidingnovel measures of preventionand therapy, new methods forobjective clinical diagnosis, andeven prospects for reconsideringthe fundamental Principles of CVD nosology.
3.	**2021** Early and Accurate Prediction of Heart Disease Using Machine LearningModel.	B Padmaja a, Chintala Srinidhi, Kotha Sindhu, Kalali Vanaja, N M Deepika, E KrishnaRao Patrof.	Random Forest(RF), Logistic Regression(LR), K-Nearest Neighbour (KNN), Support Vector Machine(SVM), Decision Tree(DT), Gradient Boosting and NaiveBayes.	TurkishJour nalof Computerand MathematicsE ducation.	The classification processuses a variety of dataminingapproaches,andthe random forest data miningclassification algorithmproduced excellent results.The smallest amount ofneeded characteristics, 14required attributes,are employed.
4.	**2020** Machine Learning Improves Cardio vascular Risk Definition for Young, Asymptomatic Individuals.	Fátima Sánchez-Cabo, PHD, Xavier Rossello, MD,PHD, ValentínFuster, MD,PHD, FernandoBenito, MSC, Jose Pedro Manzano	Study Population and assessment of the outcome, potential predictors and data preparation, Comparision of ML methods, Cardio Vascular risk scores, external validation using a WHS cohort	Journal Of The American College ofCardiology	For the prediction of generalised pre clinical atherosclerosis, EN-PESA (Progression ofEarly Sub clinical Atherosclerosis) produced a c-statistic of 0.88. Overall, 86.8% of those at greater risk based on EN-PESA had baseline SA symptoms or a substantial advancement of SA during a 3-year period.

5.	March16,2019 Machine Learning-Based Cardio Vascular Event Prediction Fo r Percutaneous Coronary Intervention.	Yijiang Zhou, Ruoyu Zhu, Xiaojun Chen, Xiaolei Xu,Qiwen Wang, Liujun Jiang, Jianhua Zhu,Jian Wu, Hui Yan,Li Zhang	Conventional cox regression and four ML algorithms, including decision trees algorithms such as extreme gradient boosting and light gradient boosting machine, support vector machine, and neural network algorithms	Journal of Big Data.	The four machine learning techniques that employed the top 20 predictors considerably improved mo del discrimination(AUC/F1 Score: extreme gradient boosting 0.70/0.84; light gradient boosting machine 0.73/0.86; neural network 0.68/0.80; support vector machine 0.62/0.84).(AUC/F1Score: 0.74/0.89) The final ensemble model shown additional improvement in prediction ability.
6.	**July-2021.** Detection of Cardiovascular Disease using Machine Learning Classification Models.	Hana H. Allawi, Manal S. Alsuwat.	Decision Tree (DT), Logistic Regression(LR), Supp ort Vector Machine (SVM), Artificial Neur al Network(ANN), Nai ve Bayes(NB), KNN, Random Forest (RF), Voting Classifier(VC), Gradient Boosting(GB).	International Journal of Engineering Research & Technology(IJERT).	At the moment, one of the leading causes of death is heart disease. Various machine learning algorithms were utilized in this study to identify heart illness utilizing datasets related to CVD and heart disorders. On the cardiovascular illness dataset with all characteristics, the gradient boosting classifier performed better, but the top classifier on the heart disease dataset was the random forest.
7.	**2020** Utilizing electronic health data and machine learning for the prediction of 30-day unplanned read mission or all-cause mortality in Heart failure.	AshleyN. Beecy, MD, Manasa Gummalla, BA,Evan Sholle, MS, Zhuoran Xu, MSc, YiyeZhang	Index admission model,Index discharge model, Feature-aggregated model	Journal of Big Data.	For the prediction of a 30- day unplanned read mission or ACM, the index admission model, index discharge model, and feature-aggregated model all had AUCs of 50.723, 5 0.754, and 50.756 respectively.
8.	17 July2020 An improved ensemble learning approach for the prediction of heart disease risk.	Ibomoiye Domor Mienye,Yanxia Sun, Professor, Zenghui Wang, Professor	Methods include k-nearest neighbor (KNN), logistic regression(LR), linear discriminant analysis(LDA), support vectormachine(SVM), classification and regression tree(CART), gradient boosting(GB),and randomforest(RF).	Informatics in Medicine Unlocked.	With accuracy of 93%, precision of 96%, sensitivity of 91%, and F1 score of 93%, the suggested technique performed best.
9.	2 April 2021 Improving patient identification for advanced cardiac imaging through machine learning-integration of clinical and coronary CT angiography data.	Jan Walter Benjamins, Ming Wai Yeung, Teemu Maaniitty, Antti Saraste, Riku Klen, Pimvander Harst, Juhani Knuuti, Luis Eduardo Juarez- Orozco	Patient population, Clinical variables, CTA and PET acquisition, Image interpretation, Machine learning, Statistical analysis	International Journal of Cardiology	ML data integration from various sources regularly enhances diagnostic performance; non etheless, it did not surpass professional CTAinter pretation.

