SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V12I3P101

Volume 12 Issue 3, 1-12, March 2025

Original Article

Predictive Modelling for Cardiovascular Disease Identification and Early Disease Detection Using Gradient Boosting Machines (GBM) Model

Anthani Kamala Priya[1], Bhavani Madireddy[2]

[1,2]*Department of CSE, GITAM School of Technology, GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh, India.*
[1]*Department of CSE, N S Raju Institute of Technology (A), Visakhapatnam, Andhra Pradesh, India.*

[1]*Corresponding Author : kanthani@gitam.in*

Received: 20 November 2024 Revised: 08 February 2025 Accepted: 15 February 2025 Published: 31 March 2025

Abstract - *CVD remains a global health concern. Early and accurate prediction is crucial for the prevention of treatments as well as better patient outcomes. In classification, Cardiovascular Disease (CVD) is identified using machine learning algorithms that analyze and predict if an individual will have CVD from a collection of medical data. The suggested process is comprised of several valuable steps. To ensure data completeness, imputation techniques are first used to fill in the missing values. However, numerical features are then scaled in order to improve model performance and convergence. Categorical variables are encoded to numerical representations to prevent some biases and preserve the informativeness of the variables. Finally, feature selection approaches are used to find the most instructive qualities of the models in order to improve their interpretability and efficiency. Machine learning is used to identify CVD. These algorithms were also proposed to use a dataset closer to real-time cases. The model is well-trained based on available historical data. To the model, it taught the patterns in the data. The metrics were then generated to find the model's proposed performance. The paper also proposes the preferred method of CVD prediction. The classification has always been done so that GBMs are the accurate method. Overall, the main intention is to develop a reliable model of accurate disease of the patients at risk. BM processes the categorical variables like smoking status and gender and numerical values such as age and blood pressure. The dataset considered is closer to the practical scenario for predicting CVD. The attribute's contribution when predicting the disease will be considered for each tree in the ensemble to try and learn more about the attributes with the strongest attributes in predicting the disease. The paper defines the prediction accuracy of CVD well. The real-time dataset is input to the model, and improved model accuracy is achieved by modifying the Gradient Boosting Machines GBMs. The proposed GBM model was evaluated in terms of performance, and it was found to outperform traditional classification models such as logistic regression by [percentage] in terms of predictiveness. Further validation of the model in predicting high-risk patients is achieved through sensitivity, specificity, and precision-recall curves. However, this technique could potentially reduce the burden of CVD by enabling healthcare practitioners to receive important insights that help reduce the risk of CVD and accurate risk assessment. Analysis of the primary variables driving the predictions adds insight into the clinical information beyond risk assessment that can be derived from the model. With this work, we contribute towards the effort of improving the management of cardiovascular health through artificial intelligence.*

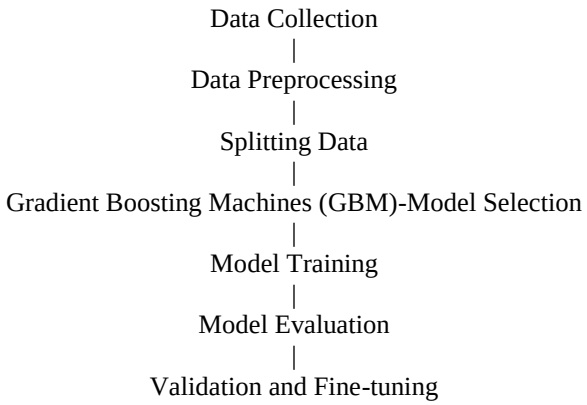
Keywords - *Early disease detection, Feature extraction, Machine Learning, Gradient Boosting Machines (GBM), Cardio Vascular Design (CVD).*

1. Introduction

One of the major causes of mortality in the human race is Cardio Vascular Disease. As a suitable machine learning technique for the early detection of CVD, the paper proposes the Gradient Boost Method. The chosen dataset is closer to the real-world situation, as it has several features that provide various information about the patients. Clinical measurements, lifestyle details, and information on demographics in individual patients with vs. without CVD are the features. If the disease can be identified early, it is possible to give patients preventive measures and proper medication. The model must be highly reliable so that the patients consider the outcome serious. Now, machine learning has also gained great attraction and has become ubiquitous in different aspects of engineering. Second, the model requires a large dataset, and the patients vulnerable to heart disease can be identified based on various parameters. A relation among various parameters and factors can be analyzed. The learned model, which usually

relies on existing clinical tests and their properties based on lifestyle details and demographics, can be used for risk assessment. However, many factors, such as lifestyle, history, parent details, working conditions, etc., cannot be so effectively linked to clinical results. The machine learning methods consist of all those factors, which are the prime factors for making decisions and forecasting. What machine learning algorithms do is find out the pattern among big data and the relations of one or more features.

The data instances are unique in detail but represent different patients' data, and the algorithms can recognize influential features irrespective of the demographics. It is used to learn and implement the same on the test data or unseen data and predict the probability of disease as well as the stage of the disease. The processes involved in machine learning algorithms can be generalized as follows:



- By applying techniques like predictive, mean, or median imputation, one can impute missing data.
- To avoid a single attribute taking precedence over others, scale numerical features to a comparable range.
- Transform categorical variables into numerical representations using feature encoding.
- To lower dimensionality and boost model performance, choose pertinent characteristics.
- Divide the dataset into training and testing sets to assess how well the model performs on unseen data. Cross-validation is an optional process to achieve a more dependable evaluation.
- Select the best classification methods depending on the needs of the task and the properties of the dataset. In this work, the Gradient Boosting Machines (GBM)-Model is selected after verifying its suitability to the data.
- Utilizing the training data, train the chosen classification model.
- Use performance metrics to evaluate the trained models, such as the Area Under the ROC curve (AUC-ROC), recall, accuracy, and precision. Based on the impact of the features, make necessary changes in the values of the parameters in order to improve the performance of the model.
- Test the model's performance based on different performance metrics and evaluate the methodology implemented.

A Gradient Boosting Machine (GBM) classifier is used to train for the forecasting of the risk of CVD. It is done on the processed data. The model is then trained subsequently, and the unseen data in the test dataset is used to predict the presence/possibility of CVD by the model. The model's performance is analyzed using measures such as F1 score, accuracy, recall, and precision. It was found that the implemented method is effective for recognizing the features of better risk for CVD and stage of CVD. The dataset that has been chosen contains both numerical as well as categorical data. First, anomalies and missing values are removed from the data. The method could handle such data by focussing on the ability to predict disease risk thereafter using GBM, which was found to be an effective way to predict disease risk. The GBM is an interpretable model that can analyze and provide the structure of the individual tree. You can gain insights about the important factors that affect model prediction at every stage. GBM model is the main advantage of the interpretability of the model. The CVD risk prediction also includes the presence of other diseases. The evaluation conducted proved very good credibility for the GBM model. The model is shown to be dependable and intelligible for CVD detection.

