Available online at sjuoz.uoz.edu.krd

Science Journal of University of Zakho

Vol. 12, No.3, pp.285–293 July-September, 2024

journals.uoz.edu.krd

p-ISSN: 2663-628X
e-ISSN: 2663-6298

THE PREDICTION OF HEART DISEASE USING MACHINE LEARNING ALGORITHMS

Snwr Jamal Mohammed[1*] , Noor Bahjat Tayfor[2]

[1]Department of Network Management, Technical College of Informatics TCI, Sulaimani Polytechnic University, Sulaimani, Kurdistan Region, Iraq, snwr.jamal@spu.edu.iq
[2]Department of Information Technology, Kurdistan Technical Institute, Sulaimani, Kurdistan Region, Iraq, noor.tayfor@kti.edu.iq

Received:10 Feb., 2024 / Accepted: 12 June., 2024 / Published: 14July., 2024. https://doi.org/10.25271/sjuoz.2024.12.3.1270

ABSTRACT:

Heart disease threatens the lives of around one individual per minute, establishing it as the foremost cause of mortality in the contemporary era. A wide range of individuals over the globe has encountered the intricacies associated with cardiovascular illness. Various factors, such as hypertension, elevated levels of cholesterol, and an irregular pulse rhythm hinder the early identification of a cardiovascular disease. In cardiology, similar to other branches of Medicine, timely and precise identification of cardiac diseases is of utmost importance. Anticipating the onset of heart failure at the appropriate moment can provide challenges, particularly for cardiologists and surgeons. Fortunately, categorisation and forecasting models can assist the medical business and provide real applications for medical data.

Regarding this, Machine Learning (ML) algorithms and techniques have benefited from the automated analysis of several medical datasets and complex data to aid the medical community in diagnosing heart-related diseases. Predicting if the patient has early-stage cardiac disease is the primary goal of this paper.

A prior study that worked on the Erbil Heart Disease dataset has proved that Naïve Bayes (NB) got an accuracy of 65%, which is the worst classifier, while Decision Tree (DT) obtained the highest accuracy of 98%. In this article, a comparison study has been applied using the same dataset (i.e., Erbil Heart Disease dataset) between multiple ML algorithms, for instance, LR (Logistic Regression), KNN (K-Nearest Neighbours), SVM (Support Vector Machine), DT (Decision Tree), MLP (Multi-Layer Perceptron), NB (Naïve Bayes) and RF (Random Forest). Surprisingly, we obtained an accuracy of 98% after applying LR, MLP, and RF, which was the best outcome. Furthermore, the accuracy obtained by the NB classifier differed incredibly from the one received in the prior work.

KEYWORDS: Machine Learning algorithms, Heart disease prediction, Cardiovascular disease prediction, Logistic Regression (LR), K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Decision Tree (DT), Multi-Layer Perceptron (MLP), Random Forest (RF) and Naïve Bayes (NB).

1. INTRODUCTION

The human heart is an essential organ. It circulates blood throughout our bodies. If it stops operating correctly, the brain and other organs will be unable to do their jobs, and the individual will die within minutes. Several heart-related disorders have seen an uptick in prevalence due to modern lifestyle factors such as increased stress at work and poor dietary choices.

Heart attacks are highly prevalent. The heart's role in the circulatory system cannot be overstated. When the heart stops working correctly, oxygen stops reaching all of the body tissues, and the entire blood system collapses. Therefore, it will result in serious health problems and, in some cases, even death. The terms heart disease and cardiovascular disease convey similar meanings. Narrowed or obstructed blood arteries are the hallmarks of cardiovascular disease, which can cause heart failure, chest pain (angina), and stroke. One of the major killers of disability worldwide is heart disease (Mohan et al., 2019). The good news is that avoiding risk factors like poor food and lifestyle choices, insufficient physical exercise leading to weight gain, and the use of toxic substances like tobacco and alcohol can minimise the risk of cardiovascular disease. Predicting heart disease has grown challenging due to the various risk factors that can contribute, such as diabetes (Kee et al., 2023), hypertension (Islam et al., 2022), excessive cholesterol (Mohi Uddin et al., 2023), irregular pulse rate (Oyeleye et al., 2022), and others. ML methods have proven effective in the healthcare business for early disease prediction (Javaid et al., 2022). Classification methods are crucial for making accurate forecasts.

The fundamental purpose of this article is to compare the performance of several ML techniques for predicting cardiovascular sickness using the Erbil Heart Disease dataset (Qadir Ahmed et al., 2022). Those ML methods are LR (Logistic Regression), KNN (K-Nearest Neighbours), SVM (Support Vector Machine), DT (Decision Tree), MLP (Multi-Layer Perceptron), NB (Naïve Bayes) and RF (Random Forest). The assessment will evaluate their accuracy, confusion matrix, precision, recall, and F-measure outcomes. Finally, we will compare our results with those obtained by (M et al., 2022) and determine the differences between both results. Fortunately, the accuracy obtained by LR, MLP, and RF is 98%, which is the highest result. However, the accuracy obtained by the NB classifier is 96%, significantly higher than the result obtained by (M et al., 2022), which is only 65%.

The subsequent sections of this article are structured in the following manner: After a brief literature review in Section 2, the following sections describe the ML algorithms utilised in Section 3, then the conduction an analysis of the study materials and techniques will be highlighted in Section 4, Section 5 will

[*] Corresponding author

provide the results and discussion, and finally a wrap up with a discussion of future research will be stated in Section 6.

