Journal of Big Data

RESEARCH

Open Access

Advanced machine learning techniques for cardiovascular disease early detection and diagnosis

Nadiah A. Baghdadi[1], Sally Mohammed Farghaly Abdelaliem[1*], Amer Malki[2], Ibrahim Gad[3], Ashraf Ewis[4,5] and Elsayed Atlam[2,3]

*Correspondence:
Smfarghaly@pnu.edu.sa

[1] Nursing Management and Education Department, College of Nursing, Princess Nourah bint Abdulrahman University, P.O. BOX 84428, Riyadh 11671, Saudi Arabia
[2] Computer Science Section, College of Computer Science and Engineering, Taibah University, Yanbu Campus, Al-Madinah, 46421, 41411 Yanbu, Saudi Arabia
[3] Computer Science Department, Faculty of Science, Tanta University, Tanta, Egypt
[4] Department of Public Health and Occupational Medicine, Faculty of Medicine, Minia University, El-Minia, Egypt
[5] Department of Public Health, Faculty of Health Sciences, AlQunfudah, Umm AlQura University, Meccah, Saudi Arabia

Abstract

The identification and prognosis of the potential for developing Cardiovascular Diseases (CVD) in healthy individuals is a vital aspect of disease management. Accessing the comprehensive health data on CVD currently available within hospital databases holds significant potential for the early detection and diagnosis of CVD, thereby positively impacting disease outcomes. Therefore, the incorporation of machine learning methods holds significant promise in the advancement of clinical practice for the management of Cardiovascular Diseases (CVDs). By providing a means to develop evidence-based clinical guidelines and management algorithms, these techniques can eliminate the need for costly and extensive clinical and laboratory investigations, reducing the associated financial burden on patients and the healthcare system. In order to optimize early prediction and intervention for CVDs, this study proposes the development of novel, robust, effective, and efficient machine learning algorithms, specifically designed for the automatic selection of key features and the detection of early-stage heart disease. The proposed Catboost model yields an F1-score of about 92.3% and an average accuracy of 90.94%. Therefore, Compared to many other existing state-of-art approaches, it successfully achieved and maximized classification performance with higher percentages of accuracy and precision.

Keywords: Heart disease, Machine learning, Feature selection, Cardiovascular diseases, Quality of life, Disease prevention, CVD

Introduction

The heart is the second-most important organ in the human body, after the brain. The heart's confusion eventually results in body turmoil. We are living in the modern era, and the world around us is undergoing significant transformations that have some impact on our day-to-day lives. Heart disease, which is claiming lives around the world, is one of the leading ailments among the top five deadly diseases [1]. Because it enables us to take the necessary steps at the right time, forecasting this disease is of the utmost importance.

Cardiovascular Diseases (CVD) are a group of heterogeneous diseases that affect the heart and circulatory system causing a variety of ailments that are typically brought on by atherosclerosis. Typically, CVD are chronic in nature and progressively manifest over time without symptoms for long periods of time before becoming advanced and showing up as symptoms of different intensity [2–4]. According to reports from the World Health Organization (WHO), CVD has been the leading cause of premature death in the world for decades, and it is expected that by 2030, CVD will be responsible for the deaths of around 23.6 million people annually.

In addition, the cost of treating cardiovascular disease and it's future consequences and early death, as measured by Disability Adjusted Life Years ("DALYS"), entails a significant economic burden [5–7]. Many factors contribute variably to the development of cardiovascular disease; these factors can be classed as modifiable and non-modifiable risk factors [5, 8]. Age, gender, and inherited variables are factors that cannot be modified. However, the other category of concerns, referred to as modifiable risk factors, comprises fasting blood sugar, high blood pressure, serum cholesterol, smoking, dietary propensity, obesity, and physical inactivity [9, 10].

Individuals will be able to avoid the development of CVD by identifying modifiable risk factors and attempting to alter lifestyle-related risk factors into healthy ones. Chest discomfort, arm pain, slowness and dizziness, weariness, and perspiration are among the early warning signs of a heart attack [11]. Individuals will be able to prevent the progression of CVD by identifying modifiable risk factors and attempting to alter lifestyle-related risk factors into healthy ones. Patients with heart disease do not have symptoms in the early stages of the disease, but they do in later stages when it is sometimes too late to manage or treat [12–14]. Therefore, despite the difficulty, Rapid recognition and prediction of CVD hypersensitivity in it seem healthy people is essential in assessing prognosis and prognosis. For early diagnosis of CVD, it will be incredibly beneficial and necessary to analyze the current significant CVD-health information contained in the huge database of hospital records. Thus, machine learning algorithms and other techniques for intelligent systems are beneficial in this field, and their findings are reliable and accurate [15–17].

The field of machine learning enables the identification of concealed patterns and the establishment of analytical structures, including clustering, classifications, regression, and correlations, through the integration and application of various techniques, such as machine learning models, neural networks, and information retrieval [18–20]. Consequently, machine learning techniques have demonstrated great potential to support clinical decision-making, aid in the development of clinical guidelines and management algorithms, and promote the establishment of evidence-based clinical practices for the management of Cardiovascular Diseases (CVDs) [21–27]. Furthermore, the early detection of CVDs using machine learning techniques can reduce the need for extensive and expensive clinical and laboratory investigations, resulting in a reduction of the financial burden on both the healthcare system and individuals [28, 29].

Cardiovascular disease is a chronic syndrome that can result in heart failure, a critical condition characterized by impaired heart function, and symptoms such as compromised blood vessel function and infarction of the coronary artery [30]. According to the American Heart Association (World Health Organization, 2021), cardiovascular

diseases are a set of heart and blood vessel abnormalities and one of the main causes of death worldwide. Accounting to statistic of almost 18 million deaths, cardiovascular disease was responsible for 32% of all deaths all over the world [31]. Heart attacks and strokes accounted for 85% of all deaths, with 38% occurring in individuals younger than 70. In the treatment and management of cardiovascular disorders, early detection is crucial, and machine learning (ML) can be a useful tool for recognizing a probable heart disease diagnosis [17, 32].

Heart disease as well identified as cardiovascular disease is a leading cause of death worldwide. The cardiac muscle is responsible for the circulation of blood around the body [33]. Although machine learning methods have demonstrated intriguing results in forecasting certain medical disorders, they have not been applied to the prediction of individual CVD survival in hypertensive patients utilizing routinely obtained big digital electronic administrative health data [34]. If a machine learning algorithm can be used to exploit the large administrative data set, it may be attainable to optimize the use of accumulated data sets to support in predicting patient outcomes, planning individualized patient care, monitoring resource utilization, and improving institutional performance. Comorbidity status, demographic information, laboratory test results, and medication information would improve prognostic evaluation and direct treatment decisions for hypertension patients [35].

In this study, we proposed a Gradient Boosting model to predict the existence of cardiovascular disease and to identify the most predictive value based on their Rough sets values. Afterward, a number of Machine Learning and Deep Learning techniques are used to analyze cardiovascular disease. Below are the main contributions of this study:

- Utilizing cross-validation and split validation, discover a machine learning algorithm with improved performance that will be applied to the detection of cardiovascular disease.
- The application of an appropriate feature selection technique can optimize prediction accuracy. Utilizing a robust machine learning algorithm can enhance early prediction of Cardiovascular Disease (CVD) development in its early stages, facilitating early intervention and promoting the selection of key features to support recovery algorithms.
- Predicting cardiovascular disease using a broadly cutting-edge Cardiovascular Diseases dataset.
- Providing reliable advise to health and medical specialists regarding significant changes in the healthcare sector.

  Sect. 2 of this paper presents related work. Section 3 proposes a methodology. Section 4 describes experimental evaluation. Section 5 analyzes discussion and comparative results. Section 6 focuses in conclusion and future work.

Related work

Many researchers examine a number of cardiac disease expectation frameworks utilizing various data mining techniques. They utilizing datasets and various calculations, in addition to test findings and future work that would be possible on the framework, and achieving more productive results. Researchers completed numerous research attempts

to accomplish efficient techniques and high accuracy in recognizing disorders associated with the heart.

Pattekari [36] study creating a model using the Naive Bayesian data mining presentation method. It's a computer program in which the user answers predetermined questions. It pulls hidden information from a dataset and evaluates client values to a preset data set. It can provide answers to difficult questions regarding heart disease diagnosis, allowing medical service providers to make more informed clinical decisions than normal choice emotionally supporting networks. It also helps reduce treatment expenses by providing effective treatments.

Tran [37] study built an Intelligent System using the Naive Bayes data mining modeling technique. It is a web application in which the user answers pre-programmed questions. It tries to find a database for hidden information and compares user values to a trained data set. It can provide answers to difficult questions about cardiac disease diagnosis, allowing healthcare professionals to make more informed clinical decisions than traditional decision support systems. It also lowers treatment costs by delivering effective care.

Gnaneswar [38] demonstrates the significance of monitoring the heart rate when cycling. Cyclists can cope with cycling meetings, such as cycling rhythm, to identify the level of activity by monitoring their pulse while accelerating. By managing their pedaling exertion, cyclists can avoid overtraining and cardiac failure. The cyclist's pulse can be used to determine the intensity of an exercise. Pulse can be measured using a sensor that can be worn. Unfortunately, the sensor does not capture all information at regular intervals, such as one second, two seconds, etc. Consequently, we will need a pulse expectation model to fill in the gaps.

Gnaneswar [38] work aims to use a Feedforward Brain Organization to construct a predictive model for pulse in consideration of cycling rhythm. On the second, pulse and rhythm are the data sources. The result is the predicted pulse for the following second. Using a feed-forward brain structure, the relationship between pulse and bicycle rhythm is represented statistically. Mutijarsa [39] expand of medical care administrations, based on these arguments. Numerous breakthroughs in remote communication have been made in anticipation of cardiac sickness. Utilizing data mining (DM) techniques for the detection and localization of coronary disease is highly useful. In their assessment, a comparative analysis of multiple single- and mixed-breed information mining calculations is conducted to determine which computation most accurately predicts coronary disease.

Yeshvendra [40] argues that the use of AI computations in the forecasting of various diseases is growing. This notion is so significant and diverse because of the ability of an AI computation to have a comparable perspective as a human for improving the accuracy of coronary disease prognosis. Patil [41] notes that a proper diagnosis of cardiac disease is one of the most fundamental biomedical concerns that must be addressed. Three information mining techniques: support vector machine, naïve bayes, and Decision tree. These techniques were used to create an emotionally supportive network for their preferred option. Tripoliti [42] argues that the identification of diseases with large prevalence rates, such as Alzheimer's, Parkinson's, diabetes, breast cancer, and coronary disease, is one of the most fundamental biomedical tests demanding immediate

attention. Gonsalves [43] attempted to forecast coronary CVD using machine learning and historical medical data. Oikonomou [44] provides an overview of the varieties of information encountered in chronic disease settings. Using multiple machine learning methods, they elucidated the extreme value theory in order to better measure chronic disease severity and risk.

