International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Machine Learning Innovations in Early Cardiovascular Disease Detection

Mukesh Kumar Saini

PhD, MBA, MCA, Technologist and Consulting Data Engineer
Email: *drmsaini78[at]gmail.com*

Abstract: ***Identifying and predicting the risk of Cardiovascular Diseases (CVD) in healthy individuals is crucial for effective disease management. Leveraging extensive health data available in hospital databases offers significant potential for early detection and diagnosis of CVD, which can greatly improve disease outcomes. Integrating machine learning techniques shows considerable promise in advancing clinical practices for CVD management. These methods enable the development of evidence-based clinical guidelines and management algorithms, potentially reducing the need for costly and extensive clinical and laboratory investigations, thereby easing financial burdens on patients and healthcare systems. To enhance early prediction and intervention for CVD, this study proposes the development of novel, robust, efficient machine learning algorithms tailored for automatic feature selection and early-stage heart disease detection. The proposed Catboost model achieves an F1-score of approximately 92.3% and an average accuracy of 90.94%. Compared to many current state-of-the-art approaches, it demonstrates superior classification performance with higher accuracy and precision.***

Keywords: Heart disease, Machine learning, Feature selection, Cardiovascular diseases, Quality of life, Disease prevention, CVD

1. Introduction

The heart, second only to the brain in importance within the human body, plays a critical role in maintaining bodily harmony. As we navigate the complexities of the modern era, profound transformations in our environment impact our daily lives significantly. Cardiovascular diseases (CVD), among the top five deadliest ailments globally, claim countless lives and necessitate timely intervention for effective management. CVD encompasses a spectrum of conditions affecting the heart and circulatory system, often stemming from factors like atherosclerosis. These diseases typically progress silently over time, manifesting symptoms only in advanced stages. According to the World Health Organization (WHO), CVD has long been the leading cause of premature death worldwide. By 2030, it is anticipated to be responsible for approximately 23.6 million deaths annually. The economic burden of treating CVD, measured in Disability Adjusted Life Years (DALYs), is considerable, emphasizing the need for proactive prevention strategies. Various factors contribute to CVD, categorized as modifiable and non-modifiable risk factors. Age, gender, and genetic predispositions fall into the non-modifiable category, whereas lifestyle-related factors such as blood sugar levels, hypertension, cholesterol levels, smoking, diet, obesity, and physical activity are considered modifiable.

Early recognition and management of CVD risk factors are crucial for prevention and effective treatment. Recognizable symptoms of a heart attack include chest discomfort, arm pain, dizziness, fatigue, and sweating. Given that symptoms may not appear until the disease is advanced, early detection through comprehensive analysis of health data stored in hospital databases is invaluable. Machine learning algorithms offer substantial promise in this regard, capable of identifying hidden patterns and facilitating predictive models for early CVD detection. These technologies can significantly augment clinical decision-making, enhance the development of evidence-based guidelines, and reduce the need for extensive diagnostic procedures, thereby alleviating financial burdens on healthcare systems and individuals alike. Cardiovascular diseases encompass chronic conditions that can lead to critical outcomes such as heart failure and coronary artery infarction. Early detection is pivotal for effective management, and machine learning tools can play a pivotal role in predicting and managing these conditions.

In this study, we propose leveraging Gradient Boosting models and other advanced machine learning techniques to predict and manage cardiovascular diseases effectively. The research aims to optimize prediction accuracy through robust algorithms and feature selection methodologies, contributing valuable insights to healthcare professionals and stakeholders.

2. Literature Survey

Numerous researchers have explored various frameworks for predicting cardiac diseases using data mining techniques. These studies utilize datasets and computational models to achieve more efficient results in diagnosing heart-related disorders.

Pattekari [36] developed a model using the Naive Bayesian data mining approach, where the system extracts hidden information from a dataset and evaluates user inputs against predefined data. This aid medical professionals in making informed clinical decisions and potentially reducing treatment costs.

Tran [37] implemented an Intelligent System based on Naive Bayes, which functions similarly to predict heart disease diagnoses, enhancing clinical decision-making and lowering treatment expenses.

Gnaneswar [38] emphasized the importance of monitoring heart rate during cycling to manage exercise intensity and avoid overtraining and cardiac stress. They proposed a feedforward neural network model to predict pulse based on cycling rhythm.

Mutijarsa [39] highlighted the role of data mining techniques in detecting and localizing coronary disease, comparing various algorithms to determine the most accurate predictors.

Yeshvendra [40] discussed the increasing use of AI algorithms in disease prediction, particularly in enhancing the accuracy of coronary disease prognosis.

Patil [41] explored tree-based data mining techniques like support vector machines, naive Bayes, and decision trees for creating predictive models in cardiac diagnosis.

Tripoliti [42] emphasized the urgent need for biomedical research in identifying diseases with high prevalence rates, including coronary disease, using advanced computational methods.

Gonsalves [43] applied machine learning to forecast coronary cardiovascular disease (CVD) using historical medical data.

Oikonomou [44] utilized machine learning to analyze chronic disease data, applying extreme value theory to assess disease severity and risk.

Ibrahim [45] highlighted the role of machine learning in predicting and diagnosing heart disease, particularly through active learning methods that improve classification accuracy.

Pratiyush et al. [46] explored ensemble classifiers within the eXplainable AI (XAI) framework to predict heart disease from comprehensive cardiovascular datasets.

Overall, these studies aim to enhance early prediction and intervention in cardiovascular disease through robust machine learning algorithms, addressing the complexities and challenges of disease prediction with advanced computational techniques.

3. Methodology

This section outlines the proposed classification framework for instances of heart disease. Initially, exploratory data analysis is conducted, encompassing a comprehensive examination of both the target variable and features. Categorical variables are converted into numerical values as part of the preprocessing stage. Various evaluation criteria are applied to compare different models under consideration. The outputs of each model are thoroughly analyzed, leading to the selection of the optimal model for the specific problem domain.

The proposed model undergoes rigorous examination, with the Optuna library employed to fine-tune model hyperparameters and assess enhancements achieved. The suggested classification scheme is structured into three main phases: (1) preprocessing, (2) training, and (3) classification, as illustrated in Figure 1. Each of these components will be further detailed in subsequent sections of this study.

3.1 Pre-Processing

Before proceeding with training, the selected models, it's essential to address missing values in Cholesterol data, initially input as 0. The approach involves segregating the data into groups based on the presence of verified cardiac conditions, and then replacing missing values with the mean of each respective group. This step ensures data integrity and prepares it for further analysis. To determine the influence of these variables in predicting heart disease, Shapley Values are utilized. Interaction terms are introduced into the models to capture potential correlations among data elements. SHAP (SHapley Additive exPlanations), a method rooted in game theory, is employed to assess the significance of each characteristic. It elucidates the contribution of each predictor to both individual model predictions and aggregated model outcomes by averaging their marginal contributions across all possible feature combinations.

