VFAST Transactions on Software Engineering
http://vfast.org/journals/index.php/VTSE@ 2022, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 10, Number 3, July-September 2022 pp: 58-70

Detection of Heart Disease Using Supervised Machine Learning

Amna Kanwal[1], Dr.Khawaja Tehseen Ahmad[2] ,Mr. Muhammad Kamran Abid[3], Dr.Naeem Aslam[4]

[1]Department of computer science, university of NFC IET, Multan, Pakistan
[2]Department of computer science, Bahauddin Zakariya University, Multan 60800, Pakistan
[3]Department of computer science, university of NFC IET, Multan, Pakistan
[4]Department of computer science, university of NFC IET, Multan, Pakistan
*Corresponding email address: 2k19mscs106@nfciet.edu.pk

ABSTRACT

One of the most prevailing and serious disease affecting human's health is Heart Disease (HD). Early diagnosis may allow for heart disease prevention or reduction, which could lower the rate of death.Machine Learning techniques have produced a variety of solutions for heart disease prediction and is capable of predicting illness at early stage . This study propose a model that includes many machine learning (ML) techniques to obtain accurate heart disease (HD) predictions. Data collection and pre-processing are used to create accurate data for the training model. Supervised Machine learning classifiers like support vector machine (SVM), decision tree (DT), logistic regression (LR), K Nearest Neighbor (KNN) and Naïve Bayes (NB) are used for predicting heart disease. Most relevant features are selected by using Relief and LASSO feature selection techniques. Various evaluating methods like, sensitivity, accuracy, specificity, MCC,confusion matrix and precision are used for the performance evaluation of model. This study did comparative analysis using supervised machine learning and feature selection techniques. *Decision tree gives highest accuracy of 85.21% with all features. On the other hand, with feature selection techniques SVM has an excellent performance. Future strategy is to use Deep learning algorithms and other feature selection techniques.*

KEYWORDS

Heart Disease, Support Vector Machine, Decision Tree, Logistic Regression, Naïve Bayes, K-Nearest Neighbor, Relief feature selection, LASSO feature selection.

JOURNAL INFO

HISTORY: Received: August 02, 2022 Accepted: September 13. 2022 Published: September 30, 2022

1. INTRODUCTION

According to the World Health Organization, heart disease is the leading cause of death worldwide, impacting 17.9 million people every year. Unhealthy habits that result in obesity, hypertension, high blood sugar, and high cholesterol raise the risk of heart disease. The American Heart Association adds gaining weight, sleep issues, foot swelling, a respiratory infection and an increased heart rate to the list of symptoms. Medical data collection is expanding, giving doctors a new chance to enhance patient diagnosis.To enhance decision-making support, practitioners have increased their use of computer technologies in the past few years. Machine learning is constantly being applied in the healthcare sector as a tool to help with patient diagnosis. Recent research has employed machine learning approaches to anticipate and diagnose different heart problems[1].Heart Disease is a general term used to describe illnesses that have an impact on the heart and blood vessels in a person. This might include arterial damage in organs such as the eyes, brain, kidneys and heart [2]. Even among young people, Heart disease is a prominent cause of mortality in many advanced and emerging countries throughout the world. But the truth is that leading a healthy lifestyle can greatly reduce the risk of it. Some types of heart disease (HD) are Coronary heart disease, Cardiomyopathy, Ischemic HD, Heart Failure, Hypertensive HD, Inflammatory HD and Valvular heart disease[3].Age, sex, cholesterol, a poor diet, high blood pressure, obesity, smoking, family history and alcohol use are all regarded to be risk factors for heart disease. Some symptoms and factors are under your control. The only solutions to lower the death rates brought on by heart diseases (HDs) are early prediction and efficient cures[4]. New predictive model technology in disease prediction are therefore appealing to the majority of medical scientists. These latest advancements in medical care have expanded electronic data availability, opened up new possibilities for decision support, and increased productivity[5].several lab tests and imaging studies can be used to determine the presence of CVD. However, the family history, patient's history, symptoms and risk factors, and physical checkup make up the bulk of the diagnosis. Using statistical data, we may coordinate the results and estimate the existence of illness from findings and treatment outcomes. Doctors can make well-informed decisions with the help of automation and deep learning (DL). Through artificial intelligence and machine learning, computers are trained to spot patterns in illness manifestation and translate them into structural data for prediction. AI is used to innovate in the areas of operations, revenue cycle, and electronic health records (EHR). It will eventually be integrated with clinical workflow, giving practitioners access to real-time data right at the point of care through the use of already-available tools[6].

Artificial intelligence is a component of machine learning, which is used in many daily life applications. The

prediction of an outcome based on past data is one common use of ML. In order to predict the outcome, the machine learns from the historical dataset and uses them on the unfamiliar dataset. For prediction, classification is best machine learning method. While some classification algorithms exhibit adequate prediction accuracy, others only show marginal accuracy[7]. As compared to other processes for data classification, machine learning techniques have the potential to provide high classification accuracy. The most practical method of creating assessments for scientific and real-world situations is ML classification. In order to distinguish patients as having or not having heart disease, persistence is also used to evaluate how various machine learning techniques behave [8].

The following are the three key contributions of our work: (1) examining pre-processing and analyzing supervised Machine learning classifiers for detecting heart disease; (2) researching the application of feature selection; and (3) developing a novel genetic training model.

The work is structured as follows: Introduction in section 1,The literature review is presented in Section 2. In this section, we looked at few research that used machine learning approaches to detect heart problems. In Section 3, we described the applied methods. Section 4 presents implementation. Section 5 presents the recommended techniques and experimental results. In Section 6 Conclusion are presented.

2. LITERATURE REVIEW

This study examined different machine learning techniques on heart disease detection at the medical center.

In [9]machine learning techniques provide only a glimpse into their ability to predict heart disease. This paper describes a method for determining substantial characteristics using ML techniques, thereby improving the precision of cardiovascular disease (CVD) prediction. The prediction was done with the help of feature combinations and supervised ML classification. Improved performance with an accuracy rate of 88.7 percent for the HD prediction model using Hybrid Random Forest with Linear Model. To improve the performance of HD prediction, it is possible to devise new feature selection techniques that provide a broader understanding of the significant features.

