Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. ICMASD (2025)

Prediction of Cardiovascular Disease using Machine Learning Approach

Rohit Kumar

SOET,CT University,India

rohitraks583@gmail.com

Article History:

Received: 12-12-2024

Revised: 25-01-2025

Accepted: 05-02-2025

Abstract:

Cardiovascular disease (CVD) In order to improve patient outcomes and lower death rates, cardiovascular disease (CVD), which is still a major worldwide health problem, requires rapid detection. In contrast to previous studies that employed the standard Cleveland dataset, which only included 303 instances, this study examines the use of sophisticated machine learning (ML) models for early CVD prediction using a large-scale dataset of 1190 instances. By integrating K-Nearest Neighbors (KNN), Gradient Boosting (GB), and Decision Trees (DT) in a hybrid strategy, the suggested model achieves an amazing accuracy of 88.78%, with precision of 89%, recall of 93%, and F1 Score of 91%. This hybrid model fills in the gaps in earlier studies by combining several machine learning methods to improve clinical decision-making and dramatically increase prediction accuracy.

Furthermore, a user-friendly recommendation system based on Tkinter is created to help healthcare professionals recognize the risks of CVD and offer preventative measures according to each patient's unique risk profile. This study highlights the value of hybrid models in promoting early CVD detection and providing useful insights for medical practitioners by concentrating on readily available clinical data. The results highlight how important ML-driven tools might be in enhancing patient care, encouraging early treatments, and eventually saving lives. The importance of hybrid machine learning models in clinical practice is highlighted by this study, which makes a substantial addition to the field of healthcare decision-making.

Index Terms- Cardiovascular Disease, Health Care, Hybrid Method, Machine Learning, Risk Prediction, Recommendation System, GUI..

1. Introduction

1.1 Introduction

Heart disease, or cardiovascular diseases, is the world's largest cause of mortality, accounting for 17.9 million deaths yearly, or almost 32% of deaths worldwide [26]. Heart and blood vessel diseases (CVDs) encompass a variety of disorders, such as rheumatic heart disease, coronary heart disease, and cerebrovascular sickness. Regretfully, 80% of fatalities linked to CVD are caused by heart attacks and strokes., with one-third of these fatalities occurring in those under the age of 70 too soon. Nearly half of all fatalities from heart disease and stroke are caused by high blood pressure, which affects one in three persons worldwide, based on the 2012 World Health Statistics report. Every 34 seconds, someone

in the US dies due to heart disease alone [15], [16]. 18 million deaths worldwide each year are caused by cardiovascular illnesses [13], [14]. 32% of deaths are the result of delayed diagnosis and inadequate monitoring technologies. Models based on ML have the potential to transform the prediction of CVD. An ML-powered prediction system using SVM, RF, DT, KNN, NB, ANN, LR, and GB algorithms is presented in this paper. By accurately predicting the beginning of CVD, it strives to improve the treatment of Patients and advance the field's understanding of accurate and effective AI predictive models for cardiovascular health.

Accurate and timely detection of cardiovascular illnesses is essential for improving patient outcomes and saving lives. In the field of healthcare, machine learning (ML) has become a potent instrument that helps physicians diagnose patients more quickly and accurately while also allowing non-specialists to more quickly and accurately identify high-risk individuals. Though ML methods have demonstrated potential in the identification of cardiovascular illness, there are still significant shortcomings and a need for development. This project aims to apply machine learning to overcome some of the major problems with current methods for diagnosing cardiovascular disease. A significant constraint noted in the extant research is the dependence on a solitary dataset, namely the Cleveland dataset, or the use of restricted parameters while utilizing several datasets. Furthermore, rather than thoroughly investigating the possibilities of feature selection and extraction methods, current studies have mostly concentrated on improving the accuracy of machine learning classification approaches.

Furthermore, there is still a knowledge vacuum about the real implementation process, especially when it comes to creating system-based suggestions using tools like the Tkinter GUI, even with the enticing label of "early detection" in certain research. The development of more efficient early detection systems is hampered by this lack of clarity. This study suggests creating a sophisticated model that gives feature selection and extraction methods precedence over conventional ML techniques in order to close these gaps. By applying optimization methods, particularly within the framework of feature selection algorithms, the suggested model seeks to greatly improve performance and accuracy. Furthermore, adding system-based suggestion mechanisms will enhance the diagnostic process's usefulness and efficacy even further. This study emphasizes how critical it is to go beyond traditional machine learning techniques for the identification of cardiovascular illness. Through the integration of optimization approaches and system-based recommendations, the proposed model aims to improve the diagnostic systems' efficiency, accuracy, and usability in clinical settings.

1.2 Decision Strategies in Clinical Practice

Clinical decision-making is included in all patient care actions, which comprise acts from a variety of possibilities. Along with administering care and providing comfort, making decisions on sickness prevention, diagnosis, and treatment are critical skills for those in the medical field. Clinicians typically base their decisions on the condition of their patients. Sometimes, patients participate completely in the decision-making process.

Each of these decisions has an effect, and some of them will eventually determine whether a patient makes a full recovery and survives their illness. Heart problems are depicted in poor countries in Figure 1.1, show in this study how early clinical judgments help experts in reducing risk factors [39]. Medical professionals need to be aware of the data that is accessible, but they also need to know how to

appropriately incorporate that data into a decision-making process that protects the privacy and interests of their patients. Making clinical decisions is getting harder. The range of options for diagnosis and therapy is ever-expanding. Medical practitioners are under pressure to weigh value when deciding between various treatment options due to rising healthcare costs. Often, there is not enough time to make decisions, especially considering how long a normal patient visit is getting shorter. Heart issues in industrialized/developing nations are seen in Figure 1.2 [39].

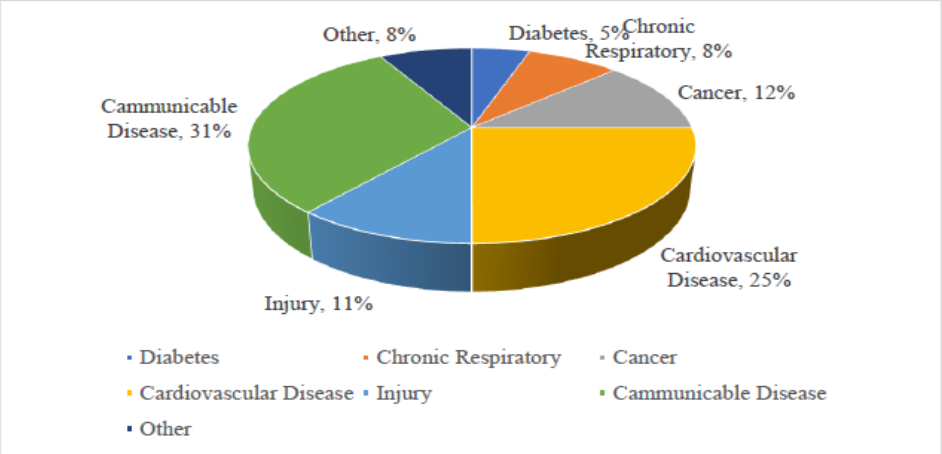


Figure 1.1 Causes of Death in Developing Countries [39]

