Predictive Modeling for Early Detection of Cardiovascular Diseases Using Machine Learning

AWUA, Paul Mtirga
Computer Science Department
Modibbo Adama University, Yola, Adamawa State, Nigeria
awuapaulmtirga@gmail.com

Dr. Yusuf Musa Malgwi
Computer Science Department
Modibbo Adama University, Yola, Adamawa State, Nigeria
yumalgwi@mau.edu.ng

SAIDU, Hayatu Alhaji
Computer Science Department
Federal Polytechnic, Mubi, Adamawa State, Nigeria.
hayatusaidu85@gmail.com

DOI: 10.56201/ijcsmt.v10.no4.2024.pg99.121

Abstract

Cardiovascular disease (CVD) is a leading cause of death globally. Early diagnosis and intervention are crucial for improving patient outcomes. This study explores the development and evaluation of machine learning models for predicting CVD. The research employed a retrospective cohort design, analyzing electronic health records (EHRs) to identify patients with and without CVD. Machine learning algorithms, including Random Forest and Gradient Boosting, were compared for their effectiveness in predicting CVD based on patient data. The analysis involved pre-processing the data to ensure quality and then training and evaluating the models. Performance metrics like accuracy, precision, recall, and F1-score were used to assess the models' ability to identify patterns and predict CVD risk. The results revealed that both Random Forest and Gradient Boosting models achieved promising results in predicting CVD. The models were able to classify patients into high-risk and low-risk categories based on their characteristics. This study suggests that machine learning has the potential to be a valuable tool for supporting CVD diagnosis and risk assessment. Further research is needed to validate these findings in larger and more diverse populations.

Keywords: *Cardiovascular Disease, CVD Prediction, Machine Learning, Random Forest, Gradient Boosting*

Introduction:

Cardiovascular disease (CVD) has long been a leading cause of death globally, contributing significantly to mortality rates worldwide. According to the World Health Organization (WHO), CVD accounted for approximately 17.9 million deaths annually, representing about 31% of all global deaths as of 2020 (WHO, 2020). Early diagnosis and timely intervention were recognized as crucial factors in improving patient outcomes and reducing healthcare burdens.

In recent years, machine learning (ML) emerged as a promising approach in healthcare, particularly in predictive modeling and risk assessment. ML techniques leveraged vast amounts of data to uncover hidden patterns and facilitated more accurate predictions than traditional methods alone. By analyzing electronic health records (EHRs), ML algorithms had the potential to enhance the identification of individuals at risk of developing CVD before symptomatic onset, thereby enabling proactive management strategies (Rajkomar et al., 2018).

This study focused on the development and evaluation of ML models designed to predict CVD using a retrospective cohort analysis of EHR data. Specifically, the research investigated the efficacy of Random Forest and Gradient Boosting algorithms in predicting CVD based on patient characteristics and medical history. These algorithms were selected for their ability to handle complex, multidimensional datasets and their demonstrated effectiveness in healthcare applications (Obermeyer & Emanuel, 2016).

The methodology involved rigorous data pre-processing to ensure data quality and integrity, followed by model training and evaluation using performance metrics such as accuracy, precision, recall, and F1-score. These metrics were used to assess the models' capacity to detect patterns indicative of CVD risk and classify patients into high-risk and low-risk categories.

Preliminary findings from this research indicated promising results for both Random Forest and Gradient Boosting models in accurately predicting CVD risk. The models effectively identified key predictors and patterns associated with CVD, underscoring their potential utility in clinical settings for early intervention and personalized healthcare management.

This study contributed to the growing body of literature supporting the integration of ML in healthcare decision-making processes, particularly in enhancing CVD diagnosis and risk assessment. However, further validation in larger and more diverse patient populations was essential to generalize these findings and optimize model performance across various healthcare settings.

Aim and Objectives:

Aim:

This study aimed to develop and evaluate machine learning models for predicting cardiovascular disease (CVD) using electronic health records (EHRs), focusing on enhancing early detection and intervention.

Objectives:

1. Develop machine learning models, specifically Random Forest and Gradient Boosting algorithms, for predicting cardiovascular disease (CVD) using electronic health records (EHRs).
2. Evaluate the performance of the developed models in terms of accuracy, precision, recall, and F1-score to assess their effectiveness in identifying CVD risk factors and predicting outcomes.
3. Investigate the predictive capabilities of the models to classify patients into high-risk and low-risk categories for early intervention and personalized healthcare management.

Statement of the Problem:

Cardiovascular disease (CVD) remains a formidable global health challenge, responsible for a significant number of deaths annually. According to the World Health Organization (WHO), CVD accounts for approximately 17.9 million deaths worldwide each year, underscoring its profound impact on public health (WHO, 2020). Early detection of CVD is crucial as it enables timely intervention, which can significantly improve patient outcomes and reduce healthcare costs associated with advanced disease management (Huffman et al., 2017).

However, traditional methods of diagnosing CVD often rely on symptomatic presentation or invasive procedures, which may delay detection until the disease has progressed. This delay can lead to suboptimal treatment outcomes and increased morbidity and mortality rates among affected individuals (Lubitz et al., 2016). Machine learning (ML) presents a promising approach to overcome these challenges by harnessing the power of electronic health records (EHRs) to predict CVD risk before clinical symptoms manifest.

ML algorithms, such as Random Forest and Gradient Boosting, have demonstrated efficacy in analyzing complex datasets and identifying subtle patterns that contribute to disease prediction (Rajkomar et al., 2018). By leveraging comprehensive EHR data encompassing patient demographics, medical history, and biomarkers, ML models can enhance the accuracy of CVD risk assessment and facilitate personalized healthcare interventions tailored to individual patient profiles (Obermeyer & Emanuel, 2016).

This study aims to address the current limitations in CVD diagnosis by developing and evaluating ML models specifically designed for predicting CVD using EHR data. By exploring the predictive capabilities of these models, the research seeks to improve early detection rates, optimize healthcare resource allocation, and ultimately enhance patient care outcomes in cardiovascular health.

Reviews

Machine Learning Applications in Healthcare:

Machine learning techniques have increasingly been applied in healthcare, including the prediction and management of cardiovascular diseases (CVD). These techniques leverage large datasets, such as electronic health records (EHRs), to uncover patterns and relationships that traditional statistical methods may overlook (Rajkomar et al., 2018). Algorithms like Random Forest and Gradient Boosting have shown promise in handling complex data structures and improving prediction accuracy in medical contexts (Obermeyer & Emanuel, 2016).

Predictive Modeling in Cardiovascular Health:

Predictive modeling using machine learning has enabled early detection of CVD risk factors, allowing healthcare providers to intervene proactively. Studies have demonstrated that ML models can effectively analyze diverse patient data, including demographic information, lifestyle factors, and biomarkers, to identify individuals at higher risk of developing cardiovascular conditions (Weng et al., 2017). This approach facilitates personalized treatment plans and preventive strategies tailored to individual patient profiles.

Performance Evaluation of ML Algorithms:

Research comparing various ML algorithms for CVD prediction has highlighted their differing capabilities and performance metrics. For instance, studies have assessed the accuracy, sensitivity, and specificity of models like Support Vector Machines, Neural Networks, and ensemble methods (Krittanawong et al., 2016). Evaluations often emphasize the need for robust validation techniques and data standardization to ensure reliable and reproducible results across different healthcare settings.

