International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 38, No. 3, June, 2024, pp. 1035-1043
Journal homepage: http://iieta.org/journals/ria

Assessment of Cardiovascular Disease Using Machine Learning

Divya Adusumilli[1*], Sree Lakshmi Damineni[1], Swathi Kailasam[2], Nagamani Tenali[3], Ramu Yadavalli[4]

[1] Department of Computer Science and Engineering, PVP Siddhartha Institute of Technology, Vijayawada 520007, India
[2] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram 522302, India
[3] Department of Computer Science and Engineering, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru 521356, India
[4] Department of Computer Science and Engineering, Shri Vishnu College of Engineering for Women, Bhimavaram 534202, India

Corresponding Author Email: adivya@pvpsiddhartha.ac.in

https://doi.org/10.18280/ria.380329

Received: 16 December 2023
Revised: 1 February 2024
Accepted: 12 March 2024
Available online: 21 June 2024

Keywords:
cardiovascular disease, machine learning, data features, accuracy, data analytics, ensemble, anticipation, enhancing

ABSTRACT

Cardiovascular disease (CVD) is a prominent contributor to global mortality rates. The principal aim of this research is to employ machine learning techniques to anticipate the early actions needed to prevent the disease from progressing. The timely identification of individuals with a heightened risk of developing CVD plays an important role in implementing early interventions to impede disease progression. Machine learning techniques have shown promise in predicting CVD risk. For this paper, we propose a comprehensive CVD prediction model using ML techniques. Our approach utilizes a substantial dataset of electronic health records (EHRs) for training and validating our model. Through the incorporation of feature engineering, feature selection, and model optimization techniques, we have reached a high level of accuracy and interpretability. To evaluate the prediction of cardiovascular disease (CVD) threats, we compare the performance of various popular ML algorithms, such as logistic regression, random forest, and Support Vector Machine. Our findings point towards that our proposed model improve on existing approaches in regard of both accuracy and efficiency. This model can efficiently recognize individuals with an elevated risk of developing CVD, enabling early interventions to prevent the onset and progression of the disease. Additionally, we perform a acatalectic analysis of the features that contribute most to the assessment of CVD risk, providing insights into the underlying mechanisms of the disease. We also evaluate the robustness of our model by testing its performance on a separate dataset. Furthermore, we discuss the clinical implications of our proposed model, highlighting the potential benefits of using ML techniques in identifying individuals at high risk of developing CVD. Our model can aid in personalized medicine and facilitate the delivery of targeted interventions to high-risk individuals, thereby improving patient outcomes and reducing healthcare costs. When it comes to treating severe stages of cardiovascular disease, preventive treatments are typically more economical. Our approach can assist in lessening the financial burden related to CVD by lowering hospital stays, ER visits, and long-term care expenses via early detection of high-risk individuals and implementation of focused therapies. In summary, our model presents a robust solution for utilizing machine learning techniques to envisage the risk of cardiovascular disease (CVD). Our study targets to provide the expanding field of research regarding the application of machine learning in healthcare. The insights extended from our findings hold significant potential for enhancing the anticipation and treatment of CVD.

1. INTRODUCTION

Cardiovascular diseases (CVD) pose a major global public health concern in the larger perspective. They include a variety of disorders that impact the heart and blood arteries, such as peripheral artery disease, heart failure, coronary artery disease, and stroke. Millions of fatalities worldwide are attributed to cardiovascular disease (CVD) each year. High blood pressure, high cholesterol, diabetes, obesity, smoking, and physical inactivity are common risk factors for cardiovascular disease (CVD). The likelihood of having cardiovascular problems is raised by the frequent coexistence and interaction of these factors. Blood clotting, inflammation, and lipid metabolism are only a few of the variables that are influenced by genetic predisposition and impact the development of CVD. Personalized prevention and treatment approaches can be informed by knowledge of the genetic markers linked to cardiovascular disease risk. Globally, a significant percentage of deaths are attributable to CVDs. Around 17.9 million fatalities worldwide in 2019 were attributed to them, according to the World Health Organization (WHO), making up almost 32% of all deaths.

The heart is an essential organ liable for circulating blood throughout the body. Any disruption in its general functioning can have severe consequences for human health. In recent times, lifestyle choices, work-related stress, and unhealthy eating habits have contributed to arise in heart-related ailments. It is therefore crucial to develop feasible and accurate methods for assessing heart disease to prevent and manage its occurrence. Healthcare organizations worldwide collect vast amounts of health-related data, which can be analyzed to gain useful insights. However, the sheer size of these datasets and their complexity make them challenging to perceive and utilize effectively.

Determining if a person has heart disease and creating the finest hybrid model to help doctors forecast the danger of heart disease are the challenges at hand. The major goals are to minimize death by creating knowledge-magnified computer-aided methods for diagnosing heart disease and to establish a system model that can quickly find out the cardiac illness in order to help doctors make timely forecasts. To create a hybrid system, we combined a prediction model with appropriate classification methods and an adaptive voting classifier.

Since heart disease is still one of the top causes of morbidity and death in the world, reliable prediction models are essential for early identification and treatment. Unfortunately, there are a number of obstacles in the way of creating reliable predictive models that limit their therapeutic utility. Heart disease is a complex ailment that is impacted by multiple hereditary, lifestyle, and environmental variables. It takes complex algorithms and reliable data integration techniques to fully represent the intricacy of these relationships in predictive models.

