Cardiovascular Disease Prediction Using Machine Learning Algorithms

Kalapraveen Bagadi
School of Electronics Enineering
VIT-AP University
Amaravati, India
ORCID: 0000-0003-1082-1972

Visalakshi Annepu
School of Comp. Science & Eng.
VIT-AP University
Amaravati, India
ORCID: 0000-0002-7199-1898

Adnan Naji Jameel AL-Tamimi
College of Technical Engineering
Al-Farahidi University
Baghdad, Iraq
adnanaji51@uoalfarahidi.edu.iq

Naga Raju Challa
Dept. of Elec. and Comm. Eng.
Bapatla Engineering College
Baptla, India
ORCID: 0000-0001-9909-3849

H.S. S. Aljibori
College of Engineering
University of Warith Al-Anbiyaa
Karbala, Iraq
Hakim.s@uowa.edu.iq

M. N. Mohammed
Mechanical Engineering Department,
College of Engineering,
Gulf University,
Sanad 26489, Bahrain
dr.mohammed.alshekhly@gulfuniversity.edu.bh

Oday I. Abdullah[1,2,3]
[1]*Dept of Energy Engineering,*
University of Baghdad, Iraq
[2]*Mechanical Engineering Dept, College of Engineering, Gulf University*, Sanad 26489, Bahrain
[3]Hamburg University of Technology, Institute of Laser and Systems Technologies, Hamburg, Germany
Oday.abdullah@tuhh.de

Rabab Alayham Abbas Helmi
School of Graduate Studies
Management and Science University
Shah Alam, Selangor, Malaysia
rabab_alayham@msu.edu.my

M. Alfiras
Electrical and Electronic Engineering Department, College of Engineering,
Gulf University
Sanad 26489, Kingdom of Bahrain
president.gu@gulfuniversity.edu.bh

***Abstract*— The fact that cardiovascular disease (CVD) is a major cause of death worldwide highlights the significance of accurate prediction for successful preventative and treatment measures. Machine learning algorithms, which use the analysis of vast patient data to reveal hidden patterns and risk variables, have recently come to light as potential techniques for CVD prediction. This study intends to analyse how machine learning (ML) techniques are used in CVD prediction and evaluate how well they perform in comparison to conventional approaches. The study makes use of a sizable patient cohort's medical history, demographic data, and clinical factors in a complete dataset. Predictive models are built using a variety of machine learning algorithms, including support vector machine (SVM), gradient boosting, K-nearest neighbours, naive Bayes classifier, and logistic regression. To find the most important factors influencing CVD risk, feature selection techniques are used. Metrics like accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve are used to assess how well the machine learning models work. The outcomes are contrasted with well-known risk prediction models and clinical guidelines in order to assess the added value of machine learning methods. This work intends to improve CVD prediction capabilities and offer useful insights for better risk assessment and management strategies by utilizing the power of machine learning.**

Keywords*— *Prediction, Health informatics, Machine learning Algorithms, Cardiovascular disease.

I. INTRODUCTION

An important worldwide health problem, cardiovascular disease (CVD) causes significant morbidity and mortality. It includes a number of ailments, including peripheral arterial disease, heart failure, and coronary artery disease. The burden on healthcare systems is reduced and patient outcomes are improved when CVD is promptly predicted and detected early. ML algorithms have become effective tools for forecasting the risk of CVD in recent years. These algorithms are computational models that, without explicit programming, can identify patterns in data and make forecasts or choices. Machine learning algorithms have the capacity to examine significant volumes of data, spot detailed patterns, and produce precise predictions for CVD risk assessment by leveraging massive datasets. This gives medical practitioners important knowledge they can use to support early intervention and preventive measures.

They are therefore suitable for creating CVD prediction models. Historically, statistical models based on age, gender, blood pressure, cholesterol levels, and smoking status have been used to predict the risk of CVD. While these risk factors offer useful information, ML systems can use a larger range of data to increase prediction accuracy, such as genetic data, medical imaging data, and lifestyle factors. The ability of ML algorithms to handle high-dimensional data and recognize nonlinear correlations between variables is one of their main advantages. They have the ability to spot tiny patterns and relationships that human experts could miss. Machine learning algorithms may generate tailored risk assessments by examining a wide range of risk indicators, enabling focused treatments and preventive measures.

Machine learning (ML) approaches are currently being employed more commonly in the engineering profession as effective solutions to a number of difficulties [1-6]. Healthcare-related ML algorithms have demonstrated encouraging outcomes in the prediction of cardiovascular illness. Numerous methods, including gradient boosting, support vector machines, decision trees, random forests, logistic regression, and neural networks, have been successfully used in this field. The best algorithm to use will

2023 IEEE 8th International Conference on Engineering Technologies and Applied Sciences (ICETAS) | 979-8-3503-2709-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICETAS59148.2023.10346353

rely on the particular dataset and research goals at hand. Each approach has distinct strengths and limitations. This research seeks to explore the possibilities and evaluate the performance of ML algorithms applied specifically for the prediction of cardiovascular illness. This work presents the advantages and disadvantages of employing these algorithms, the many kinds of data that may be used, and the performance metrics that are employed to gauge their predicted accuracy. In addition to that, this work also discuss about recent advancements that have been made as well as potential future paths in this area, such as the implementation of artificial intelligence and deep learning.