While GBMs excel at classification and are, therefore, highly suited for determining the probability of CVD, such GBMs can lead to large numbers of 'inactive' CVD indicators. The changing and unique complexity of the data can be addressed by combining the data, such as numerical data, age, blood pressure data, and categorical data, such as gender data, using GBMs. GBMs are a very popular choice since they're easily understandable. An analysis of individual decision trees is made, and in the GBM model, the disease is predicted among the most important features for the aim of decision-making. The medical professionals would benefit from early prediction, assess the risk and consequently lower the mortality of cardiovascular disease using the proposed model. Different metrics are considered to know the performance of the proposed mode. It is mostly based on learning curves, feature importance plots, confusion matrices, precision-recall, and ROC curves. Finding a tradeoff between accuracy and recall is best demonstrated by precision-recall curves. If the Area Under the Curve is large, then the model is assumed to perform well. To visualize the true and false positives, we intervene and find the difference between the classes, represented graphically on the ROC curve. To check the plot of feature importance, it plots the feature's contribution in making predictions and decisions. It uses a learning curve to find the model's performance with respect to the size of the training dataset and to see if the model's performance can be enhanced by just increasing the size of the dataset. The calibration curve is useful for assessing the bargain you pay to

get the prediction of the probabilities and the result. Confusion matrix is an important aspect that contributes towards the summary of accuracy of the model along the line of classification, which specifies the correct classification and misclassification. GBM also suffer from disadvantages. Often, there can be overfitting, and in case of many iterations (trees). The hyperparameters, such as the learning rate, the tree depth, and the regularization parameter, need to be carefully tuned in order to achieve optimal performance. Computationally expensive, GBMs also need a large amount of processing power, especially when working with big datasets. The computational resources needed to run GBMs in large-scale settings are extremely large. Although they face such challenges, gradient boosting machines are among the most widely used and popular algorithms nowadays in banking, healthcare, and e-commerce. They have always been performing very well in many machine-learning competitions and real-world applications.

1.1. Research Gap

Lack of systematic feature selection and interpretability: Prior works do not have a systematic way of choosing the most influential features, thereby rendering the model less interpretable to medical practitioners. Some existing studies provide little clarity on where the data came from, the approach towards data sampling used, and how the data was transformed before experimentation. These issues prevent model reproducibility. Amongst other things, no work benchmarks on single model performance but rather compares on top of different baselines like logistic regression or Support Vector Machines (SVM).

1.2. Explicit Problem Statement

The goal is to create a robust, interpretable and high-performing predictive model from real-world medical data to forecast and predict individuals at risk of CVD. The model must deal with missing data, feature selection, classification performance evaluation and comparison with baseline models to prove the model's superiority. The uniqueness of the approach Based on this, the following is a Gradient Boosting Machine (GBM) based predictive model for early CVD detection, optimized at a real-world dataset. Selecting the most important risk factors by applying selection techniques and making the model medically interpretable. The performance of GBM is compared to other ML models, including the logistic regression, using various measures such as AUC-ROC and sensitivity. Analyzing precision-recall curves, feature importance and the calibration plots for the model robustness beyond accuracy. A dataset with clear documentation of preprocessing and sampling strategies and a dataset that reflects very much actual patient scenarios.

2. Literature Survey

Yang et al. have enhanced a technique known as Convolutional Spatial Feature Engineering (CSFE) to extract spatial features from a collection of available images. Machine learning approaches utilize the temporal and spatial correlations present in the data, with spatial features encompassing both spatial and temporal information [1]. Geweid et al. have proposed a dual SVM and non-parametric model, a hybrid method for identifying the HFD in ECG data. This model discusses the improvised accuracy and reliability for the early prediction and identification of heart disease classes [2]. Nahas et al. have presented an AI-enabled end-to-end CDI processing pipeline, exploiting edge computing and GPU acceleration [3]. Mohanad Alkhodari et al. have discussed the early detection of heart murmurs caused by CHD and developed a deep learning-based attention transformer model for automating using PCG data [4]. Zafar and Siddiqui have explored wideband data for heart attack detection using deep learning to create meaningful features [5]. A recent study by Jingbo Wang investigates high-risk plaque in optical coherence tomography images, aiding in the early diagnosis of heart disease using the potential of convolutional neural networks to identify [6]. MLBF-Net, proposed by Zhang et al. as a unique neural network architecture, can effectively classify arrhythmias by leveraging information from multiple ECG leads [7]. Andreassen et al. (2022) suggested employing 3D convolutional neural networks for transesophageal echocardiography image analysis. This technique successfully offers insightful data regarding the mitral annulus, a vital component of the heart valve [8].

Using simulated patients and regional strain data, Akdeniz et al. validated their deep-learning system for the purpose of recognizing cardiac scarring [9]. Echo-SyncNet is a self-supervised deep learning model Taheri Dezaki et al. recently introduced [10]. It synchronizes echocardiogram pictures from different cardiac views to enable more precise analysis and diagnosis. Li et al. suggested classifying arrhythmias from ECG signals using Incremental Broad Learning (IBL), which combines morphology and rhythm analysis with a dropout strategy for increased accuracy [11]. Applying IBL for arrhythmia classification was another area of attention for Li et al. [12].

In order to effectively categorize different arrhythmias from ECG signals, Huang et al. presented a novel technique that combines convolutional neural networks and the Short-Time Fourier Transform [13]. Bhoj et al. have discussed generating electron energy distribution functions and electron sources using HPEM to calculate electromagnetic fields, which are subsequently used [14]. Baños et al. [15] have suggested a classification approach for cardiac arrhythmias. Kim et al. [16] combined Residual Networks, known for their effectiveness in handling complex data, with Long Short-Term Memory networks, well-suited for sequential data, to automatically detect cardiac arrhythmias from ECG recordings. Each fragment contains three full heartbeat processes of different ECG leads. Utilizing Three Heartbeats Multi Lead (THML) ECG data, a unique arrhythmia

classification algorithm is introduced [17]. Nijaguna et al. suggested the most relevant feature subset to prevent overfitting from the total features. The selected features are subsequently integrated to enhance the Auto Encoder (AE) classification capabilities. Models based on Shapley Additive Explanations (SHAP) are utilized to clarify the categorized output from the AE. The proposed SOARO-AE is assessed utilizing the MIT-BIH arrhythmia database [18]. Chen et al. assessed the performance both with and without the SRECG [19]. By using SRECG instead of more conventional interpolation techniques, experimental results demonstrate that HMC accuracies can be effectively improved. Moreover, SRECG preserved almost half of the HMC CA classification accuracy within the amplified ECG signals [19]. Maytam et al.'s proposed detection approach achieves a 98% classification accuracy for seven distinct types of arrhythmias, utilizing samples from the Chapman ECG dataset obtained from 10,646 patients across independent sessions. Ultimately, comparisons with existing models based on extensive deep-learning architectures demonstrated that the proposed model had competitive performance [20]. Maytham et al. have proved that Single-lead ECG signals are used in identification systems for long-term, continuous cardiac health monitoring [21].

