Vol. 06, No. 4 (2024) 1509-1522, doi: 10.24874/PES06.04.010

Proceedings on Engineering Sciences

www.pesjournal.net

COMPARATIVE MACHINE LEARNING ALGORITHM FOR CARDIOVASCULAR DISEASE PREDICTION

Ashish Mishra[1]
Jyoti Mishra
Victor Hugo
Aloísio Vieira Lira Neto

Received 10.04.2024.
Received in revised form 23.09.2024.
Accepted 16.10.2024.
UDC – 004.032.26:616.1

Keywords:

Heart Disease Prediction, Parameters, Machine Learning, Random Forest, Decision Tree

A B S T R A C T

In the present study, It used to categorise heart illness in the Cleveland UCI repository. It visually describes the dataset, operational parameters, and predictive analytics development. Machine learning (ML) begins with data preparation. The technique uses an ML model and key parameters to predict cardiovascular illness in patients. The dataset comprises 14 heart disease characteristics for this investigation. The preliminary examination and evaluation predicted heart problems. The dataset has 303 samples with 14 features. The information is presented as a percentage of truth. KNN 86%, Decision Trees 79%, Logistic Regression 85%, Naive Bayes 86%, and Support Vector Machines 87% can predict heart disease 89% accurately. The receiver working characteristics show that the random forest technique for heart disease prediction has an 89% diagnostic rate. The proposed method uses the random forest algorithm since it has been shown to be the most effective algorithm for classifying cardiovascular illness.

1. INTRODUCTION

Cardiovascular disease is a group of conditions that affect the heart and blood vessels in the body. There may also be damage to the arteries in the kidneys, heart, eyes, and brain. Cardiovascular diseases can be broken down into four main groups. Coronary Heart Disease is the first. This disease happens when blood flow to the heart muscle stops. This makes the heart work harder, which can cause angina, heart attacks, and heart failure. Here's the second group: strokes and transient ischemic attacks (Mishra, 2020). These happen when blood flow to the brain is blocked or temporarily interrupted. When blood flow to the limbs is stopped, this third type of heart disease happens. This causes a lot of pain in the legs, hair loss on the feet and legs, leg weakness, and sores that don't heal. The aorta, which is the body's largest blood vessel, is affected by the last type. This doesn't hurt, but when it bursts, it causes a life-threatening blood loss.

Heart diseases and coronary heart diseases are types of cardiovascular illness. Coronary artery diseases (CAD) include angina and myocardial infarction (also called a heart attack). In coronary heart diseases (CHD), plaque builds up inside the coronary vessels and causes heart problems. A heart attack is the top cause of death in the world, and if it is not treated quickly, it can lead to

[1] Corresponding author: Ashish Mishra
Email: ashishmishra@ggits.org

serious health problems or even death. Cardiac dysfunction has become a difficult medical issue in recent years. Myocardial infarction kills one patient every minute. Systematizing the technique and educating the patient is essential due to the difficulties of predicting cardiac illness. Worldwide cardiovascular disease risk is high. A physician's cardiovascular risk assessment must be accurate and complete to reduce attack and stroke rates and improve cardiovascular protection (Mishra, 2019). To avoid fatalities, it processes detecting these heart defects as soon as possible. Next, assess the user's cardiovascular disease risk. It solves cardiac disease detection issues, helping clinicians make better decisions. Medical specialists have collected a lot of data that can be analyzed. Hypothesis testing improves heart disease diagnosis and prognosis with machine learning. It is not unexpected that more compound algorithms (sets of rules), such as SVM and Random Forests, provided enhanced outcomes than those of simpler methods. It is crucial to emphasise that hyperparameter taking is frequently required for these procedures to produce findings that may be trusted. Simpler methods also demonstrated their value by producing respectable outcomes. Machine learning in medicine has a very bright future (Khanwalkar, 2024). Imagine a place where there are no medical professionals that specialise in heart disease. With just a little knowledge of a patient's medical history, It can predict with a high degree of accuracy whether a disease would manifest or not.

Heart disease requires cautious management due to its complexity. Failure to do so may result in cardiac damage or early death (Mishra & Bhatt, 2022). Data mining with classification, which is applied from the perspectives of medical research and data mining to analyse various forms of metabolic disorders, greatly benefits both heart disease prediction and data analysis (Medhekar et al., 2013).

The accurate identification of events connected to heart disease has also been attained using decision trees. (Kononenko, 2001).

Different information abstraction methodologies have been attempted to forecast heart illness. Through a variety of readings, a forecasting system that included not only different methodologies but also two or more techniques logically was established in this study. These freshly integrated strategies are known as hybrid techniques (Chen et al., 2011). It uses heart rate time data to illustrate neural networks. Additionally explored is the use of computer-aided decision support systems (CADSS) in research and medicine. According to earlier research, the healthcare sector can forecast diseases more quickly and accurately by applying data mining techniques (Sabarinathan & Sugumaran, 2014). It advises using the GA to detect heart conditions. This strategy makes use of Effective association rules for crossover, tournament selection, and the mutation that leads to the new proposed fitness function deduced from the GA. It will then compare our results to some of the widely used supervised learning approaches (Almarabeh & Amer 2017; Patel et al., 2015). The most effective evolutionary technique, particle swarm optimization (PSO), is presented, and certain heart disease rules are developed. Overall accuracy has enhanced due to the random implementation of the rules using encoding approaches (Soni et al., 2011). Age, sex, and pulse rate are just a few of the indicators that can be used to indicate heart disease. As demonstrated in (Das et al., 2009; Vembandasamy et al., 2015) the ML approach utilising neural networks is presented, and its conclusions are more precise and trustworthy.

The most reliable method for forecasting conditions such as cardiovascular disease and brain disease is universally acknowledged to be neural networks. The proposed technique that It employ to forecast cardiac disease consists of 13 parameters. When compared to the techniques used in earlier studies like (Tasnim & Habiba, 2021), the results demonstrate improved performance. In recent years, the Carotid Artery Stenting (CAS) procedure has grown in popularity as a therapy option. Major adverse cardiovascular events (MACE) are brought on by the CAS in senior cardiovascular sufferers. The analysis becomes crucial. To generate our results, it employs an Artificial Neural Network (ANN), which is effective at predicting cardiac disease (Fatima & Pasha, 2017; Shouman et al., 2012). It highlights the utilisation of neural network techniques, such as projected values and posterior probability from other preceding approaches. When compared to previous studies, the accuracy level of this model, which can reach 89.01 per cent, is excellent. As previously observed in (Waghulde & Patil, 2014; Wiharto et al., 2015), all tests use the Cleveland heart dataset and a neural network NN to improve the performance of cardiac problems. Numerous investigations have indeed been carried out, which have limited the features that may be used in algorithms. The HRFLM approach, in contrast, employs all highlights without any limitations on selecting features. Here, it looks into a hybrid machine-learning algorithm's feature. The results of the experiment show that, compared to other methodologies, our suggested hybrid approach has a better propensity to predict heart illness.

