Harnessing Ensemble in Machine Learning for Accurate Early Prediction and Prevention of Heart Disease

Mohammad Husain[1], Pankaj Kumar[2], Mohammad Nadeem Ahmed[3], Arshad Ali[4] (IEEE, Senior Member), Mohammad Ashiquee Rasool[5], Mohammad Rashid Hussain[6], Muhammad Shahid Dildar[7]
Department of Computer Science, Islamic University of Madinah, Kingdom of Saudi Arabia[1]
Department of Technical Education, Punjab, Sector 36, Chandigarh, India[2]
Department of Computer Science, King Khalid University, Saudi Arabia[3]
Faculty of Computer and Information Systems, Islamic University of Madinah, Saudi Arabia[4]
College of Computer Science, King Khalid University, Abha, Saudi Arabia[5]
Department of Management Information Systems, King Khalid University, Abha, Kingdom of Saudi Arabia[6, 7]

***Abstract*—Cardiovascular diseases (CVDs) remain a significant global health concern, demanding precise and early prediction methods for effective intervention. In this comprehensive study, various machine learning algorithms were rigorously evaluated to identify the most accurate approach for forecasting heart disease. Through meticulous analysis, it was established that precision, recall, and the F1-score are critical metrics, overshadowing the mere accuracy of predictions. Among the classifiers explored, the Decision Tree (DT) and Random Forest (RF) algorithms emerged as the most proficient, boasting remarkable accuracy rates of 96.75%. The DT Classifier exhibited a precision rate of 97.81% and a recall rate of 95.73%, resulting in an exceptional F1-score of 96.76%. Similarly, the RF Classifier achieved an outstanding precision rate of 95.85% and a recall rate of 97.88%, yielding an exemplary F1-score of 96.85%. In stark contrast, other methods, including Logistic Regression, Support Vector Machine, and K-Nearest Neighbor, demonstrated inferior predictive capabilities. This study conclusively establishes the combination of Decision Tree and Random Forest algorithms as the most potent and dependable approach for predicting cardiac illnesses, providing a groundbreaking avenue for early intervention and personalized patient care. These findings signify a significant advancement in the field of predictive healthcare analytics, offering a robust framework for enhancing healthcare strategies related to cardiovascular diseases.**

***Keywords*—Heart disease; machine learning; predictive modeling; cardiovascular disorders; medical diagnosis; feature selection; model evaluation; public health**

I. Introduction

A. Background and Motivation

Modern lifestyles and population change have led to widespread stress, anxiety, and health issues [1]. Sedentary living has increased mortality due to chronic diseases [2-4]. The heart's vital role in transporting nutrients makes its proper function crucial [5-6]. Machine learning extracts valuable insights from large databases. Various techniques, including clustering, association, and classification algorithms, are used to predict heart disease [45]. Cardiovascular diseases cause significant mortality, warranting urgent research [7]. Chronic illnesses, like cancer and diabetes, surpass infectious diseases in causing death and disability. The epidemiologic transition marks this shift. Globally, cardiac diseases account for 17.3 million deaths yearly, projected to rise [8-9]. Machine learning aids early heart disease diagnosis. Defining disease is complex; it generally refers to disrupted bodily functions. Heart disease is universal and results from plaque buildup in coronary arteries. Plaque narrows vessels, causing reduced blood supply, leading to heart attacks or strokes. Symptoms include chest discomfort, pain, and anxiety [46]. Disease causes, recognition, diagnosis, and risk assessment are discussed. This paper delves into heart disease's global burden, machine learning's role, problem statement, research goals, and research article structure.

B. Heart Disease Mortality Rates

Heart disease is a leading cause of death in both developing and developed countries. WHO data reveals significant mortality (region wise distribution of mortality rate is shown in Table I), with 3.8 million deaths in men and 3.4 million deaths in women attributed to heart disease [10]. In the UK, heart disease constitutes 26% of all fatalities [10]. Reports from the Australian Bureau of Statistics (ABS) and the Economic and Social Commission of Asia and the Pacific (ESCAP) indicate mortality rates ranging from 20% to 33% in 2010 [11].

TABLE I. Incidence of Heart-Related Deaths

Sr. No.	Region	Mortality Rate
1	Australia	33.7%
2	East Asia and Pacific Region	35.2%
3	Middle East and North Africa	47%
4	South Asia	10.6%
5	Sub-Sahara in Western Africa	13%
6	Europe	20-26%

Diverse regions exhibit varying heart disease-related mortality rates. Notably, heart disease accounts for 10.6% of reported fatalities [11], with 13% attributed to cardiovascular diseases. Circulatory diseases, predominantly heart diseases, dominate mortality in regions spanning Asia-Pacific, Australasia, Western Europe, and North America. Heart disease emerges as a universal cause of death, regardless of a nation's income level.

C. Global Burden of Heart Disease

Heart disease presents a significant global challenge, impacting individual mortality, family well-being, and economic costs. In the UK, heart disease costs approximately £9 billion yearly, covering premature death and disability expenses [10]. The USA spends around $312.6 billion annually on stroke and heart disease, projected to reach $1.1 trillion by 2035 [9]. China allocates over $40 billion, constituting about 4% of GNI, to heart disease treatment. South Africa's heart disease treatment costs range from 2% to 3% of GNI, a quarter of primary care expenses. Globally, heart disease treatment cost about $370 billion in 2001, accounting for 10% of global healthcare costs [12]. Eastern Europe's high blood pressure expenses reach nearly 25% of healthcare costs. The American Heart Association (AHA) devised a method to forecast medical costs for conditions like high blood pressure, coronary artery disease (CAD), and stroke [13]. By 2030, 40.8% of Americans are predicted to have heart disease. Costs are set to rise from $320 to $818 billion between 2013 and 2030. Early diagnosis is crucial to prevent worsening conditions.

D. Heart Disease Recognition and Diagnosis: Current Scenario

The surge in heart disease incidence stems from preventable factors [1], including unhealthy lifestyles and risk factors like poor diet, obesity, high blood pressure, and elevated triglyceride levels. Warning signs encompass insomnia, abnormal heartbeat, leg swelling, and rapid weight gain [2], often misinterpreted in elderly populations. Growing hospital and research data availability aids precision diagnosis and early detection. AI and ML revolutionize healthcare, enhancing diagnostics, data analysis, and risk prediction. Genetic data analysis benefits from machine learning, expanding medical evaluations and pandemic anticipation. Cardiovascular diseases account for over a third of annual deaths [6], attracting machine learning application in detecting heart disease from medical databases. Diagnostic accuracy, speed, and lifesaving insights improve through these procedures [7].

Dataset diagnosis draws on multiple patient pathology features [46], influenced by varying factors. Critical indicators often determine disease presence. Specialized feature selection enhances predictive accuracy. Addressing class imbalance and dataset rebalancing improves model reliability. Machine learning excels in complex, nonlinear problems, solving classification and prediction tasks like prenatal cardiac defect diagnosis [47] and ECG early warning systems [48]. Ensemble learning's base underlies many techniques, combining classifiers for enhanced performance. Xgboost mitigates overfitting.

Research presents numerous models for cardiac disease classification and prediction. Computerized classifiers assess congestive heart failure risk. Machine learning achieved 93.3% sensitivity and 63.5% specificity [51]. ECG-based deep neural networks improve performance [52]. Clinical decision support systems aid early heart failure detection [53]. SVM identifies diabetes and predicts heart disease with 94.60% accuracy [55]. In "curse of dimensionality," massive data's exponential growth hampers analysis, causing overfitting. Weighting characteristics reduce dataset duplication, easing processing [57, 58, 59]. Feature engineering and selection methods decrease dataset dimensionality [50].

Despite preventability efforts, heart disease persists globally. Pharmacies and health maintenance tests are crucial for rising heart disease rates. Expensive screening tests are used initially, prompting the need for cost-effective community-level alternatives. Identifying risk factors like age, alcohol, diet, smoking, and inactivity is vital to combat heart disease. Exposure to these factors increases hypertension, diabetes, dyslipidemia, obesity, and stroke risks [16].

Heart disease's high mortality demands accurate diagnosis tools. A systematic, accurate diagnostic tool based on death rates, disability rates, and costs is needed. Screening tools for cost-effective early diagnosis exist but require invasive blood sampling [16]. Main objective of this work is to study ML Algorithms (LR, KNN, SVM, RF, and Decision Tree), optimize algorithms to combat overfitting, apply ML for Classification, evaluate, and compare performance metrics.