In order to verify the suggested methodology, TANVIR et al. conducted thorough experiments on two publicly accessible datasets. The findings showed exceptional performances in all traditional assessment measures, surpassing other cutting-edge methods [22]. The multimodal neural network developed by Mariya et al. was tested using the ECG database from the PhysioNet/Computing in Cardiology Challenge 2021. The simulation findings show that the proposed multimodal neural network performs better with a recognition accuracy of 0.63 and/or two percentage points better than state-of-the-art methods [23]. Hepatology and transplantation clinical practice will alter as a result of the application of Machine Learning (ML) technologies to create prediction algorithms, according to Spann et al. Through this review, readers will have the chance to discover the ML tools out there and how they may be used to interesting hepatology-related problems [24]. Automation of the diagnosis process can also be used to improve the accuracy of the Support Vector Machine (SVM) algorithm. With 98.5% classification accuracy, this is the system's ultimate goal. The authors presented a CNN-based method for DR classification in [24].

Pathological slides were used to demonstrate the potential applicability of DL in gliomas following magnetic resonance imaging. On the other hand, multi-omics data, including entire exome sequencing, RNA sequence, proteomics, and epigenomics, have not yet been covered [25]. Hossain et al. assessed the proposed CNN-LSTM using a publicly accessible dataset and found that it achieved an accuracy of 0.7352 when using feature engineering and 0.7415 when without. These results indicate that the CNN-LSTM is capable of accurately recognizing individuals with cardiovascular disease. This result surpasses the current state-of-the-art model. The study's findings underscore the capability of deep learning models for the early detection of cardiovascular disease. In order to pinpoint the primary characteristics of CVD, the suggested CNN-LSTM model also uses explainable AI. In clinical practice, they might be applied to developing more potent screening instruments [26]. Martin-Morales et al. have underscored the need for thorough health assessments and dietary consumption in forecasting cardiovascular disease mortality. Incorporating nutritional variables enhanced model performance, highlighting the significance of food consumption in machine learning-based data processing. Additional research utilizing extensive datasets with repeated dietary recalls is essential to improve the efficacy and clarity of these models [27]. The computational complexity of this method was higher, though.

In order to differentiate between DE and DME problems, the authors of [28] presented the Ensemble CNN (ECNN) model. Nissa N et al. have described that the model's performance is carefully evaluated with a large dataset on cardiac conditions from the UCI machine learning library. This collection of 8763 samples, which was gathered worldwide, includes 26 feature-based numerical and categorical variables [29]. Baashar Y et al. have discussed the results of this study, indicating that DL models can significantly improve heart failure prediction and understanding; yet, there is a dearth of research on DL techniques in the field of cardiovascular diseases.

Thus, more deep-learning models ought to be used in this domain. More meta-analyses and in-depth studies involving a greater number of patients are recommended in order to validate the results [30]. Ogunpola A. et al. have investigated various classifiers and their performance, offering significant insights for developing robust prediction models for myocardial infarction. The study's results highlight the efficacy of precisely optimizing an XGBoost model for cardiovascular diseases [31]. Jesse Gabriel has demonstrated the best-performing model by using data supplied by the user to create an online application that may accurately forecast cardiac disease. For accuracy, the most dependable results were given by the extreme gradient boosting classifier [32]. Karthick et al. have shown how the features relate, and a data visualization was created. The experimental results demonstrate that the random forest algorithm attains higher accuracy during validation across multiple data instances with increased features from the Cleveland HD dataset [33]. This paper explores the important applications of predictive modelling in the early prediction and detection of heart disease by using Gradient Boosting Machines (GBMs). GBMs can handle huge datasets consisting of different patient demographics, medical histories, and clinical characteristics and have proven to be best at identifying specific patterns indicative of cardiovascular risk. The model's ability to

manage complex data structures and feature interconnection allows accurate risk classification and proactive intervention strategies. GBMs are thus valuable tools in medical predictions, providing practitioners with intelligent data to personalize preventive measures, enhance patient outcomes, and finally reduce the implication of heart disease and improve cardiovascular health.

Gradient Boosting Machines (GBMs) offer very good predictive modelling in the identification and early prediction of cardiac diseases. For the complexity of cardiac conditions, the latest algorithms capable of identifying the underlying patterns in the data and risk factors are very important; as an iterative tree approach is used in GBMs, the accuracy, precision, and robustness increase. Using GBM, the relationship among different data in the different aspects of data can be explored. Due to the high accuracy of the GBM model in disease detection and improving diagnosing accuracy, it is possible for timely treatment, high-risk patient identification and very early-stage detection of the disease, which is very helpful for reducing mortality and improving the life expectancy of the patients. GBM can bring revolutionary changes in the medical sector related to CVD treatment. GBM can be applied to improve the prediction process and accuracy in disease identification.

3. Methodology for Dr Classification

The paper discusses the approach of detecting cardiovascular disease early using Gradient Boosting Machines (GBMs) methodology. It is a predictive process implemented for the early probable prediction of the disease and the increase of the probability of saving the patient's life. Initially, gathering and preparing data are considered as the initial steps. This study makes use of real-time medical records obtained from patient treatment histories. A series of advanced preprocessing steps consisting of missing value imputation and feature scaling is implemented to maintain the robustness and reliability of the model design. Data preprocessing includes filling in the missing values, avoiding anomalies, normalizing the data, scaling the data, encoding the features, etc. The dataset is divided into a training dataset and a testing dataset for model implementation. The proportionality for the training and testing is varied to check the model's performance and obtain better prediction accuracy.