2. RELATED WORK

There is a tonne of work that is directly applicable huge the fields that this issue falls under. In the (Tasnim & Habiba, 2021) Comparative research of multiple machine-learning approaches for heart disease prediction to determine the optimum technique. The most efficient algorithms for heart disease prediction have been discovered to be Support Vector Machine and random forest machine learning techniques. Also, accuracy is further improved by the use of principal component analysis, this research does not include other

necessary parameters to be considered. With the introduction of ANN, the maximum level of accuracy in medical forecasts is promised (Fatima & Pasha, 2017). Heart illness is detected using ANN's backpropagation multilayer perceptron (MLP). When compared to other models that have been employed in the same field, the results are shown to be superior (García et al., 2016). NN, DT, SVM, and Naive Bayes are used to analyse the data of patients with cardiovascular illness that was gathered from the UCI research facility. For output efficiency and precision, various methodologies are contrasted. With a yield of 86.8% for the F-measure, the suggested hybrid strategy outperforms the practices now in use (Pouriyeh et al., 2017). The introduction of classification using convolutional neural networks (CNN) replaces segmentation. During the training phase of this approach, cardiac cycles with various start timings are adjusted using electrocardiogram (ECG) data. Throughout the patient testing phase, CNN can provide features in a range of positions (B. Venkata Lakshmi et al., 2014). Even though the medical sector generates a lot of data, it hasn't always been handled correctly. The novel methods described here save expenses while enhancing heart disease predictability. The many research approaches that It considered in this work for the categorization and prediction of cardiovascular sickness utilising machine learning (ML) and deep learning (DL) mechanisms are very accurate in judging how effective these strategies are (Ghumbre et al., 2011) applied. In the (Sharma et al., 2020) targeted the study of different machine learning techniques for heart disease prediction and finding out the technique that works for the same (Mishra et al., 2023). It has been found that Naive Bayes and Decision trees are not efficient for heart disease prediction whereas random forest and support vector machines are efficient for heart disease prediction. Also, this research does not include another necessary parameter to be considered during model building. It excludes various disease histories (Mishra et al., 2020). This research is based on limited data.

3. METHODOLOGY

In this work, it utilized a R studio rattle to organize the Cleveland UCI repository's heart disease data. It presents the dataset, operational parameters, and the development of predictive analytics in a straightforward visual manner. The preparation of the data is the initial step in the machine learning (ML) process. The next step outputs with higher accuracy, folloItd by feature selection based on DT entropy and segmentation of modelling performance evaluation. The process of feature selection and modelling is repeatedly done for different combinations of features. Table 1 displays the specific information and implemented characteristics of the UCI dataset. In Table 2, the data type and value range are displayed. Each model created using the 13 features and the machine learning methods applied in each iteration are evaluated. Pre-processing of the data is covered in Section A, feature selection is done in Section B using entropy, classification of the data is done in Section C using machine learning, and performance evaluation of the results is covered in Section D. To detect the risk factors for heart disease in the UCI, the three association rules of mining apriori, predictive, and Tertius are used statistically by HRFLM is shown in Figure 1.

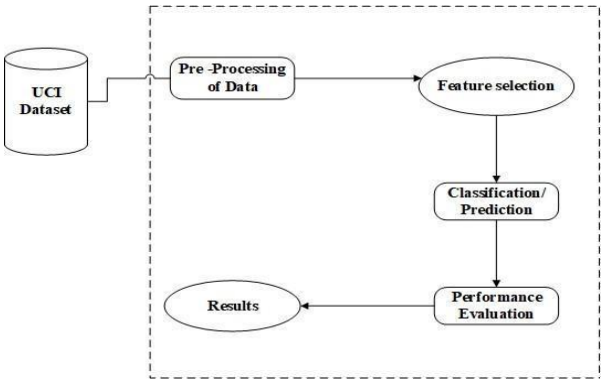

Figure 1. Flow chart of Hybrid Random Forest with Linear Model

Cleveland-specific data collection. Given the available data, it can be said that compared to men, women have a lesser risk of having heart disease. precise diagnosis is essential for the treatment of heart disease. Conventional approaches, hoItver, fall short when it comes to accurate prediction and diagnosis.

The HRFLM method feeds 14 clinical characteristics into an ANN with backpropagation. The obtained results are assessed in comparison to established methods (Amin et al., 2013). Several features are used to precisely diagnose the illness when the risk levels increase (Dangare & Apte, 2012). Due to the characteristics and sensitivity of heart illness, a suitable treatment strategy is necessary. Techniques for data mining are beneficial in therapeutic situations in the realm of medicine. DT, NN, SVM, and KNN are taken into consideration in further data mining approaches. The SVM results outperform the other methods tested in terms of disease prediction accuracy (Zriqat et al., 2017). To identify arrhythmias, the nonlinear technique with a module for monitoring heart activity is introduced. The effectiveness of this approach's performance can be evaluated depending on how accurately the conclusions drawn from the ECG data Itre reached. ANN training is used to predict possible patient anomalies and accurately detect disease (Liu et al., 2017).

KNN, LR, SVM, NN, and Vote are just a few of the data mining approaches and prediction algorithms that have recently acquired prominence for identifying and predicting heart disease (Palaniappan & Awang, 2008). In this study, both vote and a combination of LR and NB are recommended. When the proposed method was tested on the UCI dataset (Masethe & Masethe, 2014), it

was found that it could identify heart disease with an accuracy of 87.4%. Three sets of data from UCI's sites in Cleveland, Switzerland, and Hungary are used to test the probabilistic principal component analysis (PPCA) method. The method takes the vectors that are most similar to each other and uses vector projection to reduce the number of feature dimensions. A radial basis function makes kernel-based SVM possible by allowing the picking of features while minimizing the number of dimensions. According to the methodology, the UCI data sets from Cleveland, Switzerland, and Hungary had respective results of 82.18%, 85.82%, and 91.30% (Singh & Choudhary, 2017). The invention of the hybrid method, which integrates linear regression (LR), multivariate adaptive regression splines (MARS), and artificial neural networks, is the primary original contribution of this study (ANN). Using the provided method, the list of important characteristics was successfully condensed. The remainder of the properties is then added to the ANN. The effectiveness of the hybrid approach's creation is demonstrated using the heart disease datasets (Jabbar et al., 2013).