The study evaluates and compares classifiers such as decision trees, Naive Bayes, logistic regression, SVM, and random forests. It suggests an innovative ensemble classifier strategy, which combines both strong and weak classifiers. This approach is designed to accommodate diverse sample requirements for training and validation, ensuring a robust and reliable predictive model for heart problems. By harnessing the synergistic strengths of multiple classifiers, this research aims to provide a comprehensive and accurate prediction framework for cardiovascular diseases, thereby contributing significantly to the advancement of predictive healthcare analytics.

E. Research Paper Outline

The Research paper is organized into five sections including conclusion and discussion. Section I introduce the spread of heart disease, its prevalence, diagnosis, and economics of early cure including research goals. Section II provides an in-depth review of heart disease prediction using machine learning techniques. The Section III is used to describe machine learning techniques and their applications. Performance evaluation via various techniques is highlighted in Section IV. Finally, the paper is concluded in Section V with limitations of proposed technique and future research directions.

II. Literature Review

This section provides an in-depth exploration of the significant contributions made by researchers in the realm of heart disease assessment using various machine learning techniques. The focal point is on recognizing the importance of

early diagnosis and prognosis, while concurrently highlighting the gaps and limitations present within the existing literature.

Ignoring heart issues can be detrimental, with men at higher risk [10]. A pivotal dataset from 1988 combines Cleveland, Hungary, Switzerland, and Long Beach V data. 80% of heart disease can be averted through healthy living [14]. Primary, secondary, and tertiary preventions obstruct disease progression [15]. Early diagnosis reduces serious illness and cost. A reliable tool for high-risk classification is crucial. AHA's goals could prevent millions of heart disease deaths [14]. Early detection prevents severe conditions [15]. Resource constraints in LMICs require cost-effective, community-level screening for higher-risk individuals. Early prediction and cost-effective prevention strategies are essential [15].

Researchers have harnessed supervised machine learning techniques to predict heart disease. Nguyen and Davis [23] introduced the KMIX algorithm for heart ailment prognosis. Shouman, Turner, and Stocker [24] advanced Naive Bayes through K-Means clustering. Tsipouras et al. [25] innovated a fuzzy rule-based model. Aqueel and Hannan in [26] integrated SVM, genetic algorithms, rough set theory, association rules, and neural networks. Amin, Agarwal, and Beg [27] crafted a hybrid model integrating neural networks and genetic algorithms. Chaurasia and Pal in [28] envisioned heart disease forecasts by deploying Naïve Bayes, decision trees, and bagging. Bialy et al. [29] forged a hybrid model that amalgamated Bay's theorem and Perceptron. In Table II some of the recent work is listed with research methodology, limitations, and contribution. Modepalli et al. [52] embraced a hybrid approach of DT and RF. L. Sathish Kumar and A. Padmapriya [58] employed the ID3 algorithm to anticipate diseases.

Some latest results show that in one study, logistic regression exhibited notable accuracy, achieving 90.16% on the Cleveland dataset, while AdaBoost outperformed with 90% accuracy on the IEEE Dataport dataset [60]. Another comparative analysis scrutinized traditional machine learning methods against deep learning algorithms, highlighting the superiority of artificial neural networks (ANN). The ANN model demonstrated a remarkable accuracy of 93.44%, surpassing the support vector machine (SVM) model by 7.5% [61].

TABLE II. SUMMARY OF LITERATURE REVIEW

Reference No.	Methodology	Outcome	Advantages	Limitations
[17]	Machine Learning Models for CHD	Risk estimation over short and long term	Improved risk assessment	Focus on short-term forecasting
[18]	Cross-Validation and Multi-class Classification	Robust prediction model	Effective evaluation on multiple classes	Focus on cross-validation
[19]	Heart Rate Variability Analysis	CAD diagnosis using HRV	Utilization of medical domain knowledge	Focus on HRV analysis
[20]	Neural Network Model	Risk assessment using neural networks	Utilizing AI for risk assessment	Specific to neural network model
[21]	Various Machine Learning Algorithms	Heart disease prediction using ML	Comparative evaluation of algorithms	Focus on multiple machine learning models
[22]	Decision Trees and Risk Model	Risk assessment model for CHD	Effective use of decision trees	Focus on specific algorithms
[23]	KMIX Algorithm for Clustering	Improved clustering for disease prediction	Enhanced performance with KMIX algorithm	Specific to KMIX clustering method
[24]	K-Means with Naive Bayes	Enhanced prediction using K-Means and NB	Improved handling of continuous attributes	Specific to K-Means and NB
[25]	Fuzzy Rule-Based Model	CAD prediction using fuzzy rules	Improved classification with fuzzy rules	Utilization of fuzzy rules
[26]	Ensemble Techniques with Genetic Algorithms	CHD diagnosis using various algorithms	Enhanced predictive capability	Focus on ensemble and genetic algorithms
[27]	Hybrid Model (Neural Network and Genetic Algorithm)	Initial risk assessment model	Utilizing hybrid model for prediction	Specific to neural network and GA
[28]	Naïve Bayes, Decision Trees, Bagging	Accurate heart disease prediction	Effective use of multiple algorithms	Specific to certain algorithms
[29]	Ensemble Techniques with Weighted Average	CAD assessment using ensemble methods	Improved accuracy with ensemble approach	Focus on ensemble techniques
[52]	Artificial Intelligence for CHD	Severity prediction using K-Star algorithm	Utilizing AI for cardiac diagnosis	Focus on severity prediction
[58]	ID3 Algorithm in TV and Mobile Phones	Disease prediction and prevention	Effective education and prevention	Specific to certain algorithms

Furthermore, a distinct research endeavor proposed an innovative heart disease prediction model. This model incorporated embedded feature selection techniques and deep neural networks, resulting in an impressive accuracy of 98.56% on the Kaggle dataset [62]. Additionally, a neural networks model utilizing a Multilayer Perceptron (MLP) achieved commendable accuracies, recording 85.71% on the UCI Heart Disease dataset and 87.30% on the cardiovascular disease dataset [63].

A. Research Gaps

Despite the strides in heart disease prediction, the extant literature grapples with several limitations:

- Crafted models struggle with generalizability and potential sluggishness due to intricate risk rules.
- Experimentation tools entail complications and inherent limitations.
- Majority of models are circumscribed to clinical attributes, neglecting non-invasive risk elements.
- Scarce research leverages multiple feature selection techniques alongside their mean values.

A pressing need exists for further exploration into novel heart disease revelations and their effective integration into machine learning techniques. Continued research is essential to heighten diagnostic precision through machine learning methods and surmount the prevailing gaps in heart disease anticipation and detection.

III. Machine Learning Methods for Heart Disease Prediction

Machine learning techniques are used to extract hidden information in an explicit structure from these large datasets because the medical industries are overrun with noisy and incomplete data. Machine learning techniques should be used in the healthcare industry to support specialists rather than replace them [54]. The feature selection methods used to identify the significant non-invasive subset of risk attributes for the early diagnosis of heart disease are described in this Section. The machine learning methods used to create a risk evaluation model are covered in this section. Various performance measures are used in this section to assess the risk models' performance. The importance of non-invasive risk factors for the initial diagnosis and care of cardiac patients is also discussed in this section.

Exploring the heart disease dataset provides valuable insights that can significantly aid in early detection and prediction of cardiovascular conditions. In this Section, Davis' machine learning methodology was employed in the study to construct a comprehensive cardiovascular disease model. The focal point of this Section lies in outlining the research procedures, designing the study's framework, and expanding its applicability through well-defined objectives. By harnessing the power of machine learning techniques, this research effectively identifies a substantial subset of risk factors crucial for the initial prognosis of individuals with heart disorders.

A. The Methodology of Prediction

The process of transforming raw data into a dataset that can be used to produce knowledge for users is referred to as "machine learning" and a machine learning methodology is a method that uses alternative techniques to accomplish this transformation. The utilization of this methodology in particular is warranted due to the fact that it exemplifies the objectives of our research. The following is an outline in Fig. 1. The first step of the process is called data selection, and it entails selecting the pertinent information about heart disease from a variety of different sources so that it can later be entered into the standard database.

1) Data Preparation: In the first step, known as "data preparation," the dataset containing information about heart disease is analyzed and prepared so that the machine learning algorithms can derive useful insights from it and achieve the best possible results.

2) Data Task Filter: In this step, the heuristic decision rules are used to establish expected outcomes for the prognosis of heart disease in subsequent steps. The dataset that was selected is then stored in what is called the "Machine learning Task Warehouse."

