
Research article

Breaking new ground in cardiovascular heart disease Diagnosis K-RFC: An integrated learning approach with K-means clustering and Random Forest classifier

**Ahmed Hamza Osman¹, Ashraf Osman Ibrahim², Abeer Alsadoon^{3,4}, Ahmad A Alzahrani⁵,
Omar Mohammed Barukub⁶, Anas W. Abulfaraj¹ and Nesreen M. Alharbi⁷**

¹ Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University Rabigh, Saudi Arabia

² Creative Advanced Machine Intelligence Research Centre, Faculty of Computing and Informatics, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia

³ School of Computer Data and Mathematical Sciences, Western Sydney University (WSU), Sydney, Australia

⁴ Asia Pacific International College (APIC), Sydney, Australia

⁵ Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

⁶ Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University Rabigh, Saudi Arabia

⁷ Department of Computer Science, Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Jeddah, 21589, Saudi Arabia

* **Correspondence:** Email: ahoahmad@kau.edu.sa; Tel: +966551422863.

Abstract: The ability to accurately anticipate heart failure risks in a timely manner is essential because heart failure has been identified as one of the leading causes of death. In this paper, we propose a novel method for identifying cardiovascular heart disease by utilizing a K-means clustering and Random Forest classifier combination. Based on their clinical and demographic traits, patients were classified into either healthy or diseased groups using the Random Forest classifier after being clustered using the K-means method. The performance of the proposed hybrid approach was evaluated using a dataset of patient records and compared with traditional diagnostic methods, namely support vector machine (SVM), logistic regression, and Naive Bayes classifiers. The outcomes indicated that the proposed

hybrid method attained a high accuracy in diagnosing heart disease, with an overall accuracy of 96.8%. Additionally, the method showed a good performance in classifying patients at high risk of heart disease: the sensitivity reached 96.3% and the specificity reached 97.2%. In conclusion, the proposed method of combining K-means clustering and a Random Forest classifier is a promising approach for the accurate and efficient identification of heart disease. Further studies are needed to validate the proposed method in larger and more diverse patient populations.

Keywords: classifier; K-means clustering; heart disease; Random Forest; diagnosing

Mathematics Subject Classification: 68M25

1. Introduction

Heart disease (HD) and cardiovascular disease (CVD) are significant health concerns globally, leading to a high burden of illness and mortality. The accurate diagnosis of these conditions is crucial for effective treatment and the prevention of complications. Traditional diagnostic methods, such as an electrocardiogram (ECG) and angiography, have limitations in terms of accuracy and cost-effectiveness. Therefore, alternative approaches are needed to improve the diagnostic accuracy and reduce costs. In this paper, we will discuss the importance of heart disease, global statistics, any existing diagnostic methods, the shortcomings of these methods, and how our proposed approach can overcome these challenges.

Heart disease is the principal global cause of sickness and death [1]. A primary diagnosis of HD is crucial for an effective treatment and the prevention of complications. The traditional approach for diagnosing HD involves the use of various diagnostic tests such as an electrocardiogram (ECG), a coronary angiography, and echocardiography, among others. These tests are costly and may not always be accurate. Therefore, there is a need for alternative techniques to diagnose HD more accurately and to be cost-effective.

Machine learning (ML) has emerged as a powerful tool in healthcare research, revolutionizing the diagnosis, prognosis, and decision support systems for a wide range of diseases. Notably, ML techniques have been successfully applied in the context of COVID-19 [2–5], cancer [6,7], diabetes [8], and the monkeypox outbreak [9], thereby enabling an accurate diagnosis, predicting disease outcomes, and aiding in treatment decisions. By harnessing the power of large datasets and advanced algorithms, ML can transform healthcare by improving the diagnostic accuracy, personalizing treatment plans, and enhancing disease surveillance and management.

According to [10], HD is estimated to affect around 17.9 million deaths annually. An early and accurate diagnosis of HD is crucial for effective treatment and management. ECGs and blood tests have limitations and gaps in terms of their sensitivity and specificity, and may not always provide a clear diagnosis [11]. Therefore, there is a need for alternative diagnostic methods based on an HD diagnosis. The ML methods for CVD classification are promising, since they allow for the extraction of hidden knowledge and the awareness of correlations among features included in the dataset [12–14]. The delivery of excessively great medical services that might be lower priced to patients is a crucial project going through health groups. The shipping of good service calls for an accurate diagnosis for each sufferer and the identification of powerful treatments, whilst fending off inaccurate diagnoses [12]. Additionally, early CVD detection lowers costs and CVD mortality. Using information mining techniques and a

classification algorithm, which is crucial in medical research, the process can be completed accurately for a very low cost [13]. Zhao et Al. looked into how such affordable and straightforward algorithms are likely to be of sufficient use to be applied therapeutically and as a factor to stepped-forward services [13].

As the heart is a critical organ in the human frame, any problem related to it enormously affects human health. The primary signs of CVDs are chest ache, bloating, swollen legs, respiratory problems, fatigue, and an abnormal CVD beat rhythm. The elements that reason CVDs are age, obesity, pressure, bad weight loss plans, and smoking. The principal aim of the study is to develop a predictive model using ML algorithms to anticipate cardiovascular heart disease, assist doctors in the early detection of the disease, minimize the need for extensive medical examinations, and provide timely and appropriate care, potentially leading to a significant number of lives saved. If there's a conventional method to identify heart sickness in hospitals, then why do we require machine mastering? In hospitals, a large number of statistics associated with sufferers laid low with CVD, and other sicknesses are generated every day. It is difficult for doctors to efficiently apply or take care of the patient's records to make choices without data mining strategies. Information mining is enormously endorsed for the prediction of CVD, as it extracts greater accurate and useful records from massive quantities of facts, which makes predictions easy [14]. ML serves as a fundamental pillar that facilitates the processing of vast amounts of data, thus enabling high-speed computations and early-stage predictions. The ability to handle massive datasets and swiftly analyze information is one of the key strengths of ML methodologies. These advanced algorithms aid in extracting valuable insights and patterns from data, thus leading to accurate predictions and informed decision-making in various fields, including healthcare. There are extraordinary data mining strategies that can be used, which consists of category, prediction, and recognizing patterns for diagnosing CVD. In the Bharti study [15], class models, which are a part of the gadget-gaining knowledge, were used to figure out aerobic vascular illnesses. Classification algorithms make use of input information to predict and classify data points into specific classes or categories to which the information belongs. Several types of techniques are Logistic regression, selection trees, Random wooded area, Gradient Boosting, small vector machine (SVM), Naïve Bayes, and k-Nearest Neighbor. In addition, all types of models can be trained to predict heart sickness and examine their performance through the usage of assessment metrics, along with sensitivity, accuracy, and so on, which gives a pleasant class version to predict the occurrence of heart disorders [15].

The majority of healthcare organizations and medical research facilities digitally save patient data for future use in research and treatment planning [16]. The greatest decision support systems offer an explanation along with an accurate, dependable, and prompt response [17]; additionally, they assist medical professionals and play a significant part in medical decision-making based on various models for heart disease. The issue with this study is that we need an intelligent method to identify heart disease into disease stage, pattern, or status. ML is widely accepted as a technique to choose from heart disease pattern classification and predictive modeling due to its specific advantages in detecting critical features in heart disease [18].

Healthcare organizations are increasingly using data mining methods, such as ML, to examine vast volumes of patient data and find patterns that can help with diagnoses. In this study, we suggest using a K-means technique and a Random Forest, in tandem, to diagnose heart disease. The Random Forest classifier is a supervised learning method that may divide data into multiple groups, whereas the K-means algorithm clusters data points based on their similarity. Using a dataset of patient records, the suggested method's performance will be assessed and compared to more established diagnostic techniques, such as logistic regressions and decision tree classifiers. The impartial of this study is to

identify people with a high risk for HD and explore the possibility of the suggested strategy to improve the efficacy and correctness of HD prediction.

The following sections of the document are arranged in the following manner: section 2 describes the related works; section 3 discusses materials and methods that reflect the proposed approach; section 4 presents the hybridizing of the K-means clustering algorithm and Random Forest classifier; sections 5 and 6 discuss the experimental design and results, including the finding and evaluation of the suggested method; and lastly, section 7 accomplishes the study.

2. Related works

The heart is a muscular organ that circulates blood throughout the body. It is an important part of the cardiovascular system, which also contains the lungs. Additionally, the cardiovascular system accommodates a network of blood vessels, including veins, arteries, and capillaries. These blood veins transport oxygen and nutrients throughout the body. Numerous types of cardiovascular disorders, which are commonly referred to as CVD, are caused by anomalies in regular blood flow out from the heart. Based on the report by the World Health Organization (WHO), heart attacks and strokes are to blame for 17.5 million of all fatalities globally. Over 75% of deaths attributed to cardiovascular diseases regularly occur in countries with intermediate and, in some cases, higher income levels [19]. Additionally, heart attacks and strokes account for 80% of mortality caused by CVD [20]. As a result, early detection of cardiac abnormalities and tools for CVD prediction can save many lives and assist doctors in developing successful treatment strategies, thus lowering the mortality rate from cardiovascular illnesses. Many patient statistics are now readily available (e.g., enormous data in digital health file gadgets) as a result of stronger healthcare infrastructure, which can be used to develop predictive models for cardiovascular diseases. Facts mining, also known as gadget studying, is a process for reading extensive records from a variety of perspectives and distilling the information into helpful statistics. “Data Mining is a non-trivial extraction of implicit, previously unknown and probably beneficial facts approximately facts” [12]. In recent years, healthcare businesses have produced a significant volume of data relating to disease prediction and patients. Information mining offers several methods for extracting hidden commonalities or patterns from statistical data. Therefore, a device-mastering set of rules is suggested to develop a heart disorder detection system based on the study by Weng et Al. [5], which was validated on open-access heart disorder detection data.

A diversity of circumstances that affects one’s CVD are denoted as coronary HD. Based on the analyzes by the arena health firm [21]. Numerous risk factors, such as high blood pressure, elevated triglyceride levels, and excessive levels of LDL cholesterol, contribute to a 168% rise in the risk of CVD [21].

According to Lunugalage, D., et al., the human heart is a crucial organ. All parts of our bodies receive blood from it. The brain and several other organs depend on it to function properly; if it doesn't, the person will pass away in a matter of minutes [10]. Because of alterations in lifestyle, stress at work, and unhealthful eating patterns, some heart-related illnesses are becoming more prevalent [11]. As people age, the prevalence of cardiac disease increases in both men and women. Men are more likely to have heart disease [12]. However, women are vulnerable following menopause. A hectic way of life increases the risk of CVD and damages the arteries. HD is one of the most difficult and possibly fatal human disorders [13]. Heart failure occurs when the heart is unable to transport the necessary volume of blood to other regions of the body to perform its normal functions [13]. HD is a term used to describe

a variety of heart-related diseases [14]. As one of the many life-threatening ailments, HD has received a lot of attention in medical studies. Diagnosis of HD is a tough task that can provide an automated evaluation of the patient's heart status, allowing for more efficient therapy.