Initially, a gradient boosting model is trained using all variables. Subsequently, feature selection based on Shapley Values identifies predictors with values greater than 0.1, signifying substantial contributions to the model's predictive capability. These selected predictors are then employed to establish the most effective model configuration. Given the multicollinearity among interaction variables, various nonparametric tree-based methods are explored to identify the most accurate approach for predicting the risk of Cardiovascular Disease (CVD). This comprehensive approach ensures robust model selection and validation, aimed at enhancing predictive accuracy and model interpretability in assessing CVD risk.

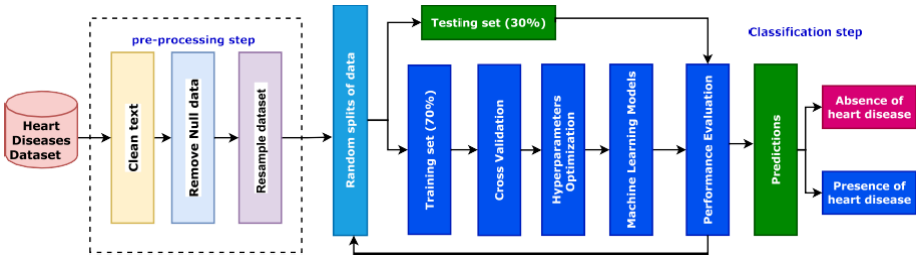


Figure 1: The main steps of the proposed methodology

3.2 Training Process

Once the datasets have been preprocessed and normalized, the machine learning algorithm can be effectively trained. Following data modification, the dataset is randomly split into a training set and a test set. Typically, 70% of the data is allocated to the training set and 30% to the test set. This ensures that the model is trained on a sufficiently large dataset while retaining a separate portion for evaluation.

To ensure robust evaluation, the k-fold cross-validation technique is employed. This method involves dividing the dataset into k subsets (or folds) of approximately equal size. The model is then trained and evaluated k times, each time using a different fold as the test set and the remaining folds as the training set. This approach provides a more reliable estimate of model performance compared to a single train-test split, as it mitigates the risk of overfitting and variance in

performance metrics. Several machine learning algorithms are evaluated to determine the optimal model for predicting heart disease. These include Support Vector Classifier (SVC), Multinomial Naive Bayes (MultinomialNB), K-Nearest Neighbors (K-Neighbor), Bernoulli Naive Bayes (BernoulliNB), Stochastic Gradient Descent (SGD), Random Forest, and Decision Tree. Each algorithm offers unique strengths and is assessed based on its performance metrics.

XGBoost (Extreme Gradient Boosting) is highlighted as a particularly effective supervised learning method for enhancing prediction accuracy. XGBoost iteratively builds decision trees using gradient boosting, where each subsequent tree aims to correct the errors of the previous trees. This iterative approach results in a powerful ensemble model capable of handling complex datasets and achieving high predictive performance. XGBoost's objective function includes a loss function as well as a regularization term, which helps to prevent overfitting.

AdaBoost (Adaptive Boosting) is another boosting algorithm that also uses decision trees as weak learners. AdaBoost assigns weights to each training example, with higher weights given to examples that were misclassified by the previous weak learner. In each subsequent iteration, a new decision tree is trained on the weighted data, with the weights updated based on the accuracy of the tree.

The *Linear Support Vector Classifier (SVC)* focuses on employing a linear decision boundary to classify data and performs exceptionally well with large datasets [54]. The Linear SVC has certain constraints, such as the requirement for standardized input and output operations. Since the Linear SVC relies on the kernel method, the kernel method cannot be modified. A Linear SVC is designed to handle data by finding the "best fit" hyperplane that separates or categorizes it. Once the hyperplane is determined, the features are fed into the classifier, which predicts their class membership.

The Naive Bayes algorithm assigns equal weight to all features or attributes. This algorithm becomes more efficient as it assumes independence among features. According to Yasin (2020), the Naive Bayes classifier (NBC) is a simple, effective, and well-known algorithm for text categorization. NBC has utilized Bayes' theorem for document classification since the 1950s, and it has a solid theoretical foundation. The classifier uses posterior estimates to determine class membership based on the highest conditional probability of features.

Bernoulli Naive Bayes is a statistical technique that generates boolean results based on the presence or absence of specific words. This classifier operates on the discrete Bernoulli Distribution. It is useful for detecting unwanted keywords or tagging specific word types within text. Unlike the multinomial approach, it produces binary outputs such as 1–0, True–False, or Yes–No.

Stochastic Gradient Descent (SGD) is a method that updates the model parameters using gradients computed on small random subsets of the data. This approach is computationally efficient because it does not require the entire dataset to compute the gradients in each iteration. SGD is particularly useful for optimizing differentiable or sub-differentiable objective functions iteratively.

Decision Tree is a well-known machine learning technique where data is recursively partitioned based on specific parameters. The tree structure includes nodes and leaves, where leaves represent decisions or outcomes, and decision nodes partition data [59]. Decision trees can be combined for ensemble learning to solve complex problems. The Random Forest algorithm addresses overfitting issues associated with decision trees by aggregating multiple decision trees. It can handle both regression and classification tasks, and evaluates numerous attributes to identify the most important ones [58].

The *K-Nearest Neighbor (K-NN) algorithm* classifies new observations based on their distances from known examples. It assigns the class based on the majority vote of its nearest neighbors using a distance function as a metric. In classification problems, K-NN returns the class membership, while in regression problems, it predicts the property value of the object. The effectiveness of K-NN can be significantly improved through normalization when dealing with features that have different physical units or scales [56].

3.3 Classification

The proposed model is based on machine learning with strong generalization capabilities and a high degree of paradigm-specific precision. In this study, we will evaluate several machine learning algorithms objectively to determine which one yields the best results. This addresses the primary purpose of using machine learning: to mitigate overfitting issues inherent in machine learning models. The curriculum also covers the fundamental concept of risk minimization.

Machine learning excels in accurately classifying data, especially in high-dimensional spaces, by identifying a hyperplane that maximizes separation. At this stage, labeled data serves as input, and significant features are extracted through a feature extraction process. Finally, the optimized model is used to classify new instances of data.

.

3.4 Experimental evaluation

In our study experiments, we utilized Google Colab as the implementation platform for our machine learning models. Google Colab provides a virtual machine running on Google's servers, offering users a Python environment equipped with popular data science libraries such as TensorFlow, PyTorch, and Scikit-Learn. It functions as a cloud-based Jupyter notebook environment that provides free access to computing resources, including a virtual machine with 12 GB of RAM and up to 100 GB of hard disk space. The memory allocation for the virtual machine can be increased to 25 GB, with options for high-RAM configurations up to 52 GB available for handling large-scale models or datasets. The virtual machine is equipped with an NVIDIA Tesla K80 GPU, facilitating efficient training of deep learning models.

Moreover, Google Colab includes a comprehensive set of preinstalled libraries and tools, simplifying the installation and use of necessary dependencies. The virtual machine operates on a stable and reliable Linux-based operating system,

specifically Linux Ubuntu, which comes preconfigured with essential system libraries and tools commonly utilized in data science projects.

3.4.1 Data Collection

The heart condition data used in this study is a synthesis of datasets sourced from the UCI Machine Learning Repository. It comprises eleven features that are utilized to predict the presence of heart failure, a prevalent cardiovascular disease that significantly increases the likelihood of cardiovascular-related mortality [60-61].