3) Selecting Appropriate Algorithms and Datasets: It is for the Task Specified in Step 3. This step involves selecting an appropriate algorithm and dataset for the task that was specified in Step 3.

4) Comparison and Evaluation: The results of the classification are compared to one another and estimated using a variety of different machine learning evaluation metrics during this phase.

In the process of developing new models, the recently finished supervised classification models have been filed away in the data warehouse in order to be ready for any upcoming issues with prediction. For each new prediction task, the procedure starts over at step three and continues through step five.

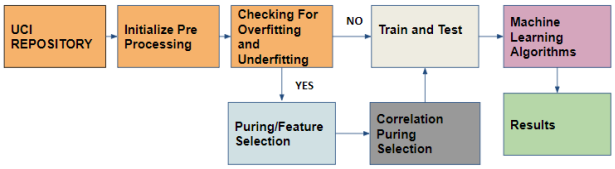

Fig. 1. Detail steps of research design.

B. Exploratory Data Analysis (EDA) Process

Fundamental statistical descriptions are conducted to enhance understanding of the myriad attribute values within the Kashmir heart disease dataset. Knowledge of these basic statistics facilitates addressing noisy values, detecting outliers, and handling missing values. The dataset contains both nominal and numerical values, all serving as risk factors for coronary heart disease. Simple mean imputation is applied to address missing numerical values, while mode imputation is used for missing values in categorical data.

C. Examination of Class Imbalances and Distributional Issues

Before engaging in any operations related to heart disease dataset, assessing class balance is crucial. Highly imbalanced data can lead to biased machine learning algorithms. Statistical analysis is applied to the data to evaluate its kurtosis, skewness, and class balance. Skewness assesses symmetry to determine if data distribution is equal on both sides of the center point. Kurtosis identifies whether data tails are light or heavy compared to a normal distribution. Skewness and Kurtosis tests reveal that the Kashmir heart disease dataset follows a normal distribution.

D. Establishing Feature Correlations

Since datasets can contain intricate interconnections between variables, quantifying the degree of attribute relationships is vital. The correlation process involves assessing the level of connection between dataset attributes. Understanding these connections is essential for data preparation before applying machine learning algorithms. Pearson's correlation method is used to explore the relationship among heart disease attributes. A heatmap depicts Pearson's correlation coefficients applied to heart disease variables (see Fig. 2).

The heatmap grid showcases associations between cardiovascular disease-related factors and associated coefficients. The matrix presents all attributes horizontally across the top and vertically down the side, offering correlations among feature combinations. The diagonal line's connection from bottom right to top left indicates perfect correlation between attributes and themselves. Correlation coefficients near zero suggest weak relationships between heart disease attributes, while values of 1 and -1 signify ideal positive and negative correlations, respectively.

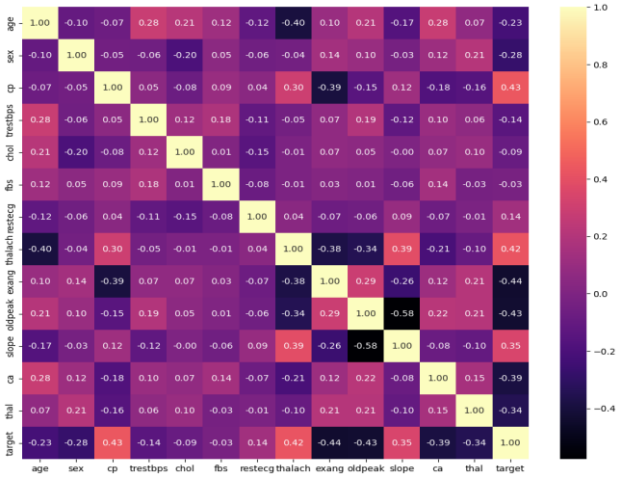

Fig. 2. Correlation in risk attributes through heat map representation.

E. Feature Selection

Feature selection is crucial as irrelevant or redundant attributes can impede classifier performance. To attain a non-invasive subset of risk attributes for precise heart disease prediction, the heart disease dataset (as listed in Table IV) undergoes five distinct Feature Elimination techniques. These techniques assign values to potential risk factors based on their disease prediction accuracy, assigning weights from 0 to 1 to each attribute associated with coronary heart disease. The final weights are determined by individual feature selection techniques, where attributes with mean values close to 1 are considered significant, while those near 0 are less significant.

These heart disease-linked characteristics are presented in descending order of mean values derived from five distinct feature selection strategies in Table V. Attributes with higher weights are more important for predicting early heart disease, while those with lower values are less significant. The model predicting the risk of heart disease development is constructed using the highly weighted significant subset of risk factors.

1) Feature selection techniques: Precise and concise prediction model subsets are identified using feature selection techniques [30]. To obtain the best non-invasive subset of risk factors for heart disease prediction, this research investigates a combination of filter, wrapper, and embedded feature selection methods.

- Extra Tree Classifier: The extra tree classifier, also known as extremely randomized trees, is an ensemble learning technique creating multiple trees without eliminating any existing ones. Decision tree nodes are divided through random splits, enhancing accuracy while significantly reducing the computational load associated with determining optimal cut-points in random forests and standard trees [31].
- Gradient Boosting Classifier: Gradient boosting is employed to address classification and regression challenges. It entails constructing decision trees in a greedy manner to optimize a loss function, adding these trees one at a time to minimize the loss function [32].
- Random Forests: Random forests involve decision tree predictors for regression and classification tasks, using multiple decision trees in a randomly selected training set to counter individual decision tree overfitting [33]. Further explanation of the random forest classifier can be found in the machine learning techniques section.
- Recursive Feature Elimination: Recursive feature elimination (RFE), a greedy optimization technique, builds the feature model until all features are used. Features are then ranked based on their elimination order [34].
- XG Boost Classifier: The XG Boost classifier employs a gradient boosting algorithm with optimized regularization to counter bias and overfitting. Its scalability enables swift learning and efficient memory usage [35].