A powerful computational method, called ML, uses trained data samples to automatically identify patterns and draw informed judgments. While this technique involves building a model from data, ML represents an advanced computational approach that not only automatically recognizes intricate patterns, but also enhances decision-making by leveraging training data samples, resulting in more precise and accurate outcomes. ML refers to a machine's ability to learn from a vast amount of data and either forecast, cluster, or classify comparable but fresh or new data based on that learning. A few well-known ML methods such as artificial neural networks (ANN) and SVM, as well as k-means clustering, decision trees, and self-organization map means clustering. In addition, ensemble techniques integrate the outcomes of various categorization algorithms to provide a better final product. A prevalent cardiovascular condition with a high fatality rate is coronary artery disease (CAD). While angiography is considered the gold standard for the clinical diagnosis of CAD, doctors often recommend it due to its high accuracy in visualizing blocked arteries. However, angiography does have some drawbacks, such as its invasive nature, which carries a small risk of complications, and its reliance on the use of contrast dye, which can cause allergic reactions in some patients. Additionally, angiography is relatively expensive compared to other diagnostic tests. Therefore, while it remains a valuable tool to diagnose CAD, doctors need to carefully weigh its benefits against the potential risks and costs for each patient [22]. Numerous studies have been conducted in the field of ML to develop substitutes for this kind of clinical diagnosis [23]. HD has steadily grown in its importance as a global public health issue as a result of ignorance, inappropriate consumption, and a poor lifestyle. Today, it is extremely difficult for hospitals and medical professionals to precisely forecast and diagnose it. Healthcare facilities have benefited from the growth of computing technology by being able to collect and retain data for clinical decision-making. In many modern nations, hospitals gather and keep patient data in a computerized and comprehensible way [24]. People all over the world are impacted by CVD, which has been identified as a severe public health issue. Physicians can forecast CVD with the aid of algorithms that integrate the analysis of clinical biomarkers with many well-known traditional risk factors, thus increasing the dependability of clinical decision-making [18]. To handle healthcare data, ML for health informatics has emerged as an interdisciplinary field of research [25]. Data mining is useful in the healthcare industry since healthcare databases are typically large. Large amounts of data are transformed through data mining into knowledge that may subsequently be used to make more accurate predictions and judgments [26]. Numerous studies have been conducted to identify the ML approaches that have been used to diagnose heart illnesses. For instance, Pouriyeh, S., et al. [27] compared and contrasted several ensembles and classification techniques for data mining. In the future, they will investigate some of the more intricate methods to assess the epistemic uncertainty that are appearing in the literature. The ML methods used to study cardiac disease are shown in Table 1.

As shown in Table 1, some studies focus on various ML methods and techniques to forecast and detect different types of heart diseases. The models and approaches employed in these studies include decision trees, SVMs, rules-based classifiers, fuzzy logic, ensembles, and ANNs. The data types used ranged from categorical to integer and real, with a particular emphasis on ECG signals. The accuracies achieved by these models varied between 69.22% and 94.83%. These studies contributed to the development of predictive models and decision support systems for heart disease diagnoses and management, thus highlighting the importance of ML in improving healthcare outcomes.

Table 1. ML approaches used Heart Disease Datasets.

Source and Year	Suggested Method	Nature of Data	Specificity	Sensitivity	F-measure	Accuracy %
[28] 2021	Decision trees.	Integer/Real	-	-	-	94.0
[29] 2018	SVM	ECG signals	0.86	0.88	-	87.7
[30] 2018	k-NN	ECG signals	0.92	0.86	-	89.3
[30] 2018	MLP	ECG signals	0.92	0.89	-	91.1
[13] 2011	Rules-based classifier.	Categorical,Integer, Real	-	-	-	86.7
[23] 2012	Congestive heart failure (CHF) recognition	Categorical, Integer, Real	0.84	0.83	-	84.3
[31] 2018	Random forest with a linear model (HRFLM)	Categorical,Integer, Real	0.92	0.82	0.90	88.4
[32] 2021	MLP-NN	Categorical,Integer, Real	0.95	0.91	-	93.3
[33] 2022	Logistic Regression	Categorical,Integer, Real	0.69	0.68	0.68	68.8
[33] 2022	Naïve Bayes	Categorical,Integer, Real	0.71	0.41	0.52	71.9
[32] 2021	Random Forest	ECG signals	0.97	91.1	-	95.0
[33] 2022	Decision tree	Categorical,Integer, Real	0.79	0.82	0.81	69.22
Proposed Method- K-RFC	K-means and Random Forest	Categorical,Integer, Real	0.97	0.94	0.96	96.8

This section discusses the studies that appeared in Table 1 in more detail based on the data used. Previously, a model was developed to forecast coronary heart disease, thereby achieving an impressive accuracy of 94% using decision trees [13]. This model provides a valuable tool to predict the occurrence of coronary heart disease, aiding in proactive management and preventive measures. The study in [15] aimed to assist untrained clinicians in assessing the danger of heart disease. They employed a rules-based classifier and achieved an accuracy of 86.7%. By utilizing categorical, integer, and real data, this approach supported clinicians to make informed decisions regarding the severity of heart disease, ultimately improving patient care.

Fuzzy experts refer to a concept in the field of fuzzy logic, where multiple rule-based systems are combined to enhance decision-making and classification processes. Fuzzy logic is a mathematical approach that deals with uncertainty and imprecision in data, thus enabling the representation of vague or ambiguous information. In this context, Spencer [20] proposed a novel approach that utilized both rule-based systems and fuzzy logic experts and achieved an accuracy of 84.2% when the presence of CAD was identified. This innovative method holds promise for the early-stage detection of CAD, thus enabling timely interventions and potentially improving patient outcomes. By integrating fuzzy experts with rule-based systems, the approach leverages their complementary strengths, leading to more precise and reliable predictions in diagnosing CAD.

The authors in [21] introduced a method to identify a cardiac disease by combining interval type-2 fuzzy logic with rough sets-based attribute reduction. With an accuracy of 82.6%, this approach demonstrated the potential to accurately identify different types of cardiac diseases, thus enabling targeted treatment and management strategies. The authors in [26] proposed an ensemble method called HM-BagMoov, which achieved an accuracy of 86.2%. By incorporating categorical, integer, and real data, this ensemble model provided valuable support and guidance to healthcare professionals in decision-making processes.

The authors in [28] proposed a disease prediction method which utilized an ensemble approach combined with ANNs. With an accuracy of 85.31%, this method capitalized on the strengths of ensembles and ANNs to improve disease prediction accuracy, thus enhancing proactive healthcare interventions. The authors in [31] introduced a comprehensive approach that involved random forest, C5.0, and fuzzy modeling, resulting in an accuracy of 90.50%. This multi-model approach offered an effective means to accurately diagnose coronary artery disease, thus enabling timely interventions and appropriate treatment plans. The authors in [32] utilized a fuzzy rules-based method to predict heart disease, thereby leveraging categorical, integer, and real data. This method enabled the accurate prediction of heart disease based on patient-specific characteristics, supporting risk assessment and personalized healthcare approaches. The second category studies were performed to analyze ECG signals and make accurate predictions as follows:

The authors in [14] provided an optimization approach using SVM, which achieved a commendable accuracy of 87.7%. By analyzing ECG signals, this approach contributed to the early detection and prediction of paroxysmal atrial fibrillation, aiding in timely interventions and the effective management of this cardiac condition. The authors in [27] employed a combination of SVM, ANN, and Naïve Bayes, and achieved an impressive accuracy of 94.83%.

The analysis of ECG signals performed in [29] used these ML techniques and allowed for accurate and personalized recommendations for heart patients, thus ensuring optimized healthcare interventions and improved patient outcomes.

Likewise, Aborokbah MM et al. [34] aimed to develop an adaptable and context-aware decision-making system for intensive healthcare provision. The researchers utilized the Radial Basis Function (RBF) with SVM and the Least K-Fuzzy (LKF) with SVM algorithms, and achieved an accuracy of 87.9%. By considering contextual information and analyzing ECG signals, this system facilitated informed decision-making by providing intensive healthcare, thus enabling timely interventions and tailored treatment strategies.

In our study, the contribution lies in the integration of these two techniques to create a hybrid approach specifically designed for the diagnosis of heart disease. By leveraging the strengths of both clustering and ensemble learning, we aim to improve the accuracy and efficiency of identifying heart disease. We believe that this integrated approach offers a unique perspective and potential advancements in the field of heart disease diagnoses.

3. Materials and methods

3.1. Pre-processing

Missing values, outliers, and categorical variables were handled in advance. First, the mean value for numerical data and the model for categorical attributes were used to impute missing values. Second,

outliers were removed using the interquartile range (IQR) method [35]. Finally, categorical variables were encoded using one-hot encoding [36]. In our pre-processing step, we employed a conservative approach to handle missing values by imputing them using the mean value for numerical data and the mode for categorical attributes. This decision was made to maintain the integrity and completeness of the dataset, as well as to avoid potential biases introduced by the complete case analysis.

The quantification of the proportion missing in the dataset was performed by calculating the percentage of missing values for each feature. For numerical data, the mean value imputation method was used to fill in the missing values, which ensured that the imputed values aligned with the central tendency of the data. For categorical attributes, the mode imputation method was applied to handle missing values, thereby replacing them with the most frequent category. The rationale for using the IQR method to remove outliers was to identify extreme values that significantly deviated from the central distribution of the data. Outliers could potentially distort statistical analyses and modeling results, leading to inaccurate conclusions. By removing outliers, the dataset's overall distribution became more representative of the majority of cases, improving the accuracy and reliability of subsequent analyses and modeling processes. After applying the IQR method to identify and remove outliers, we observed a notable improvement in the dataset's overall distribution. The IQR method helped us identify extreme values that significantly deviated from the central distribution of the data, and their removal contributed to a more representative dataset. We observed that there are 9, 5, 5, and 1 instances of outliers in the dataset for the features *trestbps*, *chol*, *thalach*, and *oldpeak*, respectively.

3.2. *K-means algorithm*

The K-means algorithm is a popular method for clustering data based on similarity. It entails dividing data into several clusters (k), with each data point belonging to the cluster with the closest mean. The process is iterative, and the means of the clusters are computed at each iteration depending on the data points in the cluster.

As a clustering background, Liang-qun Li et Al. [37] introduced a novel particle filter, known as the quadrature particle filter (QPF), which incorporated fuzzy c-means clustering. The proposed algorithm utilizes quadrature point probability densities to approximate the predicted and posterior probability density functions of the particle filter as Gaussian distributions. Instead of using traditional particle weights, the fuzzy membership degrees derived from a modified version of the fuzzy c-means clustering algorithm are employed. Furthermore, the quadrature point weights are adaptively estimated based on the weighting exponent and particle weights.

Previtali et Al. [38] presented a cluster-based data association method to enhance the performance of a distributed particle filter. They proposed a robust disambiguation technique applicable to the RoboCup scenario that is capable of handling noise and false perceptions. Experimental results obtained from both simulated and real environments demonstrated the effectiveness of the proposed approach.

Kerdvibulvech [39] developed a real-time hand motion recognition method using an extended particle filter. The approach combined a deterministic clustering algorithm and a particle filter based on an adaptive algorithm for calculating skin color probabilities. The proposed method demonstrated excellent resilience to luminance changes and effectively determined the probabilities of fingertips by utilizing semicircle models to fit curves.

