Machine Learning Ensembles for heart disease data set

Alaa aljedani & Doaa altawil & Rawan Alghamdi

King Abduaziz University, Jeddah, KSA

aaljedani0092@stu.kau.edu.sa

daltawil@stu.kau.edu.sa

ralghamdi0644@stu.kau.edu.sa

Abstract — Heart disease is a comprehensive term that refers to a different and varied set of diseases that affect the heart. It covers any heart disorder and indicates any problems in the heart itself. Some people with heart diseases do not experience symptoms unaware of their condition until they are discovered during the examination Physical, so we have many problems, the most important of which is that most Arabs do not have a background about heart disease, so there is a lack of sufficient awareness of the importance of knowing the behavior and integrity of the heart.
Computer-aided detection systems that use machine learning approaches are needed to provide an accurate diagnosis of heart disease. These systems can help detect heart disease at an early stage. When heart disease is detected early enough, the rate of survival increases because better treatment can be provided.
This paper aims to use Support Vector Machines model, K-Nearest Neighbor model, Random Forest model, Logistic Regression model, Ensemble Learning model, Artificial Neural Network model, using the Heart Disease UCI dataset, the focus of this paper is to integrate these machine-learning techniques with all features in a dataset, then compare their performances to identify the most suitable approach.
The proposed approach obtained an accuracy of 86.88%, accuracy 83.78%, cross-validation score 84.28%, F1 score 88.57%, and the average squared error of 37.57%.

Keywords: Heart disease, Machine-learning, Support Vector Machines, K-Nearest Neighbor, Random Forest, Logistic Regression, Ensemble Learning, Artificial neural network.

1. INTRODUCTION

Machine learning is one of the branches of artificial intelligence that is concerned with designing and developing algorithms and technologies that allow computers to possess the "learning" feature. In general, there are two levels of learning: inductive and deductive. The inductive draw general rules and provisions from big data. The primary task of machine learning is to extract valuable information from the data and thus is very close to data mining [1].
We want to discuss for a data set that contains data of people and their health condition and whether they have heart disease or not, the target of dataset modeling is to help people with a high risk of heart disease to get an early warning so they can get medical needs as soon as possible.

In this paper, we discuss and evaluate machine-learning models to heart disease data set, the rest of the paper is organized, follow, First, Introduction: we provide an overview of heart disease, data set and current approaches. Next Related Work: we briefly describe the similar work and make a comparison of our results with their results. Next, Background of machine learning & feature selection. Next Experimental work: we perform a complete analysis and visualization of the dataset using Histogram, Heatmaps, and Correlation Matrix. Next, Results and discussion. Finally, in conclusion, we conclude by discussing the findings, contributions, and direction of future research.

1.1 Overview of heart disease:

The symptoms of heart disease depend on which condition is affecting an individual, some of the common symptoms include:

- **Shortness of breath**

- **Pain**
  In legs or arms if the blood vessels in those parts of body are narrowed or in the neck, jaw, throat, upper abdomen or back.

- **Chest pain**
  The chest pain common too many types of heart disease are known as angina, Angina can be triggered by stressful events or physical exertion and normally lasts under 10 minutes.

- **Heart attacks**
  Heart attacks can also occur as a result of different types of heart disease, the signs of a heart attack are similar to angina except that they can occur during rest and tend to be more severe, other symptoms of a heart attack include: lightheadedness and dizzy sensations, profuse sweating nausea and vomiting.

There are many types of heart disease that affect different parts of the organ and can occur in different ways:

- **Congenital heart disease**
  This is a general term for some deformities of the heart that have been present since birth.

- **Arrhythmia**
  Is an irregular heartbeat, it is common, and all people experience them, so, they need to be taken more seriously and treated?

- **Coronary artery disease**
  Coronary arteries can become diseased or damaged, usually because of plaque deposits that contain cholesterol, and this causes the heart to receive less oxygen and nutrients.

- **Dilated cardiomyopathy**
  The heart chambers become dilated because of heart muscle weakness than cannot pump blood properly.

