Applied Computer Science, vol. 18, no. 1, pp. 70–88
doi: 10.35784/acs-2022-6

Submitted: 2022-03-12 | Revised: 2022-03-24 | Accepted: 2022-03-28

Keywords: coronary heart disease, machine learning, ensembles, outlier detection, framingham

Lubna RIYAZ [0000-0002-4502-7720]*, *Muheet Ahmed BUTT* [0000-0002-8059-0180]* , *Majid ZAMAN*[**]

IMPROVING CORONARY HEART DISEASE PREDICTION BY OUTLIER ELIMINATION

Abstract

Nowadays, heart disease is the major cause of deaths globally. According to a survey conducted by the World Health Organization, almost 18 million people die of heart diseases (or cardiovascular diseases) every day. So, there should be a system for early detection and prevention of heart disease. Detection of heart disease mostly depends on the huge pathological and clinical data that is quite complex. So, researchers and other medical professionals are showing keen interest in accurate prediction of heart disease. Heart disease is a general term for a large number of medical conditions related to heart and one of them is the coronary heart disease (CHD). Coronary heart disease is caused by the amassing of plaque on the artery walls. In this paper, various machine learning base and ensemble classifiers have been applied on heart disease dataset for efficient prediction of coronary heart disease. Various machine learning classifiers that have been employed include k-nearest neighbor, multilayer perceptron, multinomial naïve bayes, logistic regression, decision tree, random forest and support vector machine classifiers. Ensemble classifiers that have been used include majority voting, weighted average, bagging and boosting classifiers. The dataset used in this study is obtained from the Framingham Heart Study which is a long-term, ongoing cardiovascular study of people from the Framingham city in Massachusetts, USA. To evaluate the performance of the classifiers, various evaluation metrics including accuracy, precision, recall and f1 score have been used. According to our results, the best accuracy was achieved by logistic regression, random forest, majority voting, weighted average and bagging classifiers but the highest accuracy among these was achieved using weighted average ensemble classifier.

1. INTRODUCTION

Heart is a vital organ in the human body performing the crucial function of pumping blood to different parts of the body. A slight change in the normal functioning of the heart can lead to imbalance in the functioning of the whole body. Heart diseases also known as the cardiovascular diseases are the conditions that affect the structure or function of our

[*] PG Department of Computer Sciences, University of Kashmir, Srinagar, India, lubna.riyaz122@gmail.com, ermuheet@gmail.com
[**] Directorate of IT & SS, University of Kashmir, Srinagar, India, zamanmajid@gmail.com

heart (Cardiovascular (Heart) Diseases: Types and Treatments, n.d.). Heart diseases can be caused by various factors such as lifestyle changes including bad eating habits, sleep depreciation, less physical activity, smoking habits, etc. According to a report by World Health Organization, non-communicable diseases including heart diseases, strokes and other diseases are collectively responsible for death of almost 41 million people each year (Less than $1: How WHO Thinks That Can Save 7 Million Lives, n.d.). The various symptoms of heart diseases include chest pain, chest tightness, chest pressure and chest discomfort (Heart Disease – Symptoms and Causes – Mayo Clinic, n.d.). There are different kinds of heart diseases that include blood vessel disease, such as coronary artery disease, heart rhythm problems (arrhythmias), heart defects that some people are born with (congenital heart defects), heart valve disease, disease of the heart muscle, heart infection, etc. (Heart Disease – Symptoms and Causes – Mayo Clinic, n.d.). So, detection of heart disease at an early stage can help in saving numerous human lives. Detection of heart disease depends on blend of pathological and clinical data that is complex in nature. Today, medical industries produce huge quantities of data related to patients' health. This data needs to be processed to obtain useful information from it.

Heart disease is a general term for a large number of medical conditions related to heart and one of them is the coronary heart disease. Coronary heart disease is caused by the accumulation of fatty deposits on artery walls around heart and can result in severe illness and even death of the patient. According to a report by NHS (National Health Service), UK, coronary heart disease is a major cause of death in UK and worldwide (Coronary Heart Disease – NHS, n.d.). Various symptoms of coronary heart disease include shortness of breath, body pain, chest pain, feeling nausea, etc. With the passage of time, artery walls get furred up because of fatty substance known as atheroma (Coronary Heart Disease - NHS, n.d.). It can be caused by various lifestyle changes including drinking excessive alcohol, smoking etc. The build-up of fatty deposits on artery walls can block the blood flow (totally or partially) to the heart (Coronary Heart Disease | NHLBI, NIH, n.d.). It may also be caused by any injury or some disease affecting the normal working of the arteries (Coronary Heart Disease | NHLBI, NIH, n.d.). Coronary heart disease is sometimes also known as coronary artery disease (Coronary Artery Disease: Causes, Symptoms, and Treatment, n.d.) and can also lead to a heart attack (Coronary Artery Disease: Causes, Symptoms, and Treatment, n.d.). According to a survey by Centers for Disease Control and Prevention (CDC), it is the most common type of heart disease in the USA accounting for around 655,000 deaths per year (Coronary Artery Disease: Causes, Symptoms, and Treatment, n.d.).

Therefore, there is a need of some automatic diagnosis system that can process huge amounts of collected medical data and provide help with decision making. Machine learning contains large number of tools from preprocessing of data to the prediction process which can be utilized to help medical professionals in making decisions. In recent times, machine learning classifiers have been used extensively for decision making in various fields (Ashraf, Zaman & Ahmed, 2018a, 2018b, 2019, 2020; Mir et al., 2016; Mohd, Butt & Baba, 2019, 2020; Zaman, Quadri & Butt, 2012; Zaman, Kaul & Ahmed, 2020) as well as in medical science including diabetes (Kavakiotis et al., 2017; Wei, Zhao & Miao, 2018) liver diseases (El-Shafeiy, El-Desouky & Elghamrawy, 2018; Wu et al., 2019), brain diseases (Sakai & Yamada, 2019; Salvatore et al., 2014) and heart diseases. In this study various machine learning base and ensemble classifiers have been used for coronary

heart disease prediction. In the recent past, various machine learning classifiers that have been used for heart disease prediction include logistic regression, artificial neural networks, random forest, naïve bayes, k-nearest neighbor etc. In this study, the performance of various machine learning base and ensemble classifiers has been evaluated. The dataset used for this purpose is obtained from the Framingham Heart Study which is a long-term, ongoing cardiovascular study of people from Framingham city in Massachusetts (Framingham Heart Study, n.d.). The said dataset contains 4240 records with 15 attributes and one target attribute. Out of 4240 rows, there were 582 rows with missing values that were handled by removing them from our dataset. Also, the dataset contained various outliers which were also handled by removing them from our dataset. Various classifiers used include k-nearest neighbor, multilayer perceptron, multinomial naïve bayes, logistic regression, decision tree, random forest and support vector machine classifiers. Ensemble classifiers used include majority voting, weighted average, bagging and boosting classifiers. The block diagram of the proposed system is shown in Fig. 5.