Machine learning Techniques have emerged as valuable tools to analyze and make sense of massive datasets. These algorithms are capable of predicting the existence or not of heart disease with high accuracy, thereby enabling healthcare providers to target interference towards those most at risk. In this paper, we propose a ML based approach for predicting heart disease using data analytics. Our approach involves analyzing large datasets of electronic health records, demographic information, and clinical measurements to identify important risk factors related with heart disease. The results of our study have noteworthy implications for improving heart disease prevention and management strategies, leading to better health outcomes and reduced healthcare costs.

Here, we present a data analytics and machine learning approach for predicting CVD risk factors. Specifically, we investigate the use of various algorithms to analyze a huge dataset of electronic health records, demographic information, and clinical measurements. We aim to identify patterns and relationships among various risk factors, such as age, sex, blood pressure, cholesterol levels, and smoking status, and predict the feasibility of developing CVD.

Our study findings have significant implications for improving Cardio-vascular disease prevention and management strategies. By accurately identifying individuals at higher risk of developing Cardio-vascular disease, healthcare providers can target interventions and resources towards those who required them the most. Furthermore, using machine learning techniques can help optimize Cardio-vascular disease risk assessment and management, leading to better health outcomes and reduced healthcare costs.

2. RELATED WORK

Cardiovascular diseases (CVDs) have become a leads to cause of mortality throughout the world during the previous few decades posing a significant threat not only in one country but globally. Therefore, there is a persuasive need for a reliable, accurate, and feasible diagnostic system to identify and treat such diseases promptly [1]. ML algorithms and techniques have been widely employed to analyze large and intricate medical [2] datasets, automating the analysis process. Recent research has focused on utilizing various machine learning techniques to aid healthcare professionals in diagnosing heart-related diseases [3]. This paper presents a survey that examines different models based on such methods and algorithms while evaluating their performance [4]. Among the researched models, those based on supervised learning algorithms such as Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Naïve Bayes, Decision Trees (DT), Random Forest (RF), and ensemble models have gained significant popularity.

Voting is one of the modest ways to combine predictions from many ML algorithms. The adaptive voting classifier is not an unaffected classifier; rather, it is a covering for a number of dissimilar ones that are simultaneously trained and tested to take advantage of the individual features of each method. There are two categorical ways of vote: hard voting and soft voting. We decided to use the strict voting process.

Hard voting: The expected output class is the basis for voting. All classifiers in the basic model receive a set of training data. The models [5] predict the output class independently. The bulk of the models forecast this outcome class. In greater part voting, we anticipate the outcome using the manner of the classifiers' expected outputs. Let's say our classifiers provide (1,1,1,0) outputs, where 0 represents no heart disease and 1 means the mode of the set [6].

The use of information and communication technology, including gadgets like smartphones, smartwatches, smart glasses [7], and portable health monitoring devices, has caused Mobile Health (mHealth) to become a rapidly expanding field of study. Globally, coronary heart disease (CHD) is acknowledged as one of the primary reason of early death [8].

Vital signs and health metrics, including heart rate, blood pressure, and physical activity levels, can be continuously monitored thanks to mHealth technologies. Fitness trackers and smartwatches are examples of wearable technology that can gather data in real-time and provide a more complete picture of cardiovascular health than sporadic clinic visits.

mHealth-powered remote patient management solutions can help patients with CVD. With the use of these devices, medical professionals may keep an eye on their patients' health from a distance, modify treatment plans as necessary, and take quick action if something alarming happens. By being proactive, problems can be avoided and results can be enhanced.

Many obstacles must be overcome in order to effectively use the massive amounts of healthcare data, including data integration, data quality assurance, privacy issues, and the extraction of useful insights. By automating data processing, finding patterns, and making predictions, machine learning (ML) approaches are essential for addressing these obstacles.

A trustworthy cardiac monitoring system that can identify key cardiac patterns and intermittent anomalies [9] that can cause sudden death is thus becoming more and more necessary. Mobile devices have the capacity to gather seismic and electrocardiographic (ECG/ECG) data that can be effectively

[10, 11] evaluated to track a patient's cardiac activity and issue early alerts [12, 13]. This study combines the analysis of multichannel SCG data with a new approach for collecting cardiac data.

An early warning system is implemented to monitor a person's cardiac activities, and the accuracy of the system is assessed using only the ECG data [14]. The assessment demonstrates an 88% accuracy, indicating the viability and practicality [15, 16] of the proposed early warning system.

Heart disease is a universal health concern that poses a significant threat, as its symptoms may not be visible to the naked eye and can suddenly manifest when its limits are exceeded. Consequently, accurate and timely diagnosis is crucial. The healthcare industry generates vast amounts of data daily, encompassing patient and disease information [17, 18]. However, the efficient utilization of this data by researchers and practitioners remains limited. While the healthcare industry is data-rich, it lacks the necessary knowledge derived from this data.

Data mining and ML techniques and tools offer promising avenues to retrieve meaningful mastery from databases and leverage it for correctness diagnosis and decision-making [19, 20]. As research on predicting heart disease continues to grow, there is a need to summarize the existing but fragmented research in this area.

Cardiovascular disease (CVD) prediction's capabilities include its ability to recognize complex patterns, forecast with accuracy, assess risk individually, identify and intervene early, and integrate multi-omic data.

To capture minor variations in CVD risk variables and outcomes that may not be seen using traditional statistical methods, machine learning (ML) algorithms can recognize complex patterns and correlations within big and complicated datasets.

ML models are capable of achieving high levels of predictive accuracy in identifying people who are at risk of getting CVD when they are properly trained and validated. These models can produce accurate risk estimations that are customized to each patient's unique profile by utilizing a variety of data sources and sophisticated algorithms.