2. RELATED WORK

Multiple studies have demonstrated that ML models exhibit higher accuracy compared to standard statistical models when it comes to forecasting the risk of mortality and hospitalisation in cardiovascular disease patients. It was found that (Alaa et al., 2019) an automated ML model outperformed traditional statistical models in predicting cardiovascular disease risk in a large cohort of UK Biobank participants. Weng et al. (2017) discovered that ML models outperformed traditional methods in accurately forecasting the risk of cardiovascular disease using regular clinical data, (Shouval et al., 2017) found that ML systems surpassed the GRACE score in accurately predicting the likelihood of death within 30 days following an ST-elevation myocardial infarction. Moreover, Betancur et al., (2018) found that a ML model combining clinical and myocardial perfusion imaging data had better prognostic value than traditional models. Using the 14 features of the UCI Machine Learning Repository from Kaggle, an application for heart vulnerability prediction has been developed based on primary symptoms. Numbers like age, blood pressure, cholesterol levels, heart rate, blood sugar levels, body mass index, old peak, maximum heart rate achieved during exercise, number of significant vessels coloured by fluoroscopy, and serum cholesterol are examples of cardiovascular disease attributes. There are also categorical attributes like gender, chest pain type, fasting blood sugar, resting electrocardiographic results, exercise-induced angina, the peak exercise ST segment slope, thalassemia, and the presence or absence of heart disease. The study compared the accuracy of various machine-learning methods. ML techniques used are LR, NB, SVM with the Linear Kernel (LK) function, RF, and the Radial Basis Kernel (RBF) function. With an accuracy rating of 84.81%, the RF classifier outperformed all other assessed ML methods. The study used a train-test split ratio 60:40 for both the RF and SVM algorithms (Pe et al., 2021).

Another cardiovascular disease risk prediction study used data from the UC Irvine repository, which has 14 attributes and 303 samples available to the public. The dataset attributes used to forecast cardiovascular disease are as follows: numerical (continuous) attributes such as age, resting blood pressure, cholesterol, maximum heart rate, and ST depression. Some of the categorical attributes that are shown are sex, the type of chest pain, fasting blood sugar, resting electrocardiogram result, exercise-induced angina, slope of peak ST segment, number of significant vessels, thallium stress result, and cardiovascular disease presence or absence. An analysis was conducted to compare MLPro and KNN, two ML algorithms. Based on the experimental data, the MLP model achieves higher detection accuracy 82.47% and area under the curve 86.41% than the KNN model. The study used a train-test split ratio of 80:20 for both ML algorithms (Pal et al., 2022).

Statlog, VA Long Beach, Cleveland, Hungary, and Switzerland databases extracted from the UCI repository have been used in heart disease prediction. The dataset includes various features related to cardiovascular health, such as numerical attributes and categorical attributes. Numerical attributes refer to age, resting blood pressure, serum cholesterol, maximum heart rate achieved, real ST depression and the number of vessels coloured by fluoroscopy. Categorical attributes refer to sex, chest pain type, fasting blood sugar, resting ECG results, exercise-induced angina, the slope of the peak exercise, Thalasemia defect types, and diagnosis of heart disease. Performance comparisons have been conducted using the following ML approaches: Several methods are available, including RF, NB, DT, KNN, SVM, LR, Gradient Boosting (GB), AdaBoost (AB) and Bagging Method (BM). The RF classifier attained a prediction accuracy of 97.05% for cardiovascular sickness by using Principal Component Analysis (PCA) as a feature selection technique. While the Decision Tree (DT) only managed a 97.89% success rate, the Support Vector Classifier (SVC) managed a whopping 98.31%. Scientific studies have shown that when combined with other ML algorithms, the LASSO feature selection method produces a highly accurate and connected collection of features that yields optimal results across various criteria. The train-test split ratio of 80:20 was used in the study (Mahmoud et al., 2022).

The Hungarian and Statlog (heart) datasets have been employed to forecast cardiovascular disease using the Weka tool. The Hungarian database was created at the Hungarian Institute of Cardiology in Budapest and contains 294 instances. The Statlog (heart) dataset consists of 304 cases and contains 76 attributes, although only 14 of them were used in the experiments. The dataset consists of numerical attributes and categorical attributes. Numerical attributes refer to age, resting blood pressure, serum cholesterol and maximum heart rate achieved.

However, categorical attributes refer to sex, chest pain type diagnosis of heart disease and medication. The classifiers used are REP Tree, M5P Tree, Random Tree, LR, NB, J48, and JRIP. However, the performance metrics such as accuracy, mean absolute error (MAE), root mean square error (RMSE), and prediction time are calculated based on the predictions made on the testing set. Moreover, 10-fold cross-validation is a common technique used in ML to assess the performance and generalisability of predictive models. The Random Tree model in the Hungarian database study demonstrated a remarkable accuracy rate of 99.81%. Similarly, in the study utilising the Statlog (heart) database, the Random Tree model achieved a 100% accuracy rate. The results suggest that the Random Tree algorithm effectively predicts cardiovascular illness (Nadakinamani et al., 2022).

An analysis was performed on a heart disease dataset, utilising three correlation approaches to uncover robust associations between features. The dataset related to heart disease patients was obtained from the UCI Machine Learning Repository. The dataset includes 303 instances, and 14 attributes involve numerical attributes such as age, resting blood pressure, serum cholesterol, and maximum heart rate achieved and categorical attributes such as gender, chest pain type, fasting blood sugar, resting electrocardiographic results and exercise-induced angina. A comparison has been made between the accuracy of various classifiers, including KNN, LR, Gaussian Naïve Bayes, DT, and RF. The study found that when comparing accuracy and area under the ROC curve, Gaussian Naive Bayes performed better than KNN, LR, RF, and DT. Two more testing approaches, hold-out validation and 10-fold cross-validation, showed that Gaussian Naive Bayes performed better. The study also observed that LR achieved the highest performance in stratified 10-fold cross-validation and repeated random train-test splits ratio of 70:30 (Aradhana et al., 2021).

An intelligent system for predicting cardiovascular disease has been proposed to incorporate different ML algorithms and ensemble techniques.