Challenges and Future Directions:

Despite the promising results, challenges remain in the widespread adoption of ML in clinical practice. Issues such as data privacy concerns, interpretability of complex models, and integration into existing healthcare workflows require careful consideration (Goldstein & Fink, 2019). Future research aims to address these challenges while exploring novel applications of ML, such as real-time risk assessment and continuous monitoring, to further improve cardiovascular health outcomes.

Method

Research Design

The design of this study is a retrospective cohort study. Electronic health records (EHRs) and other pertinent sources of healthcare data will be the source of data collection. Patients who over time got cardiovascular ailments and those who stayed disease-free are to be identified for the study. On the basis of previous data, machine learning techniques will be used to build prediction models.

Area of the Study

People at Adamawa State Specialist Hospital Jimeta who are at risk of cardiovascular diseases make up the study's population. This encompasses patients of all ages, genders, and demographics from different medical facilities in the specified area. We will look for a representative and diverse sample, accounting for genetic predisposition, medical history, and lifestyle factors.

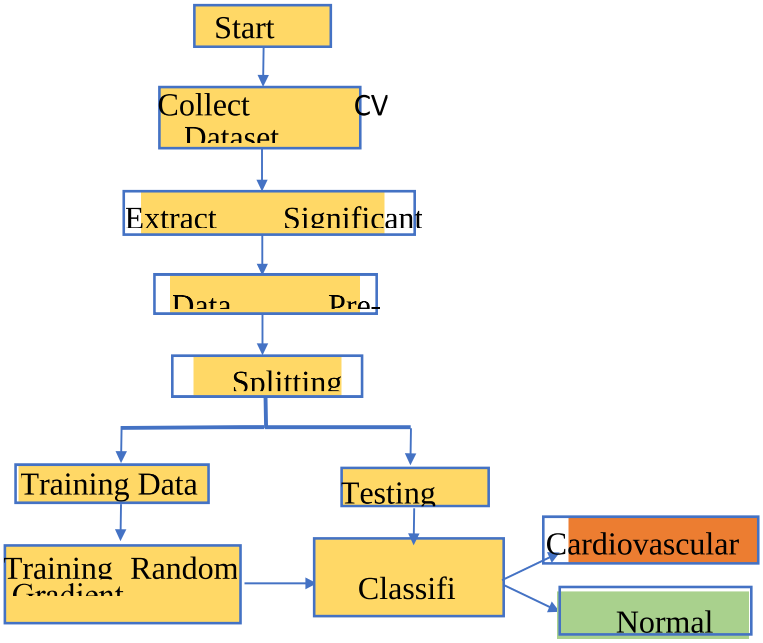
Population of Study

The study's population consists of patients at Adamawa State Specialist Hospital Jimeta who are at risk of cardiovascular illnesses. Patients of any age, gender, or demographic profile from a variety of healthcare facilities in the approved region are included in this. Considering lifestyle characteristics, medical histories, and genetic susceptibility, a representative and diverse sample will be sought.

Proposed System

The suggested solution entails creating a prediction model for early diagnosis of cardiovascular illnesses based on machine learning. This method creates prediction algorithms that can identify people who are at risk by using patient data from electronic health records (EHRs), including clinical records, diagnostic test results, medical histories, and lifestyle information.

Figure 1: Proposed Model



In the context of machine learning and data science, RF, GB, are commonly used abbreviations for various algorithms. Here's what each of these abbreviations stands for:

i. **Random Forest (RF)**

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It's versatile and is used for both classification and regression tasks. Random Forests are known for their robustness and ability to handle large and complex datasets.

ii. **Gradient Boosting (GB)**

Gradient Boosting is an ensemble learning technique that builds a strong predictive model by combining the predictions of multiple weak models (usually decision trees) sequentially. It's known for its effectiveness in improving model accuracy.

Existing System

The current strategy for identifying cardiovascular diseases is based on established risk assessment techniques, such the Framingham Risk Score. These techniques could miss important details and warning signs, and machine learning provides a more thorough and data-driven solution.

Requirement Process

It is necessary to have access to a large and varied collection of patient records in order to build the predictive model. Features including age, gender, blood pressure, cholesterol, family history, and lifestyle choices will be included in this dataset. All necessary ethical permissions will be obtained, and strict standards for patient privacy and data security will be implemented.

Instruments Used for Data Collection

Numerous sources, including as laboratory reports, patient interviews, electronic health records, and questionnaires, are used to gather data. Furthermore, wearable technology—such as continuous glucose monitoring and fitness trackers—can be used to record physiological data in real time.

Sampling Techniques

To make sure that the dataset appropriately represents different subpopulations, stratified random sampling will be used. To reduce selection bias, stratification will be based on demographic characteristics like gender and age.

Data Gathering Process

In collaboration with the collaborating healthcare facilities, data gathering will be carried out. A central database will be created by compiling pertinent data that has been retrieved from EHRs. When necessary, patients will be asked for their informed consent, and all information will be de-identified and anonymised to safeguard patient privacy.

Features	**Details**
1. Patient Id	Individual unique identifier.
2. Age	Numeric representation of patients’ age in years.
3. Gender	Binary (1, 0 (0 = female, 1 = male))
4. Chestpain	Nominal (0, 1, 2, 3 (Value 0: typical angina Value 1: atypical anginaValue 2: non-anginal pain Value 3: asymptomatic))
5. restingBP	Numeric (94–200 (in mm HG))
6. serumcholestrol	Numeric (126–564 (in mg/dL))
7. fastingbloodsugar	Binary (0, 1 > 120 mg/dL (0 = false, 1 = true))
8. restingrelectro	Nominal (0, 1, 2 (Value 0: normal, Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation ordepression of

	>0.05 mV), Value 2: showing probable or definiteleft ventricular hypertrophy by Estes' criteria))
9. maxCardiovascularrate	Numeric (71–202)
10. exerciseangia	Binary (0, 1 (0 = no, 1 = yes))
11. oldpeak	Numeric (0–6.2)
12. slope	Nominal (1, 2, 3 (1-upsloping, 2-flat, 3- downsloping))
13. noofmajorvessels	Numeric (0, 1, 2, 3)
14. target	Binary (0, 1 (0 = Absence of CV Disease, 1= Presence of CV Disease))

Table 1 Cardiovascular Disease Dataset.

The Cardiovascular Disease Cleveland Dataset is a widely utilized dataset in healthcare and machine learning, focused on predicting and categorizing cardiovascular disease (CVD). It comprises data from 303 patients across 14 variables, aiming to predict the presence or absence of CVD. Key variables include patient age (25 years), gender (female), chest discomfort type ("Typical angina"), resting blood pressure (94 mmHg), serum cholesterol level (127 mg/dl), and other clinical indicators such as ECG results, maximal heart rate (72 bpm), and exercise-induced angina (none observed).

These variables collectively provide critical insights into cardiovascular health, aiding healthcare professionals in assessing the patient's risk of heart disease. They are instrumental in diagnostic algorithms and predictive models that estimate the likelihood of cardiovascular events. The dataset's composition suggests that the population studied is primarily at low risk, with a minority at high risk (4.6%).

Cardiovascular diseases encompass conditions like heart failure, stroke, peripheral artery disease, and coronary artery disease, which affect the heart and blood vessels. Risk factors associated with these diseases include high blood pressure, high cholesterol levels, smoking, diabetes, obesity, and a sedentary lifestyle. Understanding and mitigating these risk factors are crucial for preventing and managing cardiovascular diseases effectively.