Table 1: A sample of the Heart Failure Dataset

Age	Sex	Type chest pain	BP resting	Cholesterol	BS fasting	ECG resting	HR max	Angina exercise	Old peak	ST slope	Disease of heart
41	M	ATA	142	287	0	Nor1	173	N	0.0	Upper	0
48	F	NAP	162	182	0	Nor1	157	N	1.0	Flat1	1
38	M	ATA	132	273	0	ST	98	N	0.0	Upper	0
49	F	ASY	136	224	0	Nor1	109	Y	1.5	Flat1	1
53	M	NAP	152	185	0	Nor1	123	N	0.0	Upper	0

a binary attribute that indicates a diagnosis of Heart Failure if HeartDisease is = 1 as illustrated in Table 1. Moreover, Table 2 presents the list of variables and the description of the features in the heart disease dataset. The dataset was created by combining a diverse range of datasets that were previously available independently, and were not combined before [60, 61]. In this dataset, five heart datasets are combined over 11 common features which makes it the largest heart disease dataset accessible for research purposes. The specific datasets utilized in the curation of this composite dataset are shown in Table 3.

Table 2: Symptoms, signs and laboratory investigations of the dataset of the heart disease

Variable	Interpretation
Age	Patient's Age/year
Gender	Patient's Gender, Male/Female
Type of chest pain	Type of chest pain: i. TA: Typical Angina ii. ATA: Atypical Angina iii.NAP: Non-Anginal Pain iv.ASY: Asymptomati
Resting blood pressure	Patient's Blood Pressure/mmHg.
Total Cholesterol	Patient's Cholesterol (mg/dl).
Blood Glucose level (Fasting)	Patient's fasting blood glucose level. i. glucose >120 mg/dL =1 ii. glucose below 120 mg/dL =0
ECG at rest	Electrocardiography (at rest): i. Normal ii. ST: ST segment and/or T wave abnormality iii.LVH: Probable or Definite Left Ventricular Hypertrophy
Heart Rate at Maximum	Maximum Heart Rate, heart beats per minute.
Angina on Exercising	Exercise-associated Angina, present /absent.
Old peak	Measure of ST Depression.
ST_Slope	Slope of Peak Exercise. i.Up: up sloping ii.Flat iii.Down: down sloping

Table 3: The different datasets used to create the dataset of the heart disease

Datasets	#Observations
Cleveland	303
Hungarian	294
Stalog (Heart) Data Set	270
Long Beach VA	200
Switzerland	123
Total	1190
Duplicated	272
Final dataset	918

The Heart Disease dataset has 918 observations and 12 columns [60, 61]. Table 4 summaries the main statistics for the numeric features. It is clear that, the mean value of age is 53 and the maximum is 77 as shown in Table 4. Similarly, Table 5 presents

Table 4: Summary statistics of numeric variables

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
Count	918	918	918	918	918	918	918
Max	77	200	603	1	202	6.20	1
Min	28	0	0	0	60	-2.6	0
Mean	53.51	132.39	198.79	0.23	136.81	0.89	0.55
Std	9.43	18.51	109.38	0.42	25.46	1.06	0.49
25%	47	120	173.25	0	120	0	0
50%	54	130	223	0	138	0.60	1
75%	60	140	267	0	156	1.50	1

Table 5: Summary statistics of categorical variables

	Sex	TypeChestPain	ECGResting	AnginaExercise	ST_Slope
Count	920	920	920	920	920
Unique	2	3	4	2	4
Top	M	ASY	Normals	N	Flat1
Freq	735	486	562	557	470

Table 6 The proportion of Heart Disease

Variable	Value	Total patients	Proportion of heart disease
Sex	M	725	90.2%
	F	193	9.8%
Chest Pain Type	ASY	496	77.2%
	NAP	203	14.2%
	ATA	173	4.7%
	TA	46	3.9%
Resting ECG	Normal	552	56.1%
	ST	178	23.0%
	LVH	188	20.9%
Exercise Angina	Y	371	62.2%
	N	547	37.8%
ST_Slope	Flat	460	75.0%
	Up	395	15.4%
	Down	63	9.6%

*bold numbers mean the highest frequency and percentage

the statistics of categorical attributes. From this table, the unique values in ChestPain- Type attribute are 4 and the top is "ASY".

Table 6 summaries the main details for the numeric features. It is clear that, the variable Sex has two main values male (M) and female (F) such that the proportion of Heart Disease for M is 90.2% and for F is 9.8%. Similarly, Table 6 presents the statis- tics of ChestPainType attribute, there are 4 values (ASY, NAP, ATA, and TA) and the most frequent is ASY of 77.2%.

3.4.2 Exploratory Data Analysis

Remarkably, the classifications in the heart disease attribute value are reasonably well- balanced. 508 of the 918 patients who participated in the study have been diagnosed with heart failure, while 410 have not. Patients with heart disease have a median age of 57, whereas those without heart disease have a typical age of 51. As illustrated in Fig. 2, around 63% of males have heart disease, whereas approximately 25% of females have been diagnosed with heart disease. A female has a chance of 25.91% having a Heart Dis- ease. A male has a probability of 63.17% having a Heart Disease.

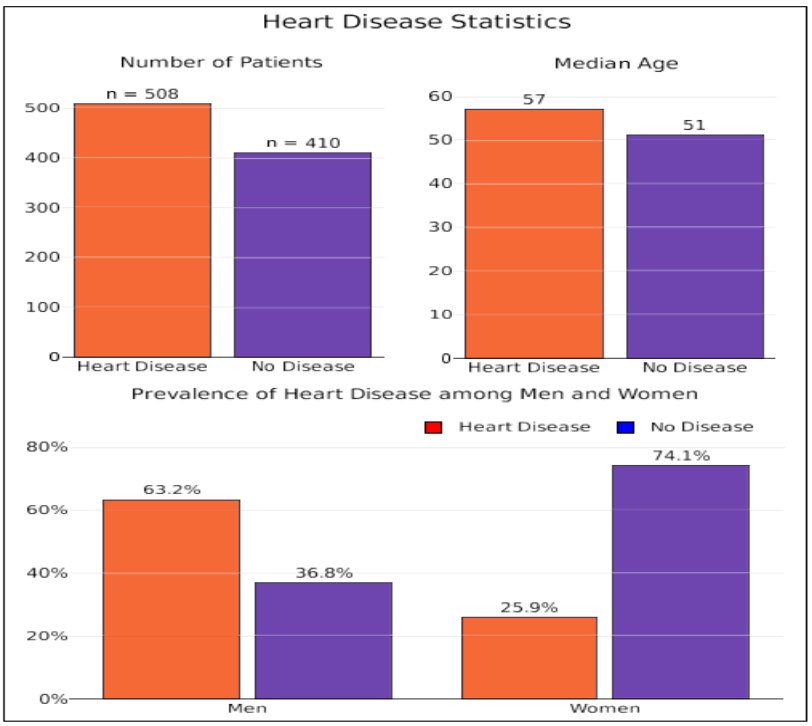


Figure 2: Prevalence of heart disease among men and women

Figure 3 displays the correlation matrix associated with the heart disease dataset. Heart disease has the strongest positive link with OldPeak (correlation = 0.4) and the strongest negative association with MaxHR (correlation = 0.4), according to the correlation matrix. Age and MaxHR also have a reasonably high link, with a correlation of 0.38. As seen in Fig. 4, heart rate tends to decrease as age increases. Results observe a weak correlation between the numerical features and the target variable based on the matrix. Oldpeak (a depression-related number) correlates positively with heart disease. Heart disease is negatively correlated with maximal heart rate. Cholesterol has an interestingly negative association with heart disease.