- **Heart failure**
  This occurs when the heart does not pump blood around the body efficiently [2].

1.2 Overview of the dataset:

The data used in this project was created and gathered from different places which are:
1- Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2- University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3- University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4- Medical Center, Long Beach, and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.
Heart Disease UCI database contains 76 attributes, but all experiments using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. [3]
The "target" field refers to the presence of heart disease in the patient (See, Figure 1, the data set value 0= no, 1=yes).

Figure 1 target attribute

Other attributes are:
Age, for patient (See, Figure 2, dataset range between 29 - 77) Age is the most important risk factor in developing cardiovascular or heart diseases.

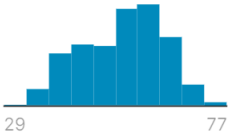

Figure 2 Age attribute

Sex, for patient (See, Figure 3, dataset value 1 = male, 0 = female) Male is greater than female

Figure 3 Sex attribute

Chest pain type (cp), Pain type and symptoms represent by 4 values (See, Figure 4, dataset value 0 = typical angina, 1 = atypical angina, 2 = non-anginal, 3 = asymptomatic).

Figure 4 cp attribute

Trestbpsresting blood pressure (trestbps), on admission to the hospital (See, Figure 5, data set a range between 94-200).

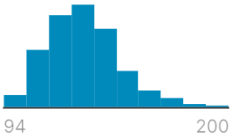

Figure 5 trestbps attribute

Cho serum Cholesterol (chol), cholesterol is often associated with heart disease. That is because low-density lipoproteins (LDL) can build up in your arteries and restrict or block blood flow. Your body still needs a little cholesterol for healthy digestion and to make vitamin D and certain hormones.

Measuring your LDL "bad" cholesterol, HDL "good" cholesterol, and triglycerides will give you a number called your total blood cholesterol, or serum cholesterol. Your serum cholesterol levels can help your doctor figure out your risk for developing heart disease in the next 10 (See, Figure 6, data set a range between 126 - 564) [4].

Figure 6 chol attribute

Fasting blood sugar (FBS), the body needs glucose to get energy, and glucose comes from the food we eat. However, the body does not use all this energy once. After eating a meal, blood sugar levels rise, usually peaking about an hour after eating.

With high blood sugar, the pancreas releases insulin. Insulin lowers and breaks blood sugar so the body can use it for energy or store it for later. Diabetics suffer from high blood sugar levels and difficulty using glucose or blood sugar.

High levels of fasting blood sugar (>120) indicate that the body has been unable to lower blood sugar levels (See, Figure 7, dataset value 0 = > 120 mg/dl, 1 = < 120 mg/dl) [5].

Figure 7 FBS attribute

Resting electrocardiographic results (restecg), Electrocardiography is the process of producing an electrocardiogram It is a graph of voltage versus time of the electrical activity of the heart. (See, Figure 8, data set 0: normal, 1: ST-T wave abnormality, 2: ventricular hypertrophy) [6]

Figure 8 restecg attribute

Maximum heart rate achieved (thalach), The normal resting heart rate for adults ranges from 60 to 100 beats per minute, generally, a lower resting heart rate means a more efficient heart function and better cardiovascular fitness. In data set recorded the highest heart rate. (See, Figure 9, data set a range between 71- 220) [7].

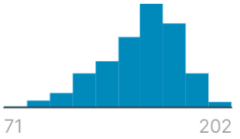

Figure 9 thalach attribute

Exercise-induced angina (exang), it's a symptom of heart disease, is chest pain that happens because there isn't enough blood going to part of your heart.(See figure 10, data set value, 1 = yes , 0 = no)[8].

Figure 10 exang attribute

ST depression induced by exercise relative to rest (oldpeak), ST depression refers to a finding on an electrocardiogram, wherein the trace in the ST segment induced by exercise relative to rest, (See, Figure 11, data set a range between 0 - 6.2) [9].