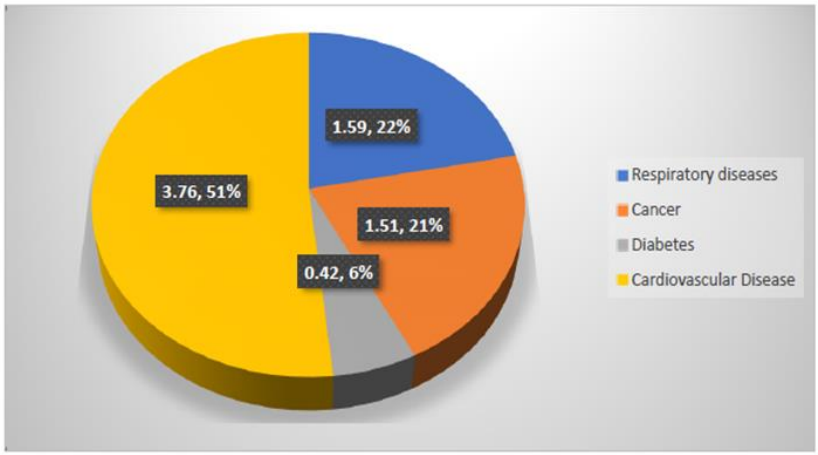


Figure 2 Causes of Death in Developing Countries [39]

The many clinical blood vascular dysfunctions are primarily the focus of the classification of heart ailments. The blood flow to the heart is influenced by fatty molecules found in blood vessels. An essential organ or part of our body is the heart.

Life depends on the heart's ability to function normally. If the heart isn't working right, it will affect other parts of the human body like the kidney, brain, etc. All the heart does is pump blood throughout the body. Problems with blood circulation can impact many organs, such as the brain, and result in death in a matter of minutes if the heart completely stops pumping. Life is completely dependent on the heart's effective operation. Coronary heart disease (CHD) is another name for heart disease.

2. Literature Review

In today's world, cardiovascular disease (CVD) remains the primary cause of global mortality, emphasizing the critical need for early detection methods to mitigate its profound impact. Recent years have witnessed substantial progress in harnessing machine learning (ML) to develop accurate and effective artificial intelligence models specifically tailored for cardiac disorders.This comprehensive study aims to delve into an extensive review of existing research that revolves around cardiovascular disease prediction systems, specifically leveraging ML-based approaches. By meticulously analyzing various studies and methodologies, the goal is to provide a comprehensive assessment of the current landscape in CVD prediction using machine learning techniques. One of the key focal points of this study involves conducting a comparative analysis.

This comparative analysis entails meticulously evaluating and juxtaposing the methodologies and outcomes of the proposed prediction systems against the backdrop of other cutting-edge approaches in the field. This comparative assessment aims to shed light on the strengths, weaknesses, and unique contributions of the different ML-based approaches employed in predicting cardiovascular diseases.

The intention behind this meticulous comparative analysis is to offer valuable insights into the efficacy, accuracy, and advancements brought forth by these ML-based models. By synthesizing and critically evaluating the findings of various studies, this research endeavours to paint a comprehensive picture of the current state of cardiovascular disease prediction systems driven by machine learning methodologies.

2.1 Comparative Analysis

Table 2.1: Comparative analysis of previous work

Author & Citation	Conference/ Journal	Algorithm	Dataset	Key Finding	Challenges
Prasanna et al.,2023[1]	IEEE Conference	Q-learning, a reinforcement learning algorithm	UCI (303)	Accuracy 87%, Precision 99%, Recall 58.3%, and F1-score 76%.	Improving recall, exploring additional features, and testing on larger datasets
Kavitha et al.,2021[2]	IEEE Conference	DT & RF and hybrid approach	UCI (303)	Accuracy 79% for DT, 81% for RF, and 88% for the Hybrid Model.	Perform better only for small dataset
Bhatt et al.,2023 [3]	MDPI Journal	RF, DT, MLP, XGB	Kaggle (70,000)	Accuracy 87.28%, Precision 88.70%, Recall 84.85%, and F1-score 86.71%, AUC 0.95(MLP)	Dimension issues, accuracy
Madhumita et al.,2021[11]	IOP Conference	RF	UCI (303)	Accuracy 86.9%, Sensitivity	Limited dataset

				90.6%, Specificity 82.7% (RF)	
Jindal et al.,2021 [12]	IOP Conference	LR, RF, KNN	UCI (303)	Accuracy 88.52% (KNN)	Limited dataset
Pal and Parija,2022 [13]	Open Medicine Journal	MLP, KNN	UCI (303)	Accuracy 82.47%, AUC 86.41 (MLP)	Limited dataset
Garg et al.,2021 [14]	IOP Conference	KNN, RF	UCI (303)	Accuracy 86.89% (KNN)	Limited dataset, testing done with less algorithm
Maiga and Hungilo ,2019[15]	IEEE Conference	LR, RF, KNN & GNB	Kaggle (70,000)	Accuracy 73%, Sensitivity 80%, Specificity 65 %	Limited features, accuracy can be improved
Shimaa Ouf,2021 [16]	Journal of Southwest Jiaotong University	LR, RF, KNN, SVM, LDA, LR, GNB & ANN	Kaggle (70,000) UCI(303)	Accuracy 71.82 % (ANN) with Kaggle dataset Accuracy 89.01 % (RF) with UCI Cleveland dataset	Performance difference due to dataset size

Author & Citation	**Conference/ Journal**	**Algorithm**	**Dataset**	**Key Finding**	**Challenges**
Theerthagiri and Vidya,2022 [17]	Expert syst Journal	GB with recursive features elimination	Kaggle (70,000)	Accuracy 89.7%, AUC 0.84(GB)	Less algorithms used for testing
Mahmud et al., 2023 [18]	IEEE Conference	SVM,KNN, LR,RF,DT & XGB	Kaggle (70,000)	Accuracy 84.03% (RF)	Feature Categorization can be improved
Boukhatem et al.,2022 [29]	ASET Conference	MLP,SVM, RF,NB	UCI (303)	Accuracy 91.67%, Precision 92.31%, Recall 88.89%, and F1-score 90.56%. (SVM)	Limited dataset
Amin et	IEEE	LR,DT,RF, SVM,KNN,	Kaggle	Accuracy 91.8%, Precision 92.5%,	Limited dataset

al.,2023 [22]	Conference	GNB	(918)	Recall 91.4%, F1-score 91.9% and AUC 90.27%(KNN)	

