Visualizing the Full Spectrum Optimization of K-Nearest Neighbors From Data Preprocessing to Hyperparameter Tuning and K-Fold Validation for Cardiovascular Disease Prediction

Jeena Joseph[*1,2], K Kartheeban[3]
[1]Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnankoil, Tamilnadu, India
[2]Department of Computer Applications, Marian College Kuttikkanam Autonomous,
Kerala, India
[3]Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, Tamilnadu, India
E-mail: jeenajoseph005@gmail.com, k.kartheeban73@gmail.com
[*]Corresponding author

Keywords: K-Nearest Neighbor, machine learning, cardiovascular disease prediction, hyperparameter tuning

Received: December 6, 2024

Cardiovascular disease (CVD) is a prominent cause of death worldwide. This alarming need requires an accurate prediction model using machine learning that can detect and help prevent or mitigate the risk. This study focuses on this issue and has come up with new dimensional capabilities to enhance the K-Nearest Neighbors (KNN) algorithm to predict cardiovascular diseases at an early stage by incorporating various techniques for data preprocessing and feature selection thereby improving the efficiency of the model. The proposed model identifies the most relevant features using Principal Component Analysis. The main innovation revolves around fine tuning the hyperparameter of K-Nearest Neighbors, specifically the choice of neighbors (K), using a data driven approach to ensure accuracy across different datasets. The performance of the optimized K-Nearest Neighbors algorithm is evaluated using the Framingham heart disease dataset. This model achieved an impressive prediction accuracy of 92.46% and outperformed methods that solely rely on traditional K-Nearest Neighbors. As machine learning techniques plays an important role in the development of prediction models for early detection and prevention of cardiovascular disease, this model can be considered as a valuable tool for healthcare professionals and researchers. The core contribution of this study lies in offering a comprehensive optimization of the traditional K-Nearest Neighbors (KNN) algorithm. This includes robust data preprocessing using the Hampel filter for outlier removal, feature selection through Principal Component Analysis (PCA), and performance enhancement using grid search for hyperparameter tuning combined with 10-fold cross-validation. Unlike prior studies that apply KNN with minimal adjustments, this research emphasizes the importance of an end-to-end machine learning pipeline. This holistic refinement significantly improves the predictive performance and reliability of KNN for cardiovascular disease prediction, achieving 92.46% accuracy on the Framingham dataset.

Povzetek: Raziskava predstavlja optimiziran KNN-algoritem za napoved srčno-žilnih bolezni, ki s PCA, čiščenjem podatkov in 10-kratno validacijo doseže zelo kvalitetno delovanje.

1 Introduction

With a huge impact on global death rates, cardiovascular disease continues to be a major health concern [1], [2]. To mitigate the risk factors associated with cardiovascular disease, early and accurate prediction models are required. Due to the technological advancements and increase in electronic health records, machine learning has become a viable tool for predictive analytics in the healthcare sector [3].

According to the World Health Organization, cardiovascular disease (CVD) accounts for approximately 17.9 million deaths annually, constituting about 32% of all global deaths. The economic impact is equally staggering, with estimated global costs projected to surpass $1 trillion by 2030. While machine learning techniques such as K-Nearest Neighbors (KNN) have been explored for disease prediction, their application to CVD data presents unique challenges. Traditional KNN models often suffer from high sensitivity to noisy data, computational inefficiency with large datasets, and reduced accuracy in high-dimensional spaces—limitations particularly evident when applied to complex medical datasets like the

Framingham Heart Study. This study seeks to overcome these challenges by proposing a fully optimized KNN pipeline tailored for CVD prediction. By integrating outlier removal using the Hampel filter, dimensionality reduction through PCA, and hyperparameter tuning with grid search and k-fold validation, this work fills a crucial gap in the literature where previous models lacked end-to-end optimization. The proposed enhancements are not generic but specifically address the data quality, dimensional complexity, and class imbalance issues inherent in the Framingham dataset, resulting in a significantly improved accuracy of 92.46%. This provides a strong foundation for clinical decision support tools and highlights the practical value of optimized KNN in real-world healthcare applications.

The K Nearest Neighbors algorithm is a type of supervised machine learning technique that involves dividing a dataset into groups or clusters based on the distances between data points. It has become quite popular because it is simple and has the capability to carry out classifications effectively [4]. Its effectiveness has been acknowledged in situations where there is a connection, between the variables or when the data cannot be easily divided in a linear manner. However, because it relies on the dataset it can be computationally expensive, especially when working with large datasets and it may also be affected by the challenge posed by high dimensional data. [5].

The main objective of this study is to create a model that can accurately identify the risk of heart disease. To achieve this, the K Nearest Neighbors (KNN) algorithm in enhanced by incorporating different optimization techniques. These techniques include data preparation methods such as outlier detection, dimensionality reduction using Principal Component Analysis (PCA), tuning hyperparameters through grid search and implementing k fold cross validation. This research focuses on assessing the practicality of using an optimized K Nearest Neighbors (KNN) model to predict heart disease. It evaluates performance metrics, like F1 score, recall, accuracy and precision. The study primarily concentrates on implementing an enhanced KNN algorithm that has shown promising advancements in predicting diseases. The goal of this approach is to provide accurate risk assessments that are relevant, to clinical settings. By overcoming the limitations of KNN models our aim is to improve treatment and reduce the healthcare costs and burdens associated with cardiovascular diseases. This study is organized into different sections. The review of literature provides a comprehensive summary and analysis of existing research and literature. The applied methodology is explained in the materials and methods. The next section discusses the data preprocessing steps. Then the optimized KNN algorithm is demonstrated and finally, the study spotlights the exploratory data analysis and results.

While previous research has shown moderate success using KNN for CVD prediction, this study distinguishes itself by addressing key limitations through systematic enhancements across the entire predictive pipeline. Specifically, this includes (1) handling missing values and outliers using robust statistical techniques like the Hampel filter; (2) applying PCA to reduce dimensionality and improve learning efficiency; and (3) optimizing the model via grid search with k-fold cross-validation to ensure generalizability. These components, when integrated, offer a fine-tuned and scalable approach that improves upon standard KNN performance, making it a practical solution for clinical use.

2 Review of literature

Machine learning algorithms have been used more frequently in a variety of healthcare applications, particularly in cardiology. In areas with limited healthcare resources, advanced prediction algorithms are especially important for identifying individuals at risk of heart failure and one of the main causes of death worldwide is heart failure [6], [7]. The study by Nagavallika discusses the prediction of heart disease using machine learning techniques, specifically the use of a hybrid random forest with a linear model (HRFLM) that achieves an accuracy of 88.7% [8]. Dimopoulos et al. assessed K-Nearest Neighbor, Random Forest, and Decision Tree, three well-liked machine learning models, using the ATTICA dataset. Results show that the Random Forest model performs much better than HellenicSCORE. It also demonstrated the model's accuracy in smaller datasets and its ability to comprehend the nuances of traits associated with CVD even with a lesser number of data points [9]. Another study discusses the use of machine learning algorithms such as SVM, KNN, RF, J.48, and MLP for predicting heart disease. It also mentions the importance of balancing the dataset for accurate prediction [10].

K-Nearest Neighbors (KNN) has been widely recognized for its simplicity, non-parametric nature, and interpretability, which are valuable in medical applications. Its ability to function without making prior assumptions about data distribution makes it particularly suitable for heterogeneous healthcare datasets. However, KNN also presents notable limitations. Its sensitivity to irrelevant features and outliers, along with computational inefficiency in high-dimensional datasets, can hinder performance—especially in complex medical data such as the Framingham Heart Study. These limitations motivate the need for preprocessing, feature reduction, and hyperparameter optimization.

Jin B et al. used sequential modelling with neural networks to create an electronic health record model that captured the sequential nature of healthcare data, including changes in lifestyle across time. With the use of word vectors and one-hot encoding, the method looks promising for heart failure prediction by looking for sequential patterns in medical data and producing diagnostic scenarios [11]. In this paper, the authors used machine learning methods and Python programming to study heart disease prediction using a dataset consisting of 12 parameters as well as 70000 unique data values, and the main goal of this study is to increase the accuracy of heart disease detection by using algorithms where the target output determines whether the subject has heart disease. From the study, it is concluded that, decision trees can lead to inaccurate results when applied to small datasets. Naive Bayes is more accurate and can be combined with K-means for better accuracy [12].

Previous studies using KNN for cardiovascular disease (CVD) prediction have shown mixed results. For example, some reported accuracy near 84% to 90%, yet lacked advanced preprocessing techniques such as normalization, outlier handling, or feature selection. Many of these models used default K values and did not tune hyperparameters or validate results through cross-validation, limiting their generalizability. This underscores the need for a more systematic and optimized application of KNN for robust medical predictions.

In a study by Shah et al., the classification algorithms like random forest, decision trees, K-nearest neighbor (KNN), and Naive Bayes were experimented on a dataset of 303 samples having 17 attributes and the KNN model achieved the highest accuracy of 90.8% [13]. Boosted decision tree algorithms have shown promise in correlating patient characteristics with mortality risk, achieving an AUC of 0.88 [14]. Pires et al. explored various machine learning methods, achieving a maximum accuracy of 87.69% for heart disease prediction [15]. However, the study's validity was constrained by the limited sample size. Ali et al. reported a 100% accuracy using Random Forest, Decision Trees, and KNN algorithms, albeit focusing on optimal cross-validation results rather than robust, conclusive findings [16]. A heart disease profiling model were developed Kahramanli et al. by using a hybrid artificial neural network with fuzzy logic. While models that incorporate both neural networks and fuzzy logic concepts have achieved an 86.8% accuracy rate, none have been permanently validated for other diseases like diabetes. They have used k-fold cross-validation for the classification purpose and also tried this model on the diabetes dataset, achieving a performance of 84.24% [17]. Faiayaz et al. improved the accuracy of KNN by 5.68%, and a hybrid Random Forest and Linear Model technique reached an 88.7% precision on a dataset of 297 records [18], [19].