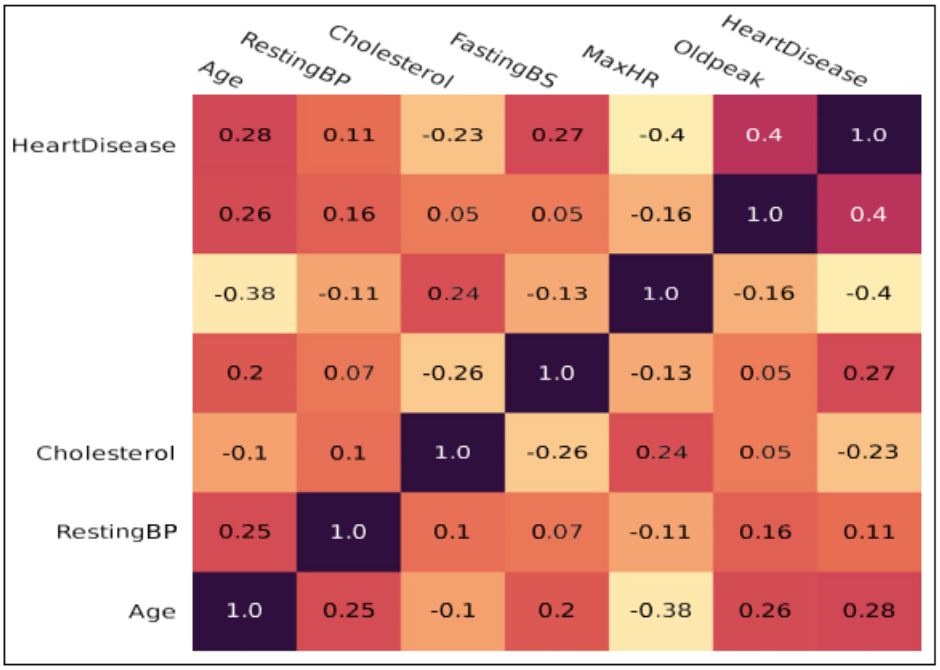
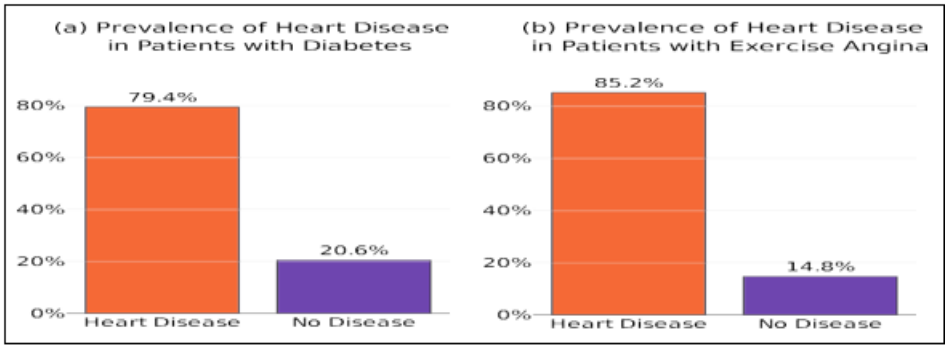


Figure 3: The correlation matrix for the Heart Disease dataset

Figure 4 illustrates the correlation between heart disease and category variables. Nearly 80% of diabetic persons suffer heart problems. Patients with exercise-induced angina have an even greater incidence of cardiovascular disease, at over 85%. Over 65% of patients diagnosed with cardiac disease had ST-T wave abnormalities in their resting ECGs, the greatest percentage across the categories. Patients with a Flat or Declining ST Slope during exercise have the highest frequency of cardiovascular dis- ease, at 82.8% and 77.8%, respectively.

Figure 5 explains data details regarding asymptomatic chest pain in heart disease at almost 77%, the absence of chest pain (asymptomatic) is the most prevalent symptom in patients with heart disease. In addition, heart disease is roughly nine times more prevalent in males than in females among patients with a cardiovascular diagnosis.



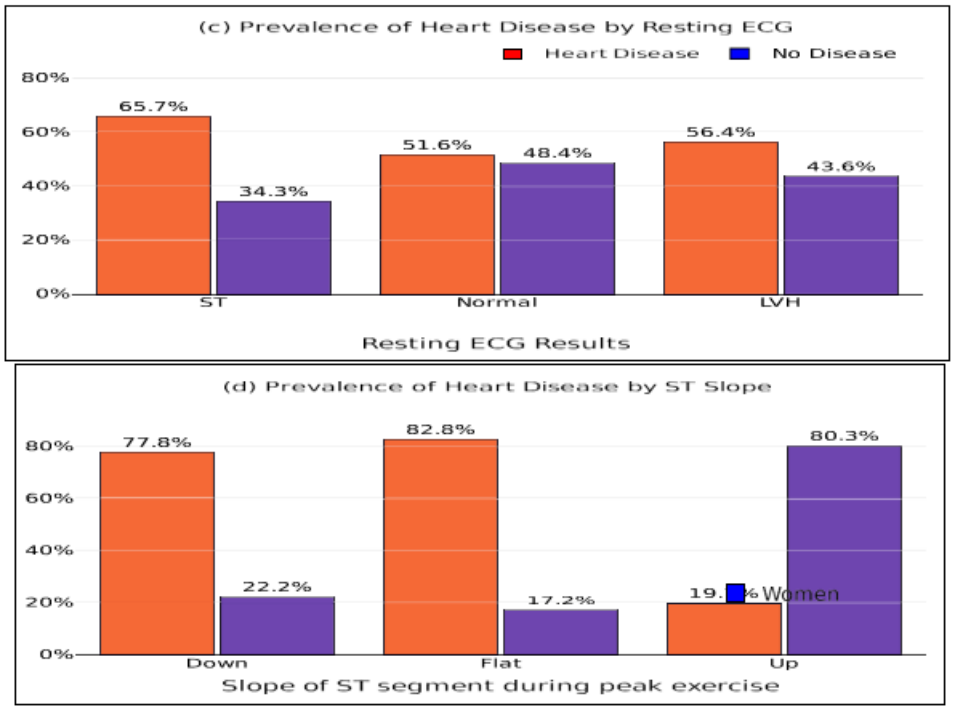


Figure 4: Prevalence of heart disease by resting ECG. (a) Prevalence of Heart Disease in Patients with Diabetes (b) Prevalence of Heart Disease in Patients with Exercise Angina. (c) Prevalence of Heart Disease by Resting ECG. (d) Prevalence of Heart Disease by ST Slope