In general, the use of machine learning algorithms to the prediction of cardiovascular diseases has a great deal of potential for enhancing risk assessment, early identification, and individualized treatment techniques. This paper introduces methods that leverage these algorithms and tap into the varied data sources available, aiming for enhanced CVD management. Such advancements can lead to improved patient outcomes and foster healthier communities. Consequently, this study endeavours to devise a new CVD prediction approach utilizing proficient machine learning methodologies, including Logistic Regression, Random Forest, Naïve Bayes, Gradient Boosting, and SVM.

II. Related Work

CVDs continue to be a significant global burden, causing high rates of mortality and morbidity. Early detection and accurate prediction of CVDs are crucial for effective preventive strategies and personalized patient care. In recent years, ML algorithms have shown promising results in predicting CVDs based on various risk factors and clinical data. This literature survey aims to explore the recent studies and advancements in the field of CVD prediction using ML algorithms.

The systematic review presented in [7] explores various machine learning algorithms employed for CVD prediction. The authors analyze the performance and accuracy of different models, such as support vector machines, random forests, and neural networks, based on the reviewed literature. In the study presented in [8], the authors develop an artificial intelligence (AI)-enabled algorithm to predict cardiac contractile dysfunction using electrocardiogram (ECG) data. The algorithm employs a convolutional neural network and achieves high accuracy in detecting individuals at risk of developing heart failure. The research presented in [9] is an interpretable machine learning model for predicting heart disease. The authors employ a random forest algorithm and develop a scoring system that assigns weights to different risk factors. The model provides insights into the importance of each factor in predicting CVD. The article [10] introduces deep learning techniques in cardiovascular medicine. The authors discuss the application of deep learning algorithms, such as convolutional neural networks and recurrent neural networks, in predicting various cardiovascular conditions, including heart failure, atrial fibrillation, and coronary artery disease. While not focused on machine learning algorithms specifically, the report [11] from the American Heart Association provides valuable statistical information on heart disease and stroke. It serves as a comprehensive reference for researchers in the field of CVD prediction and management.

Logistic regression is a commonly employed ML algorithm for binary classification tasks. Ambrish conducted a study utilizing logistic regression to predict the risk of CVD using a dataset comprising clinical and laboratory parameters. The model achieved an impressive accuracy of 84% and identified age, cholesterol levels, and blood pressure as crucial factors in CVD prediction [12]. A support vector machine (SVM) is powerful ML algorithms known for their capability to handle high-dimensional data and complex decision boundaries. Shah et al. proposed an SVM-based approach for CVD prediction in a study that incorporated demographic, clinical, and laboratory data. The model attained an accuracy of 87.3% and outperformed other ML algorithms, including k-Nearest Neighbors and Naive Bayes [13]. An ensemble learning system called Random Forest mixes various decision trees to improve prediction accuracy. A Random Forest model was used in a study published by [14, 15] to predict the occurrence of CVD using clinical and demographic variables. The model's excellent 91.23% accuracy demonstrated the importance of characteristics like age, blood pressure, and cholesterol levels in the prediction of CVD. Artificial neural networks and other deep learning approaches have drawn a lot of interest in CVD prediction because of their capacity to automatically recognize complex patterns. A deep learning model was created in a study described in [15, 16] to predict cardiovascular risk using raw electrocardiogram (ECG) signals. The model outperformed conventional risk scores with an area under the receiver operating characteristic curve (AUC-ROC) of 0.85. Recurrent neural networks of the long short-term memory (LSTM) type are particularly good at detecting temporal dependencies in sequential input. In [16, 17], an LSTM-based model was proposed for CVD prediction utilizing electronic health records. The model attained an AUC-ROC of 0.90, highlighting the importance of long-term patient history in predicting CVD events. Deep learning (DL), a subset of machine learning (ML), holds significant promise in medicine, aiding in disease classification and complex decision-making, often powered by neural networks (NN). While DL excels in areas like automatic clinical diagnosis and treatment selection, it faces challenges like the 'black-box' criticism and demands for large training data. This review provides insights into DL's role for cardiac professionals, highlighting its potential, limitations, and future opportunities [18-22].

III. Methodology

A. Data Source

Patient data covering a wide range of medical disorders has been amassed in healthcare databases. The most common cause of death worldwide and one of these, heart illnesses present a variety of problems [23-27]. The phrase "cardiovascular disease" refers to a variety of ailments that affect the heart, blood arteries, and blood flow throughout the body. The Cleveland, Hungarian, Swiss, and Long Beach VA heart disease databases, all of which are accessible in the UCI Machine Learning Repository, provided the dataset for this study that contains medical features. These datasets are used to extract patterns linked to the diagnosis of heart attacks. An equal number of records from each set are used for the training dataset and the testing dataset. The collection consists of 303 records in total, each of which has 76 medical

attributes. 14 qualities are included in Table 1 and are the main subjects of the system's examination.

