A Modified K-Nearest Neighbors Algorithm for the Detection of Heart Disease

Shakila Parveen Jan[1], Muhammad Muntazir Khan[2], Jamal Uddin[3], Basharat Ahmad Hassan[2], Anees Ur Rahman[4]

[1]Department of Information Technology, Qurtuba University, Peshawar, Pakistan

[2]Institute of Computer Science and Information Technology, ICS/IT, FMCS, the University of Agriculture, Peshawar 25130, Pakistan

[3]Computer Science, Riphah School of Computing and Innovation, Riphah International University, Lahore, Pakistan

[4]Youth affair district youth office, Upper Chitral

***Correspondence:** muntazirkhan131@gmail.com shakilaparveenjan@gmail.com, muntazirkhan131@gmail.com, jamal_din@riphah.edu.pk, basharat.94@gmail.com, aneesasra.98@gmail.com

Citation | . Jan. S. P, Khan. M M, Uddin. J, Hassan. B. A, Rahman, A. U, "A Modified K-Nearest Neighbors Algorithm for the Detection of Heart Disease", IJIST, Vol. 07, Issue. 04 pp 2863-2880, November 2025

Received | October 13, 2025 **Revised |** November 10, 2025 **Accepted |** November 17, 2025 **Published |** November 24, 2025.

The leading cause of mortality worldwide is heart disease, sometimes referred to as cardiovascular disease. It is a dangerous illness that impacts the heart and blood arteries. A significant amount of research and analysis has been done recently with the goal of improving the accuracy and dependability of heart disease data. In this discipline, machine learning is crucial since it provides medical diagnostic tools that may be used to forecast illness and enhance healthcare. In this study, heart disease detection is proposed by combining KNN with Jaccard and Cosine similarities. Further, the results of Jaccard and cosine integrated KNN are compared with the results of state-of-the-art models like KNN and decision trees. Python and its libraries are used for simulation purposes. After the simulation, it was found that Jaccard-based KNN (JKNN) had the best accuracy (97%) according to the study's analysis of the Cleveland heart disease dataset. With 91% accuracy, the Cosine-based KNN (CKNN) likewise demonstrated strong performance. In a similar vein, the decision tree is inadequate for classifying heart disease because of its poor accuracy rate as 85%. Likely, KNN shows average results in the form of accuracy, as 86%. According to the results, the JKNN technique is the best model for this task, closely followed by CKNN. The use of machine learning in the diagnosis and prognosis of heart disease is affected by these discoveries.

Keywords: Heart Disease, KNN, Jaccard, Cosine, Accuracy, Confusion Matrix

Introduction:

Coronary Heart Disease (CHD) is a condition that affects the heart, the organ responsible for pumping blood throughout the body [1][2]. Abnormalities in these organs can cause constriction of the arteries that drain blood to the heart muscle, resulting in a diminished supply of oxygen and nutrients to keep the heart working properly [3]. Atherosclerosis, or the accumulation of calcium and yellow fatty deposits, causes artery narrowing [4]. The World Health Organization (WHO) reports that cardiovascular disease is a leading cause of mortality and disability worldwide [5][2]. Globally, cardiovascular diseases cause over 17.3 million deaths annually [6][3]. Including 7.3 million due to heart disease and 6.2 million due to stroke. Key risk factors include diabetes, smoking, heavy alcohol use, high cholesterol, and hypertension [7]. Intrusion detection systems (IDS) that use machine learning can detect zero-day attacks that rule-based systems could overlook, analyze enormous volumes of network traffic in real time, and recognize intricate attack patterns [8], mitigating false positives and false negatives. These systems increase detection accuracy by limiting the misclassification of normal behavior and enabling prompt identification of actual intrusions [9]. Moreover, machine learning models prove highly effective in dynamic cybersecurity environments, as they can continuously adapt and improve by training on newly emerging threat data [10].

Diagnosis is the process of determining which illness is responsible for a patient's symptoms. However, many symptoms and clinical indicators can be ambiguous, making diagnosis one of the most challenging tasks in healthcare. Accurate identification of a disease is essential for effective treatment. Machine learning is the field that can assist in anticipating illness diagnosis based on past training data [3]. There is a significant need to enhance understanding of CHD & it's complicated variables to aid prevention, early identification, and advances in therapeutic therapy [4]. Misdiagnosis of cardiac disease can result in higher mortality [4]. In response, significant research efforts have focused on developing algorithms for automated ECG analysis [4][5]. Machine learning allows systems to learn autonomously rather than through explicit programming. Early and accurate diagnosis is essential for reducing disease-related deaths. Consequently, classification algorithms are widely used in healthcare to enhance disease detection and prediction [5]. Diseases and health issues such as liver cancer, chronic renal disease, breast cancer, diabetes, and cardiac syndrome have a substantial influence on one's health and, if left untreated, can result in death. Advancements in machine learning and artificial intelligence, such as K-Nearest Neighbor, Decision Tree, Random Forest, Support Vector Machine (SVM), Naïve Bayes, and others, can address this issue [5].

K-Nearest Neighbor (KNN) is a common classification approach in data mining & statistics because it is simple to implement and has considerable classification performance [5][6]. The KNN classifier typically uses the Euclidean distance to calculate the distance between two points on a plane. It only chooses values that are close to the K [7]. The fundamental difficulty is that Euclidean distance only works with numerical data and cannot handle categorical input. As a result, in this study, a Modified K-NN (MKNN) is presented, in which similarity metrics are used instead of Euclidean distance. Similarity measures operate on categorical data and are effectively employed in several sectors [7][11]. Similarity measures such as Jaccard measure, cosine similarity, Hamming, and straightforward Matching Coefficient are straightforward and user-friendly to apply [12].

Machine learning algorithms can examine risk variables to identify people at risk. However, conventional algorithms such as K-Nearest Neighbor, Support Vector Machine, Decision Tree, Naive Bayes, and Random Forest are insufficient for predicting heart disease owing to categorical data. The key challenge is creating an accurate machine learning model that can handle categorical data and forecast heart disease risk. The study proposes combining similarity metrics with KNN to improve the accuracy of detecting cardiac illness by

introducing two techniques: Jaccard-based KNN (JKNN) and Cosine-based KNN (CKNN). The project analyzes the Cleveland heart disease dataset, and different assessment measures such as accuracy, recall, precision, and F-measure to assess the performance of proposed and current classifiers. The acquired findings are compared with traditional SVM approaches to further validate the proposed model's superior performance.

The goal of this study is to use similarity metrics and KNN to identify heart disease using categorical variables. The key objectives of this research study are as follows.

- To improve the KNN classifier by using similarity measurements such as Jaccard and cosine.
- Heart disease will be detected using similarity-based Modified K-Nearest Neighbors classifiers such as Jaccard-based KNN (JKNN) and Cosine-based KNN (CKNN).
- To compare the accuracy, recall, precision, and other properties of the Support Vector Machine classifier to those of the proposed classifiers.

The paper is organized as follows: Section 2 discusses related work, Section 3 presents a recommended strategy, Section 4 displays the findings and discussion, and Section 5 explains the conclusions and future study prospects.

Previous Work:

Heart disease remains one of the leading global health concerns. To enhance predictive accuracy, researchers employ a range of machine learning techniques such as decision trees, Naive Bayes, neural networks, K-nearest neighbors (KNN), artificial neural networks (ANN), and various clustering algorithms.

