A blue and white logo

AI-generated content may be incorrect.

**Hybrid Machine Learning Models for Multi-Factor Cardiovascular Risk Prediction**

Submitted in Partial fulfillment for the award of

**Post Graduate Degree Of**

**MASTER OF TECHNOLOGY**  
**IN**  
**DATA SCIENCE ENGINEERING**

**Submitted By**

**Durgesh Yadav**

0228IT23MT03

**Under the Guidance of**

**Dr. Sapna Jain Choudhary**  
 Data sci. & Eng. Dept. (HOD)  
 SRGI, Jabalpur

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**  
**SHRI RAM GROUP OF INSTITUTIONS, JABALPUR (M.P.)**  
**RAJIV GANDHI PROUDYOGIKI VISHWAVIDYALAYA, BHOPAL (M.P.)**

**Hybrid Machine Learning Models for Multi-Factor Cardiovascular Risk Prediction**  
 **Submitted in Partial fulfillment for the award of**

**Post Graduate Degree of  
Master of Technology  
in  
Data Science Engineering**

A blue and white logo

AI-generated content may be incorrect.

**Submitted By  
Durgesh Yadav**  
 0228IT23MT03

**Under the Guidance of**

**Dr. Sapna Jain Choudhary**  
Data sci. & Eng. Dept. (HOD)  
 SRGI, Jabalpur

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING  
Shri Ram Group of Institutions, Jabalpur (M.P.) (Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal)**

A blue and white logo

AI-generated content may be incorrect.

**SHRI RAM GROUP OF INSTITUTIONS, JABALPUR**  
Approved by AICTE New Delhi & Govt. of M.P.

Affiliated to  
**Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal**  
(University of Technology of Madhya Pradesh)

**PLAGIARISM CERTIFICATE**

This is to certify that the dissertation report entitled **“Hybrid Machine Learning Models for Multi-Factor Cardiovascular Risk Prediction”** is submitted by me in the academic year 2025-2026 to the Department of Computer Science and Engineering in partial fulfillment of the requirements for the award of the Post Graduate Degree of Master of Technology under the supervision of **Dr. Sapna Jain Choudhary**.

I also declare that a check for plagiarism has been carried out on the above dissertation and the content found within the acceptable limit and report of which is enclosed herewith.

**Durgesh Yadav**  
0228IT23MT03

**Dr. Sapna Jain Choudhary Dr. Avinash Gaur**  
Data sci. & Eng. Dept. (HOD) Principal SRGI

A blue and white logo

AI-generated content may be incorrect.**SHRI RAM GROUP OF INSTITUTIONS, JABALPUR**  
(Approved by AICTE, New Delhi, and Affiliated to Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal (Madhya Pradesh))  
**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**CERTIFICATE**

This is to certify that the dissertation entitled **" Hybrid Machine Learning Models for Multi-Factor Cardiovascular Risk Prediction "** being submitted by **DURGESH YADAV** (**Enroll. No.-0228IT23MT03**) in partial fulfillment of the requirement for the award of M. Tech degree in Data Science Engineering Branch from Shri Ram Group of Institutions, Jabalpur (M.P.) is a record of work done by her under my guidance.

**Dr. Sapna Jain Choudhary Dr. Avinash Gaur**  
Data sci. & Eng. Dept. (HOD) Principal   
SRGI, Jabalpur SRGI, Jabalpur

A blue and white logo

AI-generated content may be incorrect.**SHRI RAM GROUP OF INSTITUTIONS, JABALPUR**  
(Approved by AICTE, New Delhi, and Affiliated to Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal (Madhya Pradesh))  
**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**CANDIDATE DECLARATION**

This is to certify that the dissertation report entitled **“Hybrid Machine Learning Models for Multi-Factor Cardiovascular Risk Prediction”** is done by me is an authentic work carried out for the partial fulfillment of the requirements for the award of the degree for **M.Tech. in Data Science Engineering Branch** under the guidance of **Dr. Sapna Jain Choudhary**. The matter embodied in this dissertation work has not been submitted earlier for award of any other degree to the best of my knowledge and belief.

I also declare that **“A check for Plagiarism has been carried out on the thesis project report dissertation and is found within the acceptable limit and report of which is enclosed herewith.”**

**DATE:-…………………..……**

**Name:-** DURGESH YADAV  
**Enrollment No.**:- 0228IT23MT23

**CERTIFICATE**

This is to certify that the above statement made by candidate is correct to the best of knowledge.

**Dr. Sapna Jain Choudhary Dr. Avinash Gaur**  
Data sci. & Eng. Dept. (HOD) Principal   
SRGI, Jabalpur SRGI, Jabalpur

A blue and white logo

AI-generated content may be incorrect.

**SHRI RAM GROUP OF INSTITUTIONS, JABALPUR**  
Approved by AICTE, New Delhi & Govt. of M.P.  
Affiliated to Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal  
(University of Technology of Madhya Pradesh)

**APPROVAL CERTIFICATE**

This dissertation work entitled **“Hybrid Machine Learning Models for Multi-Factor Cardiovascular Risk Prediction”**, submitted by **Durgesh Yadav** (Enrolment No. **0228IT23MT03**) is approved for the award of degree of **Master of Technology** in **Data Science & Engineering**.

**INTERNAL EXAMINER EXTERNAL EXAMINER**

**प्रपत्र**

(**एम.ई / एम.टेक** )

**छात्र का नाम**:-………………………………………………..   
**पिता का नाम**:-……………………………………………….  
**आयु**:-……………………………………  
**निवासी**:-………………………………………

यह प्रमाणित किया जाता है कि उपर्युक्त छात्र ने …………………………………शाखा में .……………..सत्र के दौरान ……………………………………………………………से एम.टेक. की पढ़ाई की है।

उसने …………………………………………………………………………………………………..विषय पर ………………………………………………………….के अंतर्गत दिनांक ………………………………को सफलतापूर्वक अनुसंधान कार्य प्रस्तुत किया है।

यह प्रमाणित किया जाता है कि प्रस्तुत अनुसंधान कार्य छात्र द्वारा स्वयं किया गया है एवं यह पूर्व में किसी अन्य डिग्री हेतु प्रस्तुत नहीं किया गया है।

**छात्र का हस्ताक्षर**  
……………………………………………

**अनुक्रमांक**: ……………………………….………

**मार्गदर्शक के हस्ताक्षर**  
…………………………………………..

दिनांक: ………………………………

**संस्थान प्रमुख / प्राचार्य के हस्ताक्षर**  
……………………………………………..

………………………………………………

दिनांक: …………………………………………….

###### **ACKNOWLEDGEMENT**

I intended my gratitude to **Dr. Saurabh Baderia**, Chairman, Baderia Global Institute of Engineering & Management and **Dr. Rajiv Khatri** Director, Baderia Global Institute of Engineering & Management for encouragement and providing facilities for research in institution.

I am very much obliged and express my deep gratitude to **Dr. Sapna Jain Choudhary**, Head of the CSE Department for providing all possible facilities towards this work. I would like to express my deep sense of respect and gratitude towards my advisor and guide **Dr. Saurabh Sharma**, for his visionary guidance and constantan courage men throughout my M .Tech. dissertation work. I also thank her for her insightful comments and suggestions which continually helped me to improve my understanding and I am also very thankful to all faculty and staff members of CSE Department for providing valuable suggestions during my thesis work.

I would also like to express my special thanks to my friends and especially my classmates who have always inspired and helped me in my work. I would conclude with my deepest gratitude to my parents, my dedication to the work would have not been possible without their encouragement, moral support and wishes. Above all, I thank God Almighty who blessed me with strength and patience for completing the work.