The rest of this paper is organized as; Section 2 shows the literature review of the past years related to early detection of various heart diseases. Section 3 enlightens various machine learning classifiers that have been used extensively in the past years for heart disease prediction. In section 4, proposed methodology has been described that includes source of the data, features description, classification, validation schemes and evaluation metrics. Section 5 is the results and discussion section and the last section (section 6) summarizes the conclusion.

2. LITERATURE REVIEW

Hidayet Takci et al. proposed a new method for heart attack prediction using significant features to determine which machine learning method is best for predicting heart attack. The dataset for this purpose is obtained from the Statlog heart disease dataset. According to the results, the algorithm that came out with the best prediction accuracy was support vector machine algorithm using linear kernel while as the feature selection algorithm that gave the best results was the reliefF method. Together they achieved an overall accuracy of 84.81% (Takci, 2018).

Gokulnath et al. have proposed an optimization function based on support vector machine and significant attributes are selected using objective function in genetic algorithm. Genetic algorithm (GA) and support vector machine (SVM) results are compared with other existing feature selection methods. The proposed framework is exhibited on MATLAB using Cleveland dataset and 7 significant features have been identified by GA-SVM pair. This pair together provided an average accuracy of 84.40% (Gokulnath & Shantharajah, 2019).

Benhar et al. have proposed data preparation method which is performed before data mining process for heart disease. This is done by performing systematic mapping study. A total of 58 papers are being selected from Jan 2000 up to Dec 2017. According to the results, data preparation step has been done extensively in order to increase the prediction accuracy of different algorithms and their main focus was on reducing data and performing feature selection (Benhar, Idri & Fernández-Alemán, 2019).

Bashir et al. have focused on the feature selection methods in order to increase the heart disease prediction accuracy and instead of one, many heart disease datasets have been used for improving the accuracy. For feature selection algorithms that are being used include random forest, naïve bayes, support vector machine, logistic regression, logistic regression SVM and decision tree. According to the results, logistic regression SVM and naïve bayes have achieved the highest accuracies of 84.85% and 84.24% respectively (Bashir et al., 2019).

Chandra Shekar et al. have proposed a hybrid technique of ensemble classifier using genetic algorithm with decision tree for heart disease prediction. The output achieved from GA is the optimized feature set, which is then given as an input to decision tree algorithm to give us the final result. According to results, the proposed approach achieves an accuracy of 85.37% (Chandra Shekar, Chandra & Venugopala Rao, 2019).

Latha et al. have combined multiple weak classifiers to form their ensemble for improving the accuracy of heart disease prediction. According to the results, bagging and boosting ensemble methods improve the prediction accuracy of weak classifiers up to a large extent. An overall of 7% increase in accuracy was recorded while using ensemble methods as compared to base classifiers. Also, feature selection was also done thereby resulting in an overall accuracy of 85.48% by the proposed model (Latha & Jeeva, 2019).

Thaiparnit et al. proposed a novel method for heart disease prediction using a technique known as Hoeffding Tree. The data for this purpose is obtained from from UCI Repository containing 199 records with 13 attributes. Vertical Hoeffding Decision Tree (VHDT) is used for analysing the data. As per the results, an improved accuracy of 85.43% is achieved by the proposed VHDT technique (Thaiparnit, Kritsanasung & Chumuang, 2019).

Shinde et al. have utilized k means and naïve bayes techniques to design an intelligent system for heart disease diagnosis. K means clustering has been used for grouping different combination of features whileas naïve bayes has been used for prediction purpose (Shinde et al., 2015). Riyaz et al. performed a quantitative review of various machine learning techniques used for heart disease prediction (Riyaz et al., 2022).

Otoom et al. have proposed a real time monitoring system utilizing wearable sensors and mobile technology for coronary artery disease detection using machine learning. As per results the system achieved an overall accuracy of 85.1% using support vector machine with feature selection (Otoom et al., 2015). Dun et al. proposed a heart disease diagnosis system using ensemble learning. Authors have applied various ensemble and deep learning techniques using the concept of hyper-parameter tuning. The model resulted in an average accuracy of 78% on the test set (Dun, Wang & Majumder, 2016).

3. MACHINE LEARNING CLASSIFIERS

3.1. K-Nearest Neighbors

K-Nearest Neighbor (KNN) is a simple supervised machine learning technique which compares the similarity between the new unknown data point with already existing classes. It then assigns the class to the newly arrived data point based on its resemblance with the existing classes that is most suitable for it. So whenever a new data point arrives it is being easily classified into its suitable category. Algorithm for KNN classifier is shown below:

Step 1: Select the value for *K*.
Step 2: Compute Euclidean distance for all points.
Step 3: Depending upon distance, select the *K* nearest neighbors.
Step 4: Calculate the number of data points for each category of *K* neighbors
Step 5: Assign the newly arrived point to that category whose data points are more in number in *K* nearest neighbors.
Step 6: Model is ready (K-Nearest Neighbor(KNN) Algorithm for Machine Learning – Javatpoint, n.d.).

Fig. 1 shows the pictorial representation of working of KNN.

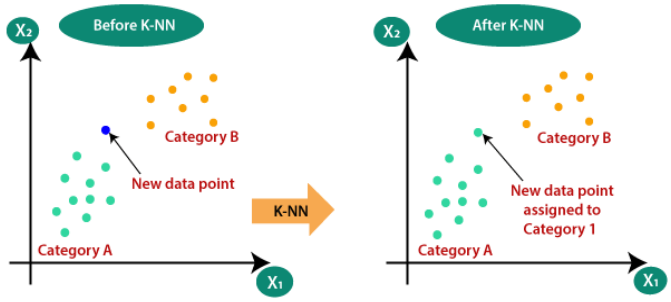

Fig. 1. Working of KNN (K-Nearest Neighbor(KNN) Algorithm for Machine Learning – Javatpoint, n.d.)

3.2. Multilayer Perceptron

Multi-layer perceptron (MLP) is an add-on of feed forward neural network. It consists of three types of layers – the input layer, hidden layer and an output layer. Input signal is fed to the input layer which processes it. The job of output layer is to perform prediction and classification. In between input and output layers, there exist one or more hidden layers considered computational engine for multilayer perceptron. The flow of data is from input to the output layer. Neurons are trained using back propagation learning algorithm (Multilayer Perceptron – an overview | ScienceDirect Topics, n.d.). Fig. 2 displays different layers of the multilayer perceptron.

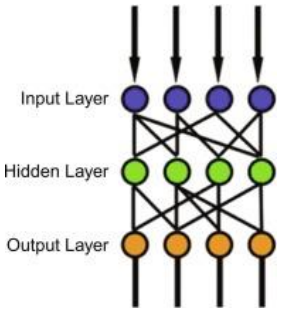

Fig. 2. Multilayer Perceptron Layers (Multilayer Perceptron – an Overview | ScienceDirect Topics, n.d.)