To forecast cardiovascular disease, it also introduces the Apriori algorithm using SVM and compare it to nine other classification methods. When compared to other existing methods, the classification method's results have demonstrated a greater level of accuracy and performance in the prediction of heart disease (Rajesh et al., 2018). To forecast cardiac disease, feature selection is crucial. For better illness prediction, ANN with backpropagation is advised. The results produced by the application of ANN are highly exact and accurate (Carroll & Miller, 2013). Recurrent Fuzzy Neural Network (RFNN), a genetic technique that integrates fuzzy NN, is used to identify heart disease.

The UCI dataset for Cleveland is used by HRFLM to computationally identify the risk factors for heart disease using the three association rules of mining: apriori, predictive, and Tertius (Kirubha & Priya, 2016). According to the information available, women are less likely than men to get heart disease. For heart disorders to be effectively treated, an accurate diagnosis is crucial. Accurate prediction and diagnosis, hoItver, are not possible using conventional methods (Sharma & Rizvi, 2017).

13 clinical characteristics are fed into an ANN using backpropagation by the HRFLM approach. The acquired results are evaluated in light of accepted methodologies (Amin et al., 2013). When the risk levels rise, a variety of characteristics are employed to precisely diagnose the illness (Dangare & Apte, 2012). The nature and complexity of cardiac disease necessitate the use of an effective treatment strategy. In the field of medicine, therapeutic circumstances benefit from the use of data mining tools. Further data mining techniques take into account DT, NN, SVM, and KNN. In terms of disease prediction accuracy, the SVM findings perform better than the other examined approaches (I. Zriqat et al., 2017). To detect arrhythmias such as bradycardia, tachycardia, atrial, atrial-ventricular flutters, and many more, the nonlinear technique with a module for tracking heart activity is introduced. The effectiveness of this approach's performance can be evaluated based on how precisely inferences It re drawn from the ECG data. Learning with ANN is used to predict probable patient anomalies and accurately detect disease (Liu et al., 2017).

KNN, LR, SVM, NN, and Vote are a few of the data mining methods and prediction algorithms that have recently grown in favour of identifying and predicting cardiac illness. Vote, a novel strategy, and a hybrid strategy combining LR and NB are suggested in this work. The suggested technique has an accuracy of 87.4% in predicting cardiovascular illness when it was evaluated using the UCI dataset (Masethe & Masethe, 2014). Principal component analysis with probability (PPCA) is evaluated using three data sets from the UCI campuses in Cleveland, Switzerland, and Hungary. The technique uses vector projection to reduce the feature dimension and extracts the vectors with a strong correlation. A radial basis function offers the feature selection with minimising dimension, enabling kernel-based SVM. The UCI data sets from Cleveland, Switzerland, and Hungary yielded values of 82.18%, 85.82%, and 91.30%, respectively, by the methodology (Singh & Choudhary, 2017). The creation of the hybrid technique, which integrates linear regression (LR), multivariate adaptive regression splines (MARS), and artificial neural networks, is the study's primary opportunity to contribute (ANN).

To more accurately forecast heart disease, Additionally, it presents the Apriori algorithm using SVM relative to nine other classification approaches. When compared to other existing methods, the classification method's results have shown a higher degree of precision and performance in the diagnosis of cardiovascular illness (Rajesh et al., 2018). To forecast cardiac disease, feature selection is crucial. For better illness prediction, ANN with backpropagation is advised. The results produced by the use of ANN are extremely beneficial, exact and accurate (Carroll & Miller, 2013). Recurrent Fuzzy Neural Network (RFNN), a genetic technique that integrates fuzzy NN, is used to identify heart disease.

The UCI data collection consists of 297 patient records, of which 252 are the remainder for training and some for testing. The analysis led to the conclusion that the results adequate. It is advised to utilize SVM and ANN to predict cardiac disease. This solution uses two separate methods to address the accuracy and testing time assumptions. According to the suggested model, the data records are split into two classes for additional SVM and ANN analysis. After the Back Propagation Neural Network (BPNN) with classification approach has been presented and tested, the precise gene

sequence for hypertension is discovered. This method's precision has grown as more recordings have been made. The effectiveness of the BPNN approaches has been evaluated using a variety of sample sizes during both the training and testing phases. As additional recordings have been made, this method's accuracy has increased.

4. RESULTS AND DISCUSSION

4.1 Libraries import

Importing all necessary libraries should come first. I'll start by utilising Numpy and Pandas. I'll style the plots with rcParams, add colours with Rainbow, and visualise them using Matplotlib's pyplotsubpackage. The sklearn package will be used by me to manage data and build machine learning models.

4.2 Dataset explanation

The dataset used for this investigation includes 14 variables. To identify whether a person is healthy or has a cardiac illness, the independent variable "diagnosis" must be predicted. Studies using the Cleveland database have concentrated on trying to differentiate betIten a disease's presence (values 1, 2, 3, and 4) and its absence (value 0). The question mark (?) denotes that several attribute values are missing. This dataset lacks a header row thus the column names must be manually entered in Table 1.

Features information:

- age - age in years
- sex - sex (1 = male; 0 = female)
- chest_pain - chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
- blood_pressure - resting blood pressure (in mm Hg on admission to the hospital)
- serum_cholestoral - serum cholesterol in mg/dl
- fasting_blood_sugar - fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
- electrocardiographic – resting electrocardiographic results (0 = normal; 1 = having ST-T; 2 = hypertrophy)
- max_heart_rate - maximum heart rate achieved
- induced_angina – exercise-induced angina (1 = yes; 0 = no)
- ST_depression - ST depression induced by exercise relative to rest
- slope - the slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)
- no_of_vessels - number of major vessels (0-3) colored by flourosopy
- thal - 3 = normal; 6 = fixed defect; 7 = reversible defect
- diagnosis - the predicted attribute - diagnosis of heart disease (angiographic disease status) (Value 0 = < 50% diameter narrowing; Value 1 = > 50% diameter narrowing)

4.3 Characteristics involve

The following characteristics presented in Table 1 can be classified (having two or more categories, and each category has a value): heartache, sexual.

Examples of ordinal qualities are fasting blood sugar, ECG, provoked angina, slope, number of vessels, thalamus, and diagnostic (variables having relative ordering or sorting betIten the values).