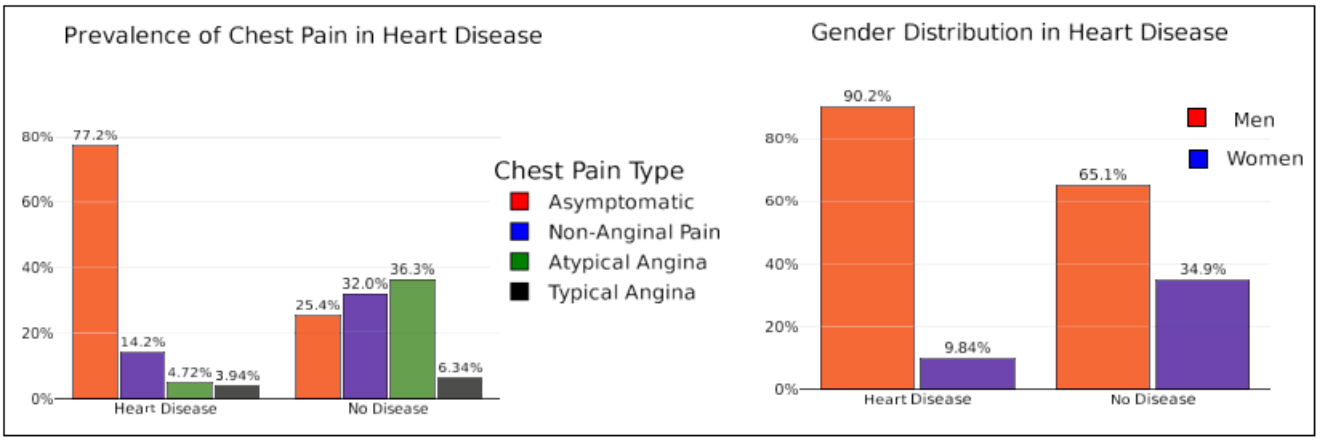


Figure 5: Prevalence of chest pain in heart disease data

patient with asymptomatic chest pain (ASY) is approximately six times more likely to suffer heart disease than a patient with atypical angina chest pain (ATA).

Overall insights obtained from the exploratory data analysis, Data for the target var- iable are near to balanced. The association between numerical features and the target variable is weak. Oldpeak (a depression-related number) correlates positively with heart disease. Heart illness is negatively correlated with maximum heart rate. Interestingly, there is a negative link between cholesterol and heart disease. Males are approximately 2.44 times more likely to suffer from heart disease than females. There are distinct vari- ances between the types of chest pain. Patients with asymptomatic chest pain (ASY) are about six times more likely to suffer heart disease than those with Atypical Angina chest pain (ATA). Resting ECG: electrocardiogram values at rest are comparable. Patients with ST-T pulse abnormalities have a higher risk of developing heart disease than those who do not. ExerciseAngina: people who have exercise-induced angina are nearly 2.4 times more likely to have heart disease than people who don't. The slope of the ST segment at maximum exertion varies. ST Slope Up has a considerably lower risk of cardiovascular disease than the other two segments. Exercise-induced angina with a 'Yes' score is nearly 2.4 times more likely to result in heart disease than exercise-induced angina with a 'No' score.

4. Results

Table 7 illustrates the classification results of the various classifiers on the dataset. The table reports the performance of various classifiers on a given dataset, measured in terms of Accuracy, Precision, Recall, and F1 score. Comparing the results of the pro- posed technique against those of other classifiers such as SVM [54], XGBoost, Ada- Boost, RandomForest [58], LinearDiscriminant [67], LightGBM, GradientBoosting, Catboost, ExtraTree, KNeighbors [56], and LogisticRegression [68]

Table 7: Comparative results on the Dataset using ML

Classifier	Accuracy	Precision	Recall	F1
XGBoost	0.8297	0.8980	0.8049	0.8489
AdaBoost	0.8659	0.9262	0.8415	0.8818
LinearDiscriminant	0.8696	0.9156	0.8598	0.8868
LightGBM	0.8732	0.9057	0.8780	0.8916
GradientBoosting	0.8768	0.9276	0.8598	0.8924
Catboost	0.8804	0.9226	0.8720	0.8966
ExtraTree	0.8804	0.9281	0.8659	0.8959
KNeighbors	0.8841	0.9074	0.8963	0.9018
SVM	0.8841	0.8976	0.9085	0.9030
LogisticRegression	0.8841	0.9231	0.8780	0.9000
RandomForest	0.8877	0.9236	0.8841	0.9034
Catboost_tuned	0.9094	0.9317	0.9146	0.9231

demonstrates the method's utility. The results of classifiers according to various metrics are displayed. The highest performing classifier based on all measures is Catboost_tuned, which achieved an accuracy of 0.9094, a precision of 0.9317, a recall of 0.9146, and F1 score of 0.9231. Other top-performing classifiers include RandomForest, LogisticRegression, SVM, and KNeighbors, with similar accuracy and precision scores, but slightly lower recall and F1 scores. In contrast, lower-performing classifiers such as XGBoost and AdaBoost exhibit moderate accuracy and precision scores, but relatively lower recall and F1 scores. Over- all, the results suggest that the choice of classifier can have a significant effect on the performance of a predictive model.

The present study employs a confusion matrix (Fig. 6) to report the performance of models in accurately predicting cardiac disease for a given set of patients, with due con-sideration to both correctly classified and misclassified instances. Specifically, the Gradi- ent Boosting model is found to exhibit the highest proportion of True Positives (TP) and True Negatives (TN) when evaluated on a test set. The computation of FN, FB, TN, and TP, values for the cardiac disease class is carried out using the Gradient Boost model, whereby the predicted values are expected to match the actual values. For instance, TP corresponds to the value at cell 1 of the confusion matrix, while FN is computed by add- ing the relevant row values, excluding TP (i.e., FN = 12). Similarly, FP is calculated as the total of column values, excluding TP, leading to a value of 11. Lastly, TN is determined by the combination of all columns and rows except the class under consideration (i.e., cardiac disease), which yields a value of 81.

5. Discussion

Despite the vast amount of data generated by healthcare systems, the field of medicine faces unique challenges compared to other data-driven industries where machine learning has thrived. The Health Insurance Portability and Accountability Act (HIPAA) mandates stringent, institution-specific Institutional Review Boards (IRBs) to govern the use of patient data, ensuring robust patient privacy protections. However, this has inadvertently led to fragmented data silos across the country [47]. Consequently, many machine learning models published in healthcare rely on locally collected datasets and lack external validation. According to the Tufts predictive analytics and comparative effectiveness cardiovascular prediction model, 58% of cardiovascular prediction models have never been externally validated [69]. Heart-related disorders remain one of the leading global causes of death and morbidity [5–7].

Individuals with heart disease often remain unaware of their condition, making early detection and intervention crucial to saving lives, reducing complications, and alleviating the global burden of disease and mortality [9]. Machine learning models offer promising capabilities in accurately predicting and diagnosing heart disorders in their early stages [12–14]. Medical machine learning presents numerous opportunities, including uncovering hidden patterns that enhance diagnostic accuracy across diverse medical datasets. Previous studies have shown the potential of machine learning in predicting cardiovascular diseases [15, 16].