2.2 Research Gap

- **Restricted Dataset Utilization:** Because of its comprehensiveness and low number of missing values, the Cleveland dataset which has 303 entries and 14 features continues to be the major dataset utilized in research. It is frequently employed either by itself or in conjunction with a small number of variables from other datasets. But depending just on one dataset reduces the diversity and dependability needed for strong model building. The robustness and generalizability of the model should be improved by taking into account a wider variety of datasets. (Prasanna et al., 2023 [1]; Jindal et al., 2021 [12]; Kavitha et al., 2021 [2]; Garg et al., 2021 [14]; Shimaa Ouf, 2021 [16]).
- **Limited Focus on Feature Extraction and Selection:** Although a lot of research has been done on classification techniques like ensemble learning, Naive Bayes, Decision Trees, Support Vector Machines, Random Forests, and Artificial Neural Networks, not enough attention has been paid to feature extraction and selection techniques. These methods should receive more focus in future study as they are essential for maximizing model performance. (Mahmud et al., 2023 [18]; Garg et al., 2021 [14]; Jindal et al., 2021 [12]).
- **Underutilization of Optimization Techniques:** It has been shown that optimization techniques can improve model accuracy when used with classification techniques. Their use is still restricted, though. To increase model robustness and accuracy, future studies should investigate the hybridization or integration of optimization methods with machine learning algorithms. (Amin et al., 2023 [22]; Mahmud et al., 2023 [18]; Shimaa Ouf, 2021 [16]).
- **Insufficient Implementation Clarity for Early Detection Systems:** Many studies support early detection systems, but they frequently don't include thorough implementation plans, particularly when it comes to system-based recommendations that make use of the Tkinter GUI. Effective early detection system development is hampered by this ambiguity. To bridge this gap, sophisticated models that give priority to feature selection and extraction approaches above conventional machine learning methods are required. (Alyas et al., 2024 [38]; Amin et al., 2023 [22]; Obayya et al., 2023 [10]).

3. Research Design & Methodology

3.1 Statement of the Research Problem

The prevalence of cardiac illnesses is rising quickly, thus early detection is crucial. Even though there are many different prediction methods available, it can still be difficult to detect heart disease. These instruments need to be able to calculate accurately and effectively. Significantly lower death rates and complications can result from early identification of heart problems. However, continuous patient monitoring is frequently impracticable since it is not possible for a doctor to consult with patients around-the-clock due to the high time, skill, and resource requirements. Machine learning and other computer-aided methods provide a way to forecast patients more quickly and accurately while cutting

expenses dramatically. Machine learning is a vast and ever-growing area, especially in the healthcare industry. The wealth of data that is currently accessible allows for the use of a variety of machine learning algorithms to find hidden patterns in the data. Medical diagnosis can then be made using these patterns. The goal of this effort is to classify people according to their risk by using machine learning algorithms to assess patient data and forecast the possibility of future heart disease.

3.2 Objectives

- To evaluate the result of machine learning algorithms for cardiovascular disease prediction
- To design a hybrid approach that combines different machine learning techniques to predict cardiovascular diseases.
- To compare the results of proposed hybrid algorithm with existing algorithms
- To design and create a patient specific recommendation interface.

3.3 Scope of Analysis

This work contributes to the development of the categorization scheme that will most accurately identify cardiac disorders.

3.4 Methodology

The research methodology involved merging five cardiovascular disease datasets (Cleveland, Switzerland, Hungarian, Long Beach VA, and Stalog) based on 12 common attributes [23],[24]. Data preparation included cleansing, analysis, and visualization to rectify inconsistencies and uncover diagnostic patterns. Data encoding and scaling ensured compatibility with machine learning algorithms and balanced distribution. Machine learning algorithms were rigorously tested on the preprocessed data, with performance evaluations and hyperparameter tuning to optimize results. A hybrid algorithm approach was employed to enhance prediction accuracy and robustness. A graphical user interface (GUI) was developed to assist healthcare professionals by providing real-time risk assessments and diagnostic recommendations based on patient data. The objective was to identify the most efficient predictive algorithm and offer a practical clinical tool through the GUI. ***Figure 3.1*** provides an overview of the complete workflow for this proposed cardiovascular disease prediction model.

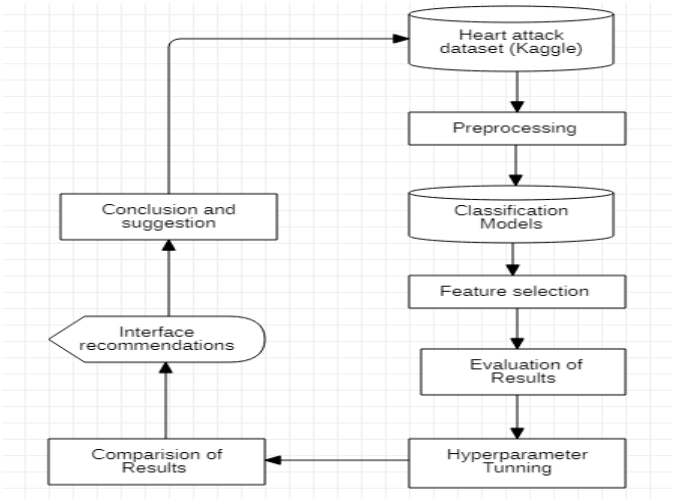


Figure 3.1: Workflow of cardiovascular disease prediction

3.4.1 Data Loading and Exploration:

The dataset was compiled by combining datasets from multiple sources, including Cleveland (303), Switzerland (123), Hungarian (294), Long Beach VA (200), and Stalog (270), resulting in a comprehensive dataset with 1190 instances [24]. Attributes include sex, age, chest pain type, cholesterol, RestingBP, Fasting BS, Maximum heart rate, resting ECG, ST slope, exercise angina, old peak, and target. The code starts by loading a dataset named "heart.csv" using Pandas. This dataset contains information about patients and whether they have heart disease. Information about the dataset is displayed using dataset.info () shown in figure 3.2. Correlation matrix and histograms are visualized to understand the relationship between variables and their distributions figure 3.3 and figure 3.4 respectively. Datasets details are illustrated in table1,

Table 3.1: Dataset Details

S.No	Attribute	Description	value
1	age	Patient's age in years	Value is continuous in range [29-77]
2	sex	Patient's sex	1 indicates male, 0 indicates female
3	cp	Type of chest pain	1->asymptomatic, 2-> atypical angina, 3->non-anginal pain, 4-> typical angina
4	trestbps	resting blood pressure (mm Hg on admission to the hospital)	Value is continuous in range [94-200]
5	chol	cholesterol measurement in mg/dl	Value is continuous in range [126-564]
6	fbs	fasting blood sugar	(> 120 mg/dl, 1 for true; 0 for false)
7	restecg	resting electrocardiographic results	0= normal, 1= having ST-T wave abnormality,2=left-ventricular hypertrophy
8	thalach	maximum heart rate achieved	Value is continuous in range [71-202] bpm
9	exang	Exercise induced angina	1 represents yes; 0 represents no
10	oldPeak	Depression in ST brought by exercise that is relative to rest	Value is continuous in range [0-6.2]
11	slope	slope of the peak exercise	1- upsloping, 2-flat, 3 - down sloping
12	target	Is the heart disease present	0=No, 1=Yes