In the context of CVD, ML models have a number of drawbacks, including interpretability, data quality and bias, overfitting, data privacy and security, validation, and reproducibility. It can be difficult to decipher the underlying causes of the predictions made by complex machine learning models, like deep learning neural networks, because these models frequently lack interpretability. The therapeutic usefulness of ML models may be limited by this lack of transparency, especially in healthcare settings where decision-making depends heavily on interpretability.

The principal objective of the research paper is to provide a comprehensive summary of recent studies in heart disease prediction, including comparative results and analytical conclusions. The outcomes indicate that Naive Bayes with Genetic algorithm, Decision Trees, and Artificial Neural Networks techniques enhance the accuracy of heart disease prediction systems in various contexts.

This paper also presents a summary of generally used data mining and ML techniques, along with their complexities.

3. SUGGESTED METHODOLOGY

Pre-processing and feature selection procedures must be used to the dataset in order to create an accurate prediction model for cardiovascular disease. After they are finished, the prediction model can be constructed using a variety of categorization algorithms.

Within this study, we have employed feature selection methods on an updated dataset to determine the most significant variables in predicting cardiovascular disease. We have then used these variables to construct a model using different classification techniques.

To ensure the definiteness of the model, we have trained it with a large and diverse dataset, which allows the model to make accurate predictions of disease likelihood based on the input data. Our proposed model design is depicted in Figure 1, which focus attention on the primary components of the model: feature selection and classification techniques.

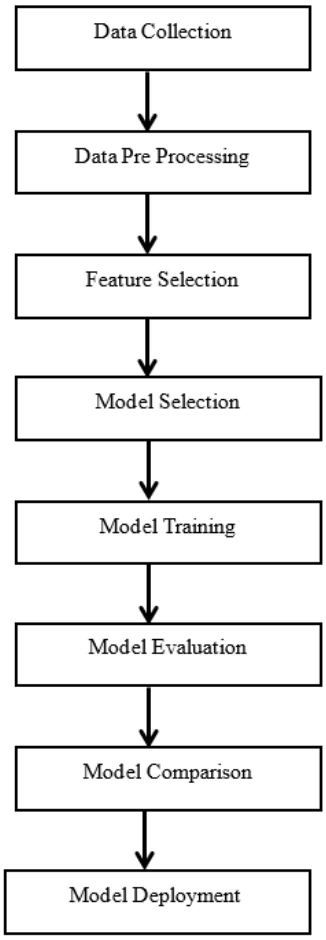


Figure 1. Flow diagram for a process

3.1 Modules

3.1.1 Data collection

The tasks of collecting data for the assessment of cardiovascular disease (CVD) involve careful planning and consideration of various data sources. Collect needed information from a variety of sources including clinical and non-clinical data from patients, such as age, sex, blood pressure, cholesterol levels, smoking status, family history, Genetic information, Life style factors etc. Respect ethical principles and protect the privacy of data.

The ethical implications of genetic data and lifestyle factors are critical while gathering data on cardiovascular disease (CVD). Participant access to genetic counseling is necessary if genetic testing is used, in order to help participants comprehend the possible hazards, psychological and social ramifications, and consequences of the test results. It is important to reassure participants that their genetic data will not be utilized in a discriminatory manner, such as preventing them from obtaining insurance or job possibilities. It is necessary to take precautions to ensure the privacy and confidentiality of the genetic and lifestyle data collected from participants. Encryption, restricted access to sensitive data, and safe data storage are required for this.

3.1.2 Data preprocessing

It is the pivotal step in preparing the data for machine learning models in the assessment of cardiovascular disease (CVD) and it is mainly used to prepare a clean, well-organized dataset that facilitates accurate model training and evaluation. Preprocess the collected data by handling missing data, outliers, and inconsistencies. Standardize and/or normalize numerical features. the features, classify variables using encoding and splitting the data into training and testing sets.

3.1.3 Feature selection

A crucial stage in the machine learning assessment of cardiovascular disease (CVD) is feature selection. Selecting the most pertinent features to add to the model entails making decisions that can boost interpretability, decrease overfitting, and improve performance. Select the most relevant features for cardiovascular disease prediction using techniques such as correlation analysis, feature importance analysis, or PCA. Conducting experiments with various techniques and evaluating their effects on model performance is frequently advantageous.

3.1.4 Model selection

Choosing a suitable machine learning model to evaluate cardiovascular disease (CVD) entails taking into account the unique needs of the problem as well as the features of the data. Select the relevant ML models for cardiovascular disease prediction, such as Support Vector Machine, Random Forest, and Logistic Regression. Both linear and non-linear classification can be successfully accomplished with SVMs. They manage complex relationships in the data and perform well in high-dimensional spaces. Random forest is an ensemble learning technique. It offers feature importance scores, manages non-linearity well, and is robust. Logistic Regression is a useful tool for binary classification issues it is also easily interpreted.

3.1.5 Model training

Training a model to assess pertinent characteristics and forecast a person's likelihood of experiencing cardiovascular problems Model development for cardiovascular disease (CVD) assessment Train the selected models on the trainset using the selected features.

3.1.6 Model evaluation

Model evaluation plays a necessary role in guaranteeing the dependability and efficiency of ML models in the context of cardiovascular applications. Assessment of performance of the trained models on the testing set using evaluation metrics such as accuracy, sensitivity, specificity, AUC, and F1-score.

3.1.7 Model comparison

To evaluate the efficacy of various models for the prediction of cardiovascular disease (CVD), a number of criteria must be taken into consideration. Assess the performance of the trained models and choose the most optimal model(s) by evaluating various metrics.