The dataset utilised in this work is formed by merging five distinct datasets: Cleveland, Hungary, Switzerland, VA Long Beach, and Statlog heart disease databases. The datasets were collected from the UCI machine learning repository. The merged dataset has about 1190 cases with 14 distinct features, the same as those mentioned in (Aradhana et al., 2021). Those classifiers applied are DT, RF, KNN, AdaBoost, and Gradient Boosting. The proposed intelligent system demonstrated exceptional accuracy in predicting cardiovascular illness, with the Random Forest Boosting Model (RFBM) attaining the best accuracy rate of 99.05%. The Relief feature selection technique successfully generated a feature set that correlated with the ML classifiers, enhancing performance. The comparison of different classifiers and hybrid techniques showed that RFBM outperformed other models, while K-Nearest Neighbors (KNN) had the lowest accuracy (Ghosh et al., 2021).

Predicting cardiovascular disease using different ML algorithms is the subject of an additional investigation. The research utilised the same dataset mentioned in (Aradhana et al., 2021) and (Ghosh et al., 2021). The study used ANN models as classifiers in addition to RF, gradient boosting, and LR. Although the RF and gradient boosting models achieved 94% accuracy, the RF model was more successful. The ANN model attained a higher percentage accuracy rate of 91% than the LR model's 79%. According to the paper's proposed cloud-based approach, early cardiovascular disease diagnosis and treatment planning could be possible with the help of these models (Kachhawa et al., 2022).

Another field of research focused on developing an AI system that can reliably detect and forecast who would develop cardiovascular disease. The study utilised a dataset of 518 heart disease patients who were randomly selected from the Lady Reading Hospital (LRM) and the Khyber Teaching Hospital (KTH) in Khyber Pakhtunkhwa (KPK), Pakistan. The dataset comprises diverse attributes pertaining to cardiovascular disease, encompassing numerical attributes such as age, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol levels, body mass index (BMI) and glucose levels. It also includes categorical attributes such as gender, smoking status, alcohol intake, physical activity level, and cardiovascular disease status (presence or absence). DT, RF, LR, NB and SVM are among the methods that have been employed. When predicting cardiovascular sickness, the RF algorithm had the best sensitivity of 92.11%, ROC curve of 87.73%, and accuracy of 85.01% when the train-test split ratio of 70:30 and 75:25 was applied. Therefore, the best approach for cardiovascular illness classification and prediction is the RF algorithm (Khan et al., 2023).

Another heart disease dataset has been compared using six different ML algorithms, including KNN, DT, SVM, LR, NB, and RF. The dataset, obtained from the Kaggle repository, has 303 instances and 14 features, encompassing numerical attributes such as age, resting blood pressure, serum cholesterol level, maximum heart rate, old peak, and the number of main vessels coloured by fluoroscopy. The categorical attributes in this dataset are sex, chest pain type, fasting blood sugar, resting electrocardiographic results, exercise-induced angina, slope, thalassemia level, and target. The accuracy of each classifier has been compared using train-test split and k-fold cross-validation methods, employing various ratios and values. For an 80:20 and 75:25 split, the best accuracy result from the logistic regression model is 82%. SVM's 82% accuracy for the 75:25 split is comparable. However, when employing the LR model with the tenfold cross-validation approach, accuracy increases to 82%. Compared to the other classifiers, LR and SVM produce more accurate results (Mengiste et al., 2020).

The performance of seven ML algorithms has been compared. Those algorithms are SVM, DT, RF, NB, LR, Adaptive Boosting (AB), and Extreme Gradient Boosting (XGB). The dataset, obtained from the UCI Machine Learning Repository, has 14 features, encompassing numerical attributes such as age, resting blood pressure, serum cholesterol level, maximum heart rate, old peak, and the number of main vessels. The categorical attributes in this dataset are sex, chest pain type, fasting blood sugar, resting electrocardiographic results, exercise-induced angina, slope, thalassemia level, and target. Moreover, it has been found that the XG-Boost and RF achieve the highest accuracy of about 86%. In addition, ML has shown to be an invaluable resource for the precise diagnosis and prediction of cardiovascular disease (Mekala et al., 2023).

A recent study compared and evaluated various ML models to predict cardiovascular sickness in people with heart issues.

The dataset utilised in the study is readily accessible on Kaggle and was gathered by the American Heart Failure Institute. The dataset includes numerical variables such as age, creatinine_phosphokinase,ejection_fraction,platelets,serum_creatinine, serum_sodium, and time, as well as categorical features such as anaemia, diabetes, high pressure, sex, smoking, and death_event. LR, SVM, RF, NB, Gradient Boosting Classifier (GBC), AdaBoost (ABC), CatBoost (CBC), DT, and KNN are some of the models that are part of the list. According to the results, LR outperformed the other classifiers with an accuracy of 82.76% in the prediction of heart illness. ML can improve healthcare outcomes, and the study suggests it could help anticipate cardiac crises (Ahmed et al., 2022).

3. MACHINE LEARNING ALGORITHMS

We used seven different ML algorithms to make predictions of cardiovascular disease. We have applied the same algorithms utilised in (M et al., 2022) to achieve the article's primary goal. Those algorithms are:

3.1 Logistic Regression (LR)

Regarding classification problems, especially binary classification tasks, logistic regression (LR) is a popular and effective ML approach. It performs well on the categorical dependent variable because of its simplicity (Sen, 2017). The approach relies on a sigmoid function, a mathematical prediction tool. In logistic regression, the likelihood of being the threshold value is either 0 or 1. The value curve used in logistic regression must be between 0 and 1 to produce an Scurve. The Sigmoid curve is another name for the logistic function (Lee et al., 2020; Ramalingam et al., 2018; Sarker, 2021). The LR formulae could be found in equation (1):

$$P = (y = 1|X) = \frac{1}{1 + e^{-wa}} \quad 1$$

Euler's number (e) is a numerical constant, and a serves as an input to the function.