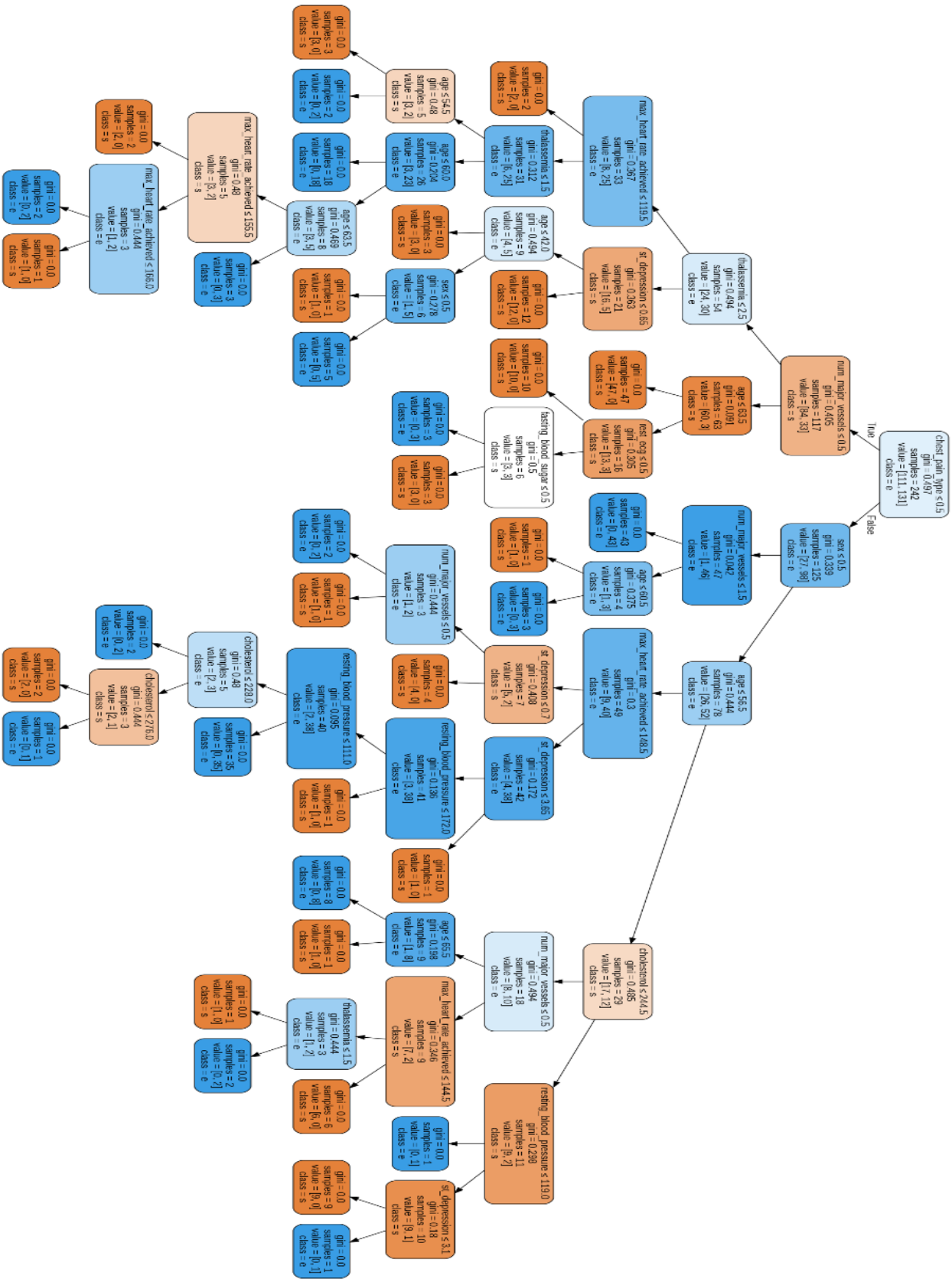

Fig. 3. Decision tree model working for heart disease prediction.

F. Predictive Analysis

Machine learning tasks capitalize on discovered patterns to learn from the machine learning process. These tasks are typically categorized into predictive and descriptive categories [36]. Predictive tasks focus on predicting the value of a dependent (target) attribute based on independent (exploratory) attributes. Descriptive tasks aim to extract patterns describing underlying relationships within data, often requiring post-processing techniques for validation and explanation due to their exploratory nature [37].

1) Machine learning techniques: Predicting heart condition from different symptoms is a stratified problem that is bound to erroneous assumptions and has impulsive effects. We use various machine learning methods to extract knowledge from the heart disease dataset. The purpose of blending machine learning methods in health care is not to take over specialists or assistants, but to give support to where they struggle [38]. Some of the popular Machine Learning algorithms are shown in Fig. 4 and described below:

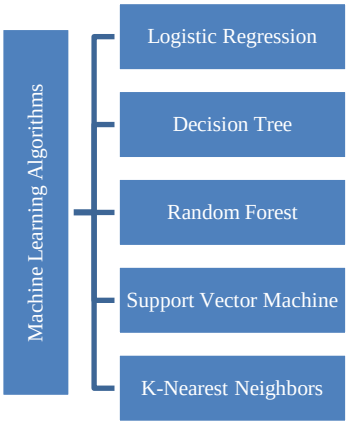

Fig. 4. Machine learning algorithms.

- Decision Tree: The decision tree is a widely used tool, especially for classification tasks [39]. It is constructed using a top-down, recursive divide-and-conquer approach, following a greedy (nonbacktracking) strategy. This is depicted in Fig. 4. There are various types of decision trees, distinguished by the mathematical model they employ to select the attribute for splitting, thereby forming decision tree rules. The attribute that effectively divides the tuples into distinct classes is chosen based on the Information Gain attribute selection measure. The Information Gain approach aims to maximize the reduction in uncertainty by selecting the splitting attribute with the lowest entropy value. The Information Gain for each attribute is determined using Eq. (1):

$$Gain\ (A)\ =\ Info\ (D)\ -\ InfoA(D) \quad (1)$$

Where:

- Info (D) represents the entropy of the entire dataset.

- InfoA (D) represents the weighted average of the entropies of subsets obtained by splitting based on attribute A.

The entropy of a set is calculated using Eq. (2):

$$Info\ (D)\ =\ \Sigma\ -p\ *\ log2(p) \quad (2)$$

where, p is the proportion of instances belonging to a specific class.

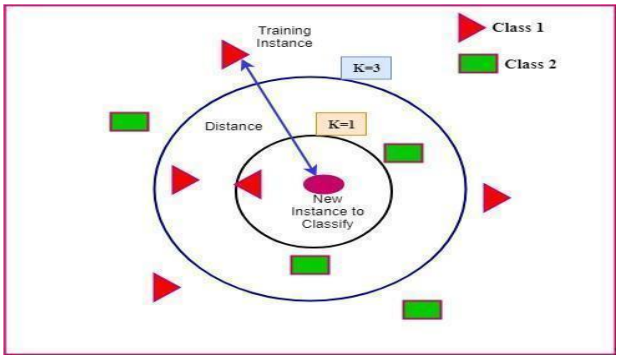

Fig. 5. K nearest neighbour classification.

- K-Nearest Neighbor (KNN): K-Nearest Neighbor (KNN) is a fundamental instance-based machine learning technique that operates in a non-parametric manner [40, 49]. It relies on learning by analogy, where a new unclassified record is compared to existing records using a distance metric. The class of the closest existing record is then assigned to the new unclassified record. Fig. 5 provides an example of KNN classification. The optimal value of k (the number of neighbors) is typically determined experimentally. This involves starting with k = 1 and gradually increasing k to account for more neighbors. The error rate of the classifier is calculated using a test set. In the KNN algorithm, a new instance is classified based on its proximity to its neighbors, determined by a distance function. Various distance measures such as Euclidean, Manhattan, and Minkowski can be utilized. In this study, due to the nature of the heart disease data, the Euclidean distance measure is used. To prevent attributes with higher values from dominating those with lower values, attribute values are normalized before applying the Euclidean distance measure. The Min-Max normalization technique is employed, which transforms a numerical attribute's value P to a value P| in the range [0, 1]. The KNN technique is used in this study for predicting heart disease.

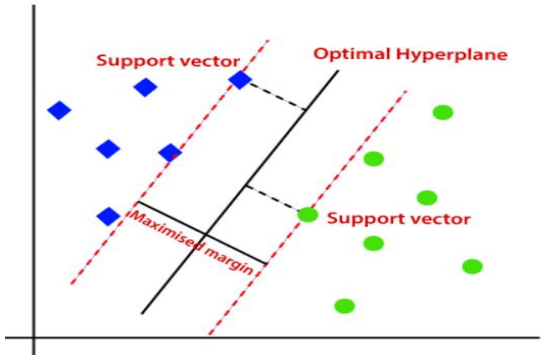

Fig. 6. Linear SVM classifier for two-class representation.

- Support Vector Machine (SVM): Support Vector Machine (SVM) is a supervised machine learning technique used for both classification and regression tasks. SVM works by translating the original training data into a higher-dimensional space through a nonlinear mapping. It seeks to find the best separating hyperplane in this new dimension. Support vectors and margins are utilized by SVM to determine this hyperplane [41]. The linear support vector machine is illustrated in Fig. 6, where red circles denote data points of class x2, and light green circles represent data points of class x1. However, if there is no obvious hyperplane in the original feature space, SVM requires moving to a higher-dimensional view known as kernelizing. The principle behind kernelizing is that the data will be mapped into higher dimensions until a hyperplane can be established to separate it. The choice of the SVM's kernel function, such as polynomial, radial basis, and Gaussian kernel functions, plays a critical role. There are other kernel functions available as well, in addition to the ones mentioned.
- Random Forests: Random Forests are an ensemble learning technique that utilizes a collection of individual decision trees for both classification and regression tasks. They are designed to address the issue of overfitting that can occur with individual decision trees. In random forest classification, the final class of a test object is determined by the majority votes from each decision tree in the forest [42]. Random Forests have significantly extended bagging, a technique that aggregates a large set of decorrelated trees. The process of the random forest algorithm is depicted in Fig. 7, where each tree is grown using a different subset of the original data. In each of the k iterations, approximately one-third of the samples are left out from the new bootstrap training set and are not used in constructing the tree. The class with the highest number of votes from the trees in the forest becomes the final classification for a given sample. The random forest algorithm is applied in this study for diagnosing and predicting heart disease, and Section IV provides further details on the outcomes.
- Naive Bayes: Naive Bayes is a classification algorithm that operates based on statistical probabilities and follows the principles of the Bayesian theorem. It is particularly effective when dealing with high-dimensional inputs. The algorithm works under the assumption of "class conditional independence," which means that the attribute values' impact on a specific class is considered unrelated to the outcomes of other attributes. This assumption is referred to as "naive" because it simplifies calculations [43]. The Naive Bayes classifier can handle both continuous and categorical variables, and it can accommodate any number of independent variables. By assuming that the probabilities are independent of each other, Naive Bayes simplifies probability calculations, leading to a fast and efficient method. In this study, the Naive Bayes algorithm is employed using the non-invasive risk attributes to predict and diagnose heart disease at its early stages. Section IV provides a detailed discussion of the Naive Bayes model's predictions for heart disease.