According to Ibrahim [45], machine learning-based systems can be utilized for predicting and diagnosing heart disease. Active learning (AL) methods enhance the accuracy of classification by integrating user-expert system feedback with sparsely labeled data. Furthermore, Pratiyush et al. [46] explored the role of ensemble classifiers over the XAI framework in predicting heart disease from CVD datasets. The proposed work employed a dataset comprising 303 instances and 14 attributes, with categorical, integer, and real type attribute characteristics, and the classification task was based on classification techniques such as KNN, SVM, naive Bayes, AdaBoost, bagging and LR.

The literature attempted to create strategies for predicting cardiac disease diagnosis. Because of the high dimensionality of textual input, many traditional machine learning algorithms fail to incorporate it into the prediction process at the same time [47–53]. As a result, this paper investigates and develops a set of robust machine learning algorithms for improving the early prediction of CVD development, allowing for prompt intervention and recovery.

Methodology

This section describes the suggested classification scheme for heart disease instances. Initially, exploratory analysis is conducted. A comprehensive analysis is undertaken on both the target and the features, and category variables are converted to numeric values. Various criteria are utilized to compare models under consideration. The outputs of each model are analyzed, and the optimal model for the problem at hand is selected. The proposed model is thoroughly examined, and the Optuna library is used to tweak the model hyperparameters to see how much they have been enhanced. The suggested model is divided into three phases: (1) pre-processing, (2) Training, and (3) classification as shown in Fig. 1. In the following sections, the Authors will examine each of these components in further depth.

Pre-processing

Before training the selected models, it is important to address the Cholesterol missing values that were initially input as 0. To accomplish this, the data is separated into groups

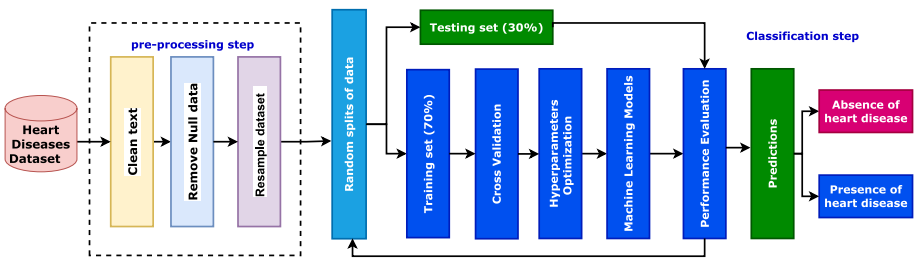

Fig. 1 The main steps of the proposed methodology

based on the presence of a verified cardiac condition, and the mean of each group is used to fill in the missing values. To assess whether these variables are influential in predicting heart disease based on their Shapley Values, interaction terms were included to the models to capture any possible correlations between the data elements. SHAP (SHapley Additive exPlanations) employs game theory to identify the significance of each characteristic and can be used to explain both individual model predictions and aggregated model results. SHAP determines the magnitude of each predictor's contribution to the model's output by averaging the marginal contributions of each feature over all feasible feature combinations.

Before doing feature selection using Shapley values, a gradient boosting model containing all variables is trained. The final predictors will be selected from the characteristics with a Shapley value greater than 0.1 that contribute significantly to the model's prediction. Then, these predictors will be used to establish the most effective model. Due to the multicollinearity between the interaction variables, a variety of nonparametric tree-based methods for predicting the risk of CVD are explored to discover the best accurate method.

Training process

The machine learning algorithm will be correctly trained after preprocessing and normalizing the datasets. Following the modification of the data, it is arbitrarily categorized into a training set and a test set, with 70% of the rows assigned to the training set and 30% to the test set. The k-fold is a common cross-validation method that entails running a large number of pertinent tests to determine the model's typical accuracy metric. This technique has existed for quite some time. To examine the proposed strategy, such AI procedures as SVC [54], MultinomialNB [55], K-Neighbor [56], BernoulliNB [55], SGD [57], Random forest [58] and Decision tree [59] are deployed for best terms of result.

XGBoost (Extreme Gradient Boosting) is a supervised learning method for improving prediction accuracy by combining multiple decision trees. XGBoost iteratively adds decision trees using gradient boosting, with each subsequent tree attempting to correct the errors of the previous trees. The final prediction is the weighted sum of all the individual tree predictions. XGBoost's objective function includes a loss function as well as a regularization term, which helps to prevent overfitting. The XGBoost objective function equation is:

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{1}$$

where l is the loss function, y_i is the true label for example i, $\hat{y}_i^{(t-1)}$ is the predicted value from the previous iteration, $f_t(x_i)$ is the prediction of the t^{th} tree for example i, and $\Omega(f_t)$ is the regularization term.

AdaBoost (Adaptive Boosting) is another boosting algorithm that also uses decision trees as weak learners. AdaBoost assigns weights to each training example, with higher weights given to examples that were misclassified by the previous weak learner. In each subsequent iteration, a new decision tree is trained on the weighted data, with the weights updated based on the accuracy of the tree. The final prediction

is the weighted sum of the predictions of all the individual trees. The equation for the prediction function of AdaBoost is:

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{2}$$

where T is the total number of trees, $h_t(x)$ is the prediction of the tth tree for input x, and α_t is the weight assigned to the tth tree.

Linear Support Vector Classifier (SVC) focus employs a straight-bit capacity to order data and operates superbly with enormous datasets [54]. The Linear SVC has more restrictions, such as standardization of consequence and misfortune work. Due to the fact that direct SVC is dependent on the bit strategy, the part strategy cannot be modified. A Direct SVC is meant to handle the data by returning the "best fit" hyper-plane that partitions or sorts it. After acquiring the hyperplane, the highlights are placed within the classifier, which predicts which class they belong to.

The Naive Bayes algorithm assigns equal weight to all features or qualities. The algorithm becomes more efficient as one property has no effect on another. According to Yasin 2020, the Naive Bayes classifier (NBC) is a simple, effective, and well-known text categorization algorithm. NBC has used the Bayes theorem to classify documents since the 1950 s, and it is theoretically sound. A posterior estimate is used to determine the class using the Naive Bayes classifier. Characteristics, for example, are categorized based on their highest conditional potential.

Bernoulli Naive Bayes is a statistical technique that produces boolean results based on the presence or absence of required text. The discrete Bernoulli Distribution is fed into this classifier. When identifying an unwanted keyword or tagging a specific word type within a text, this type of Naive Bayes classifier is useful. It is also distinct from the multinomial approach in that it generates binary output such as 1–0, True-False, or Yes–No. A stochastic system or procedure is one that has a random fit solution as part of it. Stochastic Gradient Descent (SGD) randomizes a few data samples rather than the entire dataset in each iteration. As a consequence, rather than calculating the sum of the gradients for all instances, each iteration calculates the gradient of the cost function for a single example. SGD is a method for determining the optimal smoothness properties of a differentiable or sub-differentiable objective function that is iterative.

Decision Tree is a widely known Machine Learning technique in which data is repeatedly partitioned based on specific parameters. The tree has two traversable entities: nodes and leaves. Leaves represent decisions or outcomes, whereas decision nodes partition data [59]. Decision trees can be used in combination to solve problems (ensemble learning). The Random Forest algorithm resolves the overfitting issues associated with decision tree algorithms. The algorithm is capable of dealing with regression and classification problems, as well as evaluating a large number of attributes to determine which ones are most important. Random data can learn without well-planned data alterations [58].

The K-Nearest Neighbor (K-NN) algorithm classifies new observations based on their distances from known examples. Based on the majority vote of its neighbors

and a distance function as a measuring tool, the case is designated to the class with the highest frequency among its k-nearest neighbors. In classification problems, k-NN returns the class membership. Whereas, in regression problems, it returns the object's property value. Whether k-NN is used for classification or regression has an effect on the output. Because this method relies on distance for classification, normalization can dramatically improve the training data. If the features correspond to different physical units or scales, standardization can significantly enhance the accuracy of the training data [56].

Classification

The proposed model is based on machine learning with strong generalization capabilities and a high degree of paradigm-specific precision. In this study, we will evaluate a number of machine learning algorithms and establish objectively which one delivers the greatest results. This is the primary purpose for the usage of machine learning: to combat the problem of overfitting that happens in machine learning. The curriculum also includes a structural concept of risk minimization. Machine learning can run best-described classes, particularly in higher-dimensional space, and to suggest a hyper-plane with the largest possible separation. In this stage, labeling data is used as an input, and the most significant characteristics are extracted using a feature extraction process. Finally, the optimal model is used to categorize new instances of data.

Experimental evaluation

In the experiments of the study, we utilized Google Colab as the implementation platform for machine learning models. The platform includes a virtual machine that runs on Google's servers and gives users access to a Python environment that includes popular data science libraries like TensorFlow, PyTorch, and Scikit-Learn. Google Colab is a cloud-based Jupyter notebook environment that offers free access to computing resources such as a virtual machine with 12 GB of RAM and up to 100 GB of hard disk space. The memory size allocated to the virtual machine is up to 25 GB, and it is also possible to enable high-RAM options up to 52 GB for large-scale models or data. The virtual machine runs on Google's servers and is equipped with NVIDIA Tesla K80 GPU, enabling us to train deep learning models efficiently. Additionally, Google Colab provides a wide range of preinstalled libraries and tools, making it easy to install and use the necessary dependencies. The virtual machine is powered by a Linux-based operating system, ensuring that the implementation environment is stable and reliable. Also, the operating system used by the virtual machine is Linux Ubuntu, which comes preinstalled with various system libraries and tools commonly used in data science projects.

The following subsection discussed the dataset and the results of the machine learning models.

Data collection

The Heart Condition data utilized in this study is a synthesis of data sets from the UCI Machine Learning Repository and contains eleven features that can be used to forecast the existence of heart failure, a prevalent cardiovascular disease that significantly raises the probability of a CV-related mortality [60, 61]. The target variable is

Table 1 A sample of the Heart Failure Dataset

Age	Sex	Type chest pain	BP resting	Cholesterol	BS fasting	ECG resting	HR max	Angina exercise	Old peak	ST slope	Disease of heart
41	M	ATA	142	287	0	Nor1	173	N	0.0	Upper	0
48	F	NAP	162	182	0	Nor1	157	N	1.0	Flat1	1
38	M	ATA	132	273	0	ST	98	N	0.0	Upper	0
49	F	ASY	136	224	0	Nor1	109	Y	1.5	Flat1	1
53	M	NAP	152	185	0	Nor1	123	N	0.0	Upper	0

Table 2 Symptoms, signs and laboratory investigations of the dataset of the heart disease

Variable	Interpretation
Age	Patient's Age/year
Gender	Patient's Gender, Male/Female
Type of chest pain	Type of chest pain: i. TA: Typical Angina ii. ATA: Atypical Angina iii. NAP: Non-Anginal Pain iv. ASY: Asymptomatic
Resting blood pressure	Patient's Blood Pressure/mmHg.
Total Cholesterol	Patient's Cholesterol (mg/dl).
Blood Glucose level (Fasting)	Patient's fasting blood glucose level. i. glucose >120 mg/dL =1 ii. glucose below 120 mg/dL =0
ECG at rest	Electrocardiography (at rest): i. Normal ii. ST: ST segment and/or T wave abnormality iii. LVH: Probable or Definite Left Ventricular Hypertrophy
Heart Rate at Maximum	Maximum Heart Rate, heart beats per minute.
Angina on Exercising	Exercise-associated Angina, present /absent.
Old peak	Measure of ST Depression.
ST_Slope	Slope of Peak Exercise. i. Up: up sloping ii. Flat iii. Down: down sloping

Table 3 The different datasets used to create the dataset of the heart disease

Datasets	#Observations
Cleveland	303
Hungarian	294
Stalog (Heart) Data Set	270
Long Beach VA	200
Switzerland	123
Total	1190
Duplicated	272
Final dataset	918

a binary attribute that indicates a diagnosis of Heart Failure if HeartDisease is = 1 as illustrated in Table 1. Moreover, Table 2 presents the list of variables and the description of the features in the heart disease dataset.