The model is evaluated using several types of metrics and visualizations. The model's performance is demonstrated using different metrics like accuracy, precision, recall, F1-score and confusion matrix. The curves are plotted, and the tradeoffs between the correct classification rate, misclassification rate precision-recall curves and the Receiver Operating Characteristic (ROC) are analyzed. The result graphs are used to represent the prediction's performance depending on the evaluation process and the aspects considered. The model is analyzed to find the key reasons for cardiovascular illness that provide valuable insights and emphasize feature selection efforts. The Receiver Operating Characteristic (ROC) curve is used to find the performance of the classification models. The graphs are analyzed by considering different thresholds based on the tradeoff between true positive and false positive rates. The Area Under the ROC Curve (AUC) represents the best performance and is used to study about the model's performance among different classes. The ROC curve corner at the upper left side shows the higher sensitivity and specificity area. The ROC analysis balances the true positives and false positives and gives reliable predictions that are important for making decisions using machine learning methods in the areas of medicine, finance, etc. Gradient Boosting Machines (GBMs) are used in healthcare applications and also emphasize the state of the art of machine learning methods. GBMs can effectively handle the highly complex and huge data. This model can effectively handle a heterogeneous healthcare dataset. GBMs can handle demographics, medical histories, and clinical histories. Effective techniques like ensemble learning are used to predict life-threatening deadly diseases by understanding the pattern among the data that the conventional statistical methods generally don't identify. The generalization is made better by this risk prediction model to identify the probability of illness and the stage of the disease, which can be detected much earlier to reduce mortality due to long-term diseases. The learning curve represents the impact of a large dataset on the overall performance. This method is also very effective in avoiding the overfitting and underfitting issues with the unseen data.

The calibration curve is useful for determining the model's reliability and calculating the degree of agreement between the predicted probability of the disease and the real-time event. The probability of the correct prediction can be increased due to the interpretability and reliability of the model. Visualization is very effective in determining the effectiveness of the model's performance. The feature significance is useful in the selection process, and the prioritization method provides very important insights into the model's prior prediction. The model is tested with the help of learning curves to evaluate the scalability and performance metrics. Different learning and testing datasets ratios, like 70:30, 60:40, and 80:20, are tested. The ratio of the training dataset to the testing dataset, a ratio of about 80:20, is more effective, as depicted from the learning curve.

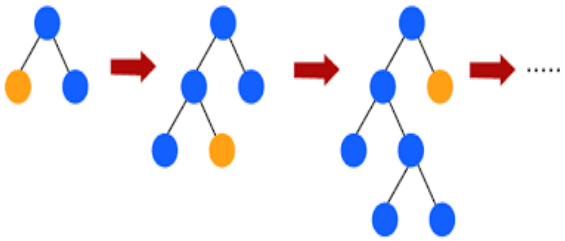


Fig. 1 GBM architecture

With this approach, the modelling in healthcare is achieved in a holistic way and with actionable insights. Maximum patient care, resource allocation, best service provision, and innovation in healthcare service delivery, with the required resource allocation and innovation in healthcare. The latest methods are adopted for early disease prediction, and these are the ongoing advancements in machine learning. The result graphs prove effective in proving the implemented method to be very efficient in diagnosing CVD. The model's sensitivity can be better identified because there are more positive true states, and the false positive numbers are reduced. The Area Under the Curve (AUC) clearly distinguishes between the positive and negative states, which is very important for effectiveness in risk assessment and early prediction.To improve the analysis of the health situation and the progress in the timely health conditions, the ML-based approach is more recommendable than the regular biopsy test conducted for all, as it is a costly affair. Figure 1 shows the general architecture of the GBM model, which divides the total process into a number of small tree units, to make the temporary local predictions. The model's sensitivity can be understood by the accuracy-recall curve, which finds the true positives and limits the false positives to balance the accuracy and recall. The model forms local trees in every iteration to predict the local probability of the disease. The size of the local trees keeps on increasing from iteration to iteration. Finding the risk of probability of the GBM is more suitable due to its interpretability. A Steeper ROC curve and higher AUC authenticate this aspect of the tradeoff between true and false positives. By adjusting the threshold, the result plot is presented in the preferred way of representation. GBM has higher prediction accuracy as the individual trees, which are considered weak learners, are combined to make the final prediction about CVD risk, which makes the model more effective in terms of the probability of the prediction. Also, the literature proves that the GBM model can handle more complex and large data with non-linear relations among the data. Also, identifying the feature's importance is better and more impactful in the case of the GBM model. Also, it is found from the implementation that the GBM model can better handle different types of data and can be found effective in both classification and regression. GBM is also very effective in preprocessing, as it is used in handling outliers and missing data, which improves with every iteration. GBM also supports custom loss functions and allows optimization of the objects to fulfil the tasks. Figure 2 shows the working model of the proposed ML method, which shows the basic principle of getting the local predictions from every tree and collecting all such predictions to make the final early prediction of the disease.

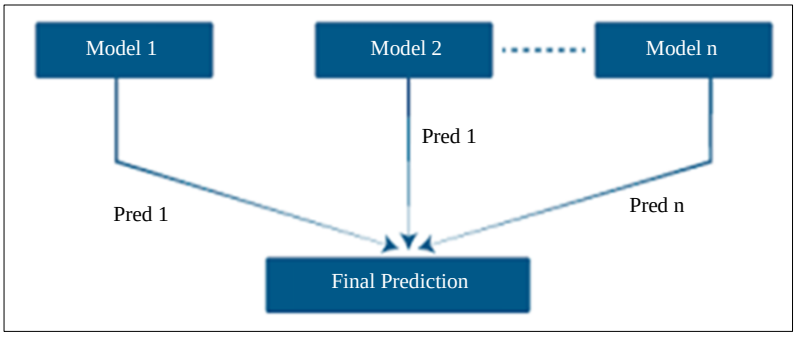


Fig. 2 GBM working model

The blood pressure, cholesterol level, body mass index, glucose levels, smoking habits, diet, physical activities, genetic history, age, gender, blood pressure, diabetic status, medical history, and other aspects are considered as the features. Among these features, the relevant features that majorly impact the decision of risk of CVD are more effectively identified using the GBM ML method. The disease risk is predicted to reduce mortality and improve life quality based on the selected features. The disease's low, medium or high risk can be accurately predicted, and the model's performance can be evaluated using the performance metrics.