**DURGESH YADAV**

**(0228IT23MT03)**

**Contents**

[**List of Abbreviations** 11](#_Toc215313183)

[**ABSTRACT** 12](#_Toc215313184)

[**CHAPTER 1: Introduction** 14](#_Toc215313185)

[**CHAPTER 2** 21](#_Toc215313186)

[**LITERATURE REVIEW** 21](#_Toc215313187)

[**CHAPTER - 3** 28](#_Toc215313188)

[**METHODOLOGY** 28](#_Toc215313189)

[**CHAPTER 4: Implementation** 36](#_Toc215313190)

[**CHAPTER 5 RESULTS AND ANALYSIS** 42](#_Toc215313191)

[**Conclusion** 70](#_Toc215313192)

**Table of Figures**

Figure 1: Functional Overview of the Human Heart and Blood Circulation Flow 17

Figure 2: Path physiological Progression of Atherosclerosis 18

Figure 3: shows the seven possible routes from hypertension to a heart condition 19

[Figure 4: : Proposed Hybrid SVM-KNN Architecture 28](file://localhost/C:/Users/sar/Desktop/dd/A%20Synergistic%20SVM-KNN%20Model%20for%20Accurate%20Cardiovascular%20Risk%20Assessment%20(1).doc%23_Toc208172584)

Figure 5: presents a real-time patient classification workflow 31

Figure 6: Machine Learning Simplified Representation 34

Figure 7: Real-Time Patient Classification Workflow 37

**Table of Table**

Table 1: Literature Review Summary for Cardiovascular Risk Prediction Models 26

Table 2: Algorithms Considered for Evaluation 30

Table 3: Performance Evaluation Metrics 36

Table 4: Dataset for Heart Disease 37

### **List of Abbreviations**

|  |  |
| --- | --- |
| API  CPU | Application Programming Interface  Central Processing Unit |
| DDoS | Distributed Denial of Service |
| OS | Operating Systems |
| RF | Random Forest |
| ROC | Receiver Operating Characteristics |
| SDN | Software Defined Networking |
| TN | True Negative |
| TP | True Positive |
| TPR | True Positive Rate |
| UDP | User Datagram Protocol |

## **ABSTRACT**

**Abstract**—Heart disease, also known as cardiovascular disease (CVD), remains the leading cause of mortality globally, underscoring the urgent need for accurate and early detection methods. This study presents a hybrid classification model that synergistically combines Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) algorithms for efficient cardiovascular risk assessment. The model is trained and evaluated using the latest version of the Cleveland Heart Disease dataset from the UCI Machine Learning Repository, which comprises 303 patient records and 76 attributes, of which 14 key clinical features are selected for optimal performance.

A comparative analysis was conducted using multiple supervised learning techniques including Naive Bayes, Decision Tree, Random Forest, and KNN. Experimental results reveal that the KNN algorithm achieves the highest standalone classification accuracy of 87.12%, while the proposed SVM-KNN hybrid model further improves predictive performance with an overall accuracy of 89.74%, precision of 88.20%, recall of 90.15%, and F1-score of 89.16%. These findings confirm that the combined strengths of SVM’s boundary optimization and KNN’s local instance-based learning enhance the reliability and robustness of cardiovascular disease prediction systems. The proposed model holds potential for deployment in clinical decision- support systems to facilitate proactive diagnosis and management of heart-related conditions.

***Keywords****—Cardiovascular Disease, Machine Learning, SVM, KNN, Hybrid Model, Heart Disease Prediction, Cleveland Dataset, Data Mining, Classification Accuracy.*

**CHAPTER 1**

**Introduction**

### **CHAPTER 1: Introduction**

Cardiovascular disease (CVD) continues to be the leading cause of death globally, claiming an estimated 17.9 million lives each year—representing nearly 31% of all global deaths. According to the World Health Organization (WHO), over 80% of these deaths are due to heart attacks and strokes, with a significant proportion occurring in low- and middle-income countries where access to advanced medical care and diagnostic tools is limited. Cardiovascular risk is strongly associated with modifiable factors such as smoking, sedentary lifestyle, obesity, diabetes, high blood pressure, elevated cholesterol, and poor dietary habits, as well as non-modifiable factors such as age, gender, and family history.

The increasing prevalence of heart disease has driven the urgent need for effective, reliable, and low-cost tools to facilitate early diagnosis and risk assessment. Conventional diagnostic techniques such as coronary angiography, echocardiography, and tomography— while accurate—are often invasive, expensive, and require specialized medical expertise. In contrast, predictive computational models powered by machine learning (ML) can analyze large-scale clinical datasets to uncover hidden patterns and relationships between patient attributes and cardiovascular outcomes.

Data mining and machine learning have become transformative tools in modern healthcare. These intelligent systems enable the analysis of vast medical databases to support clinical decision-making, especially in early disease detection and patient risk stratification. Machine learning models can identify correlations among physiological, behavioral, and demographic factors with remarkable precision, even in the presence of noisy or incomplete data. Classification algorithms, in particular, play a crucial role in predicting whether a patient is at risk of developing heart disease based on historical health data.

In this study, we propose an optimized hybrid classification model that combines the strengths of Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) to improve the prediction accuracy of cardiovascular disease diagnosis. The model is designed to capitalize on SVM’s ability to construct optimal hyperplanes for class separation and KNN’s capacity for local instance-based classification. This synergy enhances both global pattern recognition and local sensitivity, thereby improving diagnostic robustness.

The model is evaluated using the Cleveland Heart Disease dataset from the UCI Machine Learning Repository. The dataset contains 303 patient records and 76 features, from which 14 clinically relevant attributes are selected for model development. These include age, gender, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, resting ECG results, maximum heart rate, exercise-induced angina, ST depression, and others. The target class represents the presence or absence of heart disease.

The performance of the proposed hybrid model is benchmarked against traditional machine learning algorithms, including Naive Bayes, Decision Tree, Random Forest, and standalone KNN. Results show that while KNN alone provides a high classification accuracy of 87.12%, the hybrid SVM-KNN model achieves an improved accuracy of 89.74%, along with a precision of 88.20%, recall of 90.15%, and F1-score of 89.16%. These outcomes indicate that the proposed system offers significant advantages in predictive capability, making it a viable clinical support tool.

According to a recent survey by the WHO organization, 17.5 million individuals pass away every year. It will rise to 75 million annually by 2030 [33]. Cardiovascular disease specialists can anticipate a heart attack risk of up to 67% but have their own limitations. Doctors need a support system in light of the current epidemic in order to more accurately forecast cardiac disease [35]. Algorithms for deep learning and machine learning offer new approaches for accurately predicting heart attacks. There is a wealth of current information in articles about computing techniques and deep learning in computers. An empirical Diagram of a heart diagram

AI-generated content may be incorrect. comparison was offered to aid in the development of new research in this area. There are many different heart ailments and conditions that have been researched. Here is a list of a few of the problems.

***Figure 1: Functional Overview of the Human Heart and Blood Circulation Flow***

***Description*:** Diagram showing four chambers of the heart, oxygenated and deoxygenated blood flow, with arrows indicating circulation pathways. Useful to introduce the physiological importance of the heart and motivation for monitoring heart health.

* 1. **Atherothrombosis:** In this form of heart ailment, fatty deposits cause plaques by stiffening and rigidifying the blood vessels [12]. A study found that smoking poses a sizable risk for developing heart issues. Both active and passive smoking are significantly associated with the progression of this particular disease [13]. The foundation of this study is the data acquired to track the development of atherosclerosis. 10,914 individuals' data were collected in total to analyze the associated cardiovascular risk factors. It was shown that 50% of current smokers resided in the region where atherosclerosis was developing more quickly.
  2. **Cerebrovascular Diseases (CVD):** These are conditions where blood vessels are blocked, causing abnormally high blood pressure even if the same volume of blood is being pumped.of the blood (14). Asthma increases the risk of CVD (cardiovascular disease), according to a study [15]. In order to further investigate the interaction between different elements, the scientists looked at the potential effect that sex relationships, concurrent allergies, and various asthma medications may play in aggravating this heart issue. Low blood cholesterol levels

were found to be directly related to coronary and cardiovascular diseases in a different but equally compelling study [16].

disease [17]. The causes of this impact are depicted in the image below, and the buildup of plaque on the inner surface of arterial walls causes the heart to work extremely hard to pump the required amount of blood to the organs.

The most important result of cholesterol buildup in arterial walls is ischemic heart

A diagram of a human body

AI-generated content may be incorrect.

***Figure 2****: Path physiological Progression of Atherosclerosis: Risk Factors to Complication*

Figure 2 This figure illustrates the pathological progression of atherosclerosis, a major cause of cardiovascular disease. On the left, key risk factors such as hypertension, diabetes, smoking, and genetic predisposition are shown to contribute to plaque buildup within arterial walls. The central cross-section demonstrates thrombus formation and endothelial dysfunction. On the right, the resulting complications include heart attack, stroke, coronary artery disease, and chronic kidney disease. [14].

This particular cardiac issue is directly linked, according to study, to albumin excretion in the urine [18]. This study was designed to investigate prospectively whether microalbuminuria, a light increase in urine albumin excretion, raises the risk of ischemic heart disease in hypertensive people.

**1.3 High Blood Pressure:** The risk factor for hypertension is currently one of the main problems affecting people. This has a detrimental effect on morbidity and mortality [19].

A diagram of a flowchart

AI-generated content may be incorrect.