Support Vector Machine (SVM) methods have also been employed to achieve high levels of diagnostic accuracy. Recent studies examine how various machine learning algorithms can be applied to improve disease diagnosis. This section outlines several machine learning algorithms, including Naive Bayes, logistic regression, SVMs, K-means clustering, decision trees, random forests, and K-nearest neighbors, which are widely applied in disease detection and prediction. Comparisons between this method, the evaluation procedure, and the outcomes are provided. Lastly, talks regarding earlier efforts that were presented. When applying k-nearest neighbor classifiers to health datasets, the distance function is utilized by [8]. Many classification models employ the K-Nearest Neighbor classification (KNN), a predictable nonparametric classifier, as their basic classifier. The final classification output is defined by comparing the distance between each training set and the test set. KNN utilizes on Chi Square distance functions for medical data sets that include mixed, quantitative, and categorical data types. The author in [9] reported that a machine learning–based heart disease prediction system can estimate the likelihood of developing heart disease in advance. Heart disease prediction engine learning uses KNN & Tree Decision Algorithms to estimate the risk of heart disease. In addition to requiring 13 medical factors, such as age, gender, rapid blood sugar, chest discomfort, etc., the proposed system's outcome offers the chance to identify heart disease first in terms of percentage, and also shows an accuracy level higher than two algorithms. The dataset enables the prediction of heart disease risk in patients, with decision tree models attaining 81% accuracy and K-nearest neighbor models achieving 67% accuracy.

The author developed a cardiac disease prediction system using machine learning techniques. Early prediction of heart disease is essential, as its prevalence is increasing at an alarming rate. The goal of the study is to determine which patient, based on a change in a health characteristic, is more likely to suffer heart disease. The author uses several algorithms, including logistic regression and KNN, to predict and identify people with heart disease. The accuracy of the suggested model is rather high, and it could identify a person's heart disease symptoms. This prediction method for cardiac disease enhances patient care, facilitates illness diagnosis, and enables simultaneous analysis of vast amounts of data. An overview of the use

of data analytics and machine learning in the prediction of cardiac disease is presented by [13]. Machine learning approaches have been widely applied in the diagnosis of heart disease. The present review summarizes recent research on their use for predicting and estimating the risk of heart disease. This study serves as the foundation for comprehending the intricacy of the field, the instruments and strategies employed by researchers, and the degree of effectiveness attained by various contemporary approaches. Based on a survey [14] investigated heart disease prediction through machine learning techniques, examining the efficiency and effectiveness of multiple algorithms and strategies. They examine the effectiveness of different algorithms and strategies in their study. Vector support, K-Nearest Neighbors, Naive Bayes, Decision Trees, Random Forest, and other supervised learning algorithms form the basis of the models, and the collective model is highly popular among researchers. To support physicians in developing an accurate and effective prediction system. [15] Conducted a related study. In this study, various machine learning algorithms, including Naive Bayes, K-nearest neighbors, support vector machines, random forests, and decision trees, were applied to medical record parameters to forecast cardiac conditions. NB excelled utilizing the cross-validation & split-train test techniques with an accuracy of 82, 17%, 84, 28%, and 84.28%, respectively, according to several tests conducted to predict the HD using a range of UCI datasets. A comparison, performance analysis, and prediction of heart disease using supervised machine learning methods are provided [16]. According to [16], the RF approach achieved 100% sensitivity and was particularly effective when used with heart disease datasets made up of Kaggle's three classification bases based on neighboring K-nearest neighbors (KNN), decision trees (DT), and Random Forests (RF) algorithms. Thus, it is discovered that a rather basic machine learning system is being monitored. A stronger KNN classifier is created by [17] for high-dimensional and combined data. Objects labelled with a high mix of continuous and categorical variables are categorized using the KNN classifier. [17] Estimate the methodology, compare the findings with a simple approach, and employ five mixed datasets from the UCI learning repository. The results of the experiment demonstrate that the suggested method performs better in classification for large-scale mixed data.

The impact of distance functions on K-nearest neighbor classification for medical datasets was studied by [18]. To assess how these functions affect KNN performance in diverse medical applications. The objective of the study is to examine whether different distance functions affect the performance of K-nearest neighbor (KNN) classification across various medical datasets. The four dissimilar distance functions, Euclidean, Cosine, Chi-Square, and Minkowski, are employed separately to categorize KNN in these tests, which are based on three different types of medical datasets that comprise categorical, numeric, and mixed data types. An advanced study on heart disease prediction, utilizing data mining classification techniques, was performed by [19]. A sophisticated data mining technique is used to extract information from databases for medical research, particularly in the prediction of heart disease. They examined a variety of characteristics that are used to predict cardiac disease. Our neural network, decision trees, & naive Bayes are 100%, 99.62%, and 90.74% accurate, respectively. A comparison of computational intelligence methods for predicting coronary artery heart disease is presented by [20]. Logistics regression, support vector machines, neural networks, decision trees, naïve Bayes, random forest, and K-store studies are the seven computational intelligence methods that are used and comparatively illustrated. The Cleveland Heart Data, which are extracted from the UCI Repository database with a variety of calculating techniques, are used to calculate the performance of each approach. It is possible that internal neural networks achieved the best accuracy of 98.15% with sensitivity and precision of 98.67% and 98.01%, respectively. A modified KNN algorithm-based system for predicting students' academic achievement is introduced by [21]. To enable K-nearest neighbor (KNN) classification with mixed-type data, particularly nominal attributes, the authors incorporate

two similarity measures. By considering alternative voting criteria, the proposed approach also increases KNN's sensitivity to outliers. A key contribution of this study is the design of a distance function that allows classification decisions without converting nominal variables into numerical values. Experiments were conducted on an education dataset, and the results were evaluated using the KNN algorithm after one-hot encoding the nominal attributes to validate the proposed classification approach.

According to the literature analysis, there has been past research on processed categorization challenges in the healthcare sector, and each of these studies was beneficial in its own way. The research focused on numerous illnesses and used a variety of algorithms and strategies to increase classification accuracy. Despite the various methodologies, the accuracy of these automated diagnosis systems has been determined to be rather high. Overall, the literature analysis reveals that automated diagnostic systems have the potential to increase the generalization, efficiency, and accuracy of healthcare, but they should be considered as a tool to supplement and assist human knowledge rather than a replacement for it.

Material and Methods:

This section explains the materials and techniques used in the proposed research project, such as the recommended models, data gathering procedures, and preprocessing activities, including data exploration, normalization, and correlation analysis. A comprehensive description of the proposed algorithm is presented in this section, along with the performance assessment metrics utilized to evaluate the study's outcomes. Figure 1 depicts the step-by-step research flowchart for this study.

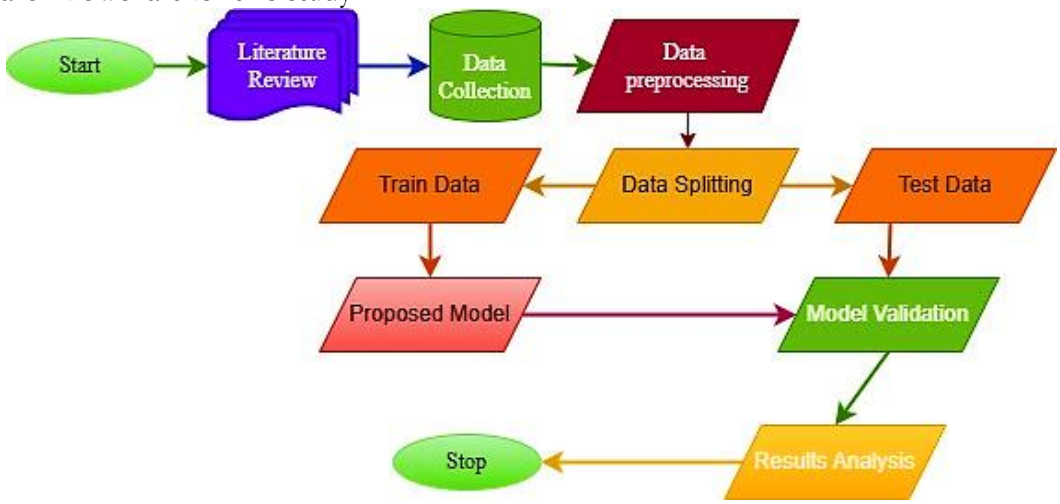

Figure 1. Step-by-step research flow diagram

Data Collection:

Data collection is a very important exercise, which directly determines the accuracy and reliability of the predictive model. It is usually done through the collection of all patient information based on trustworthy sources, including hospitals, healthcare repositories, or publicly accessible datasets like the UCI Machine Learning Repository or Kaggle. The data obtained typically contains demographic information (age, sex), clinical values (blood pressure, cholesterol level, blood sugar, heart rate), lifestyle (smokes, exercises, etc.), and diagnostic features (ECG results, types of chest pain, exercise-induced angina, etc.) [22]. The study utilized the Cleveland Heart Disease dataset from the UCI Machine Learning Repository, comprising 303 participant records. Although the dataset contains 76 features per individual, an earlier study has shown that just 13 markers may successfully detect cardiac disease. The dataset contains both categorical and numerical variables; we chose to use just categorical data for our research and excluded the numerical information. The data set utilized in this investigation had already been developed or acquired [23]. The dataset was kindly made

available by its contributor through the UCI Repository. Table 1 presents the various attributes of the proposed dataset.