Table 1. Feature Information

```
# column names in accordance with feature information
col_names=['age','sex','chest_pain','blood_pressure','serum_cholestoral','fasting_blood_sug
ar','electrocardiographic',
'max_heart_rate','induced_angina','ST_depression','slope','no_of_vessels','thal','diagnosis
']

# read the file
df=pd.read_csv("processed.cleveland.data",names=col_names,header=None,na_values="?")

print("Number of records: {}\nNumber of variables: {}".format(df.shape[0],df.shape[1]))

# display the first 5 lines
df.head()
```

```
Number of records: 303
Number of variables: 14
```

```
df.info()
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
age                     303 non-null float64
sex                     303 non-null float64
chest_pain              303 non-null float64
blood_pressure          303 non-null float64
serum_cholestoral       303 non-null float64
fasting_blood_sugar     303 non-null float64
electrocardiographic    303 non-null float64
max_heart_rate          303 non-null float64
induced_angina          303 non-null float64
ST_depression           303 non-null float64
slope                   303 non-null float64
no_of_vessels           299 non-null float64
thal                    301 non-null float64
diagnosis               303 non-null int64
dtypes: float64(13), int64(1)
memory usage: 33.2 KB
```

This dataset only has six missing values, and all variables are recognised as numbers. HoItver, it knows that the majority of the characteristics are categorical and that it is important to discern betIten them from the dataset description is presented in Table 2.

Table 2. Extract Numeric and Find Categorical Ones

```
# extract numeric columns and find categorical ones
numeric_columns=['serum_cholestoral','max_heart_rate','age','blood_pressure','ST_depression
']
categorical_columns=[cforcindf.columnsifcnotinnumeric_columns]
print(categorical_columns)
```

```
['sex', 'chest_pain', 'fasting_blood_sugar', 'electrocardiographic', 'induc
ed_angina', 'slope', 'no_of_vessels', 'thal', 'diagnosis']
```

5. ANALYZE FEATURES, IDENTIFY PATTERNS, AND EXPLORE THE DATA

5.1 Target value

The distribution of the target value must be understood to select the right accuracy metrics and, as a result, appropriately compare various machine learning models shown in Figure 2.

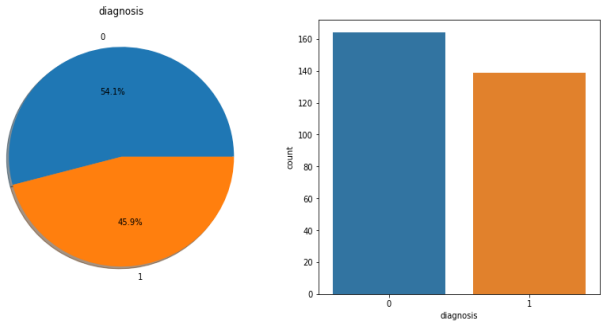

Figure 2. Analyze Features, Identify Patterns and Explore the Data

Since the values 1-4 indicate that a disease is present, it's reasonable to pull them together in Table 3.

Table 3. Count values of explained variable

```
# count values of explained variable
df. diagnosis.value_counts()
0    164
1     55
2     36
3     35
4     13
Name: diagnosis, dtype: int64
```

Now that the distribution of the goal value is almost equal, it is reasonable to use established measures for machine learning modelling, such as accuracy and AUC.

5.2 Numeric features

Let's start with the five numerical columns since there are five of them. Since the emergence of outliers in the dataset may be caused by incorrect input and produce unwanted noise, it is our responsibility to assess their significance. When a data point deviates by more than three standard deviations, it is regarded as an outlier is presented in Table 4.

All extreme (min/max) values could occur in a genuine clinical situation, which is why it was decided to leave them as-is.

By plotting each pair in a scattered form shown in Figure 3, it can develop some sense of the connections betIten the numerical properties. Pair plot technique from Seaborn library is useful for accomplishing this effectively

Table 4. Descriptive Statistics

 # view of descriptive statistics

df[numeric_columns]. describe ()

	serum_cholestoral	max_heart_rate	age	blood_pressure	ST_depression
count	303.000000	303.000000	303.000000	303.000000	303.000000
mean	246.693069	149.607261	54.438944	131.689769	1.039604
std	51.776918	22.875003	9.038662	17.599748	1.161075
min	126.000000	71.000000	29.000000	94.000000	0.000000
25%	211.000000	133.500000	48.000000	120.000000	0.000000
50%	241.000000	153.000000	56.000000	130.000000	0.800000
75%	275.000000	166.000000	61.000000	140.000000	1.600000
max	564.000000	202.000000	77.000000	200.000000	6.200000