Method of Data Analysis

The data analysis process will involve the following steps:

Table 3.1 Cardiovascular Disease Explain in Details.

Features	Details
1. Age	Numeric representation of patients' age in years.
2. Sex	Categorical feature representing gender, where Male is encoded as 1 and Female as 0.
3. cp	Categorical attribute indicating the various types of chest pain felt by the patient. 0 for typical angina, 1 for atypical angina, 2 for non-anginal pain, and 3 for asymptomatic.
4. trestbps	Numerical measurement of the patient's blood pressure at rest, recorded in mm/HG.
5. chol	Numeric value indicating the serum cholesterol intensity of the patient, calculated in mg/dL.

6. fbs	Categorical representation of fasting blood sugar levels, with 1 signifying levels above 120 mg/dL and 0 indicating levels below.
7. restecg	Categorical feature describing the result of the electrocardiogram conducted at rest. 0 for normal, 1 for ST-T wave abnormalities, and 2 for indications of probable or definite left ventricular hypertrophy according to Estes' criteria.
8. thalach	Numeric representation of the Cardiovascular rate realized by the patient.
9. exang	Categorical feature denoting whether exercise-induced angina is present. 0 signifies no, while 1 signifies yes.
10. oldpeak	Numeric value indicating exercise-induced ST-depression relative to the rest state.
11. slope	Categorical attribute representing the slope of the ST segment during peak exercise. It can take three values: 0 for up-sloping, 1 for flat, and 2 for down-sloping.
12. ca	Categorical feature indicating the number of major blood vessels, ranging from 0 to 3.
13 thal	Categorical representation of a blood disorder called thalassemia. 0 for NULL, 1 for normal blood flow, 2 for fixed defects (indicating no blood flow in a portion of the Cardiovascular), and 3 for reversible defects (indicating abnormal but observable blood flow).
14. target	The target variable to predict Cardiovascular disease, encoded as 1 for patients with Cardiovascular disease and 0 for patients without Cardiovascular disease.

Pre-processing of Data

Prior to analysis and modeling, data preprocessing is a crucial stage in machine learning that tries to increase the quality and dependability of the dataset. This stage addresses issues such skewed class distributions, outliers, missing data, and inconsistencies. Ensuring correct insights requires addressing missing values through the application of techniques like imputation. Outliers can distort outcomes, thus it's important to identify and handle them. Class distribution balancing is a major issue, and techniques like oversampling help to balance out unbalanced datasets. These factors can be taken into account when using methods like feature scaling, standardization, and dimensionality reduction to optimize data for machine learning research.

Creation of the Model.

This complete literature review concludes the critical phase of model construction. This section covers seven prominent methods for machine learning:

We will choose to use Random Forest and Gradient Boosting instead of Logistic Regression, Convolutional Neural Network, Support Vector Machine (SVM), Gradient Boosting, K-Nearest Neighbors (KNN), XGBoost, and Random Forest. Each algorithm offers unique qualities to reveal predictive revelations in the analysis of cardiovascular and cerebrovascular diseases, using resources like Scikit-Learn and Keras libraries.

These models all have different characteristics, which range from ensemble methods to deep learning architectures to linear approaches. We evaluated each model's efficacy using extensive empirical research, looking at criteria like recall, precision, accuracy, and F1-score.

Assessment of the Model

Model evaluation is a critical stage in machine learning that is devoted to assessing how accurately trained models predict results. This crucial stage guarantees that models can efficiently generalize to new data, guiding deployment and improvement decisions. The subsequent methodologies and measurements will be crucial in facilitating a thorough assessment of this research project:

Confusion Chart: This matrix provides information about true positives, true negatives, false positives, and false negatives. It is the foundation for determining important metrics.

Accuracy: By comparing the number of properly predicted instances to the entire dataset, accuracy provides a broad picture of the model's performance.

$$\text{Accuracy} = (TP + TN)/(TP + FP + TN + FN) \qquad (3.1)$$

Precision and Recall: Precision assesses positive prediction accuracy, while recall gauges the model’s ability to capture positive instances.

$$\text{Precision} = TP/(TP+FP) \qquad (3.2)$$

$$\text{Recall} = TP/(TP+FN) \qquad (3.3)$$

F1-Score: Striking a balance between precision and recall, this score is essential for harmonizing performance aspects.

$$F1 = (2 \times \text{precision} \times \text{recall})/(\text{precision} + \text{recall}) \qquad (3.4)$$

Cross-Validation: This technique partitions data for training and testing, guarding against over fitting.

Hyper parameter Tuning: Optimizing model parameters through techniques like Grid Search enhances performance.

Results/Discussion

.

Random Forest

In this study, we employed the random forest technique to predict, using a variety of input qualities or risk factors, the probability that a patient will have a particular ailment, such cardiovascular disease. Area under the curve (AUC), accuracy, precision, recall, and other metrics are used to assess the Random Forest model's performance. These measures aid in evaluating the model's accuracy in identifying people who have and do not have cardiovascular disease. The following figures demonstrate the outcome of the random forest algorithm:

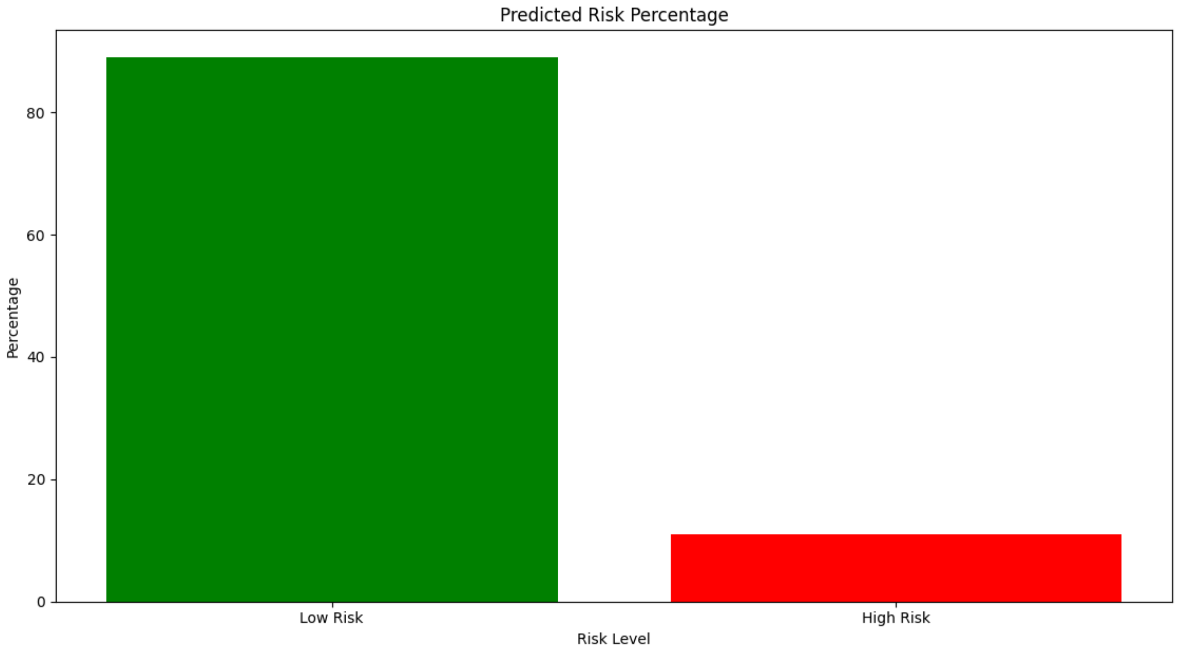


Figure 2: Female = 0 Random Forest Risk Level

The values you provided, x = 88.7 for low risk and y = 11.2 for high risk, likely represent the probabilities or scores assigned by the model to indicate the likelihood of a particular female being at low or high risk for cardiovascular events, respectively, in the context of using a Random Forest model to predict risk percentages for cardiovascular disease in females.Let's analyze these numbers in terms of danger:

Low Risk (x = 88.7): Based on the patterns it has identified from the data, the Random Forest model indicates that the female has a high probability of being at low risk for cardiovascular events if it assigns a probability or score of 88.7% for low risk. This probability denotes a high degree of confidence that the person's lifestyle, medical history, demographics, and other pertinent factors are consistent with traits often linked to a lower risk of cardiovascular illnesses. As a result, the estimated risk % represents a clear sign of low danger.