The issue of high dimensionality in health data has been addressed by employing Principal Component Analysis (PCA) for dimensionality reduction, retaining significant variance within fewer components. PCA's effectiveness was further corroborated by a study using it alongside unsupervised learning techniques, with NN classifiers, achieving a high F1 score in classifying cardiac arrhythmia with minimal components, indicating PCA's robustness in feature extraction [20], [21]. Additionally, the PCA-KNN method has been applied to medical imaging, resulting in significant accuracy for scaling diverse medical images, underscoring the adaptability of PCA in medical diagnostics [22], [23]. A deep learning technique that applies an artificial neural network algorithm with a hidden layer technique in making a heart disease prediction model was proposed by Yuda Syahidin et al., which yielded 90% accuracy [24]. Wang et al. mentioned about the center loss to enable the neural network to learn discriminative features and separate samples from different categories, which can effectively improve heart disease prediction [25].

Hybrid models, integrating the power of multiple machine learning algorithms, have also demonstrated significant performance. Sharanyaa et al. proved the higher performance of a hybrid method that combines Support Vector Machines (SVM) and Naive Bayes [26]. Moreover, in diagnosing heart disease, ensemble methods have generally better performance. It is a combination of various machine learning algorithms set up for this purpose. In another study, ensemble model, comprising multiple machine learning techniques, showed improved effectiveness [27]. Studies comparing ML classifiers like Sequential minimal optimization (SMO), naïve Bayes, and J48 decision trees in cardiovascular risk prediction found SMO to be the most accurate, suggesting its robustness [28]. Deep learning applications, including Convolutional Neural Networks (CNNs) for early heart failure detection via ECGs, have achieved a 0.78 AUC [29] and adaptive multi-layer networks have outperformed classical and hybrid models [30].

Advancements in medical imaging for brain tumor detection involve combining CNN with auto-context techniques, using multi-dimensional image patches for improved accuracy [31]. The Intelligent Deep Residual Network based Brain Tumor Detection and Classification (IDRN-BTCC) method, a novel approach for brain tumor classification using residual networks and multilayer

perceptron, has shown efficacy, enhanced by chicken swarm optimization [32]. Shankar et al.'s convolutional neural network algorithm, using structured and unstructured patient data, predicts heart disease risk with 85 to 88% accuracy [33]. Dutt et al. developed a CNN for class-imbalanced datasets, classifying 77% of positive cases and 81.8% of negative cases accurately [34]. Another study presents an Integrated Deep Learning Model with Convolution Neural Network (IDLM CNN) for heart disease prediction using various medical data sets. This model convolves the features of lungs and combines with other features to compute Disease Prone weight towards cardiac disease and the proposed model improves heart disease prediction accuracy. The False ratio is reduced with the integrated model [35].

Table 1 provides a comprehensive overview of the data sets, attributes, machine learning methods, and corresponding accuracy values employed in historical and contemporary heart disease prediction research. While ensemble and deep learning approaches offer high accuracy, they often sacrifice transparency, scalability, or require significant computational resources. Our study presents an optimized KNN framework that preserves simplicity while achieving competitive performance. By integrating PCA for dimensionality reduction, Hampel filtering for outlier removal, and grid search with k-fold cross-validation for tuning, we address known limitations of traditional KNN and demonstrate its practical value in clinical decision-making contexts.

In summary, this study addresses key limitations in the existing literature, especially the under-optimized use of KNN in CVD prediction. By implementing a robust end-to-end pipeline—from data cleaning and feature selection to hyperparameter tuning—our model not only improves prediction accuracy but also contributes a reproducible methodology for medical risk assessment. The following section elaborates on our methodological framework, which is tailored to overcome the gaps identified in prior research.

Table 1: Comprehensive summary of datasets, attributes, machine learning algorithms, and accuracy values in heart disease prediction research over time

Research Article	Algorithm Used	Dataset Used	Attributes /Parameters	Accuracy	Novelty / Limitations
[36]	Multilayer Perceptron	Multiple Datasets (Cleveland, Hungarian, Switzerland, Long Beach, StatLog)	Infinite Feature Selection	87.70%	The study proposes a novel heart disease prediction model using adaptive infinite feature selection with deep neural networks to enhance precision and sensitivity, though its accuracy (87.7%) is limited by small, diverse datasets and potential overfitting.
[37]	Sequential Minimal Optimization (SMO)	Cleveland heart dataset	Full Set and Optimized Attribute Set	85.148% using the full set of attributes and 86.468% using the optimal attribute set	The study combines multiple machine learning classifiers with attribute evaluators and hyperparameter tuning to improve heart disease prediction, but its accuracy remains moderate at 86.468% and is limited by reliance on a single dataset.
[38]	Logistic Regression, SVM, KNN, GNB, MNB, DT	Self-Augmented Datasets of Heart Patients (UCI Dataset and Local Dataset)	Anaemia, Diabetes, High_blood_pressure, Sex,Smoking, Time (Follow-up period), Death_event	Logistic regression (82.76%), SVM (67.24%), KNN (60.34 %), GNB (79.31 %), MNB (72.41%), ET (70.31%), RF (87.03%), GBC (86.21%), XGB (84.48%) LGBM (86.21%)	The study uses a self-augmented dataset approach—expanding heart disease data synthetically—and applies multiple ML models, improving prediction accuracy through enhanced data diversity. The effectiveness of synthetic data may not generalize well to real-world scenarios, and the study lacks external validation on independent datasets.
[39]	Logistic Regression with PCA and Ensemble Classifiers	Cleveland Heart Disease Dataset	Complete Set and Optimized Attribute Set	85.8%	The study improves heart disease prediction by applying Principal Component Analysis with machine learning models, achieving 85.8% accuracy, but its reliance on transformed features and a limited dataset reduces interpretability and generalizability.

[40]	Hybrid Random Forest with Linear Model (HRFLM)	UCI Cleveland dataset	13 clinical features	88.7%	The study enhances heart disease prediction by using a hybrid model combining random forest and linear models, achieving 88.7% accuracy, but its performance may be limited by dataset scope and lack of external validation.
[41]	Voting Ensemble	UCI Dataset	Optimal set of attributes	91.96%	The study proposes an ensemble model using stacking and voting techniques for heart disease prediction, achieving high accuracy (91.96%) and F1 score (91.69%) on the UCI dataset, though it performs less effectively on the Framingham dataset, indicating limited generalizability.
[42]	SVM	Framingham Heart Study	six highly correlated features	67%	The study develops a heart disease prediction model using correlation-based feature selection and achieves 67% accuracy with SVM on oversampled Framingham data, though the modest performance suggests limitations due to dataset imbalance and limited predictive power of selected features.
[43]	Logistic Regression	Framingham Heart Study	Complete set of attributes	85.063%	The study compares machine learning and deep learning models for predicting 10-year coronary heart disease risk using the Framingham dataset, with logistic regression achieving the best accuracy (85.06%), though performance differences across models were minimal and generalizability was not validated externally.
[44]	Auto Encoder-Based Kernel SVM	Framingham Heart Study	Complete set of attributes	87.14%	The study introduces an IoT-based RHMIoT framework combining deep learning and autoencoder-based ML to monitor and predict cardiovascular disease severity, achieving 87.14% accuracy, though performance may vary due to dependence on a single dataset (Framingham) and limited real-world testing.
[45]	Gradient Boosting Classifier (GBC)	Framingham Heart Study	Optimal set of attributes	87.61%	The study improves heart disease prediction by applying p-value-based backward feature elimination with several ML algorithms, achieving 87.61% accuracy using gradient boosting, though limited to a single dataset and potentially impacted by reduced feature interpretability.

3 Materials and methods

The methodology adopted in this study follows a structured and modular flow encompassing key stages: data acquisition, preprocessing (including missing value treatment and outlier handling), feature normalization and The proposed methodology used in this study is demonstrated in Figure 1. The proposed model consists of a number of rigorous steps including data acquisition, preprocessing, dimensionality reduction and feature scaling, application of optimized KNN algorithm and performance evaluation. The Kaggle repository is used to acquire the famous publicly available Framingham heart disease dataset that is used for evaluating the effectiveness of the optimized KNN algorithm. The dataset is cleaned and transformed data preprocessing for making it suitable for further analysis. The missing values are handled using mean imputation technique that replaces them with the attribute mean. The Hampel filter, based on the Median Absolute Deviation, is then used to handle outliers to minimize the probability that outliers could skew the results. Feature scaling is done using the min-max normalization technique which maintains the range of data points in the dataset. Principal component analysis (PCA) is performed for dimensionality reduction. PCA reduces the dataset to its most significant components by preserving as much of the original data variation as is possible. The amount of computational overhead for handling the dataset in the next machine- learning phases can be minimized through this process.

After data preprocessing, the study optimizes the traditional KNN algorithm and changes can boost processing capacity or forecast precision. The KNN algorithm clusters data points together in the feature space using the characteristics of their nearest neighbors. Finally, the success of the model is assessed in performance evaluation. The dataset is split into test and training sets in order to evaluate the model's predictive capacity for patient heart disease. Following its application to the test set, the model is assessed using pertinent metrics, including recall, accuracy, precision, F1 score, and others. The research leveraged the functionalities of RStudio and the R programming language.

dimensionality reduction, classification using the KNN algorithm, and evaluation using standard performance metrics. Each of these stages is described in detail in the following subsections, with a flowchart summarizing the process in Figure 1.

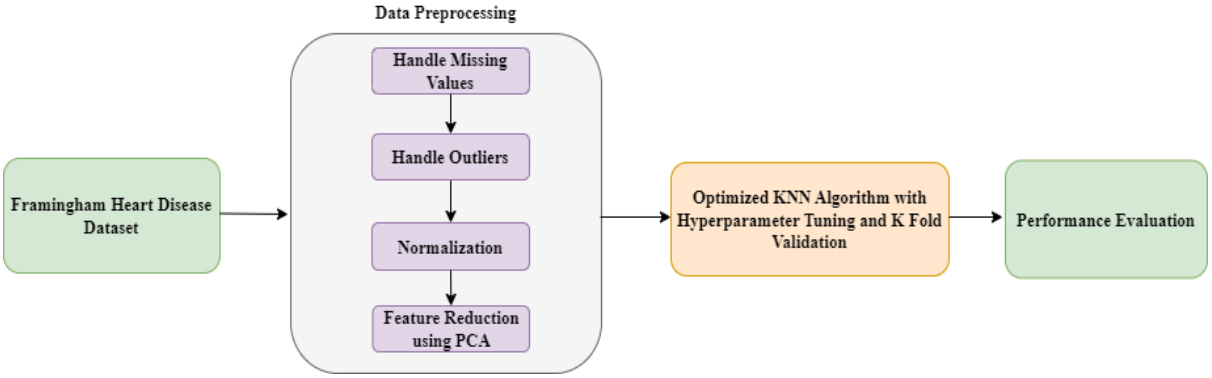

Figure 1: Proposed methodology.