TABLE I. ATTRIBUTES COLLECTED

Attribute	Description
Patient's age	>35
Gender	Value 1: male; Value 0: Female
Chest pain type	Value 1: typical type 1 angina; Value 2: typical type 2 angina; Value 3: Non-angina pain; Value 4: Asymptomatic
Fasting blood sugar	Value 1:>120 mg/dl; Value 0: 120 mg.dl
Rest ecg – resting electrographic	Value 0: normal; Value 1: having st-t wave abnormality; Value 2: definite left ventricular hypertrophy
exang – exercise included angina	Value 1: yes; Value 0: no
Slope-the slope of the peak exercise ST segment	Value 1: unsloping; Value 2: flat: Value 3: down sloping
Ca – number of major vessels colored by fluoroscopy	Value 0-3
Thal	Value 3: normal: Value 6: fixed defect; Value 7: Reversible defect
Trest blood pressure	mm hg on admission to the hospital
Serum cholesterol	mg/dl
Thalach-maximum heart rate achieved	60-200
Old peak-ST depression included by exercise	0-6
Hear disease present	0: No; 1: Yes

B. *Analysis of Data*

It is one of the important phases in performing analysis as the data in the dataset contains most of the noisy and redundant data as nulls. Data preprocessing includes replacing missing values with the mean of their respective columns and employing techniques like data cleaning and integration. Addressing missing values and eliminating noisy data are essential steps, as they can otherwise result in inaccurate outcomes.

C. *Operating Environment*

Python is a high-level, interpreted, interactive, and object-oriented programming language that serves a variety of general-purpose applications. Python provides a variety of packages related statistical computing such as NumPy. It uses Pandas to create tables. These libraries can be readily employed for data analysis and visualization, facilitating the implementation of a robust prediction system for the specified task. Good plots can be obtained from matplotlib. Implementation of R in python is also available through the seaborn library, which offers a wide variety of gradient plots. Python is an open-source programming language that is compatible with various operating systems, including UNIX and Windows. When it comes to predicting outcomes, Python often yields superior results compared to other programming languages. Although heart disease can manifest in different forms, there are common underlying risk factors that determine an individual's susceptibility to developing the condition. These fundamental characteristics need to be assessed to determine the likelihood of experiencing heart disease.

IV. PROPOSED SYSTEM

To develop a system model for CVD prediction using machine learning algorithms, several key components and steps need to be considered. Figure 1 presents system model for CVD prediction. Here is a high-level overview of the system model:

Data collection: Compile extensive databases with patient information on demographics, medical history, lifestyle choices, and pertinent biomarkers related to CVD.

Data Pre-processing: To manage missing values, outliers, and inconsistencies, clean up and pre-process the acquired data. To guarantee that the data is in an appropriate format for machine learning algorithms, this entails tasks like data normalization, feature scaling, and encoding categorical variables.

Feature Selection: Apply feature selection techniques to identify the most informative and relevant features for CVD prediction. This helps reduce dimensionality and improves the performance of the predictive models.

Algorithm Selection: Choose appropriate machine learning algorithms that are suitable for CVD prediction. Commonly used algorithms include decision trees, random forests, support vector machines, logistic regression, and neural networks. The selection process takes into account factors such as the nature of the data, the complexity of the problem, and interpretability requirements.

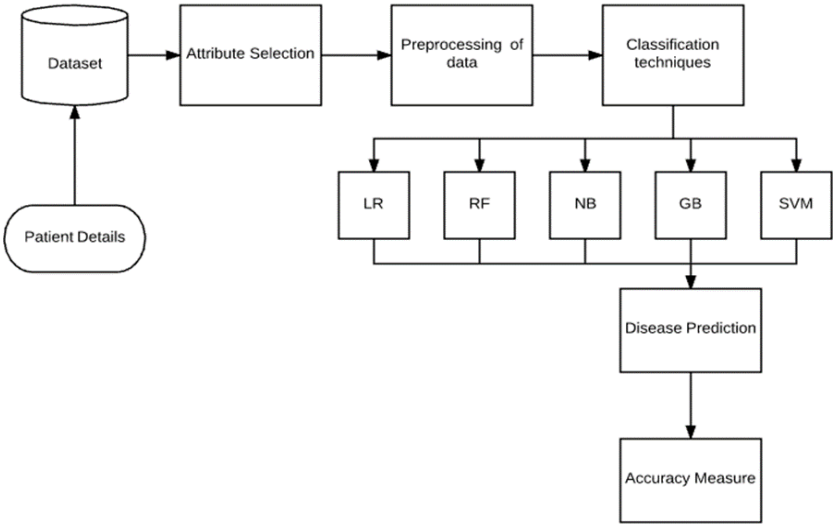

Fig.1. Proposed system model.

Model Training: Split the pre-processed dataset into training and testing sets for the model. Apply the training set to the job of CVD prediction training the chosen machine learning algorithm(s). In order to create precise predictions, the models discover patterns and relationships within the data.

Model Evaluation: Use the testing set to gauge the performance of the trained models. Accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (AUC-ROC) are common evaluation metrics for CVD prediction. These indicators shed light on how well the models are able to forecast cardiovascular illness.