In [10] different classification algorithms were utilized, and the feature selection for each dataset was determined using backward modeling and the p-value test. The attributes with a p-value exceeding 0.05 were eliminated, and the model was re-fitted with the remaining variables. This procedure was repeated numerous times until every existing model variable reached a significant level. Using Logistic Regression, our proposed method accomplishes an accuracy of 87.1% in detecting HD. Future expansion and enhancement of the project will involve automating data manipulation, feature selection (FS), and model fitting for optimal prediction accuracy. Utilizing a pipeline structure for data preprocessing could contribute to improved outcomes.

Using a ML Algorithm to Predict HD was the subject of another study [11].Using a heart disease dataset, the suggested study developed Machine Learning system for predicting HD. This study employed seven well-known ML algorithms, 3 feature selection(FS) algorithms, the cross-validation method, and 7 performance evaluation metrics for classifiers, including classification sensitivity, accuracy, specificity, Matthews' correlation coefficient, and execution time. The suggested system can easily distinguish and categorize individuals with heart disease (HD) from healthy individuals. K-NN, NB, SVM, DT, ANN, and random forest were employed, along with the feature selection algorithms Relief, mRMR, and LASSO. In comparison to mRMR and LASSO, the performance of classifiers utilizing the Relief FS algorithm for the selection of essential features is superior. Future experiments can be conducted to increase the performance of these prognostic classifiers for the prediction of heart disease by employing alternative optimization techniques and feature selection algorithms.

In this work, the Cleveland data set for cardiac diseases, which contained 303 observations, was used as the primary database for training and testing the developed system[12].to increase the amount of data, 10-Fold Cross-Validation has been utilized. The dataset has been analyzed using various classifiers, including Naive Bayes (NB), Multilayer Perceptron (MLP), Decision Tree (DT), K-Nearest Neighbor (K-NN), Radial Basis Function (RBF), and ensemble prediction of classifiers, Single Conjunctive Rule Learner (SCRL), bagging, and stacking, boosting. Experiment findings show that the Support Vector Machine (SVM) method using the boosting strategy beats the other techniques.

This study focused on Machine Learning classifiers' accuracy. It used UCI repository data set for splitting. These algorithms are decision tree, k-nearest neighbor, support vector machine (SVM), and linear regression. The best tool for implementing Python programming is the Anaconda (Jupiter) notebook, which gives more accurate results. After implementing the ML approach for testing and training, we consider that the KNN is significantly more accurate than other algorithms with an accuracy of 87%[13].

In this study [14]. the proposed ML model is checked and validated on a data set pertaining to heart disease (HD) in order to find the precision of ML algorithms. Originally referred to as a confusion matrix, which details a classification model's performance by providing details on both actual and expected data classifications made by a classifier. A data set is downloaded from the Kaggle website in order to study and apply the data on a different algorithm to evaluate the accuracy score, sensitivity, and specificity of the main attribute of heart failure patients. The KNN algorithm proven has very successful and efficient in predicting cardiac disease based on its accuracy score.

This study [15] showed a heart disease prediction system that analyses a patient's history to determine whether a patient has heart disease or not. To predict and categorizes

heart disease, they used ML algorithms like logistic regression and KNN. Using KNN and Logistic Regression, the proposed model was capable of accurately forecast the existence of heart disease(HD) in a specific individual, demonstrating greater accuracy than previously used classifiers such as naive Bayes, etc. Our model's accuracy is 87.5 percent. More training data increases the likelihood that the model will precisely forecast whether a provided individual has heart disease (HD) or not.

In [16]used Machine Learning methods and techniques to automate the study of huge and complicated medical information.

Several researchers have employed a range of machine learning technology over the years to help the medical industry and experts diagnose heart-related illnesses. This study surveys and evaluates the performance of many models based on such algorithms and approaches. Some supervised Machine Learning techniques like

KNN, SVM, NB, DT, RF and ensemble techniques are important according to studies. However, over-fitting problems are also resolved by researchers. A significant amount of study may also be undertaken on the best ensemble of algorithms for a particular sort of data.

The proposed research in [17] collected from Kaggle dataset .The diagnostic model's performance is determined using methods such as classification, accuracy, sensitivity, and specificity analysis. This research presents a methodology for predicting if a person has cardiovascular illness and providing awareness or a diagnosis. This is performed by comparing the Machine Learning classifiers' accuracy on a dataset gathered in an area to offer an appropriate cardiovascular disease prediction model. With an accuracy ranging from 58.71 percent to 77.06 percent, the machine learning algorithms were able to predict heart disease in individuals. Logistic Regression was shown to be more accurate (77.06 percent) than other Machine Learning Algorithms.

The suggested cardiac disease prediction system was created in this study [18]. To improve prediction of heart disease, feature selection algorithms and ensemble techniques are used. To extract key information from the Cleveland heart disease dataset, feature extraction methods are used. A comparison of ensemble approaches (boosting and bagging) and five classifiers is carried out on certain attributes (DT, KNN, NB, SVM, and RF). Based on the results of the experiments, the bagging ensemble learning method with DT and PCA feature extraction produced the best results.

The suggested technique in [19] proposes risk factor levels based on the patient's data to avoid heart disease through adequate health maintenance. The suggested approach in this study is to classify heart disease patients depending on the risk factors. First, use the feature selection. This is done to cut down on the number of data records and speed up model training. The five classifiers utilized are support vector machine, K-nearest neighbor, decision tree, and random forest. The suggested model categorized heart disease patients based on age-related risk variables after applying the classifiers. Random Forest is the most accurate classifier in terms of accuracy. There are some drawbacks to this study. In comparison to all other algorithms, KNN fared badly. To be helpful, features must be predictive, which necessitates correct feature selection. Because of the tiny dataset, the accuracy of each method varies greatly.

The proposed study includes a unique machine learning technique for predicting heart disease [20].The Cleveland dataset was used in the planned research, as were data mining techniques such as regression and classification. To forecast cardiac illness, the suggested study uses a TkInter Python program. Machine learning techniques such as random forest and decision tree are used. A fresh machine learning model approach has been developed. In an implementation, three machine learning algorithms are used: 1. Decision Tree, 2. Random Forest, and 3. Hybrid model. According to the testing data, the hybrid model's heart disease prediction model has an accuracy of 88.7 percent. The interface is designed to collect user input factors in order to forecast heart disease using a hybrid decision tree and random forest model. As

a consequence, utilizing deep learning algorithms to forecast heart disease may provide better outcomes. In future studies, determining the severity of the condition may also entail categorizing it as a multiclass issue.

3. RESEARCH METHDOLOGY

A broad explanation is provided for developing a machine learning model using a heart disease dataset. This is part of the framework. The process of research methodology is depicted in Figure. 1.