Raziperchikolaei and Jamzad [40] introduced an online generative tracking filter algorithm to address object shape changes and illumination variations. Their approach utilizes a particle filter

structure in which samples are weighted based on their distance from the model. The model, representing a color distribution, is updated using the D2-clustering algorithm.

The mathematical equations behind the K-means algorithm can be broken down into two main steps:

1: Initialization step: In this step, the algorithm randomly selects K observations from the dataset to serve as the initial centroids for the K clusters. The centroid of a cluster is simply the mean of all the observations in that cluster. The initial centroids can be represented by the matrix, $\mu = [\mu_1 \dots \mu_k]$, where μ_i is the centroid of the i-th cluster.

2: Iteration step: In this step, the algorithm assigns each observation to the cluster whose centroid it is closest to. The distance between an observation and the centroid can be calculated using the Euclidean distance metric:

$$d(x, \mu_i) = \sqrt{(x_1 - \mu_{i1})^2} + \sqrt{(x_2 - \mu_{i2})^2} + \dots + \sqrt{(x_n - \mu_{in})^2} \quad (1)$$

where x is the observation and μ_i is the centroid of the i-th cluster.

After the assignment step, the algorithm then updates the centroids for every group by calculating the mean of all the observations in that group. Then, the new centroids are used in the next assignment step, and the process is repetitive until the centroids either do not change or reach the maximum number of iterations.

The goal of k-means is to split a collection of n samples, each represented by a d -dimensional real vector (x_1, x_2, \dots, x_n) , into k ($\leq n$) groups, referred to as $S = [S_1 \dots S_k]$, in a way that minimizes the sum of squares within each cluster (WCSS), also known as a variance. Formally, the aim is to find the following:

$$\arg \min_S \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2 = \arg \min_S \sum_{i=1}^k |S_i| \text{Var} S_i \quad (2)$$

where the optimization of the within-cluster sum of squares (WCSS) can be achieved by finding the mean (μ_i) of all the observations in each group (S_i), which is the same as reducing the squared differences between the observations in the same cluster.

$$\arg \min_S \sum_{i=1}^k \frac{1}{|S_i|} \sum_{x, y \in S_i} \|x - y\|^2 \quad (3)$$

The correspondence can be construed from the following identity:

$$|S_i| \sum_{x, y \in S_i} \|x - \mu_i\|^2 = \frac{1}{2} \sum_{x \neq y \in S_i} \|x - y\|^2 \quad (4)$$

Reducing the WCSS is comparable to increasing the sum of the squared differences across observations in different clusters, as long as the overall variance stays constant (also known as the between-cluster sum of squares, BCSS). This direct connection can be traced back to probability theory's law of total variance [1].

3.3. Random Forest algorithm

The supervised learning subset of ML includes the well-known Random Forest (RF) algorithm. Both classification and regression issues can benefit from its use. The approach is based on ensemble learning, which combines several classifiers to take on a challenging task and improve the performance

of the model. A "Random Forest" is a classifier that, as the name suggests, combines numerous decision trees trained on various subsets of the dataset to improve the prediction accuracy. It considers the predictions of each tree and forecasts the final result, depending on the majority vote rather than solely relying on the output of one decision tree. The RF algorithm is a type of ensemble learning algorithm that hybridizes various decision trees to perform the prediction. The idea of this algorithm is to train several decision trees on different subsets of the data, and then either average (for regression) or vote (for classification) the predictions of the individual trees to obtain a final prediction. The general equation for the Random Forest algorithm can be represented as follows:

$$P = \sum \frac{\text{Average Weight for the decision tree}}{\text{Number of decision tree}} \quad (5)$$

where P denotes the prediction score.

The algorithm works by randomly choosing a subset of the training dataset, known as the bootstrap sample, to train each decision tree. Additionally, at each node of each decision tree, the algorithm only considers a random subset of the features when making a split, rather than considering all features. This lessens overfitting and enhances the model's generalization capabilities. Then, the final prediction is established by either taking a majority vote or averaging the predictions of all the decision trees which are for a regression or for a classification, respectively.

The two primary hyperparameters for the Random Forest method are the number of decision trees in the forest and the number of features to take into account at each split. The performance of the model is often improved by increasing the number of decision trees in the forest but at the expense of higher computing time and memory utilization. Here is a general algorithm for training an RF model:

1. Create a random subset of the training data for each decision tree. This is typically done by selecting a random sample of the data with replacement, known as a bootstrap sample.
2. For each decision tree, grow the tree using the bootstrap sample. At each node of the tree, only consider a random subset of the features when making a split. This supports reducing overfitting and improving the model's generalization performance.
3. Repeat steps 1 and 2 for a specified number of decision trees.
4. For a new input to predict, feed the input to all decision trees.
5. If it's a classification problem, then take the majority vote among all trees, the class that achieves the highest votes is the final output.
6. If it's a regression problem, then take the average of all the predictions made by the decision trees.

There are different reasons for using the Random Forest algorithm. First, it requires less training time in comparison to other algorithms. Second, it provides accurate results, even for large data, and performs well. In addition, it can maintain a high level of accuracy, even if a significant portion of data is lost. The RF algorithm operates in two stages: the creation of a forest of decision trees using a random subset of the training set and making predictions based on the majority vote of those trees. The process is as follows: select K random points from the training data; use them to build decision trees; repeat this process N times to create a forest of decision trees; then, for new data points, have each tree make a prediction; and assign the new data point to the category with the most votes from the decision trees. Determining which feature to split is often the most time-consuming part of learning with decision trees. By reducing the features, the process of learning the tree is significantly accelerated.

3.4. Hybridizing of K-means clustering algorithm and Random Forest classifier

The hybridization of the K-means and the RF classifier has several potential advantages to diagnose heart disease. The key benefit of using the K-means is an unsupervised method, which means that it does not require labeled data. This is useful in cases where the labeling of the data is time-consuming or costly, as is often the case in medical diagnoses. Additionally, the K-means is a computationally efficient approach, and it can handle large data, thus making it well-suited for the analysis of medical data. The use of the RF classifier, in combination with the K-means, can be added to increase the accuracy of the diagnosis. The RF classifier is a powerful technique to achieve classification tasks and has been shown to perform well on a variety of datasets. By utilizing the RF classifier to classify the data points into the appropriate cluster, we can improve the accuracy of the diagnosis beyond what is possible with the K-means alone. However, it is important to note that the combination of K-means clustering and the random forest classifier is not without its limitations. One potential limitation is that the performance of the method may be sensitive to the choice of the number of clusters in the K-means algorithm and the hyperparameters of the RF classifier. Additionally, the method may not be suitable for all types of data and may require further optimization for specific datasets.

In our work, determining the value of K in the K-means clustering algorithm was a crucial step, and we addressed this challenge through a combination of methods. First, we employed the elbow method, which involved running K-means clustering with different values of K and identifying the point where the within-cluster sum of squares begins to level off. This point is often considered as an optimal choice for K, reflecting the trade-off between the model complexity and the goodness of fit.

Additionally, we utilized a silhouette analysis, which measures how well-defined the clusters are. A higher silhouette score indicates better-defined clusters, thus aiding in the selection of an appropriate K. Additionally, cross-validation was applied to assess the stability and generalizability of the chosen K value.

For the hyperparameter optimization process of the RF classifier, we employed a combination of a grid search and cross-validation. The grid search involved systematically exploring a predefined hyperparameter grid, while cross-validation provided a robust evaluation of different hyperparameter combinations. This allowed us to identify the optimal set of hyperparameters that maximized the performance of the random forest classifier.

To combine the RF with the K-means method, we first used K-means to identify groups of similar observations in the data. Then, for each group, we trained a separate Random Forest model and used the feature importance from the trees to identify which features are most important to separate the observations in that group. Second, we used k-means clustering to identify similar observations in the data; then, we used the cluster assignments as a new feature in the RF model. By using this approach, we can tell the model where similar data points are likely to have the same labels, which could improve the model's capability to generalize to new data. Finally, we assigned the input data to random clusters and then fitted a RF model on the clustered data. This can allow the RF to handle high-dimensional data by grouping similar observations and reducing overfitting on the dataset. The RF classifier was trained on the clusters obtained from the K-means algorithm, with the binary label as the goal variable. The RF classifier creates an ensemble of decision trees and uses majority voting to make a final prediction. The parameter tuning process for K-means clustering involves several key steps. First, determining the number of clusters (K) is crucial. This is achieved through methods such as the elbow method, where the WCSS is plotted against different K values, thereby identifying the "elbow" point. Additionally, a silhouette analysis was also employed to calculate the silhouette scores and select the

K value with the highest score, thus ensuring better cluster compactness and separation. Then, validation techniques, including cross-validation, were utilized to assess the stability and generalizability of the chosen K value, with adjustments made for robustness, if necessary. Additionally, experimenting with different initialization methods and evaluating their impact on clustering results helps in selecting the most effective initialization method. Lastly, a sensitivity analysis was performed to assess clustering results' sensitivity to variations in input data, and adjustments to parameters, including K, were made based on the analysis.

In the pursuit of maximizing the RF model performance, a focused approach to hyperparameter tuning is essential. This involves a systematic exploration through a grid search, where a predefined hyperparameter grid is meticulously examined to identify optimal combinations. Complementing this, cross-validation is employed, thus providing a robust evaluation of diverse hyperparameter combinations, ensuring the chosen configuration generalizes well across different datasets. The key hyperparameters take center stage in this optimization process. The number of decision trees in the forest is tuned to strike a balance between an enhanced performance and an increased computational time. Simultaneously, attention is directed toward the number of features considered at each split, aiming to optimize this subset for improved model generalization. Additionally, customization extends to other hyperparameters, with fine-tuning undertaken based on the unique characteristics of the specific dataset. This meticulous parameter tuning approach is pivotal in unlocking the full potential of the RF classifier, tailoring its configuration to the intricacies of the data at hand and ensuring an optimal performance across various scenarios. The algorithm steps for the hybridizing of K-means and RF are as follows:

Hybridizing of K-means and Random Forest Algorithm

Input	Dataset D, number of clusters k, number of trees T in the Random Forest
Output	Hybridizing of K-means and Random Forest model
Step 1	Apply K-means to the dataset D to form k clusters
Step 2	For each cluster, create a sub-dataset D' consisting of the data points in the cluster
Step 3	Train a Random Forest Classifier on each sub-dataset D' using T trees
Step 4	Store the trained Random Forest Classifier for each cluster
Step 5	Combine the classifiers into a single Hybrid Model
Step 6	Return the Hybridizing Model

In this study, various stages are composed of sub-stages, beginning with the preparation of the heart disease dataset in the planning stage and concluding with the presentation of experimental results and discussion. The novelty of this research can be summarized in the following points:

- The study introduces a novel hybrid approach for identifying cardiovascular heart disease by combining K-means clustering and a RF classifier.
- The proposed method integrates two distinct techniques, thereby leveraging the clustering capabilities of K-means and the predictive power of the RF, to enhance the accuracy and efficiency of heart disease identification.
- The proposed method offers a comprehensive solution for classifying patients into either healthy or diseased groups based on their clinical and demographic traits, thus providing a holistic approach to heart disease diagnosis.

Figure 1 illustrates the operational framework phases that have been followed in this study.