3.2 K Nearest Neighbour (KNN)

Classification and regression are two applications of the supervised ML method known as K-Nearest Neighbours (KNN) (Thomas et al., 2016). By using the entire dataset, the KNN algorithm will make a prediction. For the trait we desire to predict, the algorithm will seek out the K dataset samples most similar to our observation (Bansal et al., 2022; Cunningham et al., 2021). The approach then uses the values of the y variables produced by these K neighbours to forecast the value of the y variable in the observation of interest. The following equation (2) represents the Euclidean distance between two observations:

$$d(x_i, y_i) = \sqrt{\left(x_{i,1} - y_{i,1}\right)^2 + \cdots + \left(x_{i,m} - y_{i,m}\right)^2} \quad 2$$

Since the K-nearest neighbour does not need training before generating predictions, it can make those predictions with much less processing time. K-nearest-neighbor can be readily implemented using only K and the distance function's value. Nevertheless, it has challenges when dealing with massive data sets and exhibits poor results when confronted with a high number of dimensions in the data (Arafat et al., 2019).

3.3 Support Vector Machine (SVM)

The support vector machine (SVM) method is a popular ML method for regression and classification applications (Gupta et al., 2017). As a result of its superior performance relative to competing algorithms, it is frequently employed as the classification method of choice. This method treats each attribute in the dataset as a coordinate and plotted on a hyperplane (Ghumbre et al., 2012). Finding the hyperplane that separates two classes allows for classification. It is a non-probabilistic binary linear classifier since it constructs a model that assigns fresh samples to each other (Aswini et al., 2022; Suresh et al., 2022).

3.3.1 Kernel Function

The kernel function transforms data points into a linear decision surface if they are currently distributed nonlinearly (i.e., not separably distributed)—examples of Kernel function: Linear Kernel (LKF), Polynomial Kernel (PKL), Sigmoid Kernel (SKL), Exponential Radial Basis Kernel (ERBKL), and Radial Basis (RBF) (Arumugam et al., 2022).

3.4 Decision Tree (DT)

Classifying massive datasets is a typical application of the DT technique. Data is organised in decision trees by linking the first "root" node to the following "leaf" nodes. It is possible to instantiate the resultant tree as a set of instructions. DTs are characterised by uncomplicated and direct rules (Deepika et al., 2020). DTs are frequently used because they are convenient, trustworthy, and straightforward. Within a decision tree, every node refers to an attribute test, while each branch indicates a test result, and the leaves reflect the class distributions, much like a flowchart (Aladeyelu et al., 2023). The jump from one node to another could be measured by the Entropy of Information gain, shown in equation (3):

$$Entropy\ (S) = \sum_{i=1}^{c} -\left(P_i \, log_2 P_i\right) \quad 3$$

and

The Information Gain could be computed as depicted in equation (4):

$$\begin{aligned} &Information\ Gain\ (S, A) \quad 4 \\ &= Entropy\ (S) \sum\nolimits_{v \in values(A)} \frac{|S_v|}{|S|}\ Entropy\ (S) \end{aligned}$$

3.5 Multi-Layer Perceptron (MLP)

MLP indicates an artificial neural network with several processing power layers. While a single perceptron can only handle linear problems, MLP models are more versatile. MLP is employed to solve complex problems. An MLP is similar to a deep feed-forward neural network due to its multiple layers (Mohanty et al., 2022). Backpropagation is used to train the feed-forward neural network. The weights are fine-tuned to train a neural network with as few errors as possible(Wu et al., 2019). The sigmoid function is the standard activation function (Pal et al., 2022) employed in the hidden layer equation (5) and the tanh activation function equation (6).

$$Sigmoid\ function, \sigma(x) = \frac{1}{1 + e^{-x}} \quad 5$$

$$\tanh(x) = 2\sigma(x)(2x) - 1 \quad 6$$

The loss function could be computed using equation (7):

$$Mean\ square\ error = \frac{1}{P}\sum_{i=1}^{p}(m_i - \hat{m})^2 \quad 7$$

Where P refers to the total number of samples, m refers to the actual observed value, and $\hat{m}$ refers to the predicted value.

Errors can be kept to a minimum by following the weight updating technique stated in equation (8).

$$* w_j = w_j - \alpha\left(\frac{dE}{dw_j}\right) \quad 8$$

where $* w_j$ represents the new weight, w_j represents the previous weight, α represents the learning rate (0< α <1) and E refers to the error term $m - \hat{m}$.

3.6 Naïve Bayes (NB)

NB is a classifier that can perform regression and classification and uses supervised learning. It uses a Bayesian approach to statistics. The theory of least squares provides the basis for two- and multi-class classification systems. The technique is primarily suited for binary classification, although it can be used for any input data type. NB is a simple framework that works well with massive datasets. In terms of accuracy, it surpasses other approaches to ML (Liu et al., 2017).

Bayes' theorem computes the probability of an event happening by considering the likelihood of a prior event. The mathematical equation (Mahmoud et al., 2022) can be written as equation (9):

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad 9$$

Where

$P(c|x)$ refers to the Posterior Probability and it could be computed using equation (10):

$$P(c|x) = P(x_1\backslash c) * P(x_2\backslash c) * P(x_3\backslash c) * \dots * P(x_n\backslash c) \qquad 10$$

P(x) refers to the Predictor Prior Probability
P(c) refers to the Class Prior Probability
P(x/c) refers to the likelihood.

3.7 Random Forest (RF)

RF refers to a ML algorithm that constructs several DTs using training data sets to generate a classification model. This technique employs a tree selection approach that yields high accuracy when dealing with large data sets (Thanh Noi et al., 2017). This technique integrates two feature selection procedures, bagging and random selection, to provide a more efficient ensemble model. Employing many trees in the RF technique mitigates the risk of overfitting and reduces training time. Additionally, it gives estimations for essential classification variables and addresses missing data, both of which contribute to enhanced accuracy (Jabbar et al., 2016).