The computations taking place at every neuron in the output and hidden layer are as follows:

$$o(x) = G\big(b(2) + W(2)h(x)\big) \tag{1}$$

$$h(x) = \emptyset(x) = s(b(1) + W(1)x) \tag{2}$$

where $b(1)$ and $b(2)$ are bias vectors; $W(1)$ and $W(2)$ as weight matrices and G and s as activation functions. The set of parameters to learn is the set $\theta = \{W(1), b(1), W(2), b(2)\}$. Typical choices for s include tanh function with $\tanh(a) = (e^a - e^{-a})/(e^a + e^{-a})$ or the logistic sigmoid function, with $\text{sigmoid}(a) = 1/(1 + e^{-a})$) (Multilayer Perceptron – an overview | ScienceDirect Topics, n.d.).

3.3. Multinomial Naïve Bayes

Multinomial Naive Bayes algorithm is a probabilistic learning algorithm being used for natural language processing but can also be used for other purposes such as classification. Multinomial Naïve Bayes algorithm is based on the Bayes theorem. Naive Bayes algorithm is a collection of many algorithms based on the principle that each feature being classified is not related in any way to other feature. The presence or absence of one feature does not affect the presence or absence of another feature (Multinomial Naive Bayes Explained: Function, Advantages & Disadvantages, Applications in 2021 | UpGrad Blog, n.d.).

Bayes theorem (given by Thomas Bayes) is based on the formula:

$$P(A|B) = P(A) * \frac{P(B|A)}{P(B)} \tag{3}$$

read as "probability of class A given B" (Multinomial Naive Bayes Explained Function, Advantages & Disadvantages, Applications in 2021 | UpGrad Blog, n.d.), where $P(B)$ is the prior probability of B, $P(A)$ is prior probability of class A, and $P(B|A)$ is the occurrence of predictor B given class A probability (Multinomial Naive Bayes Explained: Function, Advantages & Disadvantages, Applications in 2021 | UpGrad Blog, n.d.).

3.4. Logistic Regression

Logistic regression is another powerful machine learning algorithm based on supervised learning. It is mostly used for binary classification problems. In other words logistic regression is a linear regression used for classification task (Logistic Regression – an overview | ScienceDirect Topics, n.d.).

It normally uses a logistic function to model a binary output variable.

$$Logistic\ function = 1/(1 + e^{-x}) \tag{4}$$

where x is the input variable (Logistic Regression – an overview | ScienceDirect Topics, n.d.). The difference between linear regression and logistic regression is that the range of logistic regression is limited to between 0 and 1 only (Logistic Regression – an overview | ScienceDirect Topics, n.d.). In addition to this there is no need of linear relationship between input variables and output variables as far as logistic regression is concerned (Logistic Regression – an overview | ScienceDirect Topics, n.d.).

The following Fig. 3 shows logistic regression applied to the range -20 to 20 (Logistic Regression – an overview | ScienceDirect Topics, n.d.).

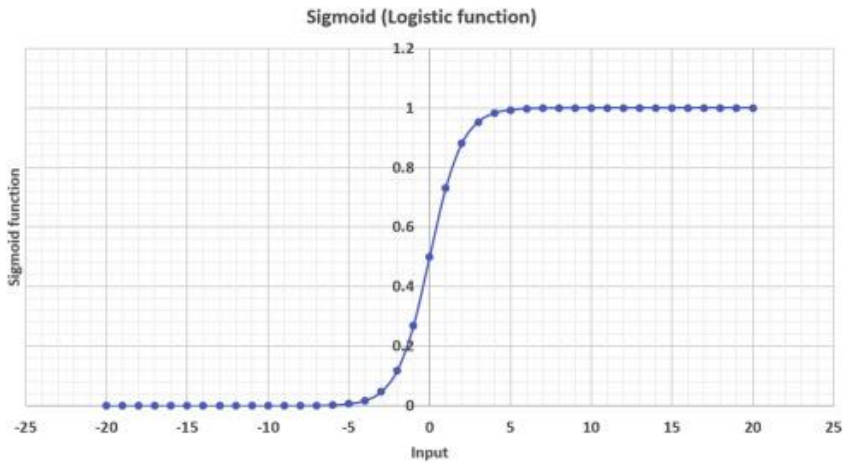

Fig. 3. Logistic regression applied to the range (-20, 20)
(Logistic Regression – an overview | ScienceDirect Topics, n.d.)

3.5. Decision Tree

Decision tree is the most powerful and popular tool for classification and prediction problems (Decision Tree – GeeksforGeeks, n.d.). It is a tree-like structure in which each internal node represents an attribute test and branches designate test outcomes. The leaf nodes represent the class label (Decision Tree – GeeksforGeeks, n.d.).

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making (Decision Trees in Machine Learning | by Prashant Gupta | Towards Data Science, n.d.). It uses a tree-like structure of decisions. In addition to being extensively used in data mining, it has also been extensively used in machine learning for making decisions (Decision Trees in Machine Learning | by Prashant Gupta | Towards Data Science, n.d.).

Following is the top-down approach of the tree construction algorithm (Data Jabberwocky: Decision Tree Mathematical Formulation, n.d.):

1. A search method to compose the tree.
2. Node splitter – to split nodes.
3. Stop criterion – rule to stop the process.
4. Split acceptor – the rule that decides either to accept the best split of a node or to make it a leaf.
5. Split prospects estimator – procedure that determines order of the nodes.
6. Split – order in which nodes need to be split.
7. Decision-making module – provides decisions for data items.
8. Optional data transformations – performs preprocessing of data (Data Jabberwocky: Decision Tree Mathematical Formulation, n.d.).

Entropy and information gain are used when dealing with decision tree. Entropy (E) is a way of measuring how mixed a column is (Entropy and Information Gain in Decision Trees | by Jeremiah Lutes | Towards Data Science, n.d.). Specifically, it is used to measure

the disorder while as information gain (IG) is used to determine the best features/attributes for split in a decision tree. Formulas for entropy and gain are as as under (Entropy: How Decision Trees Make Decisions | by Sam T | Towards Data Science, n.d.):

$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i \tag{5}$$

$$IG(Y, X) = E(Y) - E\left(\frac{Y}{X}\right) \tag{6}$$

4. METHODOLOGY

4.1. Data Source

For this study, the dataset was obtained from the Framingham Heart Study which is a long-term, ongoing cardiovascular study of people from Framingham city in Massachusetts, US to estimate the ten year risk of developing coronary heart disease in order to assess the 10-year cardiovascular disease risk. It started with 5209 adult participants in its first generation and is currently in its third generation. It is carried out by National Heart, Lung, and Blood Institute staffed by various Boston University professionals. Cerebrovascular events, peripheral artery disease and heart failure were subsequently added as disease outcomes for the 2008 framingham risk score on top of coronary heart disease. The said dataset contains 4240 records with 15 attributes and one target attribute (Framingham Heart Study, n.d.). Fig. 4 shows the matrix of correlations among various attributes of the dataset that shows how the attributes of the dataset are related to each other and with the target attribute.