10.	May2021 Deep-learning-based cardio vascular risk stratification using coronary artery calcium scores predicted from retinal photographs.	Tyler Hyungtaek Rim, Chan Joo Lee, Yih-Chung Tham, Ning Cheung, Marco Yu, Geunyoung Lee, Youngnam Kim, Daniel S W Ting, Crystal Chun Yuen Chong	Study design and population, Deep-learning Algorithm for Predicting the presence of CAC, New cardio vascular disease risk stratification system, Statistical analysis	Journal of Big Data.	Deep learning techniques and CAC scores obtained from retinal photographs out perform existing risks tratification methods for cardio vascular disease events interms of accuracy inpredicting cardio vascular events.
11.	January2022 Clinical Data Analysis for Prediction of Cardio vascular Disease Using Machine Learning Techniques.	Rajkumar Gangappa Nadakinamani, A. Reyana, Sandeep Kautish, A.S. Vibith, Yogita Gupta, SayedF. Abdelwahab, and Ali Wagdy Mohamed.	REPTree,M5P Tree,Random Tree,LinearRegression, NaiveBayes,J48,and JRIP.	Hindawi. Computational Intelligence and Neuroscience.	Two datasets—Statlog(heart) and hungarian—are included in this work to be used with machine learning classification algorithms such REP Tree, Random Tree, Linear Regression,M5PTree,Naive Bayes, J48, andJRIP. To find the most appropriate machine learning model, the performance of the proposed CDPS was assessed using a variety ofcriteria. With the greatest accuracy of 100%, the lowest MAE of 0.0011, the lowest RMSE of 0.0231,and the shortest prediction time of 0.01 when it cameto patients with cardiovascular disease, the Random Tree model did Remarkably well(secs).
12.	December2021. Supervised MachineLearning-BasedCardiovascular DiseaseAnalysisandPrediction	M.D. AmzadHossn, TahiaTazin, SumiayaKhan EvanAlam, HossainAhme dSojib, Mohammad Monirujjama nKhan, and Abdulmajeed Alsufyani	RandomForest(RF),LogisticRegression(LR),DecisionTree(DT).	Hindawi. Computational Intelligence And Neuroscience.	The article's objective was to identify which machine learning classifier,given the dataset used, would be the best at foretelling heart disease.The outputs of three different classifiers were compared. With an accuracy of 92%, the logistic regression method fared better than the other two classifiers used.

III. METHODS

We used a heart illness dataset that we downloaded from Kaggle to complete our task.As of in table 1 Age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, and goal are among the 1024 instances of 14 different variables in this dataset. Pre-processing methods are used to initially clean up and process the dataset.1024 patient records in total were displayed. With the aid of data visualisation tools, the data scientist can have a better understanding of the validity of the dataset. Gradient Booster, Random Forest, Naive , and Logistic Regression (LR, SVM, GB, RF) areamong the five machine learning classifiers that are exposed to the dataset (GB).It is possible to identify the best classifier as the one that achieves the maximum accuracy Datasets separate the disease- related trends.A training dataset and a testing dataset are created from the data. 920 records in total, together with 76 medical-related features, were obtained. The list of 14 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
10.	oldpeak	Numerical
11.	slope	Numerical
12.	ca	Numerical
13.	thal	Numerical
14.	target	Numerical

IV. DESIGN PROCESS

A. Data Collection

- The quantity and calibre of our data influence how accurate our model is.
- Typically, the outcome of this stage is a data representation that we will use for training, which Guo condenses by calling it a table.
- By utilising datasets like Kaggle, UCI, etc.,previously acquired data can still be utilised at this level.

B. Data Preparation

- Acquire data and prepare it for training. decent train/evaluation split? 80/20, 70/30, or anything comparable, depending on the field, the accessibility of the data, the particulars of the dataset, etc.
- Clean anything that might require it (remove duplicates, correct errors, deal with missing values, normalisation data type conversions, etc.)
- Randomizing the data removes the effects of the precise order in which we collected and/or otherwise prepared our data.
- Visualize data to find relevant gaps between classes or relationships between variables (bias alert!), or performadditional exploratory studies.
- Divide the sets into training and evaluation categories.

C. Choose a Model

- Select the best method for the task at hand.There are several algorithms.

D. Train the Model

- The objective of training is to provide an accurate response or prediction as frequently as feasible.
- Example of linear regression: The programme would have to learn the values of m(orW) and b.(x is input,y is output)
- Every stage of the process is a training one.