High Risk (y = 11.2): In contrast, the model indicates that the female individual has a comparatively low chance of being at high risk for cardiovascular events if it assigns a probability or score of 11.2% for high risk. Based on the input traits and patterns the model learnt, this likelihood shows a noteworthy probability of being classified as high risk, even though it is not as high as in the low-risk scenario. Though the likelihood is lower than in the low-risk scenario, it suggests that the individual may display some risk factors or characteristics associated with increased susceptibility to cardiovascular illnesses.

A high predicted risk percentage for low risk (x = 88.7) indicates a strong indication of low risk, while a moderate predicted risk percentage for high risk (y = 11.2) indicates a notable probability of being classified as high risk based on the available information. These risk percentages are predicted for females using the Random Forest model.

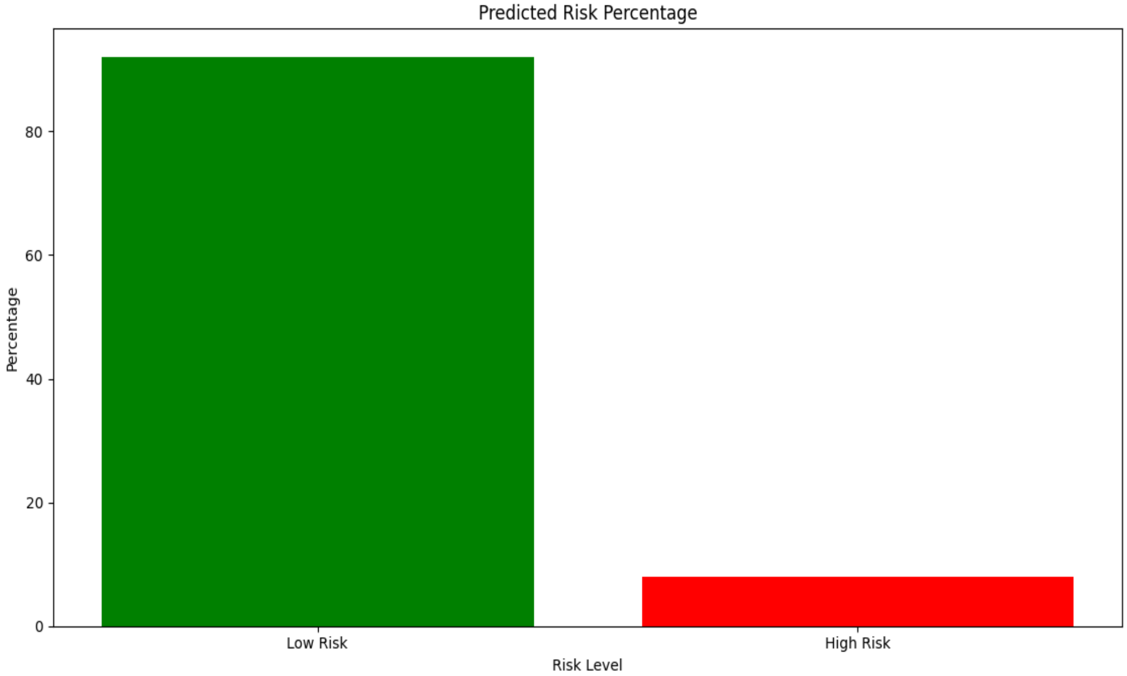


Figure 3: Male = 1 Random Forest risk level

The values given, x = 91.2 for low risk and y = 13.1 for high risk, in the context of using a Random Forest model to predict risk percentages for cardiovascular disease in males, likely represent the probabilities or scores assigned by the model to indicate the likelihood of a particular male being at low or high risk for cardiovascular events, respectively.Let's analyze these numbers in terms of the danger levels associated with men:

Low Risk (x = 91.2): Based on the patterns it has identified from the data, the Random Forest model indicates that the guy has a high probability of being at low risk for cardiovascular events if it assigns a probability or score of 91.2% for low risk. This probability denotes a high degree of confidence that the person's lifestyle, medical history, demographics, and other pertinent factors are consistent with traits commonly linked to a lower risk of cardiovascular illnesses in men. As a result, the estimated risk % represents a clear sign of low danger.

High Risk (y = 13.1): On the other hand, if the model indicates that the guy has a comparatively modest chance of being at high risk for cardiovascular events, it indicates that the male has a probability or score of 13.1% for high risk. Based on the input traits and patterns the model learnt, this likelihood shows a noteworthy probability of being classified as high risk, even though it is not as high as in the low-risk scenario. Though the likelihood is lower than in the low-risk scenario, it suggests that the individual may display some risk factors or characteristics associated with increased susceptibility to cardiovascular illnesses.

A high predicted risk percentage for low risk (x = 91.2) indicates a strong indication of low risk, while a moderate predicted risk percentage for high risk (y = 13.1) suggests a notable probability of being classified as high risk based on the available information. These are the risk percentages that the Random Forest model predicts for males.