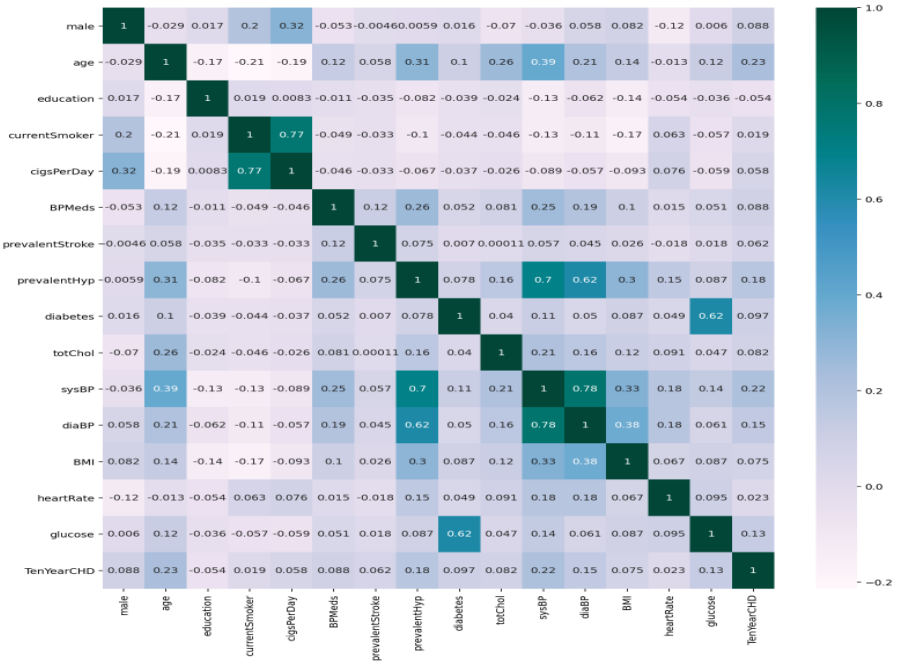

Fig. 4. Correlation Matrix

Tab. 1 shows the names and description of all 16 attributes.

Tab. 1. List of attributes along with their description

S. No.	Attributes	Description	Data type
1.	*Male*	Gender male or female	Nominal: "1" means "male", "0" means "female"
2.	*Age*	Age of the patient	Continuous
3.	*Education*	Education of patient	Categorical
4.	*Current Smoker*	Whether or not the patient is a current smoker	Nominal
5.	*CigsPerDay*	Number of cigarettes smoked per day	Continuous
6.	*BPMeds*	Whether or not patient was on blood pressure medication	Nominal
7.	*Prevalent Stroke*	Whether patient had any stroke previously	Nominal
8.	*PrevalentHyp*	Whether or not the patient was hypertensive	Nominal
9.	*Diabetes*	Whether or not the patient had diabetes	Nominal
10.	*TotChol*	Total cholesterol level	Continuous
11.	*SysBP*	Systolic blood pressure	Continuous
12.	*DiasBP*	Diastolic blood pressure	Continuous
13.	*BMI*	Body Mass Index	Continuous
14.	*Heart Rate*	Heart rate of patient	Continuous
15.	*Glucose*	Glucose level of patient	Continuous
16.	*TenYearCHD (Target variable)*	Whether the patient has 10-year future risk of Coronary Heart Disease (CHD)	Binary: "1", means "Yes", "0" means "No"

As shown in Tab. 1, there were a total of 16 attributes: *Male* – the gender of the patient. It is a binary variable named 'male' in the dataset, "1" means "male" and "0" means "female". *Age* – age of the patient (in years) at medical examination time. *Education* – a categorical variable about patient's education ("1" means "some high school", "2" means "high school/GED", "3" means "some college/vocational school", and "4" means "college"). *CurrentSmoker* – is the patient current smoker at the time of examination. *CigsPerDay* – number of cigarettes smoking each day. *BPMeds* – whether using any anti-hypertensive medication at examination. *PrevalentStroke* – whether any prevalent stroke ("0" means free of disease). *PrevalentHyp* – whether prevalent hypertensive. *Diabetes* – whether diabetic. *TotChol* – total cholesterol in mg/dL. *SysBP* – systolic blood pressure in mmHg. *DiasBP* – diastolic blood pressure in mmHg. *BMI* – body mass index (weight in kg/ height in m^2).

HeartRate – heart rate in beats/minute. *Glucose* – glucose level in mg/dL and finally the target variable *TenYearCHD* – whether risk of ten year coronary heart disease in future ("1" means "Yes" and "0" means "No").

4.2. Validation Schemes and Evaluation Metrics

4.2.1. Validation Schemes

In this study, experiments were performed using the train-test-split scheme. The 70–30 train-test data partitioning scheme was used. That is, the dataset was split into two parts using 70–30% split. Therefore, the classifiers were trained using 2561 records of patients in the first experiment and 2228 records in the next one and tested using remaining 1097 and 955 records for first and second (without outliers) experiment respectively.

4.2.2. Evaluation Metrics

To evaluate the performance of classifiers, different evaluation metrics including accuracy, precision, recall and f1-score were used. *Accuracy* is the percentage of correctly classified subjects in the test dataset. *Recall* conveys information about the percentage of correctly classified subjects while *precision* conveys information about correctly classified healthy subjects. The formulation of these evaluation metrics is given as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

where: TP denotes number of true positives, FP denotes number of false sitives, TN denotes number of true negatives and FN denotes number of false negatives.

$$Recall = TP \,/\, (TP + FN) \tag{8}$$

$$Precision = TN \,/\, (TN + FP) \tag{9}$$

4.3. Classification

Fig. 5 shows the block diagram of the proposed system. Firstly, preprocessing of the aforementioned dataset was done. The dataset contained 4240 records with 15 attributes and one target attribute. Out of 4240 rows, there were 582 rows with null values. The columns that contained missing values were *education*, *cigsPerDay*, *totChol*, *heartRate*, *BPMeds*, *BMI* and *glucose*. The number wise null values in each column were as: *education* (105 null values), *cigsPerDay* (29 null values), *totChol* (50 null values), *heartRate* (1 null value), *BPMeds* (53 null values), *BMI* (19 null values) and *glucose* (388 null values). The number and percentage wise null values in each column are shown in Tab. 2.

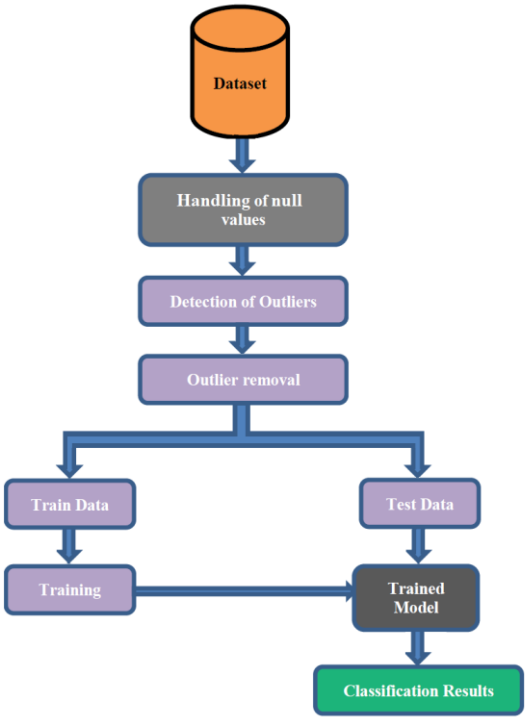

Fig. 5. Block diagram of the proposed system

Null values in a dataset can cause hindrances for the classifiers during the prediction process. Therefore, in this study null values were handled by removing all the rows with null values from the dataset.