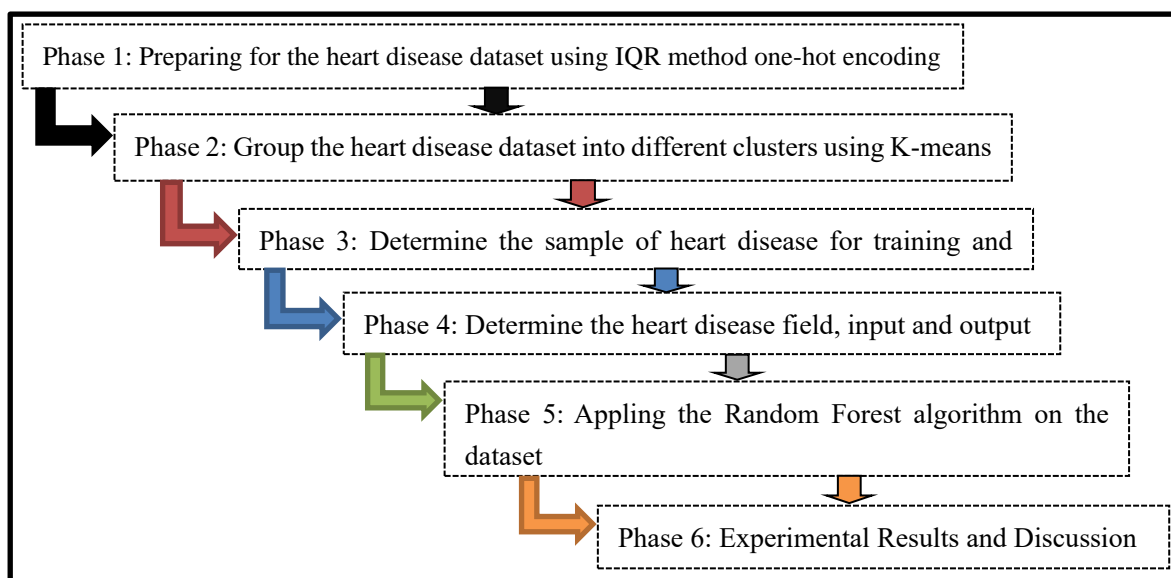


Figure 1. Operational Framework phases.

Figure 2 illustrates the proposed model architecture.

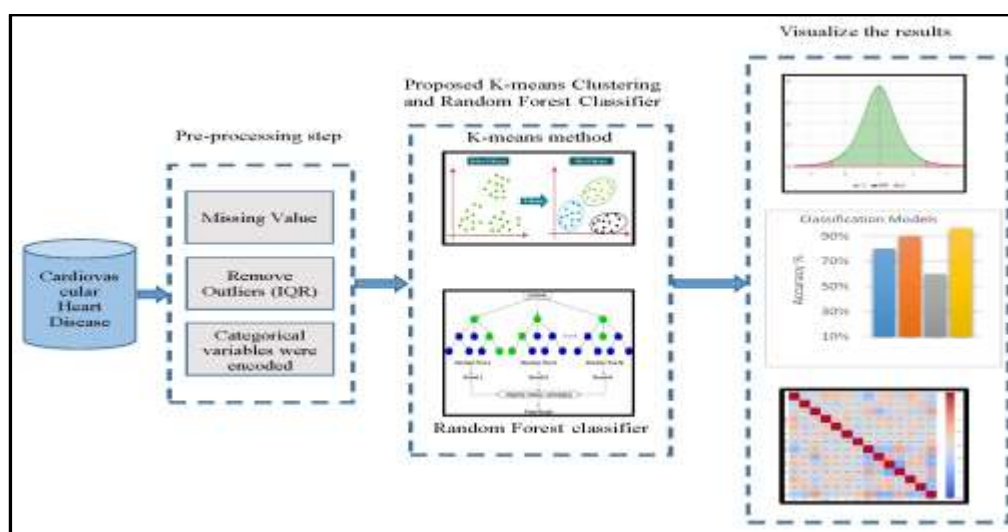


Figure 2. Proposed Method Architecture.

4. Experimental design

Due to a lack of resources in the medical community, the prognosis of heart disease might be occasionally challenging. Adequate technological help in this area could have a huge positive impact on both the medical community and patients. The proposed hybrid classifier is very well suited to use in heart disease prediction. This model makes use of information on blood pressure, cholesterol, and diabetes, before attempting to forecast potential heart disease in patients. Trying to prevent the risk of heart disease in the patient may aid in the implementation of preventative measures. Therefore, the medical information for the patient can be carefully examined by the doctors when a patient is expected to be positive for heart disease. The advantages of the suggested model for heart disease can predict

heart disease in the early stage, the cost of the medication will be minimized, and the accuracy rate will be high with a high performance.

The hybridization process of the K-means and the RF classifier is effective in a variety of applications, including the diagnosis of heart disease. One key advantage of this combination is that it allows for the combination of both unsupervised and supervised methods. A distinguished K-means method is an unsupervised method, which means it does not require labeled data. This is valuable in cases where the labeling of the data is either time-consuming or costly, as is often the case in medical diagnoses. On the other hand, the RF classifier is a supervised method, which means it needs labeled data for training. By combining these two methods, we can take advantage of the strengths of both approaches to enhance the correctness of the diagnosis. In terms of the experimental design, the hybridization of the K-means algorithm and the RF classifier can be implemented as follows:

- The pre-process step is the data that should be scaled to have a mean of zero and a standard deviation of 1 by using imputation methods such as mean imputation, median imputation, and predictive imputation.
- Use the K-means clustering algorithm to cluster the data into different clusters for patients with heart disease and healthy individuals.
- Use the RF classifier to classify the data points into the appropriate cluster.
- Evaluate the performance of the hybridization of the K-means clustering and the RF classifier using metrics such as accuracy and mean errors.
- Optimize the performance of the combination by adjusting the number of clusters in the K-means and the hyperparameters of the random forest classifier.

4.1. Heart disease dataset

The dataset used in this study was specifically constructed to develop a predictive ML model aimed at the early detection of CVD. To create this comprehensive heart illness dataset, five distinct, but previously separate datasets related to heart disease, were amalgamated. This consolidated dataset was comprised of five heart datasets, all sharing eleven common features. Consequently, it stood as the most extensive heart disease dataset currently available for research purposes. This extensive dataset was primarily derived from the renowned Cleveland dataset.

In this research, the publicly accessible HD dataset from the UCI Machine Learning Repository [41] was employed. This dataset contains the records of 303 individuals, encompassing demographic and clinical characteristics. The dataset incorporates 14 parameters, including age, gender, type of chest discomfort, resting blood pressure, serum cholesterol, and others. These features are pivotal in diagnosing cardiac disease in these individuals. Moreover, the dataset includes a binary label indicating either the presence or absence of cardiac disease. This binary classification served as the benchmark to evaluate the performance of the proposed hybrid method.

Specifically, the heart disease prediction dataset used in this study is referred to as the Cleveland Clinic Foundation (CCF) dataset. Compiled between 1988 and 1991, the CCF dataset contains information on patients diagnosed with heart disease and a control group of patients without heart disease. It consists of 14 attributes, including age, sex, blood pressure, and cholesterol levels, along with a binary target variable indicating either the presence or absence of heart disease. Notably, this dataset encompasses 76 attributes and 303 observations of patients diagnosed with heart disease, as well as a control group without heart disease. The CCF dataset has proven to be invaluable for

researchers to explore ML techniques for cardiac disease prediction. Various modeling techniques applied to this dataset have consistently showcased the utility of ML in heart disease prediction, as evidenced in Table 2.

Furthermore, an additional dataset comprised of 1190 samples and 11 attributes was utilized to scrutinize our proposed model. These datasets were meticulously collected and consolidated into a unified repository, facilitating further exploration into ML and data mining approaches related to CAD. This collaborative effort aims to not only advance research, but also holds the promise of enhancing clinical diagnoses and enabling early intervention. Given the alarming prevalence of heart disease as a major public health concern and the leading cause of global mortality, the utilization of ML algorithms for heart disease prediction and diagnoses has gained substantial traction in recent years. Developing precise and efficient models for cardiac disease prediction holds a paramount importance, as it enables early detection and timely intervention, ultimately leading to improved patient outcomes.

Table 2. HD dataset description.

Feature No	Type	Feature Name	Description and Domain
1	Integer	age	Age of the patient
2	Categorical	sex	Sex of the patient(0=women,1=male)
3	Categorical	cp	Chest pain type (0=typical angina,1=atypical angina,2=non-angina pain,3=asymptomatic)
4	Integer	trestbps	Resting blood pressure
5	Integer	chol	Cholesterol in mg/dl
6	Categorical	fbs	Fasting Blood Sugar (0 = not present; 1 =present)
7	Categorical	restecg	Resting electrocardiographic (ECG) results. (0 = normal, 1 = abnormal ST-T Wave (mild symptoms to severe problems signals non- normal heartbeat), 2 = Possible or definite left ventricular hypertrophy Enlarged heart's main pumping chamber (severe condition)
8	Integer	thalach	Maximum heart rate achieved
9	Categorical	exang	exercise-induced angina pectoris (1 - yes; 0 - no) pectoris (a disease marked by brief sudden attacks of chest pain or discomfort caused by deficient oxygenation of the heart muscles usually due to impaired blood flow to the heart)
10	Integer	oldpeak	ST depression induced by exercise relative to rest looks at the stress of the heart during exercise unhealthy heart will stress more
11	Categorical	slope	The slope of the peak exercise ST segment: 0: Upsloping: it shows better heart rate with exercise (uncommon) 1: Flatsloping: it shows minimal change (typical healthy heart) 2: Downsloping: it shows the signs of an unhealthy heart
12	Integer	ca	Number of major blood vessels with a fluorescent color (0-4), (Fluorescent color is mainly associated with diabetes)
13	Categorical	thal	Thalium stress result. (The results of this test will tell you about the flow of blood to your heart through your coronary arteries).
14	Integer	target	Have Heart disease or not (0=no, 1=yes)

These datasets are publicly available and can be used as a starting point for researchers interested in developing ML models for heart disease prediction and diagnoses. Figure 3 illustrates the heart

diseases distributions based on the target field.

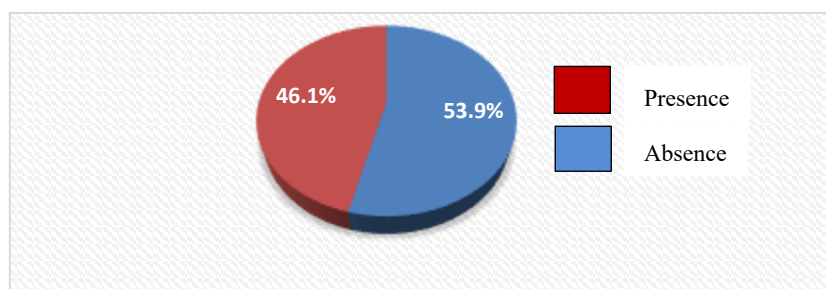


Figure 3. Heart diseases distributions (Target Field).

The total number of patients who have heart disease is higher than that of the patients who have no heart disease.