Previous research has utilized various machine learning approaches such as neural networks, Naive Bayes, Decision Trees, and SVM for diagnosing cardiac disorders, achieving varying degrees of accuracy [18, 19]. For instance, the hybrid feature selection methodology algorithm (CFS+Filter Subset Eval) attained an accuracy of 85.5% [70].

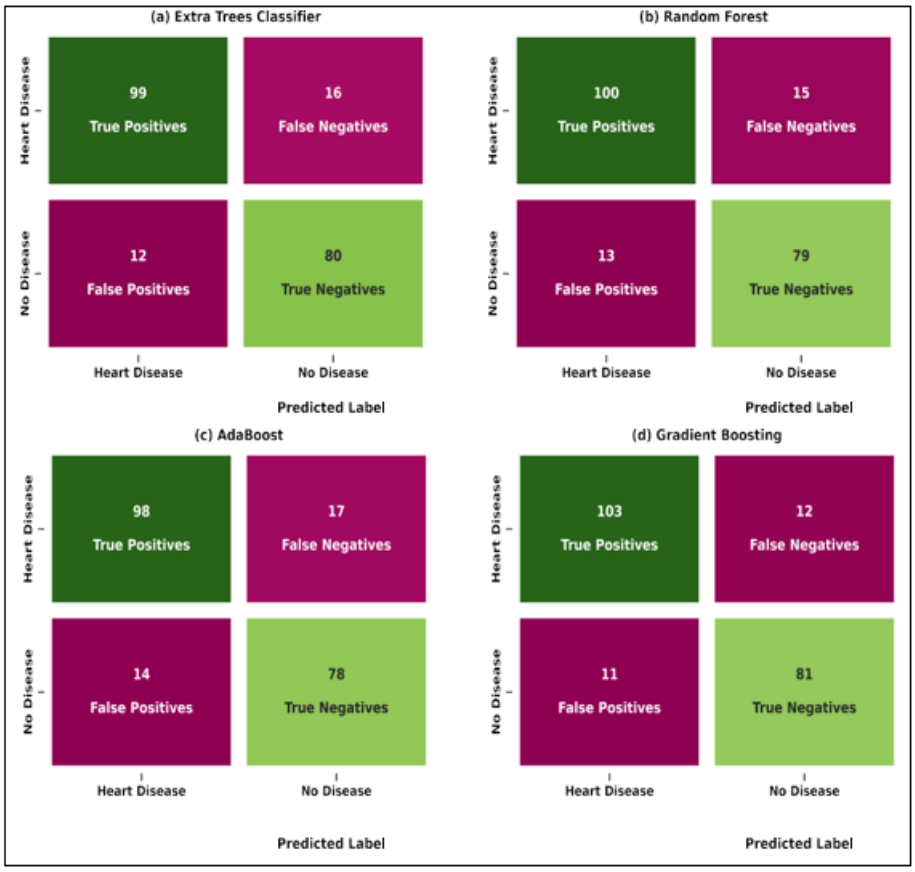


Figure 6: The confusion matrix results for Extra Trees, RandomForest, AdaBoost, and GradientBoost classifiers

methodology algorithm (CFS+Filter Subset Eval) attained an accuracy of 85.5% [70]. Shouman et al. [71] combined k-means clustering with Naive Bayes, achieving an accuracy of 84.5% in diagnosing heart disease. Rupali et al. [72] developed the Heart Disease Prediction System (HDPS) using Naive Bayesian Classification and Jelinek-Mercer smoothing, achieving an accuracy of 86%. Elma et al. [73] integrated K-nearest neighbor (cNK) and NaiveBayes classifiers, achieving an accuracy rate of 85.92%. Dulhare et al. [74] improved prediction methods using Naive Bayes and particle swarm optimization, achieving an accuracy of 87.91%.

In the current study, Shapley values were employed to develop a Gradient Boosting model for predicting cardiovascular diseases (CVDs), yielding an Area Under the Curve (AUC) of 0.927%. Shapley values identified critical signs of cardiac disease and their predictive power, revealing significant interactions among patient medical information, particularly in Age, Cholesterol, Blood Pressure, ST Slope, and Chest Pain type. The proposed Catboost model demonstrated superior performance, achieving an F1-Score of 92.3% and an accuracy of 90.94% after model optimization. Overall, this model outperformed previous approaches in diagnosing cardiac disease. However, this study has several limitations. It relied solely on secondary data from selected cardiology and internal medicine departments, resulting in some missing variables and incomplete data. The cross-sectional design also limited the ability to assess longitudinal effects of risk factors on CVD development.

Future directions should focus on enhancing prediction techniques by integrating diverse machine learning methods to improve accuracy and precision in CVD prediction and early diagnosis. This approach has shown superiority over traditional state-of-the-art methods. The suggested machine learning model for predicting heart disorders is robust, effective, and efficient, achieving higher accuracy and precision percentages while utilizing fewer features. This is crucial for clinical practice, which demands precise and straightforward diagnostic methods to support therapeutic decision-making.

Nevertheless, there are challenges to the generalizability of the CVD prediction models presented here. Further research should validate these machine learning algorithms across different population datasets to minimize variations in CVD prevalence patterns and assess their impact on clinical decision-making and patient outcomes before integrating them into clinical guidelines.

6. Conclusions

Prediction of cardiovascular diseases is crucial for assisting clinicians with early disease diagnosis. Instead of replacing clinicians, machine learning will be a supplement to the clinical portfolio, enhancing human-led decision-making and clinical practices. Furthermore, by using machine learning techniques, the cost of conducting a long list of expensive clinical and laboratory investigations will be eliminated, reducing the financial burden on patients and the healthcare system. This paper proposed new robust, effective, and efficient machine learning algorithms for predicting CVD based on symptoms, signs, and other patients' information from hospital records in order to improve the early prediction of CVD development in its early stages and to ensure early intervention with a warranted recovery. The new technique

was more accurate and precise than existing standard art-of-state algorithms for the classification and prediction of heart disease. Future research evaluating the performance of the proposed machine learning algorithms on datasets containing a greater number of modifiable and non-modifiable risk factors will be crucial for the development of a more accurate and robust system for the prediction and early diagnosis of heart diseases.

Author Contributions: Contributed to conceptualization, data curation, methodology, software, validation, visualization, and writing of the original draft.

Funding: This research did not receive any external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

References

[1] Javeed A, Rizvi SS, Zhou S, Riaz R, Khan SU, Kwon SJ. Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification. Mob Inf Syst. 2020;2020:1–11. https://doi.org/ 10.1155/2020/8843115.

[2] Eckel R, Jakicic J, Ard JD. Aha/acc guideline on lifestyle management to reduce cardiovascular risk: a report of the american college of cardiology/american heart association task force on practice guidelines. American College of Cardiology/American Heart Association Task Force on Practice Guidelines. 2014. https://doi.org/10.1161/01.cir.00004 37740.48606.d1.pmid:24222015.