3.1.8 Model deployment

Because healthcare applications are so important, there are a few special considerations like clinical validation, integration with Electronic Health Records, Feedback loop with healthcare professional etc. that must be made while deploying a cardiovascular disease (CVD) prediction model. Deploy the best model(s) in a clinical setting for cardiovascular disease prediction.

Here is a simplified flowchart diagram that illustrates the above steps:

The models used in this paper are

- Support Vector Machine
- Random Forest
- Logistic Regression

3.2 Random Forest

Random Forest is a significantly used supervised ML algorithm that excels in both regression and classification tasks. Its primary principle revolves around leveraging numerous trees to converge towards optimal decisions, thereby enhancing accuracy.

In classification tasks, Random Forest employs a voting system to determine the class, while in regression tasks, it calculates the mean of all decision tree outputs.

Random Forest possesses a remarkable capability to handle vast datasets characterized by high dimensionality. This algorithm stands out as a robust machine learning technique, employing an ensemble of decision trees to yield precise predictions in regression and classification scenarios.

Its adeptness in tackling large and intricate datasets has established it as a favored choice among data scientists and researchers alike.

Random Forest is a popular ensemble learning technique utilized for predicting cardiovascular disease. It leverages the power of multiple decision trees by combining their predictions. Each decision tree is constructed using a random subset of features and training data. By aggregating the predictions from all the decision trees, the final prediction is made. This approach allows for the analysis of numerous clinical and non-clinical features, such as age, sex, blood pressure, cholesterol levels, smoking status, and family history, to identify significant risk factors.

Figure 2 Among the benefits of Random Forest is its capacity to handle missing data and noisy features effectively. Additionally, it provides estimates of prediction uncertainty. This makes it a useful tool for cardiovascular disease prediction, as it can process extensive datasets containing known outcomes (presence or absence of cardiovascular disease) and corresponding features, including demographic and clinical data.

To utilize Random Forest for cardiovascular disease prediction, the first step is to collect the dataset with known outcomes and associated features. Next, training and testing sets are created from this dataset.

The Random Forest algorithm is tested on the testing set to gauge its prediction ability after being learned on the training set. Various evaluation metrics, such as accuracy, sensitivity, specificity, area under the curve (AUC), and F1-score, can be employed to appraise the model's effectiveness in predicting outcomes.

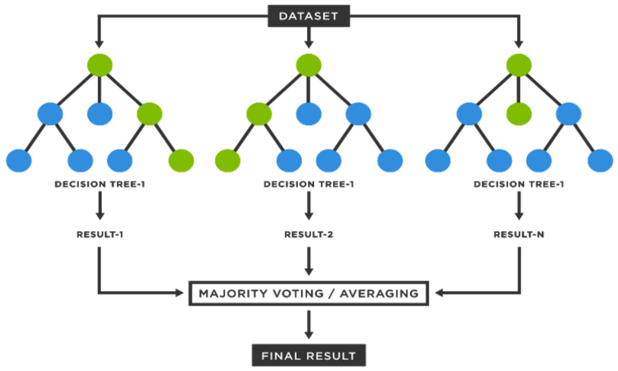


Figure 2. Random forest

3.3 Support Vector Machine

Support Vector Machine (SVM) shown in Figure 3 is a prevailing supervised machine learning technique that serves as both a classifier and a predictor for datasets with predefined target variables. It excels in classification tasks by discovering a hyperplane within the feature space, effectively separating different classes. By representing training data points as points in the feature space, SVM ensures that points belonging to distinct classes are distinctly divided by a generously wide margin. During the testing phase, the SVM algorithm projects the test data points into the same feature space and classifies them accordant with their positioning relative to the margin. By identifying an optimal hyperplane that effectively separates classes, SVM achieves highly precise outcomes, even in situations where the classes present inherent challenges in their differentiation.

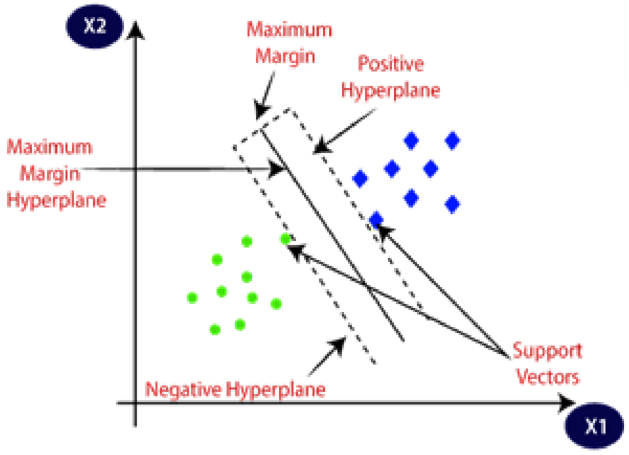


Figure 3. Support Vector Machine

Figure 3 in this scenario, we observe a two-dimensional feature space that exhibits a more intricate decision boundary, enabling greater flexibility in classification. The crucial support vectors are indicated by larger green dots, while the region between the two parallel dashed lines represents the margin. To clarify, the support vectors correspond to the points in closest proximity to the decision boundary and are denoted by larger blue dots. The objective of the SVM algorithm is to optimize the margin, maximizing its width, while ensuring accurate classification of the data points.

3.4 Logistic Regression

Logistic Regression is a mostly used supervised ML algorithm primarily employed for classification tasks. It follows a systematic approach involving dataset preparation, feature selection, model training, performance evaluation, hyperparameter tuning, and deployment for classifying new data. By minimizing errors and optimizing performance, the algorithm adapts its coefficients accordingly. Logistic Regression, a form of linear regression, predicts the probability of a binary outcome by establishing the connection between the dependent variable (the predicted outcome) and one or more independent variables (features).