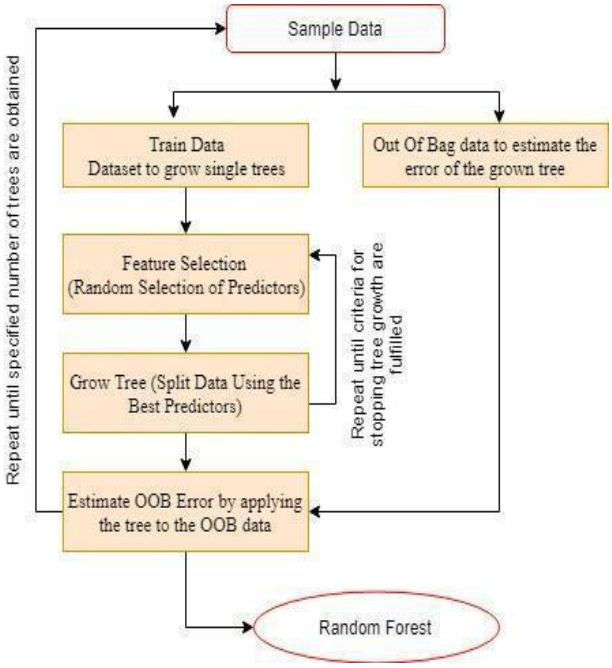

Fig. 7. Random forest algorithm working.

2) Model evaluation techniques: Model evaluation is a critical aspect of practical machine learning development. In order to interpret patterns from the provided dataset, systematic methods are required to assess the effectiveness of machine learning techniques and to compare them, helping to decide which method to use for a given problem. The performance of algorithms in classification problems can be evaluated using various metrics, including the confusion matrix, cross-validation, error rate, sensitivity, specificity, accuracy, and precision. These evaluation metrics are discussed below [44]:

a) Confusion Matrix: The confusion matrix is a fundamental tool for assessing performance in classification problems. It is particularly useful for understanding the types of classification errors that can occur in two-class classification scenarios. The confusion matrix provides insight into how well the model's predictions align with the actual outcomes. In a two-class confusion matrix, as shown in the Table III below, various classifications are categorized based on their correctness or incorrectness:

- True Positives (TP): Instances that are correctly classified as positive.

- False Negatives (FN): Instances that are actually positive but are incorrectly classified as negative.

- False Positives (FP): Instances that are actually negative but are incorrectly classified as positive.

- True Negatives (TN): Instances that are correctly classified as negative.

The confusion matrix allows for a deeper understanding of the model's performance and the types of errors it makes, such as Type We and Type II errors.

TABLE III. CONFUSION MATRIX FOR BINARY CLASSIFICATION

		Predicted Values	
		Positive	**Negative**
Actual Values	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

- Error Rate (Misclassification Rate): The error rate, also known as the misclassification rate, is a measure that quantifies the proportion of misclassified instances in a classification model. It's a combination of both training errors and generalization errors.

- Training Errors: These are the mistakes made by the model when classifying instances within the training dataset.

- Generalization Errors: These are the expected mistakes that the model will make when classifying instances that it hasn't seen before, i.e., on unseen data.

The goal of a good classification model is to have both low training and generalization errors. This indicates that the model has learned the underlying patterns in the data without overfitting to the training data.

The error rate can be calculated using the formula in Eq. (3):

$$Error\ Rate = \frac{(False\ Positives + False\ Negatives)}{(Total\ Positive + Total\ Negative)} \quad (3)$$

Where False Positives are instances that are wrongly classified as positive, False Negatives are instances that are wrongly classified as negative, and Total Positive and Total Negative are the total number of positive and negative instances, respectively.

- Cross-Validation: Cross-validation is a technique used to estimate the performance of a machine learning model on unseen data. It involves dividing the dataset into multiple subsets (folds), using some folds for training and others for testing. This process is repeated multiple times with different combinations of training and testing sets. By evaluating the model's performance across different subsets of data, cross-validation provides a more robust estimate of its generalization ability. In a common method called k-fold cross-validation, the dataset is divided into k subsets of approximately equal size. The model is trained on k-1 folds and tested on the remaining fold in each iteration. The results from all iterations are then averaged to provide an overall assessment of the model's performance. Cross-validation helps to mitigate the risk of overfitting and provides a more accurate estimation of how well the model will perform on new, unseen data.

IV. RESULTS AND DISCUSSION

In conclusion, the development of a robust risk evaluation model for cardiac disorders involves careful selection and preprocessing of data, integration of diverse information sources, and the utilization of appropriate algorithms. By prioritizing non-invasive risk factors and optimizing data quality, accurate and reliable risk assessment strategies can be implemented.

A. Dataset Selection

For the development of the risk evaluation model, we sourced a dataset from the Kaggle Machine Learning library. This dataset comprises 1025 data points, each characterized by 14 distinct attributes, encompassing 13 predictive features and 1 target class. These attributes encompass various factors such as age, sex, chest pain, high blood pressure, cholesterol levels, fasting heart rate, ECG readings, and more [5]. In order to comprehensively analyze the risk factors associated with heart disease and to construct a highly accurate model, five different algorithms are employed. The field of cardiac disorder detection encompasses various tests, some of which require invasive procedures and multiple blood tests. To implement more practical risk recognition strategies, a focus on non-invasive risk factors is essential. These factors, such as age, height, weight, and smoking habits, can be easily obtained without the need for complex equipment. While measurements like body weight and blood pressure do require devices, these tools are readily available at home or local pharmacies, eliminating the necessity for hospital-based procedures for data acquisition. Data fields are shown in Table III.

B. Data Balancing

Notably, many medical databases exhibit an imbalanced distribution of positive and negative samples. To enhance the model's reliability, it may be necessary to apply specific data processing techniques to rectify this imbalance [56]. Moreover, real-world data often contains duplicates and missing values, which can distort the analysis. Through careful data preprocessing, including techniques like smoothing, normalization, and grouping, we ensured that the input data was accurate, devoid of noise, and effective for analysis [6].

C. Data transformation

The process of transforming raw data into a more understandable format involves translation. This translation process is supplemented by steps such as smoothing, normalization, and grouping to ensure that the data is prepared optimally for analysis. Moreover, integration of information from various sources is often required to produce refined and comprehensive datasets.

D. Data Preprocessing

Within the dataset, 526 instances represent individuals with cardiac disease, while 499 instances pertain to individuals without the condition. While it can be challenging to limit the amount of data collected, it's crucial to present the data effectively to derive meaningful insights. In certain cases, specific attributes may hold a high correlation with the target variable. For instance, in analysis, the fasting blood sugar attribute displayed significant correlation, leading to eliminating the corresponding column to enhance the model's

accuracy. Table V lists out the ranking of different attributes as per the importance of attributes.

1) Splitting data into test train set: Following data preprocessing, the dataset is organized into training and validation subsets into 80:20 ratio. The performance of different algorithms is then assessed to ascertain their predictive capabilities [7]. The process of data preparation, including feature selection and data uniformity, can significantly enhance the dataset's utility and subsequently improve the accuracy of the model.