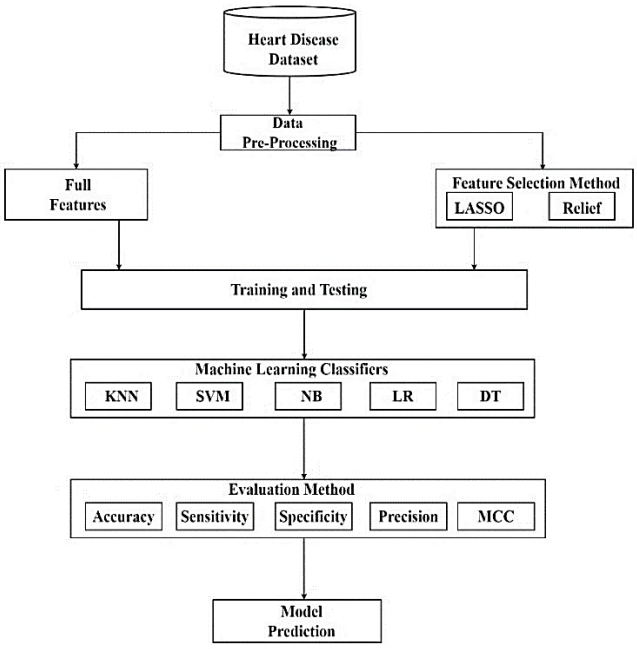


Figure 1: Research Methodology

3.1 Overview of proposed model:

In this study different ML algorithms, feature selection techniques have been performed to diagnose that patients have heart disease or not different evaluation methods like accuracy, specificity, precision, sensitivity and MCC are calculated and evaluated. Supervised ML Classifiers that are used for experiment are support vector Machine, Naïve Bayes Decision tree, k nearest Neighbor, and Logistic Regression. The dataset used in this study consists of 13 variables measured on 271 individuals. it used two Feature extraction techniques that enhance the accuracy of classifiers.

First, this study used five classifiers without feature extraction to predict patient may have heart disease or not. After computing the results, applied feature extraction techniques to check whether our accuracy increased or not by applying theses feature extraction techniques.

3.2 Performance evaluation metrics:

To identify the most effective algorithm, five (05) classification algorithms were applied to the dataset and their accuracy and other statistical metrics were compared. The algorithms used were K-nearest neighbors, Naive Bayes, Logistic Regression, support vector machine and Decision Tree Based on the metrics used to evaluate their performance, these algorithms were compared. To determine the sensitivity, specificity, Precision, MCC, and accuracy of the outcome for each method, a confusion matrix (CM) was created. All the parameters were estimated using the formulas listed below.

True Positive: According to this value, unwell subjects are correctly categorized and contain Heart Disease (HD).

True Negative: According to this value, healthy persons are correctly identified and do not have any heart diseases (HD).

False Positive: This value suggests that healthy patients are misclassified to prevent them from having heart illnesses.

False Negative: This score indicates that subjects are incorrectly categorized as unhealthy in order to prevent cardiac illnesses.

$$\textbf{Sensitivity} = \frac{TP}{TP+FN} \quad \ldots\ldots (1)$$

$$\textbf{Specificity} = \frac{TN}{TN+FN} \quad \ldots\ldots (2)$$

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \ldots\ldots (3)$$

$$\textbf{Precision} = \frac{TP}{TP+FP} \quad \ldots\ldots (4)$$

$$\boldsymbol{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \times 100 \quad \ldots\ldots (5)$$

4. Implementation:

The simulation tool is a Jupiter notebook, which is convenient for Python programming tasks. Jupyter notebook includes code as well as rich text elements like equations, links, and many other types of data. Because they integrate rich text components with code, these documents are suitable for combining an analytical description and its findings. An open-source web application called Jupyter Notebook with interactive features.

4.1 Overview of the Data

The dataset used in this study comes from the University of California Machine Learning Repository at UCI. It consists of 13 variables measured on 271 individuals. The 13th variable, called target, is a binary variable that signals the presence of heart disease or not. The variables and their descriptions are discussed in Table 1. This dataset comes from Cleveland. It has been divided into training and testing sets. For the machine learning algorithms' training input, we used an 80 percent training dataset. The Remaining 20 percent as test results for the prognosis of cardiac disease. Two alternative feature selection techniques—relief and LASSO are used during data preprocessing to address over fitting concerns and shorten execution times. Missing values and null values are found in dataset.

Table 1 Attribute Description

Attribute name	Attribute Description
Age	Age in years
Sex	1: male, 0: female
BP	resting blood pressure
Cp	chest pain type, 1: typical angina, 2: atypical angina, 3: non-angina l pain, 4: asymptomatic
Cholesterol	serum cholesterol in mg/dl
FBS over 120	fasting blood sugar > 120 mg/dl
EKG results	resting electrocardiographic result s (values 0,1,2)
Max HR	maximum heart rate achieved
Exercise angina	exercise induced angina
ST depression	oldpeak = ST depression induced by exercise relative to rest
Slope of ST	the slope of the peak exercise ST segment
Number of vessels fluro	number of major vessels (0-3) col ored by flourosopy
Thallium	thal: 3 = normal; 6 = fixed defec t; 7 = reversible
Target	Patients having heart disease or not having=1,not having=0

4.2 FEATURE SELECTION TECHNIQUES

Previous studies that deal with the datasets used in this study and are somewhat related to this study have been described previously; nevertheless, in most cases, the performance of those systems did not meet expectations. We want to create a technique that first identifies the best set of features, followed by the best algorithms to use with those features.

a)Least-Absolute-Shrinkage-Selection-Operator (LASSO)

This operator's ability to perform minimum selection and shrinking rely on adjusting the absolute value of the function coefficient. The subset of attributes can also exclude attributes with negative coefficients and features whose coefficient values are zero. When it comes to feature values with small coefficients, the LASSO performs very well. The selected feature subsets will include features with high coefficient values. With LASSO, unnecessary features might be detected[21].

b)Relief Feature Selection:

Relief is an attribute selection approach that assigns weights to each feature in the dataset. Then, these weights can be adjusted gradually. The vital features should have a substantial weight, whereas the less significant ones should have a small weight. In order to calculate feature weights, Relief employs methods similar to those used in KNN[22].