4. MATERIALS AND METHODS

The overall steps of our proposed method are illustrated in the following data flowchart shown in Figure 1:

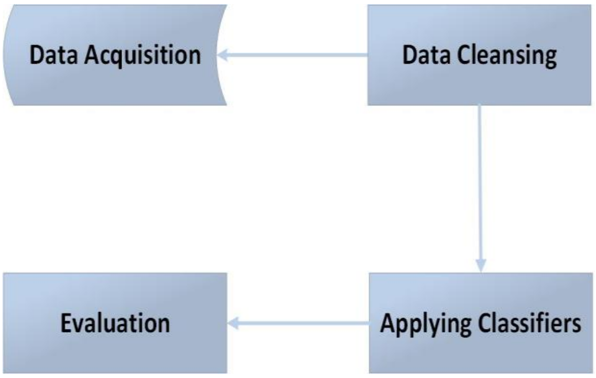

Figure 1: Data Flowchart of our Proposed Method

4.1 Data Acquisition

The data collection was compiled in a hospital setting using the information provided by actual patients. The Medical Help Center is a private heart centre in Erbil, Kurdistan Region, Iraq, where all of the data was gathered. The dataset is freely available in the Kaggle repository (Qadir Ahmed et al., 2022). The primary motivation behind compiling this dataset is to study data contributed by native patients with heart problems to make accurate diagnoses of heart diseases. The collected data is categorised into five distinct classifications. The information can be classified as demographic data, medical history, physical examinations and symptoms, medical lab tests, and diagnostic characteristics. A total of 400 patient data with 21 attributes have been collected. The dataset features are described below in Table 1:

Table 1: Attributes details and values

Attribute	Description	Value
age	The patient age	Integer value
gender	The patient sex	0=Male 1=Female
smoke	If the patient somke or not	0=No 1=Yes
years	The duration of smoking for individuals who smoke	Integer value
ldl	Patient's cholesterol ratio	Integer value
chest-pain	The kind of chest pain	1=Typical angina 2=Atypical angina 3=Non-anginal pain 4=Asymptomatic
height	The patient height	Integer value
weight	The patient weight	Integer value
family-history	If the family has heart disease in the history or not	0=No 1=Yes
active	Whether the patient is active or not	0=No 1=Yes
lifestyle	The patient lifestyle	1=City 2=Town 3=Village
cl	Whether the patient undergoes cardiac catheterisation or other cardiac intervention	0=No 1=Yes
hr	The ratio of Heart Rate	Integer value
dm	Whether the patient suffers from diabetes disease or not	0=No 1=Yes
bpsys	Systolic Blood Pressure	Integer value
bpdias	Diastolic Blood Pressure	Integer value
htn	Whether the patient suffers from hypertension or not	0=No 1=Yes
ivsd	An echo parameter (Interventricular septum thickness at end-diastole) is a measurement that is used to determine the muscle thickness of the left ventricular hypertrophy (LVH).	0 or 1
ecg-test	The ECG reading test	1=ST-Elevation 2=ST-Depression 3=T-Inversion 4=Normal
q-wave	The indications of the Q wave's presence	0=No 1=Yes
result	Whether the patient is afflicted with	0=without heart disease

	cardiovascular disease or not	1=with heart disease

4.2 Data Cleansing

The data is subjected to preprocessing to eliminate missing or noisy records. The data undergoes a cleansing procedure involving removing noisy records and outliers and identifying missing values as absent to ensure accurate results. Moreover, some features have been removed for security reasons, such as the patient's name, the number of personal identity, passport number, date of birth, place of birth, residence place, mobile number, etc. Subsequently, we have obtained 333 instances after applying the cleansing stage.

4.3 Applying Classifiers

All the ML classifiers mentioned in Section 3 have been applied to our dataset, and the comparison task will be stated in Section 5.

4.4 Performance Evaluation

The confusion matrix, accuracy, precision, recall, and f1-measure have been utilised to evaluate the efficacy of our seven ML algorithms.

4.4.1 Confusion Matrix

The confusion matrix is utilised to evaluate the accuracy of the ML method. TP, FP, TN, and FN are abbreviations for true positive, false positive, true negative, and false negative, respectively. The confusion matrix commonly evaluates these four metrics(Markoulidakis et al., 2021). The following Table 2 refers to the typical confusion matrix:

Table 2: Confusion Matrix

Confusion Matrix		**Predictive Class**	
		0	**1**
Actual Class	**0**	TP	FN
	1	FP	TN

Where

- TP: True Positive denotes heart disease patients accurately diagnosed by the ML model.
- TN: True Negative denotes individuals without heart issues and accurately categorised by the ML model.
- FP: False Positive denotes heart disease patients; the ML model incorrectly categorises them.
- FN: False Negative denotes patients without heart disease problems incorrectly caregorised by the ML model.

4.4.2 Accuracy

The accuracy of prediction is measured as a percentage of correct observations out of all observations equation (11) shows how the accuracy is measured:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \qquad 11$$

4.4.3 Precision

It is the ratio of accurately anticipated positive observations to the total number of positive observations. The precision is measured in equation (12):

$$Precision = \frac{TP}{TP + FP} \qquad 12$$

4.4.4 Recall/ Sensitivity

The computation involves dividing the count of accurately predicted positive outcomes by the overall count of observations. The response capacity is another name for sensitivity. The recall is measured in equation (13):

$$Recall = \frac{TP}{TP + FN} \qquad 13$$

4.4.5 F1-Score

It represents the weighted average of the obtained precision and recall values. The f1-score is calculated in equation (14):

$$F1 - Score = 2 \times \left[\frac{precision \times recall}{(precision+recall)}\right] \qquad 14$$

5. RESULTS AND DISCUSSION

This article evaluates LR, KNN, SVM, DT, MLP, NB and RF regarding accuracy, precision, recall and f1-score metrics. The accuracy of LR, MLP and RF outperforms the other classifiers by only 98%, which is considered an extremely significant value. Although, in the previous work (M et al., 2022), RF obtained just 93%, which is lower than our accuracy. Furthermore, NB classifier obtains 96%, significantly higher than the 65% value of NB obtained in (M et al., 2022). However, the KNN classifier reached a value of 81%, which is supposed to be the lowest value received, and it is still better than 76%, the figure gained by (M et al., 2022). Figure 2 shows the accuracies of all classifiers that have been utilised so far in this paper.