Table 1. Attribute display of the proposed dataset

Name	Type	Description
Age	Continuous	Age is converted in Categorical form as 75=060=1 50=2 40=3 30=2 etc.
Sex	Discrete	0 = female 1 = male
Cap	Discrete	Chest pain type: 1 = typical angina, 2 = atypical angina, 3 = non-angina pain 4 = symptom
Fibs	Discrete	Fasting blood sugar>120 mg/dl: 1-true 0=False
Exam Continuous Maximum heart rate achieved	Discrete	Exercise induced angina: 1 = Yes 0 = No
Slope	Discrete	The slope of the peak exercise segment; 1 = up sloping 2 = flat 3 = down sloping

Data Exploration:

Heart disease prediction data visualization is important in revealing the underlying patterns, relationships, and trends in the data. The graph discovered the distribution of such critical features as age, cholesterol level, blood pressure, and heart rate by using visual methods, including histograms, box plots, heat maps, and correlation matrices. Scatter plots and pair plots assist in determining the relationships that are non-linear, and possible outliers, whereas bar charts and pie charts can give a vivid picture of categorical variables like the type of chest pain or gender. The heat map can also be a useful tool when analyzing the correlation between features, and this can aid in choosing the most useful ones to be used when training the model. Visualization can also be used to determine the class imbalance between heart disease and non-heart disease patients, so that the preprocessing or resampling methods are selected accordingly [24]. Figure 2 display visualization of the proposed dataset as a histogram representation.

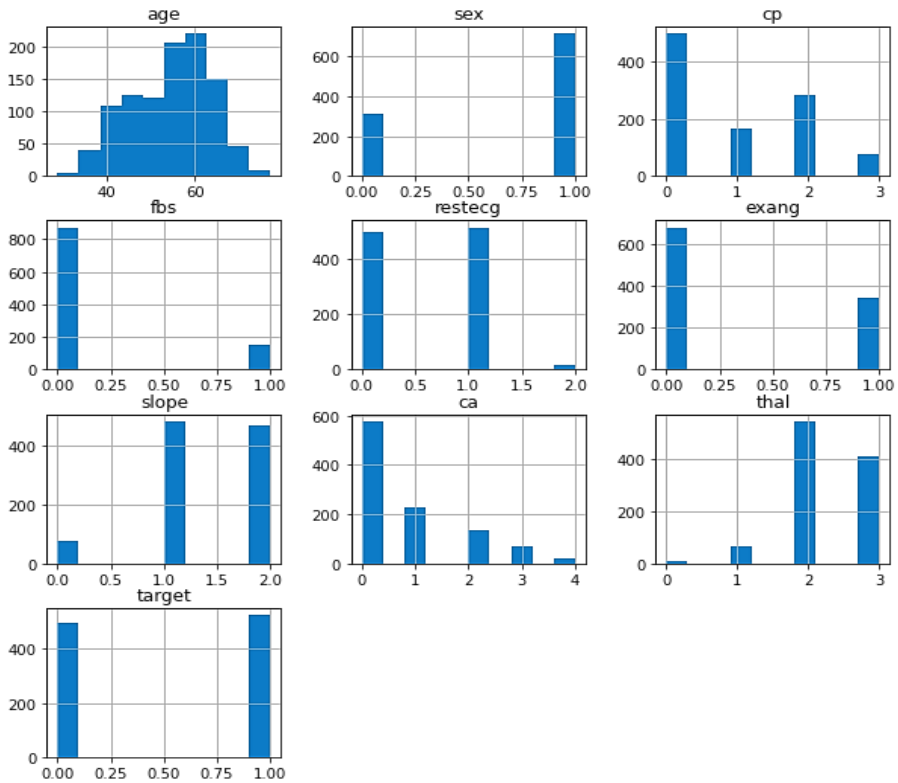

Figure 2. Histogram representation of the dataset

Figure 2 represents a combination of several histograms that show the distribution of various attributes in a heart disease data set, such as the age, sex, the type of chest pain (cp), fasting blood sugar (fbs), the results of resting electrocardiogram (restecg), exercise angina (exang), the slope, the number of major vessels (ca), thalassemia (thal), and target (presence or absence of heart disease). The subplot displays the frequency of particular values of a given feature, which is used to comprehend the dispersal, skew, and data density. To illustrate, the age distribution indicates that the majority of the patients are between 45 and 65 years of age, whereas the target variable suggests a fairly balanced dataset in diseased and non-diseased cases. This visualization can help the researchers to spot trends in the data, prevailing categories, and possible imbalance- this will make this visualization an effective tool in exploring data in the prediction of heart diseases.

Correlation Removing:

Another necessary preprocessing of heart disease prediction is correlation removal, which is required to increase the efficiency and accuracy of a model. Loosely correlated features may give redundant information, and this may cause multicollinearity and compromise the learning of specific algorithms, e.g., logistic regression or decision trees. Through calculating the correlation matrix, it is possible to determine relationships between such numerical attributes as age, cholesterol level, resting blood pressure, and maximum heart rate to identify strong interdependencies. Highly correlated variables (correlation > 0.85–0.90) are scrutinized, with redundant features removed to minimize noise and enhance generalization, ensuring that all retained features provide valuable input to the predictive model. Finally, removing highly correlated features simplifies the dataset, enhances computational efficiency, and improves both the interpretability and robustness of the model in predicting heart disease outcomes [25]. Figure 3 represents the visualization of the correlation of the proposed dataset.

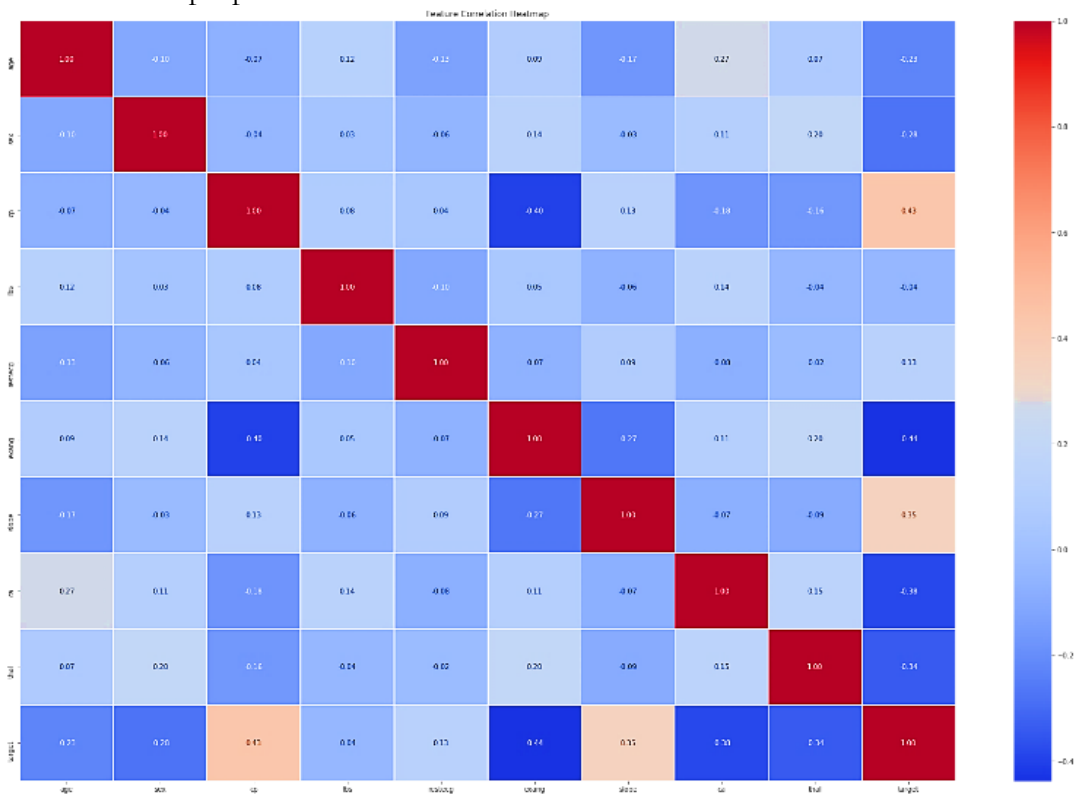

Figure 3. Correlation analysis of the proposed dataset

Figure 3 represents an image of a feature correlation heat map, which indicates the correlation among different attributes in the heart disease dataset. The correlation coefficients