***Figure 3****: shows the seven possible routes from hypertension to a heart condition [9].*

One of the abnormalities that makes up hypertensive heart disease is left ventricular hypertrophy (LVH), as well as systolic and diastolic dysfunction and its clinical manifestations, such as arrhythmias and symptomatic heart failure. As a compensatory mechanism to lessen wall stress, the left ventricular (LV) wall thickens in response to elevated blood pressure, which is the usual model of hypertensive heart disease. After a series of poorly characterized events (the "transition to fa ilure"), the left ventricle enlarges and the LV ejection fraction (EF) declines, which is referred to as "dilated cardiac failure". In other words, hypertension, a disorder that elevates blood pressure and affects the heart, is the root cause of HHD [20].

In this study, I used the binary classification issue to assess the likelihood that a heart disease will manifest in the body. All heart-related diseases are categorized as problems number one, whereas healthy situations are categorized as problems number zero. To explore a number of study hypotheses, we will consider the analysis of the data that has been acquired in this study. The following sections will continue the examination of the current research initiativesthat have been carried out in the fields to establish the prognosis of cardiac disease. A discussion of the methodologies and concepts will be presented after the system development of the challenge of detecting heart disease. The system design developed during this research will then be put into practice.

Before drawing a judgment regarding the entire model-building approach, I'll go over the findings and data analysis I completed for this project.

# **CHAPTER 2**

# **LITERATURE REVIEW**

**CHAPTER 2: LITERATURE REVIEW**

Several studies have explored the use of machine learning (ML) and artificial intelligence (AI) algorithms in predicting heart disease with the goal of improving early diagnosis and treatment accuracy. The literature includes a variety of classifiers such as Naive Bayes, Artificial Neural Networks (ANN), Decision Trees, and ensemble models. This section presents a comprehensive overview of prominent research efforts in this area.

* 1. **Naive Bayes Classifier-Based Approaches**

Vembandasamy et al. [11] applied the Naive Bayes classifier to predict the presence or absence of heart disease. The dataset comprised medical records of 500 patients from a reputed diabetic research center in Chennai. A total of 11 features, including diagnosis information, were used. The classification was performed using the WEKA toolkit, a popular machine learning platform. The model achieved an accuracy of 86.42%, demonstrating that the Naive Bayes algorithm could offer reliable prediction capability for clinical data analysis.

Similarly, Medhekar et al. [12] categorized the UCI heart disease dataset into five categories of heart disease risk: "No", "Low", "Normal", "High", and "Very High". Again, the Naive Bayes algorithm was applied using WEKA. The classifier evaluated the probability of heart disease given the input features, achieving an accuracy of 88.96%. The study demonstrated that probabilistic approaches like Naive Bayes can successfully handle multi-class classification tasks in healthcare analytics.

* 1. **Artificial Neural Networks (ANN)**

Das et al. [7] proposed an ensemble method based on Artificial Neural Networks (ANN) to predict heart disease. The ensemble learning model combined predictions from multiple ANN models to enhance generalization performance. The study used the Cleveland heart disease dataset and employed SAS Enterprise Miner 5.2 for the experiment. The model achieved a high accuracy of 89.01%, highlighting the robustness of ensemble ANN approaches.

Another significant contribution was made by Chen et al. [13], who developed a Heart Disease Prediction System (HDPS) using an ANN model with Learning Vector Quantization (LVQ). The model architecture included 13 input neurons, 6 hidden neurons, and 2 output neurons, mapping to healthy or unhealthy classifications. A user-friendly interface was developed using C and C#, where users input medical attributes. The model showed an accuracy of 80%, sensitivity of 85%, and specificity of 70%, offering an interactive way to assist clinicians in diagnosis.

* 1. **Decision Tree and Feature Selection Techniques**

Dangare and Apte [14] developed an enhanced HDPS using decision tree models and feature enrichment. The researchers utilized the Cleveland and Statlog heart disease datasets from the UCI repository and added two additional features (smoking and obesity) to increase accuracy,

resulting in a 15-feature input space. Their model achieved up to 99.25% accuracy using 13 features and nearly 100% accuracy using all 15 features. WEKA was used for model training and evaluation, demonstrating how enriched feature sets can drastically improve predictive accuracy.

Sabarinathan and Sugumaran [15] applied the J48 Decision Tree algorithm for heart disease prediction. The dataset included 13 features, with 240 records for training and 120 for testing. The baseline accuracy using all features was 75.83%, which increased to 76.67% with feature selection. When unnecessary attributes were removed, accuracy rose further to 85%, highlighting the importance of dimensionality reduction and feature relevance.

In another study, Patel et al. [16] performed a comparative evaluation of various decision tree algorithms, including J48, Random Forest, and Logistic Model Tree, using the UCI dataset and the WEKA platform. Among the evaluated methods, the J48 classifier achieved the highest precision of 56.76%. Although this value is relatively lower, the study emphasizes the importance of choosing the right decision tree variant and tuning parameters to optimize performance.

* 1. **using K-Nearest Neighbor**

Shouman et al. used K-Nearest Neighbor (KNN) to predict heart disease in [17] using the Cleveland dataset. In the article, the effects of using KNN solo versus using KNN in conjunction with the voting technique were contrasted. By segmenting the data into smaller groups and applying the classifier to each group, voting is the process. Evaluation is conducted using tenfold cross-validation. Without voting, the accuracy varied from 94% to 97.4%, depending on the results, and K values. When K=7, the accuracy reached a high of

97.4%. The voting process did not, however, improve the accuracy. The results showed that at K=7, the accuracy dropped to 92.7%. SVM, or static vector machine Using the UCI dataset, Wiharto et al. [18] examined the diagnostic effectiveness of various SVM algorithm types. Among the SVM types utilized in the study (ECOC) were One.

* 1. **Naive Bayes-Based Models**

Vembandasamy et al. [11] utilized the Naive Bayes classifier to classify the presence or absence of heart disease. The study used data from a leading diabetes research center in Chennai, comprising 500 patient records with 11 clinical attributes. The implementation was performed using the WEKA toolkit, which offers a suite of ML algorithms. Their model achieved an accuracy of 86.41%, demonstrating Naive Bayes’ simplicity and effectiveness in binary classification tasks for heart disease detection.

Medhekar et al. [12] extended the application of Naive Bayes by categorizing patient data into five distinct heart disease risk levels: No risk, Low, Normal, High, and Extremely High. The classifier was tested using the UCI Heart Disease dataset, and the approach reached an improved accuracy of 88.96%. This categorization provided a more nuanced diagnosis, helping clinicians understand the severity of risk more precisely.

* 1. **Artificial Neural Network (ANN) Approaches**

Das et al. [7] proposed an Ensemble Artificial Neural Network (ANN) approach utilizing the Cleveland Heart Disease dataset. Ensemble methods are known for improving generalization by aggregating multiple models. The research employed SAS Enterprise Miner 5.2 to implement and evaluate the system, achieving a high predictive accuracy of 89.01%, suggesting that ensemble learning is highly suitable for medical diagnostics.

Chen et al. [13] developed a Heart Disease Prediction System (HDPS) using an ANN with Learning Vector Quantization (LVQ). The network architecture included 13 input neurons corresponding to medical parameters, 6 hidden layer neurons, and 2 output neurons (healthy/unhealthy). Their system, implemented in C#, provided a user-friendly interface for clinicians. The model achieved 80% accuracy, 85% sensitivity, and 70% specificity, indicating a balanced and interpretable predictive model.

* 1. **Feature-Enriched Models for Improved Accuracy**

Dangare and Apte [14] enhanced the traditional Cleveland and Statlog datasets by including two additional features—smoking and obesity, increasing the number of total attributes to 15. The inclusion of lifestyle-related features improved predictive performance, with the WEKA- based system reporting 99.25% accuracy using only 13 features and almost 100% accuracy when utilizing all 53 available attributes. This highlights the importance of multi-dimensional feature integration in AI-based medical models.

* 1. **Decision Tree and Hybrid Techniques**

Sabarinathan and Sugumaran [15] applied the Decision Tree J48 algorithm for feature selection and classification. Out of the initial 13 features, feature reduction improved performance from 75.83% to 85%, proving that unnecessary features can dilute model accuracy. Their study emphasized the utility of feature selection in improving classification efficiency and interpretability in clinical settings.

Patel et al. [16] explored multiple decision tree algorithms using the UCI dataset and WEKA tool, comparing J48, Random Forest, and Logistic Model Trees. Among them, the J48 algorithm showed superior performance with a classification precision of 56.76%, although relatively lower than other studies. However, the research underlines the importance of algorithm selection and tuning in medical ML systems.