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1190 entries, 0 to 1189
Data columns (total 12 columns):
 #    Column     Non-Null Count  Dtype
---   ------     --------------  -----
 0    age        1190 non-null   int64
 1    sex        1190 non-null   int64
 2    cp         1190 non-null   int64
 3    trestbps   1190 non-null   int64
 4    chol       1190 non-null   int64
 5    fbs        1190 non-null   int64
 6    restecg    1190 non-null   int64
 7    thalach    1190 non-null   int64
 8    exang      1190 non-null   int64
 9    oldpeak    1190 non-null   float64
 10   slope      1190 non-null   int64
 11   target     1190 non-null   int64
dtypes: float64(1), int64(11)
memory usage: 111.7 KB
```

Figure 3.2 dataset information

Eleven characteristics in the dataset are graphically represented by the heat map, where dark blue indicates weak or negative correlations and yellow indicates a significant positive association. Weaker relationships are represented by lighter blue hues. As seen in Figure 3.3, this map provides a clear overview of feature interdependencies and aids in rapidly determining whether features have high, weak, or no association.

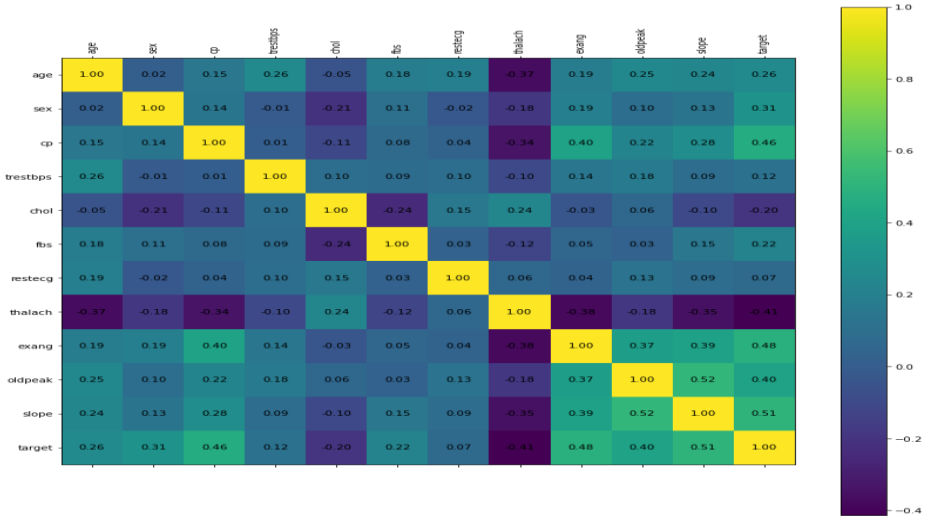


Figure 3.3 Correlation matrix or heat map

According to the statistics, there are more men than women and the majority of patients are between the ages of 50 and 60. kind 4 (classic angina) is the most prevalent kind of chest pain, followed by Types 3, 2, and 1. With a mean of 132 mm Hg, resting blood pressure (trestbps) is mostly between 125 and 135 mm Hg, while cholesterol levels (Chol) are typically 200 mg/dl. The majority of individuals do not have diabetes (fbs = 0), and the majority of resting ECG readings are 0, 2, and 1. More people do not have exercise-induced angina (exang = 0), and heart rate (thalach) normally falls between 120 and 140 bpm. In ST (oldPeak), the depression mostly ranges from -1 to 0. The target variable, which indicates the existence of heart disease, is spread almost uniformly.

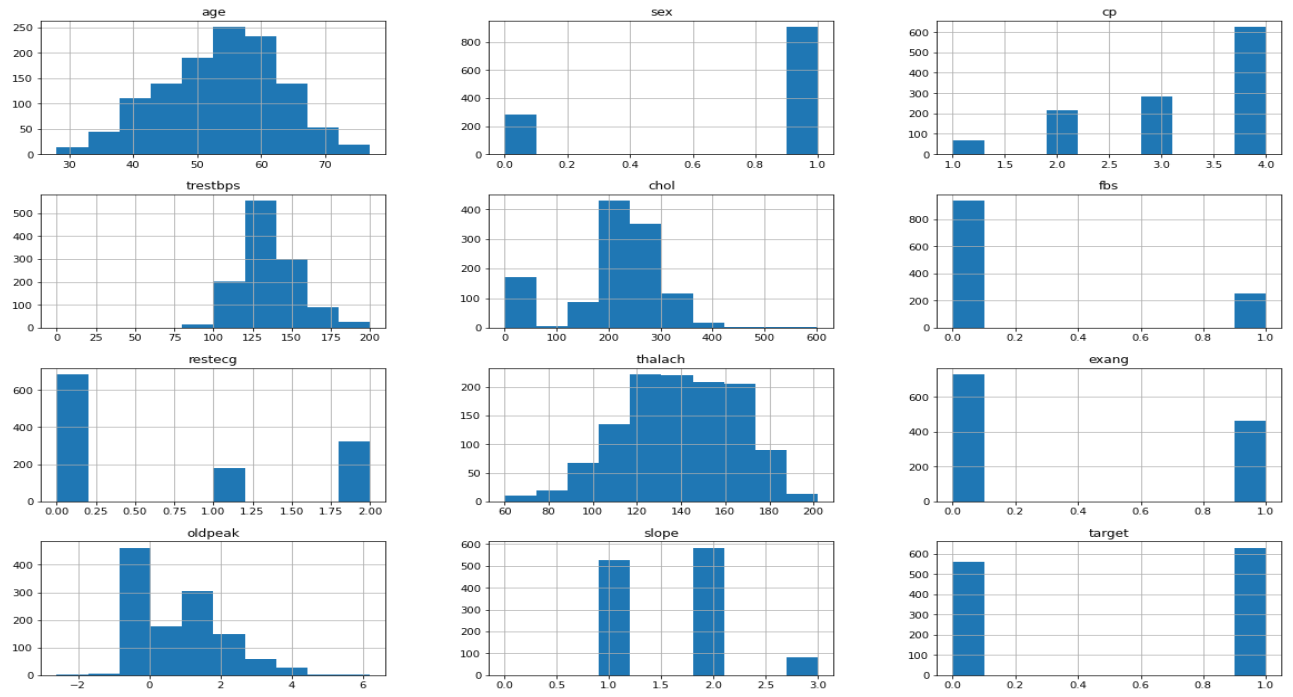


Figure 3.4 Histogram

3.4.2 Data Preprocessing:

Data cleaning, feature selection, and normalization techniques were applied to ensure data quality and consistency. Duplicate rows are checked for and removed if found. One-hot encoding is applied to categorical columns, and standardization is performed on numerical columns using StandardScaler [25].

3.4.3 Data Splitting:

The `train_test_split` function was used to divide the dataset into features (X) and target variable (y), with 70% going toward training and 30% toward testing in order to evaluate the correctness of the model [50]. Additionally, a 67:33 split ratio was investigated to enhance model performance and training data size. Several methods were used, and the accuracy of the results was assessed using a confusion matrix. To examine important distributions and trends, the dataset was also converted into histograms.

3.4.4 Verifying Data Distribution

In the dataset on cardiac illness, 47.14% of patients test positive for the condition, whereas 52.85% of patients are clear of it.The dataset may need to be balanced in order to avoid overfitting. The distribution is depicted in Figure 3, where 629 people have heart disease (1) and 561 people do not (0). This distribution aids in the identification of pertinent patterns for prediction using machine learning algorithms.

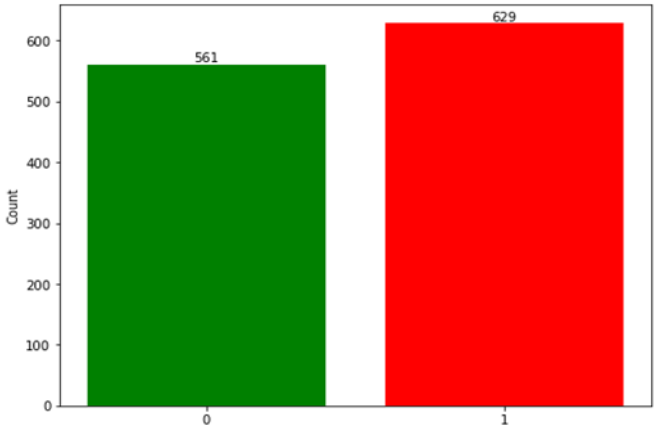


Figure 4.5 Distribution of data (1: Positive, 0: Negative)

3.4.5 Exploratory Data Assessment

This study examined the distribution and correlations between components and the goal variable, using statistical analysis to give important insights. Figure 3.6 provides a summary of the cardiovascular disease dataset, which comprises 12 variables with information on count, mean, standard deviation, and range. There are 561 people with heart disease and 629 people without heart disease in the sample.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	target
count	1190	1190	1190	1190	1190	1190	1190	1190	1190	1190	1190	1190
mean	53.720168	0.763866	3.232773	132.153782	210.363866	0.213445	0.698319	139.732773	0.387395	0.922773	1.624370	0.528571
std	9.358203	0.424884	0.935480	18.368823	101.420489	0.409912	0.870359	25.517636	0.487360	1.086337	0.610459	0.499393
min	28	0	1	0	0	0	0	60	0	-2.6	0	0
25%	47	1	3	120	188	0	0	121	0	0	1	0
50%	54	1	4	130	229	0	0	140.5	0	0.6	2	1
75%	60	1	4	140	269.75	0	2	160	1	1.6	2	1
max	77	1	4	200	603	1	2	202	1	6.2	3	1

Figure 3.6 Analysis of data of all numeric parameters in the dataset

3.4.6 Model Initialization and Training:

The test data (X_test, y_test) is used to assess the accuracy of a number of classifiers, such as KNN, SVM, Decision Tree, Random Forest, Naive Bayes, ANN, Logistic Regression, and Gradient Boosting, which are trained on the training data (X_train, y_train) [35].

3.4.7 Hybrid Model Creation and Evaluation:

KNN, Random Forest, and Gradient Boosting are all combined in a hybrid model that uses a Voting Classifier. Using the test data, it is trained and its accuracy assessed.

3.4.8 Hyperparameter Tuning and Evaluation:

Hyperparameter tuning is performed for each classifier using GridSearchCV. Grid search is applied to find the best combination of hyperparameters that maximizes accuracy on the validation set.Results before and after tuning are stored in dictionaries for comparison [42].

3.4.9 Performance Metrics:

Recall (True Positive Rate): Recall measures how well the model detects actual positive cases. In medical diagnosis, it's critical to prevent overlooking individuals that require care.

Recall = TP / (TP + FN)

TP: True Positives, FN: False Negatives

F1 Score: The F1 score balances precision and recall, especially useful for imbalanced datasets. In medical diagnosis, it's critical to prevent overlooking individuals that require care.

$$F1\ Score = 2 \times (Precision \times Recall) / (Precision + Recall)$$

Precision (Positive Predictive Value): The fraction of true positive predictions among all positive predictions

Precision = TP / (TP + FP)

4. Result Analysis

According to this study, the best model for predicting cardiovascular disease (CVD) was a hybrid machine learning approach shown in fig 4.1, that used K-Nearest Neighbors (KNN), Decision Trees (DT), and Gradient Boosting (GB). The hybrid model outperformed other models and became the best performer for CVD prediction with an astounding accuracy of 88.78%.KNN outperformed Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting (GB), all of which had accuracies above 85%, with the best accuracy of any standalone method at 87.46%. The Naive Bayes (NB) method was the least successful, with an accuracy of 77.56%, while Decision Tree (DT), Artificial Neural Networks (ANN), and Logistic Regression (LR) demonstrated intermediate performance with accuracies of about 84%.The pre-processing stage eliminated 272 duplicates from the 1,190 occurrences that made up the dataset utilized in this study. In order to handle the dataset's complexity, eight different machine learning techniques were evaluated. With five neighbors, KNN yielded the best accuracy. But in addition to having the highest accuracy, the hybrid model which combines KNN, DT, and GB also performed exceptionally well in important performance criteria, such as 89% precision, 93% recall, and 91% F1 score, can see the differences in Table 4. In addition to its excellent accuracy, this hybrid model was chosen because it can adapt to datasets with complicated global patterns, which are best suited for DT and GB, and local structures, which are best suited for KNN. A graphical user interface (GUI), which simplifies implementation and offers tailored health risk recommendations based on vital health data, was created to increase the system's usefulness. The overall efficacy and accuracy of the CVD prediction system were greatly increased by this user-friendly interface, which also provided trustworthy, customized information for those looking to evaluate and reduce their cardiovascular risks. Strong machine learning models and an easy-to-use graphical user interface (GUI) combine to improve the accuracy, accessibility, and actionability of

CVD prediction, which helps with early identification and individualized treatment plans., as seen in Figure 4.2

In terms of test accuracy (0.89), precision (0.89), recall (0.93), and F1 score (0.91), hybrid models (KNN + DT + GB) fared better than individual models (GB and KNN), indicating improved generalization and dependability. Individual models performed poorly on tests due to overfitting, even though they had good training accuracy (e.g., Decision Tree: 0.84, Random Forest: 0.86). In addition to reducing overfitting, hybrid models are excellent at reducing false positives, enhancing recall, and guaranteeing balanced performance. They are a more dependable and scalable option for real-world medical diagnostics because of their resilience to noise and dataset imbalance.

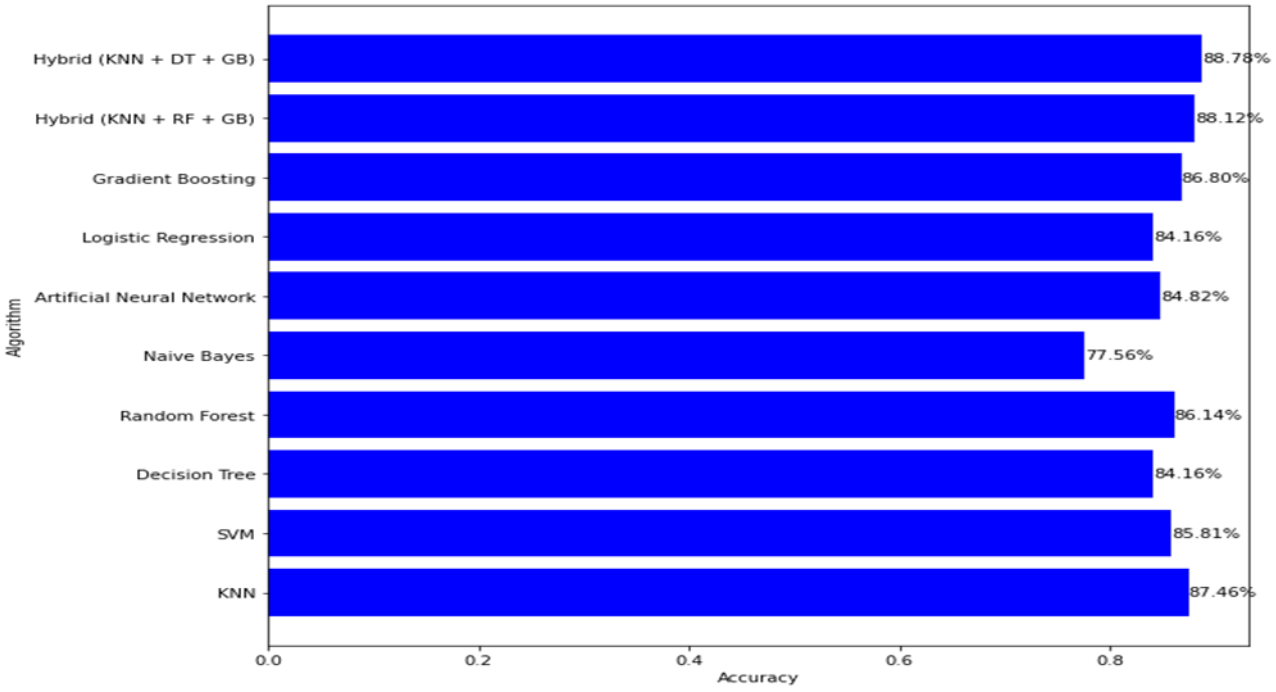
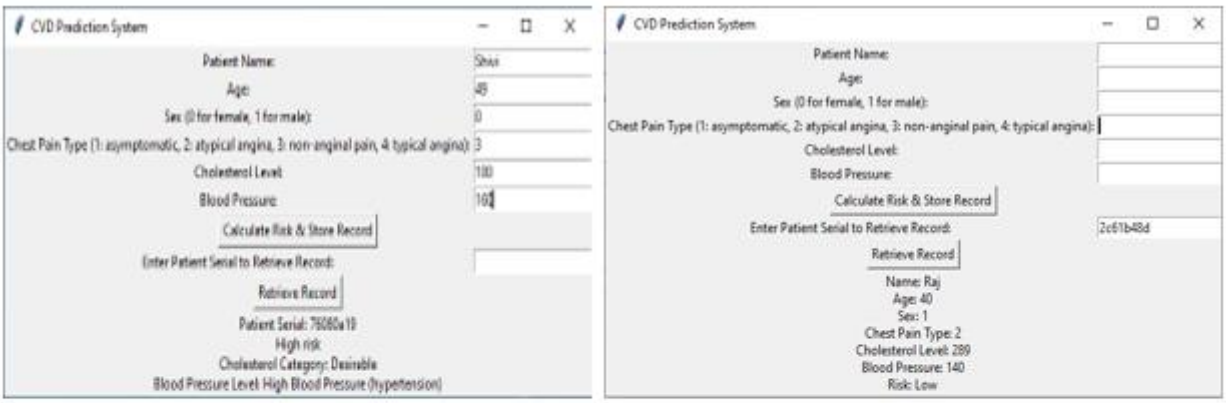


Figure 4.1 Accuracy Score of All algorithm including Hybrid Combination

Table 4: Performance Matrices with training and testing Accuracy

Model	Train Accuracy	Test Accuracy	Train Precision	Test Precision	Train Recall	Test Recall	Train F1 Score	Test F1 Score