Tab. 2. Percentage of missing values for each attribute

Attribute	Number of missing values	Percentage of missing values (in %)
Male	0	0.00
Age	0	0.00
Education	105	2.48
Current Smoker	0	0.00
Cigarettes per day	29	0.68
BPMeds	53	1.25
Prevalent stroke	0	0.00
Prevalent hypertensive	0	0.00
Diabetes	0	0.00
Total cholesterol	50	1.18
SysBP	0	0.00
DiasBP	0	0.00
BMI	19	0.45
Heart rate	1	0.02
Glucose	388	9.15

After removing all the rows with null values, our dataset was now left with 3658 records of patients only.

In the next step, the task of outlier detection was performed. Out of all attributes, *age*, *totChol*, *sysBP*, *diaBP*, *BMI*, *heartRate* and *glucose* were the only attributes with continuous values. Therefore the outliers were detected for these columns only. The figure (Fig. 6) shows the plot representation of outliers for these attributes.

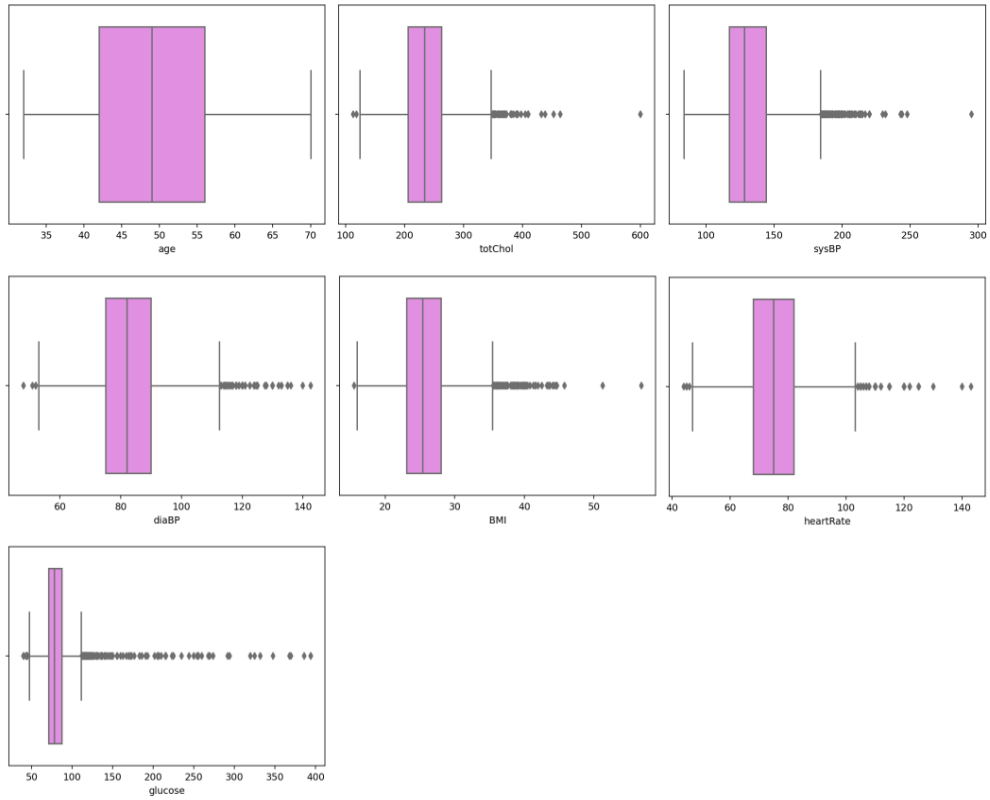

Fig. 6. Plot showing outliers

As seen in the figure, attributes *totChol*, *sysBP*, *diaBP*, *BMI*, *heartRate* and *glucose* contained outliers. Fig. 7 shows the distribution of values for these columns.

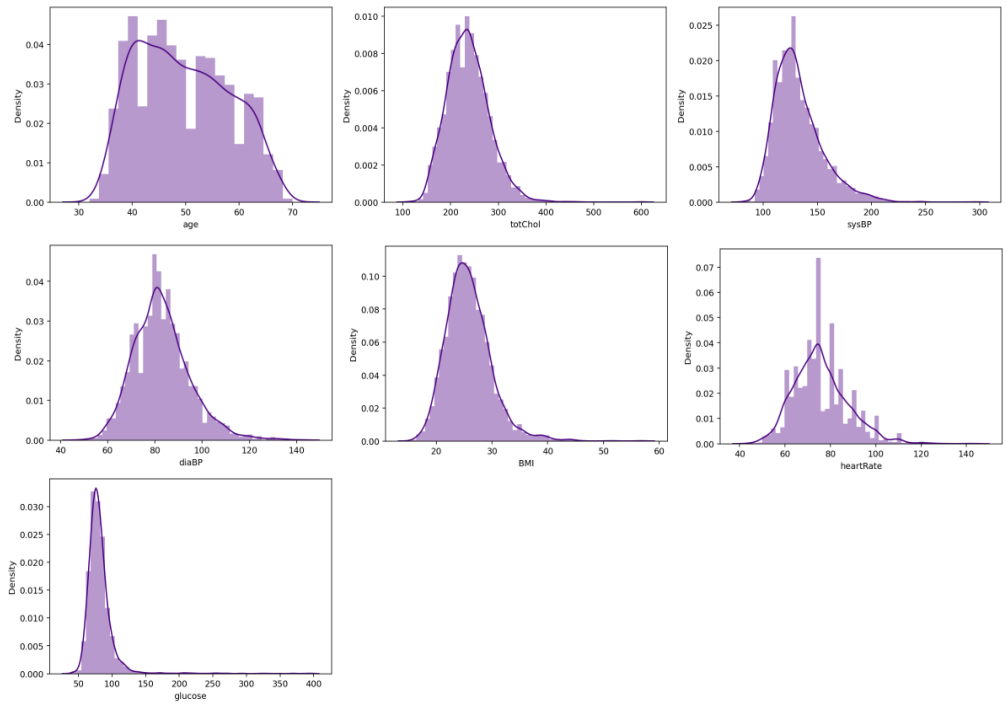

Fig. 7. Plot showing distributions for the Continuous values

The number of outliers in each of these columns was as; *totChol* (46 outliers), *sysBP* (116 outliers), *diaBP* (32 outliers), *BMI* (66 outliers), *heartRate* (60 outliers) and *glucose* (155 outliers) (Tab. 3). Therefore, the outliers were handled by removing all the outliers from *totChol*, *sysBP*, *diaBP*, *BMI*, *heartRate* and *glucose* columns.