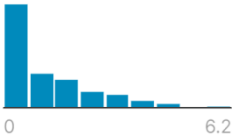

Figure 11 oldpeak attribute

(Slope), the slope of the peak exercise ST segment (See, Figure 12, data set value 0 = upsloping, 1 = flat, 2 = downsloping).

Figure 12 Slope attribute

(Ca), the number of major vessels (0-3) colored by fluoroscopy (See, Figure 13, data set a range between 0 -4).

Figure 13 Ca attribute

(thal), patients with fixed defect have a significantly higher incidence of heart disease. (See, Figure 14, dataset value 0 = normal, 1 = normal, 2 = fixed defect, 3= reversable defect)

Figure 14 thal attribute

1.3 Current approaches:

The tests to diagnose heart disease depend on the type of heart disease; the doctor will likely perform a physical exam and ask about personal and family medical history before doing any tests. Besides blood tests and a chest X-ray, tests to diagnose heart disease can include:

- **Electrocardiogram (ECG)**
  An ECG records these electrical signals and can help your doctor detect irregularities in your heart's rhythm and structure.

- **Holter monitoring**
  A Holter monitor is a portable device you wear to record a continuous ECG, usually for 24 to 72 hours, Holter monitoring is used to detect heart rhythm irregularities.

- **Echocardiogram**
  This noninvasive exam, which includes an ultrasound of your chest, shows detailed images of your heart's structure and function.

- **Stress test**
  This type of test involves raising your heart rate with exercise or medicine while performing heart tests and imaging to check how your heart responds.

- **Cardiac computerized tomography (CT)**
  This test is often used to check for heart problems.

- **Cardiac magnetic resonance imaging (MRI)**
  For this test, the magnetic field produces pictures to help your doctor evaluate your heart [10].

2. RELATED WORK

In this section, some of the related works previously done on breast heart disease by researchers using different machine learning approaches are discussed.

Vitalii Mokin [11], he uses a different ML algorithm to show the best model using the Heart Disease UCI database, he split data on 20% in the test dataset, the remaining 80% - in the training dataset and then tested a lot of the model, we will focus on 4 models and show the accuracy result of them.
The evaluation result produced In Logistic Regression he concluded is Positive coefficients increase the log-odds of the response (and thus increase the probability), and negative coefficients decrease the log-odds of the response (and thus decrease the probability), cp is the highest negative coefficient, trestbps, thal, oldpeak are the largest numbers by absolute value, in Voting Classifier he combines between (Logistic Regression model, Random Forest model, AdaBoost model and Ensemble model).
Experimental results revealed that Logistic Regression accuracy is 84.71% , ANN = 66.12% , KNN = 78.1% , Voting Classifier (with hard voting) = 90.5%, Voting Classifier (with soft voting) = 94.63% , the best is voting classifier (with soft voting) model .

Rajesh Kumar Jha [12], He built 2 models (Artificial neural network model, Random Forest model) then compare the accuracy between them, he uses the Heart Disease UCI database, He built the ANN model, uses two layers, a hidden layer, and an output layer, in the hidden layer using the sigmoid function for activations. The output layer has only one node and is used for the regression, the output of the node is the same as the input of the node, then he Fitting the model and choose the number of epochs = 100 (epochs, number of times the dataset will pass through the network), Finally, the accuracy of ANN model = 86.88%, Random Forest model, he found the accuracy of this model = 80.32%, the best is ANN model.

Bruno Aldo Lunardi [13], he split the data using the stratified method, to keep the data more approximately distributed as the original, then create a simple pipeline, for imputing future values. Then he creates function (classification model) use it into the different model to see the results of the model (Precision, recall, F1 score, and Cross-Validation Score).
Experimental results revealed that Logistic Regression Cross-Validation Score is 83.866%, Random Forest Classifier = 82.650% , XGB Classifier = 82.668% , Voting Classifier (He use the higher Cross-Validation Score model: Logistic Regression XGBClassifier) = 83.892% , the best is Voting Classifier model.

3. Background of machine learning & feature selection

Machine learning helps many problems in human life to make life easier, so this paper is designed to conduct a review of some of the widely used classification algorithms and their application in heart disease, and we selected most of the features in the dataset in our different modeling techniques.