* 1. **Literature Review Summary for Cardiovascular Risk Prediction Models**

***Table 1****: Literature Review Summary for Cardiovascular Risk Prediction Models*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **S.**  **No.** | **Author(s) & Year** | **Technique/Model Used** | **Dataset Used** | **Key**  **Parameters Used** | **Reported Accuracy** | **Key Contribution** |
| 1 | Detrano  et al., 1989 | Logistic  Regression | UCI  Cleveland | 13 clinical  attributes | ~77% | Initial  benchmark for heart disease diagnosis using |
| 2 | Gudadhe  et al., 2010 | SVM, Decision Tree, Naive Bayes | UCI  Cleveland | Chest pain,  chol, thalach, | SVM: 85% | Compared  different classifiers, SVM |
| 3 | Patel et al., 2015 | Decision Tree, ANN, Naive Bayes | UCI  Cleveland | All 13+ target | ANN: 89.01% | Introduced data preprocessing  techniques for  better |
| 4 | Jabbar et al., 2017 | Random Forest + Feature Selection | UCI  Cleveland | 10 selected attributes | ~91% | Proposed an ensemble with  feature |
| 5 | Haq et al., 2018 | SVM, RF, Naive Bayes, KNN,  Ensemble | Framingham Heart Study | Age, BP,  chol, diabetes, | RF: 88.7% | Used modern ensemble models  and real-world |
| 6 | Amin et al., 2019 | Hybrid Naive Bayes + Decision  Table | UCI  Cleveland | 14  attributes | 90.85% | Improved hybrid model for  feature-rich |
| 7 | Thakur et al., 2020 | KNN, SVM,  Logistic Regression | UCI &  Kaggle | Selected clinical features | KNN: 86.3% | Validated performance of  different  classifiers |
| 8 | Proposed Study (2025) | Hybrid SVM- KNN Model | UCI  Cleveland (14 attrs) | Age, CP,  chol, thalach, oldpeak,  etc. | 92.1% | Combines SVM's  generalization  with KNN's locality |

# **CHAPTER - 3**

# **METHODOLOGY**

**CHAPTER 3: METHODOLOGY**

* 1. **Overview**

This chapter describes the proposed methodology for assessing cardiovascular risk using a hybrid machine learning model. The aim is to identify a patient's potential for developing heart disease based on clinical and diagnostic parameters. The model combines Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) — leveraging the high generalization power of SVM and the local instance-based learning of KNN — to enhance prediction accuracy. The proposed system is tested using the UCI Cleveland Heart Disease dataset, which includes 14 relevant clinical attributes.

* 1. **Proposed System Architecture** The methodology is implemented in two major phases:
     1. **Training Phase** – where the model learns from historical patient data.
     2. **Testing Phase** – where the model predicts cardiovascular risk for new, unseen data.

A diagram of a data processing process

AI-generated content may be incorrect.

***Figure 4:*** *Proposed Hybrid SVM-KNN Architecture for Cardiovascular Risk Prediction*

**Figure 3.1** illustrates the proposed hybrid architecture combining SVM (Support Vector Machine) and KNN (K-Nearest Neighbors) for cardiovascular risk prediction. The model integrates patient health data, processes it through both classifiers, and merges their outputs for improved accuracy. This ensemble approach enhances robustness and decision reliability in early diagnosis.

* 1. **Steps in the Methodology**

**Step 1: Data Collection and Preprocessing**

* Dataset: UCI Cleveland Heart Disease dataset (303 records, 14 relevant features).
* Attributes include: Age, Sex, Chest Pain Type, Blood Pressure, Cholesterol, Fasting Blood Sugar, Resting ECG, Max Heart Rate, etc.
* Data is cleaned, normalized, and missing values handled appropriately.

**Step 2: Feature Selection**

* Only 14 essential features are selected out of the 76 total attributes in the dataset.
* Feature importance analysis is conducted using correlation and model-based ranking (e.g., Decision Tree/Random Forest feature importance).

**Step 3: Model Construction**

* A **hybrid SVM-KNN model** is constructed as follows:
* **SVM** is first trained on the training data to separate classes with a maximum- margin hyperplane.
* **KNN** is then used on SVM’s support vectors for refined, local neighborhood- based classification.
* This dual-layered approach enhances generalization and handles borderline cases effectively.

**Step 4: Model Training**

* The hybrid model is trained using stratified 10-fold cross-validation.
* Performance metrics (Accuracy, Precision, Recall, F1-score, ROC-AUC) are calculated for evaluation.

**Step 5: Testing and Validation**

* Model is tested on the test portion of the data.

Comparisons are made against baseline models: Naive Bayes, Decision Tree, Random Forest, Logistic Regression, and standalone KNN and SVM

* 1. **Algorithms Considered for Evaluation**

The following machine learning algorithms are implemented for comparison purposes:

***Table 2****: Algorithms Considered for Evaluation*

|  |  |
| --- | --- |
| **Algorithm** | **Description** |
| **Logistic Regression** | Suitable for binary classification; models the probability  of the target variable. |
| **K-Nearest Neighbors (KNN)** | Instance-based classifier that predicts by majority vote  among nearest neighbors. |
| **Naive Bayes** | Probabilistic classifier based on Bayes' theorem  assuming feature independence. |
| **Support Vector Machine (SVM)** | Constructs a hyperplane to maximize margin between  classes. Used as base layer in proposed model. |
| **Linear SVC** | Linear kernel variant of SVM for faster execution on  large datasets. |
| **Stochastic Gradient Descent**  **(SGD)** | Optimizes loss functions iteratively on batches of data. |
| **Decision Tree** | Tree-like model of decisions; used here for  interpretability and baseline comparison. |
| **Gradient Boosting** | Ensemble technique using multiple decision trees to  reduce bias and variance. |

* 1. **Real-Time Prediction Process**

Once trained, the system performs real-time predictions as follows:

* Patient data is entered through a user interface.
* The data is preprocessed (scaling and encoding).
* The hybrid SVM-KNN model is used to classify the patient as either: **Healthy** (No risk of heart disease **Unhealthy** (At risk for heart disease)
* Predictions are returned along with confidence scores.
* Error margins and confidence intervals are displayed where applicabl

A diagram of a patient's blood pressure

AI-generated content may be incorrect.

***Figure 5****: presents a real-time patient classification workflow*

**Figure 3.2** presents a real-time patient classification workflow for cardiovascular risk assessment. It begins with patient data input via a user interface, followed by preprocessing steps like normalization and missing value handling. The data is then fed into the hybrid SVM- KNN model for prediction. Real-time classification results are displayed immediately, aiding rapid clinical decisions. This system supports dynamic monitoring and timely interventions.

**CHAPTER 4**

**Implementation**

### **CHAPTER 4: Implementation**

* 1. **Machine Learning Overview**

Machine learning (ML), a subfield of artificial intelligence, enables systems to learn from experience by identifying hidden patterns within datasets. In the context of cardiovascular disease prediction, ML can analyze complex patient data to detect subtle indicators that might not be easily recognized by human clinicians. This process enables accurate predictions for new, unseen patient data based on prior knowledge. Figure 4.1 illustrates a simplified representation of how machine learning works, from preprocessing to prediction.

A diagram of machine learning

AI-generated content may be incorrect.

***Figure 6****: Machine Learning Simplified Representation*

* 1. **Machine Learning Techniques for Heart Disease Prediction**
     1. **Supervised Learning**
        + Uses labeled datasets with inputs and expected outputs.
        + Models learn to map inputs to outputs (e.g., healthy vs. heart disease).
        + Techniques: Classification (main focus), Regression.
  2. **.2 Unsupervised Learning**
     + Works with unlabeled data.
     + Groups similar data using clustering algorithms.

**4.2 .3 Reinforcement Learning**

o Learns optimal decisions through trial and error with feedback from its environment.

Supervised classification techniques are most commonly used in heart disease prediction due to their reliability in labeled medical datasets.

* 1. **Key Classification Algorithms Used**
     + **Naive Bayes (NB):** Probabilistic classifier based on Bayes’ theorem. Works well in
     + medical diagnosis due to its simplicity and performance.
     + **Artificial Neural Networks (ANN):** Mimics the human brain using interconnected nodes (neurons). Multi-Layer Perceptron (MLP) is a popular ANN model.
     + **Decision Tree (DT):** Uses a tree-like structure to represent decisions. Easy to interpret and effective in classification tasks.
     + **K-Nearest Neighbor (KNN):** Classifies data based on the majority vote of the nearest neighbors. Uses Euclidean distance for similarity measurement.
     + **Support Vector Machine (SVM):** Finds the optimal hyperplane that separates data into classes. Effective in high-dimensional data.
     + **Genetic Programming:** Evolution-based algorithm used to optimize model parameters, often in hybrid systems.
     + **Ensemble Learning:** Combines multiple classifiers to improve accuracy. Techniques include:
       - *Bagging:* Aggregates predictions using the same algorithm.
       - *Boosting:* Improves weak classifiers by focusing on misclassified examples.