Model Optimization: Fine-tune the hyper-parameters of the chosen algorithm(s) using techniques like cross-validation or grid search to improve the model's performance.

Deployment and Integration: Integrate the trained and optimized model into a practical system or application. This could involve creating an interface where healthcare providers or patients can input relevant information, and the system provides predictions or risk scores.

Monitoring and Maintenance: Continuously monitor the system's performance and update the model periodically as new data becomes available or as improvements in algorithms and techniques emerge.

Ethical Considerations: Ensure that the system addresses ethical considerations related to patient privacy, data security, and bias mitigation to maintain fairness and trustworthiness in the predictions.

The above steps provide a general framework for developing a system model for CVD prediction using machine learning algorithms. However, the specific implementation and choice of algorithms may vary depending on the available resources, expertise, and the specific requirements of the healthcare environment.

Figure 1 shows the functioning of the system:

(1) The UCI dataset provides the attributes and patient's data.
(2) Attribute selection is performed to identify the attributes that are relevant for estimating heart disease.
(3) The obtained data is further processed, including data classification and removal of noise, such as missing values.
(4) Various supervised classification algorithms are applied to the preprocessed data to determine the likelihood of developing CVD.
(5) The accuracy of the algorithms is estimated, and the values are compared among all the algorithms to evaluate their performance.

V. Machine Learning Techniques

A. Logistic Regression

A statistical model that is frequently used to address binary classification issues is logistic regression. It belongs to the class of supervised learning algorithms that attempts to forecast the likelihood of a binary outcome or the probability of an event based on one or more independent factors. In logistic regression, the dependent variable is binary, with just two possible outcomes commonly represented as 0 and 1. The predictor variables, often referred to as the independent variables, can be continuous or categorical. Finding the link between the independent variables and the likelihood of the binary result is the goal of logistic regression. The logistic function, often known as the sigmoid function, is used to do this. A linear combination of the independent variables is transformed by the logistic function into a probability value between 0 and 1. The logistic regression is expressed as:

$$P(Y=1)=\frac{1}{1+e^{-z}} \tag{1}$$

The above equation represents the logistic function used in logistic regression. In this equation, $P(Y = 1)$ represents the probability of the positive outcome, 'e' denotes the base of the natural logarithm, and 'z' represents the linear combination of the independent variables.

The maximum likelihood estimation (MLE) technique is used to determine the coefficients of the logistic regression model. The objective of this strategy is to identify the coefficients that increase the probability of witnessing the provided data. Optimization techniques like gradient descent, Newton's method, or the L-BFGS algorithm are frequently

used to train the logistic regression model. These algorithms iteratively update the coefficients based on the difference between predicted probabilities and actual outcomes, minimizing the logistic loss function. Once trained, the logistic regression model can predict the probability of a positive outcome for new instances. By setting a threshold value, predicted probabilities can be converted into binary predictions. For example, a threshold of 0.5 classifies probabilities above it as positive outcomes and those below as negative outcomes. Logistic regression offers several advantages. Understanding the effect of each independent variable on the likelihood of the result is made possible by the model's relative simplicity and interpretability. It is resilient to outliers and can handle both continuous and categorical variables. Logistic regression, on the other hand, presupposes a linear relationship between the independent variables and the outcome's log-odds. Logistic regression might not work well if the relationship is nonlinear. Furthermore, it presumes that independent variables are not multi-collinear and that observations are independent of one another. Numerous industries, including healthcare, finance, marketing, and social sciences, use logistic regression. It is very helpful for issues involving binary categorization, such as predicting customer turnover, recognizing diseases, or identifying spam emails.

B. Naïve Bayes

A popular machine learning technique known as naive Bayes bases its assumptions on the Bayes theorem and feature independence given the class label. It is frequently used for classification jobs, particularly when working with huge datasets. The eq. (2) is used by the algorithm to determine the likelihood of a class label based on the feature values:

$$P(C|F) = (P(C) * P(F|C)) / P(F) \quad (2)$$

where C represents class and F represents features. A well-liked machine learning technique called Naive Bayes employs the Bayes theorem and presumes feature independence in light of the class label. It is frequently employed for classification jobs, especially when working with sizable datasets. Using the above formula, the algorithm determines the likelihood of a class label based on the feature values. $P(F|C)$ is the conditional probability of the features given the class, $P(F|C)$ is the prior probability of the features given the class, and $P(F)$ is the probability of the features. $P(C|F)$ reflects the likelihood of a class given the feature values. Given the class label, the "nave" assumption in Nave Bayes is that all the features are conditionally independent of one another. By independently multiplying the probability of each feature, this assumption simplifies the computation of $P(F|C)$. The algorithm calculates the conditional probabilities of the features given each class and estimates the prior probability of each class using the training data in order to train a Naive Bayes model. Typically, smoothing techniques like Laplace smoothing or maximum likelihood estimation are used to estimate these probabilities. Using the aforementioned algorithm, Nave Bayes determines the likelihood of each class label given the feature values during the prediction phase. The projected class is given as the one with the highest likelihood. Naive Bayes is renowned for being straightforward, efficient in computing, and able to handle high-dimensional datasets. It excels at a variety of practical tasks like text classification, spam filtering, sentiment analysis, and document classification. The naive assumption of feature independence, however, may not hold in some situations, which can affect the algorithm's accuracy.