KNN	0.89	0.87	0.89	0.88	0.92	0.91	0.90	0.89
SVM	0.91	0.86	0.90	0.85	0.94	0.91	0.92	0.88
DT	1.00	0.84	1.00	0.86	1.00	0.86	1.00	0.86
RF	1.00	0.86	1.00	0.86	1.00	0.91	1.00	0.88

NB	0.79	0.78	0.93	0.92	0.67	0.67	0.77	0.77
ANN	1.00	0.85	1.00	0.86	1.00	0.87	1.00	0.87
LR	0.87	0.84	0.88	0.84	0.89	0.90	0.88	0.87
GB	0.96	0.87	0.96	0.87	0.97	0.90	0.96	0.89
Hybrid (KNN + RF + GB)	0.96	0.88	0.96	0.88	0.97	0.93	0.97	0.90
Hybrid (KNN + DT + GB)	0.96	0.89	0.96	0.89	0.97	0.93	0.97	0.91



	Patient Serial	Name	Age	Sex	Chest Pain Type	Cholesterol Level	Blood Pressure	Risk
0	2c61b48d	Raj	40	1	2	289	140	Low

	Patient Serial	Name	Age	Sex	Chest Pain Type	Cholesterol Level	Blood Pressure	Risk
0	76080a19	Shivi	49	0	3	180	160	High

Figure 4.2 Patient Specific Interface Using Tkinter GUI

5. Conclusion:

This study underscores the transformative potential of hybrid machine learning (ML) models in early cardiovascular disease (CVD) prediction, demonstrating how AI can assist medical practitioners in making more accurate, timely decisions. By integrating K-Nearest Neighbors (KNN), Decision Trees (DT), and Gradient Boosting (GB) into a hybrid model, we achieved an impressive accuracy of 88.78%, with 89% precision, 93% recall, and a 91% F1 score, outperforming traditional individual models. Hybrid models, particularly KNN + DT + GB, not only reduce overfitting but also enhance generalization, minimize false positives, and improve clinical outcomes. The integration of a user-

friendly Tkinter-based interface further supports healthcare professionals by offering personalized CVD risk assessments, making this system both scalable and accessible. This research highlights the immense value of ML-driven tools in refining clinical decision-making, advancing early CVD detection, and ultimately saving lives, establishing hybrid models as a reliable, efficient solution for modern medical diagnostics.

References

[1] Prasanna, K. S. L., Challa, N. P., & Nagaraju, J. (2023). Heart Disease Prediction using Reinforcement Learning Technique. 2023 3rd International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies, ICAECT 2023. https://doi.org/10.1109/ICAECT57570.2023.10118232

[2] Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., & Suraj, R. S. (2021). Heart Disease Prediction using Hybrid machine Learning Model. Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021, 1329–1333. https://doi.org/10.1109/ICICT50816.2021.9358597

[3] Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms, 16(2). https://doi.org/10.3390/a16020088

[4] Alty, S. R., Millasseau, S. C., Chowienczyk, P. J., & Jakobsson, A. (2003). CARDIOVASCULAR DISEASE PREDICTION USING SUPPORT VECTOR MACHINES. Midwest Symposium on Circuits and Systems, 1, 376–379. https://doi.org/10.1109/MWSCAS.2003.1562297

[5] Nikam, A., Bhandari, S., Mhaske, A., & Mantri, S. (2020). Cardiovascular Disease Prediction Using Machine Learning Models. 2020 IEEE Pune Section International Conference, PuneCon 2020, 22–27. https://doi.org/10.1109/PuneCon50868.2020.9362367

[6] R. J. P. Princy, S. Parthasarathy, P. S. Hency Jose, A. Raj Lakshminarayanan, and S. Jeganathan, "Prediction of Cardiac Disease using Supervised Machine Learning Algorithms," in Proceedings of the 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 570-575. https://doi.org/10.1109/ICICCS48265.2020.9121169

[7] Q. He, A. Maag, and A. Elchouemi, "Heart disease monitoring and predicting by using machine learning based on IoT technology," in Proceedings of the 2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA), Sydney, Australia, 2020, pp. 1-10. https://doi.org/10.1109/CITISIA50690.2020.9371772

[8] M. M. El-Gayar, E. M. El-Daydamony, and S. E. A. Ashri, "HDPF: Heart Disease Prediction Framework Based on Hybrid Classifiers and Genetic Algorithm," IEEE Access, vol. 9, pp. 146797-146809, 2021. https://doi.org/10.1109/ACCESS.2021.3071109

[9] D. Sivabalaselvamani, D. Selvakarthi, L. Rahunathan, S. Nandhini Eswari, M. Pavithraa, and M. Sridhar, "Investigation On Heart Disease Using Machine Learning Algorithms," in Proceedings of the International Conference on Computer Communication Informatics (ICCCI), 2021.

[10] M. Obayya, J. M. Alsamri, M. A. Al-Hagery, A. Mohammed, and M. A. Hamza, "Automated Cardiovascular Disease Diagnosis Using Honey Badger Optimization with Modified Deep

Learning Model," IEEE Access, vol. 11, pp. 64272-64281, 2023. https://doi.org/10.1109/ACCESS.2023.3271342

[11] M. Pal and S. Parija, "Prediction of heart diseases using random forest," Journal of Physics: Conference Series, vol. 1817, no. 1, Mar. 2021, Art. no. 012009. https://doi.org/10.1088/1742-6596/1817/1/012009