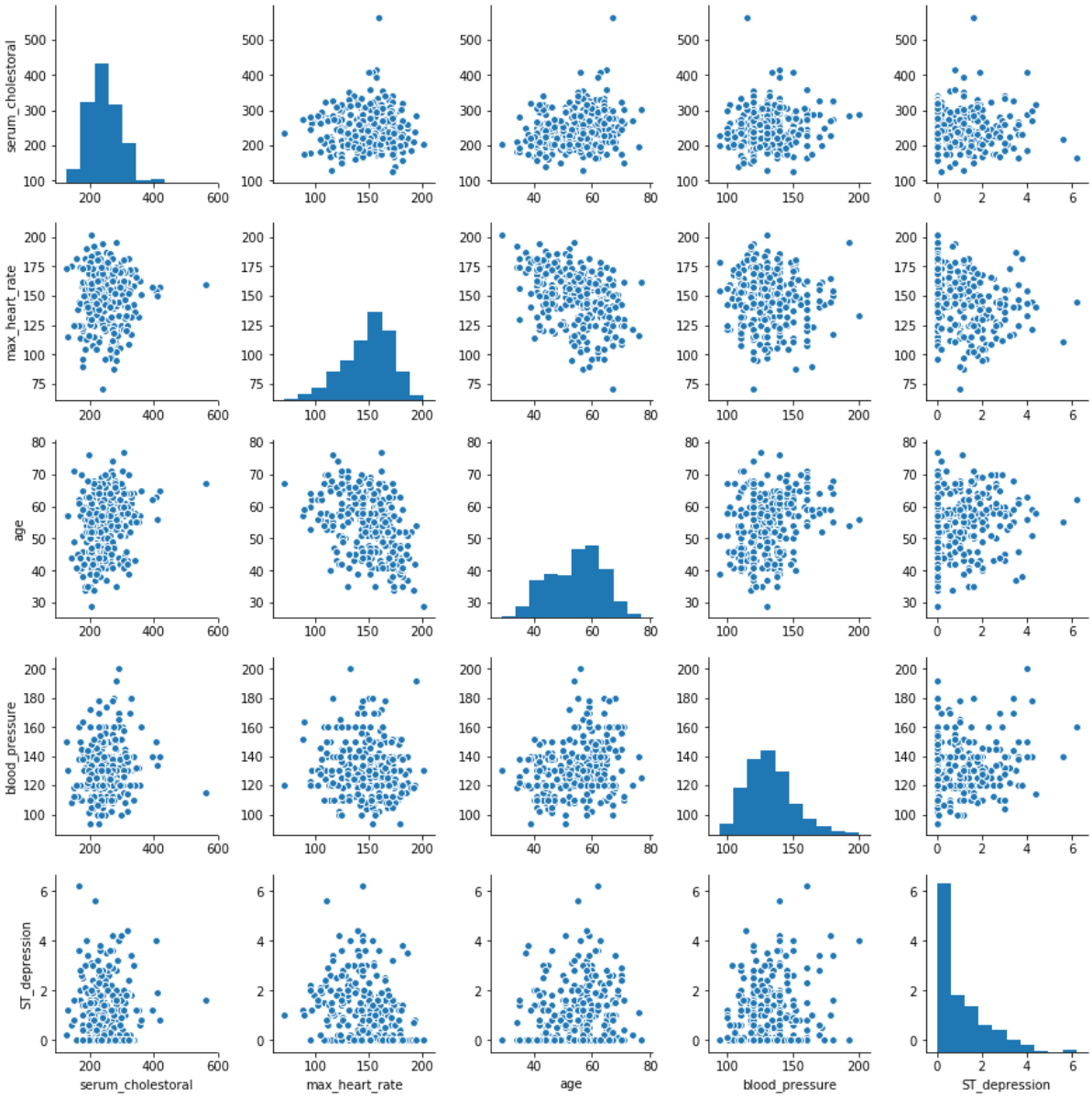

Figure 3. Plotting Each Pair in a Scattered Form

The aforementioned graphs lead me to conclude that none of the value pairings is explicitly showing a high correlation, hence there is no immediate need to drop any features. In addition, I see an intuitively-seeming negative association betIten "age" and "max heart rate" and a positive correlation betIten "age" and "blood pressure."

It can verify whether the information above is accurate using a correlation matrix. Shown in Figure 4.

In addition to the two relationships indicated above, there is another significant dependency betIten "max heart rate" and "ST depression" shown in Figure 5. It is concluded that both the characteristic "age" and the feature "max heart rate" will be crucial in predicting heart disease. Check out their distributions now.

Table 5. Correlation heatmap

```
# create a correlation heatmap
sns.heatmap(df[numeric_columns].corr(),annot=True,cmap='terrain',linewidths=0.1)
fig=plt.gcf()
fig.set_size_inches(8,6)
plt.show()
```

Figure 4. Correlation Matrix

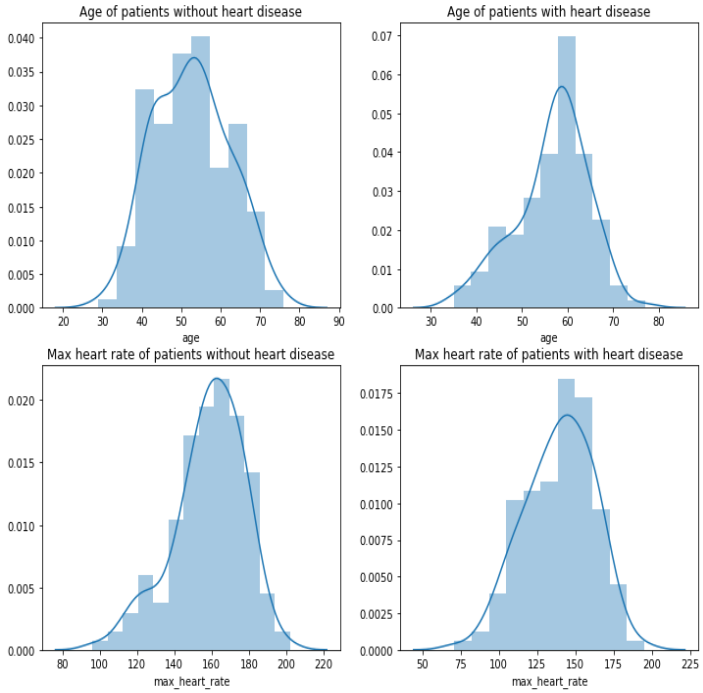

Figure 5. Two Relationships Dependency betIten "Max Heart Rate" and "ST Depression"

Patients will have a significantly larger range of ages than those who are ill. In their sixties, the latter group is most at risk. Although there are not many differences in the distribution of maximum heart rates shown in Figure 6, the risk is greatest betIten 150 and 170. Greater values are more typical in healthy patients. The graphics below will give us an alternative viewpoint.

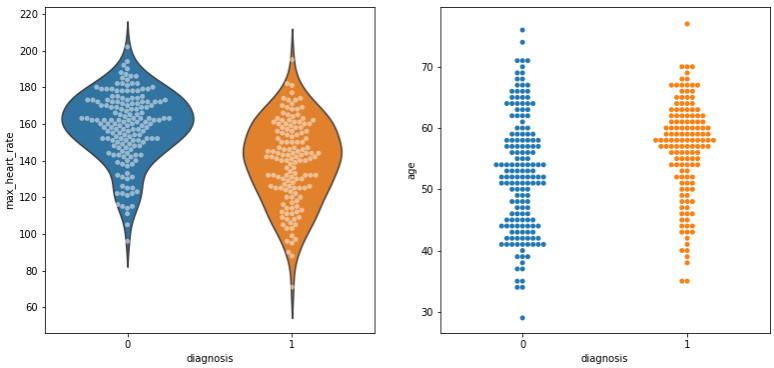

Figure 6. Distribution of Maximum Heart Rates

5.3 Categorical features

Let's examine category variables in more detail and see how they affect our aim shown in Figure 7.