between the features (age, sex, chest pain type (cp), fasting blood sugar (fbs), resting electrocardiographic results (restecg), exercise-induced angina (exang), slope, number of major vessels (ca), thalassemia (thal) and the target variable are given visually in the form of a heat map. The color gradient (with the dark blue (negative correlation) to the dark red (positive correlation)) shows how strongly and in which direction relationships between variables exist. As an example, the target variable and chest pain type (cp) have a moderately positive relationship (0.43), indicating that the type of chest pain is useful in predicting the presence of heart diseases, and exang and ca are not correlated with the target, indicating an inverse relationship. Most variables exhibit weak correlations, indicating minimal multicollinearity and supporting the selection of these features for predictive analysis. Overall, the heat map serves as a valuable visualization tool for understanding inter-feature relationships and guiding feature selection in heart disease prediction.

Important Features Selection:

Selection of features is a very essential task that focuses on the selection of the most valuable attributes that have a significant impact on the prediction. The choice of meaningful features assists in minimizing the model complexity, overfitting, and enhancing prediction accuracy. Dimensions that are frequently studied in heart disease data are age, sex, type of chest pain (cp), resting blood pressure (restbps), cholesterol level (chol), fasting blood sugar (fbs), resting electrocardiographic findings (restecg), highest heart rate (thalach), induced angina during exercise (exang), slope, number of major vessels (ca), and thalassemia (thal). Statistical methods (correlation analysis, chi-square tests, ANOVA) and machine learning techniques, including Recursive Feature Elimination and tree-based feature importance, are typically used to select the most relevant predictors. The characteristics of the chest pain type, max heart rate, the number of large vessels, thalassemia, and angina experienced after exercise have a way of becoming the most effective predictors of heart disease. Finally, a useful feature selection can increase model interpretability, decrease computation time, and also make sure that the predictive model concentrates on the most informative and clinically meaningful variables [23]. Figure 4 represents the visualization of important features and the value of their importance regarding the proposed dataset.

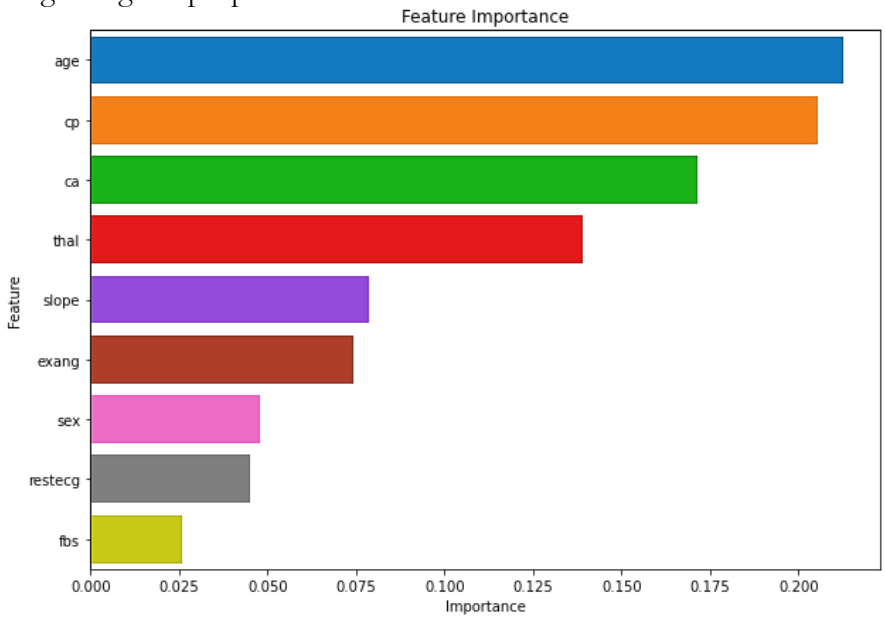

Figure 4. Visualization of important features

Figure 4 shows a feature importance bar chart showing the proportional contribution of different attributes in predicting heart disease. The features are ordered by their importance, with age being the most important predictor, and closely after it comes the type of chest pain

(cp), number of major vessels (ca), and thalassemia (thal). The features are very important in the aspect of identifying the probability of a patient having heart disease since they are directly connected to physiological and diagnostic attributes. Other characteristics include slope, exercise-induced angina (exang), sex, resting electrocardiographic outcomes (restecg), and the fasting blood sugar (fbs), with lower significance but still with an effect on the overall performance of a model. The visualization is also useful in that it shows the strongest predictive variables, which allows researchers to give priority to the most important clinical variables as well as optimize model input to achieve better models in predicting heart diseases through improved accuracy and interpretability.

Data Splitting:

To assess model performance and reduce the risk of overfitting in machine learning, data are typically divided into training and testing sets. The model is trained on the training set, and its ability to generalize to new, unseen data is evaluated using the testing set. A popular split ratio is 80:20 or 70:30; this might vary depending on the complexity of the problem and the size of the dataset. In unbalanced datasets, stratified sampling guarantees that classes are represented proportionally. Furthermore, cross-validation approaches like k-fold cross-validation improve model dependability by offering many train-test splits. The K-fold method is used in this study. The following are covered:

K-FOLD:

K-fold cross-validation is a trustworthy technique for evaluating machine learning models. It splits the dataset into k equal-sized subsets, or "folds." The model is trained on k-1folds and tested on the remaining fold, ensuring that each fold serves as a test set exactly once. This procedure is repeated ktimes. This method reduces performance estimate volatility and yields a more accurate assessment of model generalization by averaging results across all folds. Stratified k-fold, which maintains the class distribution in each fold, and leave-one-out cross-validation (LOOCV), where k is the number of samples, are common variations. K-fold cross-validation is particularly useful for small datasets since it minimizes overfitting and maximizes data utilization, which ultimately improves model selection and hyperparameter tuning.

Proposed Model:

This study is based on heart disease prediction. For said purpose, this study suggested a well-known machine learning model K-nearest neighbor model. Further, to enhance the performance of KNN, this study suggested the integration of KNN with two will-known method like Jaccard and cosine similarity. Both of the integration methods are discussed below:

Jaccard-Based KNN (J-KNN):

The combination of the Jaccard similarity measure with the K-Nearest Neighbor (KNN) algorithm is an efficient hybrid method of heart disease prediction, particularly when one has categorical or mixed-type medical data. A similarity measurement in the Jaccard method compares two sets and calculates the size of the intersection of the two sets and the size of the union between the sets. Patient records may be expressed in the form of a set of symptoms, conditions, or diagnostic properties in the medical dataset context. The Jaccard similarity is especially handy when the features are binary or nominal, e.g., presence or absence of chest pain, hypertension, or diabetes, because it does not concentrate on the distance between two features, but rather on common features. This renders it appropriate in the comparison of patient profiles on the basis of shared medical conditions.