TABLE IV. THE HEART DISEASE DATASET ATTRIBUTES

Variable Name	Role	Type	Units	Missing Values
Systolic BP	Feature	Integer	mm Hg	no
Diastolic BP	Feature	Integer	mm Hg	no
BMI	Feature	Integer	Number	no
Age	Feature	Integer	Years	no
Healthy Diet	Feature	Categorical	No Unit	no
Hereditary	Feature	Categorical	No Unit	no
Smoking	Feature	Categorical	Binary	no
Physical Activity	Feature	Categorical	Binary	no
Socio-Economic Level	Feature	Categorical	No unit	no
Sex	Feature	Binary	No Unit	no
Alcohol Consumption	Feature	Categorical	Number	no
CHD	Target	Integer	No unit	no

TABLE V. MEAN RANKING OF WEIGHTAGE OF ATTRIBUTES

Sr. No.	Attributes	Mean ranking of attributes
1	Systolic BP	0.82
2	Diastolic BP	0.80
3	BMI	0.78
4	Age	0.76
5	Healthy Diet	0.54
6	Hereditary	0.42
7	Smoking	0.28
8	Physical Activity	0.24
9	Socio-Economic Level	0.16
10	Sex	0.14
11	Alcohol Consumption	0.12

E. Experimental Results of the Proposed Machine Learning Techniques

The existing models employed for assessing the risk of heart disease have demonstrated inherent flaws that undermine their effectiveness. These models often yield inconsistent results when applied to diverse datasets, thereby compromising their reliability. In this study, the focus is on leveraging machine learning techniques, specifically Decision Tree (DTC), K-Nearest Neighbor (KNN), Random Forest (RFC), Support Vector Machine (SVM), and Naive Bayes (NBC), to extract objective and dependable outcomes from the cardiovascular disease dataset. To achieve this, a range of performance metrics relevant to the medical domain, including sensitivity, specificity, accuracy, and precision, are employed to ensure the generation of accurate and reliable results. The following subsections elucidate the experimental findings yielded by various models in the context of disease assessment.

The central aim of this study revolves around predicting the likelihood of an individual developing heart disease. To fulfill this objective, a variety of supervised classification approaches, including Support Vector Machine, Random Forest, K-Nearest Neighbor, and Logistic Regression, are explored. The experimentation encompasses the utilization of diverse computational models, particularly Decision Trees, facilitated by the SkLearn package. The experimental setup utilized a 6th generation Intel Core i3 processor with a 3300H CPU, operating at up to 2.1 GHz, and 4 gigabytes of RAM. A prompt data analysis procedure was employed to swiftly provide a comprehensive accuracy assessment for the adopted methods. The dataset partitioning involved allocating 55% (563 instances) of the data for training purposes and 45% (462 instances) for testing purposes. The subsequent graph depicts the distribution of training and testing activities undertaken during the study:

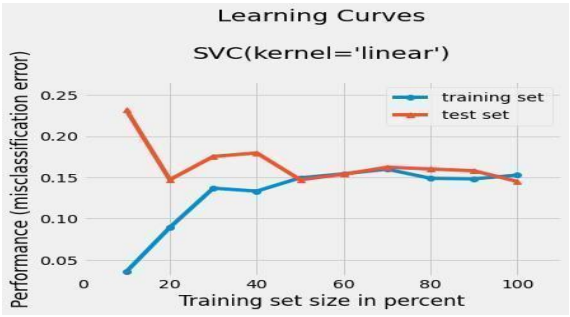

Fig. 8. SVM train-test split.

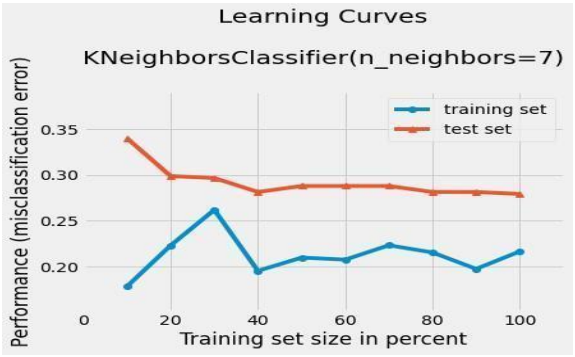

Fig. 9. LR train-test split.

Fig. 10. KNN train-test split.

TABLE VI. CONFUSION-MATRIX DT

	Predicted(N)	Predicted(P)
Actual(N)	223 (TN)	5 (FP)
Actual(P)	10 (FN)	224 (TP)

TABLE VII. CONFUSION-MATRIX SVM

	Predicted(N)	Predicted(P)
Actual(N)	179	49
Actual(P)	19	215

TABLE VIII. CONFUSION-MATRIX RF

	Predicted(N)	Predicted(P)
Actual(N)	216	10
Actual(P)	5	231

TABLE IX. CONFUSION-MATRIX LR

	Predicted(N)	Predicted(P)
Actual(N)	183	48
Actual(P)	19	214

TABLE X. CONFUSION-MATRIX KNN

	Predicted(N)	Predicted(P)
Actual(N)	165	58
Actual(P)	35	204

TABLE XI. MODEL EVALUATION RESULTS IN %

Algorithm Used	Accu-racy	Precision	Recall	F1-score
LR Classifier	85.93	81.68	91.85	86.47
SVM Classifier	85.28	81.44	91.88	86.35
KNN Classifier	79.87	77.86	85.36	81.44
DT Classifier	96.75	97.81	95.73	96.76
RF Classifier	96.75	95.85	97.88	96.85

Fig. 8 to 10 showcase the division of the test dataset and the train dataset's performance, indicating that optimal performance was achieved within the 60%–80% split range. Several options were experimented to find out the best test train split. The binary label's confusion matrix for each tested method is depicted in Tables VI to X. While accuracy is a valuable metric, we place greater significance on precision, recall, and the F-1 score, all of which can be found in Table XI. Among the methods tested, K-Nearest Neighbor (KNN) yields the least favorable results, while regression and Support Vector Machine (SVM) methods perform moderately. Notably, Decision Tree and Random Forest methods exhibit the highest accuracy and F-1 score, as evidenced by this dataset. Thus, it can be inferred that Random Forest is a versatile method capable of achieving substantial accuracy with ease.

The table presents the results of different classification algorithms employed to predict heart disease, showcasing their performance metrics. These metrics are vital in assessing the accuracy and effectiveness of each algorithm in identifying individuals at risk of cardiovascular or heart-related ailments.

Logistic Regression (LR) Classifier achieved an accuracy of 85.93%, meaning it correctly predicted the presence or absence of heart disease in individuals 85.93% of the time. Its precision, which measures the accuracy of its positive predictions, stood at 81.68%, indicating that 81.68% of the cases it classified as positive were indeed true positives. With a recall rate of 91.85%, this model identified 91.85% of the actual positive cases. The F1-score, a balance between precision and recall, was 86.47%.

The Support Vector Machine (SVM) Classifier, on the other hand, achieved an accuracy of 85.28%. It exhibited a precision of 81.44%, indicating that 81.44% of its positive predictions were accurate, while its recall rate was 91.88%. The F1-score for this model was 86.35%.

The K-Nearest Neighbor (KNN) Classifier exhibited an accuracy of 79.87%, with a precision rate of 77.86%, suggesting that 77.86% of its positive predictions were correct. It had a recall rate of 85.36%. The F1-score for KNN was 81.44%.

Now, the Decision Tree (DT) Classifier stood out with a remarkable accuracy of 96.75%. Its precision rate was an impressive 97.81%, indicating a high accuracy in positive predictions. The model captured 95.73% of the actual positive cases (recall), resulting in a high F1-score of 96.76%.

Lastly, the Random Forest (RF) Classifier shared the same accuracy as the Decision Tree at 96.75%. It had a precision rate of 95.85% and an outstanding recall rate of 97.88%. The F1-score for Random Forest was 96.85%.

In summary, the Decision Tree and Random Forest classifiers exhibited the highest accuracy and strong F1-scores among the algorithms, signifying their effectiveness in predicting heart disease. These models excelled in both accurately identifying positive cases and capturing a substantial portion of actual positive cases. These results emphasize the potential of these algorithms in aiding the diagnosis and prediction of cardiac conditions with a high degree of accuracy.

While other machine learning algorithms, including Logistic Regression, SVM, K-Nearest Neighbor, and Random Forest, were explored, they were found to be comparatively less effective in predicting instances of cardiac illness. In essence, this research article underscores the combination of the Decision Tree and Random Forest algorithms as the most accurate approach for forecasting heart disease. This amalgamation offers a dependable means of predicting the potential development of cardiovascular or heart-related disorders in the future. While other algorithms were assessed, such as Logistic Regression, SVM, K-Nearest Neighbor, and Random Forest, they were not found to be as potent as the methods discussed in this study for predicting cardiac conditions.

V. Conclusion and Future Work

In the pursuit of advancing predictive analytics for cardiovascular diseases, this study meticulously examined various machine learning algorithms, aiming to identify the most effective approach for accurate and early prediction. The results presented in this research, encompassing a thorough analysis of different classifiers, unveil valuable insights into the realm of cardiac health forecasting.