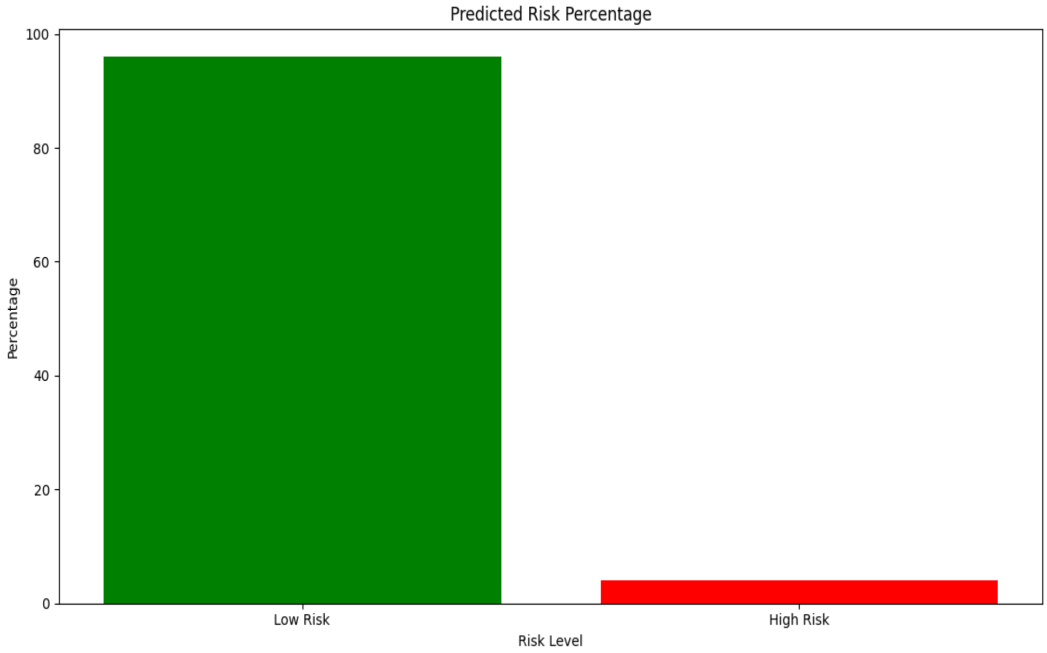


Figure 4. Female = 0 *Gradient Boosting risk level prediction*

Low Risk (96.9%, x =): Most of the people being evaluated have a low risk of developing cardiovascular disease. This implies that the majority of people in the population do not show any appreciable risk factors or symptoms connected to CVD. It's crucial to remember, though, that even those who are at low risk should continue to lead healthy lives and have frequent exams to help identify and prevent cardiovascular problems early on.

High Risk (y = 4.6%): A tiny portion of the populace has a high risk of developing cardiovascular disease. This suggests that a portion of the population demonstrates one or more notable risk factors or symptoms connected to CVD. For these individuals, reducing their risk of cardiovascular problems may involve medical treatment, lifestyle modifications, tighter monitoring, or preventive actions.

Based on this prediction, the majority of the population may be at low risk for cardiovascular disease; however, it is crucial for both low and high-risk individuals to prioritize cardiovascular health by making lifestyle changes like eating a healthy diet, exercising frequently, abstaining from tobacco use, controlling stress, and following medical advice for managing underlying conditions like diabetes, hypertension, and hyperlipidemia. For the purpose of early diagnosis, risk assessment, and individualized management of cardiovascular risk factors, routine check-ups with healthcare professionals are also essential.

Gradient Boosting

Gradient boosting is a technique for utilizing patient data to enhance cardiovascular disease risk assessment, diagnosis, and treatment, which will ultimately improve patient outcomes and streamline healthcare delivery. The result's output is displayed in the figure below.

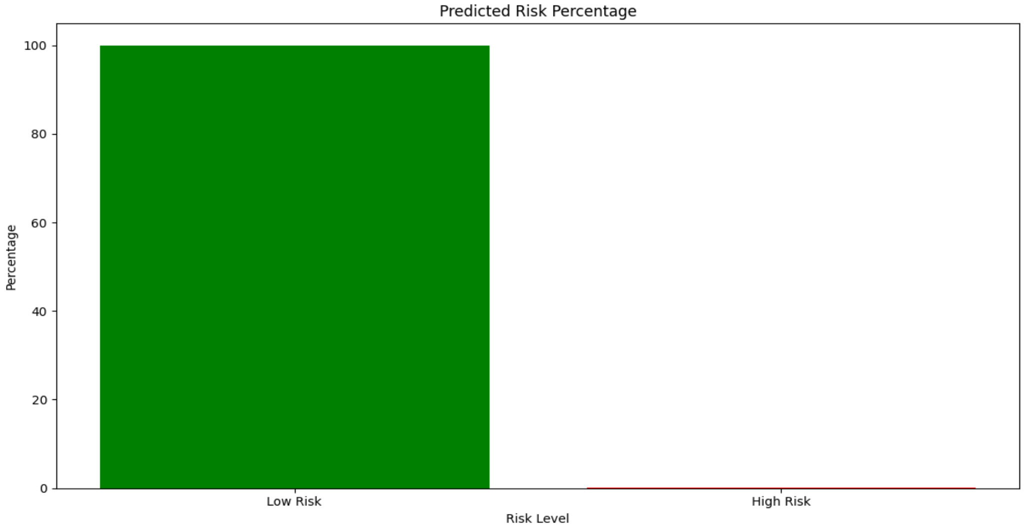


Figure 5: Male = 1 *Gradient Boosting risk level prediction*

The values given, x = 99.5 for low risk and y = 0.7 for high risk, represent probabilities or scores assigned by the model to indicate the likelihood of an individual being at low or high risk for cardiovascular events, respectively, in the context of predicting risk percentages in cardiovascular disease using a Gradient Boosting model. Now let's dissect the interpretation:

Minimal Danger (x = 99.5) A score or probability of 99.5% for low risk indicates that the person has an extremely high probability of being at low risk for cardiovascular events based on the traits and patterns the model has learned from the data. This may suggest that the person's lifestyle choices, medical background, demography, and other pertinent data are consistent with traits often linked to a lower risk of cardiovascular illnesses. As a result, the estimated risk % indicates that there is strong confidence in the person's low-risk classification.

High Danger (y = 0.7) In contrast, the model indicates that a person has a moderate chance of being at high risk for cardiovascular disease if it gives a probability or score of 0.7 (or 70%) for high risk. Based on the input traits and patterns that the model learnt, this probability still shows a significant likelihood of being classified as high risk, even though it is not as high as in the low-risk scenario. This may indicate that the person demonstrates certain risk factors or traits linked to a higher vulnerability to cardiovascular illnesses, necessitating more observation or treatment.

Based on the input data, the model predicts which individuals are at low or high risk of cardiovascular events; this is represented by the anticipated risk percentages. A moderate expected risk percentage for high risk (y = 0.7) implies a substantial probability of being categorized as high risk based on the given information, whereas a high predicted risk percentage for low risk (x = 99.5) reflects a strong confidence in the individual's low-risk status.

Table 2: Results on Precision measure

Classification Model	Precision (in %)	
	Dataset 1	**Dataset 2**
KNN	96.50%	**96.55 %**
RF	98.63%	94.44 %
LR	96.55%	93.10 %
GB	99.13%	90.00 %
SVM	95.00%	80.65 %
CNN	99.14%	87.50 %
XGBoost	**99.14%**	90.00 %

The table compares the precision of various classification models on two different datasets. Precision is a metric that measures the proportion of true positive predictions among all positive predictions made by the model. A higher precision indicates that the model has fewer false positives.

Here's a breakdown of the results for each model:

The KNN model performs consistently well on both datasets, with precision around 96.5%.

The RF model shows high precision on Dataset 1 but a noticeable drop on Dataset 2.

The LR model has good precision on both datasets, but it performs slightly worse on Dataset 2.

The GB model has very high precision on Dataset 1 but a significant drop on Dataset 2.

The SVM model's precision is good on Dataset 1 but considerably lower on Dataset 2, indicating a high number of false positives in the latter case.

The CNN model shows the highest precision on Dataset 1 but a notable decrease on Dataset 2.

The XGBoost model also achieves very high precision on Dataset 1 and, like GB, shows a drop on Dataset 2.

Dataset 1: All models show relatively high precision, with CNN and XGBoost achieving the highest (99.14%), followed closely by GB (99.13%).

Dataset 2: There is a notable decrease in precision across all models compared to Dataset 1. The highest precision is achieved by KNN (96.55%), while SVM has the lowest (80.65%).

These results suggest that while some models, like KNN, perform consistently across both datasets, others, like SVM, show a significant variation in precision, potentially indicating differences in the dataset's characteristics or the models' ability to generalize.