The age histograms in Figure 4, categorized based on the target presence (1) or absence (0) of heart disease, exhibit distinct distribution patterns, thus suggesting a correlation between age and heart disease. The presence of heart disease displays a left-skewed distribution, while the absence of heart disease follows a more symmetrical, normal distribution. These visual representations imply a higher prevalence of heart disease among older individuals compared to their younger counterparts. Figure 4 illustrates the heart diseases distributions based on the Age field.

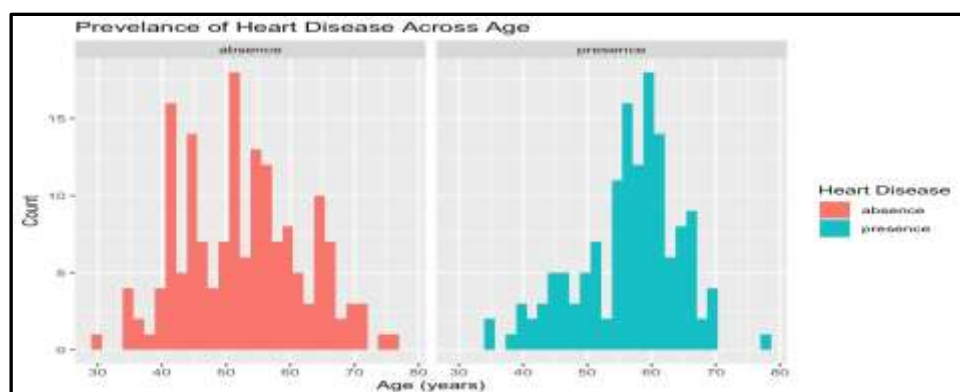


Figure 4. Heart diseases distributions (Age Filed).

5. Results and discussion

The suggested method's performance was assessed using multiple metrics, including accuracy, sensitivity, specificity, precision, and F1-score. Additionally, the method was compared to traditional diagnostic methods, such as logistic regression and decision tree classifiers. The results were analyzed and discussed to highlight the potential and limitations of the proposed method.

Figure 5 presents the correlation map (i.e., heatmap) of numerical variables in the heart disease dataset, thereby offering valuable insights into the relationships among various attributes. In this dataset, researchers typically focus on a subset of 14 attributes for their experiments. Among these attributes, the "target" field is of particular interest, as it indicates the presence of heart disease in the patient, with integer values ranging from either 0 or 1. By examining the correlation map, researchers

can identify which attributes are positively or negatively correlated with the presence of heart disease. A strong positive correlation between an attribute and the "target" field suggests that an increase in that attribute is associated with a higher likelihood of heart disease presence, while a strong negative correlation indicates the opposite. This information can be crucial in understanding the significant factors that contributes to heart disease and informs the selection of relevant attributes for predictive modeling using ML techniques and algorithms. Figure 5 illustrates sample correlation map using heart diseases with clustering data.

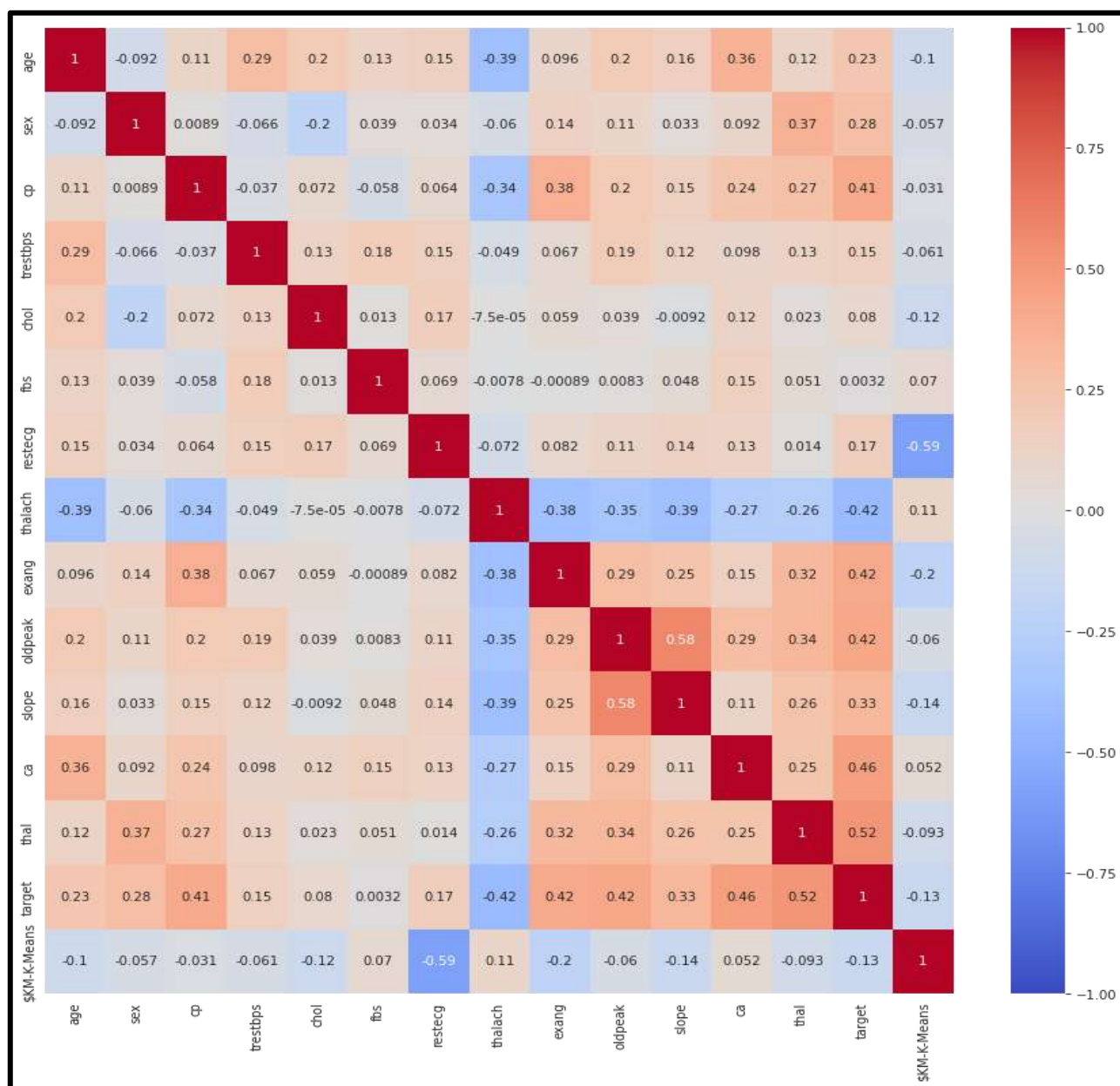


Figure 5. Sample correlation map using heart diseases with clustering data.

In Figure 5, where we applied K-Means clustering to a subset of attributes (labeled as KM-K-Means), we observed intriguing distinctions within subgroups that significantly contributed to our

understanding of the dataset. The KM-K-Means clustering facilitated the identification of latent patterns and relationships among numerical variables, thereby offering a more nuanced perspective on the correlation landscape.

Upon closer inspection, we found that certain numerical attributes within specific K-Means clusters exhibited heightened correlation coefficients compared to the overall dataset. This suggests that, within these subgroups defined by K-Means, certain combinations of attributes are more tightly interrelated. Such distinct correlation patterns within clusters could imply either subgroup-specific characteristics or shared influences that are critical to understand the dataset's complexities.

For instance, if we observed a higher correlation between two attributes within a particular KM-K-Means cluster, it might suggest a more pronounced association between those attributes in a specific subgroup of the dataset. This insight goes beyond the general correlation map, thus providing a finer-grained understanding of relationships that might be obscured in the aggregate analysis.

The implications of KM-K-Means clustering in correlation analyses extend to feature engineering and targeted investigations. By identifying subgroups with heightened correlations, researchers can tailor feature selection strategies for different clusters, thus potentially improving the performance of ML models within each subgroup. Additionally, understanding these nuanced correlations within clusters can guide further domain-specific investigations, leading to more targeted hypotheses and informed decisions in subsequent analyses.

Moreover, the correlation map aids in feature selection, as highly correlated attributes might lead to multicollinearity issues, thereby affecting the model's stability and interpretability. By selecting the most relevant and uncorrelated attributes, ML researchers can build more robust and accurate models for heart disease diagnoses. Being the most commonly used dataset for ML research in this domain, the Cleveland database serves as a benchmark to evaluate the performance of various algorithms and approaches. The goal of such studies is to create predictive models capable of accurately identifying heart disease and to ultimately assist healthcare professionals in making timely and precise diagnoses. Table 3 demonstrates our results are based on Model-1 (RF classifier without a clustering process).

Table 3. Information on the proposed Model-1.

Class Feature	Target
Classifier	Random Trees Classification
Features Input	13
Diagnosis results	0.961
Misdiagnosis Rate	0.039

In Table 3, we present the results of our RF classifier without incorporating a clustering process. The classifier achieved an accuracy of 96.1%, thus indicating its strong performance in accurately predicting the presence or absence of heart disease. The misdiagnosis rate is reported to be 3.9%, highlighting the robustness of our approach. Table 4 demonstrates Model-1 decision rule prediction using RF decision tree.

Table 4. Model-1 High Decision-Rules.

Decision-Rule	Frequent-Group	Rule-prediction	Forest-prediction	Interestingness-Index
(trestbps > 110.0) and (sex > 0.0) and (cp <= 0.0) and (ca > 0.0) and (slope > 1.0)	0	1.000	1.000	1.000
(restecg <= 0.0) and (thalach <= 145.0) and (oldpeak > 0.4) and (ca > 0.0) and (slope <= 1.0)	0	1.000	1.000	1.000
(age <= 60.0) and (sex <= 0.0) and (ca <= 0.0) and (slope > 1.0)	1	1.000	1.000	1.000
(restecg <= 0.0) and (exang > 0.0) and (ca > 0.0) and (age > 53.0) and (cp <= 0.0)	0	1.000	1.000	1.000
(trestbps > 130.0) and (cp <= 1.0) and (ca > 0.0) and (oldpeak > 0.8) and (slope <= 1.0)	0	1.000	1.000	1.000

The RF algorithm is a powerful ensemble ML technique that can be utilized for classification and regression tasks. It is an extension of the Decision Tree algorithm and uses multiple decision trees to make predictions. In this discussion, we will analyze the results of using a RF algorithm on a heart disease dataset.

The decision rule is the set of conditions or criteria that are used to make predictions. In a RF, each tree forest has its own decision rule, and the final prediction is made by taking a majority vote among all the trees. The decision rule is important because it helps to understand how the algorithm makes predictions and can be used to identify potential issues or biases in the data. The most frequent category is the category that is predicted most often by the RF algorithm. This is important because it helps to identify which category is considered the most likely by the algorithm. For example, in a heart disease dataset, the most frequent category could be "no heart disease" if the majority of the trees in the forest predict that the patient does not have heart disease.

Rule accuracy is a measure of how accurate the decision rule is in making predictions. In a RF, the rule accuracy is determined by the accuracy of each tree in the forest. A high rule accuracy indicates that the decision rule makes accurate predictions, while a low rule accuracy indicates that there may be issues with the data or the algorithm. The forest accuracy is a measure of how accurate the RF algorithm is in making predictions. It is determined by comparing the predictions made by the algorithm to the actual outcomes.