The process for Logistic Regression commences with dataset preparation, encompassing the selection of pertinent features and splitting the data into training and testing sets. Feature selection holds paramount importance as it identifies the key features essential for accurate predictions. Subsequently, the Logistic Regression model is trained using the training set. Throughout the training process, the algorithm regulates its coefficients to minimize the disparity between the predicted and actual outcomes of the training data. Performance evaluation is then conducted using the testing dataset. If necessary, the model's hyperparameters are tuned to optimize its performance. Once trained and fine-tuned, the model becomes deployable, enabling it to classify new, unseen data points

Logistic regression is a highly utilized technique applied across various domains, such as medical diagnosis, fraud detection, and sentiment analysis. Its remarkable ability to estimate the probability of an event occurrence makes it a valuable tool for decision-making and classification tasks. By looking the relationship between a set of independent variables (e.g., demographic and clinical data) and a binary outcome variable (e.g., presence or absence of cardiovascular disease), logistic regression provides statistical insights.

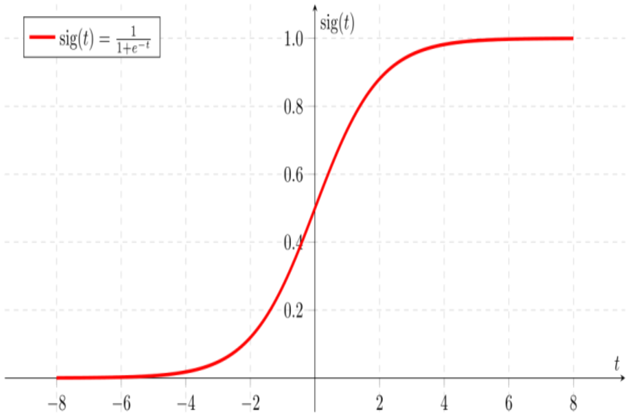


Figure 4. Logistic regression

In the specific context of predicting cardiovascular disease, logistic regression enables us to evaluate the likelihood of an individual having the condition based on their clinical and non-clinical characteristics. This modeling approach accommodates both continuous and categorical independent variables and can account for interactions between these variables, enabling a comprehensive analysis.

To employ logistic regression for cardiovascular disease prediction, a dataset containing known outcomes (e.g., presence or absence of cardiovascular disease) along with corresponding features (e.g., demographic and clinical data) is initially collected. This dataset is then split up into training and testing subsets. The logistic regression algorithm is trained using the training set and subsequently evaluated on the testing set to determine its predictive performance. By examining the model coefficients, we can estimate the relative significance of every feature in forecasting the outcome.

Different performance metrics are commonly employed to evaluate the efficacy of logistic regression models. These metrics consists of accuracy, sensitivity, specificity, area under the curve (AUC), and F1-score. Moreover, logistic regression models can be enhanced by integrating them with other ML algorithms like Support Vector Machines (SVM) shown in Figure 4 and random forests, aiming to bolster the accurateness of predictions.

4. ANALYSIS AND DISCUSSION OF THE RESULT

This section presents the experimental findings applied to a dataset employing five popular ML algorithms on relevant features. This study report also presents a comparison between the proposed model and previous studies.

4.1 Examination of experiments

In the experimental setup, we utilized a Quad-core i5 system with 4GB of RAM within the Collaborator web application environment for model development. Our focus was on employing pandas and SciPy libraries. As part of performance evaluation, we employed the Confusion matrix, which serves as a valuable measurement tool. It represents a multidimensional matrix of size N*N, effectively summarizing the classification performance of a given classifier on specific test data.

The confusion matrix consists of the following dimensions: "Actual" and "Predicted," and encompasses key metrics such as "True Positives (TP)," "True Negatives (TN)," "False Positives (FP)," and "False Negatives (FN)" for both dimensions. These metrics collectively contribute to assessing the effectiveness and exactness of the classifier. For a more comprehensive overview, please refer to Table 1, which presents the evaluation metrics acquired from the experiment.

Table 1. Evaluation metrics

METRICS	DEFINITION
Precision	$\frac{True\ Positives}{True\ Positives + False\ Positives}$
Recall	$\frac{True\ Positives}{True\ Positives + False\ Negatives}$
F1 – Score	$\frac{2 * Precision * Recall}{Precision + Recall}$
Sensitivity	$\frac{True\ Positives}{Positives}$
Specificity	$\frac{True\ Negatives}{Negatives}$
Accuracy	$\text{Sensitivity} * \frac{Positives}{Positives + Negatives} +$ $\text{Specificity} * \frac{Negatives}{Positives + Negatives}$

Evaluation metrics are employed to evaluate the effectiveness of machine learning models in predicting cardiovascular disease outcomes. These metrics provide a quantitative measure of how well the model is performing and isuseful for comparing different models and select the best one for a given task. Here are some generally used evaluation metrics in cardiovascular disease prediction:

4.1.1 Accuracy

When assessing the effectiveness of predictive models, such as those used to forecast cardiovascular disease (CVD), accuracy is a frequently used indicator. The percentage of accurately identified instances—both true positives and true negatives—out of all instances is known as accuracy. The percentage of cases among all the persons that were correctly classified. This metric measures the as a whole performance of the model and is sensitive to class imbalance. Make sure your model can be understood and that it can explain the predictions it makes. Both patients and healthcare providers benefit from this increased trust [20].