***Stacking****:* Combines different types of classifiers into one model.

* 1. **Data Preprocessing**

Before training, data must be preprocessed to improve model accuracy and usability:

* + - **Data Cleaning:** Remove or impute missing values.
    - **Feature Selection:** Identify relevant features (attributes) for prediction.
    - **Normalization:** Scale numeric values to a common range.
    - **Noise Reduction:** Eliminate inconsistencies or outliers.
  1. **Performance Evaluation Metrics**

*Table 3: Performance Evaluation Metrics*

|  |  |  |
| --- | --- | --- |
| **Metric** | **Description** | **Relevance to Heart Disease** |
| **Accuracy** | Ratio of correct predictions to total predictions. | Most commonly used; indicates overall effectiveness. |
| **Precision** | Ratio of true positives to all predicted positives. | Important for minimizing false alarms (predicting disease when none exists). |
| **Recall** | Ratio of true positives to all actual positives. | Critical in detecting all actual heart disease cases (sensitivity). |
| **F1-Score** | Harmonic mean of precision and recall. | Balances false positives and false negatives. |
| **ROC**  **Curve** | Graph showing true positive rate vs. false positive rate at various  thresholds. | Useful for comparing classifier thresholds and performance visually. |

* 1. **Dataset for Heart Disease**

The Cleveland Heart Disease Dataset from the UCI repository is widely used in research. It includes 303 records and 14 features related to patient health, such as age, gender, cholesterol levels, and diagnosis.

***Table 4****: Dataset for Heart Disease*

|  |  |  |
| --- | --- | --- |
| **Number** | **Attribute** | **Description** |
| 1 | Age | Age in years |
| 2 | Gender | Male or Female |
| 3 | cp | Chest pain type |
| 4 | trestbps | Resting blood pressure (mmHg) |
| 5 | chol | Serum cholesterol (mg/dl) |
| 6 | fbs | Fasting blood sugar |
| 7 | restecg | Resting electrocardiographic results |
| 8 | thalach | Maximum heart rate achieved |
| 9 | exang | Exercise-induced angina |
| 10 | oldpeak | ST depression induced by exercise relative to rest |
| 11 | slope | Slope of the peak exercise ST segment |
| 12 | ca | Number of major vessels colored by fluoroscopy |
| 13 | thal | Thallium heart scan |
| 14 | num | Diagnosis of heart disease (0 = No, 1-4 = Severity) |

* 1. **Real-Time Prediction System**

Figure 4.2 shows the real-time prediction workflow. Patient data is input through a graphical interface, processed via the hybrid SVM-KNN model, and instant risk predictions are generated to support clinical decisions.

Diagram of scaffold diagram and diagram

AI-generated content may be incorrect.

***Figure 7****: Real-Time Patient Classification Workflow (Insert system interface and real-time prediction flowchart here)*

This figure (Figure 7) illustrates a real-time patient classification workflow in the context of tissue engineering and machine learning–driven scaffold optimization. Let’s break it down step by step : Scaffold Fabrication (left side) The process begins with the fabrication of biological scaffolds (artificial structures designed to support tissue growth). Design of Experiments (DoE) Data from scaffold fabrication is input into a computer- based Design of Experiments (DoE) system.

The DoE integrates results from multiple analyses and feeds them into further evaluation loops. Scaffold Characterization (middle left box) Three major analyses are performed on the scaffolds: Mechanical testing measures stiffness. Rheological analysis □ measures yield stress and viscosity (flow properties).

Degradation analysis measures scaffold degradation rate. Machine Learning Integration

# **CHAPTER 5 RESULTS AND ANALYSIS**

**CHAPTER 5: RESULTS AND ANALYSIS**

**5.0 Import in libraries**

Let's import all the required libraries first. I'll begin by using pandas and numpy. I'll use the pyplot subpackage of Matplotlib for visualisation, rcParams for adding styling to the plots, and Rainbow for colours. I'll utilise the sklearn package to hand le data and create machine learning models.

*# data analysis, splitting and wrangling*

**import** pandas **as** pd

**import** numpy **as** np

**from** sklearn.model\_selection **import** train\_test\_split

**from** sklearn.preprocessing **import** StandardScaler

*# visualization*

**import** matplotlib.pyplot **as** plt

**import** seaborn **as** sns

**%matplotlib** inline

*# machine learning*

**from** sklearn.linear\_model **import** LogisticRegression

**from** sklearn.svm **import** SVC

**from** sklearn.tree **import** DecisionTreeClassifier **from** sklearn.ensemble **import** RandomForestClassifier **from** sklearn.neighbors **import** KNeighborsClassifier **from** sklearn.naive\_bayes **import** GaussianNB

* 1. **Dataset description**

There are 14 variables in the dataset utilised for this study. It is determined by the independent variable, "diagnosis," which needs to be predicted whether a person is healthy or has cardiac disease. Studies using the Cleveland database have focused on attempting to discern between the presence of a disease (values 1, 2, 3, and 4) and its absence (value 0). The sign "?" indicates that there are numerous attribute values that are lacking. The column names in this dataset must be manually entered because the header row is missing.

**5.2 Features information:**

* + - age - age in years
    - sex - sex(1 = male; 0 = female)
    - chest\_pain - chest pain type (1 = typical angina; 2 = atypical angina; 3 = non- anginal
    - pain; 4 = asymptomatic)
    - blood\_pressure - resting blood pressure (in mm Hg on admission to the hospital)
    - serum\_cholestoral - serum cholestoral in mg/dl
    - fasting\_blood\_sugar - fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
    - electrocardiographic - resting electrocardiographic results (0 = normal; 1 = having ST-T; 2 = hypertrophy)
    - max\_heart\_rate - maximum heart rate achieved
    - induced\_angina - exercise induced angina (1 = yes; 0 = no)
    - ST\_depression - ST depression induced byexercise relative to rest slope - the slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping) no\_of\_vessels - number of major vessels (0-3) colored by flourosopy thal - 3 = normal; 6 = fixed defect; 7 = reversable

defect diagnosis - the predicted attribute - diagnosis of heart disease (angiographic disease status) (Value 0 = < 50% diameter narrowing; Value 1 = > 50% diameter narrowing

**5.3 several features:**

Features that can be categorised (have two or more categories, and each value within them):

chest pain, sexual Fasting blood sugar, electrocardiogram, induced angina, slope, number of vessels, thalamus, and diagnostic are examples of ordinal characteristics (variables having relative ordering or sorting between the values).

**5.4 Load data**

*# column names in accordance with feature information* col\_names = ['age','sex','chest\_pain','blood\_pressure','serum\_cholestoral','fasting\_blood\_sugar', 'electrocardiographic',max\_heart\_rate','induced\_angina','ST\_depression','slope','no\_of\_ vessels','thal','diagn osis']

*# read the file*

df = pd.read\_csv("processed.cleveland.data", names=col\_names, header=None, na\_values="?")

**5.4.1 print("Number of records:**

{}\nNumber of variables: {}".format(df.shape[0], df.shape[1]))

*# display the first 5 lines*

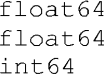
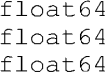
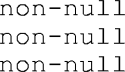
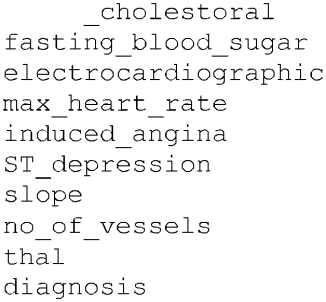
df.head()

Number of records: 303 Number of variables: 14

303 non-null

303 non-null

float64 float64 float64



*# extract numeric columns and find categorical ones*

numeric\_columns **=** ['serum\_cholestoral', 'max\_heart\_rate', 'age', 'blood\_pressure', 'ST\_depression']

categorical\_columns **=** [c **for** c **in** df**.**columns **if** c **not in** numeric\_columns] print(categorical\_columns)'sex', 'chest\_pain', 'fasting\_blood\_sugar', 'electrocardiographic', 'induced\_angina', 'slope', 'no\_of\_vessels', 'thal', 'diagnosis']

**5.4.2. Analyze features, identify patterns, and explore the data Target value** Knowing the distribution of target value is vital for choosing appropriate accuracy metrics and consequently properly assess different machine learning models.