V. EXISTING SYSTEMS

The data set for the patient is taken into account, and data cleaning is completed. The process of identifying and fixing erroneous or corrupt records from a record set is known as data cleaning.Fitting this data with some machine learning models and finding out the most accurate one. Training and testing the data with the best model and thus predicting the disease. Additionally, itdetermines algorithm correctness and assesses algorithm

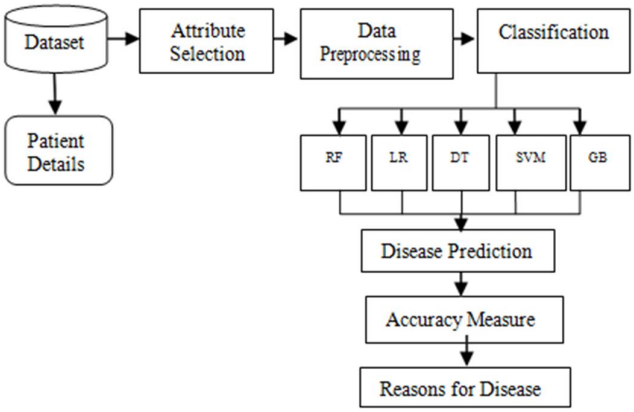

Fig 1:Flow Chart for the evaluation of model

accuracy across the board. The framework may predict thecoronary sickness using cutting-edge developments and othermachine learning calculations by using a lesser number of characteristics.Top resent an accurate model of predicting cardio vascular disease.To verify the lab test reports.Deficiency is The hospital's poor reputation is a result of misdiagnosis and incorrect treatment, which can even endanger the patient's life.The different algorithms explored in this paper are listed as below:

A. Random Forest

Regression and classification both employ Random Forest methods.The data is organised into a tree, and predictions are based on that tree. Even with a substantial number of record values missing,the Random Forest technique may still provide the same results when applied to huge data sets.

B. Logistic Regression

For binary classification issues,the logistic regression approach is typically utilised. Instead of fitting a straight line or hyperplane, the logistic regression technique uses the logistic function to limit the output of a linear equation to the range between 0 and 1.

C. Decision Tree

The decision tree method's core node displays the characteristics of the dataset, while the outside branches show the outcome. Decision trees are employed because they require minimum data preparation, are dependable,rapid, and easy to grasp.

D. Support Vector Machine(SVM)

Support one of the most well-liked algorithms for supervised learning is the vector machine,which may be used to solve both classification and regression issues. n- dimensional space maybe divided into classes by drawing the optimal line or decision boundary, according to the SVM method.

E. Gradient Booster(GB)

A approach that stands out for its prediction speed and accuracy,especially with big and complicated datasets, is gradient booster.Errors come in two flavours: bias error and variance error. We can reduce the model's bias error with the aid of the gradient boost approach.

VI. CONCLUSION

Heart disease is one of the conditions that causes death most frequently today. This study employed a range of machine learning techniques to identify cardiac diseases, including Support Vector Machine (SVM), Decision Tree Classifier (DT), Logistic Regression (LR), Random Forest (RF), and Gradient Boosting Classifier (GB). This review paper contributes to the comparative study and implementation of various machine learning techniques in Python software by offering a quick approach for the user to use the machine learning algorithms in Python software to forecast cardiovascular illnesses. This unethical study applies currently available machine learning techniques using Python's Jupiter notebook software.The identification of the processing of raw healthcare data of cardiac information would enable the long-term preservation of human life and the early

diagnosis of anomalies in heart diseases. Machine learning techniques were used in this workto analyse the raw data and provide a new and unique perspective on cardiac disease. The fatality rate can be considerably decreased, though, if the condition is caught early and preventative measures are put in place as soon as is practical. Further improvement of this study is highly desired in order to concentrate the research on real-world datasets rather than merely theoretical frameworks and simulations.

REFERENCES

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

[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, X. Rossello, V. Fuster, F. Benito, J. P. Manzano, J. C. Silla, & E. Lara-Pezzi (2020). Machine learning enhances the definition of cardiovascular risk in young, asymptomatic individuals. American College of Cardiology Journal, 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] Hana H. Allawi and S. Alsuwat Manal, "Detection of Cardiovascular Disease UsingMachine Learning Classification Models." The ISSN (2021) for the International Journal of Engineering Research & Technology is 2278-0181.

[7] Ashley N. Beecy, Manasa Gummalla, Evan Sholle, Zhuoran Xu, Yiye Zhang, Kelly Michalak, Kristina Dolan, et al. Using machine learning and electronic health data to predict 30- day unplanned readmission or all-cause mortality in heart failure Journal of cardiovascular digital health 1, no. 2 (2020): 71–79.

[8] Sun, Y., Mienye, I. D., and Wang, Z. (2020). An enhanced ensemble learning strategy for heart disease risk prediction. Medical InformaticsUnlocked, 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] Nadakinamani, R.G., et al. , 2022. Clinical Data Analysis for Machine Learning-Based Cardiovascular Disease Prediction.Neuroscience and computational intelligence, 2022.

[12] M. D. Hossen, et al. . "Supervised analysis and prediction of cardiovascular disease using machine learning." Engineering Mathematical Issues 2021 (2021)