4.3 PROPOSED MACHINE LEARNING CLASSIFICATION MODEL

This section covers the machine learning methodologies utilized in this study to build an intelligent heart disease prediction system. Numerous supervised learning calculations exist, including neural networks (NN), and logistic regression support vector machines (SVMs), and Naive Bayes classifiers and decision tree.

a)Support Vector Machine

Support vector machines (SVMs) are fantastic but flexible machine learning (ML) algorithms that are used for both regression and classification. SVMs [23] differ from other machine learning algorithms in that they execute in a unique way. They've risen to prominence in recent years because to their capacity to handle a wide range of continuous and categorical data. A SVM model is a multidimensional hyperplane representation of many classes.

b) K-NEAREST NEIGHBORS

One of the familiar categorization techniques in the field of machine learning is K-Nearest Neighbors. It was once employed to treat heart disease. K Nearest Neighbor is regarded as non-parametric because it does not make any assumptions about how data will be distributed. KNN takes into account whether the new data and the current data are equivalent, and it assigns the new data to the class that is closest to the existing classes. KNN is used to solve both recognition and regression problems. The lazy learner algorithm is another name for it [24].

c)Decision Tree

A decision tree is a structure that resembles a tree and is used to classify instances by ordering them according to the values of the variables. In a decision tree, each node denotes a variable, and each branch denotes a possible value for the node. Beginning at the root node, instances are categorized and arranged according to the values of the variables. The root node of the tree would be the variable that divides the dataset most effectively. The decision-making portion, known as internal nodes (or split nodes), is responsible for selecting a choice based on several algorithms and for visiting subsequent nodes. Once the leaf has met a userdefined criteria, the split procedure is over. Classes are represented by the routes from root nodes to leaf nodes[25].

d)Naïve Bayes

The Naive Bayes classification technique may be applied in accordance with Naïve Bayes' theorem. Assumed independence among the different predictors. It implies that there shouldn't be any co-relation between qualities or attributes with regard to one another[26].

e)Logistic Regression

The primary ML model utilized is logistic regression (LR). It is a classification algorithm that is used to assign observations to various classes [27]. LR is a potent and well-known technique for supervised machine learning classification. It can be viewed as an extension of conventional regression and can only model a dichotomous variable, which often indicates whether an event will occur or not.

5. Result Analysis

The dataset comes from the University of Used in this study California Irvine's Machine Learning Repository at UCI. It consists of 13 variables measured on 271 individuals. The 13th variable, called target, is a binary variable that signals the presence of heart disease or not.

A	B	C	D	E	F	G	H	I	J	K	L	M	N
Age	Sex	cp	BP	Cholester	FBS over 1	EKG result	Max HR	Exercise a	ST depres	Slope of S	Number o	Thallium	target
70	1	4	130	322	0	2	109	0	2.4	2	3	3	1
67	0	3	115	564	0	2	160	0	1.6	2	0	7	0
57	1	2	124	261	0	0	141	0	0.3	1	0	7	1
64	1	4	128	263	0	0	105	1	0.2	2	1	7	0
74	0	2	120	269	0	2	121	1	0.2	1	1	3	0
65	1	4	120	177	0	0	140	0	0.4	1	0	7	0
56	1	3	130	256	1	2	142	1	0.6	2	1	6	1
59	1	4	110	239	0	2	142	1	1.2	2	1	7	1
60	1	4	140	293	0	2	170	0	1.2	2	2	7	1
63	0	4	150	407	0	2	154	0	4	2	3	7	1
59	1	4	135	234	0	0	161	0	0.5	2	0	7	0
53	1	4	142	226	0	2	111	1	0	1	0	7	0
44	1	3	140	235	0	2	180	0	0	1	0	3	0
61	1	1	134	234	0	0	145	0	2.6	2	2	3	1
57	0	4	128	303	0	2	159	0	0	1	1	3	0
71	0	4	112	149	0	0	125	0	1.6	2	0	3	0
46	1	4	140	311	0	0	120	1	1.8	2	2	7	1
53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
64	1	1	110	211	0	2	144	1	1.8	2	0	3	0
40	1	1	140	199	0	0	178	1	1.4	1	0	7	0
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1