[3] Anderson KM, Wilson PW, Odell PM, Kannel WB. An updated coronary risk profile. A statement for health profession- als. Circulation. 1991;83(1):356–62. https://doi.org/10.1161/01.cir.83.1.356.

[4] Azmi J, Arif M, Nafis MT, Alam MA, Tanweer S, Wang G. A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data. Med Eng Phys. 2022;103825.

[5] Day TE, Goldlust E. Cardiovascular disease risk profiles. Am Heart J. 2010;160(1):3. https://doi.org/10.1016/j.ahj.2010. 04.019.

[6] Alwan A. Global status report on noncommunicable diseases. World Health Organization, 2011;293–298.

[7] 7.....Tsao CW, Aday AW, Almarzooq ZI, Anderson CAM, Arora P, Avery CL, Baker-Smith CM, Beaton AZ, Boehme AK,

[8] Buxton AE, Commodore-Mensah Y, Elkind MSV, Evenson KR, Eze-Nliam C, Fugar S, Generoso G, Heard DG, Hiremath S, Ho JE, Kalani R, Kazi DS, Ko D, Levine DA, Liu J, Ma J, Magnani JW, Michos ED, Mussolino ME, Navaneethan SD, Parikh NI, Poudel R, Rezk-Hanna M, Roth GA, Shah NS, St-Onge M-P, Thacker EL, Virani SS, Voeks JH, Wang N-Y, Wong ND, Wong SS, Yaffe K, Martin SS. Heart disease and stroke statistics-2023 update: a report from the American heart association. Circulation. 2023. https://doi.org/10.1161/CIR.0000000000001123.

[9] Wilson P, DAgostino RB, Levy D, Belanger A, Silbershatz H, Kannel W. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97(12):1837–47. https://doi.org/10.1161/01.CIR.97.18.1837.

[10] Mythili T, Mukherji D, Padalia N, Naidu A. A heart disease prediction model using svm-decision trees-logistic regres- sion (sdl). Int J Comput Appl. 2013;68(16):11–5. https://doi.org/10.1161/01.CIR.97.18.1837.

[11] Frieden TR, Jaffe MG. Saving 100 million lives by improving global treatment of hypertension and reducing cardio- vascular disease risk factors. J Clin Hypertens. 2018;20(2):208.

[12] Haissaguerre M, Derval N, Sacher F, Deisenhofer I, de Roy L, Pasquie J, Nogami A, Babuty D, Yli-Mayry S. Sudden cardiac arrest associated with early repolarization. N Engl J Med. 2008;58(19):2016–23.

[13] Kumar PM, Lokesh S, Varatharajan R, Babu GC, Parthasarathy P. Cloud and iot based disease prediction and diagnosis system for healthcare using fuzzy neural classifier. Future Gener Comput Syst. 2018;68:527–34.

[14] Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning technique. IEEE Access. 2019;7:81542–54.

[15] Kwon JM, Lee Y, Lee S, Park J. Effective heart disease prediction using hybrid machine learning technique. J Am Heart Assoc. 2018;7(13):1–11.

[16] Esfahani HA, Ghazanfari M, Ecardiovascular disease detection using a new ensemble classifier. in,. IEEE 4th interna- tional conference on knowledge-based engineering and innovation (KBEI). Tehran, Iran. 2017;2017:488–96.

[17] Gandhi M, Singh SN. Cardiovascular disease detection using a new ensemble classifier. in 2015 International Confer- ence on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), Greater Noida, India, 2015;520–525

[18] Krittanawong C, Virk HUH, Bangalore S, Wang Z, Johnson KW, Pinotti R, Zhang H, Kaplin S, Narasimhan B, Kitai T, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. Sci Rep. 2020;10(1):16057.

[19] Shouman TT, Stocker R. Integrating clustering with different data mining techniques in the diagnosis of heart disease. J Comput Sci Eng 2013;20(1).

[20] Motur S, Rao ST, Vemuru S. Frequent itemset mining algorithms: a survey. J Theor Appl Inf Technol 2018;96(3).

[21] Javeed A, Khan SU, Ali L, Ali S, Imrana Y, Rahman A. Machine learning-based automated diagnostic systems devel- oped for heart failure prediction using different types of data modalities: A systematic review and future directions. Comput Math Methods Med. 2022;2022:1–30. https://doi.org/10.1155/2022/9288452.

[22] Malki Z, Atlam E, Dagnew G, Alzighaibi AR, Ghada E, Gad I. Bidirectional residual lstm—based human activity recog- nition. J Comput Inf Sci. 2020;13(3):1–40.

[23] Malki Z, Atlam E-S, Hassanien AE, Dagnew G, Elhosseini MA, Gad I. Association between weather data and COVID- 19 pandemic predicting mortality rate: machine learning approaches. Chaos Solitons Fractals. 2020;138: 110137.