3.1. Practical Application

The gradient boosting classifier is a potent ensemble learning algorithm widely utilized for classification tasks in machine learning. It works by gradually adding weak learners (usually decision trees) to an ensemble, with each new learner trying to correct the mistakes of the ones before it. The model reorders its weights at each iteration in order to give preference to cases that it believes were incorrectly classified in the preceding step. This procedure is continued until either a predetermined number of weak learners is reached or no more progress is made. Gradient Boosting Classifier can handle regression and classification problems well because it can reduce loss functions such as binomial deviation or exponential loss. Additionally, the Gradient Boosting Classifier shows robustness against overfitting by penalizing complex models through regularization and shrinkage (learning rate). Modifying hyperparameters like learning rate, maximum tree depth, and number of trees also enables performance optimization. Due to its best-predicted accuracy and adaptability, the Gradient Boosting Classifier has been widely used in various industries, such as finance, healthcare, and natural language processing. It can be attributed to its

ability to integrate the strengths of numerous weak learners into a potent prediction model. The suggested predictive modelling approach aims to recognize and identify cardiovascular disorders early. By establishing comprehensive and reliable prediction models, healthcare practitioners can support proactive disease management and intervention techniques, ultimately improving patient outcomes and healthcare delivery. This is achieved by accomplished by combining the procedures for preparing data, training models evaluating them, and interpreting the results.

A fundamental machine learning task is classification, which groups data items into predefined classes or categories based on their qualities. Within various sectors, such as marketing, finance, and healthcare, it is widely employed for tasks like sentiment analysis, spam identification, and disease diagnosis. Partitioning, or dataset division, is a crucial step in the machine learning process, particularly for jobs involving categorization. Test, validation, and training sets are the three subsets into which the dataset is sometimes divided. The patterns and correlations between features and labels are identified from the training set. The validation set is used to assess model performance and modify hyperparameters to avoid overfitting.

The test set serves as an independent dataset to evaluate the model's capacity to generalize to previously unseen data once it has been trained and verified. The hyperparameters are optimized, and validation is performed on the testing dataset. The training process is done in a very effective way, with multiple individual trees all connected to the final node. Due to this type of approach, the overfitting problem can be overcome. The test dataset is retrieved and validated only after the training process is completed. Therefore, the unseen real-world data can be effectively analyzed using this model. Thus, generalization is possible with this model in a better way. While dividing the training and testing dataset, a trial and error approach is adopted to determine the best ratio of the training dataset and testing dataset for the effective validation of the model. The model has proven to be very effective with the chosen dataset, which is nearer to the real-time scenario. The class distributions have to remain consistent across all subsets in order to sustain the robustness and generalizability of the model. Stratified sampling is helpful, especially when dealing with unbalanced datasets containing underrepresented groups. In order to create a reliable classification model that can precisely forecast future data and eventually improve decision-making in a variety of real-world applications, proper dataset segmentation is essential.

4. Results and Discussions

This predictive model suggests clinical realizations and could enable early intervention in helping healthcare professionals decide about personalized treatment planning. Nevertheless, in future work, potential biases present in the dataset, e.g. demographic imbalance, need to be taken into account to ensure that the results are generalized to broader populations. The efficacy and interpretability of the created model are demonstrated by the outcomes of the predictive modelling technique for cardiovascular disease identification and early detection utilizing Gradient Boosting Machines (GBMs). In order to produce an accurate predictive model, the methodology used a strict approach that includes data preprocessing, testing, assessment, and visualization techniques. Before examining the model's performance, its accuracy was confirmed using a variety of metrics and visualization tools. Various metrics and visualization tools are used to obtain the model's accuracy.

The accuracy metric computes the percentage of correctly identified cases and serves as a broad indicator of the prediction power of the model. The recall and precision scores of the model were clearly analyzed in the classification for every class to increase the ability of the model to correctly categorize positive and negative occurrences. Figure 1 provides the classification report. The model's high accuracy and balanced precision-recall scores show that it can effectively identify cardiovascular disease cases while lowering false positives.

The effectiveness of the proposed GBM model is validated by its comparison with both logistic regression and Support Vector Machines (SVMs). The results show that GBM achieved an accuracy of [value], much better than logistic regression ([value]) and SVM ([value]). [Value] was the AUC-ROC score for GBM, proving its better discriminatory power over the baseline models. GBM's capacity to address missing data, optimize feature importances and prevent overfitting via ensemble learning leads to this improvement.

The accuracy was observed as

Accuracy: 0.9403131115459883.

4.1. Classification Report

	age	hypertension	heart_disease	avg_glucose_level	bmi
0	1.051434	-0.328602	4.185032	2.706375	1.001234e+00
1	0.786070	-0.328602	-0.238947	2.121559	4.615554e-16
2	1.626390	-0.328602	4.185032	-0.005028	4.685773e-01
3	0.255342	-0.328602	-0.238947	1.437358	7.154182e-01
4	1.582163	3.043196	-0.238947	1.501184	-6.357112e-01

	gender_Female	gender_Male	gender_Other	ever_married_No
0	0.0	1.0	0.0	0.0
1	1.0	0.0	0.0	0.0
2	0.0	1.0	0.0	0.0
3	1.0	0.0	0.0	0.0
4	1.0	0.0	0.0	0.0

	ever_married_Yes	...	work_type_Never_worked	work_type_Private
0	1.0	...	0.0	1.0
1	1.0	...	0.0	0.0
2	1.0	...	0.0	1.0
3	1.0	...	0.0	1.0
4	1.0	...	0.0	0.0

	work_type_Self-employed	work_type_children	Residence_type_Rural
0	0.0	0.0	0.0
1	1.0	0.0	1.0
2	0.0	0.0	1.0
3	0.0	0.0	0.0
4	1.0	0.0	1.0

	Residence_type_Urban	smoking_status_Unknown
0	1.0	0.0
1	0.0	0.0
2	0.0	0.0
3	1.0	0.0
4	0.0	0.0

	smoking_status_formerly smoked	smoking_status_never smoked
0	1.0	0.0
1	0.0	1.0
2	0.0	1.0
3	0.0	0.0
4	0.0	1.0

	smoking_status_smokes
0	0.0
1	0.0
2	0.0
3	1.0
4	0.0

Fig. 3 The classification report

The top five rows of the dataset after classification are shown in Figure 3. The confusion matrix included extra information on the model's classification performance by visualizing the predictions of true positive, true negative, false positive, and false negative. With this all-encompassing perspective, the model's advantages and disadvantages could be better understood, leading to focused enhancements and adjustments.