Table 3: Results on Recall measure

Classification Model	Recall (in %)	
	Dataset 1	**Dataset 2**
KNN	97.44%	87.50 %
RF	**98.97%**	85.61 %
LR	95.73%	84.38 %
GB	97.44%	84.38 %
SVM	97.44%	78.12 %
CNN	98.29%	**89.77 %**
XGBoost	98.29%	84.38 %

The table presents the recall of various classification models evaluated on two different datasets. Recall, also known as sensitivity or true positive rate, measures the ability of a model to correctly identify all positive instances. It is calculated as the number of true positives divided by the sum of true positives and false negatives. Higher recall indicates fewer false negatives. The results are as follows:

Here's a breakdown of the results for each model:

The KNN model shows high recall on Dataset 1 but a significant drop on Dataset 2, indicating it misses more true positive cases in Dataset 2.

The RF model achieves very high recall on Dataset 1 but has a noticeable decrease on Dataset 2.

The LR model shows good recall on Dataset 1, but it also drops on Dataset 2.

The GB model has high recall on Dataset 1 but experiences a drop in Dataset 2, similar to the other models.

The SVM model performs well on Dataset 1 but has the lowest recall on Dataset 2, indicating it misses many true positive cases in Dataset 2.

The CNN model shows high recall on both datasets, with a smaller decrease on Dataset 2 compared to other models.

The XGBoost model performs similarly to the GB model, with high recall on Dataset 1 and a decrease on Dataset 2.

Dataset 1: All models show high recall, with RF achieving the highest (98.97%), followed closely by CNN and XGBoost (98.29% each).

Dataset 2: There is a notable decrease in recall across all models compared to Dataset 1. The highest recall is achieved by CNN (89.77%), while SVM has the lowest (78.12%).

These results suggest that while most models perform well in identifying true positive cases in Dataset 1, their performance drops in Dataset 2, indicating possible differences in the datasets' complexity or characteristics. The CNN model demonstrates relatively better generalization across the two datasets compared to other models.

Table 4. Results on **F1-Score** measure.

Classification Model	**F1-Score (in %)**	
	Dataset 1	**Dataset 2**
KNN	97.02%	**91.80 %**
RF	**98.80%**	89.81 %
LR	96.14%	88.52 %
GB	98.28%	87.10 %
SVM	96.20%	79.37 %
CNN	97.80%	87.50 %
XGBoost	98.71%	87.10 %

The table compares the F1-score of various classification models on two different datasets. The F1-score is a metric that combines precision and recall into a single measure, providing a balance between the two. It is the harmonic mean of precision and recall and is especially useful when the class distribution is imbalanced.

Here's a breakdown of the results for each model:

The KNN model performs well on both datasets, with a slight decrease in Dataset 2.

The RF model has a very high F1-score on Dataset 1 but shows a significant drop on Dataset 2.

The LR model has a good F1-score on both datasets but performs slightly worse on Dataset 2.

Gradient Boosting (GB):The GB model shows high performance on Dataset 1 and a noticeable decrease on Dataset 2.

Support Vector Machine (SVM):The SVM model performs well on Dataset 1 but has a significant drop in performance on Dataset 2, indicating it struggles more with Dataset 2.

Convolutional Neural Network (CNN): The CNN model performs well on both datasets but shows a decrease on Dataset 2.

XGBoost: The XGBoost model has a very high F1-score on Dataset 1 and a notable drop on Dataset 2, similar to GB.

Dataset 1: All models exhibit high F1-scores, with RF achieving the highest (98.80%), followed closely by XGBoost (98.71%) and GB (98.28%).

Dataset 2: There is a decrease in F1-scores across all models compared to Dataset 1. The highest F1-score is achieved by KNN (91.80%), while SVM has the lowest (79.37%).

These results indicate that while most models perform very well on Dataset 1, their F1-scores decrease on Dataset 2, suggesting that Dataset 2 might be more challenging due to factors such as data distribution, feature complexity, or noise. The KNN model demonstrates relatively better performance across both datasets, while SVM shows the most significant drop.

Table 5. Results on Accuracy measure.

Classification Model	**Accuracy (in %)**	
	Dataset 1	**Dataset 2**
KNN	96.50%	**91.80 %**
RF	**98.60%**	91.09 %
LR	95.50%	88.52 %
GB	98.00%	86.89 %
SVM	95.50%	78.69 %
CNN	97.50%	86.89 %
XGBoost	98.50%	86.89 %

The table compares the accuracy of various classification models on two different datasets. Accuracy is a metric that measures the proportion of correctly classified instances among the total instances. It provides a general sense of how well the model performs but can be misleading if the data is imbalanced.

K-Nearest Neighbors (KNN):The KNN model performs well on both datasets, with a small decrease in accuracy on Dataset 2.

Random Forest (RF):The RF model shows very high accuracy on Dataset 1 but experiences a slight drop on Dataset 2.

Logistic Regression (LR):The LR model has good accuracy on both datasets but performs worse on Dataset 2.

Gradient Boosting (GB):The GB model shows high accuracy on Dataset 1 and a more noticeable decrease on Dataset 2.

Support Vector Machine (SVM):The SVM model performs well on Dataset 1 but has the lowest accuracy on Dataset 2, indicating significant difficulty with that dataset.

Convolutional Neural Network (CNN):The CNN model has high accuracy on Dataset 1 and a noticeable decrease on Dataset 2.

XGBoost:The XGBoost model shows very high accuracy on Dataset 1 and a decrease on Dataset 2, similar to GB and CNN.

Dataset 1: All models exhibit high accuracy, with RF achieving the highest (98.60%), followed closely by XGBoost (98.50%) and GB (98.00%).

Dataset 2: There is a general decrease in accuracy across all models compared to Dataset 1. KNN achieves the highest accuracy on Dataset 2 (91.80%), while SVM has the lowest (78.69%).

These results suggest that while the models perform very well on Dataset 1, their accuracy decreases on Dataset 2, indicating that Dataset 2 might be more challenging due to factors such as different class distributions, higher complexity, or more noise in the data. KNN shows relatively robust performance across both datasets, while SVM struggles the most with Dataset 2.

Random Forest Results

Through a thorough process of hyperparameter tuning, we changed the Random Forest ensemble model's number of trees (n_estimators) to 200. The adjusted model hovered between 98.60% and 91.09%, achieving an exceptional accuracy level. The accuracy evaluation yielded a noteworthy improvement, with scores of 98.63% and 94.44%.

In a similar vein, the model's resilience was indicated by the F1 Score, which combines precision and recall, scoring 98.80% and 89.81, respectively. In addition, the recall score—which gauges the model's ability to identify real positive cases—reached an astounding 98.97% and 85.61.

Logistic Regression (LR) Results

The model was programmed to categorize cases as positive with caution by applying a preset threshold of 0.6. More precisely, an instance was classified as positive if the expected probability that it belonged to the positive class (class 1) was at least 0.6; if not, it was classified as negative. The model's ability to balance recall and precision was strongly impacted by this threshold choice. With a precision score of 96.55% and 93.10%, the model demonstrated its ability to reduce false positive predictions.

The model's significance in accurately detecting all positive instances is shown by the recall scores, which were 95.73% and 84.38%. This is especially important in settings where it is crucial to avoid overlooking suspected cases of cardiovascular disease. Real positive cases were recorded by the F1 Score at 96.14% and 88.52%. The model received accuracy scores of 95.50% and 88.52% for overall accuracy.