4.1.2 Sensitivity

Sensitivity is calculated using the formula

Sensitivity (Recall) =True Positive (TP)/True Positive (TP)+False Negative (FN)

The percentage of real positive cases that the model accurately detects is what it measures.

In situations where overlooking positive examples is expensive or has major repercussions, high sensitivity is preferred. For instance, high sensitivity is essential in medical diagnosis in order to identify as many true positive instances (i.e., patients with a disease) as possible accurately, albeit at the expense of a greater false positive rate.

The percentage of true positive instances (i.e., patients with cardiovascular disease appropriately identified by the model) out of the total quantity of positive instances (i.e., all patients with cardiovascular disease). This metric measures the model's capacity to correctly identify patients with cardiovascular disease.

4.1.3 Specificity

The prediction of cardiovascular disease (CVD) is estimating a person's chance of experiencing cardiovascular events over a specific time frame, such as heart attacks or strokes. One of the most important measures for assessing how well prediction models work is specificity. A model's specificity is determined by how well it can distinguish between people who truly do not have the illness and those who do not (true negatives). The percentage of true negative instances (i.e., patients without cardiovascular disease correctly identified by the model) out of the total number of negative instances (i.e., all patients without cardiovascular disease). This metric measures the capacity of the model to finely identify patients without cardiovascular disease.

4.1.4 Positive predictive value (PPV)

Precision, or Positive Predictive Value (PPV), is a statistic used to evaluate a predictive model's effectiveness, especially when it comes to the prediction of cardiovascular disease (CVD). The percentage of true positive instances out of the total count of instances predicted as positive by the model.

PPV=True positives/ True positives+False positives

This metric measures the probability that a patient predicted to have cardiovascular disease actually has the disease.

4.1.5 Negative predictive value (NPV)

Particularly when it comes to the prediction of cardiovascular disease (CVD), the metric known as Negative Predictive Value (NPV) is employed to evaluate the efficacy of a predictive model. In all cases predicted as negative by the model, NPV quantifies the percentage of actual negative predictions (i.e., accurately detected instances where CVD is absent).The proportion of true negative instances out of the total quantity of instances predicted as negative by the model. The implications of incorrect negative predictions determine whether or not NPV is clinically relevant. A high NPV is preferred in CVD prediction to reduce the possibility of false negative results, which could mean missing a chance for preventive interventions.

NPV=True Negatives/ True Negatives +False Negatives

This metric measures the most likely that a patient predicted not to have cardiovascular disease actually does not have the disease.

4.1.6 F1-score

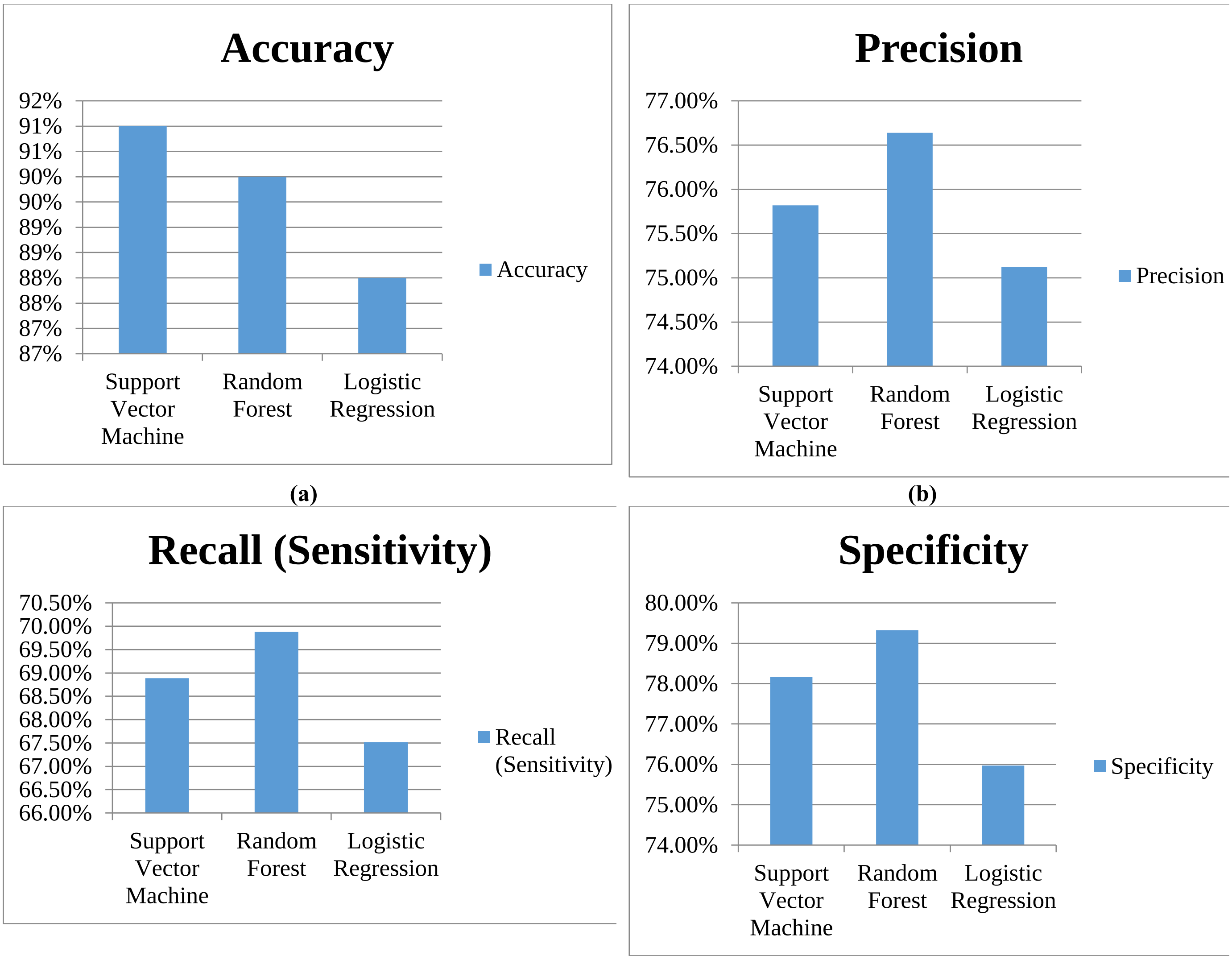
In binary classification issues, such as those involving the prediction of cardiovascular illnesses, the F1-score is a measure that is frequently employed. When it comes to predicting diseases, including cardiovascular disease, the F1-score offers a compromise between recall and precision.