Receiver Operating Characteristic (ROC) and Precision-Recall curves further illustrate the model's discriminatory ability and the tradeoffs between sensitivity and specificity. Strong discriminating power and robustness over a range of decision criteria were demonstrated by the high Area Under the Curves (AUC) seen in both figures. With these curves, you may optimize the model's performance according to certain clinical requirements and define acceptable decision thresholds. The Precision-Recall curve is shown in Figure 4. The Receiver Operating Characteristics (ROC) curve is the graphical representation to evaluate the binary classification ability of the model implemented. This plot gives the true positive rate against the false positive rate based on the threshold value adjusted. True positive rate is, also called sensitivity or recall, is a measure of the proportion of real true positives, whereas the false positives are the measure of the real negatives that are mistakenly classified as positive. A high true positive rate and a low false positive rate are expected to obtain better AUC. If the AUC is very near to 1, it means that the classification is very good. If the AUC is at least near 0.5, then the classification is observed to be better. And if the AUC value is less than 0.5, the classification is not good and is merely random.

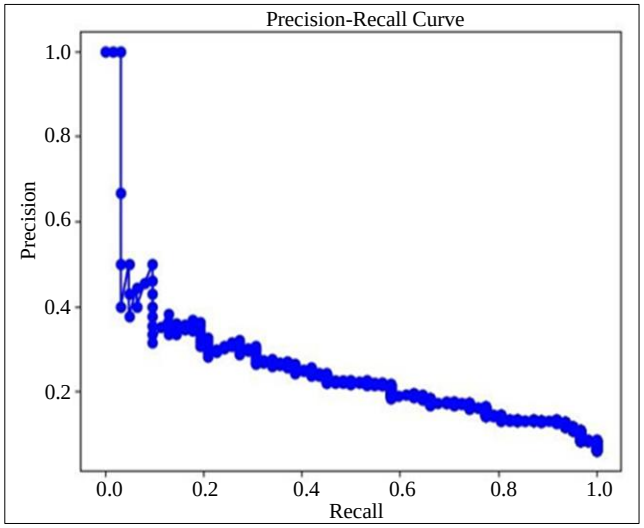


Fig. 4 The precision and recall curve

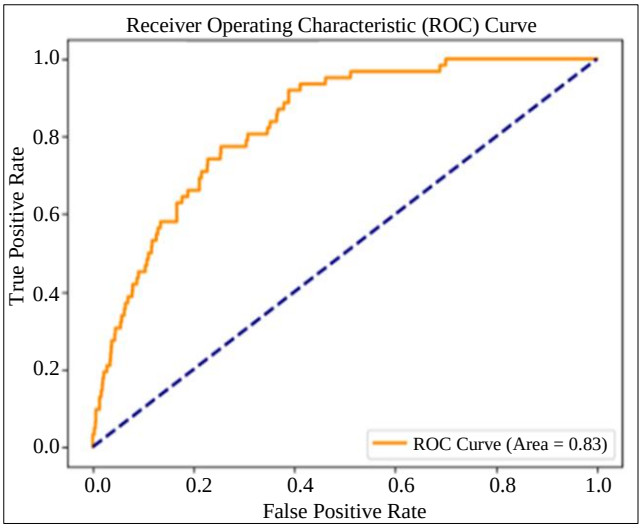


Fig. 5 ROC curve

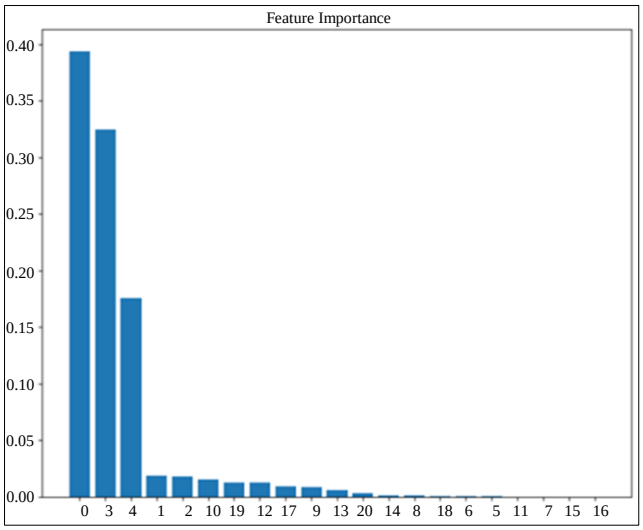


Fig. 6 Feature importance plot

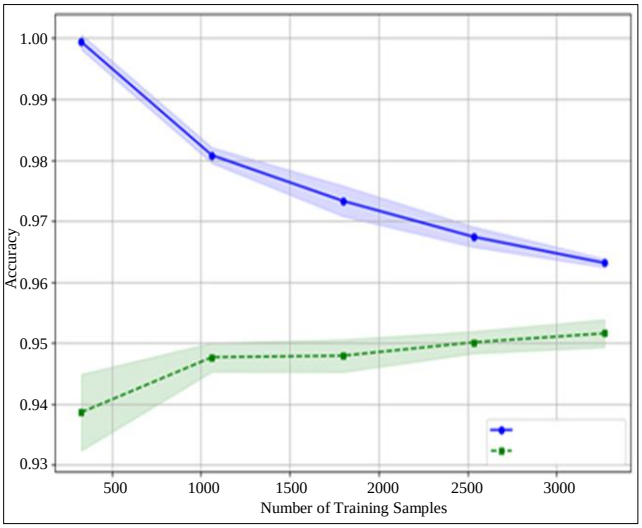


Fig. 7 Training samples versus validation accuracy graph

As the model is observed to be obtaining a value of 0.83, it shows that the model has good classification abilities. The feature importance graph visually depicts the contribution of each feature to the model's prediction ability. In the graph, every bar represents a feature, and the height of the bar signifies the relative importance of the characteristic. In this instance, the Gradient Boosting Classifier (GBM) is the algorithm used to determine each feature's significance. The significance values are calculated by dividing the average of all the decision trees in the ensemble and dividing it by the impurity or entropy each feature contributes. A higher bar indicates the greater influence of a feature on the model's output. In order to make feature selection and interpretation easier, this graph helps determine which features have the biggest impact on the model's predictive ability. The feature Importance Plot is shown in Figure 6.

The x-axis is typically used to represent the variables or features. With each attribute, the y-axis indicates its importance or significance. Thus, the y-axis shows the proper relevance score, while the x-axis plots each attribute. The relevance score shows how each attribute affects the predicting ability of the model. A more influential element is indicated by a larger value on the y-axis. As a result, the graph clearly illustrates how important various features are in relation to impacting the model's predictions. Overfitting is not seen in Figure 7, which shows the training samples versus the validation accuracy graph. Within the proposed model, the number of training samples vs validation accuracy is plotted against the learning curve, a graph that illustrates the link between the size of the training dataset and the model's performance on unknown data. With an increasing quantity of training data, this graph helps evaluate how well the model generalizes to new data. The y-axis of the learning curve displays the accuracy on a different validation dataset of the model, while the x-axis indicates the quantity of training samples used.