The dataset was created by combining a diverse range of datasets that were previously available independently, and were not combined before [60, 61]. In this dataset, five heart datasets are combined over 11 common features which makes it the largest heart disease dataset accessible for research purposes. The specific datasets utilized in the curation of this composite dataset are shown in Table 3.

The Heart Disease dataset has 918 observations and 12 columns [60, 61]. Table 4 summaries the main statistics for the numeric features. It is clear that, the mean value of age is 53 and the maximum is 77 as shown in Table 4. Similarly, Table 5 presents

Table 4 Summary statistics of numeric variables

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
Count	918	918	918	918	918	918	918
Max	77	200	603	1	202	6.20	1
Min	28	0	0	0	60	-2.6	0
Mean	53.51	132.39	198.79	0.23	136.81	0.89	0.55
Std	9.43	18.51	109.38	0.42	25.46	1.06	0.49
25%	47	120	173.25	0	120	0	0
50%	54	130	223	0	138	0.60	1
75%	60	140	267	0	156	1.50	1

Table 5 Summary statistics of categorical variables

	Sex	TypeChestPain	ECGResting	AnginaExercise	ST_Slope
Count	920	920	920	920	920
Unique	2	3	4	2	4
Top	M	ASY	Normals	N	Flat1
Freq	735	486	562	557	470

Table 6 The proportion of Heart Disease

Variable	Value	Total patients	Proportion of heart disease
Sex	M	725	90.2%
	F	193	9.8%
ChestPainType	ASY	**496**	**77.2%**
	NAP	203	14.2%
	ATA	173	4.7%
	TA	46	3.9%
RestingECG	Normal	**552**	**56.1%**
	ST	178	23.0%
	LVH	188	20.9%
ExerciseAngina	Y	371	**62.2%**
	N	**547**	37.8%
ST_Slope	Flat	**460**	**75.0%**
	Up	395	15.4%
	Down	63	9.6%

bold numbers mean the highest frequency and percentage

the statistics of categorical attributes. From this table, the unique values in ChestPainType attribute are 4 and the top is "ASY".

Table 6 summaries the main details for the numeric features. It is clear that, the variable Sex has two main values male (M) and female (F) such that the proportion of Heart Disease for M is 90.2% and for F is 9.8%. Similarly, Table 6 presents the statistics of ChestPainType attribute, there are 4 values (ASY, NAP, ATA, and TA) and the most frequent is ASY of 77.2%.

Exploratory data analysis

Remarkably, the classifications in the heart disease attribute value are reasonably well-balanced. 508 of the 918 patients who participated in the study have been diagnosed with heart failure, while 410 have not. Patients with heart disease have a median age of 57, whereas those without heart disease have a typical age of 51. As illustrated in Fig. 2, around 63% of males have heart disease, whereas approximately 25% of females have been diagnosed with heart disease. A female has a chance of 25.91% having a Heart Disease. A male has a probability of 63.17% having a Heart Disease.

Figure 3 demonstrates the heart disease ranges for Age, Systolic Blood Pressure, Cholesterol, Heart Rate, and ST Segment Depression. The boxplot of heart disease patients fall between the ages of 51 and 62, as depicted by the Age boxplot. There are also a few younger outliers below the lower margin in this category. Non-cardiovascular disease-free individuals have an age range that is slightly more variable but more evenly distributed, and there are no outliers. The vast majority of patients falling into this category are quite young, with ages ranging from 43 to 57 [62].

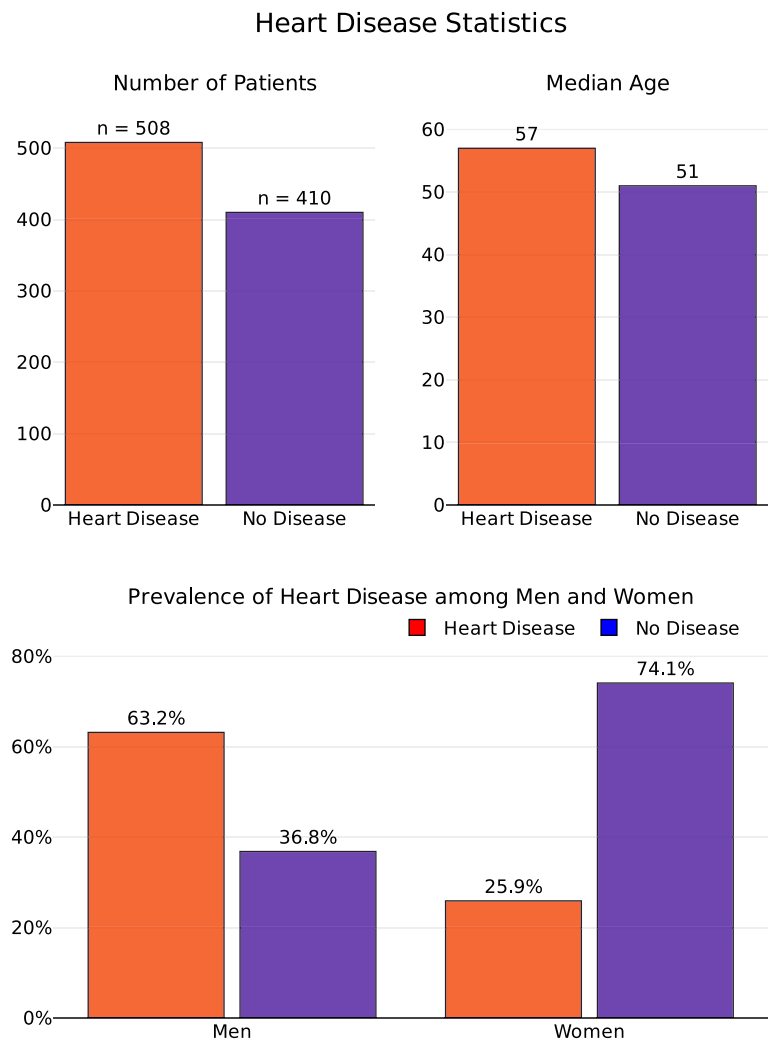

Fig. 2 Prevalence of heart disease among men and women

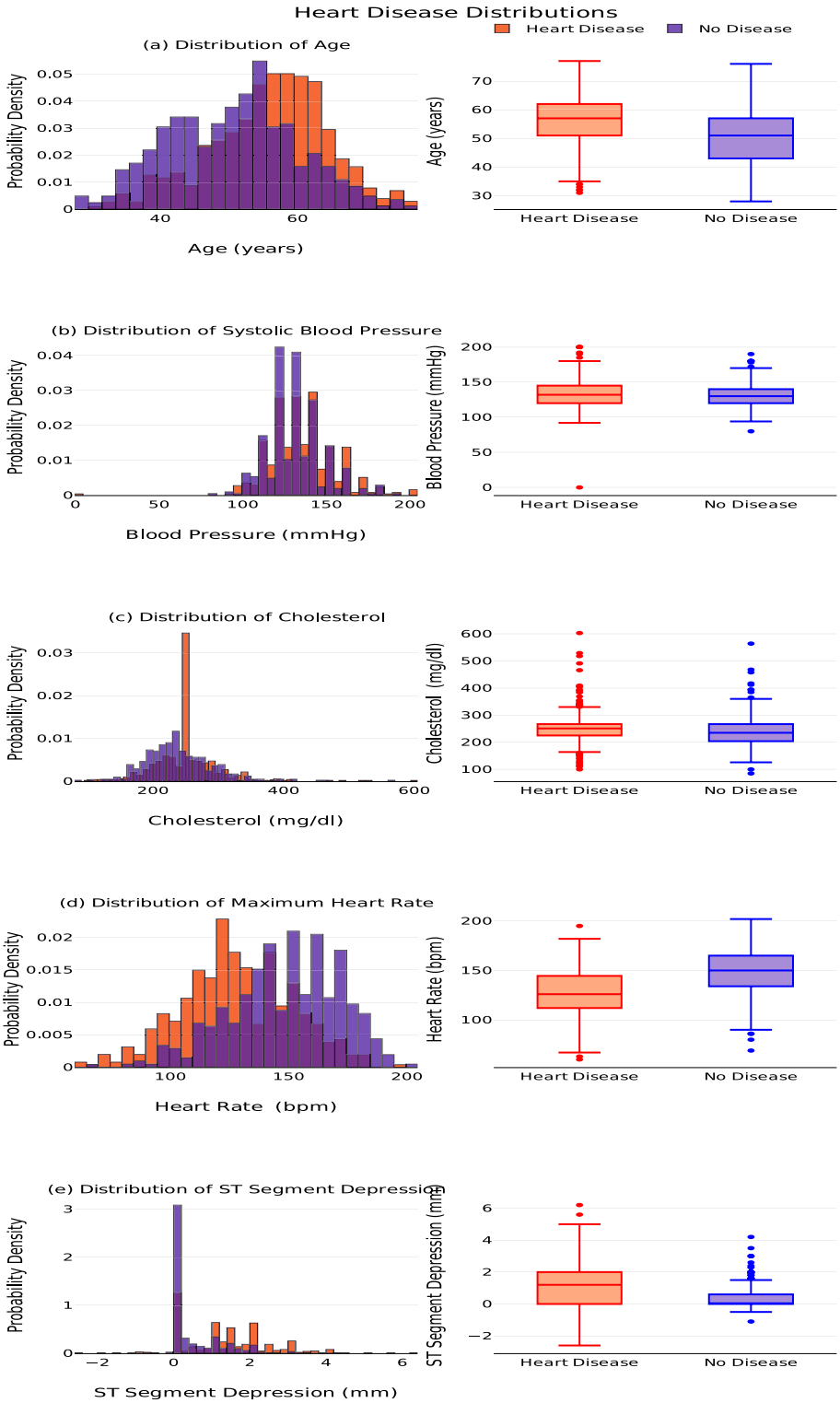

Fig. 3 The distributions of heart disease for age, systolic blood pressure, cholesterol, heart rate and ST segment depression

Furthermore, the boxplots between the groups for the Pulse Pressure Pressure variable are extremely similar. Both have upper and lower outliers, with the vast majority of patients' blood pressure falling between 120 and 145 mmHg. As demonstrated in Fig. 3, the median blood pressure in both groups is roughly 130 mmHg. Also, for the Cholesterol variable, the distribution of cholesterol appears to be skewed to the right, particularly among individuals with heart disease, where a substantial number of observations were reported with cholesterol values of 0. As illustrated in Fig. 3, those without heart illness have a median heart rate of 150 beats per minute, but those with heart disease have a median heart rate of 126 beats per minute.