In this paper, we use 6 different models from open source to solve many problems in our dataset.

3.1 Support Vector Machines

The support vector machines algorithm (SVM) The objective that finds if we have space in a figure of any N-dimensional (N= number of feature) that classifies data point for two categories and our goal is to find the distance between the gap sense of who has the furthest gap between the data point on both distinguish layer that makes a hyperplane, and between the example of different categories that increase in the size or decrease in the gap of the data [14]

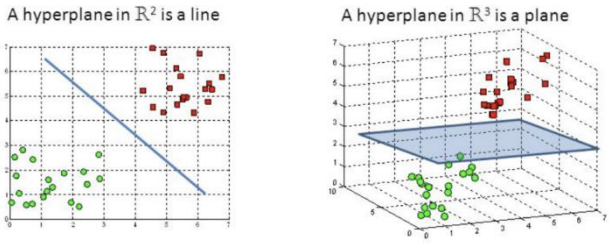

Figure 15 Hyperplanes in 2D and 3D feature space [14]

In large margin Intuition of SVM, we take the output of the linear support vector machine (linear kernel) and if that output is greater than 1, we identify it with one class and if the output is -1, we identify is with another class. That is mean our y ∈ [1, -1] and the X ∈ $\boldsymbol{R^2}$ which acts as margin [14].
The only information is available where the number of data training pairs and is therefore equal to the size of the training data set (1)[15]:

$$D = \{(\mathbf{Xi}.\mathbf{Yi}) \in \boldsymbol{X} * \boldsymbol{Y}\},\ i=1 \qquad (1)$$

Finds the parameters of $w = [w_1\ w_2\ ...\ w_n]^T$ and the B of a discriminant or decision function d (x,w,b) given as (2) [15]:

$$d(x.w.b) = w^T x + b = \sum_{i=1}^{n} w_i x_i + b \qquad (2)$$

Maximization margin (M) gives the weight vector and the bias value. This is defined by (3):

$$M=\frac{2}{||w||} \qquad (3)$$

Minimization of margin given as (4) [15]:

$$\frac{1}{2} w^T w \qquad (4)$$

To find optimal hyperplane having maximal margin that should minimize $||w^2||$ constrained optimization problem which is also known as the primal problem can be formulated (5)[15]:

$$L=(w,b,\alpha)=\frac{1}{2}\, w^T w \;-\; \sum_{i=1}^{l} \alpha_i \{y_i[w^T x + b] - 1\} \qquad (5)$$

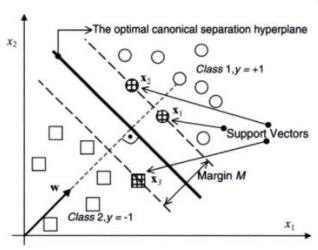

Figure 16 the optimal canonical separation hyperplane

3.2 K-Nearest Neighbor

K-Nearest Neighbor (KNN) is one of the simplest Machine Learning algorithms based on Supervised Learning technique that take the similarity between the new data and available data and put the new data into the category that is most similar to the available categories stores all the available data and classifies a new data point based on the similarity [16].
K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems and also non-parametric algorithm, means it doesn't make any assumption on underlying data, another named called a lazy learner algorithm [16].

K-NN work by some step can be explained based on the below algorithm:
Step-1: Select the number K of the neighbors
Step-2: Calculate the Euclidean distance of K number of neighbors.
Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.
Step-4: Among these k neighbors, count the number of the data points in each category.
Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.
Step-6: Our model is ready [16].