Tab. 3. Number of outliers present

Attribute	Number of Outliers
totChol	46
sysBP	116
diaBP	32
BMI	66
heartRate	60
glucose	155

After removing outliers from our data, the number of instances in our data got reduced from 3658 to 3183 records. Fig. 8 shows the distribution of continuous values after outlier treatment for these columns.

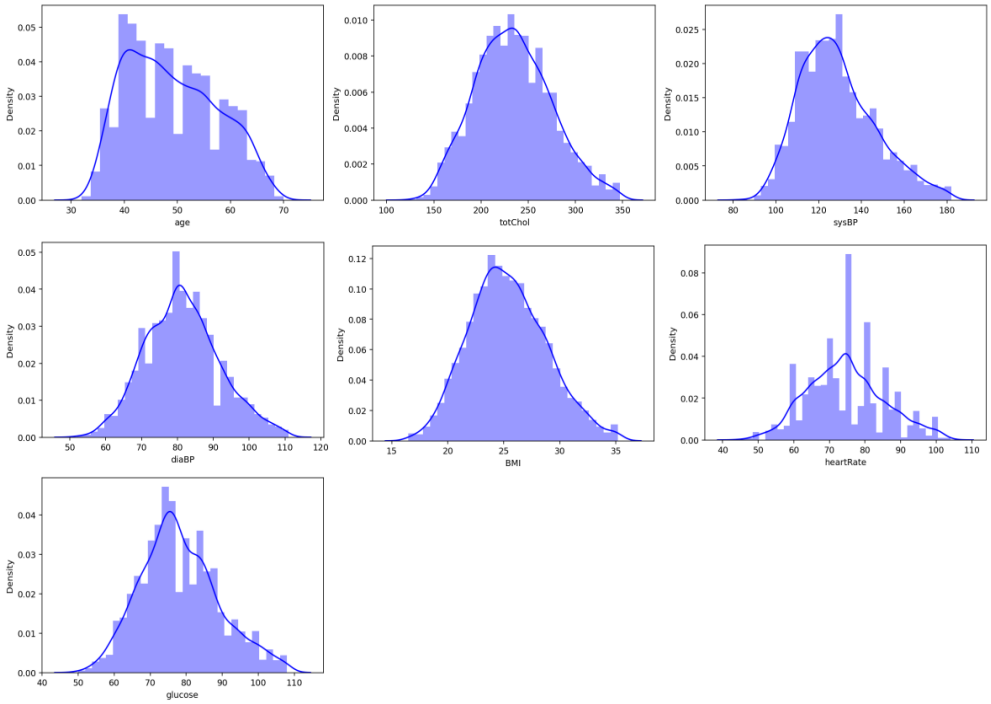

Fig. 8. Plot showing distributions after Outlier Treatment

After treatment of outliers, the distribution looked normal (as shown in Fig. 8). After preprocessing of data, the dataset was then divided into 2 parts: training set and test set where 70% of the data was used for training purpose and remaining 30% was used for testing. In this study, no feature selection technique was used therefore the whole feature set was used for training and testing of classifiers. The machine learning classifiers that were used for this study include: k-nearest neighbor, multi-layer perceptron, support vector machine, multinomial naïve bayes, logistic regression, decision tree and random forest algorithms. Tab. 4 shows the performance metrics of each of these machine learning classifiers.

The same data was then used for training ensemble classifiers. Tab. 5 shows the performance metrics of these classifiers after outlier treatment.

5. RESULTS AND DISCUSSION

To evaluate the performance of the classifiers, the pre-processed data was supplied to all the aforementioned classifiers in order to calculate their efficiencies. All the computations were done on Intel Core I3 processor on windows 10 operating system. Code was written in *python* language using Anaconda3 *Spyder* platform.

In the first experiment, 2561 records were used to train all the mentioned 13 machine learning classifiers without removal of outliers and were then tested using the remaining 1097 records to evaluate their efficiencies. Tab. 4 shows the performance achieved by various classifiers without outlier treatment.

Tab. 4. Performance achieved by various classifiers without outlier treatment

Classifier	Performance Metrics			
	Precision	**Recall**	**F1 Score**	**Accuracy**
K Nearest Neighbor (KNN)	0.78	0.83	0.79	0.8324
Multilayer Perceptron (MLP)	0.70	0.84	0.76	0.8361
Multinomial Naïve Bayes (MNB)	0.78	0.82	0.79	0.8151
Logistic Regression (LR)	0.83	0.84	0.79	0.8443
Decision Tree (DT)	0.77	0.84	0.77	0.8361
Random Forest (RF)	0.77	0.84	0.77	0.8361
Support Vector Machine (SVM)	0.82	0.65	0.70	0.6503
Majority Voting	0.80	0.84	0.78	0.8397
Weighted Average	0.80	0.84	0.77	0.8379
Bagging	0.81	0.84	0.79	0.8424
AdaBoost	0.78	0.83	0.79	0.8270
Gradient Boosting	0.79	0.84	0.78	0.8370
XGBoost	0.76	0.82	0.78	0.8179

Tab. 5. Performance achieved by classifiers after outlier treatment

Classifier	Performance Metrics			
	Precision	**Recall**	**F1 Score**	**Accuracy**
K Nearest Neighbor (KNN)	0.78	0.85	0.80	0.8503
Multilayer Perceptron (MLP)	0.79	0.86	0.80	0.8597
Multinomial Naïve Bayes (MNB)	0.78	0.74	0.76	0.7382
Logistic Regression (LR)	0.82	0.86	0.80	0.8639
Decision Tree (DT)	0.77	0.86	0.80	0.8565
Random Forest (RF)	0.81	0.86	0.81	0.8628
Support Vector Machine (SVM)	0.84	0.66	0.71	0.6597
Majority Voting	0.80	0.86	0.80	0.8628
Weighted Average	0.88	0.86	0.80	0.8649
Bagging	0.82	0.86	0.81	0.8639
AdaBoost	0.79	0.85	0.81	0.8450
Gradient Boosting	0.79	0.85	0.81	0.8482
XGBoost	0.78	0.84	0.80	0.8419

Then in next experiment, the same set of classifiers was again trained using the training set but this time the outliers were first removed from the data. A total of 475 rows were detected that contained null values and after being removed, the dataset was left with 3183 records of patients only. Therefore in the second experiment, only 2228 instances (70%) of patients were used for training of classifiers and remaining 955 (30%) records were used for evaluating their performance. Tab. 5 shows the performance achieved by various machine learning classifiers after outlier treatment. Fig. 9 shows performance comparison of all classifiers with and without outlier treatment using bar-plot (Fig. 9).