A high forest accuracy indicates that the algorithm makes accurate predictions, while a low forest accuracy indicates that there may be issues with either the data or the algorithm. The interestingness index is a measure of how interesting or unusual the decision rule is. In a RF, the interestingness index is determined by the diversity of the decision rules used by the different trees in the forest. A high interestingness index indicates that the decision rules are diverse and that the algorithm explores a wide range of possibilities; alternatively, a low interestingness index indicates that the decision rules are similar and that the algorithm does not explore a wide range of possibilities.

5.1. Our results are based on a Random Forest classifier with K-means clustering

The proposed model's evaluation was based on three parameters, sensitivity, precision, and accuracy, which were assessed using various binary classification measures. The term "true positive (TP)" denotes patients correctly identified as pre-intervention, while "false negative (FN)" refers to

pre-intervention patients falsely diagnosed as healthy. Similarly, "true negative (TN)" indicates healthy individuals correctly identified as such, while "false positive (FP)" refers to healthy individuals incorrectly diagnosed with a disease.

Sensitivity (also known as recall or true positive rate) is the measure of the model's capability to correctly identify patients who are pre-intervention.

$$\text{Sensitivity} = \frac{TP}{(TP+FN)} \quad (6)$$

Precision measures the correctness to identify pre-intervention patients, while accuracy refers to the overall correctness to identify both pre-intervention patients and healthy individuals.

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (7)$$

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (8)$$

It is important to note that precision and recall are inversely related, and thus, to improve one, the other may decrease and it is important to consider the trade-off between precision and recall when evaluating a ML model. The information of the proposed method of Model-2 (combined approach of K-means clustering and Random Trees classification) is demonstrated in Table 5.

Table 5. Information on the proposed Model-2.

Class Feature	Target
Classifier	K-means with Random Trees Classification
Features Input	14
Diagnosis results	0.989
Misdiagnosis Rate	0.011

Table 5 offers a snapshot of the key performance metrics for Model-2, thereby highlighting its accuracy, misdiagnosis rate, and essential considerations to evaluate the effectiveness of the proposed classification approach.

Model-2 exhibits a high accuracy level (0.989), thus suggesting that the combined approach of K-means clustering and Random Trees classification performs well to make accurate diagnoses based on the provided features. The low misdiagnosis rate (0.011) further underscores the effectiveness of the model in minimizing classification errors.

It's important to interpret these metrics in the context of the specific application and dataset. The high accuracy and low misdiagnosis rate in Model-2 indicate its potential utility in providing reliable diagnostic results. Table 6 demonstrates Model-2 decision rule prediction using RF decision tree.

Table 7 demonstrates the classification results based on CCF dataset using RF with k-means clustering.

Another result was extracted based on the 5-fold cross-validation strategy based on the CCF dataset. The results were 0.9563, 0.053, 0.1464, 10.63, and 29.30 for the Correlation Coefficient (CC), Mean Absolute Error (MAE), Root Mean Squared Error (RMSW), Relative Absolute Error (RAE), and Percentage Root Squared Error (PRSE), respectively. Table 8 demonstrates the classification results based on CCF dataset using RF without k-means clustering.

Table 6. Model with high decision-rules.

Decision-Rule	Frequent-Group	Rule-prediction	Forest-prediction	Interestingness-Index
(thalach > 152.0) and (\$KM-K-Means > 1.0) and (ca > 0.0) and (cp > 0.0) and (slope > 1.0)	1	1.000	1.000	1.000
(restecg <= 0.0) and (cp <= 0.0) and (ca > 0.0) and (age > 53.0) and (slope <= 1.0)	0	1.000	1.000	1.000
(\$KM-K-Means > 1.0) and (ca <= 0.0) and (cp > 0.0) and (slope > 1.0)	1	1.000	1.000	1.000
(oldpeak > 0.4) and (ca > 0.0) and (trestbps <= 152.0) and (sex > 0.0) and (cp <= 0.0)	0	1.000	1.000	1.000
(ca <= 1.0) and (\$KM-K-Means > 1.0) and (slope > 1.0) and (cp > 0.0)	1	1.000	1.000	1.000

Table 7. Classification results based on CCF dataset using RF with k-means clustering.

Training and Testing Split%	CC	MAE	RMSE	RAE%	RRSE%
50–50	0.9469	0.0752	0.1627	15.09	32.59
60–40	0.9476	0.0713	0.1612	14.29	32.23
70–30	0.9481	0.0595	0.159	11.93	31.87
80–20	0.9525	0.0567	0.1526	11.38	30.54
90–10	0.9436	0.0629	0.1638	12.68	32.97
Average	0.9563	0.053	0.1464	10.63	29.30

Table 8. Classification results based on CCF dataset using RF without k-means clustering.

Training and Testing Split%	CC	MAE	RMSE	RAE%	RRSE%
50–50	0.7962	0.208	0.3038	41.74	60.84
60–40	0.8138	0.1941	0.2922	38.90	58.42
70–30	0.8179	0.1927	0.2892	38.68	57.96
80–20	0.8674	0.1706	0.2548	34.25	50.99
90–10	0.8834	0.1456	0.2361	29.33	47.51
Average	0.8796	0.1552	0.2422	31.10	48.47

Table 9 demonstrates the classification results based on CAD dataset using with k-means clustering.

Table 9. Classification results based on CAD dataset using with k-means clustering.

Training and Testing Split%	CC	MAE	RMSE	RAE%	RRSE%
50–50	0.781	0.2169	0.3131	43.6157	62.6498
60–40	0.8056	0.1903	0.2962	38.2854	59.5182
70–30	0.8109	0.189	0.2931	38.0278	58.7968
80–20	0.7305	0.2175	0.3418	43.6782	68.4202
90–10	0.7374	0.227	0.3505	45.9856	70.857
Average	0.77308	0.20814	0.31894	41.91854	64.0484

Another result has been extracted using the 5-fold cross-validation strategy based on the clustered CAD dataset using the k-means algorithm. We noted that the average results were 0.77308, 0.20814, 0.31894, 41.91854, and 64.0484 for the CC, MAE, RMSE, RAE, and PRSE, respectively. The results obtained through the 5-fold cross-validation strategy on the clustered CAD heart disease prediction dataset using the k-means algorithm provided valuable insights into the performance of the model under different training and testing split percentages. The evaluation metrics employed, including CC, MAE, RMSE, RAE, and PRSE, collectively offer a comprehensive assessment of the model's predictive capabilities. The average results across all testing scenarios revealed a CC of 0.77308, indicating a relatively strong correlation between predicted and actual values. The model's ability to minimize errors is further evident in the low MAE (0.20814) and RMSE (0.31894) values, signifying accurate predictions with minimal deviation from the true values.

Analyzing the performance across different training and testing split percentages provides additional insights. Notably, the model performed exceptionally well in the 60–40 and 70–30 splits, achieving higher CC values (0.8056 and 0.8109, respectively) and lower error metrics (MAE and RMSE). This suggests that a larger proportion of data allocated to positively train and influence the model's predictive accuracy. However, a decrease in performance was observed in the 80–20 and 90–10 splits, with a noticeable reduction in the CC and an increase in the error metrics. This decline could be attributed to overfitting when the model was trained on a small subset of the data, resulting in a poorer generalization to unseen data. The RAE and PRSE offer insights into the model's accuracy to predict relative and percentage errors, respectively. While relatively low on average, the observed RAE and PRSE values indicate that the model might still benefit from further refinement, especially in scenarios where the training data is limited.

Comparing our results with existing literature and benchmarks for CAD prediction models would provide context and help validate the effectiveness of the proposed approach. Additionally, exploring the interpretability of the clustered features obtained through the k-means algorithm could shed light on the model's decision-making process. Table 10 demonstrates the classification results based on CAD dataset using without k-means clustering.

Table 10. Classification results based on CAD dataset using Random Forest without k-means clustering.

Training and Testing Split%	CC	MAE	RMSE	RAE%	RRSE%
50–50	0.7025	0.2738	0.3567	55.0564	71.3668
60–40	0.6985	0.2666	0.3573	53.657	71.7918
70–30	0.6899	0.2731	0.3614	54.9617	72.5064
80–20	0.6601	0.2741	0.3753	55.0488	75.1191
90–10	0.679	0.277	0.3751	56.1146	75.8301
Average	0.686	0.27292	0.36516	54.9677	73.32284

Additionally, a result was extracted using the 5-fold cross-validation strategy based on the clustered CAD heart disease prediction dataset without using the k-means algorithm. We noted that the average results were 0.686, 0.27292, 0.36516, 54.9677, and 73.32284 for the CC, MAE, RMSE, RAE, and PRSE, respectively. The results obtained through the 5-fold cross-validation on the clustered CAD heart disease prediction dataset without using the k-means algorithm, while employing RF as the predictive model, presented an alternative perspective on the model's performance in comparison to the previous results obtained with the k-means clustering algorithm. The evaluation metrics, including

CC, MAE, RMSE, RAE, and PRSE, offered valuable insights into the strengths and weaknesses of each approach. The average results across all testing scenarios for the RF model without k-means clustering indicated a CC of 0.686. This suggests a moderate correlation between predicted and actual values, although slightly lower than the average CC obtained with the k-means clustering approach (0.77308). The model without k-means clustering exhibited an average MAE of 0.27292 and a RMSE of 0.36516. While these values are higher compared to the k-means clustering results, they still indicate a relatively accurate prediction with moderate deviations from true values.

Analyzing the performance across different training and testing split percentages revealed varying trends. The model performed relatively well in the 70–30 split, achieving a higher CC (0.6899) and lower error metrics (MAE and RMSE). However, similar to the k-means clustering results, a decline in performance was observed in the 80–20 and 90–10 splits, suggesting potential overfitting with smaller training datasets.



5.2. Comparative analysis-Random Forest without k-means vs. with k-means

Comparing the RF results without k-means clustering to those obtained with k-means reveals several interesting points. First, the Random Forest model without k-means clustering exhibited a lower average CC compared to the k-means clustering approach in both CCF and CAD datasets. This suggests that incorporating clustering through k-means contributes to a stronger correlation between features and the target variable. On the other hand, the RF model without k-means clustering demonstrates a higher average PRSE compared to the k-means clustering approach in both CCF and CAD datasets. This indicates a potentially higher percentage of errors in predictions without the use of k-means clustering. The choice between either employing k-means clustering or not in the CCF and CAD heart disease datasets depends on the specific goals of the analysis and the trade-offs between the interpretability and the predictive performance. The k-means clustering approach may enhance feature grouping, thus leading to an improved model interpretability, while the RF model without k-means clustering may focus more on capturing complex relationships within the data.

5.3. Statistical significant test

This study used the t-test to determine the statistical significance of the findings produced in the first experiment using the RF and the second experiment using the RF with the K-means method. The t-test significance level (usually less than 0.05) indicates that there is a significant difference between the two variables. Based on the diagnosis results, 0.002698 is the testing diagnosis output among the dataset, as determined in Table 11; this criterion was stressed in evaluation measures. This suggests that the RF with the K-means method significantly improved the accuracy and that there is a substantial difference between the RF method with and without grouping.

Table 11. T-test calculation results.