[12] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," in Proceedings of the IOP Conference Series: Materials Science and Engineering, vol. 1022, 2021.

[13] M. Pal, S. Parija, G. Panda, K. Dhama, and R.K. Mohapatra, "Risk prediction of cardiovascular disease using machine learning classifiers," Open Medicine, vol. 17, no. 1, pp. 1100–1113, Jun. 2022. https://doi.org/10.1515/med-2022-0175

[14] A. Garg, B. Sharma, and R. Khan, "Heart disease prediction using machine learning techniques," in Proceedings of the IOP Conference Series: Materials Science and Engineering, vol. 1022, 2021, Art. no. 012046. https://doi.org/10.1088/1757-899X/1022/1/012046

[15] J. Maiga, G. G. Hungilo, and Pranowo, "Comparison of machine learning models in prediction of cardiovascular disease using health record data," in Proceedings of the International Conference on Information, Multimedia, Cyber, and Information Systems (ICIMCIS), Oct. 2019, pp. 45–48.

[16] S. Ouf and A. I. B. ElSeddawy, "A proposed paradigm for intelligent heart disease prediction system using data mining techniques," Journal of Southwest Jiaotong University, vol. 56, no. 4, pp. 220–240, Aug. 2021. https://doi.org/10.35741/issn.0258-2724.56.4.22

[17] P. Theerthagiri and J. Vidya, "Cardiovascular disease prediction using recursive feature elimination and gradient boosting classification techniques," Expert Systems, vol. 39, no. 9, Nov. 2022, Art. no. e13064. https://doi.org/10.1111/exsy.13064

[18] T. Mahmud, A. Barua, M. Begum, E. Chakma, S. Das, and N. Sharmen, "An improved framework for reliable cardiovascular disease prediction using hybrid ensemble learning," in Proceedings of the International Conference on Electrical, Computer, and Communication Engineering (ECCE), Feb. 2023, pp. 1–6.

[19] N. Khateeb and M. Usman, "Efficient heart disease prediction system using k-nearest neighbor classification technique," in Proceedings of the International Conference on Big Data and Internet of Thing (BDIOT), New York, NY, USA: ACM, 2017, pp. 21–26. https://doi.org/10.1145/3152424.3152461

[20] C. Beulah Christalin Latha and S. Carolin Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Informatics in Medicine Unlocked, vol. 16, p. 100203, 2019. https://doi.org/10.1016/j.imu.2019.100203

[21] K. Shah, K. Sharma, and D. Saxena, "Editorial: Health technology assessment in cardiovascular diseases," Frontiers in Cardiovascular Medicine, vol. 10, Jan. 2023, Art. no. 1108503. https://doi.org/10.3389/fcvm.2023.1108503

[22] A. Amin et al., "Enhancing Prognosis Accuracy for Ischemic Cardiovascular Disease Using K Nearest Neighbor Algorithm: A Robust Approach," IEEE Access, vol. 11, pp. 97879-97895, 2023. https://doi.org/10.1109/ACCESS.2023.3312046

[23] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "UCI machine learning repository-heart disease data set," School of Information and Computer Science, University of California, Irvine, CA, USA, 1988. https://doi.org/10.24432/C52P4X

[24] S. Ulianova, "Cardiovascular Disease Dataset Kaggle," 2019. [Online]. Available: https://www.kaggle.com/datasets/sulianova/cardiovasculardisease-dataset

[25] "Sklearn.Preprocessing. LabelencoderScikitLearn1.2.2Documentation," Scikit-Learn, 2007. [Online].Available:https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

[26] World Health Organization, "Cardiovascular Diseases," Modified 2015. [Online]. Available: http://www.who.int/classifications/icd

[27] A. Srivastava and A. K. Singh, "Heart disease prediction using machine learning," in Proceedings of the 2nd International Conference on Advanced Computing, Innovations, and Technologies in Engineering (ICACITE), Apr. 2022, pp. 2633–2635.

[28] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, Logistic Regression, Cham, Switzerland: Springer, 2002.

[29] C. Boukhatem, H. Y. Youssef, and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," in Proceedings of the 2022 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, pp. 1-6.

[30] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," Journal of Chemometrics, vol. 18, no. 6, pp. 275–285, Jun. 2004. https://doi.org/10.1002/cem.862

[31] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, Oct. 2001. https://doi.org/10.1023/A:1010933404324

[32] W. S. Noble, "What is a support vector machine?" Nature Biotechnology, vol. 24, no. 12, pp. 1565–1567, 2006. https://doi.org/10.1038/nbt1206-1565

[33] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," IEEE Access, vol. 7, pp. 81542-81554, 2019. https://doi.org/10.1109/ACCESS.2019.2924677

[34] Peterson, "K-nearest neighbor," Scholarpedia, vol. 4, no. 2, p. 1883, 2009. https://doi.org/10.4249/scholarpedia.1883