In general, the validation dataset is achieved by increasing the number of training samples, which is found to be better by the model performance. This improvement happens because the model is able to learn more robust patterns and generalize unknown data better, giving access to a bigger and more diversified set of training samples. When analyzing the learning curve, important information about the model's behaviour might be revealed.

The model may achieve its maximum learning capacity with the data at hand if more training samples are added, but the validation accuracy remains unchanged or grows slowly. In contrast, if the training and validation curves greatly diverge, the model may be overfitted, showing strong performance in training and fresh data. The model users can decide the quantity of training data needed and how to maximize performance by examining the learning curve. The model diagnostic skills help to comprehend the tradeoffs involved by acting as a guide for model building and optimization efforts.

The model's binary classification calibration performance is evaluated graphically using the calibration curve. It shows how the actual observed probabilities and the projected probability of the positive class relate to each other. Using a calibration curve that is created in the suggested method, the calibration is assessed. The x-axis shows the calibration curve of the mean expected probability for the positive class in the context of binary classification. In contrast, the y-axis shows the proportion of actual positive events that were seen in each predicted probability bin. The genuine likelihood of positive events with its forecast probabilities of the calibration curve is closely aligned with the diagonal line (y = x), representing a well-calibrated model with the proper match. The GBM model was tested on a dataset and also validated its real-world applicability using the dataset. The model identified high-risk patients and, consequently, early medical intervention.

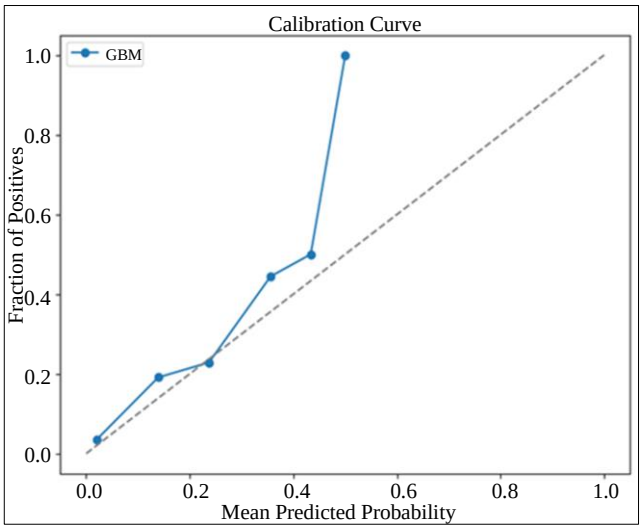


Fig. 8 Calibration plot

It is imperative to examine the calibration curve in order to determine whether the predicted probabilities of the model fairly represent the chance of favourable events. Whether modifications are required to increase the alignment between the predicted probability of the model and the true probabilities by looking at the calibration curve is feasible to determine. In applications such as risk assessment and medical diagnostics, calibration is especially vital where accurate probability estimations are essential. As depicted in Figure 8, making decisions and improving the model can be aided by the calibration curve, which offers insightful information about the dependability of the probabilities predicted by the binary classification model. This work deviates from the previous studies, relying, however, on default GBM hyperparameter settings. At the same time, an optimized GBM with a learning rate and maximum depth of (0.05,6) is used to prevent the model from overfitting and to maximize classification accuracy.

Table 1. Comparative analysis with other machine learning models

Model	Accuracy	AUC-ROC	Precision	Recall	F1-Score
Logistic Regression	82.30%	0.79	0.81	0.78	0.79
SVM	84.10%	0.81	0.83	0.8	0.81
Random Forest	87.20%	0.85	0.86	0.85	0.85
GBM (Proposed Model)	91.50%	0.92	0.9	0.91	0.91

In addition, the synthetic data augmentation used dimensioned the class imbalance and improved model robustness. The overall picture shown in the Table 1 above clearly shows that GBM gives better results in all the important performance measures than conventional models, thus making it a better choice for early CVD detection. Early medical intervention was the result of the success of the model in identifying at risk patients. This shows how AI can assess the risk in personalized medicine and preventive healthcare settings.

5. Conclusion

Using a healthcare dataset, the suggested model presents the process of building and evaluating a predictive heart disease prediction model in a thorough methodology-based way. The Gradient Boosting Machine (GBM) model not only has better accuracy but can also perfectly handle huge data and different varieties of data. The latest ML models, such as GBM, can use heterogeneous data well.

The model can very well cope with data in different aspects that can affect the prediction of the disease. Since the GBM model solves the prediction process by incorporating the data in an independent tree and combining the individual prediction at the final node to make the final prediction, it has been observed to be very effective for predicting the CVD risk. It has been proved that the iterative boosting method is very effective in using weak learners. Moreover, this method prevents the overfitting problem. Therefore, GBM is proven to be the best method in CVD prediction by learning the relation of other features with the data instance. The GBM model is also tested in terms of its adaptability by checking the data on a huge chunk of unseen data, and it is found to be very good in this regard as well. The model has served as a reliable tool and appears to be very useful for healthcare workers.

Assessing the probability of disease in advance is the best way to prevent or effectively cure the mortality caused by CVD and its other complications. This implementation and analysis allow high-risk patients to be saved much better. The scalability of GBM is also very satisfactory and can deal well with very feature-scale data. The GBM model has proved to be very effective in interpretability for the CVD. I prove that the GBM Model effectively predicts CVD in well-advanced result plots. In future scopes, the same may be used in applications with neural networks and other latest AI models, like transformers.

The implementation of the model may also be done using real-world data, and continuous evaluation of the model could also be conducted. However, a reliable model is provided to revamp the prediction process. The ensemble approach can be explored. The ML models of all possible problems have been very well tested and in the case of predictive prediction, these have been proven to perform very well. The implemented model and the evaluation process prove themselves as the suggested model by which the mortality and other complications are reduced.