Figure 2: Attribute Representation

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 270 entries, 0 to 269
Data columns (total 14 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Age                     270 non-null    int64
 1   Sex                     270 non-null    int64
 2   cp                      270 non-null    int64
 3   BP                      270 non-null    int64
 4   Cholesterol             270 non-null    int64
 5   FBS over 120            270 non-null    int64
 6   EKG results             270 non-null    int64
 7   Max HR                  270 non-null    int64
 8   Exercise angina         270 non-null    int64
 9   ST depression           270 non-null    float64
 10  Slope of ST             270 non-null    int64
 11  Number of vessels fluro 270 non-null    int64
 12  Thallium                270 non-null    int64
 13  target                  270 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 29.7 KB
```

Figure 3: Concise summary of the Data Frame

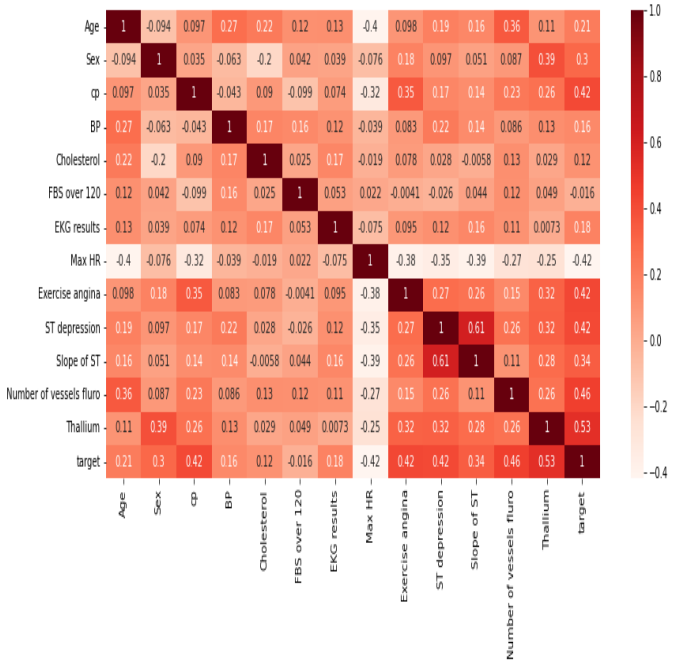


Figure 4: Heat map for correlation of attributes

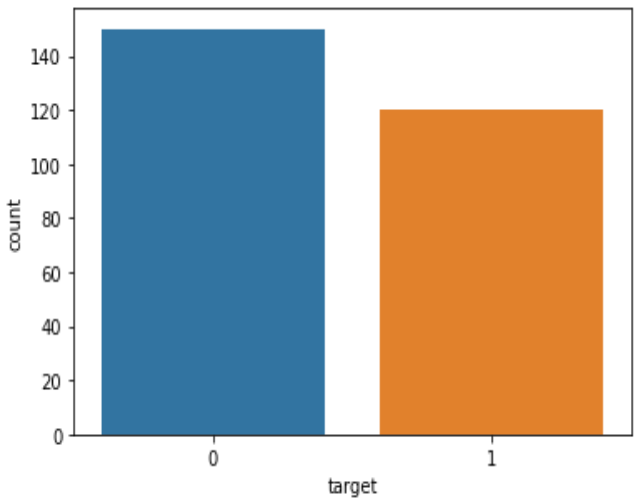


Figure 5. analyzing the target variable

Obtaining the data set including the characteristics of a human with a cardiac condition and a person who is not, as well as the conclusion as to whether the person has the condition or not, is the first stage in the setup. Python was employed as the experiment's programming language. The data collection has thirteen attributes, which are utilized. The data must next be analyzed. The info() function is applied to the Pandas library's data set to obtain a quick summary of the Data Frame.

5.1 Checking correlation between columns

After verifying that the data is balanced, the correlation between the variables are calculated and shown as a heat map using the Seaborn library. The heat map undeniably demonstrates the positive link between the target attribute and variables. After verifying the correlation, we must create dummy variables out of categorical variables. The Pandas library may be used to do this.figure 4 show heat map.

5.2 Exploratory Data Analysis (EDA)

For exploratory data analysis (EDA) and visualization, Jupyter notebook was utilized as a tool, and Python version 3.10.5 was employed. Because the success of a machine learning methodology is dependent on how effectively the information is prepared and presented. First, analyzing the target variable:

Heart Disease Absence in patients: 50.17
Heart Disease presence in patients: 40.13

5.3 Model Fitting

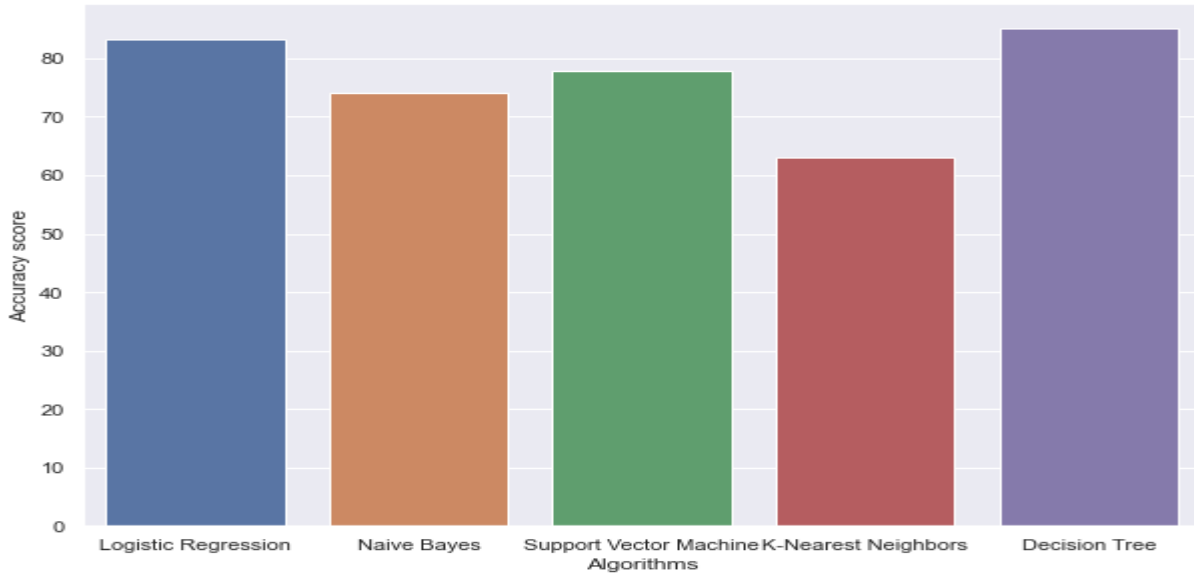


Figure 6: Machine Leraning classifier comparison

Table 2: Machine Learning Classifier Accuracy

Classifier	Accuracy
Logistic Regression	83.33 %
Naive Bayes	74.07 %
Support Vector Machine	77.78 %
K-Nearest Neighbors	62.96 %
Decision Tree	85.21 %

This figure shows DT achieves the highest accuracy of 85% among all .so, we can easily see that SVM predict the heart disease at its earlier stages with higher accuracy by using a dataset containing the attributes between the range of 10-13.

5.4 Feature Selection Using Least-Absolute-Shrinkage-Selection-Operator (LASSO)

Taking the names of attributes for further usage in plotting the graph between coefficient and attributes.

	Attributes	Coefficients
0	Age	0.0003
1	Sex	1.0613
2	cp	0.2044
3	BP	0.0078
4	Cholesterol	0.0006
5	FBS over 120	0.1767
6	EKG results	-0.1959
7	Max HR	0.0011
8	Exercise angina	0.2794
9	ST depression	0.0991
10	Slope of ST	0.2994
11	Number of vessels fluro	0.0781
12	target	1.2503

Figure 7: Attribute Co-efficient using LASSO

By seeing the above table, we can understand that when alpha is 0.1 there is no much difference between the coefficients of features w.r.t the output feature. But, alpha with 0.01 is having considerable a difference between the coefficients of features. So, it becomes easy to select the required features from the graph to further use in the classifier models.The LASSO considers closely related attributes to be true and the others to be false. After LASSO technique applying, sex received the highest rank score (1.0613), but maximal heart rate (EKG findings) received a relatively low score. Table 3 displays the LASSO score for the six

most important criteria for identifying heart disease.able 3: Attributes Selected from LASSO

Attributes	Coefficients
Sex |	1.0613
Exercise angina	0.2794
cp	0.2044
Slope of ST	0.2994
FBS over 120	0.1767
ST depression	0.0991

Table 4. Machine Learning Classifiers with LASSO feature selection

Models	Accuracy	Sensitivity	Specificity	Precision	MCC
Logistic Regression	83.3333	84.6154	82.1429	81.4815	66.7