[35] J. Tan, J. Yang, S. Wu, G. Chen, and J. Zhao, "A critical look at the current train/test split in machine learning," 2021, arXiv:2106.04525. [Online]. Available: https://arxiv.org/abs/2106.04525

[36] R. Kumar, P. Thakur, and S. Chauhan, "Special Disease Prediction System Using Machine Learning," in Proceedings of the 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), Faridabad, India, pp. 42-45.

[37] N. K. Kumar, A. Kalyan Kumar, G. Thorani, L. Sahithi, and P. Pujitha, "Improving Cardiovascular Disease Prediction: Machine Learning and Cross-Fold Validation," in Proceedings of the 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), Gwalior, India, 2024, pp. 1-6, https://doi.org/10.1109/IATMSI60426.2024.10503072

[38] S. U. Warsi, S. Mohsin, M. Asif, A. Hassan, R. Khan, and T. Alyas, "A Hybrid Approach for Heart Disease Prediction using Genetic Algorithm and SVM," in Proceedings of the 2024 5th International Conference on Advancements in Computational Sciences (ICACS), Lahore, Pakistan, 2024, pp. 1-6, https://doi.org/10.1109/ICACS60934.2024.10473308

[39] World Health Organization, "Leading causes of death in the developing countries," Modified 2015. [Online]. Available: http://www.who.int/classifications/icd

[40] IBM, "What is big data?" Accessed: Jun. 20, 2024. [Online]. Available: https://www.ibm.com/analytics/hub/big-data-analytics

[41] Tkinter. [Online]. Available: https://docs.python.org/3/library/tkinter.html

[42] scikit-learn,"GridSearchCV Documentation." https://scikitlearn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[43] scikit-learn, "VotingClassifier Documentation." [Online]. Available: https://scikit learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html

[44] A. Shestel, "Python in Healthcare," Belitsoft, Mar. 18, 2021, https://belitsoft.com/custom-application-development-services/healthcare-software-development/python-healthcare

[45] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," Neurocomputing, vol. 415, pp. 295–316, Nov. 2020. https://doi.org/10.1016/j.neucom.2020.06.048

[46] K. P. Murphy, "Naive Bayes classifiers," University of British Columbia, vol. 18, no. 60, pp. 1–8, 2006.

[47] S. R. Fardibha, P. H. Basha, and V. Sriharsha, "Prediction of Heart Diseases using Advanced Learning and Data Analytics Approach," in Proceedings of the 2024 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, pp. 460-465.

[48] A. Jabbar, S. Naseem, T. Mahmood, T. Saba, F. S. Alamri, and A. Rehman, "Brain tumor detection and multi-grade segmentation through hybrid caps-VGGNet model," IEEE Access, vol. 1, pp. 72518–72536, 2023. https://doi.org/10.1109/ACCESS.2023.3314697

[49] N. S, V. K, I. B, and J. N. Kalshetty, "Heart Disease Prediction Using Artificial Intelligence Ensemble Network," in Proceedings of the 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, pp. 1-6.

[50] Z. Reitermanova, "Data splitting," in Proceedings of the 2010 10th Annual International Conference on World Wide Web Applications, vol. 10, Prague, Czechia: Matfyzpress Prague, pp. 31–36.

[51] A. Sharma, B. Gupta, and S. Singh, "Comparative Analysis of Machine Learning Techniques for Heart Disease Prediction," in Proceedings of the 2023 International Conference on Computational Intelligence and Sustainable Technologies (ICCIST), Jaipur, India, pp. 1-6.

[52] R. Menon, S. Agarwal, and P. Patel, "Early Detection of Cardiovascular Diseases using Deep Learning Models," in Proceedings of the 2022 IEEE International Conference on Innovations in Information Technology (ICIIT), Dubai, UAE, pp. 125-130.

[53] S. Verma, R. Kumar, and V. Singh, "Heart Disease Risk Assessment Using Convolutional Neural Networks," in Proceedings of the 2023 4th International Conference on Advances in Computational Techniques (ICACT), Varanasi, India, pp. 45-50.

[54] B. Mathur, A. Jain, and S. Mishra, "Predictive Modeling of Cardiac Health Parameters using Machine Learning Algorithms," in Proceedings of the 2022 IEEE International Symposium on Sustainable Systems and Technology (ISSST), Bangalore, India, pp. 1-5.

[55] A. Tiwari, S. Gupta, and N. Sharma, "Heart Disease Prediction Based on Genetic Algorithm and Support Vector Machine," in Proceedings of the 2023 5th International Conference on Computing, Communication and Security (ICCCS), Chandigarh, India, pp. 78-83.

[56] S. Kumar, P. Singh, and R. Gupta, "Enhancing Cardiovascular Disease Prediction Accuracy Using Ensemble Learning Techniques," in Proceedings of the 2023 International Conference on Machine Learning, Big Data, and Artificial Intelligence (ICMLBDAI), New Delhi, India, pp. 1-8.