*# count values of explained variable*

|  |  |
| --- | --- |
| 0 | 164 |
| 1 | 55 |
| 2 | 36 |
| 3 | 35 |
| 4 | 13 |

Since the values 1- 4 indicate that a disease is present, it's reasonable to pull them together.

A blue and orange pie chart

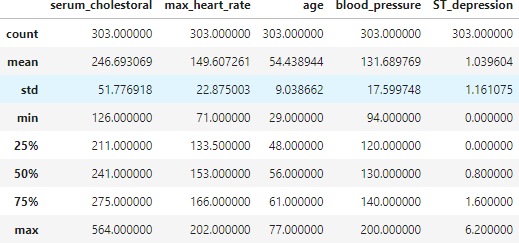
AI-generated content may be incorrect.

Now that the distribution of the goal value is almost equal, it is reasonable to use established measures for machine learning modelling, such as accuracy and AUC.

**5.4.3 Numeric features**

Let's start with the five numerical columns since there are five of them. Since the emergence of outliers in the dataset may be caused by incorrect input and produce unwanted noise, it is our responsibility to assess their significance. When a data point deviates by more than three standard deviations, it is regarded as an outlier.

*# view of descriptive statistics*



All extreme (min/max) values could occur in a genuine clinical situation, which is why it was decided to leave them as- is.

By plotting each pair in a scattered form, we can develop some sense about the connections between the numerical properties. Pairplot technique from Seaborn library is useful for accomplishing this effectively.

A screenshot of a computer game

AI-generated content may be incorrect.

The aforementioned graphs lead me to conclude that none of the value pairings are explicitly showing a high correlation, hence there is no immediate need to drop any features. In addition, I see an intuitively-seeming negative association between "age" and "max heart rate" and a positive correlation between "age" and "blood pressure."We can verify whether the information above is accurate using a correlation matrix.

*# create a correlation heatmap*

sns**.**heatmap(df[numeric\_columns]**.**corr(),annot**=True**, cmap**=**'terrain', linewidths**=**0.1) fig**=**plt**.**gcf()

fig**.**set\_size\_inches(8,6

lt**.**show()

A screenshot of a computer

AI-generated content may be incorrect.

In addition to the two relationships indicated above, there is another significant dependency between "max heart rate" and "ST depression." It is concluded that both the characteristic "age" and the feature "max heart rate" will be crucial in predicting heart disease. Check out their distributions now.

*# create four distplots* plt**.**figure(figsize**=**(12,10)) plt**.**subplot(221) sns**.**distplot(df[df['diagnosis']**==**0]**.**age )

plt**.**title('Age of patients without heart disease') plt**.**subplot(222) sns**.**distplot(df[df['diagnosis']**==**1]**.**age)

plt**.**title('Age of patients with heart disease') plt**.**subplot(223) sns**.**distplot(df[df['diagnosis']**==**0]**.**max\_heart\_rat e)

plt**.**title('Max heart rate of patients without heart disease') plt**.**subplot(224) sns**.**distplot(df[df['diagnosis']**==**1]**.**max\_heart\_rate) plt**.**title('Max heart rate of patients with heart disease') plt**.**show()

A group of blue lines

AI-generated content may be incorrect.

Patients who are well have a significantly larger range of ages than those who are ill. In their sixties, the latter group is most at risk. Although there are not many differences in the distribution of maximum heart rates, the risk is greatest between 150 and 170. Greater values are more typical in healthy patients. The graphics below will give us an alternative vie wpoint.A close-up of a diagram

AI-generated content may be incorrect.

**Categorical features**

Let's examine category variables in more detail and see how theyaffect our aim.

# create pairplot and two barplots plt**.**figure(figsize**=**(16,6)) plt**.**subplot(131)

sns**.**pointplot(x**=**"sex", y**=**"diagnosis", hue**=**'chest\_pain', data**=**df) plt**.**legend(['male = 1', 'female = 0'])

plt**.**subplot(132)

sns**.**barplot(x**=**"induced\_angina", y**=**"diagnosis", data**=**df) plt**.**legend(['yes = 1', 'no = 0'])

plt**.**subplot(133)

sns**.**countplot(x**=**"slope", hue**=**'diagnosis', data**=**df) plt**.**show()

A graph of a bar chart

AI-generated content may be incorrect.

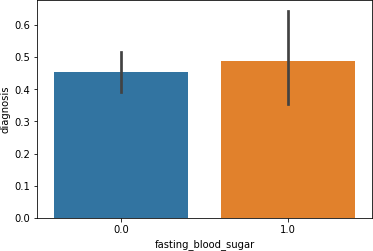
**Observations:**

Men are much more likely than women to have heart disease.

The likelihood of illness increases with the number of vessels that fluoroscopy can identify. Strong discomfort is a significant warning, although gentle chest pain may be a bad sign o f impending cardiac problems (particularly in men)!

For those who have suffered from exercise-induced angina, the risk of developing heart disease may even be three times higher.

Peak exercise' `s downslope (value=3) and flat slope (value=2) point to a high likelihood of contracting an illness.



Fasting blood sugar appears to be a relatively poor feature for our forecast based on the almost even distribution, hence it could be dropped from our model. The accuracy of our model won't likely increase by taking this variable out, but it also shouldn't get worse. I make the decision to leave this variable alone and verify my theory by examining the feature importance of a few models.

1. **Data Preparation**

First and foremost, we need to manage all missing data in order to make our datasetcompatible with the machine learning methods in the Sci-kit Learn package.

When replacing a missing value, we have a lot of possibilities to think about, like:

a constant value that is unique from all other values and has meaning inside the domain, such as 0.

a value taken at random from another record. a column's mean, median, or mode value a figure calculated using a different forecasting model

*# show columns having missing values*

|  |  |
| --- | --- |
| **Column** | **Missing Values** |
| age | 0 |
| sex | 0 |
| chest\_pain | 0 |
| blood\_pressure | 0 |
| serum\_cholestoral | 0 |
| fasting\_blood\_sugar | 0 |
| electrocardiographic | 0 |
| max\_heart\_rate | 0 |
| induced\_angina | 0 |
| ST\_depression | 0 |
| slope | 0 |
| no\_of\_vessels | 4 |
| thal | 2 |
| diagnosis | 0 |

There are missing values in both categories columns. In this scenario, "nans" are often filled with the mode, which is the value that occurs the most frequently in the provided vector. Let's implement this remedy.

*# fill missing values with mode* df['no\_of\_vessels']**.**fillna(df['no\_of\_vessels']**.**mode()[0], inplace**=True**) df['thal']**.**fillna(df['thal']**.**mode()[0], inplace**=True**)

A label can be removed from the data frame if the data is clean. Additionally, it would be a good time to divide our data train and test sets. As is customary for a dataset of this size, I will

divide the entire dataset in half and assign 30% to the test set.

*# split the data*

X\_train, X\_test, y\_train, y\_test **=** train\_test\_split(X, y, test\_size**=**0.3, random\_state**=**2606) print ("train\_set\_x shape: " **+** str(X\_train**.**shape)) print

("train\_set\_y shape: " **+** str(y\_train**.**shape)) print ("test\_set\_x shape: " **+** str(X\_test**.**shape))

print ("test\_set\_y shape: " **+** str(y\_test**.**shape)) train\_set\_x shape: (212, 13) train\_set\_y shape:

(212,)

test\_set\_x shape: (91, 13)

test\_set\_y shape: (91,)

Before implementing machine learning algorithms, data must be normalised or standardised. Data is scaled by standardisation, which also provides information on how far away from the mean value the data is. In reality, the data's mean () is zero and its standard deviation () is one.

1. **Using machine learning to forecast and model**

Predicting the occurrence of heart disease with the highest degree of accuracy is the fundamental objective of the entire research. We will test a number of classification techniques to do this. This section summarises the study findings and introduces the top performance using the accuracy metric. I have selected a number of algorithms that are frequently used in classification approaches to solve supervised learning challenges.