KNN	81.4815	76.9231	85.7143	81.4815	66.7
SVM	85.1852	80.7692	89.2857	87.5	70.4
Naïve Bayes	79.6296	80.7692	78.5714	77.7778	59.2
Decision tree	81.4815	0.8462	0.7857	0.7857	0.63

This table shows SVM achieves the highest accuracy of 81% among all .so, we can easily see that SVM predict the heart disease at its earlier stages with higher accuracy by using a dataset contains the attributes between the range of 10-13

5.5 Feature Selection Relief Feature Selection Technique

Relief, a feature selection algorithm, chooses important features based on data weight.Figure 9 lists the six most essential input features chosen by Relief. According to the data,the most relevant factor for Relief Feature Selection Technique chooses siattributes with highest scores for better results

.

Attributes	Scores
Age	4.9
Sex	0.37
cp	1
BP	8.53
Cholesterol	8.16
FBS over 120	0.26
EKG results	0.96
Max HR	8.54
Exercise angina	0.46
ST depression	1.266
Slope of ST	0.57
Number of vessels fluro	1.07
Thallium	2.07
target	

Figure 8: Attribute Scores using relief feature selection

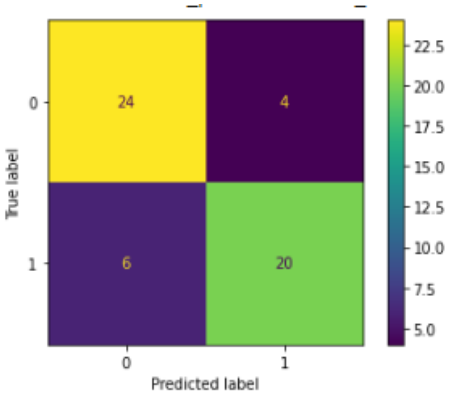
Table 5. Attribute selected with Relief feature selection

Attributes	Coefficients
Max HR	8.54
BP	8.53
Cholesterol	8.16
Age	4.9
ST depression	2.07
Thallium	1.266

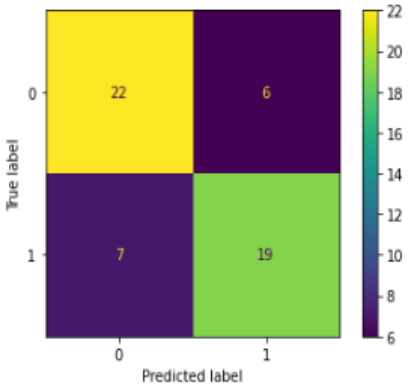
Table 6: Machine Learning Classifiers with relief Feature Selection

Models	Accuracy	Sensitivity	Specificity	Precision	MCC
Logistic Regression	79.6296	69.2308	89.2857	85.7143	59.976
KNN	70.3704	65.3846	75	70.8333	40.6084
SVM	81.4815	76.9231	85.7143	83.3333	62.9844
Naïve Bayes	77.7778	76.9231	78.5714	76.9231	55.4945
Decision tree	75.9259	73.0769	78.5714	76	51.7551

this table shows SVM achieves the highest accuracy of 81% among all .so, we can easily see that SVM predict the heart disease at its earlier stages with higher accuracy by using a dataset contains the attributes between the range of 10-

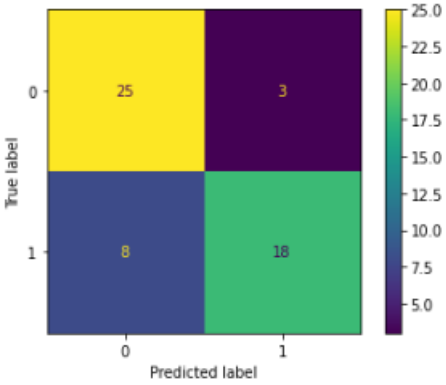


Support vector Machine

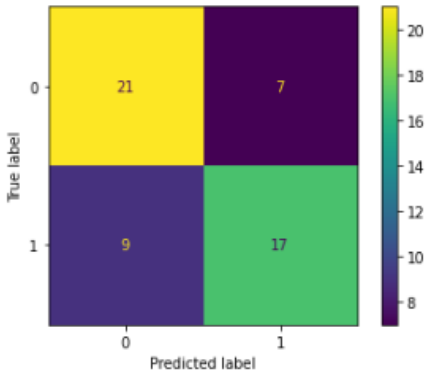


Decision Tree

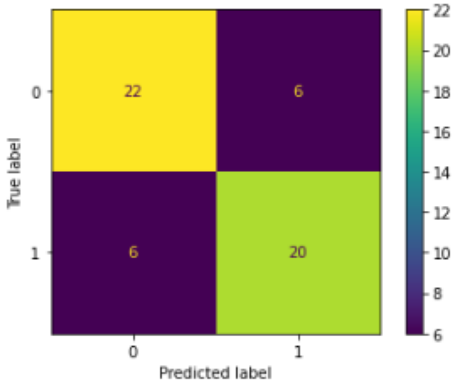
Confusion Matrix's of ML Classifiers with Relief Feature Selection:



Logistic Regression



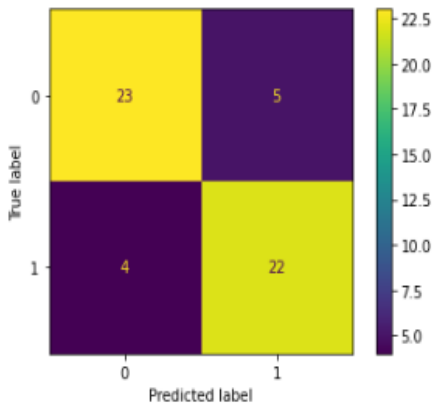
k Nearest neighbor



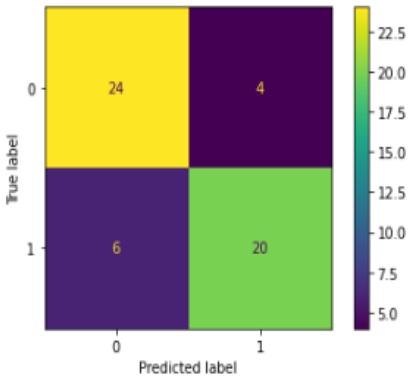
Naïve Bayes

Figure 9: Confusion Matrix's with Relief Feature Selection

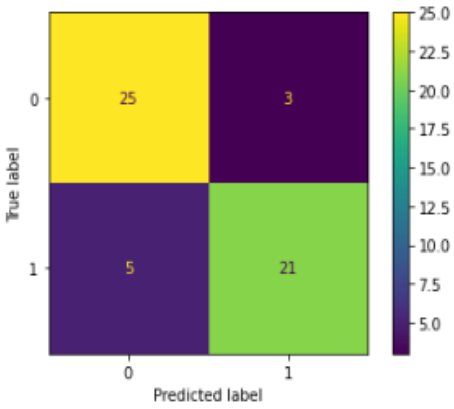
Confusion Matrix's of Machine Learning Classifiers with LASSO Feature Selection:



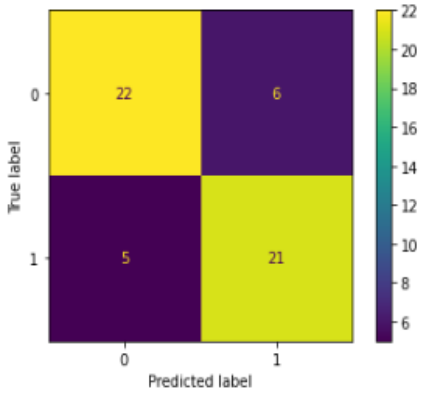
Logistic Regression



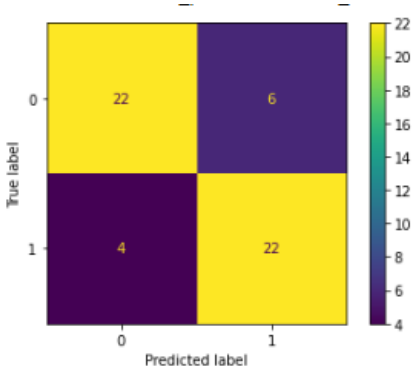
K Nearset Neighbor



Support vector Machine



Naïve Bayes



Decision Tree

Figure 10: Confusion Matrix's with LASSO Feature Selection

5.6 Results comparisons of Machine Learning Classifiers with Feature Selection Techniques

Table 7 Results comparisons of Machine Learning Classifiers with Feature Selection Techniques

Classifiers	Accuracy With all features	Accuracy with LASSO Technique	Accuracy with Relief technique
Logistic Regression	83.33 %	83.3333	79.6296
KNN	62.96 %	81.4815	70.3704
SVM	77.78%	85.1852	81.4815
Naïve Bayes	74.07 %	79.6296	77.7778
Decision tree	85.21 %	81.4815	75.9259

The Table shows the results accuracies on dataset. First, we applied ML classifiers on dataset without the Feature Extraction techniques and after these feature selection techniques are applied to enhance our accuracies. From this table we can clearly see when we applied feature selection techniques, accuracy of the each classifier improved. Here our comparison is between the results that are computed without feature selection and with feature selection techniques.

5.7 COMPARISON OF VARIOUS ALGORITHMS

It compares the results of several classification algorithms with various input attributes. First, all attributes of the heart disease dataset were subjected to five machine learning classifiers. Second, the Least Absolute Shrinkage and Selection Operator Features Selection Algorithm was used to select certain important attributes, followed by the same five machine learning classifiers. Finally, the Relief model's most essential attributes were utilized as input to the classifiers. To assess the projected results, many performance measures are used.Our dataset has 13 distinct features to determine the disease's result. After applying the Relief feature selection algorithm 6 features: age, cholesterol, maximum heart rate, ST depression, Thallium and Blood pressure have been
selected based on highest the score. Six relevant features: Sex, Exercise angina, cp, FBS over 120, slope of ST, and ST depression were selected with LASSO feature selection algorithm. SEX feature has highest the score.