Table 1 [16]

Advantage	Disadvantage
It is simple to implement. It is robust to the noisy training data It can be more effective if the training data is large.	Always needs to determine the value of K which may be complex some time. The computation cost is high because of calculating the distance between the data points for all the training samples

The k-nearest neighbor algorithm is a classification algorithm that assigns to a given data point the majority class among the k-nearest neighbors. The distance between two points is measured by a metric. The dimensionality and position of a data point in the space are determined by its qualities. Many dimensions can result in the accuracy of the k-NN algorithm. Reducing the dimensions of qualities of smaller can increase

accuracy, similarly, to increase accuracy further, distances for each dimension should be scaled according to the importance of the quality of that dimension [16].
We should instead use a different metric to measure the distance Instead of the angle, one could take the cosine of the angle cos(0), and then use the well-known dot product formula to calculate the cos(0)[16].
Let a-($a_x . a_y$), b-($b_x . b_y$), then instead this formula (6) [16]:

$$|a||b| \cos(\theta) = a * B = a_x * b_x + a_y * b_y \quad (6)$$

On derives (12) [7]:

$$\cos(\theta) = \frac{a_x * b_x + a_y * b_y}{|a||b|} \quad (7)$$

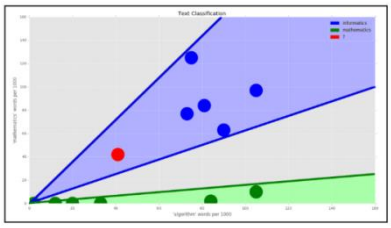

Figure 17 [16]

3.3 Random Forest

Random forest, like her name, consists of many individual decision trees that working as a group. Each tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction (See, Figure 18) [17]

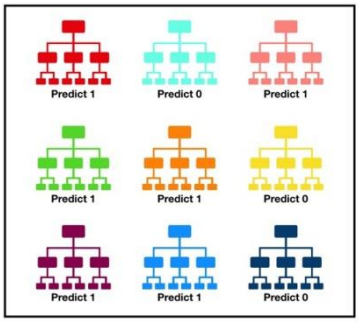

Figure 18

The basic concept behind random forest it is simple but powerful, the reason that the random forest model works so well is:
Many relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models [5].
Uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. The reason for this effect is that the trees protect each other (if they do not constantly all err in the same direction). but they are some trees that may be in the wrong direction, also many other trees will be the right direction, so as a group the trees can move in the correct direction. So, the prerequisites for a random forest to perform well are [17]:

- There needs to be some actual signal in our features to be better than random guessing.
- The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other .

In random forest If our tree did not have any behavior to follow, so we are following two methods [17]:

- **Bagging**
  Decisions trees are very sensitive to the data they are trained on any small changes to the training set that can lead to a hugely different and new tree structure. this by taking each individual tree to randomly sample from the dataset with replacement, resulting in different trees. and it designed to improve stability and reduce variance and accuracy.

- **Feature Randomness**
  In a normal decision tree, when we want to split a node, we take every possible feature and pick the one that produces the most separation between the observations in the left node vs the right node. In contrast, each tree can pick only from a random subset of features. This forces even more variation amongst the trees in the model and ultimately results in lower correlation across trees and more diversification.

The number of planes planted trees with n non-root and non-labeled is equal to (8):

$$\frac{1}{n+1}\binom{2n}{n} \tag{8}$$

Finally, the observed data are the original unlabeled data and the synthetic data are drawn from a reference distribution.

3.4 Logistic Regression

Logistic regression is the analysis regression to conduct when the dependent variable is dichotomous (binary) and is a predictive analysis we can use to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
There should be no outliers in the data, which can be assessed by converting the continuous predictors to standardized scores, and removing values below -3.29 or greater than 3.29, This can be assessed by a correlation matrix among the predictors. Mathematically, logistic regression estimates a multiple linear regression function defined as (9)[19]:

$$\text{Log (p)} = \log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 * x_i + \beta_2 * x_i \ldots \beta_n * x_m \tag{9}$$

- **Overfitting**
  The model for the logistic regression analysis. Adding independent variables to a logistic regression model will always increase the amount of variance (typically expressed as R^2). adding more and more variables to the model can result in overfitting, which reduces data on the model is fit [19]**.**

- **Reporting the R2:**
  R2 values have been developed for binary logistic regression. These should be interpreted with extreme caution as they have many computational issues that cause them to be artificially high or low. A better approach is to present any of the goodness of fit tests available; use Hosmer-Lemeshow is a commonly used measure of goodness of fit based on the Chi-square test [19].