Utilized with KNN, the Jaccard similarity works as the distance measure rather than the conventional Euclidean or Manhattan distance. In this method, the Jaccard similarity between the test record and all the training records is calculated in each instance of the test. The k nearest similar patients (neighbors) having the highest Jaccard scores are then selected

by the algorithm. The majority vote among these neighbors is then used to predict the class label, e.g., heart disease present or absent. This modification enables KNN to run effectively even when medical characteristics are categorical, sparse, or are left to represent sets of conditions, which can be problematic for distance-based measures that assume continuous values.

Compared to traditional methods, the hybrid model is more interpretable and accurately predicts heart disease by effectively identifying relationships among patient characteristics. The Jaccard-KNN technique is strong in terms of dealing with missing or irrelevant attributes because the dissimilarities are determined by the attributes that are common to the two instances. Besides, since it is based on similarity and not simple numerical variances, the model does not have the problem of scaling and offers greater insights into healthcare data analysis. Generally speaking, a combination of the Jaccard similarity and KNN is a hybrid of the set-based similarity and the instance-based learning, which provides superior classification accuracy and decision support for medical diagnosis.

Jaccard-KNN integration is a method based on the Jaccard similarity that determines the degree of similarity between two records of patients based on similar symptoms or medical conditions. The Jaccard similarity is defined to be.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

Similarly, the Jaccard distance will be defined as

$$d_J(A,B) = 1 - J(A,B) \quad (2)$$

This distance is used in place of the traditional Euclidean distance to locate the k nearest patients in K-Nearest Neighbor (KNN). In the case of predicting heart diseases, the method is effective since medical data are usually binary or categorical (e.g., the presence or absence of chest pain, diabetes, etc.). The Jaccard algorithm also only considers common positive attributes, and does not count irrelevant nos. Mathematically, since dJ is a valid metric (non-negative, symmetrical, triangle inequality) has the advantage of allowing the same distance to be measured. On the whole, Jaccard based on KNN leads to improved prediction since it gives more priority to patients with similar health profiles.

Cosine-based KNN (C-KNN):

Cosine similarity and the K-Nearest Neighbor (KNN) algorithm are effective methods to predict heart disease, particularly in the case of medical data that is in the form of numerical values, i.e., heart rate, cholesterol level, or blood pressure. Cosine similarity compares the similarity between two feature vectors; it does not compare their magnitude directly, but the angle between the two feature vectors. It emphasizes data patterns, allowing the model to recognize similar medical tendencies even when patients' numeric values differ. In this technique, the data of every patient is given as a feature-vector, and cosine similarity is applied to find the resemblance between two patients about the feature vectors.

The cosine similarity between two patients x and y is mathematically defined as.

$$\text{Cosine Similarity}(x,y) = \frac{x.y}{||x||\,||y||} = \frac{\sum_{i=1}^{n} x_i\, y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\ \sqrt{\sum_{i=1}^{n} y_i^2}} \quad (3)$$

xi and yi are the values of the features (such as blood pressure or cholesterol) of the patients x and y. The value is within the range of -1 and 1, but in the case of non-negative medical data, the value is within the range of 0 to 1. A score of close to 1 indicates that the two patients share health patterns. To apply this to KNN, we transform it to a cosine distance:

$$d_{cos}(x,y) = 1 - \text{Cosine Similarity}(x,y)$$

Only the k patients having the least cosine distance are then chosen, and the majority class of the chosen patients is used to predict the presence of heart disease with the aid of the algorithm.

Results evaluation:

After testing the model with Python 3.6, this study calculated the classification performance as well as outcomes for a specific dataset. Following that, a broad evaluation was conducted using the following standards:

Confusion Matrix:

The confusion matrix, also referred to as an error matrix or contingency table, may be used to coordinate every classification or comparison study with a variety of constraints. The number of classes that need to be constructed defines the size of the confusion matrix M (n*n). The collection (total) of all recovered positive results, including both true positives (TP) and false positives (FP), is considered the most accurate identification approach. True positives are statements indicating that a component is linked with a class that genuinely belongs to that class, whereas false positives indicate that the element is unrelated to a class that does belong to that class. The overall classification errors, defined as the sum of false positives and false negatives (FP + FN), encompass all incorrectly classified cases. Table 1 presents the confusion matrix data as supplied by [23][26][27].

Table 2. Confusion Matrix

		Predicted Class	**Predicted Class**
		Related (P)	**Related (P)**
Actual Class	**Related (P)**	True Positive (TP)	False Negative (FN)
	Not Related (N)	False Positive (FP)	True Negative (TN)

The confusion matrix is obtained by combining the following values:

True Positive (TP):

When applied in the context of the prediction of heart disease, a true positive (TP) would indicate a situation in which the predictive model is able to diagnose a patient as having heart disease when, in reality, the patient does have the illness. It is an indicator of the model's capacity to identify instances of the disease at a good level of accuracy as a positive response, which is imperative in medical diagnostics, where early detection can save lives. When the true positives are high, it means that the model is successfully identifying patients at real risk, thus helping to obtain timely medical care and treatment c.

False positive (FP)

A false positive (FP) is used in heart disease prediction when the model misclassifies a healthy person as one with heart disease. This error occurs when the algorithm incorrectly predicts the presence of a disease that is not actually present. Although false positives do not cause direct bodily harm, they may cause undue stress to the patient, require further diagnostic procedures, and incur higher healthcare expenses. False positives directly impact the accuracy measure in model measurement [26].

True Negative (TN):

A true negative (TN) in the context of making predictions about heart disease is a situation in which the model used in prediction correctly warns about the possibility of heart disease when a patient is actually healthy. True negatives are also important since they will show how the model can correctly exclude those who do not need medical intervention, and this will save unnecessary anxiety, treatment, and healthcare expenses. A large number of true negatives means that there is high reliability of the model in its ability to differentiate between sick and healthy patients. True negatives get calculated mathematically in order to determine specificity [27].

False Negative:

A false negative (FN) is used in the context of heart disease prediction to describe a situation whereby the model with predictive values misclassifies a heart disease victim as normal. This kind of error is especially critical in medical use since it implies that a person who, in fact, needs medical assistance is ignored, and there is a potential complication to late

diagnosis and severe health effects. False negatives affect the model's recall (sensitivity) directly, in that the more false negatives, the lower the recall value, meaning that false negatives are included in the model. False negative minimization is paramount to the prediction of heart diseases, as correct diagnosis through automated diagnostic systems will guarantee that patients receive timely treatment, improve their outcomes, and put more trust in the system [28].

Accuracy:

Accuracy is one of the important performance measures used in predicting heart diseases because it measures how well a predictive model can take into consideration both diseased and healthy individuals. It is the proportion of cases that are correctly classified, including both true positives (TP) and true negatives (TN), to the counts of cases that are evaluated. The large value of accuracy means that the model is valuable in differentiating between patients with heart disease and those without it. In medical data, however, where there is often an imbalance in classes (i.e., fewer positive cases than negative cases), accuracy itself may not be an adequate measure of diagnostic performance. Hence, though accuracy gives a rough estimate of the reliability of the model, it must be used in conjunction with other measures like precision, recall, and F1-score to make a complete assessment of the heart disease prediction models. Accuracy is defined mathematically by [28] as.

$$\boldsymbol{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Precision:

In heart disease prediction, the measures of precision are used to determine the percentage of patients who have been accurately predicted to have heart disease out of the total number of patients who have been predicted to have the disease. It is also an indication of how effective the model is in preventing misleading alarms, since the majority of the positive forecasts made are, in reality, positive. High precision means that few false positives occur as a result of the model, hence it is reliable in identifying the real patients who actually need medical care. This is vital in clinical practice because the inaccuracy of low precision may result in unnecessary anxiety, extra testing, and high healthcare costs for healthy people. According to [28][29], Precision is mathematically very much as:

$$\boldsymbol{Precision} = \frac{TP}{TP+FP} \quad (5)$$

Recall:

In the prediction of heart disease, the parameter of recall (sensitivity) assesses how the model identifies the actual patients with heart disease correctly. It shows the effectiveness of the model in terms of capturing the real positive cases of all the real disease cases in the dataset. In medical diagnosis, high recall is necessary since failure to identify a patient with heart disease (false negative) may result in delayed treatment and possibly important health effects. Thus, a model that has high recall will make sure that the majority of patients at risk are properly identified and sent to further medical care. Mathematically, recall is defined by [30] as.

$$\boldsymbol{Recall} = \frac{TP}{TP+FN} \quad (6)$$

F1-Score:

To obtain a single performance statistic, the F1 score, recall, and accuracy are averaged by calculating the harmonic mean of the two metrics. Since it has a balance between false negatives and false positives, it provides a more in-depth view of the performance of a classifier compared to the accuracy itself. It is particularly useful when one has an unbalanced dataset. F1 score is a powerful evaluation method as it emphasizes accuracy as well as memory, especially where it is important to strike a balance between reducing false alarms and identifying positives [26].

$$\boldsymbol{F1-Score} = \frac{2*recall*precision}{Recall+Precision} \quad (7)$$

Result and Discussion:

This section compares the performance of the proposed modified KNN (integrated with cosine similarity and Jaccard similarity measure) with state-of-the-art machine learning models like KNN and decision trees. The Cleveland heart dataset, offered online at the University of California, Irvine (UCI), including 303 records of participants, is used in the proposed study. Although the dataset contains 76 features per individual, an earlier study has shown that just 13 markers may successfully detect cardiac disease. The dataset contains both categorical and numerical variables, but only categorical data is preferred for the research study, and the numerical information is excluded.

Preliminaries:

The studies were carried out on a 2.0 GHz Intel Core i5 CPU with 8 GB of RAM. The operating system is Windows 10. The Keras Python library is used for the training and testing of the model across all datasets. To explore the suggested modified KNN method, the algorithms KNN and decision tree are compared based on accuracy, recall, precision, and f-measure.

Experimental Results:

Four models were assessed using accuracy, precision, recall, and F-measure: Jaccard-based KNN (J-KNN), Cosine-based KNN (C-KNN), K-Nearest Neighbor (KNN), and Decision Tree.

Result outcomes:

Figure 5 shows that there is a significant performance difference between the assessed machine learning classifiers on the Cleveland heart disease dataset, and the modified KNN strategies perform best compared to traditional ones. After the simulation, it was found that J-KNN achieved the accuracy of 97.88%, C-KNN achieved the accuracy of 91.01%, K-Nearest Neighbor achieved the accuracy of 86.24%, and Decision Tree achieved the performance of 85.71%. So, Jaccard-based and cosine-based versions prove to be more predictive, with stronger and better classification robustness being demonstrated when it is evident that improvement of the similarity measurement by a great deal makes classification robustness much better than the traditional KNN and decision tree models. This tendency is also supported by the results of the integration, with Jaccard-based integration always being more effective than the cosine-based one, which proves the efficiency of this approach in predicting heart diseases. In general, these performance trends can be visualized in Figure 5, and both bar and line visualizations helped to better see the improvement of the modified methods.

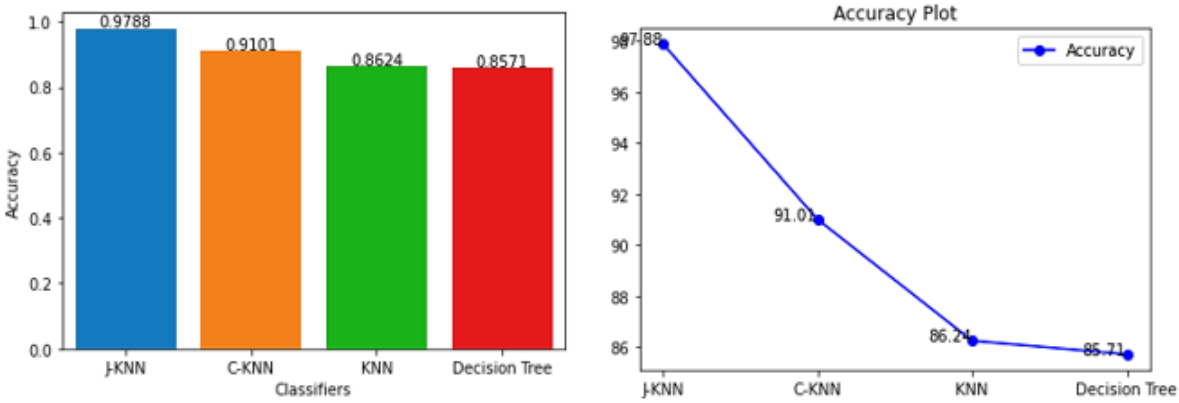

Figure 5. display of the accuracy of various models

Figure 6.Shows the overall performance of multiple machine learning models in terms of accuracy, precision, recall, and f-measures in the form of a confusion matrix after exhibiting their accuracy rates in Figure 5.

In Figure 6, the comparative perspective of how each classifier treats the differences among classes by their confusion matrices is presented, with the primary highlight on the strengths and weaknesses of the models instead of the actual numbers. The default KNN has a moderate discriminative capability, and there is a discernible degree of confusion between the two classes, implying that it struggles to separate mutable patterns within the dataset. This limitation shows that better or hybrid variants are necessary in the case of complex medical data. The decision tree, in contrast, has a less biased and more dependable segregation of the classes, and it better fits the basic form of the data, but still has the over-prediction of positive class tendency. The improved KNN versions- J-KNN and C-KNN have much greater discrimination of classes, of which J-KNN has a constant sensitivity-specificity balance, and C-KNN has specificity with a specific separation and minimal false classification. Combined, these interpretations highlight the enhanced robustness and generalization capacity of the altered KNN methods, as well as establish their appropriateness in heart disease forecasting within the proposed research context.