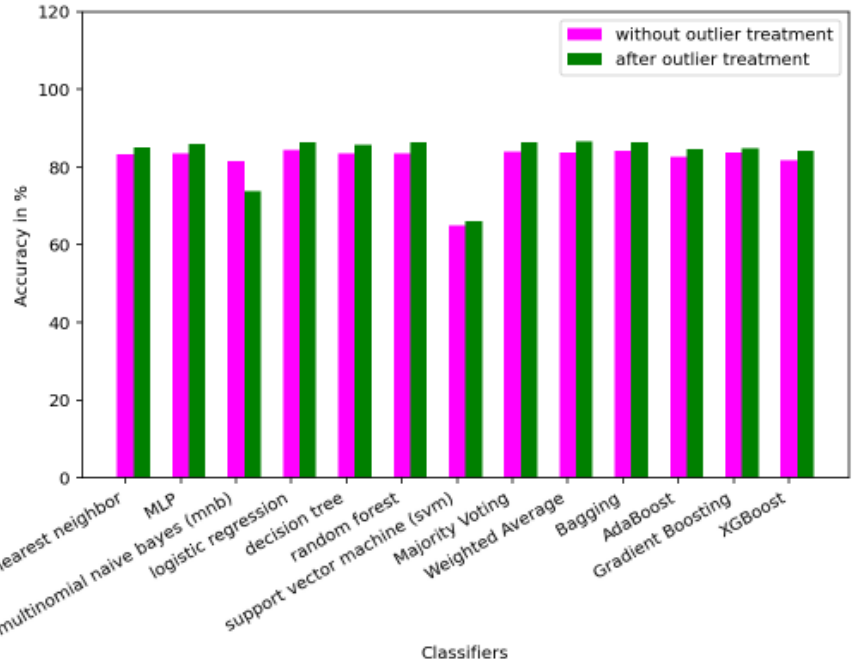

Fig. 9. Bar-plot showing differences in accuracies of classifiers with and without outlier treatment

As shown in the figure (Fig. 9), it can be clearly seen that there was a difference in the accuracies achieved by each of these machine learning classifiers with and without the outlier treatment excluding the one that is multinomial naïve bayes classifier for which accuracy got decreased upon removal of outliers. All the rest of the classifiers showed an accuracy improvement when outliers were removed from the data.

With removal of outliers, the increase in the percentage of accuracies for each of these classifiers was as k-nearest neighbor (1.79%), multilayer perceptron (2.36%), logistic regression (1.96%), decision tree (2.04%), random forest (2.67%), support vector machine (0.94%), majority voting (2.31%), weighted average (2.70%), bagging (2.15%), adaboost (1.80%), gradient boosting (1.12%) and xgboost (2.4%) respectively. Therefore, it was concluded that outliers in data affect the overall efficiency of the machine learning classifiers.

6. CONCLUSION

Heart is a vital organ of the human body performing the crucial function of pumping blood to different body parts. A slight change in the normal functioning of the heart can lead to imbalance in the functioning of the whole body. Heart diseases are the main reason behind deaths in the world today. So, detection of heart disease at an early stage can help in saving numerous human lives. Heart disease is a general term for various types of heart conditions and one such kind is the coronary heart disease. Coronary heart disease is caused by blockage of heart arteries by fat deposits around heart. The aim of this paper is prediction of coronary heart disease using machine learning. Various machine learning base (such as k-nearest neighbor, multilayer perceptron, multinomial naïve bayes, logistic

regression, decision tree, random forest, support vector machine) and ensemble classifiers have been used in this study. The dataset was obtained from the Framingham Heart Study database comprising 4240 instances with 15 attributes for each instance. Firstly, the data was preprocessed by removing all the rows containing null values from the data. Then in the next step the data was checked for presence of any outliers and hence removed accordingly. Then finally in the last step, various machine learning base and ensemble classifiers were trained and tested using the given dataset for predicting the coronary heart disease first including outliers and then in the second experiment without outliers. As per our results, classifiers performed better in the second experiment where the outliers were first removed from the data as compared to the previous experiment where the outliers were also included. In addition to this, the classifier that came out with the best perfromance among all was weighted average ensemble classifier achieving an accuracy of almost 86.50% in the second experiment. Therefore, it was concluded that outliers in data affect the overall efficiency of the machine learning classifiers.

In future, accuracy can further be improved by reducing the number of features under consideration.

Acknowledgments

The authors of this paper would like to extend their gratitude to all those people/organizations that provided their expertise and assistance at different points during the course of this study.

Conflicts of Interest

The authors have no conflicts of interest to declare.

REFERENCES

Ashraf, M., Zaman, M., & Ahmed, M. (2018a). Using ensemble stackingc method and base classifiers to ameliorate prediction accuracy of pedagogical data. *Procedia Computer Science*, *132*(Iccids), 1021–1040. https://doi.org/10.1016/j.procs.2018.05.018

Ashraf, M., Zaman, M., & Ahmed, M. (2018b). Performance analysis and different subject combinations: an empirical and analytical discourse of educational data mining. *Proceedings of the 8th International Conference Confluence 2018 on Cloud Computing, Data Science and Engineering, Confluence 2018* (pp. 287–292). IEEE. https://doi.org/10.1109/CONFLUENCE.2018.8442633

Ashraf, M., Zaman, M., & Ahmed, M. (2019). To ameliorate classification accuracy using ensemble vote approach and base classifiers. In *Advances in Intelligent Systems and Computing* (vol 813). Springer Singapore. https://doi.org/10.1007/978-981-13-1498-8_29

Ashraf, M., Zaman, M., & Ahmed, M. (2020). An intelligent prediction system for educational data mining based on ensemble and filtering approaches. *Procedia Computer Science*, *167*(2019), 1471–1483. https://doi.org/10.1016/j.procs.2020.03.358

Bashir, S., Khan, Z. S., Hassan Khan, F., Anjum, A., & Bashir, K. (2019). Improving Heart Disease Prediction Using Feature Selection Approaches. *Proceedings of 2019 16th International Bhurban Conference on Applied Sciences and Technology,* (pp. 619–623). IEEE. https://doi.org/10.1109/IBCAST.2019.8667106

Benhar, H., Idri, A., & Fernández-Alemán, J. L. (2019). A Systematic Mapping Study of Data Preparation in Heart Disease Knowledge Discovery. *Journal of Medical Systems*, *43*(1), 17. https://doi.org/10.1007/s10916-018-1134-z

Cardiovascular (Heart) Diseases: Types and Treatments. (n.d.). Retrieved January 8, 2022 from https://www.webmd.com/heart-disease/guide/diseases-cardiovascular

Chandra Shekar, K., Chandra, P., & Venugopala Rao, K. (2019). An Ensemble Classifier Characterized by Genetic Algorithm with Decision Tree for the Prophecy of Heart Disease. In *Lecture Notes in Networks and Systems* (Vol. 74). Springer Singapore. https://doi.org/10.1007/978-981-13-7082-3_2

Coronary artery disease: Causes, symptoms, and treatment. (n.d.). Retrieved December 22, 2021 from https://www.medicalnewstoday.com/articles/184130

Coronary heart disease – NHS. (n.d.). Retrieved December 22, 2021 from https://www.nhs.uk/conditions/coronary-heart-disease/

Coronary Heart Disease | NHLBI, NIH. (n.d.). Retrieved December 22, 2021 from https://www.nhlbi.nih.gov/health-topics/coronary-heart-disease

Data Jabberwocky: Decision Tree Mathematical Formulation. (n.d.). Retrieved December 26, 2021 from http://fiascodata.blogspot.com/2018/08/decision-tree-mathematical-formulation.html

Decision Tree – GeeksforGeeks. (n.d.). Retrieved December 26, 2021 from https://www.geeksforgeeks.org/decision-tree/

Decision Trees in Machine Learning | by Prashant Gupta | Towards Data Science. (n.d.). Retrieved December 26, 2021 from https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

Dun, B., Wang, E., & Majumder, S. (2016). Heart Disease Diagnosis on Medical Data Using Ensemble Learning. *Computer Science*, *1*(1), 1–5.