In the case of the ST Segment Depression (OldPeak) variable, there is a variance between the distribution of ST segment depression groups. ST depression is more variable in patients with heart disease, with numerous larger outliers. The majority of these patients exhibit ST depressions between 0 and 2 mm, with a mean of 1.2 mm. In patients without heart disease, the range is narrower, between 0 and 0.6 mm, with a median ST depression of 0 mm, however the distribution of this group is more skewed overall, as illustrated in Fig. 3.

Figure 4 displays the correlation matrix associated with the heart disease dataset. heartdisease has the strongest positive link with OldPeak (correlation = 0.4) and the strongest negative association with MaxHR (correlation = −0.4), according to the correlation matrix. Age and MaxHR also have a reasonably high link, with a correlation of −0.38. As seen in Fig. 4, heart rate tends to decrease as age increases. Results observe a weak correlation between the numerical features and the target variable based on the matrix. Oldpeak (a depression-related number) correlates positively with heart disease. Heart disease is negatively correlated with maximal heart rate. Cholesterol has an interestingly negative association with heart disease.

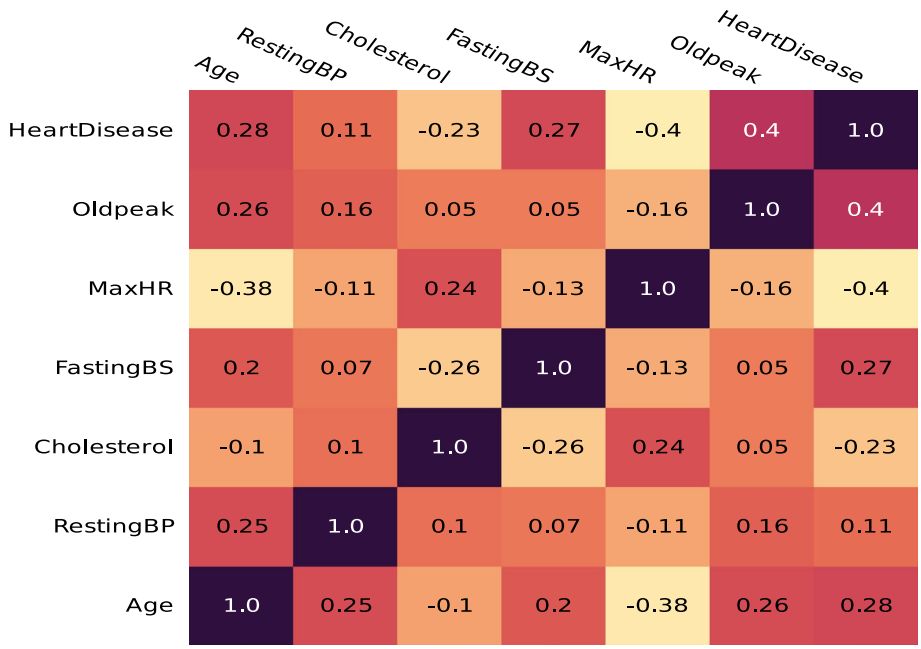

Fig. 4 The correlation matrix for the Heart Disease dataset

Figure 5 illustrates the correlation between heart disease and category variables. Nearly 80% of diabetic persons suffer heart problems. Patients with exercise-induced angina have an even greater incidence of cardiovascular disease, at over 85%. Over 65% of patients diagnosed with cardiac disease had ST-T wave abnormalities in their resting ECGs, the greatest percentage across the categories. Patients with a Flat or Declining ST Slope during exercise have the highest frequency of cardiovascular disease, at 82.8% and 77.8%, respectively.

Figure 6 explains data details regarding asymptomatic chest pain in heart disease at almost 77%, the absence of chest pain (asymptomatic) is the most prevalent symptom in patients with heart disease. In addition, heart disease is roughly nine times more prevalent in males than in females among patients with a cardiovascular diagnosis. A

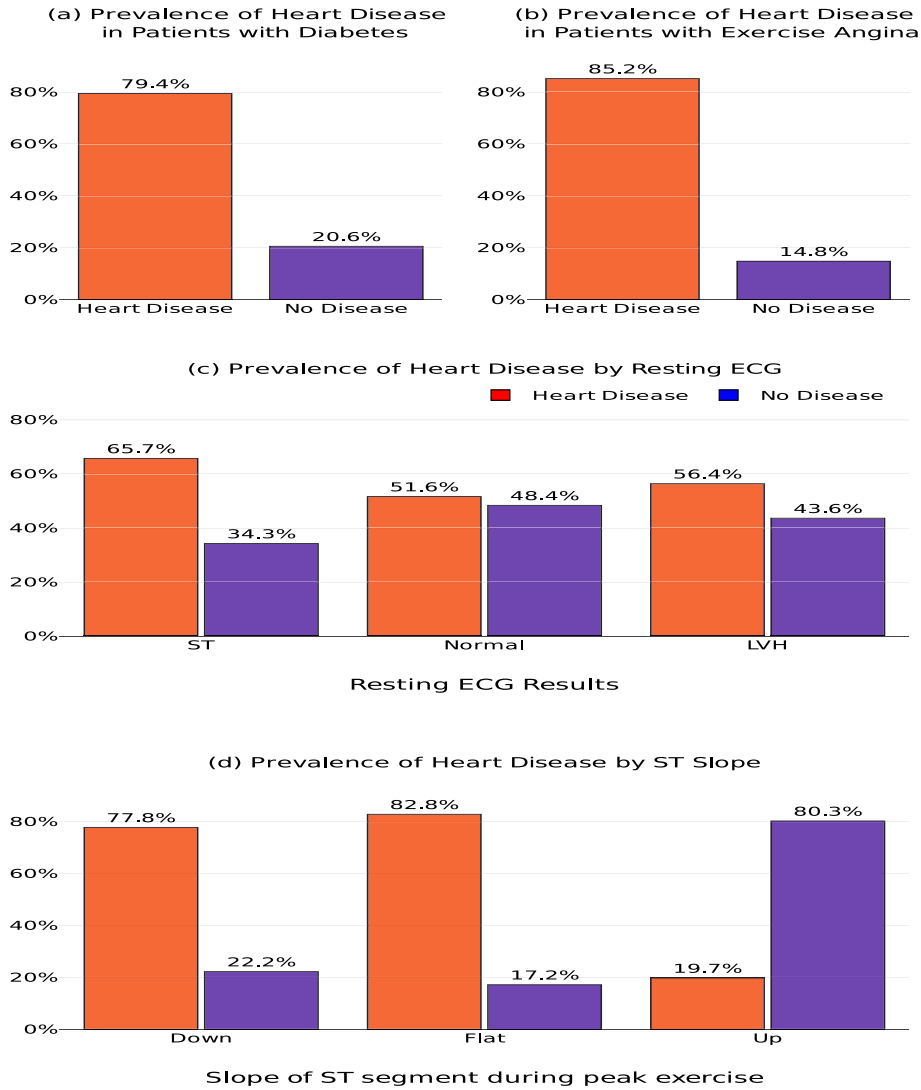

Fig. 5 Prevalence of heart disease by resting ECG. (a) Prevalence of Heart Disease in Patients with Diabetes. (b) Prevalence of Heart Disease in Patients with Exercise Angina. (c) Prevalence of Heart Disease by Resting ECG. (d) Prevalence of Heart Disease by ST Slope

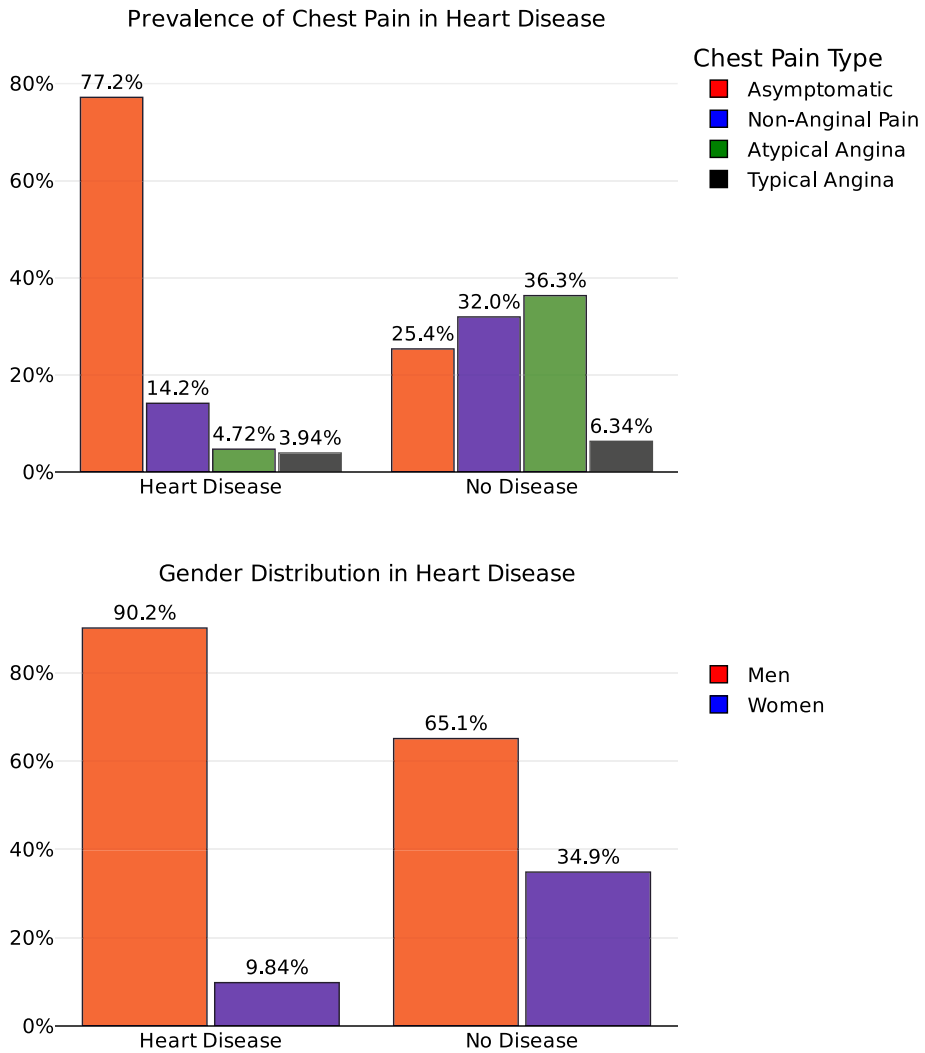

Fig. 6 Prevalence of chest pain in heart disease data

patient with asymptomatic chest pain (ASY) is approximately six times more likely to suffer heart disease than a patient with atypical angina chest pain (ATA).