The experiments demonstrated that the optimal performance was achieved within the 60%–80% split range of the test and train dataset. This meticulous evaluation led to the conclusion that precision, recall, and the F-1 score are pivotal metrics, often surpassing the significance of mere accuracy. Among the array of methods explored, K-Nearest Neighbor (KNN) emerged with comparatively less favorable outcomes, while regression and Support Vector Machine (SVM) methods exhibited moderate performance.

However, the spotlight of this research undoubtedly falls upon the Decision Tree (DT) and Random Forest (RF) classifiers. The DT Classifier showcased an exceptional accuracy of 96.75%, coupled with an impressive precision rate of 97.81% and a robust recall rate of 95.73%. This translated into an outstanding F1-score of 96.76%, underlining its proficiency in positive predictions. Equally noteworthy, the RF Classifier mirrored the DT's accuracy at 96.75% while achieving a remarkable precision rate of 95.85% and an outstanding recall rate of 97.88%, resulting in an exemplary F1-score of 96.85%. These results clearly indicate that the Decision Tree and Random Forest classifiers possess the highest accuracy and robust F1-scores, making them exceptionally effective in forecasting heart disease.

In contrast, Logistic Regression, SVM, K-Nearest Neighbor, and even Random Forest, despite its overall competence, fell short when compared to the superior predictive capabilities of Decision Tree and Random Forest classifiers. This study unequivocally establishes the amalgamation of Decision Tree and Random Forest algorithms as the most potent and dependable approach for predicting instances of cardiac illness. This combination not only accurately identifies positive cases but also captures a substantial portion of the actual positive instances, emphasizing their potential in aiding the diagnosis and prediction of cardiac conditions with an unparalleled degree of accuracy.

A. Research Limitations

However, as with any research endeavor, certain limitations must be acknowledged. The study's predictive approach focuses on a subset of non-invasive attributes, potentially missing out on the broader spectrum of factors that influence heart disease risk. Additionally, while the model's performance is evaluated through metrics, usability testing of the prediction tools remains unexplored, leaving room for understanding user interaction and practical implementation challenges. Moreover, the study's reliance on a specific dataset categorized by a particular ethnic group might restrict the generalizability of the findings to other populations.

B. Future Scope

The research's future trajectory offers opportunities for refinement and expansion. Further investigations could explore the efficiency of other robust machine learning techniques, such as genetic algorithms, neural networks, and hybrid models, to provide a comparative analysis of predictive performance. Expanding the model's scope to include additional non-invasive characteristics like socioeconomic status, depression severity, and ethnicity could enrich its accuracy and applicability. This might illuminate the relative importance of controlled non-invasive factors across various age and gender groups.

Furthermore, embracing diverse real-world datasets featuring multiple population groups and attributes can enhance the model's robustness and generalizability. An exciting future direction lies in the development of a comprehensive and universally applicable risk model. This model could not only predict cardiac disorders but also offer personalized treatment plans, amplifying its utility for medical professionals and patients alike. Through iterative refinement and continuous exploration, machine learning techniques hold the potential to revolutionize the landscape of heart disease prediction and prevention.

Acknowledgement

The researchers wish to extend their sincere gratitude to the Deanship of Scientific Research at the Islamic University of Madinah for the support provided to the Post-Publishing Program 2.

References

[1] Omran AR (2005). The epidemiologic transition: A Theory of the Epidemiology of Population Change. Milbank Mem Fund Q, volume 49 on page 509.

[2] World Health Organization (2010). Global status report on noncommunicable diseases 2010. https://www.who.int/nmh/publications/ncd_report_full_en.pdf

[3] World Health Organization (2011a). The Top Ten Causes Of Death. Accessed 18 August 2017, http://www.who.int/mediacentre/factsheets/fs310_2008.pdf

[4] National Center for Chronic Disease Prevention and Health Promotion (2013). Know the facts about heart disease. http://www.cdc.gov/heartdisease/docs/consumered_heartdisease.pdf

[5] European Public Health Alliance (2013). Cardiovascular Health Takes Center Stage in Brussels. Accessed 12 March 2016, from http://www.epha.org/a/5899

[6] Heart and Circulatory Disease Statistics (2019). British Heart Foundation. https://www.bhf.org.uk

[7] Shahwan-Akl, L. (2010). Cardiovascular Disease Risk Factors among Adult Australian Lebanese in Melbourne. International Journal of Research in Nursing, 1(1), 1-7.

[8] Economic and Social Survey of Asia and the Pacific (2010). http://www.unescap.org/stat/data/syb2009/9.Health-risks-causes-of-death.asp

[9] Huang Yanzhong, Moser Patricia, and Roth Susann (2015). Health in the Post-2015 Development Agenda for Asia and the Pacific.

[10] World Health Organization (2013c). Deaths from Coronary Heart Disease. http://www.who.int/cardiovascular_diseases/en/cvd_atlas_14_deathHD.pdf

[11] Gregory A. Roth (2017). Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. Journal of the American College of Cardiology. DOI: 10.1016/j.jacc.2017.04.052

[12] Ensminger, M. E., and Ensminger, A. H. (1993). Foods & nutrition encyclopaedia (Second Edition), Volume 1. CRC Press, ISBN 9780849389818

[13] Colin D. Mathers and Dejan Loncar (2006). Updated Projections of global mortality and burden of disease, from 2002 to 2030. Published online November 28. DOI: 10.1371/journal.pmed.0030442

[14] World Health Organization Press (2014). Global Status Report on Non-Communicable Diseases. https://www.who.int/nmh/publications/ncd-status-report-2014/en/

[15] Din, S., Rabbi, F., Qadir, F., and Khattak, M. (2007). Statistical Analysis of Risk Factors for Cardiovascular Disease in Malakand Division. Pakistan Journal of Statistics and Operation Research, 3(2), 117-124.

[16] Reynolds Risk Score (2015). About the Reynolds Risk Score. Accessed 10 November 2016, from http://www.reynoldsriskscore.org/home.aspx

[17] Colombet, A. Ruelland, G. Chatellier, F. Gueyffier, P. Degoulet, and M. C. Jaulent (2000). Models to predict cardiovascular risk: comparison of CART, multilayer perceptron, and logistic regression. Proc. AMIA Symp. PP. 156–60.

[18] H. Yan (2003). Development of a Decision Support System for Heart Disease Diagnosis Using Multilayer Perceptron. IEEE Int. Symp. Circuits Syst., vol. 5, pp. 709–712.

[19] Kiyong Noh, Heon Gyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu (2006). Associative Classification Approach for Diagnosing Cardiovascular Disease. ICIC, 2006, LNCIS 345, pp. 721 – 727, 2006.

[20] K. U. Rani (2011). Analysis of Heart Diseases Dataset using Neural Network Approach. Int. J. Data Min. Knowl. Manag. Process, vol. 1, no. 5, pp. 1–8.

[21] M. Kumari and S. Godara (2011). Comparative Study of Machine learning Classification Methods in Cardiovascular Disease Prediction. Int. J. Comput. Sci. Trends Technol., vol. 2, no. 2, pp. 304– 308.

[22] V. Chaurasia (2013). Early Prediction of Heart Diseases Using Machine learning. Caribb. J. Sci. Technol., vol. 1, no. December, pp. 208–217, 2013.

[23] Thuy Nguyen Thi Thu, and Darryl.N. Davis (2007). A Clustering Algorithm for Predicting Cardiovascular Risk. World Congress on Engineering 2007: 354-357

[24] M. Shouman, T. Turner, and R. Stocker (2012a). Integrating Naive Bayes and K Means Clustering with different Initial Centroid Selection methods in the diagnosis of heart disease patients. airccj.org, pp. 431–436, 2012.

[25] Markos G. Tsipouras, Themis P. Exarchos, Dimitrios I. Fotiadis, Anna P. Kotsia, Konstantinos V. Vakalis, Katerina K. Naka, and Lampros K. Michalis (2008). Automated diagnosis of coronary artery disease based on machine learning and fuzzy modeling. IEEE Trans. Inf. Technol. Biomed., vol. 12, no. 4, pp. 447–58, 2008.

[26] Aqueel and S. A. Hannan (2012). Machine learning Techniques to find out Heart Diseases : An Overview. Int. J. Innov. Technol. Explore. Eng., vol. 1, no. 4, pp. 18–23, 2012.

[27] S. U. Amin, K. Agarwal, and R. Beg (2013). Genetic Neural Network Based Machine learning in Prediction of Heart Disease Using Risk Factors. ICT 2013 - Proc. 2013 IEEE Conf. Inf. Commun. Technol., no. ICT, pp. 1227–1231, 2013.