C. Random Forest

A flexible and effective machine learning method used for both classification and regression applications is the Random Forest algorithm. By using several different decision trees to combine into one forecast, it overcomes the shortcomings of individual decision trees, such as overfitting and instability. In a Random Forest, several decision trees are constructed and trained using various subsets of the training data. A random subset of features is taken into account for splitting at each node of the trees. As a result, the trees become more diverse and unpredictable, which helps to prevent overfitting. Each tree makes an individual prediction of the class or value of the input during prediction. In order to arrive at the final forecast, the predictions are combined and either voted on by a majority (for classification) or averaged (for regression).

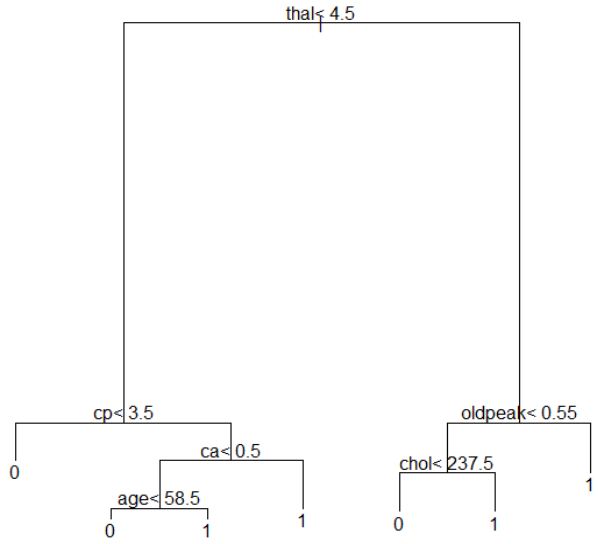

Fig. 3. Random forest tree.

Random Forest incorporates two key concepts: bagging and feature randomization. Bagging involves randomly selecting subsets of the training data with replacement to train each tree, creating diverse models and reducing variance. Feature randomization is achieved by considering only a random subset of features at each node, preventing any single feature from dominating the splits and encouraging different trees to focus on different subsets of features. Random Forest excels at handling large datasets with many dimensions, locating key features, and making reliable predictions. Compared to individual decision trees, it is less prone to overfitting and can handle missing data well. Numerous industries, including banking, healthcare, and image processing, frequently employ Random Forest. It provides decent accuracy, noise resistance, and interpretability via feature importance measures. The interpretability of Random Forest, however, declines as the number of trees in the forest rises, and it can be

computationally expensive for huge datasets. Figure 3 displays the several decision trees connected by links in the Random Forest.

D. Support Vector Machine

An efficient machine learning technique used for classification and regression applications is the Support Vector Machine (SVM). It seeks to locate an ideal hyperplane that efficiently divides data into different classes or forecasts desired values. SVM seeks to locate a hyperplane in binary classification that maximizes the margin, or the distance between the hyperplane and the closest data points from each class. The support vectors, which stand in for the data points closest to the border of the decision, are very important in choosing the best hyperplane. By using kernel functions, SVM can handle both linearly and non-linearly separable data. These kernel functions make it easier to map data into a higher-dimensional feature space where a linear separation can be found. SVM can capture complicated relationships and produce precise classification or regression predictions by altering the data. SVM's strength is in its capacity to work with a variety of data sources and deliver reliable results even with a small number of training samples. It is less prone to overfitting and can handle high-dimensional data successfully. SVM has been extensively used in several fields, including bioinformatics, text categorization, and image classification. When working with huge datasets, SVM's computational complexity can be a problem, thus choosing the right kernel function and fine-tuning the hyper-parameters call for careful thought in order to get the best results. The SVM hyperplane is defined mathematically as:

$$w \times x + b = 0 \tag{3}$$

Here, w represents the weights or coefficients associated with each feature, x denotes the feature vector, and b is the bias term. The objective of SVM is to find the optimal values of w and b that minimize the classification error while maximizing the margin. This can be formulated as an optimization problem. For linearly separable data, the problem can be written as:

Minimize: $1/2 * ||w||^2$

Subject to: $y_i * (w*x_i + b) >= 1$ for all data points (x_i, y_i)

where y_i is the class label of data point x_i.

SVM uses kernel functions such the polynomial kernel, Gaussian (RBF) kernel, or sigmoid kernel to handle non-linearly separable data. The data must be transformed into a higher-dimensional space in order for it to demonstrate linear separability, and these kernel functions are crucial in this process. The data can be successfully separated into multiple classes by finding a hyperplane by mapping the data into this higher-dimensional feature space. Even for complex and non-linear data patterns, this transformation enables SVM to capture complex relationships and produce precise predictions. The chosen kernel function implicitly defines the feature space in which the kernel optimization problem is solved. SVM determines the sign of the decision function ($w \times x + b$) during prediction in order to categorize brand-new examples into the appropriate classes. The instance is assigned to one class if the value is positive; the other class if the value is negative. SVM functions well in situations where the classes are separable or nearly separable and is effective when dealing with high-dimensional data. It performs well in generalization and is less prone to overfitting. However, SVM can be computationally taxing, especially when working with huge datasets, and it might call for rigorous hyper-parameter adjustment. Despite these factors, SVM continues to be widely used in a variety of fields, such as image classification, text categorization, bioinformatics, and finance, where it reliably produces precise results. Its adaptability and efficiency in solving a variety of problems have helped it become widely used and well-liked in the machine learning community.