Overall insights obtained from the exploratory data analysis, Data for the target variable are near to balanced. The association between numerical features and the target variable is weak. Oldpeak (a depression-related number) correlates positively with heart disease. Heart illness is negatively correlated with maximum heart rate. Interestingly, there is a negative link between cholesterol and heart disease. Males are approximately 2.44 times more likely to suffer from heart disease than females. There are distinct variances between the types of chest pain. Patients with asymptomatic chest pain (ASY) are about six times more likely to suffer heart disease than those with Atypical Angina chest pain (ATA). Resting ECG: electrocardiogram values at rest are comparable. Patients with ST-T pulse abnormalities have a higher risk of developing heart disease than those who do not. ExerciseAngina: people who have exercise-induced angina are nearly 2.4 times more likely to have heart disease than people who don't. The slope of the ST segment at

maximum exertion varies. ST Slope Up has a considerably lower risk of cardiovascular disease than the other two segments. Exercise-induced angina with a 'Yes' score is nearly 2.4 times more likely to result in heart disease than exercise-induced angina with a 'No' score.

Performance evaluation

When dealing with imbalanced datasets, classification accuracy alone may not be the most suitable performance metric. Therefore, authors often use additional performance metrics to address this issue [63]. The confusion matrix is frequently employed for expressing a classifier's classification results, with diagonal elements indicating correctly classified samples as positive or negative and off-diagonal elements indicating misclassification. As a consequence, performance improvement metrics such as accuracy, precision, recall (sensitivity), F1-score, and ROC curve are employed. F1-score accuracy, recall, and precision can be calculated using the Eqs. 3, 4, 5, and 6, respectively. These formulas are based on the numbers of False Positive (FP), False Negative (FN), True Positive (TP), and True Negative (TN) samples in the test dataset [64].

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{3}$$

$$F1 - score = 2\frac{precision \star recall}{precision + recall} \tag{4}$$

$$Recall = positivePredictivevalue = \frac{TP}{TP + FP} \tag{5}$$

$$Precision = \frac{TP}{TP + FN} \tag{6}$$

Machine learning models

Studies are carried out using the collected dataset, which has approximately 918 rows. The final version of the updated data was split into training and testing sets in order to fit the model, with 70% of the data used for the learning set and 30% for the testing set. Table 7 shows the shapes of three datasets: training, validation, and test. The training set has 504 rows and 19 columns, while the validation and test sets both have 207 rows and 19 columns. AdaBoost, Gradient Boost, Random Forest (RF), k-nearest neighbor (KNN), Support Vector Machine (SVM), and Decision tree classifiers are used in this study [64–66].

Table 7 Dataset shapes

Dataset	Shape
Training	(504, 19)
Validation	(207, 19)
Test	(207, 19)

In order to develop a robust classifier with high precision, it is vital to use an appropriate evaluation approach. One such method is the K-fold Cross-Validation, which generates diverse data samples to determine the average correctness of a model. The strategy of k-fold is a commonly used cross-validation technique, where a specified value of k is chosen, such as five, and the data is divided into k subsets of equal size. In each iteration, one of the k subsets is used as the test set, and the remaining $k-1$ subsets are used for model learning. This process is repeated until all subsets have been used as the test set once.

The k-fold cross-validation method employs computed values average as a performance metric. This approach provides a reliable estimate of the model's generalization ability, which is particularly useful when the data is limited and cannot be split into separate learning and testing sets.

Finally, the best hyperparameter values for each algorithm are determined through experimentation and optimization of the model, often through methods such as grid search and Bayesian optimization. The best hyperparameter values can serve as a starting point for developing new models or improving existing ones, as they provide insight into the values that have yielded the best performance for each algorithm. The results of hyper-parameter optimization of Machine learning models are shown in Table 8.

Table 8 presents the results of hyper-parameter optimization for four machine learning models: Extra Trees, Random Forest, AdaBoost, and Gradient Boosting. For each model, a range of hyper-parameters was explored using cross-validation, and the

Table 8 The results of hyper-parameter optimization of Machine learning models

Model	Parameters	Best parameters	Accuracy	AUC
Extra Trees	n_estimators: [100, 105, ..., 500], criterion :('gini', 'entropy'), max_depth: [5, 10, 15, 20], min_samples_split: [2, 4, 6], min_samples_leaf: [4, 5, 6]	criterion='entropy', max_depth=15, min_samples_leaf=4, n_estimators=300	84.54%	0.920
Random Forest	n_estimators: [100, 105, ..., 500], criterion :('gini', 'entropy'), max_depth: [3, 7, 14, 21], min_samples_split: [2, 5, 10], min_samples_leaf: [3, 5, 7], max_features: [None, 'sqrt'], max_leaf_nodes: [None, 5, 10, 15, 20], min_impurity_decrease': [0.001, 0.01, 0.05, 0.1], bootstrap: [True, False]	max_depth=14, max_features='sqrt', max_leaf_nodes=15, min_impurity_decrease=0.001, min_samples_leaf=3, min_samples_split=10, n_estimators=200	85.52%	0.924
AdaBoost	n_estimators: [100, 105, ..., 500], learning_rate: [0.25, 0.5, 0.75, 0.9]	learning_rate=0.25, n_estimators=100	84.06%	0.897
Gradient Boosting	boosting_type: ['gbdt', 'dart'], num_leaves: [20, 27, 34, ...,50], max_depth : [-1, 3, 7, 14, 21], learning_rate: [0.0001, 0.001, 0.01, 0.1, 0.5, 1], n_estimators': [100, 105, ..., 500], min_split_gain: [0.00001, 0.0001, 0.001, 0.01, 0.1], min_child_samples: [3, 5, 7], subsample: [0.5, 0.8, 0.95], colsample_bytree: [0.6, 0.75, 1]	boosting_type='dart', colsample_bytree=1, learning_rate=0.5, max_depth=3, min_child_samples=7, min_split_gain=1e-05, num_leaves=30, subsample=0.5	88.9%	0.925

best parameters were selected based on the highest accuracy and AUC scores. The accuracy and AUC scores were calculated using a hold-out test set.

For the Extra Trees model, the best parameters were found to be criterion='entropy', max_depth=15, min_samples_leaf=4, and n_estimators=300, resulting in an accuracy of 84.54% and an AUC score of 0.920. Similarly, for the Random Forest model, the best parameters were max_depth=14, max_features='sqrt', max_leaf_nodes=15, min_impurity_decrease=0.001, min_samples_leaf=3, min_samples_split=10, and n_estimators=200, resulting in an accuracy of 85.52% and an AUC score of 0.924.

The AdaBoost model achieved an accuracy of 84.06% and an AUC score of 0.897 with the best parameters of learning_rate=0.25 and n_estimators=100. Finally, the Gradient Boosting model achieved the highest accuracy of 88.9% and the highest AUC score of 0.925 with the best parameters of boosting_type='dart', colsample_bytree=1, learning_rate=0.5, max_depth=3, min_child_samples=7, min_split_gain=1e-05, num_leaves=30, and subsample=0.5. Overall, the results indicate that hyper-parameter optimization can significantly improve the performance of machine learning models, and the Gradient Boosting model performed the best on this particular dataset.

The results of the Chi-Squared test are presented in Table 9. Based on the p-values, which are less than 0.05, all discrete variables are included in the models as predictors.

The summary plot of Shapley values of feature importance in a machine learning model provides insights into the relative importance of different features in making predictions. The Shapley value is a concept from cooperative game theory that provides a way to allocate the main contribution of each feature to the final prediction. In a machine learning environment, the Shapley value of a feature represents the average contribution of that feature to the model output across all possible subsets of features. The calculation of Shapley values requires the evaluation of the model output for all possible subsets of features, which can be computationally expensive for high-dimensional datasets. However, there are several efficient algorithms for approximating the Shapley values, such as the KernelSHAP algorithm, which is based on sampling.

As shown in Fig. 7, the summary plot of Shapley values displays the top 20 predictors of heart disease in order of relevance. Each point on the graph represents a training set observation. When the points are to the right of the 0 lines, this suggests a greater risk of being diagnosed with heart disease, whereas points to the left of the 0 line indicate a lower likelihood. The values of each feature are represented by the color of the points, with light orange indicating high feature values and dark blue indicating low feature values. The shape of the points in each row is determined by the number of observations that overlap for that feature. Along with three independent features, Cholesterol,

Table 9 The results of Chi-Squared test

	Chi statistic	p-value
ExerciseAngina	222.26	0.00000
ChestPainType	268.07	0.00000
ST_Slope	355.92	0.00000
Sex	84.15	0.00000
RestingECG	10.93	0.00423

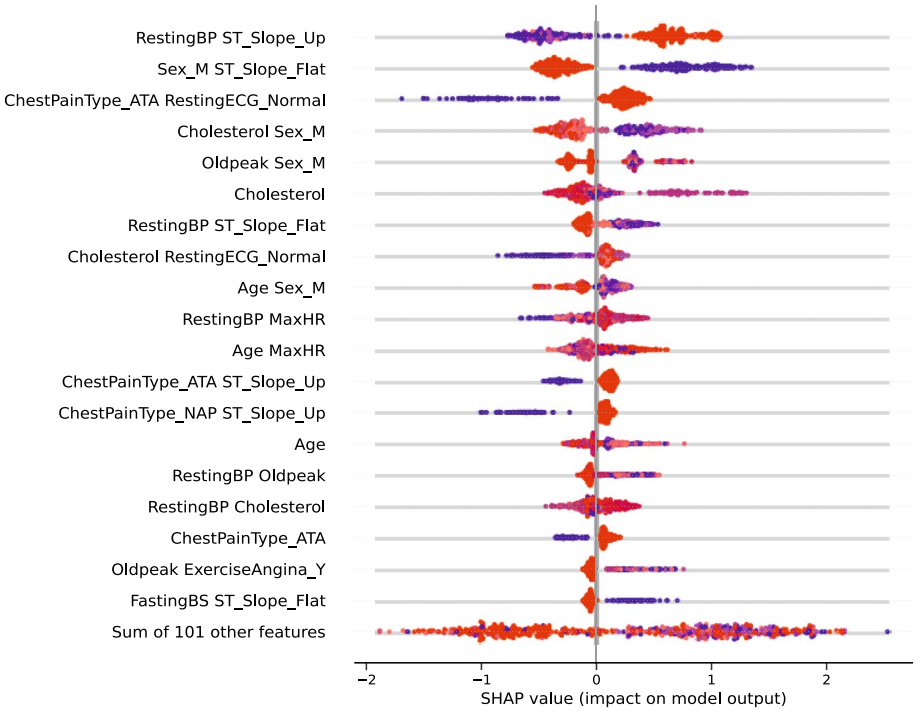

Fig. 7 The summary plot of Shapley values of features importance

Age, and typical chest pain, nearly all of the variables in the plot are interaction terms that were included to the model. The variable in the first row represents the interaction between SBP at Rest and ST Slope Up. People with an upward ST slope and high blood pressure have a lower risk of heart disease, according to the Shapley values. The Shapley values for the second variable, "Sex M ST slope flat," show that male patients with a flat ST Slope are more likely to develop cardiovascular disease. The fourth variable in the scatter plot, Cholesterol Sex M, indicates that men with high cholesterol are more likely to be diagnosed with cardiovascular disease.