El-Shafeiy, E. A., El-Desouky, A. I., & Elghamrawy, S. M. (2018). Prediction of Liver Diseases Based on Machine Learning Technique for Big Data. *Advances in Intelligent Systems and Computing*, *723*, 362–374. https://doi.org/10.1007/978-3-319-74690-6_36

Entropy: How Decision Trees Make Decisions | by Sam T | Towards Data Science. (n.d.). Retrieved December 26, 2021 from https://towardsdatascience.com/entropy-how-decision-trees-make-decisions-2946b9c18c8

Entropy and Information Gain in Decision Trees | by Jeremiah Lutes | Towards Data Science. (n.d.). Retrieved December 26, 2021 from https://towardsdatascience.com/entropy-and-information-gain-in-decision-trees-c7db67a3a293

Framingham Heart Study. (n.d.). Retrieved September 9, 2021 from https://framinghamheartstudy.org/

Gokulnath, C. B., & Shantharajah, S. P. (2019). An optimized feature selection based on genetic approach and support vector machine for heart disease. *Cluster Computing*, *22*(s6), 14777–14787. https://doi.org/10.1007/s10586-018-2416-4

Heart disease – Symptoms and causes - Mayo Clinic. (n.d.). Retrieved January 8, 2022 from https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint. (n.d.). Retrieved December 26, 2021 from https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, *15*, 104–116. https://doi.org/10.1016/J.CSBJ.2016.12.005

Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, 100203. https://doi.org/10.1016/j.imu.2019.100203

Less than $1: How WHO thinks that can save 7 million lives. (n.d.). Retrieved January 9, 2022 from https://www.downtoearth.org.in/news/health/less-than-1-how-who-thinks-that-can-save-7-million-lives-80679

Logistic Regression - an overview | ScienceDirect Topics. (n.d.). Retrieved December 26, 2021 from https://www.sciencedirect.com/topics/computer-science/logistic-regression

Mir, N. M., Khan, S., Butt, M. A., & Zaman, M. (2016). An experimental evaluation of Bayesian classifiers applied to intrusion detection. *Indian Journal of Science and Technology*, *9*(12), 1–13. https://doi.org/10.17485/ijst/2016/v9i12/86291

Mohd, R., Butt, M. A., & Baba, M. Z. (2020). GWLM–NARX: Grey Wolf Levenberg–Marquardt-based neural network for rainfall prediction. *Data Technologies and Applications*, *54*(1), 85–102. https://doi.org/10.1108/DTA-08-2019-0130

Mohd, R., Butt, M. A., & Baba, M. Z. (2019). SALM-NARX: Self adaptive LM-based NARX model for the prediction of rainfall. *Proceedings of the International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), I-SMAC 2018* (pp. 580–585). IEEE. https://doi.org/10.1109/I-SMAC.2018.8653747

Multilayer Perceptron – an overview | ScienceDirect Topics. (n.d.). Retrieved December 26, 2021 from https://www.sciencedirect.com/topics/computer-science/multilayer-perceptron

Multinomial Naive Bayes Explained: Function, Advantages & Disadvantages, Applications in 2021 | upGrad blog. (n.d.). Retrieved December 26, 2021 from https://www.upgrad.com/blog/multinomial-naive-bayes-explained/

Otoom, A. F., Abdallah, E. E., Kilani, Y., & Kefaye, A. (2015). Effective Diagnosis and Monitoring of Heart Disease. *International Journal of Software Engineering and Its Applications*, *9*(1), 143–156.

Riyaz, L., Butt, M. A., Zaman, M., & Ayob, O. (2022). Heart Disease Prediction Using Machine Learning Techniques: A Quantitative Review. *Advances in Intelligent Systems and Computing* (pp. 81–94). Springer. https://doi.org/10.1007/978-981-16-3071-2_8

Sakai, K., & Yamada, K. (2019). Machine learning studies on major brain diseases: 5-year trends of 2014–2018. *Japanese Journal of Radiology*, *37*, 34–72. https://doi.org/10.1007/s11604-018-0794-4

Salvatore, C., Cerasa, A., Castiglioni, I., Gallivanone, F., Augimeri, A., Lopez, M., Arabia, G., Morelli, M., Gilardi, M. C., & Quattrone, A. (2014). Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and Progressive Supranuclear Palsy. *Journal of Neuroscience Methods*, *222*, 230–237. https://doi.org/10.1016/J.JNEUMETH.2013.11.016

Shinde, R., Arjun, S., Patil, P., & Waghmare, P. J. (2015). An Intelligent Heart Disease Prediction System Using K-Means Clustering and Naïve Bayes Algorithm. *International Journal of Computer Science and Information Technolog*, *6*(1), 637–639.

Takci, H. (2018). Improvement of heart attack prediction by the feature selection methods. *Turkish Journal of Electrical Engineering and Computer Sciences*, *26*(1), 1–10. https://doi.org/10.3906/elk-1611-235

Thaiparnit, S., Kritsanasung, S., & Chumuang, N. (2019). A Classification for Patients with Heart Disease Based on Hoeffding Tree. *JCSSE 2019 – 16th International Joint Conference on Computer Science and Software Engineering: Knowledge Evolution Towards Singularity of Man-Machine Intelligence* (pp. 352–357). IEEE. https://doi.org/10.1109/JCSSE.2019.8864158

Wei, S., Zhao, X., & Miao, C. (2018). A comprehensive exploration to the machine learning techniques for diabetes identification. *IEEE World Forum on Internet of Things, WF-IoT 2018 - Proceedings*, (pp. 291–295). IEEE. https://doi.org/10.1109/WF-IOT.2018.8355130

Wu, C. C., Yeh, W. C., Hsu, W. D., Islam, M. M., Nguyen, P. A., Poly, T. N., Wang, Y. C., Yang, H. C., & Li, Y. C. (2019). Prediction of fatty liver disease using machine learning algorithms. *Computer Methods and Programs in Biomedicine*, *170*, 23–29. https://doi.org/10.1016/J.CMPB.2018.12.032

Zaman, M., Kaul, S., & Ahmed, M. (2020). Analytical comparison between the information gain and gini index using historical geographical data. *International Journal of Advanced Computer Science and Applications*, *11*(5), 429–440. https://doi.org/10.14569/IJACSA.2020.0110557

Zaman, M., Quadri, S. M. K., & Butt, M. A. (2012). Information translation: A practitioners approach. *Lecture Notes in Engineering and Computer Science*, *1*, 45–47.