[28] V. Chaurasia and S. Pal (2014). Machine learning Approach to Detect Heart Diseases. Int. J. Adv. Comput. Sci. Inf. Technol. Vol. 2, no. 4, pp. 56–66, 2014.

[29] Omar Karam, Mostafa A. Salama and Randa El Bialy (2016). An ensemble model for Heart disease data sets : a generalized model. ACM, pp. 191–196, 2016.

[30] Arabasadi Z, Alizadehsani R, Roshanzamir M, Moosaei H, and Yarifard AA (2017). Computer-Aided Decision Making For Heart Disease Detection Using Hybrid Neural NetworkGenetic Algorithm. Computer Methods and Programs in Biomedicine 141 (2017) 19-26. https://doi.org/10.1016/j.cmpb.2017.01.004

[31] Geurts. Pierre, Ernst. Damien and Wehenkel. Louis (2006). Extremely Randomized Trees. Machine Learn 63: 3-42. DOI 10.1007/s10994-006-6226-1

[32] Natekin. Alexey and Knoll. Alois (2013). Gradient Boosting Machines-A Tutorial. Frontiers in Neuro Robotics Volume7| Article21.

[33] Biau Gerard (2012). Analysis of a Random Forests Model. Journal of Machine Learning Research 1063-1095.

[34] Khaing T. Kyaw (2010). Enhanced Features Ranking and Selection using Recursive Feature Elimination (RFE) and K- Nearest Neighbor Algorithms in Support Vector Machine for Intrusion Detection System. International Journal of Network and Mobile Technologies VOL 1/ ISSUE1/ JUNE.

[35] Chen. Tianqi and Guestrin. Carlos (2016). XG BOOST: A Scalable Tree Boosting System. KDD‟16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Machine learning Pages 785-794 at San Francisco, California, USA.

[36] Matkovsky, I., and Nauta, K. (1998). Overview of machine learning techniques. Presented at the Federal Database Colloquium and Exposition, San Diego, CA.

[37] Richard J. Roiger (2017). Machine learning: A Tutorial - Based Primer (Second Edition). CRC Press Taylor & Francis Group New York.

[38] Jure Leskovec, Anand Rajaraman, and Jeffrey Ullman (2014). Mining of Massive Datasets (Second Edition). Cambridge University Press. ISBN-10: 1107077230.

[39] Esposito, F., Malerba, D., Semeraro, G., and Kay, J. (1997). A comparative analysis of methods for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(5), 476-491.

[40] I.Ketut Agung Enriko, Muhammad Suryanegara, and Dadang Gunawan (2018). Heart Disease Diagnosis System with k-Nearest Neighbors Method Using Real Clinical Medical Records. Published in: Proceeding ICFET '18 Proceedings of the 4th International Conference on Frontiers of Educational Technologies Pages 127-131.

[41] Purnami, S., Zain, J., and Embong, A. (2010). A New Expert System for Diabetes Disease Diagnosis Using Modified Spline Smooth Support Vector Machine. Computational Science and Its Applications – ICCSA 2010 (Vol. 6019, pp. 83-92): Springer Berlin Heidelberg.

[42] Abdullah, A. S., and Rajalaxmi, R. R. (2012). A Machine learning Model for Predicting The Coronary Heart Disease using Random Forest Classifier. IJCA Proceedings on International Conference on Recent Trends in Computational Methods, Communication, and Controls (ICON3C 2012), ICON3C (3), 22-25.

[43] Mudasir M Kirmani and Syed Immamul Ansarullah (2016) Classification models on cardiovascular disease detection using Neural Networks, Naive Bayes and J48 Machine learning Techniques". International Journal of Advanced Research in Computer Science. Volume 7, No. 5, September-October 2016.

[44] Rajul Parikh, Annie Mathai, Shefali Parikh, G Chandra Sekhar and Ravi Thomas (2008). Understanding and using Sensitivity, Specificity, and Predictive Values. Indian Journal of Ophthalmology, Jan- Feb; 56(1): 45-50.

[45] Sonam Nikhar, A.M. Karandikar (2016) "Prediction of Heart Disease Using Machine Learning Algorithms" in International Journal of Advanced Engineering, Management and Science (IJAEMS)

[46] Deeanna Kelley "Heart Disease: Causes, Prevention, and Current Research" in JCCC Honors Journal.

[47] Ponrathi Athilingam, Bradlee Jenkins, Marcia Johansson, Miguel Labrador (2017) "A Mobile Health Intervention to Improve Self-Care in Patients With Heart Failure: Pilot Randomized Control Trial" in JMIR Cardio.

[48] DhafarHamed, Jwan K. Alwan, Mohamed Ibrahim, Mohammad B. Naeem (2017) "The Utilisation of Machine Learning Approaches for Medical Data Classification" in Annual Conference on New Trends in Information & Communications Technology Applications.

[49] Mai Shouman, Tim Turner, and Rob Stocker (2013) Applying kNearest Neighbour in Diagnosing Heart Disease Patients International Journal of Information and Education Technology, Vol. 2,

[50] Joo, G.; Song, Y.; Im, H.; Park, J. (2020) Clinical Implication of Machine Learning in Predicting the Occurrence of Cardiovascular Disease Using Big Data (Nationwide Cohort Data in Korea). IEEE Access, 8, 157643–157653.

[51] Amudhavel, J., Inbavalli, P., Bhuvaneswari, B., Anandaraj, B., Vengattaraman, T., Premkumar, K., (2015) "An effective analysis on

harmony search optimization approaches", International Journal of Applied Engineering Research, 10 (3), pp. 2035-2038.

[52] Modepalli, K.; Gnaneswar, G.; Dinesh, R.; Sai, Y.R.; Suraj, R.S. (2021) Heart Disease Prediction using Hybrid machine Learning Model. In Proceedings of the 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 20– 22 January .

[53] Li, J.; Haq, A.; Din, S.; Khan, J.; Khan, A.; Saboor, (2020) 0A. Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare. IEEE Access , 8, 107562–107582.

[54] Ali, F.; El-Sappagh, S.; Islam, S.M.R.; Kwak, D.; Ali, A.; Imran, M.; Kwak, K. (2020) A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Inf. Fusion

[55] Rahim, A.; Rasheed, Y.; Azam, F.; Anwar, M.; Rahim, M.; Muzaffar, A. (2021) An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases. IEEE Access

[56] Ishaq, A.; Sadiq, S.; Umer, M.; Ullah, S.; Mirjalili, S.; Rupapara, V.; Nappi, M. (2021) Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Machine learning Techniques. IEEE Access

[57] Khurana, P.; Sharma, S.; Goyal, (2021) A. Heart Disease Diagnosis: Performance Evaluation of Supervised Machine Learning and Feature Selection Techniques. In Proceedings of the 8th International Conference on Signal Processing and Integrated Networks, SPIN 2021, Matsue, Japan

[58] Amudhavel, J., Padmapriya, S., Nandhini, R., Kavipriya, G., Dhavachelvan, P., Venkatachalapathy, V.S.K., (2016) "Recursive ant colony optimization routing in wireless mesh network", Advances in Intelligent Systems and Computing, 381, pp. 341-351.

[59] Nabil Alshurafa, Costas Sideris, Mohammad Pourhomayoun, Haik Kalantarian, Majid Sarrafzadeh (2015) "Remote Health Monitoring Out-come Success Prediction using Baseline and First Month Intervention Data" in IEEE Journal of Biomedical and Health Informatics

[60] Samineni, Peddakrishna. (2023). Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization. Processes, doi: 10.3390/pr11041210

[61] Mohammed, Ali, Shaik., T., V., Akshay. (2023). Improving Accuracy of Heart Disease Prediction through Machine Learning Algorithms. doi: 10.1109/ICIDCA56705.2023.10100244

[62] Raniya, Rone, Sarra., Ahmed, Musa, Dinar., Mazin, Abed, Mohammed. (2022). Enhanced accuracy for heart disease prediction using artificial neural network. Indonesian Journal of Electrical Engineering and Computer Science, doi: 10.11591/ijeecs.v29.i1.pp375-38

[63] Dengqing, Zhang., Yunyi, Chen., Yuxuan, Chen., Shengyi, Ye., Wenyu, Cai., Junxue, Jiang., Yechuan, Xu., Gongfeng, Zheng., Ming, Chen. (2021). Heart Disease Prediction Based on the Embedded Feature Selection Method and Deep Neural Network.. Journal of Healthcare Engineering, doi: 10.1155/2021/626002