In addition, the order of relevance in the summary plot is determined by the feature's average absolute Shapley value, which quantifies the average amount by which the characteristic affects the projected chance of heart disease. There are 18 features that contribute at least 0.1 on average to the model's prediction. Table 10 provides a listing of the final predictors chosen and the feature importance of each. "RestingBP ST Slope" appears in five of the top 19 of most significant predictors.

Model performance on the validation set

The ROC Curves (Fig. 8) illustrate the performance of the models at various thresholds. The y-axis indicates the True Positive Rate or Sensitivity of the models, which is a measure of how well the model identifies patients with heart disease (true positives), while the x-axis indicates the number of patients that the model incorrectly classifies as false positives. A model with a curve at the upper left corner of the graph, with a higher true positive rate and a lower false positive rate, shows a greater capacity to differentiate between the classes. On the test set, all of the models depicted in the above scatter plot produce strong results. Overall, Gradient Boosting has the highest Area Under the

Table 10 The list of the final predictors selected and their feature importance

Feature	Importance
Cholesterol ST_Slope_Flat	0.941
RestingBP ST_Slope_Up	0.569
Sex_M ST_Slope_Flat	0.512
ChestPainType_ATA RestingECG_Normal	0.341
Cholesterol Sex_M	0.319
Oldpeak Sex_M	0.260
Cholesterol	0.252
RestingBP ST_Slope_Flat	0.160
Cholesterol RestingECG_Normal	0.159
Age Sex_M	0.153
RestingBP MaxHR	0.151
Age MaxHR	0.150
ChestPainType_ATA ST_Slope_Up	0.146
ChestPainType_NAP ST_Slope_Up	0.144
Age	0.127
RestingBP Oldpeak	0.124
RestingBP Cholesterol	0.118
ChestPainType_ATA	0.115
Oldpeak ExerciseAngina_Y	0.100

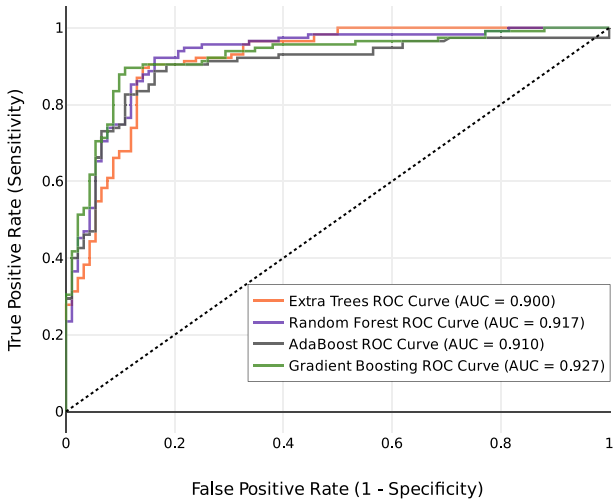

Fig. 8 ROC curve comparison on the test set

Curve at 0.927, but at specific thresholds, the Random Forest model offers somewhat superior results, since the curve surpasses that of Gradient Boosting.

By using Shapley features greater than 0.1, the Extra Trees classifier achieves an AUC of 0.89. After tweaking the model's hyperparameters, the classifier achieves an average accuracy of 88%, an F1-score of 89.5%, and a standard deviation of 6.7% on the validation set. With an AUC of 0.917, the Random Forest model outperforms the Extra Trees classifier across all three criteria. On the validation set, the model achieves an average precision of 88.7% and an F1-score of almost 90%. Evidently, there is a minor performance reduction in the AdaBoost model. The AUC declined to 0.91, and the overall accuracy

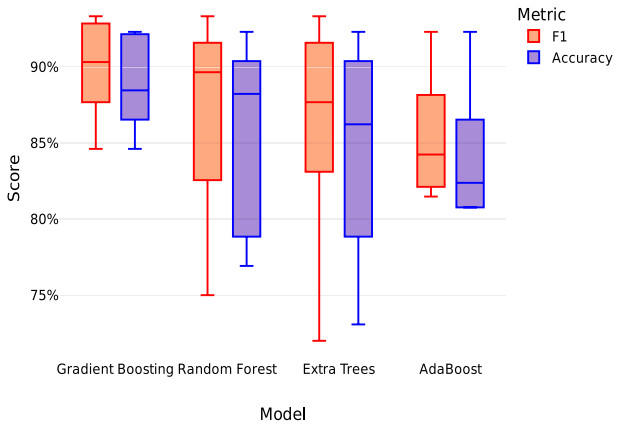

Fig. 9 Model performance on the validation set

Table 11 Classification report for Catboost_tuned model

	Precision	Recall	F1-score	Support
0	0.88	0.90	0.89	112
1	0.93	0.91	0.92	164
Accuracy			0.91	276
Macro avg	0.90	0.91	0.91	276
Weighted avg	0.91	0.91	0.91	276

and F1-score fell to 86.5% and 88% respectively. Despite the fact that the model yields a smaller standard deviation than the others. At 0.927, the Gradient Boosting model has the greatest Area Under the Curve among the classifiers. In addition, the model improves the validation set's accuracy to about 87% and the F1-score to 89%.

Comparing the cross-validation results shown in the boxplots of Fig. 9, it is clear that the Gradient Boosting model has the highest median F1-score of 90.3% and the highest median accuracy of 88.5%. It also has the smallest standard deviation of the distribution, at around 3. The Random Forest model comes in a close second with a median F1-score of 89.7% and a median accuracy of 88.2%, albeit with slightly greater score variability.

The recall, precision, and accuracy values for the Catboost model are shown in Table 11. Catboost is a model that can determine whether a patient has heart disease. Furthermore, the Catboost algorithm is biased because it is extremely sensitive to major class. When calculating the comprehensive performance measurement, the F1-score is also used to compare the algorithm's precision. Classification of heart disease was significantly improved by the Catboost method. The model achieved 93% accuracy in the "heart illness" category and 88% accuracy in the "No disease" category. The accuracy of the Catboost classification model was 91%.

Table 12 illustrates the classification results of the various classifiers on the dataset. The table reports the performance of various classifiers on a given dataset, measured in terms of Accuracy, Precision, Recall, and F1 score. Comparing the results of the proposed technique against those of other classifiers such as SVM [54], XGBoost, AdaBoost, RandomForest [58], LinearDiscriminant [67], LightGBM, GradientBoosting,

Table 12 Comparative results on the Dataset using ML

Classifier	Accuracy	Precision	Recall	F1
XGBoost	0.8297	0.8980	0.8049	0.8489
AdaBoost	0.8659	0.9262	0.8415	0.8818
LinearDiscriminant	0.8696	0.9156	0.8598	0.8868
LightGBM	0.8732	0.9057	0.8780	0.8916
GradientBoosting	0.8768	0.9276	0.8598	0.8924
Catboost	0.8804	0.9226	0.8720	0.8966
ExtraTree	0.8804	0.9281	0.8659	0.8959
KNeighbors	0.8841	0.9074	0.8963	0.9018
SVM	0.8841	0.8976	0.9085	0.9030
LogisticRegression	0.8841	0.9231	0.8780	0.9000
RandomForest	0.8877	0.9236	0.8841	0.9034
Catboost_tuned	0.9094	0.9317	0.9146	0.9231

Catboost, ExtraTree, KNeighbors [56], and LogisticRegression [68] demonstrates the method's utility. The results of classifiers according to various metrics are displayed. The highest performing classifier based on all measures is Catboost_tuned, which achieved an accuracy of 0.9094, a precision of 0.9317, a recall of 0.9146, and F1 score of 0.9231. Other top-performing classifiers include RandomForest, LogisticRegression, SVM, and KNeighbors, with similar accuracy and precision scores, but slightly lower recall and F1 scores. In contrast, lower-performing classifiers such as XGBoost and AdaBoost exhibit moderate accuracy and precision scores, but relatively lower recall and F1 scores. Overall, the results suggest that the choice of classifier can have a significant affect on the performance of a predictive model.

The present study employs a confusion matrix (Fig. 10) to report the performance of models in accurately predicting cardiac disease for a given set of patients, with due consideration to both correctly classified and misclassified instances. Specifically, the Gradient Boosting model is found to exhibit the highest proportion of True Positives (TP) and True Negatives (TN) when evaluated on a test set. The computation of FN, FB, TN, and TP, values for the cardiac disease class is carried out using the Gradient Boost model, whereby the predicted values are expected to match the actual values. For instance, TP corresponds to the value at cell 1 of the confusion matrix, while FN is computed by adding the relevant row values, excluding TP (i.e., FN = 12). Similarly, FP is calculated as the total of column values, excluding TP, leading to a value of 11. Lastly, TN is determined by the combination of all columns and rows except the class under consideration (i.e., cardiac disease), which yields a value of 81.

Discussion

Despite the vast amount of data produced by healthcare systems, medicine faces unique obstacles in comparison to other data-driven businesses where machine learning has flourished. The Health Insurance Portability and Accountability Act (HIPAA) mandates strict, center-specific Institutional Review Boards (IRBs) to govern the usage of patient data. This significantly preserves patient privacy, but it has unwittingly created data silos across the nation [47]. Consequently, the majority of

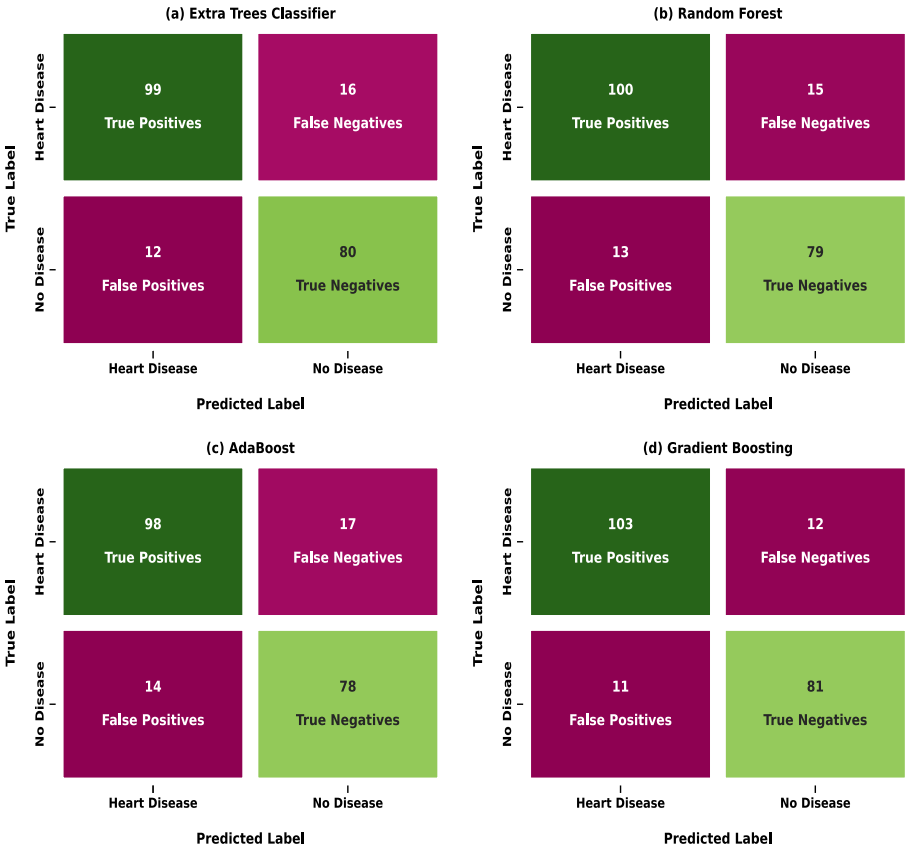

Fig. 10 The confusion matrix results for Extra Trees, RandomForest, AdaBoost, and GradientBoost classifiers

published healthcare machine learning models rely on locally acquired datasets and lack external validation. 58% of cardiovascular prediction models, according to the Tufts predictive analytics and comparative effectiveness cardiovascular prediction model have never been externally verified [69]. Heart-related disorders are one of the leading causes of deaths and morbidity on a global scale [5–7].