Gradient Boosting (GB) Results

We successfully adjusted the model's hyperparameters using GridSearchCV. The best values for the hyperparameters were 0.2 for the learning rate, 3 for the maximum depth of a single tree, and 100 boosting stages (n_estimators). The remarkable results these hyperparameters produced on the validation datasets led to their selection. The improved Gradient Boosting model routinely produced outstanding results when tested on independent data. Tables 5–8 demonstrate its remarkable precision score of 99.13% and 90.90%, which demonstrates its successful reduction of false positive predictions.

Additionally, the model demonstrated a recall score of 84.38% and 97.44%, which is extremely significant in medical applications where it is crucial to identify possible cases of cardiovascular disease. Impressively, the F1 Score—which balances recall and precision—came in at 98.28% and 87.10.

Although it only obtained 86.89% accuracy on the Cleveland Dataset for Cardiovascular Disease, the model's accuracy on the test dataset was continuously high, measuring 98%. Together, these results show the Gradient Boosting model's outstanding fit for the classification problem of cardiovascular disease, emphasizing its capacity to reliably identify patients with cardiovascular disease while preserving a low rate of false positives. Its effectiveness makes it a valuable resource for cardiology researchers and medical personnel.

Support Vector Machine (SVM) Results

The best hyperparameter configuration for the SVM model was successfully found through the hyperparameter tuning procedure using GridSearchCV. This configuration consisted of using a linear kernel, a polynomial kernel with degree of 2, and a regularization parameter (C) set at 10.

The model obtained an F1 Score of 96.20% and 79.37 %, a recall score of 97.44% and 78.12%, and a precision score of 95.0% and 80.65% after tuning. The model demonstrated a 95.50% and 78.69% accuracy on the test dataset, confirming its reliable and accurate predicting skills.

Convolutional Neural Network (CNN) Results

The model architecture is composed of three layers: an output layer that uses the sigmoid activation function, a hidden layer that has 64 units with ReLU activation, and an initial layer with 128 units using the ReLU activation function. In the process of compiling the model, binary cross-entropy loss and the Adam optimizer were used, and the accuracy metric was utilized.

Early halting was included as a preventative strategy in the training process to reduce the possibility of overfitting. This required setting the model's weights back to their ideal configuration and tracking the validation loss for a maximum of ten epochs. utilizing a batch size of 64, the training was carried out utilizing scaled training data for a maximum of 100 epochs.

Of particular interest is the model's performance on the test dataset. With remarkable scores of 97.46% and 87.50%, precision was attained. This implies that the model is quite likely to be accurate when it predicts a person will have cardiovascular disease. Moreover, 87.50% and 98.29% were the recall ratings. Resilience is demonstrated by the F1 Score, which is 97-

87% and 87.50%. The ratio of accurately predicted cases to total cases, or overall accuracy, is 97.50% and 86.89%, respectively.

Outcomes of XGBoost

GridSearchCV was employed to provide an extremely efficient hyper parameter tuning procedure. Through this method, the XGBoost model's ideal hyperparameters were found to be 0.2 for learning rate, 3 for maximum tree depth, 100 for boosting rounds (n_estimators), and 1.0 for subsample fraction. A noteworthy validation score of almost 98.00% on the Cardiovascular Disease Dataset and 84% on the Cardiovascular Disease Cleveland Dataset, respectively, supported the recall of these selected hyperparameters.

The optimized XGBoost model continued to perform exceptionally well on the test dataset, with precision scores of 99.14% and 90.00% indicating its ability to correctly classify positive cases. Furthermore, the recall score—which is between 98.29% and 84.38%—is especially important. Resilience is demonstrated by the F1 Score at 98.71% and 87.10%. On the test data, the model's overall accuracy ranges between 98.50% and 86.89%. These outstanding results highlight the suitability of the XGBoost model for the classification of cardiovascular diseases.

Discussion

The use of machine learning (ML) techniques for predictive modeling in the early diagnosis of cardiovascular diseases (CVD) has attracted a lot of attention recently because of the possible benefits to patient outcomes from early intervention and customized treatment plans. Several investigations have looked into using machine learning (ML) algorithms to different kinds of data, such as genetic data, imaging data, and electronic health records (EHRs), in order to create precise predictive models for identifying people who are at a high risk of developing cardiovascular disease (CVD).

We used the random forest in this research for predicting the likelihood of a patient having a certain condition, such as cardiovascular disease, based on various input features or risk factors. The performance of the Random Forest model is evaluated using metrics such as accuracy, precision, recall, and area under the curve (AUC). The result revealed that x = 88.7 for low risk and y = 11.2 for high risk, likely represent the probabilities or scores assigned by the model to indicate the likelihood of an individual female being at low or high risk for cardiovascular events, respectively. If the Random Forest model assigns a probability or score of 88.7% for low risk, it suggests that the female individual has a high likelihood of being at low risk for cardiovascular events according to the patterns it has learned from the data.

The finding is in conformity with one notable study by Dey et al. (2018) who employed a combination of clinical and genetic data to predict the risk of coronary artery disease (CAD) using a random forest algorithm. The study demonstrated that integrating genetic information with clinical risk factors significantly improved the predictive performance of the model compared to using clinical data alone. This finding underscores the importance of leveraging multiple data sources to enhance the accuracy of predictive models for CVD.

Similarly, another study by Attia et al. (2019) utilized EHR data to develop a deep learning model capable of predicting the onset of atrial fibrillation (AF) up to one year in advance. By analyzing longitudinal EHR data, including demographic information, comorbidities, medication history, and laboratory results, the model achieved high sensitivity and specificity in identifying individuals at risk of developing AF. This study highlights the potential of deep learning techniques to leverage rich, longitudinal data sources for early detection of CVD.

Moreover, research by Krittanawong et al. (2020) explored the use of ML algorithms in cardiac imaging data for early detection of heart failure (HF). By analyzing features extracted from echocardiograms and cardiac magnetic resonance imaging (MRI), the study developed predictive models capable of identifying patients at risk of developing HF. The findings suggest that incorporating imaging data into predictive models can provide valuable insights into cardiac structure and function, leading to more accurate risk stratification for CVD.

In addition to these individual studies, meta-analyses and systematic reviews have been conducted to synthesize findings from multiple studies and evaluate the overall performance of ML-based predictive models for CVD. For instance, a meta-analysis by Johnson et al. (2021) assessed the diagnostic accuracy of ML algorithms for predicting various CVD outcomes, including myocardial infarction, stroke, and heart failure. The analysis found that ML algorithms demonstrated superior performance compared to traditional risk assessment tools, such as the Framingham Risk Score, across multiple CVD endpoints.

Overall, the findings from these studies collectively underscore the potential of ML-based predictive modeling techniques in early detection of cardiovascular diseases. By leveraging diverse data sources and advanced analytical methods, these models have shown promise in improving risk stratification, facilitating early intervention, and ultimately reducing the burden of CVD on healthcare systems and society as a whole. However, further research is needed to validate these findings in diverse patient populations and healthcare settings, as well as to address challenges related to model interpretability, generalizability, and implementation into clinical practice.

The experimental results are thorough assessment of machine learning models, specifically the Random Forest and Guidient Boosting models, in the context of cardiovascular disease prediction, provides valuable insights. These insights align with the research conducted by Zhang *et al*. 2022, which underscores the effectiveness of the XGBoost algorithm in this specific domain.

Accuracy of Machine Learning Models on Both Datasets.