E. Gradient Boosting

A powerful machine learning approach called gradient boosting combines several weak learners, often decision trees, to create a reliable predictive model. As an ensemble approach, it uses an iterative construction process in which each new model is created to address flaws in the prior models. Gradient boosting's main goal is to minimize a loss function's value during the course of an iterative procedure in order to optimize a loss function. An initial weak model is fitted to the data at the beginning of the method. The residuals, which indicate the differences between the actual values and the predictions provided by the prior models, are then used to train a new model in each iteration. As a result, the new model may concentrate particularly on identifying the patterns that the old one missed. The final prediction is created by combining the predictions from all the models, with the weighting of each model's contribution determined by how well it performed individually.

The optimization in gradient boosting involves minimizing the loss function with respect to the model's predictions. Different loss functions can be employed, such as mean squared error (MSE) for regression problems or log loss (or cross-entropy) for binary classification problems. By calculating the gradient (partial derivative) of the loss function with respect to the predictions, the algorithm adjusts the predictions in a manner that minimizes the loss. The learning rate is another crucial parameter in gradient boosting. It controls the impact of each individual model on the final prediction. A smaller learning rate results in a slower convergence but can enhance the model's generalization ability. On the other hand, a larger learning rate speeds up convergence but may lead to overfitting the training data. Gradient boosting finds applications in various machine learning tasks, including regression, classification, and ranking. It is renowned for its high predictive accuracy and robustness. Notably, there are popular implementations of gradient boosting algorithms, such as XGBoost, LightGBM, and CatBoost, which have achieved state-of-the-art performance in numerous machine learning competitions and real-world applications.

To summarize, gradient boosting is an ensemble algorithm that combines weak learners to create a powerful predictive model. Through iterative training and residual-based learning, it continuously improves the model's predictive capabilities. Gradient boosting is widely utilized

and has demonstrated exceptional performance in both competitive settings and practical scenarios.

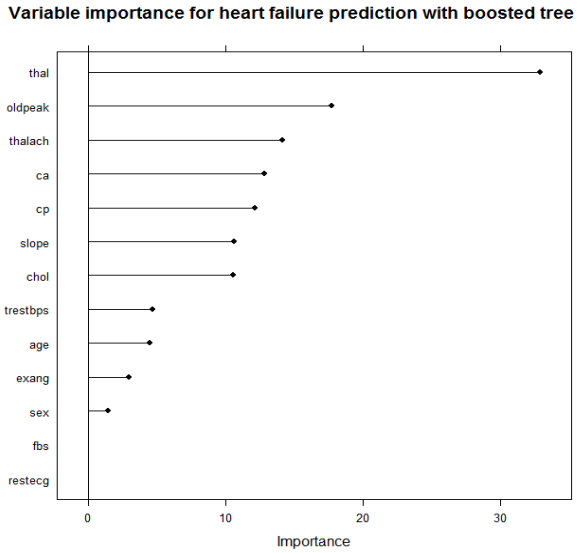

Fig. 4. Variable importance graph.

VI. Accuracy Module

The accuracy of CVD prognosis is predicted by the module using machine learning techniques. It establishes the highest likelihood of getting a CVD based on the algorithms' maximum accuracy. Depending on the factors taken into account, which are connected to the development of CVD, each algorithm produces a varied accuracy rate. As seen below, the module derives specificity and sensitivity from the confusion matrix:

$$\left.\begin{aligned}\text{True Negative Rate, specificity} &= \frac{P}{P+Q}\\ \text{False Positive Rate, } 1-\text{specificity} &= \frac{Q}{P+Q}\end{aligned}\right\}\ \text{sum to 1} \quad (4)$$

$$\left.\begin{aligned}\text{True Positive Rate, sensitivity} &= \frac{R}{R+S}\\ \text{False Negative Rate} &= \frac{S}{R+S}\end{aligned}\right\}\ \text{sum to 1} \quad (5)$$

where P is number of true negatives,
Q is number of false positives
R is number of true positives,
S is number of false negatives
These metrics are important for evaluating the performance of the prediction model and understanding its ability to accurately identify both healthy individuals and those at risk of CVD.

VII. Data Visualization

In this system, a powerful data visualization feature has been incorporated using the Seaborn library. Seaborn offers a wide range of visualization techniques, including boxplots, mosaic plots, and scatterplots, to effectively represent the relationships and characteristics of various attributes. These visualizations enhance the understanding of the data and provide valuable insights into the nature of the attribute relationships.