It is common for those with heart disease to be unaware of their condition, and it is difficult to predict their health condition and diagnose their disease in its early stages in order to save their lives, minimize their complications and suffering, and reduce the global burden of disease and mortality [9]. Machine learning models are capable of accomplishing this difficult task and can be of tremendous assistance in the early diagnosis and prediction of heart disorders [12–14]. Medical machine learning offers a vast array of opportunities, including the discovery of hidden patterns that can be utilized to generate diagnostic accuracy on any medical dataset.

Previous research has demonstrated that machine learning can aid in the prediction of cardiovascular illness [15, 16]. For the diagnosis of cardiac disorders, this prior research employed various machine learning approaches, such as neural networks, Naive Bayes, Decision Tree, and SVM, and obtained varying degrees of accuracy [18, 19]. The accuracy of the proposed feature selection methodology algorithm (CFS+Filter Subset Eval), a hybrid method that combines CFS and Bayes theorem,

was 85.5%, according to [70]. Shouman et al. [71] presented an integrated k-means clustering with the Naive Bayes approach for enhancing the accuracy of Naive Bayes in diagnosing patients with heart disease, with an accuracy of 84.5%. Using both Naive Bayesian Classification and Jelinek-Mercer smoothing techniques. Rupali et al. [72] developed decision support for the Heart Disease Prediction System (HDPS), with Laplacian smoothing for approximating important patterns in the data while avoiding noise; their accuracy was 86%.

Elma et al. [73] created a classifier for predicting heart illness that merged the distance-based approach K-nearest neighbor with a statistically-based NaiveBayes classifier (cNK) and achieved an 85.92% accuracy rate. Dulhare et al. [74] improved cardiac disease prediction methods using Naive Bayes and particle swarm optimization, attaining an accuracy of 87.91%.

To accurately predict CVDs in the present study, Shapley values were used to create a Gradient Boosting model with an Area Under the Curve of 0.927% for predicting the risk of a heart disease diagnosis. Using Shapley values, Authors discovered critical cardiac disease signs and their predictive power for a positive diagnosis. Interaction effects between a patient's medical information were some of the most relevant predictors in the model, particularly in features such as Age, Cholesterol, Blood Pressure, ST Slope, and Chest Pain kind. The proposed Catboost model offered the strongest results overall and can be utilized for the early identification and diagnosis of heart disease, with an overall F1-Score of 92.3% and an accuracy of 90.94%, when picking the optimal model. Overall, the proposed model is superior to earlier approaches for diagnosing cardiac disease.

However, this study is important but many Limitations exist. First, this research depends solely on secondary data using the available data at the selected cardiology and internal medicine departments. Hence, there were some missing data and some variables could not be included in the analysis. The cross-sectional design of the study is the second limitation that could not examine the longitudinal effects of the risk factors on the development of the CVDs.

The possible future orientation of this study is to improve prediction techniques by combining various machine learning techniques and increase the accuracy and precision of CVD prediction and early diagnosis, which has been shown to be superior to the majority of traditional state-of-the-art methods. Based on machine learning techniques, the suggested model for the prediction of heart disorders is a robust, effective, and efficient method for the prediction and early detection of heart ailments. It obtained and maximized classification performance with greater accuracy and precision percentages than other current models. One of the most significant outcomes of our proposed machine learning algorithms is that they achieved good accuracy while displaying fewer feature sets. This is crucial for clinical medical practice, which requires the most precise and straightforward methods for confirming a diagnosis in order to make a final therapeutic decision. Nonetheless, there are obstacles to the generality of the CVD prediction models reported in this study. Before being implemented into the clinical guidelines, the suggested machine learning algorithm must investigate different population datasets to minimize variation in CVD prevalence patterns and evaluate the possible impact on physicians' decision making or patient outcomes.

Conclusion

Prediction of cardiovascular diseases is crucial for assisting clinicians with early disease diagnosis. Instead of replacing clinicians, machine learning will be a supplement to the clinical portfolio, enhancing human-led decision-making and clinical practices. Furthermore, by using machine learning techniques, the cost of conducting a long list of expensive clinical and laboratory investigations will be eliminated, reducing the financial burden on patients and the healthcare system. This paper proposed new robust, effective, and efficient machine learning algorithms for predicting CVD based on symptoms, signs, and other patients' information from hospital records in order to improve the early prediction of CVD development in its early stages and to ensure early intervention with a warranted recovery. The new technique was more accurate and precise than existing standard art-of-state algorithms for the classification and prediction of heart disease. Future research evaluating the performance of the proposed machine learning algorithms on datasets containing a greater number of modifiable and non-modifiable risk factors will be crucial for the development of a more accurate and robust system for the prediction and early diagnosis of heart diseases.

Acknowledgement
Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R 293), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author contributions
All authors participated in the research idea, conceptualization, data collection, analysis and preparation of the manuscript for publication.

Funding
Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R293), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Availability of data and materials
The datasets generated and/or analyzed during the current study are not publicly available due to data privacy but are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 14 August 2022 Accepted: 30 August 2023
Published online: 17 September 2023

References

1. Javeed A, Rizvi SS, Zhou S, Riaz R, Khan SU, Kwon SJ. Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification. Mob Inf Syst. 2020;2020:1–11. https://doi.org/10.1155/2020/8843115.
2. Eckel R, Jakicic J, Ard JD. Aha/acc guideline on lifestyle management to reduce cardiovascular risk: a report of the american college of cardiology/american heart association task force on practice guidelines. American College of Cardiology/American Heart Association Task Force on Practice Guidelines. 2014. https://doi.org/10.1161/01.cir.0000437740.48606.d1.pmid:24222015.
3. Anderson KM, Wilson PW, Odell PM, Kannel WB. An updated coronary risk profile. A statement for health professionals. Circulation. 1991;83(1):356–62. https://doi.org/10.1161/01.cir.83.1.356.
4. Azmi J, Arif M, Nafis MT, Alam MA, Tanweer S, Wang G. A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data. Med Eng Phys. 2022;103825.

5. Day TE, Goldlust E. Cardiovascular disease risk profiles. Am Heart J. 2010;160(1):3. https://doi.org/10.1016/j.ahj.2010.04.019.
6. Alwan A. Global status report on noncommunicable diseases. World Health Organization, 2011;293–298.
7. ...Tsao CW, Aday AW, Almarzooq ZI, Anderson CAM, Arora P, Avery CL, Baker-Smith CM, Beaton AZ, Boehme AK, Buxton AE, Commodore-Mensah Y, Elkind MSV, Evenson KR, Eze-Nliam C, Fugar S, Generoso G, Heard DG, Hiremath S, Ho JE, Kalani R, Kazi DS, Ko D, Levine DA, Liu J, Ma J, Magnani JW, Michos ED, Mussolino ME, Navaneethan SD, Parikh NI, Poudel R, Rezk-Hanna M, Roth GA, Shah NS, St-Onge M-P, Thacker EL, Virani SS, Voeks JH, Wang N-Y, Wong ND, Wong SS, Yaffe K, Martin SS. Heart disease and stroke statistics-2023 update: a report from the American heart association. Circulation. 2023. https://doi.org/10.1161/CIR.0000000000001123.
8. Wilson P, DAgostino RB, Levy D, Belanger A, Silbershatz H, Kannel W. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97(12):1837–47. https://doi.org/10.1161/01.CIR.97.18.1837.
9. Mythili T, Mukherji D, Padalia N, Naidu A. A heart disease prediction model using svm-decision trees-logistic regression (sdl). Int J Comput Appl. 2013;68(16):11–5. https://doi.org/10.1161/01.CIR.97.18.1837.
10. Frieden TR, Jaffe MG. Saving 100 million lives by improving global treatment of hypertension and reducing cardiovascular disease risk factors. J Clin Hypertens. 2018;20(2):208.
11. Haissaguerre M, Derval N, Sacher F, Deisenhofer I, de Roy L, Pasquie J, Nogami A, Babuty D, Yli-Mayry S. Sudden cardiac arrest associated with early repolarization. N Engl J Med. 2008;58(19):2016–23.
12. Kumar PM, Lokesh S, Varatharajan R, Babu GC, Parthasarathy P. Cloud and iot based disease prediction and diagnosis system for healthcare using fuzzy neural classifier. Future Gener Comput Syst. 2018;68:527–34.
13. Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning technique. IEEE Access. 2019;7:81542–54.
14. Kwon JM, Lee Y, Lee S, Park J. Effective heart disease prediction using hybrid machine learning technique. J Am Heart Assoc. 2018;7(13):1–11.
15. Esfahani HA, Ghazanfari M, Ecardiovascular disease detection using a new ensemble classifier. in,. IEEE 4th international conference on knowledge-based engineering and innovation (KBEI). Tehran, Iran. 2017;2017:488–96.
16. Gandhi M, Singh SN. Cardiovascular disease detection using a new ensemble classifier. in 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), Greater Noida, India, 2015;520–525
17. Krittanawong C, Virk HUH, Bangalore S, Wang Z, Johnson KW, Pinotti R, Zhang H, Kaplin S, Narasimhan B, Kitai T, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. Sci Rep. 2020;10(1):16057.
18. Shouman TT, Stocker R. Integrating clustering with different data mining techniques in the diagnosis of heart disease. J Comput Sci Eng 2013;20(1).
19. Motur S, Rao ST, Vemuru S. Frequent itemset mining algorithms: a survey. J Theor Appl Inf Technol 2018;96(3).
20. Javeed A, Khan SU, Ali L, Ali S, Imrana Y, Rahman A. Machine learning-based automated diagnostic systems developed for heart failure prediction using different types of data modalities: A systematic review and future directions. Comput Math Methods Med. 2022;2022:1–30. https://doi.org/10.1155/2022/9288452.
21. Malki Z, Atlam E, Dagnew G, Alzighaibi AR, Ghada E, Gad I. Bidirectional residual lstm—based human activity recognition. J Comput Inf Sci. 2020;13(3):1–40.
22. Malki Z, Atlam E-S, Hassanien AE, Dagnew G, Elhosseini MA, Gad I. Association between weather data and COVID-19 pandemic predicting mortality rate: machine learning approaches. Chaos Solitons Fractals. 2020;138: 110137. https://doi.org/10.1016/j.chaos.2020.110137.
23. Atlam E-S, El-Raouf MMA, Ewis A, Ghoneim O, Gad I. A new approach to identify psychological impact of covid-19 on university students academic performance. Alex Eng J. 2021;61(7):5223–33.
24. Malki Z, Atlam E-S, Ewis A, Dagnew G, Reda A, Elmarhomy G, Elhosseini MA, Hassanien AE, Gad I. ARIMA models for predicting the end of COVID-19 pandemic and the risk of a second rebound. J Neural Comput Appl. 2020;33(7):2929–2948.https://doi.org/10.21203/rs.3.rs-34702/v1
25. Almars MM, Almaliki M, Noor TH, Alwateer MM, Atlam E. Hann: hybrid attention neural network for detecting covid-19 related rumors. IEEE Access. 2022;10:12334–44.
26. Malki Z, Atlam E-S, Ewis A, Dagnew G, Ghoneim OA, Mohamed AA, Abdel-Daim MM, Gad I. The covid-19 pandemic: prediction study based on machine learning model. J Environ Sci Pollut Res. 2021;28(30):40496–506.
27. Manjunatha MFDH, Ibrahim Gad E-SA, Ahmed A, Elmarhomy G, Elmarhoumy M, Ghoneim OA. Parallel genetic algorithms for optimizing the sarima model for better forecasting of the ncdc weather data. Alexandria Eng J. 2020;60:1299–316.
28. Khan MA, Algarn F. A healthcare monitoring system for the diagnosis of heart disease in the iomt cloud environment using msso-anfis. IEEE Access. 2020;8:122259–69.
29. Javeed A, Zhou S, Yongjian L, Qasim I, Noor A, Nour R. An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection. IEEE Access. 2019;7:180235–43. https://doi.org/10.1109/access.2019.2952107.
30. Meter W. World Meter. Accessed: October 2020 (2020). https://www.worldometers.info/coronavirus/.
31. Coronavirus: Who (2020) coronavirus (2020). www.who.int/health-topics/.
32. Ali L, Rahman A, Khan A, Zhou M, Javeed A, Khan JA. An automated diagnostic system for heart disease prediction based on χ^2 statistical model and optimally configured deep neural network. IEEE Access. 2019;7:34938–45. https://doi.org/10.1109/access.2019.2904800.
33. Health M. Ministry of Health, COVID-19. Accessed: October 2020. 2020. https://covid19.moh.gov.sa/.
34. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, Gomes AS, Folsom AR, Shea S, Guallar E, et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. Circ Res. 2017;121(9):1092–101.
35. Feng Y, Leung AA, Lu X, Liang Z, Quan H, Walker RL. Personalized prediction of incident hospitalization for cardiovascular disease in patients with hypertension using machine learning. BMC Med Res Methodol. 2022;22(1):1–11.
36. Adam P, Parveen A. Prediction system for heart disease using naïve bayes. J Adv Comput Math Sci. 2012;3(3):290–4.
37. Tran H. A survey of machine learning and data mining techniques used in multimedia system. no 113 13–21 2019.