https://doi.org/10.1016/j.chaos.2020.110137.
[24] Atlam E-S, El-Raouf MMA, Ewis A, Ghoneim O, Gad I. A new approach to identify psychological impact of covid-19 on university students academic performance. Alex Eng J. 2021;61(7):5223–33.
[25] Malki Z, Atlam E-S, Ewis A, Dagnew G, Reda A, Elmarhomy G, Elhosseini MA, Hassanien AE, Gad I. ARIMA models for predicting the end of COVID-19 pandemic and the risk of a second rebound. J Neural Comput Appl. 2020;33(7): 2929–2948.https://doi.org/10.21203/rs.3.rs-34702/v1
[26] Almars MM, Almaliki M, Noor TH, Alwateer MM, Atlam E. Hann: hybrid attention neural network for detecting covid- 19 related rumors. IEEE Access. 2022;10:12334–44.
[27] Malki Z, Atlam E-S, Ewis A, Dagnew G, Ghoneim OA, Mohamed AA, Abdel-Daim MM, Gad I. The covid-19 pandemic: prediction study based on machine learning model. J Environ Sci Pollut Res. 2021;28(30):40496–506.
[28] Manjunatha MFDH, Ibrahim Gad E-SA, Ahmed A, Elmarhomy G, Elmarhoumy M, Ghoneim OA. Parallel genetic algorithms for optimizing the sarima model for better forecasting of the ncdc weather data. Alexandria Eng J. 2020;60:1299–316.
[29] Khan MA, Algarn F. A healthcare monitoring system for the diagnosis of heart disease in the iomt cloud environ- ment using msso-anfis. IEEE Access. 2020;8:122259–69.
[30] Javeed A, Zhou S, Yongjian L, Qasim I, Noor A, Nour R. An intelligent learning system based on random search algo- rithm and optimized random forest model for improved heart disease detection. IEEE Access. 2019;7:180235–43. https://doi.org/10.1109/access.2019.2952107.
[31] Meter W. World Meter. Accessed: October 2020 (2020). https://www.worldometers.info/coronavirus/.
[32] Coronavirus: Who (2020) coronavirus (2020). www.who.int/health-topics/.
[33] Ali L, Rahman A, Khan A, Zhou M, Javeed A, Khan JA. An automated diagnostic system for heart disease prediction based on χ^2 statistical model and optimally configured deep neural network. IEEE Access. 2019;7:34938–45. https:// doi.org/10.1109/access.2019.2904800.
[34] Health M. Ministry of Health, COVID-19. Accessed: October 2020. 2020. https://covid19.moh.gov.sa/.
[35] Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, Gomes AS, Folsom AR, Shea S, Guallar E, et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. Circ Res. 2017;121(9):1092–101.
[36] Adam P, Parveen A. Prediction system for heart disease using naïve bayes. J Adv Comput Math Sci. 2012;3(3):290–4.
[37] Tran H. A survey of machine learning and data mining techniques used in multimedia system. no 113 13–21 2019.
[38] Gnaneswar B, Jebarani ME. A review on prediction and diagnosis of heart failure. In 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 17-18 March, Coimbatore, India, 2017;1–3. https://doi.org/10.1109/ICIIECS.2017.8276033
[39] Kusprasapta M, Ichwan M, Utami DB. Heart rate prediction based on cycling cadence using feedforward neural network. In 2016 International Conference on Computer, Control, Informatics and its Applications (IC3INA), IEEE, 2016;72–76. https://doi.org/10.1109/IC3INA.2016.7863026
[40] Singh KY, Sinha N, Singh KS. Heart disease prediction system using random forest. In International Conference on Advances in Computing and Data Sciences, Advances in Computing and Data Sciences. ICACDS 2016. Communi- cations in Computer and Information Science, Singapore. 2017;721:613–623. https://doi.org/10.1007/978-981-10- 5427-3_63
[41] Priya RP, SKinariwala A. Automated diagnosis of heart disease using random forest algorithm. Int J Adv Res Ideas Innovat Technol 2017;3(2).
[42] Tripoliti E, Fotiadis ID, Manis G. Automated diagnosis of diseases based on classification: dynamic determination of the number of trees in random forests algorithm. EEE Trans Inf Technol Biomed 2012;16(4).
[43] Gonsalves AH, Thabtah F, Mohammad RMA, Singh G. Prediction of coronary heart disease using machine learning: an experimental analysis. In: Proceedings of the 2019 3rd International Conference on Deep Learning Technologies, 2019;51–56.
[44] Oikonomou EK, Williams MC, Kotanidis CP, Desai MY, Marwan M, Antonopoulos AS, Thomas KE, Thomas S, Akoumi- anakis I, Fan LM, et al. A novel machine learning-derived radiotranscriptomic signature of perivascular fat improves cardiac risk prediction using coronary ct angiography. Eur Heart J. 2019;40(43):3529–43.
[45] El-Hasnony IM, Elzeki OM. Multi-label active learning-based machine learning model for heart disease prediction. Sensors. 2022;22(3):1184–8. https://doi.org/10.3390/s22031184.
[46] Guleria P, Srinivasu PN, Ahmed S. Ai framework for cardiovascular disease prediction using classification techniques. Electronics. 2022;11(24):1184–8. https://doi.org/10.3390/electronics11244086.
[47] Javaid A, Zghyer F, Kim C, Spaulding EM, Isakadze N, Ding J, Kargillis D, Gao Y, Rahman F, Brown DE, et al. Medicine 2032: the future of cardiovascular disease prevention with machine learning and digital health technology. Am J Prevent Cardiol, 2022;100379
[48] Alaa AM, Bolton T, Di Angelantonio E, Rudd JH. Van der Schaar M. Cardiovascular disease risk prediction using auto- mated machine learning: a prospective study of 423,604 uk biobank participants. PloS One 2019;14(5):0213653.
[49] Chunhu Zhang DL. Xiaojian Shao: knowledge-based support vector classification based on c-svc. Proc Comput Sci. 2013;17:1083–90. https://doi.org/10.1016/j.procs.2013.05.137.
[50] Zhang Z. Introduction to machine learning: k-nearest neighbors. Ann Transl Med. 2016. https://doi.org/10.21037/ atm.2016.03.37.
[51] Mohandoss DP, Shi Y, Suo K. Outlier prediction using random forest classifier. In: 2021 IEEE 11th Annual

Computing and Communication Workshop and Conference (CCWC). IEEE. 2021. https://doi.org/10.1109/ccwc51732.2021.93760 77.
[52] Dimovski AS, Apel S, Legay A. A decision tree lifted domain for analyzing program families with numerical fea- tures. In: Fundamental Approaches to Software Engineering, pp. 67–86. Springer. 2021. https://doi.org/10.1007/ 978-3-030-71500-7_4.
[53] Fedesoriano: Heart Failure Prediction Dataset. Retrieved [Date Retrieved] from. Accessed: September 2021 (Septem- ber 2021). https://www.kaggle.com/fedesoriano/heart-failure-prediction.
[54] UCI: Heart Failure Prediction Dataset. UCI Machine Learning Repository. Accessed: September 2021 (September 2021). https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/.
[55] Castelli WP, Anderson K. A population at risk prevalence of high cholesterol levels in hypertensive patients in the Framingham study. Am J Med. 1986;80(2A):23–32. https://doi.org/10.1016/0002-9343(86)90157-9.
[56] Saini, Mukesh Kumar; Singh, Jaibir. "Big data analytics and machine learning: Personalized, predictive health and boost exactitude medicine research" In Big data analytics and machine learning, 2021 International Journal of Management IT and Engineering, pp. 47-56. International Journal of Management IT and Engineering 2021.
[57] Ghosh J, Shuvo SB. Improving classification model's performance using linear discriminant analysis on linear data. 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India 2019. https://doi.org/10.1109/ICCCNT45670.2019.8944632.
[58] Imon AHMR, Roy MC, Bhattacharj SK. Prediction of rainfall using logistic regression. Pak J Stat Oper Res. 2012. https:// doi.org/10.1234/pjsor.v8i3.535.
[59] Wessler BS, Paulus J, Lundquist CM, Ajlan M, Natto Z, Janes WA, Jethmalani N, Raman G, Lutz JS, Kent DM. Tufts pace clinical predictive model registry: update 1990 through 2015. Diagn Prognostic Res. 2017;1:1–8.
[60] Peter J, Somasundaram K. Study and development of novel feature selection framework for heart disease predic- tion. Int J Sci Res Publ. 2012;10(2):1–7.
[61] Shouman M, Turner T, Stocker R. Integrating decision tree and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients. Int J Inf Educ Technol. 2012;2(3):220–3.
[62] MS. RR. Heart disease prediction system using naive based and jelmeck mercer smoothing. IJARCCE 2014;3:6787–6792.
[63] Ferdousy EZ, Islam MM, Matin MA. ombination of naïve bayes classifier and k-nearest neighbor (cnk) in the classifica- tion based predictive models. Comput Inf Sci. 2013;6(3):48–56.
[64] N. DU. Prediction system for heart disease using naive bayes and particle swarm optimization. Biomedical Research- Tokyo 2018;29:2646–2649.

Author Profile

Dr. Mukesh Kumar Saini holds PhD, MBA, and MCA degrees in Computer Science and Information Technology, and brings over two decades of industry experience. He specializes in Artificial Intelligence, Machine Learning, Data Engineering and their applications in clinical settings. His research focuses on personalized healthcare, with an emphasis on clinical applications designed to optimize healthcare provider costs.