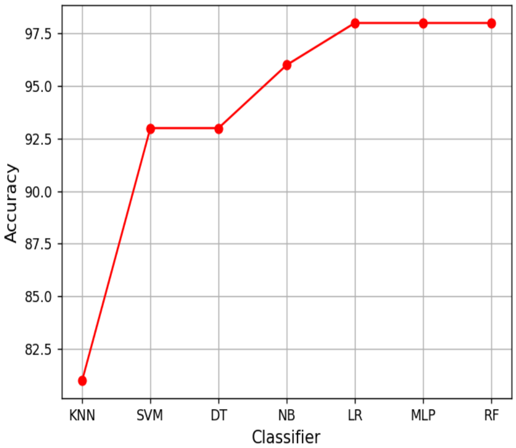

Figure 2: The Accuracy of ML Algorithms

Additionally, to revise the accuracy of the classifiers applied in the prior work, Figure 3 represents the accuracy of NB, KNN, RF, SVM and DT classifiers obtained by (M et al., 2022).

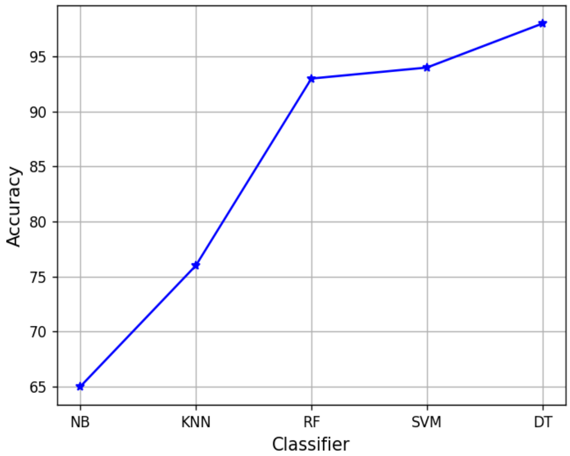

Figure 3: The Accuracy of ML Algorithms obtained from (M et al., 2022)

All the classifiers achieved high TP rates, ranging from 66 to 69, in correctly classifying individuals with heart disease. In addition, the TN rates are consistently high for nearly all classifiers, ranging from 27 to 29, indicating the number of correctly diagnosed patients without cardiac disease. The exception is the KNN classifier, which only correctly classified 13 patients. The FP numbers for the misdiagnosed heart disease patients are 2, 3, 4, and 5. Conversely, the FN values for misclassified patients without heart disease in the case of all classifiers are 0, 1, and 2, except KNN, which has a value of just 16. All the values of the confusion matrix are represented in Table 3: The Results of the Confusion Matrix for all Classifiers:

Table 3: The Results of the Confusion Matrix for all Classifiers

	TP	**FP**	**FN**	**TN**
LR	69	2	0	29
KNN	68	3	16	13
SVM	66	5	2	27
DT	66	5	2	27
MLP	69	2	0	29
NB	67	4	0	29
RF	69	2	0	29

Notably, the confusion matrix results acquired by the previous work(M et al., 2022) differ considerably from our results. Table 4 shows the values of the confusion matrix obtained by (M et al., 2022):

Table 4: The Results of the Confusion Matrix for the Classifiers applied in (M et al., 2022)

	TP	**FP**	**FN**	**TN**
NB	52	13	31	4
KNN	51	25	19	5
RF	50	43	1	6
SVM	50	44	0	6
DT	54	44	0	2

As previously stated, the accuracy values of all classifiers are remarkably significant. Therefore, it is strongly advised that precision, recall, and f1-score metrics be provided for each classifier to assess their performance based on these metrics. The precision, recall, and f1-score metrics for KNN are significantly lower than other methods because of the low accuracy value of KNN, as indicated in Figure 2 and previously described. Conversely, the remaining classifiers obtained elevated values in the abovementioned criteria. Table 5 represents the precision, recall, and f1-score values for all the ML classifiers that have been applied in this article.

Table 5: Precision, Recall, and F1-Score Values for all Classifiers

		0	**1**
LR	**Precison**	1.00	0.94
	Recall	0.97	1.00
	F1-Score	0.99	0.97
KNN	**Precison**	0.81	0.81
	Recall	0.96	0.45
	F1-Score	0.88	0.58
SVM	**Precison**	0.97	0.84
	Recall	0.93	0.93
	F1-Score	0.95	0.89
DT	**Precison**	0.97	0.84
	Recall	0.93	0.93
	F1-Score	0.95	0.89
MLP	**Precison**	1.00	0.94
	Recall	0.97	1.00
	F1-Score	0.99	0.97
NB	**Precison**	1.00	0.88
	Recall	0.94	1.00
	F1-Score	0.97	0.94
RF	**Precison**	1.00	0.94
	Recall	0.97	1.00
	F1-Score	0.99	0.97

Finally, we applied a t-test to each pair of our classifiers. A T-test is a well-known statistical hypothesis test for comparing two samples. We have chosen the significance level = 0.05. The difference between any pair of classifiers must be less than the specified significance level to be statistically significant. Otherwise, the differences will be statistically insignificant. Table 6 depicts the differences between each pair of classifiers. The values with superscript ***** are statistically significant difference pairs of classifiers.