4 About the dataset

The "Framingham" heart disease dataset contains more than 4,240 entries, with 16 columns and 15 characteristics. It aims to forecast whether a patient is at risk of developing coronary heart disease (CHD) within the next ten years. Each attribute within the dataset represents a potential risk factor, spanning various demographic, behavioral, and medical factors. The Framingham Heart Study is a long-term cardiovascular cohort study initiated in 1948 in Framingham, Massachusetts, involving over 5,000 men and women aged 30–62. It aimed to investigate factors contributing to cardiovascular disease development. While it has generated valuable clinical insights, the dataset is predominantly composed of white, middle-class individuals, which may limit its generalizability to diverse populations. However, its depth, quality, and widespread use make it a reliable benchmark for model evaluation. Although other datasets such as the UCI Cleveland dataset are available, they are smaller and less comprehensive, justifying the use of the Framingham dataset in this study.

Attributes:

Demographic:

- Gender: Categorized as either male or female (Nominal)

- Age: Represents the individual's age, considered a continuous variable (Even though ages are rounded to the nearest whole number, age itself is continuous)
- Education: Specific details about education are not provided.

Behavioral:

- Current Smoker: This variable denotes whether the patient is currently engaged in smoking (Categorical).
- Cigarettes Per Day: This variable quantifies the average number of cigarettes an individual smoke daily. (Treated as continuous since it can encompass any numerical value, including fractional quantities.)

Information on medical history:

- BP Meds: Shows whether the patient is currently using medication for controlling blood pressure or not (Categorical)
- Prevalent Stroke: Indicates whether the patient has a history of stroke (Categorical)
- Prevalent Hyp: Indicates whether the patient has been diagnosed with hypertension (Categorical)
- Diabetes: Indicates whether the patient has been diagnosed with diabetes (Categorical)

Information about the patient's current health status:

- Total Cholesterol (Tot Chol): This represents the continuous measurement of the total cholesterol level.
- Systolic Blood Pressure (Sys BP): This indicates the continuous measurement of systolic blood pressure.
- Diastolic Blood Pressure (Dia BP): This records the continuous measurement of diastolic blood pressure.
- Body Mass Index (BMI): BMI is calculated and measured as a continuous variable, reflecting the patient's body mass in relation to their height.
- Heart Rate: The heart rate is recorded as a continuous variable, acknowledging its wide range of potential values in medical studies.
- Glucose: The continual monitoring of the patient's glucose levels in their healthcare records.

Target variable:

The target variable portrays the ten-year likelihood of cardiovascular disease and it is represented in binary notation. A value of 0 implies a negative risk and a value of one a positive risk.

5 Data preprocessing

In data preprocessing, the raw data is cleansed and transformed into information that is useful for model training, and it is a pivotal step before using machine learning techniques. Data preprocessing has to be done carefully because it has adverse effects on the performance and calibre of the machine learning models.

5.1 Handling missing values

Missing values in a dataset may produce biased reults and it has adverse impact on the performance and reliability of machine learning models.Therefore, handling missing values is considered as an essential step in the data preprocessing pipeline. The renowned Framingham dataset has 4,240 records with 15 distinct features that are often utilized in cardiovascular research. The analysis revealed that 645 records were incomplete, indicating missing values within the data. Figure 2 presents a meticulous mapping of these gaps, shedding light on the magnitude and pattern of the missing information across various attributes. Attributes such as 'TenYearCHD,' 'diaBP,' 'sysBP,' 'diabetes,' 'prevalentHyp,' 'prevalentStroke,' 'currentSmoker,' 'age,' and 'gender' exhibit complete data (0% missing entries). The 'heartRate' attribute shows an insignificantly small fraction of missing data, at 0.02%. Other attributes like 'BMI,' 'cigsPerDay,' 'totChol,' and 'BPMeds' show a larger incidence of missing data, between 0.45% and 1.25%. The 'education' attribute has 2.48% of its values missing. Notably, the 'glucose' feature experiences the highest level of missing entries, standing at 9.15%. To address these gaps in the dataset, multiple techniques are available. For this study, the strategy of mean imputation has been applied. Mean imputation is a statistical technique used to fill in missing values in a dataset by replacing them with the mean (average) of the available values for a specific variable. Figure 3 shows all features having 0% missing data, which suggests that mean imputation has been used to fill in all the missing values for each feature with the mean value of that feature. The result is a dataset with no apparent missing data. Mean imputation was chosen for its simplicity and effectiveness in cases of low missingness, particularly where the feature distribution is symmetric or approximately normal. Although methods like KNN imputation or multiple imputation offer more sophistication, they are computationally intensive and less suitable when missingness is minimal. For instance, attributes such as glucose (9.15% missing) and education

(2.48%) were imputed using the mean to preserve sample size while avoiding bias.

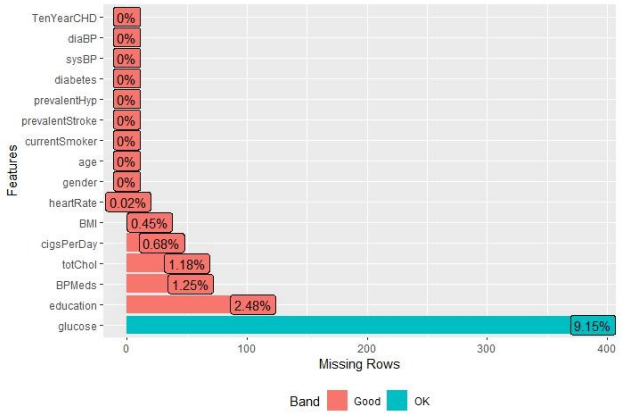

Figure 2: Visual representation of missing values in various features of the Framingham dataset

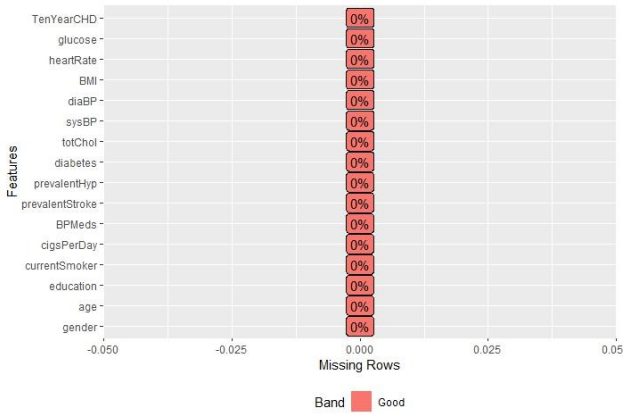

Figure 3: Visual representation of missing values after performing mean imputation.

5.2 Handling outliers

Addressing outliers in the data is crucial as they can considerably influence the outcomes and effectiveness of both statistical analyses and machine learning models. Outliers refer to data points that deviate substantially from the majority of the dataset.

Figure 4 shows a box plot of six variables: sysBP, totChol, diaBP, BMI, heartRate, and glucose. The variable sysBP (Systolic Blood Pressure) has a wide interquartile range (IQR), the distance between the first and third quartiles, indicating variability in the data. Several outliers above the upper whisker suggest that some individuals have unusually high systolic blood pressure readings. The total cholesterol (totChol) levels also display a relatively wide IQR, suggesting variability. There are outliers on both ends, indicating that there are individuals with unusually high and low cholesterol levels. The diastolic blood pressure (diaBP) levels show a smaller IQR compared to systolic blood pressure, indicating less variability. There are a few outliers, particularly on the higher end. The BMI box plot shows a moderate IQR. There are several outliers on the upper side, indicating that there are individuals with a BMI much higher than the average. When compared with sysBP and diaBP, the attrbute "heartRate" has a lower IQR. Although, there exist outliers,they are not very noticable. The values of glucose are closer to the median thereby indicating the smallest IQR. But there exist number of high outliers showcasing the elevated sugar levels. There is notable variance in the total cholestrol, systolic blood pressure and diastolic blood pressure.

In this study, outliers are identified and removed using the famous statistical approach called Hampel filter based on the Median Absolute Deviation (MAD). When compared to the standard deviation, MAD is less prone to outliers. So, it is especially helpful when dealing with data that may not be regularly distributed or in situations where the existence of outliers might cause the standard deviation to be skewed. The Hampel filter was selected over traditional z-score methods due to its robustness against non-normal data distributions, which are common in medical datasets. An outlier was defined as any data point beyond ±3 times the median absolute deviation (MAD) from the median. Rather than removing these values, we clipped them to the calculated bounds to prevent loss of potentially valuable information and maintain the dataset's integrity.

The steps that are performed during this method include:

- Sliding Window: For every data points of interest, the hampel filter select a window of data points around them. The sliding window is normally symmetric and contains many data points before and after the current point.
- Calculation of the Median: The filter calculates the median of the data points within this window.
- Calculation of the MAD: The absolute differences between the values of an attribute and the median of that attribute is detemined for the calculation of MAD.
- Setting the Bounds: The range of data variation that is considered as normal is defined by setting the upper and lower boundaries. The bounds are set at three times the MAD below and above the median. The lower bound is calculated as the median minus three times the MAD. The upper bound is the median plus three times the MAD.
- Identification of Outliers: It identifies which attribute values fall outside these bounds.

- Replacement of Outliers: It replaces the attribute values identified as outliers with the nearest boundary value.