VIII. Accuracy Module

In this section, the results are presented and the accuracy outputs of the algorithms are displayed. A comparison is made between the algorithms based on their accuracy. When building predictive models in the domain of CVD prediction using ML algorithms, the training and testing data split ratio can vary based on a number of factors. In this work, the dataset sizes are 80 and 20 for training and testing respectively. The accuracy generated by each algorithm is shown in Table II as follows:

TABLE II. Comparison of Accuracies

Algorithm	Accuracy	Overall Accuracy
Logistic regression	0.9160	0.8652
Random forest	0.8952	0.8087
Naïve Bayes	0.9094	0.8416
Gradient boosting	0.9069	0.8416
SVM	0.8825	0.7972

The above results are obtained by using the in-built classifier present in the Sklearn – a machine learning library in python. By performing the train-test split and hyper parameter tuning, the model can be trained and tested on the given dataset. Additionally, this approach allows for obtaining the optimal parameter that yields higher efficiency. User interface can also be provided to obtain the values from the user to provide results directly to the interface used by the user.

IX. Conclusions

In the sphere of healthcare, the application of ML algorithms to the prediction of CVD has shown considerable promise. These algorithms can efficiently examine and find patterns that can be suggestive of CVD risks by utilizing enormous volumes of patient data. Researchers and healthcare professionals have been able to create precise models for predicting cardiovascular illness by implementing different ML approaches such decision trees, random forests, SVMs, and NNs. These models give a thorough evaluation of a person's risk by taking into account a wide range of variables, such as demographics, medical history, lifestyle decisions, and biomarkers. The quality of patient care and the results can be greatly impacted by the ability to forecast CVD accurately. Early identification of high-risk people enables prompt interventions, dietary changes, and appropriate medical care, perhaps delaying or reducing the onset of CVD. Additionally, ML has the ability to help healthcare professionals make better judgments, enhance patient management techniques, and maximize resource allocation. However, it's critical to recognize the difficulties and restrictions that come with employing ML algorithms to forecast CVD. Further study and research are needed in the areas of obtaining high-quality and diverse datasets, dealing with bias issues, guaranteeing algorithm transparency and interpretability, and integrating these algorithms into current healthcare systems. Overall, the use of ML algorithms to the prediction of CVD shows significant promise for increasing patient outcomes and healthcare delivery. Realizing the full potential of these algorithms and implementing them in clinical practice will

depend heavily on ongoing developments in this area and interdisciplinary partnerships between researchers, healthcare professionals, and data scientists.

REFERENCES

[1] V. Annepu, D. R. Sona, C. V. Ravikumar, K. Bagadi, M. Alibakhshikenari, A. A. Althuwayb, B. Alali, C. S. Virdee, G. Pau, I. Dayoub, C. H. See, and F. Falcone, "Review on Unmanned Aerial Vehicle Assisted Sensor Node Localization in Wireless Networks: Soft Computing Approaches," IEEE Access, vol. 10, pp. 132875-132894, December 2022.

[2] N. K. Vaegae, K. K. Pulluri, K. Bagadi, and O. O. Oyerinde, "Design of an Efficient Distracted Driver Detection System: Deep Learning Approaches," IEEE Access, vol. 10, pp. 116087-116097, November 2022.

[3] V. Annepu, A. Rajesh, and K. Bagadi, "Radial basis function-based node localization for unmanned aerial vehicle-assisted 5G wireless sensor networks", Neu. Comp. and Appl., vol. 33, pp. 12333-12346, March 2021.

[4] Kala Praveen Bagadi, Visalakshi Annepu, Susmita Das, "Recent trends in multiuser detection techniques for SDMA–OFDM communication system," Phy. Comm., vol. 20, pp. 93–108, September 2016.

[5] K. P. Bagadi, and Susmita Das, "Multiuser Detection in SDMA-OFDM Wireless Communication System Using Complex Multilayer Perceptron Neural Network," Wir. Pers. Comm., vol. 77, pp. 21 – 39, 2014.

[6] K. P. Bagadi, and Susmita Das, "Efficient complex radial basis function model for multiuser detection in a space division multiple access/multiple-input multiple-output-orthogonal frequency division multiplexing system", IET Comm., vol. 7, pp. 1394-1404, September 2013.

[7] M. Alshahrani, T. Alkhalifah, and M. Javaid, "Cardiovascular disease prediction using machine learning algorithms: A systematic review," J. of Health. Eng., vol. 6656361, May 2021.

[8] Z. I. Attia, S. Kapa, F. Lopez-Jimenez, et al. "Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram," Nat. Med., vol. 25, pp. 70-74, January 2019.

[9] B. Auffarth, F. Eichinger, and F. Scheipl, "An interpretable machine learning model for the prediction of heart disease," Eur. J. of Ope. Res., vol. 287, pp. 592-603, 2020.

[10] C. Krittanawong, H. Zhang, and Z. Wang, "Deep learning for cardiovascular medicine: A practical primer," Eur. Heart J., vol. 41, pp. 2555-2564, July 2020.

[11] D. Mozaffarian, and E. J. Benjamin, A. S. Go, "Heart disease and stroke statistics—2016 update," A report from the Amer. Heart Asso. Cir., 133, e38-e360, 2016.