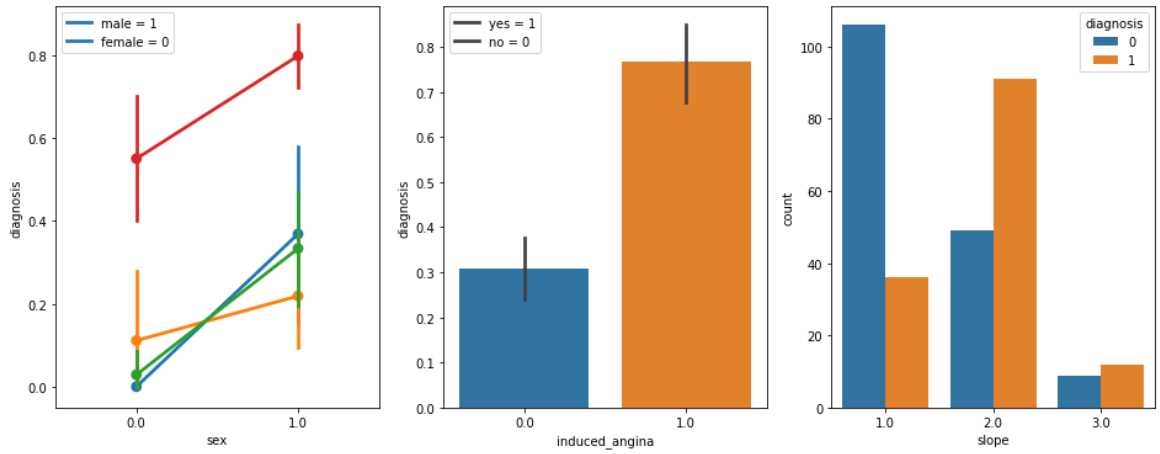

Figure 7. Examine Category

5.4 Observations

Coronary artery disease is far more familiar in male than in female.

The likelihood of illness increases with the number of vessels that fluoroscopy can identify.

Strong discomfort is a significant warning, although gentle chest pain may be a bad sign of impending cardiac trouble (particularly in men)!

For those who have suffered from exercise-induced angina, the risk of developing heart disease may even be three times higher.

Peak exercise's downslope (value=3) and flat slope (value =2) point to a high likelihood of contracting an illness.

Fasting blood sugar appears to be a relatively poor feature for our forecast based on the almost even distribution hence it could be dropped from our model shown in Figure 8. The accuracy of our model won't likely increase by taking this variable out, but it also shouldn't get worse. I decide to leave this variable alone and verify my theory by examining the feature importance of a few models.

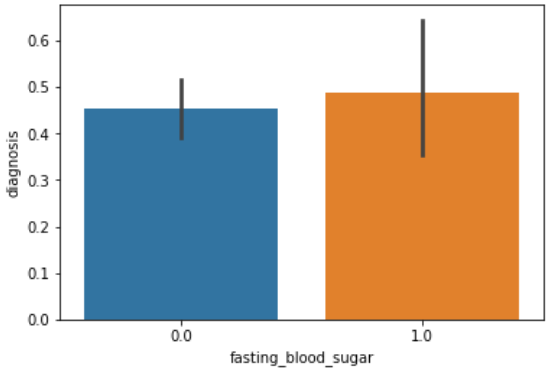

Figure 8. Graph of Fasting Blood Sugar

5.5 Data preparation

First and foremost, it needs to manage all missing data to adapt our dataset to the Sci-kit Learn package's machine learning techniques.

When replacing a missing value is presented in Table 6, it have a lot of possibilities to think about, like:

- a constant value, such as 0, that is distinct from all other values and has significance inside the domain. a value is taken at random from another record.
- a column's mean, median, or mode value
- a figure calculated using a different forecasting model

Table 6. Columns having a missing value

```
# show columns having missing values
df.isnull().sum()
age                     0
sex                     0
chest_pain              0
blood_pressure          0
serum_cholestoral       0
fasting_blood_sugar     0
electrocardiographic    0
max_heart_rate          0
induced_angina          0
ST_depression           0
slope                   0
no_of_vessels           4
thal                    2
diagnosis               0
dtype: int64
```

There are missing values in both categories' columns. In this scenario, "nans" are often filled with the mode, which is the value that occurs the most frequently in the provided vector. Let's implement this remedy.

```
# fill missing values with mode
df['no_of_vessels'].fillna(df['no_of_vessels'].mode()[0],inplace=True)
df['thal'].fillna(df['thal'].mode()[0],inplace=True)
```

A label can be removed from the data frame if the data is clean. Additionally, it would be a good time to divide our data train and test sets. As is customary for a dataset of this size, I will divide the entire dataset in half and assign 30% to the test set.

In Table 7 Before implementing machine learning algorithms, data must be normalised or standardised. Data is scaled by standardisation, which also provides information on how far away from the mean value the data is. In reality, the data's mean () is zero and its standard deviation () is one.

Table 7. Data Splitting

```
# split the data
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3,random_state=2606)
print("train_set_x shape: "+str(X_train.shape))
print("train_set_y shape: "+str(y_train.shape))
print("test_set_x shape: "+str(X_test.shape))
print("test_set_y shape: "+str(y_test.shape))
train_set_x shape: (212, 13)
train_set_y shape: (212,)
test_set_x shape: (91, 13)
test_set_y shape: (91,)
```

6. USING MACHINE LEARNING TO FORECAST AND MODEL

The overall project's primary goal is to provide extremely precise forecasts regarding the propensity to acquire heart disease. To do this, it will put a variety of classification strategies to the test. This section provides a summary of the study's findings and introduces the best accuracy metric performance. To address supervised learning problems, I have chosen a variety of algorithms that are frequently applied in classification approaches.

6.1 K-Nearest Neighbours (KNN)

K-Nearest Neighbours algorithm finding a predetermined number of training samples that are physically close to the new site and predicting the label from some of these forms the basis of nearest-neighbour strategies.

```
# KNN
Train accuracy: 88.21%
Test accuracy: 86.81%
```

Even if it's simple, the result is positive. Let's try out different "n neighbours" inputs to see if KNN can perform any better.

By learning straightforward decision rules derived from the properties of the data, the Decision Trees DT method develops forecasts for the value of a goal variable. It is easy to comprehend and analyse, and it is feasible to ascertain just how crucial a particular quality was to the development of our tree.

```
# Decision Tree
Train accuracy: 100.00%
Test accuracy: 75.82%
```

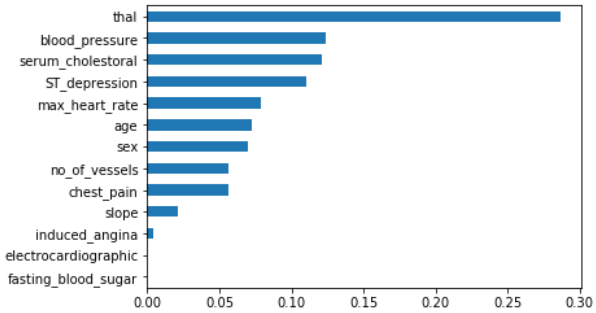

Figure 9. Decision Tree Model

Variable "thal" turns out to be an incredibly important quality.