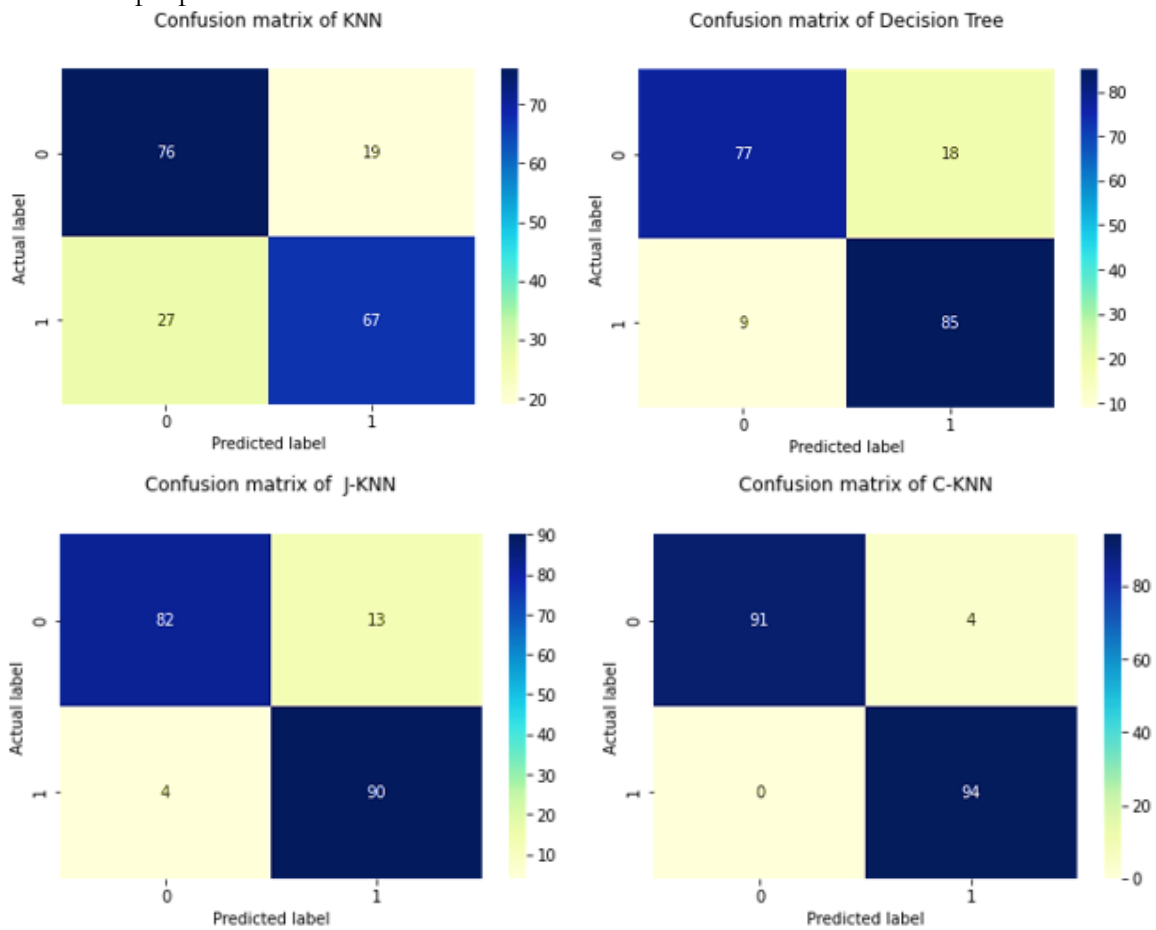

Figure 6. Visualization of the performance achieved by each model

Discussion:

The given paper is aimed at the prediction of heart disease with the help of KNN with the addition of Jaccard and cosine similarity indices, and the comparison of the results with basic KNN and decision tree classifiers in terms of the Cleveland dataset provided by the UCI repository. The comparison, which is conducted in terms of accuracy, precision, recall, F-measure, and confusion matrices, shows a distinct favor to the modified methods. The Jaccard-based KNN is always the best performer, which shows more consistent and stable classification behavior in all measures. The cosine-based form also demonstrates the competitive results, which suggests the similarity strategy refinement to enhance the ability of

the model to detect meaningful patterns in the data. Conversely, the classical KNN and decision tree frameworks have a lower predictive ability, and the decision tree provides an intermediate result, while the simple KNN fails, especially in terms of precision and classification ratio. All these trends have in common the fact that the more sophisticated the similarity measures that are integrated into the classifier, the greater the ability of the latter to make a distinction between healthy and diseased cases. These insights can be visually reinforced by Figure 7, which compares the models in all the performance measures, and Table 2 represents a brief numerical summary of the results.

Results Validation:

The performance of the four classification models is compared, and it is evident that the J-KNN has the most predictive ability, which is stable and balanced in all evaluation criteria. Its high level of accuracy, sensitivity, and overall reliability shows that the Jaccard-based similarity measure is useful in ensuring that the model will be able to effectively identify cases of heart disease and reduce the cases of misclassification. C-KNN is also competitive, especially in balancing well between precision and recall, thus indicating that cosine-based similarity at least is stable but not as effective as Jaccard. As a comparison, the conventional KNN has significantly worse generalization, and this is a result of the fact that it is poorly suited to work with complex decision boundaries within the data. The decision tree has a relatively lower level of performance compared to the normal KNN, but is still not as strong as the modified variants of KNN. On the whole, it is possible to note that J-KNN provides the most reliable results, and it is the most efficient model among the ones that are considered.

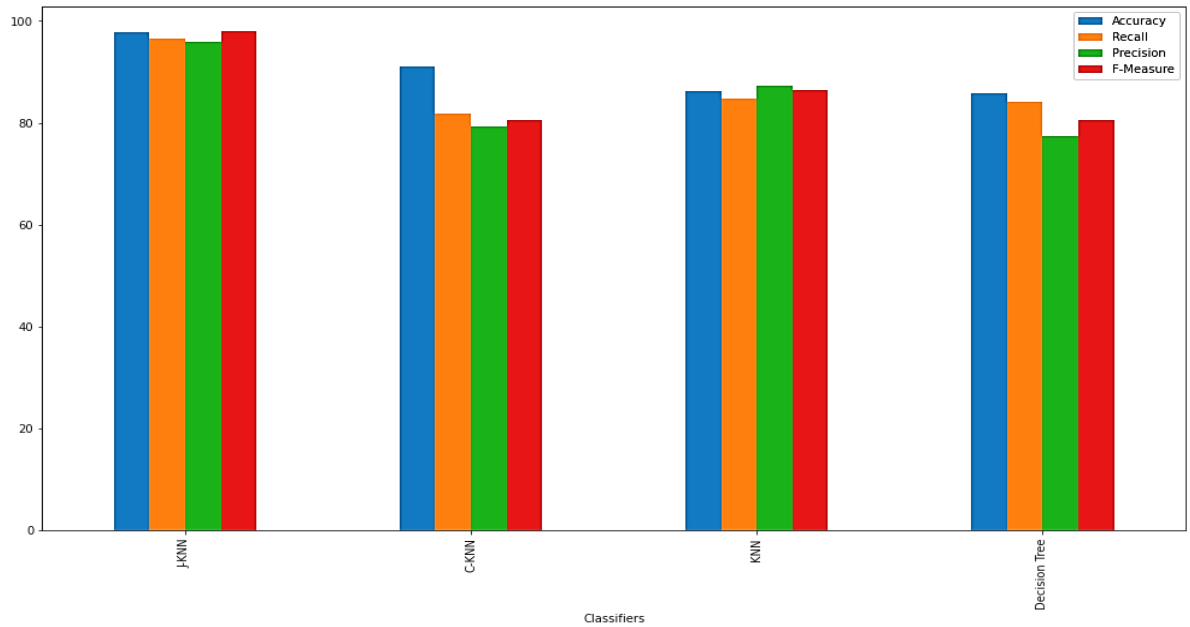

Figure 7. Performance evaluation of various models

Table 3. Tabular representation of the performance of various models for intrusion detection

Model	Accuracy %	Precision %	Recall %	F-Measures %
J-KNN	97.88	94.78	96.77	97.88
C-KNN	91.01	98.45	95.04	97.61
KNN	86.24	73.64	80.33	76.24
Decision Tree	85.71	81.47	89.66	84.61

Conclusion:

The foundation of this work is the prediction of heart disease by the combination of KNN with cosine similarity and Jaccard. Additionally, the effectiveness of these integrated

models is contrasted with decision trees and K-nearest neighbors. For testing and training, a publicly accessible dataset named the Cleveland heart disease dataset was gathered from the UCI online repository. In conclusion, this paper proves that machine learning methods can be successfully used to predict heart diseases, and the suggested J-KNN model was the most effective in terms of overall performance, as opposed to C-KNN, regular KNN, and Decision Tree classifiers. High accuracy, recall, and F-measure of J-KNN demonstrate its capacity to provide accurate and consistent predictions that make it a good contender to be utilized in clinical decision-support. The comparative analysis also demonstrates the fact that tailored or hybrid variants of KNN can substantially improve the predictive results as compared to the conventional models. These results highlight the possibilities of optimized machine learning solutions that can help healthcare providers in early diagnosis and risk with the final results in better patient management and preventive care methods.

Acknowledgement: We would like to extend our sincere gratitude to our department, FMCS, and the AUP Peshawar.

Author's Contribution: Shakila Parveen, Jan conceptualization, and primary writing, Muhammad Muntazir Khan (M. Mutazir Khan). Literature preparation. Methodology, implementation, Basharat Ahmad Hassan, funding sources, and final checking. Anees Ur Rahman's results analysis and review. Jamal Uddin validation. All authors have read and agreed to the published version of the manuscript.

Conflict of Interest: All authors have read and agreed to the published version of the manuscript in IJIST.

References:

[1] K. S. and C. J. A Ishak, A Ginting, "Clasiffication of Heart Disease using Decision Tree Algorithm," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 003, 20201, doi: 10.1088/1757-899X/1003/1/012119.

[2] R. Katarya and S. K. Meena, "Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis," *Heal. Technol. 2020 111*, vol. 11, no. 1, pp. 87–97, Nov. 2020, doi: 10.1007/S12553-020-00505-7.