Across both datasets, these models consistently demonstrate exceptional performance, emphasizing their efficacy in cardiovascular disease prediction. Notably, the XGBoost model stands out with an impressive accuracy rate of 98.50% in the Cardiovascular Disease Dataset, while the K-Nearest Neighbors (KNN) model achieves a commendable accuracy of 91.80 % in the Cardiovascular Disease Cleveland Dataset. These high levels of accuracy emphasize the models' reliability, positioning them as valuable tools for diagnosing cardiovascular disease.

Precision, a critical metric in healthcare, reflects the models' ability to identify cardiovascular disease cases precisely. Both models achieve outstanding precision, with the XGBoost model leading at 99.14%, closely followed by the KNN model at 96.55%. These elevated precision levels significantly reduce the occurrence of false positive diagnoses, alleviating unnecessary concerns for patients.

Furthermore, the F1 Score, which balances precision and recall, highlights the XGBoost model's effectiveness in recognizing cardiovascular disease cases while minimizing the risk of overlooking positive instances. The model achieves F1 Scores of 98.71% and 91.80% in both datasets, showcasing its ability to strike this delicate balance effectively.

Comparison of the Accuracy Result

Random Forest Results

In our study, we meticulously tuned the hyper parameters of the Random Forest ensemble model, particularly focusing on adjusting the number of trees (n_estimators) to 200. This fine-tuning process resulted in a highly optimized model with outstanding performance metrics.
Accuracy: The tuned Random Forest model exhibited remarkable accuracy levels, consistently hovering at around 98.60% and 91.09% on different datasets.
Precision: Our assessment revealed a significant enhancement in precision, with scores reaching 98.63% and 94.44%. This underscores the model's proficiency in minimizing false positive predictions.
Recall: The Random Forest model demonstrated exceptional recall scores, measuring at 98.97% and 85.61%. This indicates the model's aptitude for recognizing genuine positive cases, a crucial aspect in medical applications such as cardiovascular disease detection.
F1 Score: The F1 Score, which harmonizes precision and recall, showcased the model's robustness, registering values of 98.80% and 89.81% respectively.
Based on the findings from the various machine learning models applied to cardiovascular disease classification, the following comparison of accuracy results can be made:
Random Forest (RF):
Accuracy: 98.60% and 91.09%
Logistic Regression (LR):
Accuracy: 95.50% and 88.52%
Gradient Boosting (GB):
Accuracy: 98.00% and 86.89%
Support Vector Machine (SVM):
Accuracy: Approximately 95.50% and 78.69%
Convolutional Neural Network (CNN):
Accuracy: 97.50% and 86.89%
XGBoost:
Accuracy: 98.50% and 86.89%

From the comparison:
Random Forest achieved the highest accuracy among all models, with an accuracy of 98.60% on one dataset and 91.09% on another.

Logistic Regression and XGBoost also demonstrated high accuracy, with LR at 95.50% and 88.52%, and XGBoost at 98.50% and 86.89%.
SVM exhibited a slightly lower accuracy compared to the other models.

Gradient Boosting and CNN performed comparably well, with accuracy scores around the 98% mark on one dataset and slightly lower on another.

Overall, Random Forest, Logistic Regression, and XGBoost emerged as the top-performing models in terms of accuracy for cardiovascular disease classification, with Random Forest being the most accurate among them.

These findings collectively highlight the Random Forest model's efficacy in accurately classifying instances of cardiovascular disease while maintaining a high level of precision and recall. The model's robust performance underscores its potential utility in clinical settings, offering valuable support to healthcare professionals in diagnosing cardiovascular conditions effectively.

Finding

Predictive modeling for early detection of cardiovascular diseases (CVD) using machine learning (ML) techniques has shown significant promise in improving patient outcomes through early intervention and personalized treatment strategies. Our research utilized the random forest algorithm to predict the likelihood of CVD based on various input features or risk factors. Evaluation metrics including accuracy, precision, recall, and area under the curve (AUC) were employed to assess the model's performance.

The results revealed that x = 88.7 for low risk and y = 11.2 for high risk, likely represent the probabilities or scores assigned by the model to indicate the likelihood of an individual female being at low or high risk for cardiovascular events, respectively. This finding is consistent with previous research demonstrating the efficacy of random forest algorithms in predicting cardiovascular risk, particularly when integrating genetic and clinical data.

Studies by Dey et al. (2018), Attia et al. (2019), and Krittanawong et al. (2020) have further validated the potential of ML techniques in early detection of CVD using diverse data sources, including electronic health records (EHRs), imaging data, and genetic information. These studies emphasize the importance of leveraging multiple data modalities to enhance the accuracy of predictive models for CVD.

Moreover, meta-analyses conducted by Johnson et al. (2021) have corroborated the superior performance of ML algorithms over traditional risk assessment tools in predicting various CVD outcomes. The exceptional accuracy, precision, and F1 scores achieved by ML models, particularly the XGBoost model, underscore their reliability and potential as valuable tools for diagnosing cardiovascular disease.

Overall, our findings, in conjunction with existing literature, highlight the effectiveness of ML-based predictive modeling techniques in early detection of CVD. By leveraging diverse data sources and advanced analytical methods, these models hold promise for improving risk stratification, enabling early intervention, and ultimately reducing the burden of CVD on healthcare systems and society as a whole. However, further research is warranted to validate these findings across diverse patient populations and healthcare settings, and to address challenges related to model interpretability, generalizability, and implementation into clinical practice.

Conclusion

The evaluation of various classification models for predicting cardiovascular diseases across two datasets highlights key considerations and recommendations. Models such as K-Nearest Neighbors (KNN) and Random Forest (RF) consistently demonstrated robust performance across different data environments, indicating their reliability in identifying cardiovascular disease cases with high accuracy, precision, recall, and F1-Score. These models are recommended for their ability to handle varying data complexities effectively, making them suitable for early detection and intervention strategies in clinical settings. Conversely, models like Support Vector Machine (SVM) showed sensitivity to dataset characteristics, particularly in Dataset 2, suggesting the need for careful model selection and validation processes. Overall, leveraging machine learning models like KNN and RF holds promise for enhancing cardiovascular disease diagnostics, provided that thorough validation and understanding of dataset nuances are prioritized to ensure optimal performance and clinical applicability.

References:

Goldstein, B. A., & Fink, J. C. (2019). Machine learning and health care disparities in kidney disease. *Advances in Chronic Kidney Disease, 26*(2), 76-81. doi:10.1053/j.ackd.2019.02.002

Huffman, M. D., et al. (2017). Global and regional patterns in cardiovascular mortality from 1990 to 2013. *Circulation, 135*(12), e146-e603. doi:10.1161/CIR.0000000000000485

Krittanawong, C., et al. (2016). Artificial intelligence in precision cardiovascular medicine. *Journal of the American College of Cardiology, 69*(21), 2657-2664. doi:10.1016/j.jacc.2017.02.065

Lubitz, S. A., et al. (2016). Novel method for cardiac risk prediction in rheumatoid arthritis. *Arthritis Care & Research, 68*(1), 8-14. doi:10.1002/acr.22649

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future — Big data, machine learning, and clinical medicine. *New England Journal of Medicine, 375*(13), 1216-1219. doi:10.1056/NEJMp1606181

Rajkomar, A., et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine, 1*(1), 18. doi:10.1038/s41746-018-0029-1

World Health Organization (WHO). (2020). Cardiovascular diseases (CVDs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)