[12] G. Ambrish, G. Bharathi, G. Anitha, S. Chetana, Dhanraj, and M. Kiran, "Logistic regression technique for prediction of cardiovascular disease," Glo. Tran. Pro., vol. 3, pp. 127-130, June 2022.

[13] S. M. S. Shah, F. A. Shah, S. A. Hussain, and S. Batool, "Support Vector Machines-based Heart Disease Diagnosis using Feature Subset, Wrapping Selection and Extraction Methods," Comp. & Elec. Eng., vol. 84, June 2020.

[14] S. Dhanka and S. Maini, "Random Forest for Heart Disease Detection: A Classification Approach," 2021 IEEE 2nd International Conference On Electrical Power and Energy Systems (ICEPES), Bhopal, India, 2021, pp. 1-3.

[15] V. Annepu, D. R. Sona, C. V. Ravikumar, K. Bagadi, M. Alibakhshikenari, A. A. Althuwayb, B. Alali, C. S. Virdee, G. Pau, I. Dayoub, C. H. See and F. Falcone, "Review on Unmanned Aerial Vehicle Assisted Sensor Node Localization in Wireless Networks: Soft Computing Approaches," *IEEE Access*, vol. 10, pp. 132875-132894, 2022, doi: 10.1109/ACCESS.2022.3230661.

[16] L. Brunese, F. Martinelli, F. Mercaldo, and A. Santone, "Deep learning for heart disease detection through cardiac sounds," Proc. Comp. Sci., vol. 176, pp. 2202-2211, 2020.

[17] K. Bagadi, et al., "Precoded Large Scale Multi-User-MIMO System Using Likelihood Ascent Search for Signal Detection", *Rad. Sci.*, vol. 57, no. 12, December 2022. https://doi:10.1029/2022RS007573

[18] M. G. Veerabaku, J. Nithiyanantham, S. Urooj, A. Q. Md, A. K. Sivaraman, and K. F. Tee, "Intelligent Bi-LSTM with Architecture Optimization for Heart Disease Prediction in WBAN through Optimal Channel Selection and Feature Selection," Biomedicines, vol. 11, pp. 1167, April 2023.

[19] K. Bagadi, "Detection of Signals in MC–CDMA Using a Novel Iterative Block Decision Feedback Equalizer," *IEEE Access*, vol. 10, pp. 105674-105684, October 2022, https://10.1109/ACCESS.2022.3211392.

[20] N. F. A. Alhadeethy, A. Zaki, and A. Shah "Deep learning model for predicting and detecting over lapping symptoms of Cardiovascular Disease in Hospitals of UAE," *Turk. J. of comp. and Math. Edu.*, vol. 12, no. 14, pp. 5212-5224. November 2021.

[21] K. Bagadi, T. Abrao and F. Benedetto, "A Novel Machine Learning Approach for Intelligent Spectrum Management in Cognitive Radio Networks," *IEEE Net. Let*., doi: 10.1109/LNET.2023.3300274.

[22] Mohammed, M.N., Dionova, B.W., Al-Zubaidi, S., Bahrain, S.H.K., Yusuf, E., An IoT-based smart environment for sustainable healthcare management systems, Healthcare Systems and Health Informatics: Using Internet of Things, 2022, pp. 51–74

[23] Mohammed, M. N., Desyansah, S. F., Al-Zubaidi, S., & Yusuf, E. (2020, February). An internet of things-based smart homes and healthcare monitoring and management system. In *Journal of physics: conference series* (Vol. 1450, No. 1, p. 012079). IOP Publishing.

[24] Mohammed, M. N., Alfiras, M., Al-Zubaidi, S., Al-Sanjary, O. I., Yusuf, E., & Abdulrazaq, M. (2022, May). 2019 Novel Coronavirus Disease (Covid-19): Toward a Novel Design for Smart Waste Management Robot. In 2022 IEEE 18th International Colloquium on Signal Processing & Applications (CSPA) (pp. 74-78). IEEE.

[25] Desyansah, S. F., et al. "Bradykinesia Detection System Using IoT Based Health Care System for Parkinson's Disease Patient." 2021 IEEE International Conference on Automatic Control & Intelligent Systems (I2CACIS). IEEE, 2021.

[26] Mohammed, M. N., Al-Rawi, O. Y. M., Arif, A. S., Al-Zubaidi, S., Yusof, K. H., Rusli, M. A., ... & Abdulrazaq, M. (2022, May). 2019 Novel Coronavirus Disease (Covid-19): Toward a Novel Design for Nasopharyngeal and Oropharyngeal Swabbing Robot. In 2022 IEEE 18th International Colloquium on Signal Processing & Applications (CSPA) (pp. 69-73). IEEE.

[27] Mohammed, M. N., Hazairin, N. A., Arif, A. S., Al-Zubaidi, S., Alkawaz, M. H., Sairah, A. K., ... & Yusuf, E. (2021, August). 2019 novel coronavirus disease (covid-19): Toward a novel design for disinfection robot to combat coronavirus (covid-19) using iot based technology. In 2021 IEEE 12th Control and System Graduate Research Colloquium (ICSGRC) (pp. 211-216). IEEE.