[3] A. M. A. Ibrahim Mahmood Ibrahim, "The Role of Machine Learning Algorithms for Diagnosing Diseases," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 1, 2021, [Online]. Available: https://jastt.org/index.php/jasttpath/article/view/79

[4] K. Z. Xinyao LI, Linlin ZHANG, Xuehua BI, Ying ZHANG, Guanglei YU, "The Classificatied Prediction of Coronary Heart Disease Based on Patient Similarity Analysis," *Res. Sq.*, 2021, [Online]. Available: https://www.researchsquare.com/article/rs-724235/v1

[5] S. Grampurohit and C. Sagarnal, "Disease prediction using machine learning algorithms," *2020 Int. Conf. Emerg. Technol. INCET 2020*, Jun. 2020, doi: 10.1109/INCET49848.2020.9154130.

[6] X. Z. and R. W. S. Zhang, X. Li, M. Zong, "Efficient kNN Classification With Different Numbers of Nearest Neighbors," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 5, pp. 1774–1785, 2018, doi: 10.1109/TNNLS.2017.2673241.

[7] D. A. . A. and N. C. Aziz, "Implementation of K-Nearest Neighbors Algorithm for Predicting Heart Disease Using Python Flask," *Iraqi J. Sci.*, vol. 62, no. 9, 2021, doi: 10.24996/ijs.2021.62.9.33.

[8] S.-W. K. & C.-F. T. Li-Yu Hu, Min-Wei Huang, "The distance function effect on k-nearest neighbor classification for medical datasets," *Springerplus*, vol. 5, no. 1304, 2016, doi: https://doi.org/10.1186/s40064-016-2941-7.

[9] A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," *Int. Conf. Electr. Electron. Eng. ICE3 2020*, pp. 452–457, Feb. 2020, doi: 10.1109/ICE348803.2020.9122958.

[10] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN Comput. Sci. 2020 16*, vol. 1, no. 6, pp. 1–6, Oct. 2020, doi: 10.1007/S42979-020-00365-Y.

[11] M. Alobed, A. M. M. Altrad, and Z. B. A. Bakar, "A Comparative Analysis of Euclidean, Jaccard and Cosine Similarity Measure and Arabic Wordnet for Automated Arabic Essay Scoring," *Proc. - CAMP 2021 2021 5th Int. Conf. Inf. Retr. Knowl. Manag. Digit. Technol. IR 4.0 Beyond*, pp. 70–74, Jun. 2021, doi: 10.1109/CAMP51653.2021.9498119.

[12] M. Besta *et al.*, "Communication-Efficient Jaccard similarity for High-Performance Distributed Genome Comparisons," *Proc. - 2020 IEEE 34th Int. Parallel Distrib. Process. Symp. IPDPS 2020*, pp. 1122–1132, May 2020, doi: 10.1109/IPDPS47924.2020.00118.

[13] M. A. M. Marimuthu, "A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach," *Int. J. Comput. Appl.*, vol. 81, no. 18, pp. 975–8887, 20181, doi: 10.5120/ijca2018917863.

[14] A. A. A. & O. O. Micheal Olaolu Arowolo, Marion Olubunmi Adebiyi, "Optimized hybrid investigative based dimensionality reduction methods for malaria vector using KNN classifier," *J. Big Data*, vol. 8, no. 29, 2021, doi: https://doi.org/10.1186/s40537-021-00415-z.

[15] H. El Hamdaoui, S. Boujraf, N. E. H. Chaoui, and M. Maaroufi, "A Clinical support system for Prediction of Heart Disease using Machine Learning Techniques," *2020 Int. Conf. Adv. Technol. Signal Image Process. ATSIP 2020*, Sep. 2020, doi: 10.1109/ATSIP49331.2020.9231760.

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

[17] S. L. B. Pauline Rothmann-Brumm, "An Improved K Nearest Neighbor Classifier for High-Dimensional and Mixture Data," *J. Phys. Conf. Ser.*, 2021, doi: 10.1088/1742-6596/1813/1/012026.

[18] R. D. Canlas, "Data Mining in Healthcare : Current Applications and Issues By," *Comput. Sci.*, 2010.

[19] Chaitrali S., Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques," *Int. J. Comput. Appl.*, vol. 47, no. 10, 2012, [Online]. Available: https://research.ijcaonline.org/volume47/number10/pxc3880076.pdf

[20] S. I. Ayon, M. M. Islam, and M. R. Hossain, "Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques," *IETE J. Res.*, vol. 68, no. 4, pp. 2488–2507, 2022, doi: 10.1080/03772063.2020.1713916;WGROUP:STRING:PUBLICATION.

[21] V. S. Moohanad Jawthari, "Predicting students' academic performance using a modified kNN algorithm," *Pollack Period.*, 2021, doi: https://doi.org/10.1556/606.2021.00374.

[22] S. S. Yadav, S. M. Jadhav, S. Nagrale, and N. Patil, "Application of Machine Learning for the Detection of Heart Disease," *2nd Int. Conf. Innov. Mech. Ind. Appl. ICIMIA 2020 - Conf. Proc.*, pp. 165–172, Mar. 2020, doi: 10.1109/ICIMIA48430.2020.9074954.

[23] R. K. A. Vijay Verma, "A New Similarity Measure Based on Simple Matching Coefficient for Improving the Accuracy of Collaborative Recommendations," *Int. J. Inf. Technol. Comput. Sci.*, vol. 11, no. 6, 2019, doi: https://doi.org/10.5815/ijitcs.2019.06.05.

[24] S. R. K. Maheswari, A. Balamurugan, P. Malathi, "Hybrid clustering algorithm for an

efficient brain tumor segmentation," *Mater. Today*, 2021, doi: https://doi.org/10.1016/j.matpr.2020.08.718.

[25] C. Fan, "Correlation Coefficients of Refined-Single Valued Neutrosophic Sets and Their Applications in Multiple Attribute Decision-Making," *J. Adv. Comput. Intell. Intell. Informatics*, vol. 23, no. 3, pp. 421–426, May 2019, doi: 10.20965/JACIII.2019.P0421.

[26] M. B. M. S. Asfandyar Khan, Abdullah Khan, Muhammad Muntazir Khan, Kamran Farid, Muhammad Mansoor Alam, "Cardiovascular and Diabetes Diseases Classification Using Ensemble Stacking Classifiers with SVM as a Meta Classifier," *Diagnostics*, vol. 12, no. 11, p. 2595, 2022, doi: https://doi.org/10.3390/diagnostics12112595.

[27] E. A. Muhammad Muntazir Khan, Muhammad Zubair Rehman, Abdullah Khan, "Anomaly detection in network traffic with ELSC learning algorithm," *Electron. Lett.*, vol. 16, no. 14, 2024, doi: https://doi.org/10.1049/ell2.13235.

[28] M. I. & M. A. Yousef Alhwaiti, Muntazir Khan, Muhammad Asim, Muhammad Hameed Siddiqi, "Leveraging YOLO deep learning models to enhance plant disease identification," *Sci. Rep.*, vol. 15, p. 7969, 2025, doi: https://doi.org/10.1038/s41598-025-92143-0.

[29] S. A. Lashari, M. M. Khan, A. Khan, S. Salahuddin, and M. N. Ata, "Comparative Evaluation of Machine Learning Models for Mobile Phone Price Prediction: Assessing Accuracy, Robustness, and Generalization Performance," *J. Informatics Web Eng.*, vol. 3, no. 3, pp. 147–163, Oct. 2024, doi: 10.33093/JIWE.2024.3.3.9.

[30] M. Imran, J. Usman, M. Khan, and A. Khan, "A Hybrid Deep Learning VGG-16 Based SVM Model for Vehicle Type Classification," *J. Informatics Web Eng.*, vol. 4, no. 1, pp. 152–167, Feb. 2025, doi: 10.33093/JIWE.2025.4.1.12.

Copyright © by authors and 50Sea. This work is licensed under the Creative Commons Attribution 4.0 International License.