This MAD-based method is a good choice for outlier detection when the data may not be normally distributed because it is based on the median, which is a robust measure of central tendency that is not affected by extreme values as much as the mean.

Figure 5 shows a grid of six histograms, each depicting the distribution of values for a different biomedical metric. These histograms are useful for identifying the range of values and potential outliers within each category. The categories presented are BMI, diaBP (diastolic blood pressure), glucose, heartRate, sysBP (systolic blood pressure), and totChol (total cholesterol). The observations based on the histograms are:

- BMI: The distribution is somewhat right-skewed, indicating that most individuals have a BMI within the normal to overweight range, but there are some with high BMIs indicative of obesity.
- Diastolic Blood Pressure (diaBP): The distribution appears approximately normally distributed, with most values centering around the median. There are a few potential outliers on the higher end.
- Glucose: This histogram is heavily right-skewed, with most individuals having glucose levels in the normal range, but there is a long tail to the right, indicating some individuals with very high glucose levels, which may suggest diabetes or other metabolic disorders.
- Heart Rate (heartRate): Most of the heart rate values are clustered in the middle range. The distribution is almost normal with a slight right skew.
- Systolic Blood Pressure (sysBP): The distribution is right-skewed, with a peak in what might be considered the high-normal range and some individuals with particularly high systolic blood pressure values, potentially indicating hypertension.
- Total Cholesterol (totChol): This distribution is roughly normal but with a slight right skew, suggesting that while most individuals have cholesterol levels within the normal range, there are some with high cholesterol levels.

Each histogram is labeled with "count" on the y-axis, representing the number of observations within each bin of the histogram, and "Value" on the x-axis, representing the range of values for the metric in question. These visualizations help in understanding the overall health profile of a population or a sample of individuals, particularly in pinpointing common ranges for these health metrics and identifying outliers that might warrant further investigation or intervention.

The outlier handling procedure has determined specific lower and upper bounds for key variables in the dataset. For "totChol," the calculated bounds are 150 as the lower limit and 318 as the upper limit. Similarly, for "sysBP," the lower bound is set at 89, while the upper bound is established at 167. The variable "diaBP" is subject to lower and upper bounds of 59.5 and 104.5, respectively. "BMI" adheres to limits of 17.94 as the lower bound and 32.88 as the upper bound. The heart rate variable, denoted as "heartRate," is constrained between 54 and 96. Finally, "glucose" follows boundaries of 59 (lower limit) and 101 (upper limit). These bounds serve as thresholds for identifying and handling outliers in the respective variables, contributing to the robustness of data analysis and model building.

The updated box plots in Figure 6 show that the outliers have been handled, as there are no longer points beyond the whiskers. The scale of the y-axis has changed for some metrics, indicating that the maximum values are lower, which is consistent with the removal of high outliers.

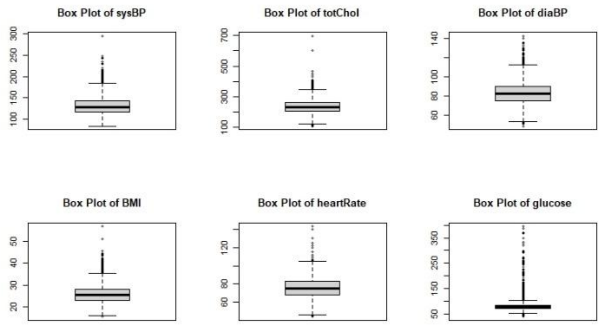

Figure 4: Box Plot before handling outliers in the dataset

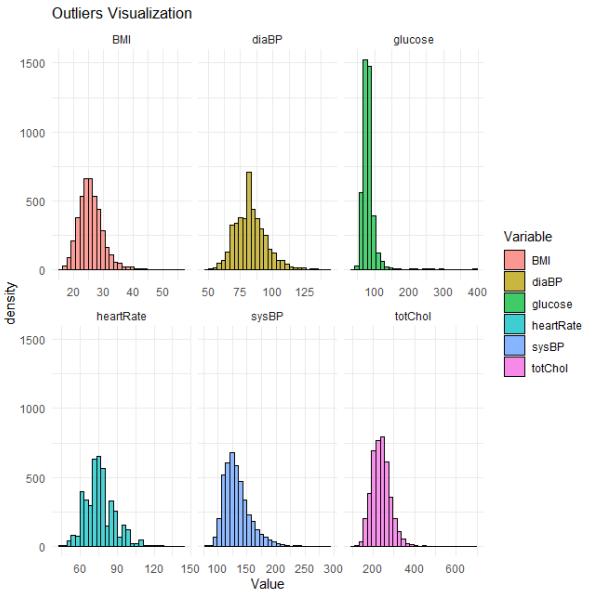

Figure 5: Histogram visualization of outliers

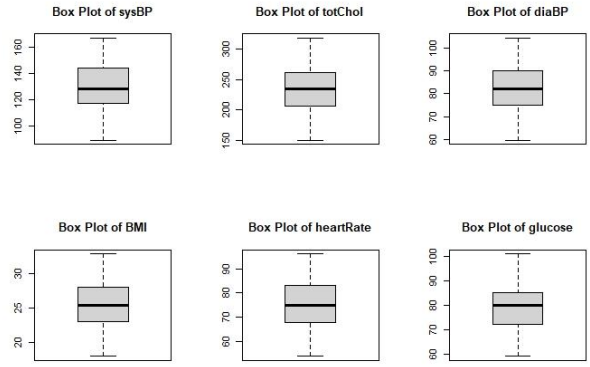

Figure 6: Box plot after handling outliers in the data

5.3 Normalization

Normalization serves as a data preprocessing method aimed at scaling and standardizing the features or variables within a dataset. Its primary objective is to place all variables on a uniform scale, facilitating comparisons and frequently enhancing the performance of machine learning algorithms. Min-Max normalization, often called feature scaling or min-max scaling, is employed in this study and using this method, the values of numerical variables are converted into a predetermined range, usually between 0 and 1. By reducing the influence of outliers, Min-Max normalization helps to make data more comparable by guaranteeing that different variables have an identical scale. For normalizing a single variable, the following formula is used:

$$X_{normalized} = (X-X_{min})/(X_{max}-X_{min})$$

Where:

$X_{normalized}$ indicates the normalized value of the variable X.

X is the original value of the variable.

X_{min} denotes the minimal value of the variable X in the dataset.

X_{max} denotes the maximal value of the variable X in the dataset.

After applying Min-Max normalisation, the transformed values will lie between 0 and 1, where 0 denotes the variable's lowest value in the dataset and 1 its highest value. The relative positions of each data point inside the initial range are represented by values ranging from zero to one. Figure 7 shows box plots for six biomedical metrics before and after normalization. After normalization, all values are adjusted to fit within a similar scale, between 0 and 1. Min-max normalization was preferred over standardization (z-score normalization) because it preserves the original distribution shape and maps features to a common scale [0,1], which is particularly advantageous for distance-based algorithms like KNN. It helps prevent features with larger numeric ranges from dominating the distance calculations, thereby improving classification accuracy.

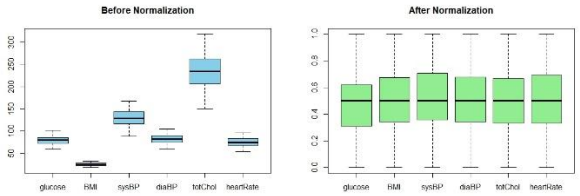

Figure 7: Dataset before and after normalization

5.4 Principal component analysis

Principal Component Analysis (PCA) is a crucial method in both data analysis and machine learning, designed to reduce the dimensionality of datasets with numerous variables. It achieves this by identifying orthogonal axes, termed principal components, which effectively capture the primary sources of variation in the data. The first principal component accounts for the highest variance, the second for the second-highest, and so forth. PCA proves valuable in streamlining data, facilitating visualization and processing, and is frequently employed as a preprocessing measure to improve the efficacy of machine learning models.

The dataset originally consisted of 15 attributes, but the 'education' column was excluded on the grounds that it has no impact on heart disease. Following the application of Principal Component Analysis (PCA) and the determination of PCA scores, eight attributes were chosen. The output of PCA shows the loadings (also known as

eigenvectors) for each principal component. Loadings are coefficients that represent how much weight each original variable contributes to each principal component. PC1 to PC8 are the principal components. Within Principal Component Analysis (PCA), the initial principal component (PC1) captures the highest variance present in the data, and each succeeding component captures the majority of the remaining variance while maintaining orthogonality to the preceding components. In this study, cigsPerDay, BPMeds, totChol, sysBP, diaBP, BMI, heartRate, and glucose are the original variables that have been transformed into principal components.

In PC1, "cigsPerDay" exhibits a notable positive loading, indicating a significant impact, while "sysBP" and "diaBP" contribute negatively. Compared to positive loadings from "glucose" and "totChol," PC2 is mostly affected by negative loadings from "heartRate" and "cigsPerDay." The PC3 loadings for "totChol" and "cigsPerDay" are positive, whereas "glucose" has a notable negative loading. The loadings from "sysBP" and "diaBP" are positive; however, "BPMeds" significantly decreases PC4. While "glucose" and "totChol" exhibit negative loadings in PC5, "cigsPerDay" and "BMI" have positive loadings. Positive loadings from "heartRate" and "cigsPerDay" are positively correlated with PC6, whereas negative loadings from "glucose" and "BMI" stabilise the situation. PC7 shows positive loadings from "cigsPerDay" and "sysBP," in addition to negative loadings from "heartRate" and "BMI." Finally, a notable negative loading from "sysBP", predominating PC8, contrasts with a strong positive loading from "diaBP". The aforementioned analysis offers insights into the key elements that contribute to each principal component, which aids in the understanding of the deeper trends and patterns in the dataset. PCA was used to retain eight components which collectively explain over 97.3% of the variance in the dataset, as shown in Table 2 and the accompanying bi plot (Figure 8). The retained components capture most of the essential variability while reducing redundancy and noise. PCA was selected over feature selection methods like decision tree-based importance due to its ability to decorrelate variables and improve computational efficiency in high-dimensional data.