The harmonic mean of precision and recall, which balances the exchange between sensitivity and specificity. This metric is useful when the classes are imbalanced and one class is of more interest than the other.

These evaluation metrics can be utilized to evaluate the effectiveness of many ML models and select the best one for cardiovascular disease prediction. Comprehending accuracy as well as recall is essential for evaluating the model's efficacy in cardio vascular disease, where false positives and false negatives might have disparate effects.

4.2 Result analysis

Our intended results and accuracy levels through the application of several machine learning techniques are shown in Table 2. For each algorithm the parameter values are shown graphically in Figure 5 (a), 5(b), 5(c) and 5(d). For Random Forest we got an accuracy of 90%, Precision of 76.64%, Recall 69.88%, Specificity of 79.32%. The second model, that is, Support Vector Machine gave an accuracy of 91%, Precision of 75.82%, Recall of 68.89% and Specificity of 78.16%. The Logistic Regression shows the accuracy of 88%, Precision of 75.12%, Recall of 67.52% and Specificity of 75.97%.



(a) **(b)**

(c) **(d)**

Figure 5. Precision Graph-three algorithms

Table 2. Metric measures for different models

	Accuracy	Precision	Recall (Sensitivity)	Specificity
Support Vector Machine	91%	75.82%	68.89%	78.16%
Random Forest	90%	76.64%	69.88%	79.32%
Logistic Regression	88%	75.12%	67.52%	75.97%

In the prediction of cardiovascular disease (CVD), Random Forest may perform better than SVM or Logistic Regression. For complex nonlinear interactions between predictor variables and the result (existence of CVD), Random Forest is an effective tool. Because the association between risk variables (such as blood pressure, cholesterol, and lifestyle factors) and the development of CVD can be very nonlinear, this is especially helpful in the prediction of CVD. Class imbalance is a common problem with CVD datasets, meaning that there are much more examples of one class (people without CVD, for example) than of the other class (those with CVD). Random Forest can manage situations like these since it is comparatively resilient to class imbalance. The Random Forest algorithm yields a measure of feature relevance that helps identify the variables that have the greatest impact on CVD prediction. Finding the main risk factors for CVD and directing future studies or treatments can both benefit from this information.

To be useful in a real-world context, CVD models must perfectly interface with current electronic health record (EHR) systems and clinical procedures. In order to operate in real-time or almost real-time within clinical workflows, models need to be computationally efficient. This entails minimizing processing time by optimizing algorithms and architectures, particularly when dealing with big datasets or intricate models. For instance, Random Forest can be computationally demanding; nevertheless, approximation approaches and model reduction can assist lower computational costs without appreciably compromising performance.

5. CONCLUSION

This project focuses on developing a machine learning-based model for predicting cardiovascular disease. By utilizing machine learning techniques, the accuracy of predictions can be enhanced while reducing false positive rates, enabling earlier detection and treatment of cardiovascular disease. Our analysis encompasses a range of machine learning algorithms commonly applied in this domain, such as logistic regression, random forest, Support Vector Machine, and ensemble models. The probable of machine learning to transform cardiovascular disease prediction is evident, and we anticipate ongoing advancements in this field in the future.

Continuously monitor physiological data and biomarkers linked to cardiovascular health, such as blood pressure, heart rate variability, and blood glucose levels, by utilizing machine learning approaches. Machine learning algorithms have the ability to identify early warning indications of cardiovascular events or exacerbations, thereby providing patients and healthcare providers with timely interventions and preventive measures.

6. FUTURE SCOPE

The future scope is Predicting heart failure, identifying coronary artery disease risk factors, and estimating the risk of stroke are some of the specialized machine learning applications in cardiovascular disease evaluation. Enhancing the generalizability and performance of models in real-world clinical settings requires careful collaboration with healthcare experts, ethical considerations, and model validation on representative and heterogeneous datasets. With continuous improvements in technology and healthcare, the potential application of ML in the assessment of cardiovascular disease is significant.

Examine how to predict CVD using deep learning architectures like recurrent neural networks (RNNs) and convolutional neural networks (CNNs). These models can identify intricate patterns in longitudinal patient data, such as time-series data from wearable devices, or complicated patterns in medical imaging data, such as MRI and CT scans, which may include important prediction information.

Model intricate interactions among biomarkers, clinical outcomes, and cardiovascular risk factors by employing graph-based learning techniques. You can improve predictability and interpretability of medical knowledge by expressing it as graphs (disease networks, biological pathways, etc.) and using graph-based regularization techniques like graph neural networks.