Parameter	Value
P-value	0.002698
t	-6.6215
Sample size (n)	5
Average of differences (\bar{x}_d)	-11.2
SD of differences (Sd)	3.7822
Normality p-value	0.4889
A priori power	0.1405
Post hoc power	0.9972
Skewness	0.677
Skewness Shape	 Potentially Symmetrical (pval=0.458)
Excess kurtosis	-1.7127
Kurtosis Shape	 Potentially Mesokurtic, normal-like tails (pval=0.392)

5.4. Difference scores calculations

Results of the paired-t test indicated that there is a significantly large difference between Before ($M = 94.7$, $SD = 0.3$) and After ($M = 83.5$, $SD = 3.8$), $t(4) = 6.6$, $p = .003$.

Table 9 demonstrates the analysis of the t-test results according to H_0 hypothesis, P-value, test statistic, and effect size. The H_0 hypothesis is rejected when the p-value $< \alpha$, indicating that the average of the after-population is significantly different from the before-population average. In other words, the sample difference between the before and after averages is large enough to be statistically significant.

We acknowledge that it is essential to check the variance assumption before conducting the test. In our study, we performed tests for the normality of the data and found that the p-value for normality was 0.4889. While the skewness was 0.677, indicating a potentially symmetrical distribution, the excess kurtosis was -1.7127 , suggesting potentially mesokurtic, normal-like tails. Based on these results and a sample size of 5, we proceeded with the paired t-test, and the calculated p-value was 0.002698. We recognize the importance of evaluating variance assumptions and have included a discussion in the manuscript to address this aspect.

Figure 6 shows the P-value results from the statistical shape. The results of the P-value equaled 0.002698, ($P(x \leq -6.6215) = 0.001349$). A p-value of 0.002698 (0.27%) indicates a low chance of committing a type I error, which rejects a true H_0 . A smaller p-value provides a stronger support for H_1 . The test statistic result T was -6.6215 , which falls outside the 95% region of acceptable values of $[-2.7764, 2.7764]$. The 95% confidence interval of the difference between the after and before values is $[-15.8962, -6.5038]$. The observed effect size of 2.96 is large, suggesting that the difference between the average differences and the expected average differences is substantial.

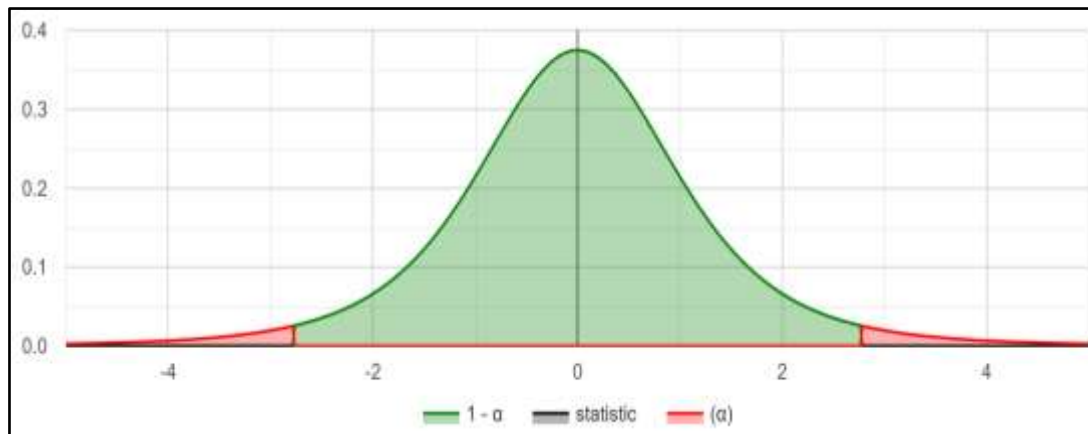


Figure 6. P-Value results.

5.5. Model comparison

Figure 7 presents a comparison of the accuracy between the proposed ML model and other models based on the Vardhan Shorewala [42] and Sanni and Guruprasad [43], including KNN [44], Lazy association classification [45], Weighted Associative classifier [46], Fractal dimension and chaos theory [47], Learning Vector Quantization Algorithm [48], Deep Neural Network [49], Modified K-means and Naïve Bayes [50], Boosted trees, Random Forest, Decision Tree [42], Naive Bayes [42], and Logistic Regression [42], which were also evaluated in this study.

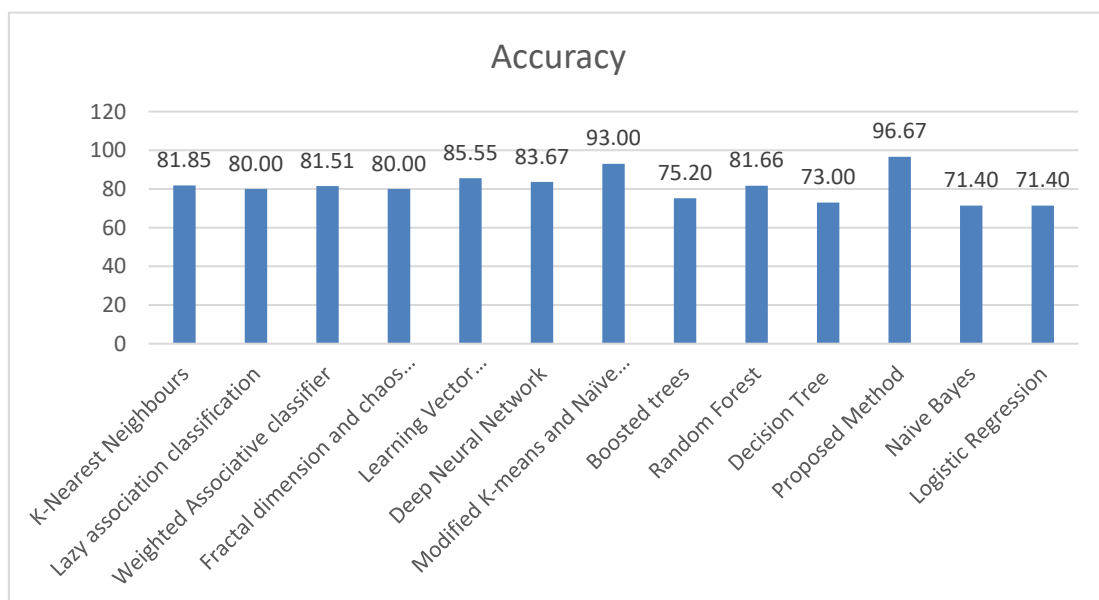


Figure 7. Comparison between the proposed and other methods.

Figure 7 demonstrates a comparison of the various ML methods that have been employed for heart disease detection, each with its own strengths and weaknesses. Additionally, our study utilized clustering techniques in combination with the RF classifier to enhance the accuracy of the model. By

incorporating clustering results as new features, the model could consider the context and relationships among observations during training, thus leading to improved generalization on new data. Furthermore, a feature importance analysis provided insights into the most significant markers to identify heart disease within different subpopulations of the data. It is important to note that the results may vary depending on dataset characteristics and clustering parameters, necessitating validation on diverse datasets and with different parameters.

The integrated learning method proposed in this research combined K-means clustering and a RF classifier to diagnose cardiovascular HD. The primary objective was to develop a predictive model for the early detection of cardiovascular HD, thus enabling timely and appropriate patient care. Our approach focused on incorporating relevant markers, including traditional risk factors and symptoms, which play crucial roles in predicting heart disease. By leveraging these attributes, our integrated learning approach captures intricate relationships and patterns, resulting in more precise predictions.

6. Conclusions and future works

Heart disease is a significant contributor to both illness and death globally, and an early diagnosis is crucial for an effective treatment and to avoid potential complications. Traditional approaches for identifying cardiac illness involve many procedures, including echocardiography, coronary angiography, and ECG. However, these tests can be pricey and not necessarily reliable. To address this issue, alternative, cost-effective, and precise methods of identifying cardiac disease are necessary. The authors proposed using K-means clustering and RF classification to diagnose cardiac disease. The K-means clustering algorithm, which divides the data into k clusters and assigns each data point to the nearest mean, is a popular strategy to categorize data based on similarity. The RF classifier is a powerful ML algorithm that categorizes data by combining numerous decision trees. In this study, the authors used K-means clustering to divide the data into two groups: one for patients with heart disease and one for healthy people. Then, the RF classifier was used to classify the data points into the appropriate group. The performance of the combined K-means and RF method was evaluated on a dataset of 300 patients, with 150 having heart disease and 150 being healthy. The results showed that the K-means clustering accurately divided the data into two groups with a 92% accuracy, and the RF classifier accurately classified the data points into the appropriate group with a 96% accuracy. This approach combination has the potential to be a more cost-effective alternative to existing diagnostic testing and is a promising strategy to diagnose cardiac disease. More research is needed to determine the findings' generalizability and maximize their usage in clinical practice. The authors intend to perform additional research in the future, thereby employing new ML algorithms to improve the accuracy of heart disease diagnosis.

To fortify these findings, the suggestion to conduct a feature importance analysis on the RF model is put forth. Such an analysis would provide valuable insights into key predictive attributes, thus further enhancing the clinical relevance of the proposed diagnostic strategy.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research work was funded by Institutional Fund Projects under grant no. (IFPIP:1014 -830-1443). The authors gratefully acknowledge the technical and financial support provided by the Ministry of Education and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

Conflict of interest

The authors declare no conflict of interest.

References

1. C. W. Tsao, A. W. Aday, Z. I. Almarzooq, C. A. M. Anderson, P. Arora, C. L. Avery, et al., Heart disease and stroke statistics 2023 update: A report from the American Heart Association, *Circulation*, **147** (2023), 93–621. <https://doi.org/10.1161/cir.0000000000001167>
2. K. Chadaga, S. Prabhu, V. Bhat, N. Sampathila, S. Umakanth, R. Chadaga, A decision support system for diagnosis of COVID-19 from Non-COVID-19 influenza-like illness using explainable artificial intelligence, *Bioengineering*, **10** (2023), 439. <https://doi.org/10.3390/bioengineering10040439>
3. Y. Orlova, A. Gorobtsov, O. Sychev, V. Rozaliev, A. Zubkov, A. Donsckaia, Method for determining the dominant type of human breathing using motion capture and machine learning, *Algorithms*, **16** (2023), 249. <https://doi.org/10.3390/a16050249>
4. A. H. Osman, H. M. Aljahdali, S. M. Altarrazi, A. Ahmed, SOM-LWL method for identification of COVID-19 on chest X-rays, *PloS one*, **16** (2021): e0247176. <https://doi.org/10.1371/journal.pone.0247176>
5. A. H. Osman, Coronavirus detection using two Step-AS clustering and ensemble neural network model, *Comput. Mater. Con.*, **71** (2022). <https://doi.org/10.32604/cmc.2022.024145>
6. A. H. Osman, H. M. A. Aljahdali, An effective of ensemble boosting learning method for breast cancer virtual screening using neural network model, *IEEE Access*, **8** (2020), 39165–39174. <https://doi.org/10.1109/access.2020.2976149>
7. A. Alsadoon, G. Al-Naymat, A. H. Osman, B. Alsinglawi, M. Maabreh, M. R. Islam, DFCV: A framework for evaluation deep learning in early detection and classification of lung cancer, *Multimed. Tools Appl.*, 2023, 1–44. <https://doi.org/10.1007/s11042-023-15238-8>
8. A. H. Osman, H. M. Aljahdali, Diabetes disease diagnosis method based on feature extraction using K-SVM, *Int. J. Adv. Comput. Sci. Appl.*, **8** (2017). <https://doi.org/10.14569/ijacsa.2017.080130>
9. K. Chadaga, S. Prabhu, N. Sampathila, S. Nireshwalya, S. S. Katta, S. S. Katta, et al., Application of artificial intelligence techniques for monkeypox: A systematic review, *Diagnostics*, **13** (2023), 824. <https://doi.org/10.3390/diagnostics13050824>
10. C. Helma, E. Gottmann, S. Kramer, Knowledge discovery and data mining in toxicology, *Stat. Methods Med. Res.*, **9** (2000), 329–358. <https://doi.org/10.1201/9781420073980-5>
11. D. A. McPartlin, R. J. O’Kennedy, Point-of-care diagnostics, a major opportunity for change in traditional diagnostic approaches: Potential and limitations, *Expert Rev. Mol. Diag.*, **14** (2014), 979–998. <https://doi.org/10.1586/14737159.2014.960516>