Detailed information on the significance of the principal components is provided in Table 2. For every principal component, it reveals the variation in proportion, cumulative variance proportion, and standard deviation. The standard deviation, which expresses the variance that each main component captures, is the square root of its eigenvalue. If a component has a larger standard deviation, it is considered to account for more volatility in the dataset. Variance is the proportion of the total variance of the dataset that each primary component explains. It is calculated by squaring the component's standard deviation and dividing the result by the total of all the eigenvalues. Cumulative Proportion is the total variance captured by all the principal components up to and including the current one. It is a running total of the 'Proportion of Variance' and shows how much of the total variance is explained by the combined effect of all the principal components up to that point. PC1 captures the most variance by far, with about 28.41% of the variance. This is a significant amount, suggesting that PC1 represents a meaningful underlying pattern in the data. PC2 accounts for an additional 13.77% of the variance, bringing the cumulative total to 42.18%. PC3 adds another 13.08% of the variance, resulting in a cumulative proportion of 55.25%. PC4 through PC7 gradually contribute less and less, with PC4 adding 11.46%, PC5 adding 11.38%, PC6 adding 10.19%, and PC7 adding 9.092% of the variance, respectively. PC8 contributes the least to the variance (2.631%), and it is often the case that later components account for less variance as the most significant patterns are captured by the initial components.

Each principal component comprises a combination of attributes contributing to cardiovascular risk. For example, PC1 is significantly influenced by cigarette consumption and blood pressure, which are known predictors of heart disease. Descriptive statistics such as variance, standard deviation, and range for each attribute are provided in Table 2, offering insights into their original distributions.

Table 2: Importance of Principal Components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.5076	1.0495	1.0228	0.9574	0.9539	0.9029	0.85284	0.45882
Proportion of Variance	0.2841	0.1377	0.1308	0.1146	0.1138	0.1019	0.09092	0.02631
Cumulative Proportion	0.2841	0.4218	0.5525	0.6671	0.7809	0.8828	0.97369	1.00000

Figure 8 shows a Bi plot representing the distribution of data after Principal Component Analysis (PCA) has been conducted. In the scatter plot, the axes are labeled `PC1` and `PC2`, which stand for Principal Component 1 and Principal Component 2, respectively. These two principal components are the new axes in a two-dimensional feature space onto which the original data has been projected. The points on the plot represent individual data items in terms of their `PC1` and `PC2` scores, which are the coordinates of each point in the new feature space. Red lines emanating from the origin point to the position of the original variables (like `cigsPerDay`, `BPMeds`, `totChol`, `sysBP`, `diaBP`, `BMI`, `heartRate`, `glucose`) on this plane. The direction and length of these lines indicate how each variable correlates with the principal components: the longer the line, the more the variable influences that principal component. The angle between the lines suggests whether the correlation between variables is positive (lines more closely directed), negative (lines more divergent), or neutral (lines are perpendicular).

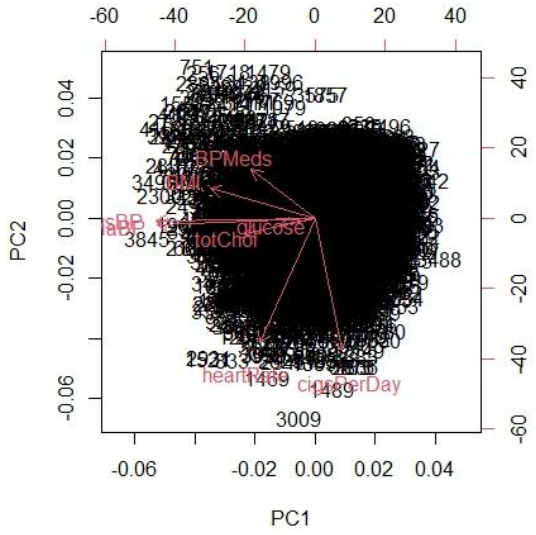

Figure 8: Biplot of PCA

6 Classification using optimized K-Nearest Neighbors (KNN) algorithm

The novelty of this study does not rest solely on the use of the KNN algorithm, which is a well-established classification method, but on how the algorithm has been carefully adapted and optimized for medical data. The proposed model systematically incorporates advanced preprocessing techniques, such as Hampel filter-based outlier removal, Min-Max normalization, and PCA for dimensionality reduction. Moreover, the study introduces a grid search-based strategy for hyperparameter tuning, complemented by 10-fold cross-validation, to empirically identify the most effective K value. These innovations collectively enhance both the accuracy and interpretability of the KNN model in cardiovascular disease prediction.

The K-Nearest Neighbors (KNN) algorithm is a straightforward and easy-to-understand machine-learning classification method suitable for both supervised and unsupervised tasks. Operating as a non-parametric and instance-based learning approach, KNN refrains from assuming any specific characteristics about the data distribution, relying on predictions derived from the similarity between data points. The key steps involved in the optimized KNN classification algorithm are:

- Loading the Data set: Load the dataset for data analysis, visualization, and modeling.
- Handling Missing Data: Identify missing values in the dataset and impute missing values for the columns by replacing them with their mean.
- Outlier Detection and Handling: Apply the Hampel Filter and median absolute deviation (MAD) to detect outliers in each variable and subsequently bound these outliers with the calculated Lower and Upper bound values for outlier handling.
- Normalization: Normalize the dataset using Min-Max normalization to bring all variables to a common scale.
- Feature Selection using Principal Component Analysis (PCA): Apply Principal Component Analysis (PCA) for feature selection and

understand the importance of each principal component.

- K-Nearest Neighbors (KNN) Classification: Split the dataset into training and testing sets in the ratio 60:40, then train the KNN model on the training set using hyperparameter tuning, considering k values ranging from 1 to 20, and determine the optimal value of k through 10-fold cross-validation.
- Evaluate KNN Model: After training the K-nearest neighbors (KNN) model on the testing set, predict the accuracy and subsequently compute and present the evaluation metrics, including the confusion matrix, precision, recall, and F1-score.

KNN utilizes a distance metric to gauge the resemblance between data points by calculating the distance of each data point in the dataset to the point intended for classification. This study employed the Euclidean distance metric, defined as the square root of the sum of squared differences between corresponding feature values. Euclidean distance is commonly used in KNN due to its simplicity and effectiveness on normalized, continuous data. Alternative metrics like Manhattan or Minkowski distance were considered, but Euclidean showed more stable results across cross-validation folds. The choice of distance metric directly influences the neighborhood formation and thus affects model accuracy. Following this, the algorithm proceeds to find the K Nearest Neighbors, pinpointing the K data points with the smallest distances to the target point, constituting the "nearest neighbors. To accomplish classification tasks, the algorithm tallies the occurrences of each class and subsequently conducts a majority vote among neighboring instances. This process allows for the consideration of weighted voting or tie-breaking methods. The class with the highest count is determined to be the anticipated class at the target point. For making predictions, the procedure associates the anticipated class with the target point by considering the majority class that occurs most often among its K nearest neighbors. To evaluate the effectiveness of the algorithm, the dataset is often divided into two separate sets, including a training set and a testing set. Subsequently, KNN is applied to the testing set, and its accuracy and other pertinent metrics are assessed to determine the algorithm's effectiveness. 'K,' or the optimal parameter, has to be identified for performance optimisation to be effective. This implies that K's value must be adjusted throughout the evaluation process. One hyperparameter that must be set before the algorithm begins is the number of nearest neighbors considered for predictions, represented by the selected value of K. Remarkably, the choice of K has an enormous effect on the performance of the algorithm. A popular approach for determining K is to compute the square root of the total number of observations in the training dataset. This technique yields an accuracy of 85.08% (K=65) and gives an initial estimate; however, the best value for "K" will vary depending on the specific dataset and should be discovered by the method termed as hyperparameter tuning. Hyperparameter tuning involves selecting the set of optimal hyperparameters for a learning algorithm. For KNN, the primary hyperparameter is the number of neighbors (K). The study utilized the Grid Search method, a more systematic approach that defines a grid of hyperparameters and exhaustively tries all combinations. For KNN, this study combined grid search with cross-validation, and the steps are:

- Define Parameter Grid: Create a grid of 'K' values you want to explore.
- Cross-Validation: Use k-fold cross-validation to estimate the effectiveness of each 'K.' This involves splitting your training set into 'k' smaller sets (folds), then training the model 'k' times, each time using a different fold as the validation set and the remaining as the training set. This study used 10-fold cross-validation.
- Search: Apply grid search to systematically work through the grid of 'K' values, training and validating the model for each.
- Best Model: The grid search process keeps track of the performance for each 'K' value and ultimately selects the one with the best cross-validated performance.

6.1 Hyperparameter tuning: K-Selection process

Hyperparameter tuning, especially the selection of the optimal number of neighbors (K), is critical in improving the performance of the K-Nearest Neighbors (KNN) algorithm. In this study, a data-driven approach was implemented to select the most suitable K value. The selection process involved evaluating multiple K values

based on their predictive performance using 10-fold cross-validation. We considered values of K ranging from 1 to 25, to strike a balance between underfitting and overfitting. Each value was assessed by measuring the average accuracy across 10 cross-validation folds on the training data. The model with the highest average cross-validation accuracy was selected as optimal. This method ensures better generalizability and avoids bias due to any single train-test split. Though computationally more intensive than a single evaluation, the use of cross-validation provides a more reliable estimate of model performance. Given the modest size of the Framingham dataset (4,240 instances), the grid search over K values was completed efficiently within seconds using RStudio, making this approach practical for real-world medical datasets. The complete step-by-step procedure for selecting the optimal K using grid search and 10-fold cross-validation is:

- Step 1: Split dataset D into training (60%) and testing (40%) sets.
- Step 2: For each K in range [1 to 25]:
- Initialize accuracy list Acc = []
- Perform 10-fold cross-validation:
- Divide training data into 10 folds.
- For each fold:
    - Train KNN on 9 folds.
    - Validate on the remaining fold.
    - Record the accuracy and append to Acc.
- Compute average accuracy AvgAcc(K) = mean (Acc)
- Step 3: Select K with the highest AvgAcc(K) as the optimal K.
- Step 4: Train the final KNN model using the full training set and optimal K.
- Step 5: Evaluate the model on the test set.