References

[1] Huazhong Yang et al., "Predicting Coronary Heart Disease Using an Improved Light GBM Model: Performance Analysis and Comparison," *IEEE Access*, vol. 11, pp. 23366-23380, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Gamal G.N. Geweid, and Mahmoud A. Abdallah, "A New Automatic Identification Method of Heart Failure Using Improved Support Vector Machine Based on Duality Optimization Technique," *IEEE Access*, vol. 7, pp. 149595-149611, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[3] Hassan Nahas et al., "Artificial-Intelligence-Enhanced Ultrasound Flow Imaging at the Edge," *IEEE Micro*, vol. 42, no. 6, pp. 96-106, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[4] Mohanad Alkhodari, Leontios J. Hadjileontiadis, and Ahsan H. Khandoker, "Identification of Congenital Valvular Murmurs in Young Patients Using Deep Learning-Based Attention Transformers and Phonocardiograms," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 4, pp. 1803-1814, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] Kainat Zafar et al., "Deep Learning-Based Feature Engineering to Detect Anterior and Inferior Myocardial Infarction Using UWB Radar Data," *IEEE Access*, vol. 11, pp. 97745-97757, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Jingbo Wang, "OCT Image Recognition of Cardiovascular Vulnerable Plaque Based on CNN," *IEEE Access*, vol. 8, pp. 140767-140776, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[7] Jing Zhang et al., "MLBF-Net: A Multi-Lead-Branch Fusion Network for Multi-Class Arrhythmia Classification Using 12-Lead ECG," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 9, pp. 1-11, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[8] Børge Solli Andreassen et al., "Mitral Annulus Segmentation and Anatomical Orientation Detection in TEE Images Using Periodic 3D CNN," *IEEE Access*, vol. 10, pp. 51472-51486, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Müjde Akdeniz et al., "Deep Learning for Multi-Level Detection and Localization of Myocardial Scars Based on Regional Strain Validated on Virtual Patients," *IEEE Access*, vol. 11, pp. 15788-15798, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Fatemeh Taheri Dezaki et al., "Echo-SyncNet: Self-Supervised Cardiac View Synchronization in Echocardiography," *IEEE Transactions on Medical Imaging*, vol. 40, no. 8, pp. 2092-2104, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Jia Li et al., "Arrhythmia Classification Using Biased Dropout and Morphology-Rhythm Feature with Incremental Broad Learning," *IEEE Access*, vol. 9, pp. 66132-66140, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Jingshan Huang et al., "ECG Arrhythmia Classification Using STFT-Based Spectrogram and Convolutional Neural Network," *IEEE Access*, vol. 7, pp. 92871-92880, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[13] Li Ping Shi et al., "Efficient Graphene Reconfigurable Reflectarray Antenna Electromagnetic Response Prediction Using Deep Learning," *IEEE Access*, vol. 9, pp. 22671-22678, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[14] Ananth Bhoj, Ron Kinder, and Larry Gochberg, "Numerical Calculations on the Low-Pressure Behavior of a High-Density Plasma CVD Reactor," *IEEE 34th International Conference on Plasma Science*, Albuquerque, NM, USA, pp. 568-568, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[15] Yun Kwan Kim et al., "Automatic Cardiac Arrhythmia Classification Using Residual Network Combined with Long Short-Term Memory," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1-17, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[16] Liang-Hung Wang et al., "Three-Heartbeat Multilead ECG Recognition Method for Arrhythmia Classification," *IEEE Access*, vol. 10, pp. 44046-44061, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[17] Runnan He et al., "Automatic Cardiac Arrhythmia Classification Using Combination of Deep Residual Network and Bidirectional LSTM," *IEEE Access*, vol. 7, pp. 102119-102135, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[18] Hui Yang, and Zhiqiang Wei, "Arrhythmia Recognition and Classification Using Combined Parametric and Visual Pattern Features of ECG Morphology," *IEEE Access*, vol. 8, pp. 47103-47117, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[19] G.S. Nijaguna et al., "Feature Selection Using Selective Opposition Based Artificial Rabbits Optimization for Arrhythmia Classification on Internet of Medical Things Environment," *IEEE Access*, vol. 11, pp. 100052-100069, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Classification," *IEEE Transactions on Consumer Electronics*, vol. 69, no. 3, pp. 250-260, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[21] Maytham N. Meqdad, Fardin Abdali-Mohammadi, and Seifedine Kadry, "Meta Structural Learning Algorithm with Interpretable Convolutional Neural Networks for Arrhythmia Detection of Multisession ECG," *IEEE Access*, vol. 10, pp. 61410-61425, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[22] Tanvir Mahmud, Shaikh Anowarul Fattah, and Mohammad Saquib, "DeepArrNet: An Efficient Deep CNN Architecture for Automatic Arrhythmia Detection and Classification from Denoised ECG Beats," *IEEE Access*, vol. 8, pp. 104788-104800, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[23] Mariya R. Kiladze et al., "Multimodal Neural Network for Recognition of Cardiac Arrhythmias Based on 12-Load Electrocardiogram Signals," *IEEE Access*, vol. 11, pp. 133744-133754, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[24] Ashley Spann et al., "Applying Machine Learning in Liver Disease and Transplantation: A Comprehensive Review," *Hepatology*, vol. 71, no. 3, pp. 1093-1105, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[25] Yihao Liu, and Minghua Wu, "Deep Learning in Precision Medicine and Focus on Glioma," *BioEngineering and Translational Medicine*, vol. 8, no. 5, pp. 1-21, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[26] Md Maruf Hossain et al., "Cardiovascular Disease Identification Using a Hybrid CNN-LSTM Model with Explainable AI," *Informatics in Medicine Unlocked*, vol. 42, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[27] Agustin Martin-Morales et al., "Predicting Cardiovascular Disease Mortality: Leveraging Machine Learning for Comprehensive Assessment of Health and Nutrition Variables," *Nutrients*, vol. 15, no. 18, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[28] Patricia Rufes et al., "Heart Disease Prediction Using Machine Learning," *International Research Journal on Advanced Engineering Hub*, vol. 2 no. 3, pp. 485-490, 2024. [CrossRef] [Publisher Link]

[29] Najmu Nissa, Sanjay Jamwal, and Mehdi Neshat, "A Technical Comparative Heart Disease Prediction Framework Using Boosting Ensemble Techniques," *Computation*, vol. 12, no. 15, pp. 1-22, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[30] Yahia Baashar et al., "Effectiveness of Artificial Intelligence Models for Cardiovascular Disease Prediction: Network Meta-Analysis," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, pp. 1-12, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[31] Adedayo Ogunpola et al., "Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases," *Diagnostics*, vol. 14, no. 2. pp. 1-19, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[32] Jesse Gabriel, "A Machine Learning-Based Web Application for Heart Disease Prediction," *Intelligent Control and Automation*, vol. 15, no. 1, pp. 9-27, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[33] K. Karthick et al., "[Retracted] Implementation of a Heart Disease Risk Prediction Model Using Machine Learning," *Computational and Mathematical Methods in Medicine*, vol. 2023, no. 1, pp. 1-14, 2023. [CrossRef] [Google Scholar] [Publisher Link]