Remember how I stated the "fasting blood sugar" feature is meagre? This is adequately supported by the graph above.

The decision tree model completely overfits the data while simultaneously learning the train set flawlessly shown in Figure 9, which leads to poor prediction. Different settings for the "max depth" parameter need to be tried.

6.2 Logistic regression

A fundamental statistical analysis technique called logistic regression makes an effort to forecast a data value based on previous observations. A dependent factor, one or more additional dependent variables are compared using the logistic regression approach.

```
# Logistic Regression
Train accuracy: 85.85%
Test accuracy: 85.71%
```

If there is little variance betIten the train and test scores, the model is likely operating at its peak. Even though the real result is slightly less favourable than KNN, it is still respectable.

6.3 Gaussian Naive Bayes

Naive A series of simple probabilistic classifiers called Bayes classifiers are utilised in machine learning. They are founded on the use of the Bayes theorem along with fervent (naive) individuality hypotheses regarding the relationships betIten the attributes.

```
#Gaussian Naive Bayes
Train accuracy: 85.38%
Test accuracy: 86.81%
```

This model's output matched the top KNN method exactly. Although the data are somewhat under-fitted by this model, there are no hyperparameters that might be changed to improve performance.

6.4 Support vector machine

The support vector machine is unquestionably among the best machine learning techniques. They are the best option for a high-performing algorithm with a minor modification. Let's first test it out with the default settings.

```
# Support Vector Machines
Train accuracy: 92.92%
Test accuracy: 82.42%
```

By no means are the aforementioned numbers notable. I shall change the values of the parameters "C" and "kernel" to fully utilise the capabilities of SVM.

6.5 Random forests

Random forests are a technique used in ensemble learning for classification, regression, and other issues shown in Figure 10 which is generated in vast numbers during the training phase.

```
# Random Forests
model=train_model(X_train,y_train,X_test,y_test,RandomForestClassifier,ra
ndom_state=2606)
pd.Series(model.feature_importances_,X.columns).sort_values(ascending=Tru
e).plot.barh()
Train accuracy: 99.06%
Test accuracy: 83.52%
```

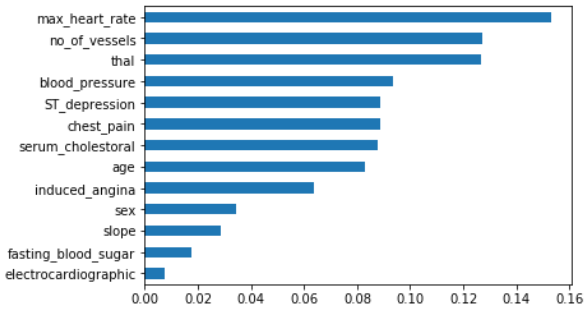

Figure 10. Random Forests

7. COMPARISION BETWEEN DIFFERENT MACHINE LEARNING ALGORITHM

This evaluates multiple machine learning algorithms and predicts a person's likelihood of developing cardiac disease based on their characteristics and symptoms. Analyzing and identifying heart health data will aid in the early detection of atypical cardiac conditions, and the quickest life-saving measures feasible Itre implemented for the prediction of heart disease. For the prediction of heart disease, the experimental results Itre as follows: KNN 86%, Decision Trees 79%, Logistic Regression 85%, Naive Bayes 86%, Support Vector Machines 87%, and random forest 89% shown in Figure 11.

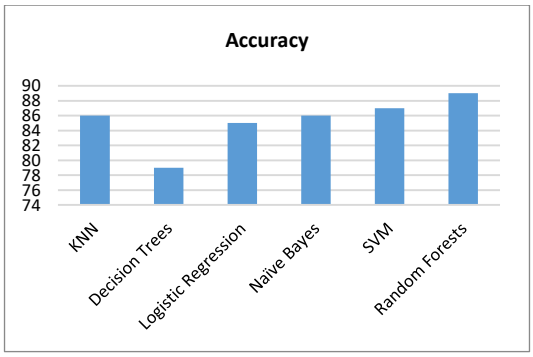

Figure 11. Comparisons between different Machine Learning Algorithm

8. CONCLUSION AND FUTURE WORK

In this research is to discuss numerous data mining approaches that can be used to forecast cardiac disease. The goal with us is to make precise and successful forecasts while using feItr qualities and tests. In this work, various pre-processing approaches and machine learning algorithms are employed to conduct in-depth analysis and obtain findings. To train and test the developed system, three datasets are integrated. If the obtained data was used, the DT classifier outperformed the ML technique for the 14 attributes included in the dataset. The random forest method must be the most efficient algorithm for cardiovascular disease categorization hence it is employed in the proposed approach. Due to the limited sample size, it is difficult to generalize these findings on cardiovascular disease. Because of the limited size of the sample, it is difficult for us to generalize our findings on heart disease. This is our primary limitation. To prepare for future advancements, we intend to apply our method to a more extensive dataset and carry out the study of another disease with a different feature selection methodology.

References:

Almarabeh, H., & Amer, E. (2017). A study of data mining techniques accuracy for healthcare. *International Journal of Computer Applications*, *168*(3), 12-17.

Amin, S. U., Agarwal, K., & Beg, R. (2013, April). Genetic neural network based data mining in prediction of heart disease using risk factors. In *2013 IEEE conference on information & communication technologies* (pp. 1227-1231). IEEE.

Carroll, W., & Miller, G. E. (2013). Disease among Elderly Americans: Estimates for the US civilian non institutionalized population, 2010. *Med. Expend. Panel Surv., no. June*, 1-8.

Chen, A. H., Huang, S. Y., Hong, P. S., Cheng, C. H., & Lin, E. J. (2011, September). HDPS: Heart disease prediction system. In *2011 computing in Cardiology* (pp. 557-560). IEEE.

Dangare, C. S., & Apte, S. S. (2012). Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications*, *47*(10), 44-48.

Dangare, C., & Apte, S. (2012). A data mining approach for prediction of heart disease using neural networks. *International Journal of Computer Engineering and Technology (IJCET)*, *3*(3), 30–40.

Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. *Expert systems with applications*, *36*(4), 7675-7680.

Fatima, M., & Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, *9*(01), 1-16.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. *Big data analytics*, *1*, 1-22.