The average cross-validation accuracies for each K value in the range of 1 to 25 are summarized in Table 3, highlighting the performance trend and identifying the optimal K.

Table 3: Hyperparameter tuning results for K

K Value	Average Cross-Validation Accuracy (%)
1	86.02
2	86.9
3	88.12
4	88.95
5	89.91
6	90.18
7	90.45
8	90.89
9	91.05
10	91.12
11	91.26
12	91.4
13	91.3
14	91.72
15	91.85
16	91.93
17	92.03
18	92.15
19	92.36
20	92.46
21	92.4
22	92.32
23	92.1
24	91.87
25	91.65

The implementation of the proposed solution was carried out using R programming language in RStudio. A range of libraries were utilized to perform specific tasks: tidyverse for data manipulation, DataExplorer for exploratory data analysis and visualization of missing values, psych and lattice for descriptive statistics, car for boxplot visualization, caret for model training and evaluation, caTools for dataset splitting, and class for applying the K-Nearest Neighbors (KNN) algorithm. For performance evaluation, metrics such as accuracy, precision, recall, F1-score, and AUC were computed using the pROC and caret packages. Principal Component Analysis (PCA) was conducted using the base prcomp () function. This structured pipeline ensures transparency, reproducibility, and scientific rigor in the analysis.

7 Results and discussion

Heart disease prediction using the K-nearest neighbor (KNN) algorithm has been extensively studied in the literature. Figure 9 is a violin plot that provides a more detailed representation of the distribution of the accuracy of different machine learning algorithms in predicting cardiovascular diseases in the scientific literature and the optimized KNN algorithm.

Similar to a box plot, the violin plot provides a detailed view of the accuracy distribution of several algorithms by including markers for the mean and median. Garg et al., examined the diagnosis of cardiovascular diseases by the application of machine learning (ML) methods, such as K-Nearest Neighbor (KNN) and Random Forest. For cardiovascular disease, the two models' respective prediction accuracy was 86.885% and 81.967% [46]. In a study of supervised machine learning algorithms for predicting and diagnosing heart disease, the random forest technique beats the other four algorithms—decision tree, logistic regression, KNN, and random forest—when applied to a 70,000 sample dataset from Kaggle, with a 92% F1 score and a 95% AUC ROC [47]. In another study, Poojitha et al. examine the K Nearest Neighbor and Novel Random Forest methodologies to see the extent to which data mining algorithms predict heart disease. With a 90.16% success rate for forecasting cardiovascular disease compared to 67.21% for K Nearest Neighbor, it is concluded that the Novel Random Forest approach performs much better in terms of accuracy [48].

The following machine learning methods for classification revealed the following accuracies in an investigation using the Framingham dataset of 4240 observations: Random Forest (RF) leading with an accuracy of 85.05%, K-Nearest Neighbors (KNN) at 83.95%, Support Vector Machine (SVM) at 84.5%, Decision Tree (DT) at 84.82%, and Logistic Regression (LR) at 84.89% [49]. Using a common dataset, Aviral Chanchal et al. investigate the predictive power of many machines learning models for cardiovascular disorders, contrasting the performance of Decision Tree, KNN, Naïve Bayes, SVM, XGBoost, and Random Forest. Despite having lower accuracy percentages, Naïve Bayes, XGBoost, and Random Forest beat the other models in predicting cardiac illnesses. This was discovered by a deeper study utilizing the ROC curve and AUC values, even though KNN, SVM, and RFC exhibited high accuracy scores (85.33%) [50]. Ahmed et al. utilized algorithms such as KNN and SVM, demonstrating that KNN and SVM individually achieved accuracies of approximately 75% and 76%, respectively; a hybrid model integrating both algorithms significantly improved accuracy to 81%. This increase highlights the potential of hybrid machine-learning models in enhancing diagnostic precision in medical applications [51].

Pallathadka et al. emphasize the importance of developing accurate heart disease prediction models using data mining methods like ANN, KNN, and CNN and report that CNNs have shown the most promise in terms of utility and consistency in predicting CHD using the UCI Cleveland database [52]. A study by Gupta et al. explores the application of supervised machine-learning techniques, with Logistic Regression emerging as the superior model in terms of performance metrics, boasting the highest accuracy of 92.30% and lower false negatives compared to other classifiers, demonstrating its potential for prompt disease management. Apart from the higher performance of Logistic Regression, the research also shows that K-Nearest Neighbor (KNN) achieved competitive accuracy rates, with k values of 7 and 14 obtaining around 86.81% and 90.11%, correspondingly [53]. Multi-layer perceptron (MLP) and K-nearest neighbor (K-NN) machine learning techniques were assessed for the prediction of cardiovascular disease (CVD) in research by Pal et al. Both diagnosis rate (86.41%) and accuracy (82.47%) were better with MLP than with K-NN (73.77%) [54]. Bhatt et al. investigated several machine learning techniques and presented a model employing k-modes clustering with Huang initialization using a real-world dataset of 70,000 cases from Kaggle. Combining the multilayer perceptron with cross-validation yielded the most accurate result, outperforming previous approaches with an accuracy of 87.28% [55].

The optimized KNN model works exceptionally well for predicting cardiovascular disease (CVD), having been improved and verified for this purpose. The model achieves 0.9246 and 0.9608 F1-score metrics and accuracy with a strategically selected hyperparameter, k=20. With an overall accuracy of 92.46%, the classification model produced promising outcomes. The outcome has a greater impact on the elements crucial for minimizing the risk of CVD. The positive predictive value (precision) was 92.46%, indicating the proportion of predicted cases that were correctly classified.

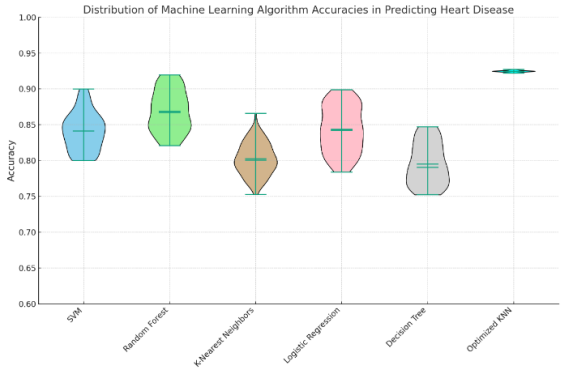

Figure 9: Violin plot showcasing the distribution of accuracies of different machine learning algorithms in predicting heart disease in the scientific literature and optimized KNN.

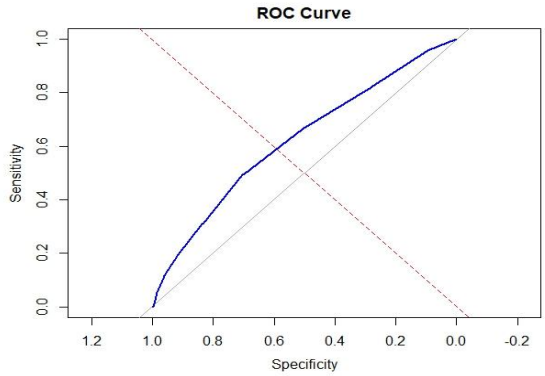

Figure 10: Receiver Operating Characteristic (ROC) Curve

Figure 10 illustrates the ROC curve, a technique employed for evaluating the effectiveness of a binary classification model. The y-axis displays the true positive rate, denoting the percentage of true positives accurately identified by the model. Meanwhile, the x-axis showcases the false positive rate, indicating the percentage of true negatives incorrectly classified as positives. This curve plots the trade-off between sensitivity and specificity (1 - false positive rate) at different thresholds. A model with perfect discrimination (no overlap between the two distributions of the binary classifier) would have a curve that goes straight up the y-axis and straight across at a true positive rate of 1. The Area Under the Curve (AUC=0.6216) condenses the entire ROC curve into a singular metric. A perfect test is denoted by a value of 1, while a worthless test is indicated by 0.5. A higher AUC signifies better discrimination between positive and negative classes. The ROC curve of a random classifier is represented by the diagonal dashed line, corresponding to an AUC of 0.5. The curve above this line indicates that the classifier has a better-than-random ability to discriminate between the two classes. The steepness of the curve at different points can indicate how thresholds can be adjusted to optimize for either sensitivity or specificity. A steep initial rise indicates that a small decrease in specificity will gain a large increase in sensitivity. Based on the ROC curve, the model is evaluated to have a good performance in distinguishing between the positive and negative classes, but there is room for improvement.

In the context of cardiovascular disease (CVD) prediction, the choice of evaluation metrics plays a crucial role in assessing a model's clinical relevance. While our optimized KNN model demonstrated strong performance with an accuracy of 92.46% and F1-scores of 0.9246 and 0.9608, relying solely on accuracy can be misleading due to the class imbalance present in the Framingham dataset. In such datasets, a model may perform well on the majority class (non-CVD) while failing to identify the minority class (CVD), thus inflating the overall accuracy.

To address this, we incorporated additional metrics such as precision, recall, and F1-score, which provide a better understanding of the model's ability to correctly identify positive cases. The F1-score, as the harmonic mean of precision and recall, is especially useful when the cost of false negatives is high—as in medical diagnosis. Our model's high F1-score indicates a good balance between sensitivity and specificity.

However, a discrepancy arises with the AUC-ROC score, which is 0.6216. This value, while better than random guessing (AUC = 0.5), indicates that the model's ability to distinguish between positive and negative classes is moderate. This is likely due to class imbalance and the nature of KNN, which does not output calibrated probability scores. While the model may classify well at a specific threshold, its probability estimates do not align closely with the true likelihood of disease, limiting its usefulness for clinical risk stratification.

To compute these metrics, the model first constructs a confusion matrix from true and predicted labels to determine true positives, false positives, true negatives, and false negatives. From this, accuracy is calculated as the proportion of correct predictions; precision is the ratio of true positives to all predicted positives; recall is the ratio of true positives to all actual positives; F1-score combines both precision and recall; and the AUC-ROC represents the area under the curve plotting the true positive rate against the false positive rate across different thresholds.