REFERENCES

[1] Akay, Y.M., Welkowitz, W., Kostis, J. (1994). Noninvasive detection of coronary artery disease. IEEE Engineering in Medicine and Biology Magazine, (1994). 761-764. https://doi.org/10.1109/51.334639

[2] Shorewala, V. (2021). Early detection of coronary heart disease using ensemble techniques. Informatics in Medicine Unlocked, 26(2021): 100655. https://doi.org/10.1016/j.imu.2021.100655

[3] Sahoo, P.K., Thakkar, H.K., and Lee, M.Y. (2017). A cardiac early warning system with multichannel SCG and ECG monitoring for mobile health. Sensors, 17(4): 711. https://doi.org/10.3390/s17040711

[4] Ullah, T., Ullah, S.I., Ullah, K., Ishaq, M., Khan, A., Ghadi, Y.Y., Algarni, A. (2024). Machine learning-based cardiovascular disease detection using optimal feature selection. IEEE Access, 12: 16431-16446. https://doi.org/10.1109/ACCESS.2024.3359910

[5] Bhatt, C., Patel, P., Ghetia, T., Mazzeo, P.L. (2023). Effective heart disease prediction using machine learning techniques. Algorithms, 16(2): 88. https://doi.org/10.3390/a16020088

[6] Diker, A., Aykut, et al. (2018). Intelligent system based on Genetic Algorithm and Support Vector Machine for detection of myocardial infarction from ECG signals. 2018 26th Signal processing and communications applications conference (SIU), Kuala Lumpur, Malaysia. https://doi.org/10.1109/SIU.2018.8404299

[7] Muthulakshmi, P., Parveen, M., Rajeswari, P. (2023). Prediction of heart disease using ensemble learning. Indian Journal of Science and Technology, 16(20), 1469-1476. https://doi.org/10.17485/IJST/v16i20.2279

[8] Akkaya, B., Sener, E., Gursu, C. (2022). A comparative study of heart disease prediction using machine learning techniques. 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). https://doi.org/10.1109/HORA55278.2022.9799978

[9] Rahmat, D., Putra, A.A., Setiawan, A.W. (2021). Heart disease prediction using K-nearest neighbor. 2021 International Conference on Electrical Engineering and Informatics (ICEEI), Kuala Terengganu, Malaysia, 1-6. https://doi.org/10.1109/ICEEI52609.2021.9611110

[10] Ahsan, Md M., Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. Artificial Intelligence in Medicine, 128. https://doi.org/10.1016/j.artmed.2022.102289

[11] Grundy, S. M., Pasternak, R., Greenland, P., Smith Jr, S., Fuster, V. (1999). Assessment of cardiovascular risk by use of multiple-risk-factor assessment equations: a statement for healthcare professionals from the American Heart Association and the American College of Cardiology. *Circulation*, *100*(13), 1481-1492https://www.ahajournals.org/doi/epub/10.1161/01.CIR.100.13.1481

[12] Adhishayaa, P.V., Gomathi, V., Mahendran, K. (2023). Review on cardiovascular disease prediction using machine learning algorithm. 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India. https://doi.org/10.1109/ICCCI56745.2023.10128403

[13] Cuevas-Chávez, A., Hernández, Y., Ortiz-Hernandez, J., Sánchez-Jiménez, E., Ochoa-Ruiz, G., Pérez, J., González-Serna, G. (2023). A systematic review of machine learning and IoT applied to the prediction and monitoring of cardiovascular diseases. Healthcare, 11(16), 2240. https://doi.org/10.3390/healthcare11162240

[14] Dalal, S., Adhishayaa, P.V., Bhardwaj, A. (2023). Application of machine learning for cardiovascular disease risk prediction. Computational Intelligence and Neuroscience, 2023. https://doi.org/10.1155/2023/9418666

[15] Ogunpola, A., Hernández, Y., Ortiz-Hernandez, J., Sánchez-Jiménez, E., Ochoa-Ruiz, G., Pérez, J., González-Serna, G. (2023). Machine learning-based predictive models for detection of cardiovascular diseases. Diagnostics, 14(2): 144. https://doi.org/10.3390/diagnostics14020144

[16] Balakrishnan, M., Arockia Christopher, A.B., Ramprakash, P., Logeswari, A. (2021). Prediction of cardiovascular disease using machine learning. Journal of Physics: Conference Series, 1767(1): 012013. https://doi.org/10.1088/1742-6596/1767/1/012013

[17] Divyasri, P., SreeLakshmi, D., Sathvika, P., Teja, P., Charan, T.V. (2023). Cardiovascular disease prediction using machine learning. In 2023 6th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, pp. 1-6. https://doi.org/10.1109/ISCON57294.2023.10112052

[18] Singh, T. (2023). Prediction of heart disease diagnosis using deep learning fusion algorithm for patient monitoring system. In 2023 3rd Asian Conference on Innovation in Technology (ASIANCON), Ravet IN, India, pp. 1-7. https://doi.org/10.1109/ASIANCON58793.2023.10269897

[19] Gudadhe, M., Wankhade, K., Dongre, S. (2010). Decision support system for heart disease based on Support Vector Machine and artificial neural network. In 2010 International Conference on Computer and Communication Technology (ICCCT), Allahabad, India, pp. 741-745. https://doi.org/10.1109/ICCCT.2010.5640377

[20] Divya, A., Deepika, B., Akhila, C.H.D., Devi, A.T., Lavanya, B., Teja, E.S. (2022). Disease prediction based on symptoms given by user using machine learning. SN Computer Science, 3: 504. https://doi.org/10.1007/s42979-022-01399-0