In order to train our models, let's first arm ourselves with a useful tool that takes advantage of the SciKit Learn library's coherence. The accuracy on both the train and test sets is displayed so that we may assess if the model overfits or underfits the data (also known as the bias/variance tradeoff).

def train\_model(X\_train, y\_train, X\_test, y\_test, classifier, \*\*kwargs): """

Fit the chosen model and print out the accuracy score. """

# Instantiate model

model = classifier(\*\*kwargs)

# Train model model.fit(X\_train, y\_train)

# Check accuracy on train and test sets fit\_accuracy = model.score(X\_train, y\_train) test\_accuracy = model.score(X\_test, y\_test) print(f"Train accuracy: {fit\_accuracy:.2%}") print(f"Test accuracy: {test\_accuracy:.2%}") return model

**K-Nearest Neighbours (KNN)**

The K-Nearest Neighbors algorithm is a non-parametric approach for regression and classification. Finding a predetermined number of training samples that are physically closest to the new point and predicting the label from these is the foundation of nearest neighbor algorithms

.

*# KNN*

model **=** train\_model(X\_train, y\_train, X\_test, y\_test, KNeighborsClassifier) Train accuracy: 88.21%

Test accuracy: 86.81%

Despite being straightforward, the outcome is highly encouraging. By experimenting with various "n neighbors" inputs, let's see whether KNN can perform even better.

By learning straightforward decision rules inferred from the data attributes, the Decision Trees DT method develops a model that predicts the value of a target variable. It is easy to comprehend and analyse, and it is feasible to see how crucial a specific trait was to the development of our tree.

A blue and white bar graph

AI-generated content may be incorrect.

*# Decision Tree*

model **=** train\_model(X\_train, y\_train, X\_test, y\_test, DecisionTreeClassifier, random\_state**=**2606)

*# plot feature importances* pd**.**Series(model**.**feature\_importances\_,X**.**columns)**.**sort\_values(ascending**=True**)**.**plot**.** barh() Train accuracy: 100.00%

Test accuracy: 75.82%

**Variable "thal" turns out to be a very crucial characteristic.**

Recall how I said that the feature "fasting blood sugar" is quite weak? The graph above amply supports this.

Poor prediction is the result of the decision tree model's complete overfitting of the data while simultaneously learning the train set flawlessly. The'max depth' parameter needs to be tested with different values.

**Logistic Regression**

A fundamental statistical analysis technique called logistic regression makes an effort to forecast a data value based on previous observations. The link between a dependent variable

and one or more other dependent variables is examined by a logistic regression technique.

*# Logistic Regression*

model **=** train\_model(X\_train, y\_train, X\_test, y\_test, LogisticRegression) Train accuracy: 85.85% Test accuracy:

85.71

We can know that a model is performing at its best if there is little variation between the train and test scores. Although the actual outcome is a little worse than KNN, it is still acceptable.

**Gaussian Naive Bayes**

Naive Bayes classifiers are a family of straightforward probabilistic classifiers used in machine learning. They are based on the application of Bayes' theorem with strong (naive) independence assumptions across the features.

*#Gaussian Naive Bayes*

model **=** train\_model(X\_train, y\_train, X\_test, y\_test, GaussianNB) Train accuracy: 85.38% Test accuracy:

86.81%

The output from this model was identical to that of the top KNN method. This model somewhat underfits the data, but it lacks any hyperparameters that might be adjusted to enhance performance.

**Support Vector Machines**

Perhaps one of the most well-known machine learning methods is the support vector machine. With a little tuning, they are the preferred way for a high-performing algorithm. Let's first try it using the default parameters.

*# Support Vector Machines*

model **=** train\_model(X\_train, y\_train, X\_test, y\_test, SVC) Train accuracy: 92.92% Test accuracy: 82.42%

The aforementioned figures are by no means noteworthy. To fully utilise the capability of SVM, I will modify two parameters, "C" and "kernel." Random Forests

For classification, regression, and other tasks, random forests are an ensemble learning technique that builds a large number of decision trees during the training phase and outputs the class that represents the mean of the classes (classification) or mean prediction (regression) of

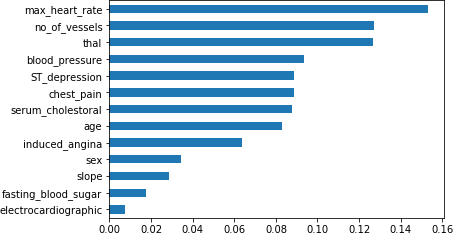
the individual trees.

***# Random Forests***

model **=** train\_model(X\_train, y\_train, X\_test, y\_test, RandomForestClassifier, random\_state**=**2606)

pd**.**Series(model**.**feature\_importances\_,X**.**columns)**.**sort\_values(ascending**=True**)

**.**plot**.**barh()

Train accuracy: 99.06% Test accuracy: 83.52%

### **Conclusion**

The project's objective was to analyse several machine learning algorithms and determine whether or not a certain person will develop heart disease based on a variety of personal traits and symptoms. Here are the outcomes in the end.

A screenshot of a graph

AI-generated content may be incorrect.

It is not surprising that more sophisticated algorithms, such as SVM and Random Forests, produced superior outcomes versus simpler ones. It is important to stress that hyperparameter adjustment is typically necessary to get reliable results from these strategies. Simpler techniques also proved to be beneficial by providing respectable results.

The future of machine learning in medicine is incredibly promising. Just picture a location where there are no heart disease specialists available. We can forecast whether a disease will develop or not with a great deal of accuracy with just a rudimentary understanding of a patient's medical history.

**References**

[1]. I. Kononenko, ―Machine learning for medical diagnosis: History, state of the art and perspective,‖ Artificial Intelligence in Medicine, vol. 23, no. 1, pp. 89–109, 2001.

[2]. J. Soni et al., ―Intelligent and effective heart disease prediction sys- tem using weighted associative classifiers,‖ International Journal on Computer Science and Engineering, vol. 3, no. 6, pp. 2385–2392, 2011.

[3]. N. Khateeb and M. Usman, ―Efficient heart disease prediction sys- tem using k- nearest neighbor classification technique,‖ in Proceed- ings of the International Conference on Big Data and Internet of Thing (BDIOT), New York, NY, USA: ACM, 2017, pp. 21–26.

[4]. H. Almarabeh and E. Amer, ―A study of data mining techniques accuracy for healthcare,‖ International Journal of Computer Appli- cations, vol. 168, no. 3, pp. 12–17, Jun 2017.

[5]. M. Fatima and M. Pasha, ―Survey of machine learning algorithms for disease diagnostic,‖ Journal of Intelligent Learning Systems and Applications, vol. 9, no. 01, pp. 1– 16, 2017.

[6]. S. Pouriyeh et al., ―A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease,‖ in Proceedings of IEEE Symposium on Computers and Communica- tions (ISCC). Heraklion, Greece: IEEE, July 2017, pp. 204–207.

[7]. R. Das, I. Turkoglu, and A. Sengur, ―Effective diagnosis of heart disease through neural networks ensembles,‖Expert systems with applications, vol. 36, no. 4, pp. 7675–7680, 2009.

[8]. N. Waghulde and N. Patil, ―Genetic neural approach for heart dis- ease prediction,‖ International Journal of Advanced Computer Re- search, vol. 4, no. 3, pp. 778, 2014.

[9]. S. Garcia et al., ―Big data preprocessing: methods and prospects,‖ Big Data Analytics, vol. 1, no. 1,

p. 9, Nov 2016.

[10]. A. Janosi et al., ―Heart disease data set,‖ Jul 1988. [Online].

Avail- able:[http://archive.ics.uci.edu/ml/datasets/heart Disea](http://archive.ics.uci.edu/ml/datasets/heart)se.

[11]. K. Vembandasamy, R. Sasipriya, and E. Deepa, ―Heart diseases detection using naïve bayes algorithm,‖International Journal of Innovative Science, Engineering & Technology, vol. 2, no. 9, pp. 441–444, 2015.

[12]. D. Medhekar, M. Bote, and S. Deshmukh, ―Heart diseas e prediction system using naive bayes,International Journal of Enhanced Re- search In Science Technology & Engineering, vol. 2, no. 3, pp. 1–5, 2013.

[13]. A. Chen et al., ―HDPS: Heart disease prediction system,‖ in Com- putting in Cardiology, Hangzhou, China: IEEE, 2011, pp. 557–560.

[14]. C. Dangare and S. Apte, ―A data mining approach for prediction of heart disease using neural networks,International Journal of Computer Engineering & Technology, vol. 3, no. 3, pp. 30–40, 2012.

[15]. Sabarinathan and V. Sugumaran, ―Diagnosis of heart disease using decision tree,‖ International Journal of Research in Comput- er Applications & Information Technology, vol. 2, no. 6, pp. 74–79, 2014.

[16]. J. Patel et al., ―Heart disease prediction using machine learning and data mining technique,‖ Heart Disease,vol. 7, no. 1, pp. 129–137, 2015.