To improve AUC and overall discrimination, future enhancements could include probability calibration methods (like Platt scaling), resampling techniques such as SMOTE to handle class imbalance, and the use of precision-recall curves to better evaluate model performance under imbalance.

Figure 11 is a histogram based on the probability of predictions used to evaluate the binary classifier. It

illustrates the data distribution by creating bins across the data range and subsequently using bars to represent the quantity of observations within each bin. The data is categorized into two groups represented by different colors: red for "factor(Actual) 0" and teal for "factor(Actual) 1." The horizontal axis (X-axis) represents predicted probabilities, ranging from 0 to approximately 0.6. The vertical axis (Y-axis) shows the count of occurrences for each probability bin. The red bars show a high frequency of predicted probabilities around 0.1, indicating that for the factor level 0, the model predicted a low probability. The teal bars, which are fewer in number, also show predictions mostly in the lower probability range but are more spread out than the red bars. For the red group (0), the model has high confidence in its predictions as the probabilities are clustered around a peak. The more evenly dispersed probabilities for group 1 suggest a lower or more diverse level of confidence. The dispersion of the teal distribution can signal less confidence in assigning a positive class (1), whereas the concentration of red at lower probabilities shows that the model is confident in giving a negative class (0). A typical problem that may impact the effectiveness of classification algorithms is a class imbalance in the dataset, shown by more red bars than teal (i.e., more factor 0 than factor 1).

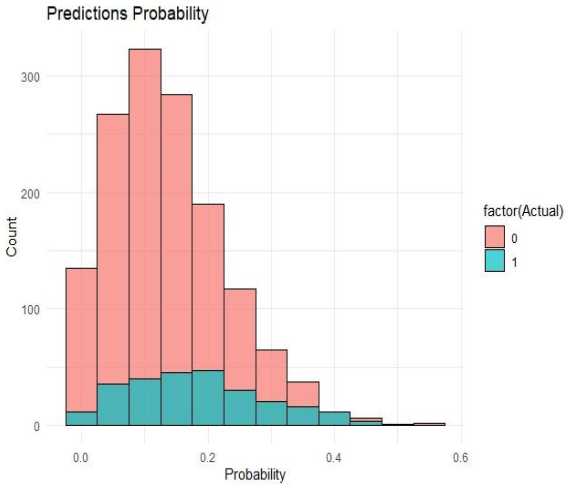

Figure 11: Probability of predictions

Although the overall accuracy of the model is impressive, it has difficulties in correctly identifying instances of the positive class. To increase the model's ability to forecast positive instances, further optimization and maybe even a solution to the class imbalance problem could be needed. The results of the study spotlight the significance of model refinement methods such as feature selection and hyperparameter tuning in improving the classification accuracy. The optimized KNN algorithm analyzes the Framingham dataset of 191 KB in size and 4240 observations in 7.92 seconds. As the size of the dataset increases there is a notable increase in the execution time. But, in medical prognosis precision is more crucial than speed.

8 Conclusion

In the medical industry, early detection and precise diagnosis of cardiovascular disorders are essential since they can greatly enhance patient outcomes. This study presents an optimized approach for K-Nearest Neighbors and shows a significant increase in cardiovascular disease prediction when applied. The meticulous combination of intricate feature selection methods, principal component analysis (PCA) for dimensionality reduction, and hyperparameter tuning yields an exceptionally accurate and efficient model. The optimized KNN model performs better in early CVD detection than typical KNN models, as evidenced by its remarkable metrics and prediction accuracy of 92.46%. The complexity of medical data can be accommodated by customizing machine learning algorithms, as demonstrated by this study. Improving preventive health tactics and possibly saving lives requires the integration of these cutting-edge techniques into clinical procedures. More widespread applications in healthcare are possible as a result of the study's foundational principles and methods, which can be applied to other complex illness projections. Future research might concentrate on correcting the dataset's class imbalance in order to improve the KNN model's capacity to identify instances of the positive class more precisely. Advanced tactics may be used to further enhance the prediction performance of the model and lessen the effects of class imbalance. Typical examples of these strategies include resampling methods and the use of stacked and ensemble approaches. The true innovation of this study lies in presenting a robust and reproducible framework for enhancing KNN-based classification in the context of medical diagnosis. Rather than introducing a novel algorithm, this research demonstrates how existing algorithms can be significantly improved through systematic optimization strategies. The combination of Hampel filtering, PCA-based feature selection, and cross-validated hyperparameter tuning delivers a highly accurate and computationally efficient model. Future studies can further extend this work by integrating ensemble-based or hybrid learning strategies and addressing class imbalance through advanced resampling techniques such as SMOTE.

References

[1] E. Maini, B. Venkateswarlu, B. Maini, and D. Marwaha, "Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India," *Med. J. Armed Forces India*, vol. 77, no. 3, pp. 302–311, Jul. 2021, doi: 10.1016/j.mjafi.2020.10.013.

[2] A. Rajdhan, A. Agarwal, M. Sai, D. Ravi, and D. P. Ghuli, "Heart Disease Prediction using Machine Learning," *Int. J. Eng. Res.*, vol. 9, no. 04. doi: 10.17577/IJERTV9IS040614.

[3] D. A. Anggoro, "Comparison of Accuracy Level of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) Algorithms in Predicting Heart Disease," *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 5, pp. 1689–1694, May 2020, doi: 10.30534/ijeter/2020/32852020.

[4] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, Madurai, India: IEEE, May 2019, pp. 1255–1260. doi: 10.1109/ICCS45141.2019.9065747.

[5] H. Yepdjio and S. Vajda, "Optimization Strategies for the k-Nearest Neighbor Classifier," *SN Comput. Sci.*, vol. 4, Nov. 2022, doi: 10.1007/s42979-022-01469-3.

[6] M. Muzammal, R. Talat, A. H. Sodhro, and S. Pirbhulal, "A multi-sensor data fusion enabled ensemble approach for medical data from body sensor networks," *Inf. Fusion*, vol. 53, pp. 155–164, Jan. 2020, doi: 10.1016/j.inffus.2019.06.021.

[7] H. Yang and J. M. Garibaldi, "A hybrid model for automatic identification of risk factors for heart disease," *Suppl. Proc. 2014 I2b2UTHealth Shar.-Tasks Workshop Chall. Nat. Lang. Process. Clin. Data*, vol. 58, pp. S171–S182, Dec. 2015, doi: 10.1016/j.jbi.2015.09.006.

[8] V. Nagavallika, 'Heart disease prediction using machine learning techniques', Int. J. Sci. Res. (Raipur), vol. 10, no. 11, pp. 630–633, Nov. 2021, doi: 10.21275/SR21918142603.

[9] A. C. Dimopoulos *et al.*, "Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk," *BMC Med. Res. Methodol.*, vol. 18, no. 1, p. 179, Dec. 2018, doi: 10.1186/s12874-018-0644-1.

[10] Prof. Madhavi Tota, Manthan Moon, Pranit Nagrale, Akshay Pandav, and Gunjan Das, "Heart Diseases Prediction System using ML," *Int. J. Adv. Res. Sci. Commun. Technol.*, pp. 337–345, Dec. 2022, doi: 10.48175/IJARSCT-7798.

[11] B. Jin, C. Che, Z. Liu, S. Zhang, X. Yin, and X. Wei, "Predicting the Risk of Heart Failure With EHR Sequential Data Modeling," *IEEE Access*, vol. 6, pp. 9256–9261, 2018, doi: 10.1109/ACCESS.2017.2789324.

[12] A. S. S. Kotia, M. Rastogi, and R. A. Bhongade, "Use of machine learning techniques for effective prediction of heart disease," *CARDIOMETRY*, no. 26, pp. 315–321, Mar. 2023, doi: 10.18137/cardiometry.2023.26.315321.

[13] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN Comput. Sci.*, vol. 1, no. 6, p. 345, Oct. 2020, doi: 10.1007/s42979-020-00365-y.

[14] E. D. Adler *et al.*, "Improving risk prediction in heart failure using machine learning," *Eur. J. Heart Fail.*, vol. 22, no. 1, pp. 139–147, Jan. 2020, doi: 10.1002/ejhf.1628.

[15] I. M. Pires, G. Marques, N. M. Garcia, and V. Ponciano, "Machine learning for the evaluation of the presence of heart disease," *Procedia Comput. Sci.*, vol. 177, pp. 432–437, 2020, doi: 10.1016/j.procs.2020.10.058.

[16] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Comput. Biol. Med.*, vol. 136, p. 104672, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104672.

[17] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," *Expert Syst. Appl.*, vol. 35, no. 1–2, pp. 82–89, Jul. 2008, doi: 10.1016/j.eswa.2007.06.004.

[18] A. Kondababu, V. Siddhartha, BHK. B. Kumar, and B. Penumutchi, "A comparative study on machine learning based heart disease prediction," *Materials Today: Proceedings*, Feb. 2021, doi: 10.1016/j.matpr.2021.01.475.

[19] S. Faiayaz Waris and S. Koteeswaran, "Heart disease early prediction using a novel machine learning method called improved K-means neighbor classifier in python," *Materials Today: Proceedings*, Mar. 2021, doi: 10.1016/j.matpr.2021.01.570.

[20] R. Gopal and V. Ranganathan, "Evaluation of effect of unsupervised dimensionality reduction techniques on automated arrhythmia classification," *Biomed. Signal Process. Control*, vol. 34, pp. 1–8, Apr. 2017, doi: 10.1016/j.bspc.2016.12.017.

[21] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, *Feature extraction. Foundations and applications. Papers from NIPS 2003 workshop on feature extraction, Whistler, BC, Canada, December 11–13, 2003. With CD-ROM*, vol. 207. 2006. doi: 10.1007/978-3-540-35488-8.

[22] N. R. Ratnasari, A. Susanto, I. Soesanti, and Maesadji, "Thoracic X-ray features extraction using thresholding-based ROI template and PCA-based features selection for lung TB classification purposes," in *2013 3rd International Conference on Instrumentation, Communications, Information Technology and Biomedical Engineering (ICICI-BME)*, Nov. 2013, pp. 65–69. doi: 10.1109/ICICI-BME.2013.6698466.