Ghumbre, S., Patil, C., & Ghatol, A. (2011). Heart disease diagnosis using support vector machine, in *International conference on computer science and information technology.* Pattaya, Thailand: Planetary Scientific Research Centre, pp. 84–88.

Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using artificial neural network and feature subset selection. *Global Journal of Computer Science and Technology Neural & Artificial Intelligence*, *13*(3), 4-8.

Kirubha, V., & Priya, S. M. (2016), Survey on Data Mining Algorithms in Disease Prediction, *International Journal of Computer Trends and Technology (IJCTT)*, *38*(3), 124–128.

Khanwalkar, P. (2024). Patient model based personalized remote health care for chronic disease. *Proceedings on Engineering Sciences*, *6*(3), 897-902. doi: 10.24874/PES06.03.001

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, *23*(1), 89-109. doi: 10.1016/s0933-3657(01)00077-x. PMID: 11470218.

Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Q., & Wang, Q. (2017). A hybrid classification system for heart disease diagnosis based on the RFRS method. *Computational and mathematical methods in medicine*, *2017*(1), 8272091.

Masethe, H., & Masethe, M. (2014). Prediction of heart disease using classification algorithms, in *Proceedings of the world congress on Engineering and Computer Science, San Francisco, USA: International Association of Engineers (IAENG)*, pp. 22–24.

Medhekar, D., Bote, M., Deshmukh, S. (2013). Heart disease prediction system using naive Bayes, *International Journal of Enhanced Research in Science, Technology and Engineering*, *2*(3), 1–5.

Mishra, A., & Bhatt, N. (2022). A review of predicting heart disease using machine learning model. *Ann For Res*, *65*(1), 7516-7520.

Mishra, A., & Lin, JCW. (2023). *Industry 4.0 and Healthcare: Impact of Artificial Intelligence.* Springer Singapore. ISBN: 978-981-99-1948-2.

Mishra, J. (2019). Modified Chua chaotic attractor with differential operators with non-singular kernels. *Chaos, Solitons & Fractals*, *125*, 64-72. https://doi.org/10.1016/j.chaos.2019.05.013.

Mishra, J. (2020). A study on the spread of COVID 19 outbreak by using mathematical modeling. *Results in Physics*, *19*, 103605. https://doi.org/10.1016/j.rinp.2020.103605.

Mishra, A., Suseendran, G., & Nghia, T. (2020). *Soft Computing Applications and Techniques in Healthcare.* Taylor & Francis Group, CRC Press, USA, ISBN: 978-0-367-42387-2.

Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques, in *IEEE/ACS International Conference on Computer Systems and Applications. Doha, Qatar*, pp. 108–115.

Patel, J., Upadhyay, P. and Patel, D. (2016) Heart Disease Prediction Using Machine learning and Data Mining Technique. Journals of Computer Science & Electronics, 7, 129-137.

Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017, July). A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In *2017 IEEE symposium on computers and communications (ISCC)* (pp. 204-207). IEEE.

Rajesh, N., Maneesha, T., Hafeez, S., & Krishna, H., (2018). Prediction of Heart Disease Using Machine Learning Algorithms, *International Journal of Engineering & Technology*, *7*(2), 363-366.

Sabarinathan, V., & Sugumaran, V. (2014). Diagnosis of heart disease using decision tree. *International Journal of Research in Computer Applications & Information Technology*, *2*(6), 74-79.

Sharma, H., & Rizvi, M. A. (2017). Prediction of heart disease using machine learning algorithms: A survey. *International Journal on Recent and Innovation Trends in Computing and Communication*, *5*(8), 99-104.

Sharma, V., Yadav, Y., & Gupta, M. (2020). Heart Disease Prediction using Machine Learning Techniques. 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2020, pp. 177-181.

Shouman, M., Turner, T., & Stocker, R. (2012). Applying k-nearest neighbour in diagnosing heart disease patients. *International Journal of Information and Education Technology*, *2*(3), 220-223.

Singh, M. S., & Choudhary, P. (2017). August. Stroke prediction using artificial intelligence. In 8th *Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)*, pp. 158-161.

Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Intelligent and effective heart disease prediction system using weighted associative classifiers. *International Journal on Computer Science and Engineering*, *3*(6), 2385-2392. ISSN: 0975-3397.

Tasnim, F., & Habiba, S. U. (2021, January). A comparative study on heart disease prediction using data mining techniques and feature selection. In *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)* (pp. 338-341). IEEE. https://doi.org/10.1109/ICREST51555.2021.9331158.

Vembandasamy, K., Sasipriya, R., & Deepa, E. (2015). Heart diseases detection using Naive Bayes algorithm. *International Journal of Innovative Science, Engineering & Technology*, *2*(9), 441-444.

Venkata Lakshmi, B., & Shivsankar, M. (2014). Heart disease diagnosis using predictive data mining, *International Journal of Innovative Research in Science, Engineering and Technology*, *3*(3), 1873–1877.

Waghulde, N. P., & Patil, N. P. (2014). Genetic neural approach for heart disease prediction. *International Journal of Advanced Computer Research*, *4*(3), 778.

Wiharto, W., Kusnanto, H., & Herianto, H. (2015). Performance analysis of multiclass support vector machine classification for diagnosis of coronary heart diseases. *International Journal on Computational Science & Applications*, *5*(5), 27–37.

Zriqat, I. A., Altamimi, A. M., & Azzeh, M. (2017), A comparative study for predicting heart diseases using data mining classification methods, *International Journal of Computer Science and Information Security (IJCSIS)*, *14*(12), 868–879.

Ashish Mishra
Department of CSE Gyan Ganga Institute of Technology and Sciences, Jabalpur (M.P.), India
Research Scholar, Federal Institute of Education, Science and Technology of Ceara, Fortaleza, Brazil
ashishmishra@ggits.org
ORCID 0000-0002-2604-3370

Jyoti Mishra
Department of Mathematics Gyan Ganga Institute of Technology and Sciences, Jabalpur (M.P.), India
Research Scholar, Federal Institute of Education, Science and Technology of Ceara, Fortaleza, Brazil
jyotimishra@ggits.org
ORCID 0000-0003-4938-6461

Victor Hugo
De Albuquerque, Senior Member of IEEE; Department of Teleinformatics Engineering (DETI)
Federal University of Ceará, Brazil
victor120585@yahoo.com.br
ORCID 0000-0003-3886-4309

Aloísio Vieira Lira Neto
Graduation Program in Telecommunication Engineering, Federal Institute of Ceará, Fortaleza, CE, Brazil
aloisio.lira@prf.gov.br