5.8 COMPARISON OF DIFFERENT METHODS

In this study, we apply five classifiers. We used the five distinct approaches on the initial 13 input features, then the LASSO approach's six input features, then the Relief method's six input features. Figure 9 depicts the accuracy of several types of classifiers. The Classifier produced the most accurate prediction when 13 characteristics were used, while the Decision Tree produced an accuracy of 85.21 percent. Feature selection Techniques, on the other hand, produce much improved results. When just six characteristics are evaluated (LASSO), the Naive Bayes Classifier produces the lowest accuracy (79.62%). With the 6 LASSO features, we get 83.33%, 81.45 %, 85.18%, and 81 % accuracy for the LR, KNN, SVM, and DT classifiers, respectively. SVM has an excellent performance of 85.18 percent. When the accuracy of these five techniques with the Relief characteristics was compared, the support vector machine showed an exceptional accuracy of 81.18 percent.

5.9 Comparison of proposed Model and Existing Systems:

Table 8 show Comparison of proposed Model and Existing Systems.

Table 8. Comparison of proposed Model and Existing System

Our Work				Others Work			
Models	**Accuracy of Model with all features(13)**	**Accuracy of Model with LASSO**	**Accuracy of Model with Relief**	**Dataset Used**	**Existing Systems Accuracy**	**Dataset Used**	**Existing Systems Accuracy**
DT	85.21%	81.45%	76%	UCI Repository	70.97%[10]	Kaggle Dataset	82%[14]
LR	83.33%	84%	79%	Kaggle Dataset	77% [17]	Cleveland heart disease	82.9% [25]
KNN	62.96%	82%	70%	Kaggle Dataset	58% [30]	Cleveland heart disease	59.1% [29]
SVM	78%	85%	81%	UCI Machine Learning Repository	74% [28]	Kaggle dataset	71.5%[30]
NB	74%	80%	78%	UCI Machine Learning Repository	58% [28]	Cardiovascular Kaggle Dataset	69.72% [30]

6. CONCLUSION

Regardless of social or cultural background, accurately predicting the risk of heart disease could possibly have a significant impact on death rate of people. A vital first step in achieving that objective is early diagnosis. Several research has previously utilized machine learning to try to predict cardiac disease. Typically, accuracy is most significant method for assessing machine learning (ML) algorithms. Five classifiers are used on the 13 initial input features, the six features chosen using the LASSO approach, and the six features chosen using the Relief method, we applied the five different methods. This study demonstrates that several machine learning (ML) algorithms can be used with the Relief feature selection to select correlated feature set. The Classifier made the most accurate prediction out of 13 features, while the Decision Tree's accuracy is 85.21 %. However, Feature selection Techniques produce noticeably better results. The Nave Bayes Classifier produces the lowest accuracy when only evaluating six chosen features (LASSO) (79.62 percent). With the 6 LASSO features, we achieve accuracy of 83.33 %, 81.45 %, 85.18 %, and 81 % for the LR, KNN, SVM, and DT classifiers, respectively. SVM performed exceptionally well, scoring 85.18 percent. In the future, this study wants to further generalize the model to make it compatible with other feature selection techniques and more resistant to datasets with significant amounts of missing data. Another future strategy is to use Deep Learning algorithms. The main goal of this research was to make system practical and simple to use in real-world situations.

AUTHOR CONTRIBUTIONS

Amna Kanwal: Conceptualization, Methodology, Writing-Original draft preparation, Software implementation. **Dr Khawaja Tehseen Ahmad:** Supervision. **Muhammad Kamran Abid:** Software Validation, Writing- Reviewing and Editing. **Dr Naeem Aslam:** Visualization, Investigation.

COMPLIANCE WITH ETHICAL STANDARDS

It is declared that all authors are consented in submission to this Journal. It is also declared that we all author don't have any conflict of interest.

REFERENCES

[1] Hajjam, E. Hassani, E. Andr, and A. K. G, "Informatics in Medicine Unlocked Classification models for heart disease prediction using feature selection and PCA," vol. 19, 2020, doi: 10.1016/j.imu.2020.100330.

[2] http://www.nhs.uk/conditions/cardiovascular-disease/

[3] L. Yahaya, N. D. Oye, and E. J. Garba, "A Comprehensive Review on Heart Disease Prediction Using Data Mining and A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques," no. October, 2020, doi: 10.11648/j.ajai.20200401.12.

[4] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Comput. Biol. Med.*, vol. 136, no. May, p. 104672, 2021, doi: 10.1016/j.compbiomed.2021.104672.