[23] P. Kamencay, R. Hudec, M. Benco, and M. Zachariasova, "Feature extraction for object recognition using PCA-KNN with application to medical image analysis," in *2013 36th International Conference on Telecommunications and Signal Processing (TSP)*, Jul. 2013, pp. 830–834. doi: 10.1109/TSP.2013.6614055.

[24] Yuda Syahidin, Aditya Pratama Ismail, and Fawwaz Nafis Siraj, "Application of Artificial Neural Network Algorithms to Heart Disease Prediction Models with Python Programming," *J. E-Komtek Elektro-Komput.-Tek.*, vol. 6, no. 2, pp. 292–302, Dec. 2022, doi: 10.37339/e-komtek.v6i2.932.

[25] Yichun Wang, "Heart disease prediction with discriminative deep neural network," presented at the Proc.SPIE, May 2023, p. 126401P. doi: 10.1117/12.2673756.

[26] S. .S, S. Lavanya, M. R. Chandhini, R. Bharathi, and K. Madhulekha, "Hybrid Machine Learning Techniques for Heart Disease Prediction," *Int. J. Adv. Eng. Res. Sci.*, vol. 7, pp. 44–48, Jan. 2020, doi: 10.22161/ijaers.73.7.

[27] D. M and R. Abirami, "Heart Disease Prediction System using Ensemble of Machine Learning Algorithms," *Recent Patents on Engineering*, vol. 15, pp. 130–139, Mar. 2021, doi: 10.2174/1872212113666190328220514.

[28] R. R. K. AL-Taie, B. J. Saleh, A. Y. Falih Saedi, and L. A. Salman, "Analysis of WEKA data mining algorithms Bayes net, random forest, MLP and SMO for heart disease prediction system: A case study in Iraq," *IJECE*, vol. 11, no. 6, p. 5229, Dec. 2021, doi: 10.11591/ijece.v11i6.pp5229-5239.

[29] O. Akbilgic *et al.*, "ARTIFICIAL INTELLIGENCE APPLIED TO ECG IMPROVES HEART FAILURE PREDICTION ACCURACY," *ACC.21*, vol. 77, no. 18, Supplement 1, p. 3045, May 2021, doi: 10.1016/S0735-1097(21)04400-4.

[30] O. W. Samuel *et al.*, "A new technique for the prediction of heart failure risk driven by hierarchical neighborhood component-based learning and adaptive multi-layer networks," *Future Gener. Comput. Syst.*, vol. 110, pp. 781–794, Sep. 2020, doi: 10.1016/j.future.2019.10.034.

[31] S. Alagarsamy, K. Kamatchi, K. Selvaraj, A. Subramanian, L. R. Fernando, and R. Kirthikaa, "Identification of Brain Tumor using Deep Learning Neural Networks," in *2019 IEEE International Conference on Clean Energy and Energy Efficient Electronics Circuit for Sustainable Development (INCCES)*, Dec. 2019, pp. 1–5. doi: 10.1109/INCCES47820.2019.9167685.

[32] K. Kartheeban, K. Kalyani, S. K. Bommavaram, D. Rohatgi, M. N. Kathiravan, and S. Saravanan, "Intelligent Deep Residual Network based Brain Tumor Detection and Classification," in *2022 International Conference on Automation, Computing and Renewable Systems (ICACRS)*, Dec. 2022, pp. 785–790. doi: 10.1109/ICACRS55517.2022.10029146.

[33] V. Shankar, V. Kumar, U. Devagade, V. Karanth, and K. Rohitaksha, "Heart Disease Prediction Using CNN Algorithm," *SN Comput. Sci.*, vol. 1, no. 3, p. 170, May 2020, doi: 10.1007/s42979-020-0097-6.

[34] A. Dutta, T. Batabyal, M. Basu, and S. T. Acton, "An efficient convolutional neural network for coronary heart disease prediction," *Expert Syst. Appl.*, vol. 159, p. 113408, Nov. 2020, doi: 10.1016/j.eswa.2020.113408.

[35] S. A. H. Fazlur and S. K. Thillaigovindan, "Integrated Deep Learning Model for Heart Disease Prediction Using Variant Medical Data Sets," *Int. J. Online Biomed. Eng. IJOE*, vol. 18, no. 09, pp. 178–191, Jul. 2022, doi: 10.3991/ijoe.v18i09.30801.

[36] M. Sudipta, E. Abdel-Raheem, and L. Rueda, *Heart Disease Prediction Using Adaptive Infinite Feature Selection and Deep Neural Networks*. 2022, p. 240. doi: 10.1109/ICAIIC54071.2022.9722652.

[37] K. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua, and S. Pranavanand, "Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators," *Appl. Sci.*, vol. 11, no. 18, 2021, doi: 10.3390/app11188352.

[38] S. Ahmed *et al.*, "Prediction of Cardiovascular Disease on Self-Augmented Datasets of Heart Patients Using Multiple Machine Learning Models,"

J. Sens., vol. 2022, p. 3730303, Dec. 2022, doi: 10.1155/2022/3730303.

[39] K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, and H. N. Chua, "Heart Disease Risk Prediction using Machine Learning with Principal Component Analysis," in *2020 8th International Conference on Intelligent and Advanced Systems (ICIAS)*, Jul. 2021, pp. 1–6. doi: 10.1109/ICIAS49414.2021.9642676.

[40] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[41] A. B. Ambrews, E. Gubin Moung, A. Farzamnia, F. Yahya, S. Omatu, and L. Angeline, "Ensemble Based Machine Learning Model for Heart Disease Prediction," in *2022 International Conference on Communications, Information, Electronic and Energy Systems (CIEES)*, Nov. 2022, pp. 1–6. doi: 10.1109/CIEES55704.2022.9990665.

[42] S. P. Patro, N. Padhy, and R. D. Sah, "Classification model for heart disease prediction using correlation and feature selection techniques," in *2022 OITS International Conference on Information Technology (OCIT)*, Dec. 2022, pp. 29–34. doi: 10.1109/OCIT56763.2022.00016.

[43] M. I. Ahmed and F. Shefaq, "A Study on Machine Learning and Supervised and Deep Learning Algorithms to Predict the Risk of Patients: Ten Year Coronary Heart Disease," *Int. J. Pract. Healthc. Innov. Manag. Tech. IJPHIMT*, vol. 9, no. 1, pp. 1–12, 2022, doi: 10.4018/IJPHIMT.305127.

[44] S. Patro and Dr. N. Padhy, "An RHMIoT Framework for Cardiovascular Disease Prediction and Severity Level Using Machine Learning and Deep Learning Algorithms," *Int. J. Ambient Comput. Intell.*, vol. 13, pp. 1–37, Jan. 2022, doi: 10.4018/IJACI.311062.

[45] R. Aggrawal and S. Pal, "Elimination and Backward Selection of Features (P-Value Technique) In Prediction of Heart Disease by Using Machine Learning Algorithms," *Turk. J. Comput. Math. Educ. TURCOMAT*, vol. 12, pp. 2650–2665, Apr. 2021, doi: 10.17762/turcomat.v12i6.5765.

[46] A. Garg, B. Sharma, and R. Khan, "Heart disease prediction using machine learning techniques," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, p. 012046, Jan. 2021, doi: 10.1088/1757-899X/1022/1/012046.

[47] S. Yousefi and M. Poornajaf, "Analysis of Accuracy Metric of Machine Learning Algorithms in Predicting Heart Disease," *Front. Health Inform. Vol 12 2023 Contin. Vol. - 1030699fhiv12i0402*, Apr. 2023, [Online]. doi: 10.30699/fhi.v12i0.402.

[48] T. Poojitha and R. Mahaveerakannan, "Prediction Analysis of Novel Random Forest Algorithm and K Nearest Neighbor Algorithm in Heart Disease Prediction with an Improved Accuracy Rate," *CARDIOMETRY*, no. 25, pp. 1554–1561, Feb. 2023, doi: 10.18137/cardiometry.2022.25.15541561.

[49] W. A. Mahmoud and D. M. Aborizka, "Heart Disease Prediction Using Machine Learning and Data Mining Techniques: Application of Framingham Dataset," 2021. Available at: https://www.idosr.org/wp-content/uploads/2021/11/IDOSR-JCAS-6166-73-2021..pdf.

[50] A. Chanchal, A. S. Singh, and K. Anandhan, "A Modern Comparison of ML Algorithms for Cardiovascular Disease Prediction," in *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Sep. 2021, pp. 1–5. doi: 10.1109/ICRITO51393.2021.9596228.

[51] R. Ahmed, M. Bibi, and S. Syed[3], "Improving Heart Disease Prediction Accuracy Using a Hybrid Machine Learning Approach: A Comparative study of SVM and KNN Algorithms," *Int. J. Comput. Inf. Manuf. IJCIM*, vol. 3, p. 2023, Jun. 2023, doi: 10.54489/ijcim.v3i1.223.

[52] H. Pallathadka, M. Naved, K. Phasinam, and M. M. Arcinas, "A Machine Learning Based Framework for Heart Disease Detection," *ECS Trans.*, vol. 107, no. 1, pp. 8667–8673, Apr. 2022, doi: 10.1149/10701.8667ecst.

[53] C. Gupta, A. Saha, N. V. Subba Reddy, and U. Dinesh Acharya, "Cardiac Disease Prediction using Supervised Machine Learning Techniques.," *J. Phys. Conf. Ser.*, vol. 2161, no. 1, p. 012013, Jan. 2022, doi: 10.1088/1742-6596/2161/1/012013.

[54] M. Pal, S. Parija, G. Panda, K. Dhama, and R. K. Mohapatra, "Risk prediction of cardiovascular disease using machine learning classifiers," *Open Med.*, vol. 17, no. 1, pp. 1100–1113, Jun. 2022, doi: 10.1515/med-2022-0508.

[55] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," *Algorithms*, vol. 16, no. 2, p. 88, Feb. 2023, doi: 10.3390/a16020088.