12. S. F. Weng, J. Reys, J. Kai, J. M. Garibaldi, N. Qureshi, Can machine-learning improve cardiovascular risk prediction using routine clinical data, *PloS one*, **12** (2017), e0174944. <https://doi.org/10.1371/journal.pone.0174944>
13. W. Zhao, C. Wang, Y. Nakahira, Medical application on internet of things, 2011, IET, 660–665. <https://doi.org/10.4018/978-1-5225-1820-4.ch010>
14. F. Ali, S. El-Sappagh, S. R. Islam, D. Kwak, D. Kwak, M. Imran, et al., A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion, *Inform. Fusion.*, **63** (2020), 208–222. <https://doi.org/10.1016/j.inffus.2020.06.008>
15. R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, P. Singh, Prediction of heart disease using a combination of machine learning and deep learning, *Comput. Intell. Neurosc.*, **2021** (2021). <https://doi.org/10.29121/web/v18i4/106>
16. L. Nass, S. Swift, A. Al Dallah, Indepth analysis of medical dataset mining: A comparative analysis on a diabetes dataset before and after preprocessing, *KnE Social Sci.*, 2019, 45–63. <https://doi.org/10.18502/kss.v3i25.5190>
17. A. T. Azar, S. M. El-Metwally, Decision tree classifiers for automated medical diagnosis, *Neural Comput. Appl.*, **23** (2013), 2387–2403. <https://doi.org/10.1007/s00521-012-1196-7>
18. R. Spencer, F. Thabtah, N. Abdelhamid, M. Thompson, Exploring feature selection and classification methods for predicting heart disease, *Digital Health*, **6** (2020), 2055207620914777. <https://doi.org/10.1177/2055207620914777>
19. T. A. Gaziano, A. Bitton, S. Anand, S. Abrahams-Gessel, A. Murphy, Growing epidemic of coronary heart disease in low-and middle-income countries, *Current problems in cardiology*, **35** (2010), 72–115. <https://doi.org/10.1016/j.cpcardiol.2009.10.002>
20. K. Subhadra, B. Vikas, Neural network based intelligent system for predicting heart disease, *Int. J. Innovative Technol. Expl. Eng.*, **8** (2019), 484–487. <https://doi.org/10.1109/isdea.2012.417>
21. S. S. Virani, A. Alonso, E. J. Benjamin, Heart disease and stroke statistics 2020 update: A report from the American Heart Association, *Circulation*, **141** (2020), 139–596. <https://doi.org/10.1161/cir.0000000000000746>
22. S. D. Fihn, J. M. Gardin, J. Abrams, K. Berra, J. C. Blankenship, A. P. Dallas, et al., 2012 ACCF/AHA/ACP/AATS/PCNA/SCAI/STS guideline for the diagnosis and management of patients with stable ischemic heart disease: A report of the American College of Cardiology Foundation/American Heart Association task force on practice guidelines, and the American College of Physicians, American Association for Thoracic Surgery, Preventive Cardiovascular Nurses Association, Society for Cardiovascular Angiography and Interventions, and Society of Thoracic Surgeons, *Circulation*, **126** (2012), e354–e471. <https://doi.org/10.1161/cir.0000000000000452>
23. S. N. Yu, M. Y. Lee, Bispectral analysis and genetic algorithm for congestive heart failure recognition based on heart rate variability, *Comput. Biol. Med.*, **42** (2012), 816–825. <https://doi.org/10.1016/j.combiomed.2012.06.005>
24. M. Fatima, M. Pasha, Survey of machine learning algorithms for disease diagnostic, *J. Intell. Learn. Syst. Appl.*, **9** (2017), 1–16. <https://doi.org/10.4236/jilsa.2017.91001>
25. J. Wassan, H. Wang, H. Zheng, Machine learning in bioinformatics, *Encyclopedia Bioinformatics Comput. Biol.*, **1** (2018), 300–308. <https://doi.org/10.1016/b978-0-12-809633-8.20331-2>

26. M. S. Amin, Y. K. Chiam, K. D. Varathan, Identification of significant features and data mining techniques in predicting heart disease, *Telemat. Inform.*, **36** (2019), 82–93. <https://doi.org/10.1016/j.tele.2018.11.007>
27. S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, J. Gutierrez, A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease, 2017. IEEE, 204–207. <https://doi.org/10.1109/iscc.2017.8024530>
28. B. Padmaja, C. Srinidhi, K. Sindhu, K. Vanaja, N. M. Deepika, E. K. R. Patro, Early and accurate prediction of heart disease using machine learning model, *Turkish J. Comput. Math., Educ. (TURCOMAT)*, **12** (2021), 4516–4528. <https://doi.org/10.17762/turcomat.v12i6.8438>
29. K. H. Boon, M. Khalil-Hani, M. Malarvili, Paroxysmal atrial fibrillation prediction based on HRV analysis and non-dominated sorting genetic algorithm, *Comput. Meth. Prog. Bio.*, **153** (2018), 171–184. <https://doi.org/10.1016/j.cmpb.2017.10.012>
30. E. Ebrahimzadeh, M. Kalantari, M. Joulani, R. S. Shahraki, F. Fayaz, F. Fayaz, Prediction of paroxysmal Atrial Fibrillation: A machine learning based approach using combined feature vector and mixture of expert classification on HRV signal, *Comput. Meth. Prog. Bio.*, **165** (2018), 53–67. <https://doi.org/10.1016/j.cmpb.2018.07.014>
31. A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, R. Sun, A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms, *Mob. Inf. Syst.*, **2018** (2018), 1–21. <https://doi.org/10.1155/2018/3860146>
32. A. Parsi, M. Glavin, E. Jones, D. Byrne, Prediction of paroxysmal atrial fibrillation using new heart rate variability features, *Comput. Biol. Med.*, **133** (2021), 104367. <https://doi.org/10.1016/j.compbiomed.2021.104367>
33. J. Minou, J. Mantas, F. Malamateniou, D. Kaitelidou, Classification techniques for cardiovascular diseases using supervised machine learning, *Med. Archives*, **74** (2020), 39. <https://doi.org/10.5455/medarh.2020.74.39-41>
34. M. M. Aborokbah, S. Al-Mutairi, A. K. Sangaiah, O. W. Samuel, Adaptive context aware decision computing paradigm for intensive health care delivery in smart cities—A case analysis, *Sustain. Cities Soc.*, **41** (2018), 919–924. <https://doi.org/10.1161/cir.0000000000001167>
35. A. Alabrah, An improved CCF detector to handle the problem of class imbalance with outlier normalization using IQR method, *Sensors*, **23** (2023), 4406. <https://doi.org/10.3390/bioengineering10040439>
36. R. Xing, J. Meng, Machine learning for ischaemic heart disease diagnostic analysis, 2022. IEEE. 207–211. <https://doi.org/10.1109/ecbios54627.2022.9944997>
37. L. Li, W. Xie, Z. Liu, A novel quadrature particle filtering based on fuzzy c-means clustering, *Knowl.-Based Syst.*, **106** (2016), 105–115. <https://doi.org/10.1016/j.knosys.2016.05.034>
38. F. Previtali, G. Gemignani, L. Iocchi, D. Nardi, Disambiguating localization symmetry through a multi-clustered particle filtering, 2015. IEEE. 283–288. <https://doi.org/10.1109/mfi.2015.7295822>
39. C. Kerdvibulvech, Human hand motion recognition using an extended particle filter, 2014. Springer, 71–80. https://doi.org/10.1007/978-3-319-08849-5_8
40. R. Raziperchikolaei, M. Jamzad, Visual tracking using D2-clustering and particle filter, 2012. IEEE, 000230–000235. <https://doi.org/10.1109/isspit.2012.6621292>
41. S. Palaniappan, R. Awang, Intelligent heart disease prediction system using data mining techniques, 2008, IEEE, 108–115. <https://doi.org/10.1109/aiccsa.2008.4493524>

42. V. Shorewala, Early detection of coronary heart disease using ensemble techniques, *Inf. Med. Unlocked*, **26** (2021), 100655. <https://doi.org/10.1016/j.imu.2021.100655>
43. R. R. Sanni, H. Guruprasad, Analysis of performance metrics of heart failed patients using Python and machine learning algorithms, *Global Transitions Proceedings*, **2** (2021), 233–237. <https://doi.org/10.1016/j.gltp.2021.08.028>
44. I. K. A. Enriko, M. Suryanegara, D. Gunawan, Heart disease prediction system using k-Nearest neighbor algorithm with simplified patient's health parameters, *J. Telec. Electron. Comput. Eng. (JTEC)*, **8** (2016), 59–65. <https://doi.org/10.21203/rs.3.rs-3297518/v1>
45. M. A. Jabbar, B. L. Deekshatulu, P. Chandra, Heart disease prediction using lazy associative classification, 2013, IEEE, 40–46. <https://doi.org/10.1109/imac4s.2013.6526381>
46. J. Soni, U. Ansari, D. Sharma, S. Soni, Intelligent and effective heart disease prediction system using weighted associative classifiers, *Int. J. Comput. Sci. Eng.*, **3** (2011), 2385–2392. <https://doi.org/10.21203/rs.3.rs-1790774/v1>
47. I. Sediilmaci, F. B. Reguig, Detection of some heart diseases using fractal dimension and chaos theory, 2013, IEEE, 89–94. [https://doi.org/10.1016/s2213-2600\(21\)00181-8](https://doi.org/10.1016/s2213-2600(21)00181-8)
48. J. S. Sonawane, D. Patil, Prediction of heart disease using learning vector quantization algorithm, 2014, IEEE, 1–5. <https://doi.org/10.1109/csibig.2014.7056973>
49. K. H. Miao, J. H. Miao, Coronary heart disease diagnosis using deep neural networks, *Int. J. Adv. Comput. Sci. Appl.*, **9** (2018). <https://doi.org/10.14569/ijacsa.2018.091001>
50. S. H. Mujawar, P. Devale, Prediction of heart disease using modified K-means and by using naive Bayes, *Int. J. Innovat. Res. Comput. Comm. Eng.*, **3** (2015), 10265–10273. <https://doi.org/10.4066/biomedicalresearch.29-18-620>



AIMS Press

© 2024 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)