[5] G. Battineni, G. G. Sagaro, and N. Chinatalapudi, "Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis," 2020.

[6] M. Swathy and K. Saruladha, "A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques," *ICT Express*, vol. 8, no. 1, pp. 109–116, 2022, doi: 10.1016/j.icte.2021.08.021.

[7] B. C. Latha and S. C. Jeeva, "Informatics in Medicine Unlocked Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics Med. Unlocked*, vol. 16, no. November 2018, p. 100203, 2019, doi: 10.1016/j.imu.2019.100203.

[8] A.K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Comput. Appl.*, 2016, doi: 10.1007/s00521-016-2604-1.

[9] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction using Hybrid Machine Learning Techniques," *IEEE Access*, vol. PP, p. 1, 2019, doi: 10.1109/ACCESS.2019.2923707.

[10] P. S. Kohli and A. L. Regression, "Application of Machine Learning in Disease Prediction," *2018 4th Int. Conf. Comput. Commun. Autom.*, pp. 1–4, 2018.

[11] A.U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms," vol. 2018, 2018.

[12] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease," no. Iscc, 2017.

[13] A.Singh, "Heart Disease Prediction Using Machine Learning Algorithms," pp. 452–457, 2020.

[14] N. Basha, "Early Detection of Heart Syndrome Using Machine Learning Technique," vol. 90, pp. 4–8, 2019.

[15] H. Jindal, S. Agrawal, R. Khera, and R. Jain, "Heart disease prediction using machine learning algorithms Heart disease prediction using machine learning algorithms," 2021, doi: 10.1088/1757-899X/1022/1/012072.

[16] V. V Ramalingam, A. Dandapath, and M. K. Raja, "Heart disease prediction using machine learning techniques: A survey Heart disease prediction using machine learning techniques : a survey," no. March 2018, 2019, doi: 10.14419/ijet.v7i2.8.10557.

[17] M. Srivenkatesh, "Prediction of Cardiovascular Disease using Machine Learning Algorithms," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 3, pp. 2404–2414, 2020, doi: 10.35940/ijeat.b3986.029320.

[18] X. Y. Gao, A. Amin Ali, H. Shaban Hassan, and E. M. Anwar, "Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/6663455.

[19] A.V. Vidyapeetham, A. V. Vidyapeetham, and A. V. Vidyapeetham, "Earlier Prediction on the heart disease based on supervised machine learning techniques," no. Icicc, pp. 1696–1703, 2021.

[20] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021.

[21] Zhou and A. Wieser, "Modified Jaccard index analysis and adaptive feature selection for location fingerprinting with limited computational complexity," *J. Locat. Based Serv.*, vol.

13, no. 2, pp. 128–157, 2019, doi: 10.1080/17489725.2019.1577505.

[22] P. Ghosh *et al.*, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.

[23] M. Srivenkatesh, "Prediction of Cardiovascular Disease using Machine Learning Algorithms," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 3, pp. 2404–2414, 2020, doi: 10.35940/ijeat.b3986.029320.

[24] F. M. Javed Mehedi Shamrat, P. Ghosh, M. H. Sadek, M. A. Kazi, and S. Shultana, "Implementation of Machine Learning Algorithms to Detect the Prognosis Rate of Kidney Disease," *2020 IEEE Int. Conf. Innov. Technol. INOCON 2020*, 2020, doi: 10.1109/INOCON50539.2020.9298026.

[25] A.Akella and S. Akella, "Machine learning algorithms for predicting coronary artery disease: Efforts toward an open source solution," *Futur. Sci. OA*, vol. 7, no. 6, 2021, doi: 10.2144/fsoa-2020-0206.

[26] A.Kondababu, V. Siddhartha, B. B. Kumar, and B. Penumutchi, "A comparative study on machine learning based heart disease prediction," *Mater. Today Proc.*, no. xxxx, pp. 1–5, 2021, doi: 10.1016/j.matpr.2021.01.475.

[27] S. Ambesange, A. Vijayalaxmi, S. Sridevi, Venkateswaran, and B. S. Yashoda, "Multiple heart diseases prediction using logistic regression with ensemble and hyper parameter tuning techniques," *Proc. World Conf. Smart Trends Syst. Secur. Sustain. WS4 2020*, pp. 827–832, 2020, doi: 10.1109/WorldS450073.2020.9210404.

[28] Rubini P. E., Dr. C. A. Subasini, Dr. A. Vanitha Katharine, V. Kumaresan, S. Gowdham Kumar, T. M. Nithya,"A Cardiovascular Disease Prediction using Machine Learning Algorithms", Annals of RSCB,vol. 25, no. 2, pp. 904–912, Mar. 2021.

[29] A.Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," vol.no 8. Ml, 2020, doi: 10.1109/ACCESS.2020.3001149.

[30] C.Zhou and A. Wieser, "Modified Jaccard index analysis and adaptive feature selection for location fingerprinting with limited computational complexity," *J. Locat. Based Serv.*, vol. 13, no. 2, pp. 128–157, 2019, doi: 10.1080/17489725.2019.1577505