Table 6: Statistical Analysis

	LR	SVM	KNN	DT	MLP	NB	RF
LR	-	**0.019***	**0.0008***	**0.0008***	0	**0.0260***	**0.030***
SVM	-	-	0.050	0.050	**0.019***	**0.035***	**0.0255***
KNN	-	-	-	0	**0.0008***	**0.0031***	**0.0005***
DT	-	-	-	-	**0.0008***	**0.003***	**0.00058***
MLP	-	-	-	-	-	**0.0260***	**0.0300***
NB	-	-	-	-	-	-	**0.0300***
RF	-	-	-	-	-	-	-

CONCLUSIONS

In the end, cardiovascular disease has become one of the leading causes of death worldwide. Predicting cardiovascular disease using ML algorithms holds promise for early detection and treatment alternatives. By evaluating several risk variables and using sophisticated classification techniques, these models may assist healthcare practitioners in identifying individuals who are more susceptible to developing heart disease. It allows for prompt implementation of preventive measures and interventions.

This research presents various ML algorithms and methodologies for classifying heart disorders using a real dataset obtained from the Erbil Heart Disease Centre in Erbil (Qadir Ahmed et al., 2022), for example, LR, KNN, SVM, DT, MLP, NB and RF. As a result, applying LR, MLP, and RF obtained the most accurate early-stage prediction for heart disease by 97%. Likewise, we have compared our results with those obtained from prior work (M et al., 2022), and we have reached that our results outperformed those obtained in (M et al., 2022).

In future research, the experiment can be enhanced by utilising a more extensive dataset and comparing the findings with the previously obtained outcomes in this work to forecast the survival rates of patients with heart disease.

REFERENCES

Ahmed, S., Shaikh, S., Ikram, F., Fayaz, M., Alwageed, H. S., Khan, F., & Jaskani, F. H. (2022). Prediction of Cardiovascular Disease on Self-Augmented Datasets of Heart Patients Using Multiple Machine Learning Models. *Journal of Sensors, 2022*. doi: 10.1155/2022/3730303

Alaa, A. M., Bolton, T., Angelantonio, E. Di, Rudd, J. H. F., & van der Schaar, M. (2019). Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLoS ONE*, *14*(5). doi: 10.1371/journal.pone.0213653

Aladeyelu, A. C., & Adekunle, G. T. (2023). *Predicting Heart Disease Using Machine Learning APPLICATION OF GIS IN ELECTRICITY DISTRIBUTION OF IKEDC: A CASE STUDY OF RESIDENTIAL AREA (OMOLE ESTATE PHASE ONE) AND THE INDUSTRIAL AREA OF IKEJA. View project Predicting Heart Disease Using Machine Learning*. *10*. Retrieved from https://www.researchgate.net/publication/370583760

Aradhana, S., Jankisharan, P., Virendra, S. K., & Ashish, M. (2021). Cardiovascular diseases prediction using various machine learning techniques. *IOP Conference Series: Materials Science and Engineering*, *1022*(1). doi: 10.1088/1757-899X/1022/1/012003

Arafat, M. Y., Hoque, S., Xu, S., & Farid, D. M. (2019). *Machine learning for mining imbalanced data*.

Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A., & Gonzales-Yanac, T. (2022). Multiple disease prediction using Machine learning algorithms. *Materials Today: Proceedings*. doi: 10.1016/j.matpr.2021.07.361

Aswini, J., Yamini, B., Jatothu, R., Nayaki, K. S., & Nalini, M. (2022). An efficient cloud-based healthcare services paradigm for chronic kidney disease prediction application using boosted support vector machine. *Concurrency and Computation: Practice and Experience*, *34*(10), e6722.

Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. *Decision Analytics Journal*, *3*, 100071.

Betancur, J., Otaki, Y., Motwani, M., Fish, M. B., Lemley, M., Dey, D., Gransar, H., Tamarappoo, B., Germano, G., Sharir, T., Berman, D. S., & Slomka, P. J. (2018). Prognostic Value of Combined Clinical and Myocardial Perfusion Imaging Data Using Machine Learning. *JACC: Cardiovascular Imaging*, *11*(7), 1000–1009. doi: 10.1016/j.jcmg.2017.07.024

Cunningham, P., & Delany, S. J. (2021). k-Nearest neighbour classifiers-A Tutorial. *ACM Computing Surveys (CSUR)*, *54*(6), 1–25.

Deepika, P., & Sasikala, S. (2020). Enhanced model for prediction and classification of cardiovascular disease using decision tree with particle swarm optimization. *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 1068–1072.

Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. M. J. M., Ignatious, E., Shultana, S., Beeravolu, A. R., & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques. *IEEE Access*, *9*, 19304–19326. doi: 10.1109/ACCESS.2021.3053759

Ghumbre, S. U., & Ghatol, A. A. (2012). Heart disease diagnosis using machine learning algorithm. *Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) Held in Visakhapatnam, India, January 2012*, 217–225.

Gupta, N., Ahuja, N., Malhotra, S., Bala, A., & Kaur, G. (2017). Intelligent heart disease prediction in cloud environment through ensembling. *Expert Systems*, *34*(3), e12207.

Islam, S. M. S., Talukder, A., Awal, M. A., Siddiqui, M. M. U., Ahamad, M. M., Ahammed, B., Rawal, L. B., Alizadehsani, R., Abawajy, J., Laranjo, L., Chow, C. K., & Maddison, R. (2022). Machine Learning Approaches for Predicting Hypertension and Its Associated Factors Using Population-Level Data From Three South Asian Countries. *Frontiers in Cardiovascular Medicine*, *9*. doi: 10.3389/fcvm.2022.839379

Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2016). Intelligent heart disease prediction system using random forest and evolutionary approach. In Journal of Network and Innovative Computing (Vol. 4). Retrieved from www.mirlabs.net/jnic/index.html

Javaid, M., Haleem, A., Pratap Singh, R., Suman, R., & Rab, S. (2022). Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks*, *3*, 58–73. doi: 10.1016/j.ijin.2022.05.002

Kachhawa, A., & Hitt, J. (2022). An Intelligent System for Early Prediction of Cardiovascular Disease using Machine Learning. *Journal of Student Research*, *11*(3), 1–10. Retrieved from www.JSR.org

Kee, O. T., Harun, H., Mustafa, N., Abdul Murad, N. A., Chin, S. F., Jaafar, R., & Abdullah, N. (2023). Cardiovascular complications in a diabetes prediction model using machine learning: a systematic review. In Cardiovascular Diabetology (Vol. 22, Issue 1). BioMed Central Ltd. doi: 10.1186/s12933-023-01741-7

Khan, A., Qureshi, M., Daniyal, M., & Tawiah, K. (2023). A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction. *Health & Social Care in the Community*, *2023*, 1–10. doi: 10.1155/2023/1406060

Lee, H., & Kim, H.-S. (2020). Logistic regression and least absolute shrinkage and selection operator. *Cardiovascular Prevention and Pharmacotherapy*, *2*(4), 142–146.

Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Q., & Wang, Q. (2017). A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method. *Computational and Mathematical Methods in Medicine*, *2017*. doi: 10.1155/2017/8272091

M, M., K, Y., & Kumar, M. M. (2022). Cardiovascular Disease Prediction using Machine Learning Methods. *IJARCCE*, *11*(9). doi: 10.17148/IJARCCE.2022.11920

Mahmoud, S., Gaber, M., Farouk, G., & Keshk, A. (2022). Prediction of Cardiovascular Disease Using Machine Learning Techniques. *International Journal of Computers and Information (IJCI)*, *9*(2), 25–44. Retrieved from https://ijci.journals.ekb.eg/

Markoulidakis, I., Kopsiaftis, G., Rallis, I., & Georgoulas, I. (2021). Multi-Class Confusion Matrix Reduction method and its application on Net Promoter Score classification problem. *ACM International Conference Proceeding Series*, 412–419. doi: 10.1145/3453892.3461323

Mekala, S., Balarama Krishna, Y., Nagaraju, M., Abhiram, P., Pavan Mahith, K., & Balarama Krishna, Y. (2023). *CARDIOPATHY-HEART DISEASE PREDICTION USING MACHINE LEARNING*. Retrieved from https://www.researchgate.net/publication/371760631

Mengiste, B. K., Tripathy, H. K., & Rout, J. K. (2020). *Analysis and Prediction of Cardiovascular Disease Using Machine Learning Techniques*. 133–141. Retrieved from http://www.springer.com/series/7818

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, *7*, 81542–81554. doi: 10.1109/ACCESS.2019.2923707

Mohanty, M. D., & Mohanty, M. N. (2022). Verbal sentiment analysis and detection using recurrent neural network. *Advanced Data Mining Tools and Methods for Social Computing*, 85–106. doi: 10.1016/B978-0-32-385708-6.00012-6

Mohi Uddin, K. M., Ripa, R., Yeasmin, N., Biswas, N., & Dey, S. K. (2023). Machine learning-based approach to the diagnosis of cardiovascular vascular disease using a combined dataset. *Intelligence-Based Medicine*, *7*. doi: 10.1016/j.ibmed.2023.100100

Nadakinamani, R. G., Reyana, A., Kautish, S., Vibith, A. S., Gupta, Y., Abdelwahab, S. F., & Mohamed, A. W. (2022). Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine Learning Techniques. *Computational Intelligence and Neuroscience*, *2022*. doi: 10.1155/2022/2973324

Oyeleye, M., Chen, T., Titarenko, S., & Antoniou, G. (2022). A Predictive Analysis of Heart Rates Using Machine Learning Techniques. *International Journal of Environmental Research and Public Health*, *19*(4). doi: 10.3390/ijerph19042417

Pal, M., Parija, S., Panda, G., Dhama, K., & Mohapatra, R. K. (2022). Risk prediction of cardiovascular disease using machine learning classifiers. *Open Medicine (Poland)*, *17*(1), 1100–1113. doi: 10.1515/med-2022-0508

Pe, R., Subasini, .C.A., Vanitha Katharine, A., Kumaresan, V., GowdhamKumar, S., & Nithya, T. M. (2021). A Cardiovascular Disease Prediction using Machine Learning Algorithms. *Annals of the Romanian Society for Cell Biology*, *25*(2), 904–912. Retrieved from https://www.researchgate.net/publication/350312435

Qadir Ahmed, H., Othman Amen, S., Qasim Rassol, B., & Ismael Hamad, I. (2022). Erbil Heart Disease Dataset. In Kaggle.

Ramalingam, V. V, Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology*, *7*(2.8), 684–687.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 160.

Sen, S. K. (2017). Predicting and diagnosing of heart disease using machine learning algorithms. *International Journal of Engineering and Computer Science*, *6*(6), 21623–21631.

Shouval, R., Hadanny, A., Shlomo, N., Iakobishvili, Z., Unger, R., Zahger, D., Alcalai, R., Atar, S., Gottlieb, S., Matetzky, S., Goldenberg, I., & Beigel, R. (2017). Machine learning for prediction of 30-day mortality after ST elevation myocardial infraction: An Acute Coronary Syndrome Israeli Survey data mining study. *International Journal of Cardiology*, *246*, 7–13. doi: 10.1016/j.ijcard.2017.05.067

Suresh, T., Assegie, T. A., Rajkumar, S., & Kumar, N. K. (2022). A hybrid approach to medical decision-making: diagnosis of heart disease with machine-learning model. *Int J Elec Comp Eng*, *12*(2), 1831–1838.

Thanh Noi, P., & Kappas, M. (2017). Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors (Basel, Switzerland)*, *18*(1). doi: 10.3390/s18010018

Thomas, J., & Princy, R. T. (2016). Human heart disease prediction system using data mining techniques. *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, 1–5.

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can Machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE*, *12*(4). doi: 10.1371/journal.pone.0174944

Wu, C.-C., Yeh, W.-C., Hsu, W.-D., Islam, M. M., Nguyen, P. A. A., Poly, T. N., Wang, Y.-C., Yang, H.-C., & Li, Y.-C. J. (2019). Prediction of fatty liver disease using machine learning algorithms. *Computer Methods and Programs in Biomedicine*, *170*, 23–29.