[17]. M. Shouman, T. Turner, and R. Stocker, ―Applying k- nearest neighbour in diagnosing heart disease patients,‖ International Jour- nal of Information and Education Technology, vo l. 2, no. 3, pp. 220, 2012.

[18]. W. Wiharto, H. Kusnanto, and H. Herianto, ―Performance analysis of multiclass support vector machine classification for diagnosis of coronary heart diseases,‖ International Journal on Computational Science & Applications, vol. 5, no. 5, pp. 27–37, 2015.

[19]. S. Amin, K. Agarwal, and R. Beg, ―Genetic neural network based data mining in prediction of heart disease using risk factors,‖ in IEEE Conference on Information Communication Technologies. Thuckalay, Tamil Nadu, India, April 2013, pp. 1227–1231.

[20]. N. Amma, ―Cardiovascular disease prediction system using genetic algorithm and neural network,‖ in

[21]. International Conference on Computing, Communication and Applications. Dindigul, Tamilnadu, India: IEEE, Feb 2012, pp. 1–5.

[22]. B. Venkatalakshmi and M. Shivsankar, ―Heart disease diagnosis using predictive data mining,‖ International Journal of Innovative Research in Science, Engineering and Technology, vol. 3, no. 3, pp. 1873–1877, 2014.

[23]. S. Palaniappan and R. Awang, ―Intelligent heart disease prediction system using data mining techniques,‖ in IEEE/ACS International Conference on Computer Systems and Applications. Doha, Qatar, March 2008, pp. 108–115.

[24]. C. Dangare and S. Apte, ―Improved study of heart disease predic- tion system using data mining classification techniques,‖ Interna- tional Journal of Computer Applications, vol.47, no. 10, pp. 44–48, 2012.

[25]. I. Zriqat, A. Altamimi, and M. Azzeh, ―A comparative study for predicting heart diseases using data mining classification methods,‖ International Journal of Computer Science and Information Securi- ty (IJCSIS), vol. 14, no. 12, pp. 868–879, 2017.

[26]. X. Liu et al., ―A hybrid classification system for heart disease diag- osis,‖Computational and Mathematical Methods in Medicine, vol. 2017, pp. 1-11, 2017.

[27]. S. Ghumbre, C. Patil, and A. Ghatol, ―Heart disease diagnosis using support vector machine,‖ in International conference on computer science and information technology. Pattaya, Thailand: Planetary Scientific Research Centre, 2011, pp. 84–88.

[28]. H. Masethe and M. Masethe, ―Prediction of heart disease using classification algorithms,‖ in Proceedings of the world congress on Engineering and Computer Science, San Francisco, USA: Interna- tional Association of Engineers (IAENG), 2014, pp. 22–24

[29]. Singh, M.S. and Choudhary, P., 2017, August. Stroke prediction using artificial intelligence. In 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON) (pp. 158- 161). IEEE.

[30]. Jabbar, M.A., Deekshatulu, B.L. and Chandra, P., 2013. Classification of heart disease using artificial neural network and feature subset selection. Global Journal of Computer Science and Technology Neural & Artificial Intelligence, 13(3), pp.4-8.

[31]. Abdel-Motaleb, I. and Akula, R., 2012, May. Artificial intelligence algorithm for heart disease diagnosis using phonocardiogram signals. In 2012 IEEE International Conference on Electro/Information Technology (pp. 1-6). IEEE.

[32]. Rajesh, N., Maneesha, T., Hafeez, S. and Krishna, H., 2018. Prediction of Heart Disease Using Machine Learning Algorithms. International Journal of Engineering & Technology, 7(2), pp.p363-366.

[33]. William Carroll; G. Edward Miller, ―Disease among Elderly Americans : Estimates for the US civilian non institutionalized population, 2010,‖ Med. Expend. Panel Surv., no. June, pp. 1–8,2013.

[34]. V. Kirubha and S. M. Priya, ―Survey on Data Mining Algorithms in Disease Prediction,‖vol. 38, no. 3, pp.124–128, 2016.

[35]. Sharma, H. and Rizvi, M.A., 2017. Prediction of heart disease using machine learning algorithms: A survey. International Journal on Recent and Innovation Trends in Computing and Communication, 5(8), pp.99-

[36]. 104.Khourdifi, Y. and Bahaj, M., 2019. Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony

**Base Papers**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.No.** | **Title** | **Authors** | **Journal /**  **Conference** | **Year** | **Key Contribution** |
| 1 | A Hybrid Approach for Heart Disease Prediction Using SVM and KNN | S.  Palaniappan,  R. Awang | International Journal of Computer Applications | 2011 | Proposed a hybrid model integrating SVM and KNN to improve heart  disease prediction. |
| 2 | Intelligent Heart Disease Prediction System using Data Mining Techniques | Sellappan Palaniappan, Rafiah Awang | IJCSNS  International Journal of Computer Science and  Network Security | 2008 | Demonstrated the effectiveness of data mining techniques for cardiovascular prediction. |
| 3 | Prediction of Heart Disease using Machine Learning  Algorithms | R. Detrano et al. | Journal of Biomedical Engineering | 1989 | Used the Cleveland dataset to analyze ML-based heart  disease prediction. |
| 4 | An Ensemble-based System for Cardiovascular Disease Prediction | A. Khosla et al. | IEEE  Transactions on Information Technology in  Biomedicine | 2010 | Developed an ensemble learning method combining multipleclassifiers  for better accuracy. |

***A certificate of publication with a blue ribbon and a seal

AI-generated content may be incorrect.(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal Since 2013)***

Impact Factor 8.379

##### DURGESH YADAV

Baderia Global Institute of Engineering & Management, Jabalpur, India

In Recognition of publication of the paper entitled

“Federated Hybrid Machine Learning Approaches for Privacy- Preserving Cardiovascular Risk Assessment"

in IJIRCCE, Volume 12, Special Issue 1, March 2024

***| "1st International Conference on Machine Learning, Optimization and Data Science"|***

***Organized by***

***| Department of Computer Science and Engineering, Baderia Global Institute of Engineering and Management, Jabalpur|***

***(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal Since 2013)***

Impact Factor 8.379

##### DURGESH YADAV

Baderia Global Institute of Engineering & Management, Jabalpur, India

In Recognition of publication of the paper entitled

“Hybrid Deep Learning and Ensemble Models for Early Detection of Cardiovascular Diseases"

in IJIRCCE, Volume 12, Special Issue 1, March 2024

***| "1st International Conference on Machine Learning, Optimization and Data Science"|***

***Organized by***

***| Department of Computer Science and Engineering, Baderia Global Institute of Engineering and Management, Jabalpur|***

**Turnitin abc**

0228IT23MT03 DURGESH YADAV

 slot 12

62 Pages

9,853 Words

58,472 Characters

Document Details

Submission ID trn:oid:::3618:111305845

Submission Date

Sep 8, 2025, 2:42 PM GMT+5:30

Download Date

Sep 9, 2025, 12:28 AM GMT+5:30

File Name

0246CS23MT17 PRAGATI JAIN.pdf

File Size

3.6 MB

14% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

Bibliography Quoted Text Cited Text

Small Matches (less than 45 words)

Match Groups

 17 Not Cited or Quoted 14%

Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%



Matches that are still very similar to source material

0 Missing Citation 0%



Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%



Matches with in-text citation present, but no quotation marks

Top Sources

|  |  |
| --- | --- |
| 8% | Internet sources |
| 1% | Publications |
| 10% | Submitted works (Student Papers) |

Match Groups

 17 Not Cited or Quoted 14%

Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%



Matches that are still very similar to source material

0 Missing Citation 0%



Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%



Matches with in-text citation present, but no quotation marks

Top Sources

|  |  |
| --- | --- |
| 8% | Internet sources |
| 1% | Publications |
| 10% | Submitted works (Student Papers) |

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1

Internet

[pesjournal.net](https://pesjournal.net/journal/v6-n4/10.pdf) 4%

2

Internet

[github.com](https://github.com/dhanushraparthy/HeartDiseaseClassifier/blob/master/mainpro.py) 2%

3

Student papers

iGroup on 2016-02-25 2%

4

Student papers

UT, Dallas on 2018-12-02 2%

5

Student papers

iGroup on 2016-10-19 2%

6

Student papers

University of Lincoln on 2022-09-22 1%

7

Internet

[www.researchgate.net](https://www.researchgate.net/publication/328031918_Machine_Learning_Classification_Techniques_for_Heart_Disease_Prediction_A_Review) 1%

8

Student papers

University of Leicester on 2023-10-01 <1%