38. Gnaneswar B, Jebarani ME. A review on prediction and diagnosis of heart failure. In 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 17-18 March, Coimbatore, India, 2017;1–3. https://doi.org/10.1109/ICIIECS.2017.8276033
39. Kusprasapta M, Ichwan M, Utami DB. Heart rate prediction based on cycling cadence using feedforward neural network. In 2016 International Conference on Computer, Control, Informatics and its Applications (IC3INA), IEEE, 2016;72–76. https://doi.org/10.1109/IC3INA.2016.7863026
40. Singh KY, Sinha N, Singh KS. Heart disease prediction system using random forest. In International Conference on Advances in Computing and Data Sciences, Advances in Computing and Data Sciences. ICACDS 2016. Communications in Computer and Information Science, Singapore. 2017;721:613–623. https://doi.org/10.1007/978-981-10-5427-3_63
41. Priya RP, SKinariwala A. Automated diagnosis of heart disease using random forest algorithm. Int J Adv Res Ideas Innovat Technol 2017;3(2).
42. Tripoliti E, Fotiadis ID, Manis G. Automated diagnosis of diseases based on classification: dynamic determination of the number of trees in random forests algorithm. EEE Trans Inf Technol Biomed 2012;16(4).
43. Gonsalves AH, Thabtah F, Mohammad RMA, Singh G. Prediction of coronary heart disease using machine learning: an experimental analysis. In: Proceedings of the 2019 3rd International Conference on Deep Learning Technologies, 2019;51–56.
44. Oikonomou EK, Williams MC, Kotanidis CP, Desai MY, Marwan M, Antonopoulos AS, Thomas KE, Thomas S, Akoumianakis I, Fan LM, et al. A novel machine learning-derived radiotranscriptomic signature of perivascular fat improves cardiac risk prediction using coronary ct angiography. Eur Heart J. 2019;40(43):3529–43.
45. El-Hasnony IM, Elzeki OM. Multi-label active learning-based machine learning model for heart disease prediction. Sensors. 2022;22(3):1184–8. https://doi.org/10.3390/s22031184.
46. Guleria P, Srinivasu PN, Ahmed S. Ai framework for cardiovascular disease prediction using classification techniques. Electronics. 2022;11(24):1184–8. https://doi.org/10.3390/electronics11244086.
47. Javaid A, Zghyer F, Kim C, Spaulding EM, Isakadze N, Ding J, Kargillis D, Gao Y, Rahman F, Brown DE, et al. Medicine 2032: the future of cardiovascular disease prevention with machine learning and digital health technology. Am J Prevent Cardiol, 2022;100379
48. Alaa AM, Bolton T, Di Angelantonio E, Rudd JH. Van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 uk biobank participants. PloS One 2019;14(5):0213653.
49. Ward A, Sarraju A, Chung S, Li J, Harrington R, Heidenreich P, Palaniappan L, Scheinker D, Rodriguez F. Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. NPJ Digital Med. 2020;3(1):125.
50. Jamthikar A, Gupta D, Khanna N.N, Araki T, Saba L, Nicolaides A, Sharma A, Omerzu T, Suri HS, Gupta A, et al. A special report on changing trends in preventive stroke/cardiovascular risk assessment via b-mode ultrasonography. Cognitive Inf Comput Modelling Cognitive Sci 2020;291–318.
51. Suri JS, Bhagawati M, Paul S, Protogeron A, Sfikakis PP, Kitas GD, Khanna NN, Ruzsa Z, Sharma AM, Saxena S, et al. Understanding the bias in machine learning systems for cardiovascular disease risk assessment: the first of its kind review. Comput Biol Med. 2022;105204.
52. Vulli A, Srinivasu PN, Sashank MSK, Shafi J, Choi J, Ijaz MF. Fine-tuned densenet-169 for breast cancer metastasis prediction using fastai and 1-cycle policy. Sensors. 2022. https://doi.org/10.3390/s22082988.
53. Chatzimichail T, Hatjimihail AT. A software tool for calculating the uncertainty of diagnostic accuracy measures. Diagnostics. 2021. https://doi.org/10.3390/diagnostics11030406.
54. Chunhu Zhang DL. Xiaojian Shao: knowledge-based support vector classification based on c-svc. Proc Comput Sci. 2013;17:1083–90. https://doi.org/10.1016/j.procs.2013.05.137.
55. Md Yasin Kabir SM. Coronavis: A real-time covid-19 tweets data analyzer and data repository. arXiv. 2020.
56. Zhang Z. Introduction to machine learning: k-nearest neighbors. Ann Transl Med. 2016. https://doi.org/10.21037/atm.2016.03.37.
57. Wilbur WJ, Kim W. Stochastic gradient descent and the prediction of mesh for pubmed records. AMIA Annu Symp Proc 2014;1198–1207.
58. Mohandoss DP, Shi Y, Suo K. Outlier prediction using random forest classifier. In: 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC). IEEE. 2021. https://doi.org/10.1109/ccwc51732.2021.9376077.
59. Dimovski AS, Apel S, Legay A. A decision tree lifted domain for analyzing program families with numerical features. In: Fundamental Approaches to Software Engineering, pp. 67–86. Springer. 2021. https://doi.org/10.1007/978-3-030-71500-7_4.
60. Fedesoriano: Heart Failure Prediction Dataset. Retrieved [Date Retrieved] from. Accessed: September 2021 (September 2021). https://www.kaggle.com/fedesoriano/heart-failure-prediction.
61. UCI: Heart Failure Prediction Dataset. UCI Machine Learning Repository. Accessed: September 2021 (September 2021). https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/.
62. Castelli WP, Anderson K. A population at risk prevalence of high cholesterol levels in hypertensive patients in the Framingham study. Am J Med. 1986;80(2A):23–32. https://doi.org/10.1016/0002-9343(86)90157-9.
63. Luque A, Carrasco A, Martín A. de Las Heras A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition 2019;91:216–231.
64. Gad I, Hosahalli D. A comparative study of prediction and classification models on NCDC weather data. Int J Comput Appl. 2020. https://doi.org/10.1080/1206212x.2020.1766769.
65. Clarin JA. Academic analytics: predicting success in the licensure examination of graduates using CART decision tree algorithm. J Adv Res Dyn Control Syst. 2020. https://doi.org/10.5373/jardcs/v12sp1/20201057.
66. Hosahalli D, Gad I. A generic approach of filling missing values in NCDC weather stations data. In: 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 143–149. IEEE. 2018. https://doi.org/10.1109/icacci.2018.8554394.

67. Ghosh J, Shuvo SB. Improving classification model's performance using linear discriminant analysis on linear data. 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India 2019. https://doi.org/10.1109/ICCCNT45670.2019.8944632.
68. Imon AHMR, Roy MC, Bhattacharj SK. Prediction of rainfall using logistic regression. Pak J Stat Oper Res. 2012. https://doi.org/10.1234/pjsor.v8i3.535.
69. Wessler BS, Paulus J, Lundquist CM, Ajlan M, Natto Z, Janes WA, Jethmalani N, Raman G, Lutz JS, Kent DM. Tufts pace clinical predictive model registry: update 1990 through 2015. Diagn Prognostic Res. 2017;1:1–8.
70. Peter J, Somasundaram K. Study and development of novel feature selection framework for heart disease prediction. Int J Sci Res Publ. 2012;10(2):1–7.
71. Shouman M, Turner T, Stocker R. Integrating decision tree and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients. Int J Inf Educ Technol. 2012;2(3):220–3.
72. MS. RR. Heart disease prediction system using naive based and jelmeck mercer smoothing. IJARCCE 2014;3:6787–6792.
73. Ferdousy EZ, Islam MM, Matin MA. ombination of naïve bayes classifier and k-nearest neighbor (cnk) in the classification based predictive models. Comput Inf Sci. 2013;6(3):48–56.
74. N. DU. Prediction system for heart disease using naive bayes and particle swarm optimization. Biomedical Research-Tokyo 2018;29:2646–2649.

Publisher's Note