On the other hand, a logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to linear regression, but the curve is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group [20]

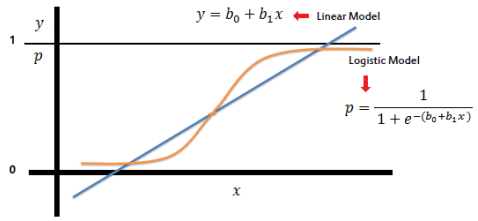

Figure 19 [20]

3.5 Ensemble Learning (Voting)

Ensemble models in machine learning combine the decisions from multiple models to improve the overall performance, the main reason to have causes of error in learning models are due to noise, bias, and variance [21].

Ensemble models have many techniques such as:

- **Taking the mode of the results**
  The mode is a statistical term that refers to the most frequently occurring number found in a set of numbers, in this technique, multiple models are used to make predictions for each data point. The predictions by each model are considered as a separate vote. The prediction which we get from most of the models is used as the final prediction.

- **Taking the average of the results**
  We take an average of predictions from all the models and use it to make the final prediction:

AVERAGE= sum of numbers /Total of numbers (10)

Taking the weighted average of the results
This is an extension of the averaging method. All models are assigned different weights defining the importance of each model for prediction called weighted average [21].

Types of ensemble model:

- **Bootstrap Aggregating**
  It is an ensemble method. First, we create random samples of the training data set with replacement (subsets of training data set). Then, we build a model (classifier or Decision tree) for each sample. Finally, the results of these multiple models are combined using average or majority voting [21].

- **Boosting**
  is a sequential technique in which, the first algorithm is trained on the entire data set and the subsequent based on the last classification algorithms are built by fitting the residuals of the first algorithm, thus giving higher weight to those observations that were poorly predicted by the previous model [21].

	Bagging	**Boosting**
Similarities	• Uses voting • Combines models of the same type	
Differences	Individual models are built separately	Each new model is influenced by the performance of those built previously
	Equal weight is given to all models	Weights a model's contribution by its performance

Figure 20 [21]

Table 2 [21]

Advantage	Disadvantage
• More accurate prediction results • Stable and more robust model • Ensemble models can be used to capture the linear as well as the non-linear relationships in the data	• Reduction in model interpret-ability • Computation and design time is high • he selection of models for creating an ensemble is an art that is hard to master.

3.6 Artificial neural network

Neural networks (also called “perceptron's or ANN) one of the main tools used in machine learning. like there name “neural” part of their, they are brain-inspired systems. Neural networks consist of input and output layers, as well as (in most cases) a hidden layer consisting of units that transform the input into something that the output layer can use. They are excellent tools for finding patterns that are far too complex or numerous for a human programmer to extract and teach the machine to recognize it is only in the last several decades where they have become a major part of artificial intelligence. This is due to the arrival of a technique called “backpropagation" [22].

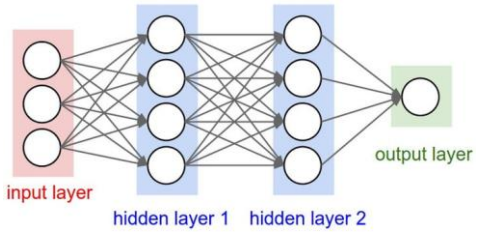

Figure 21 [22]

To get the idea of how a deep learning neural network learns, imagine a factory line and the way how to be. After the raw materials (the data set) are input, they are then passed down the conveyer belt, with each subsequent stop or layer extracting a different set of high-level features. If the network is intended to recognize an object, the first layer might analyze the brightness of its pixels [22].
Non-Linear functions are those which have a degree more than one and they have a curvature
Types of Activation Functions:

- **Threshold Activation Function — (Binary step function)**
  A Binary step function is a threshold-based activation function. If the input value is above or below a certain threshold, the neuron is activated and sends the same signal to the next layer [23].

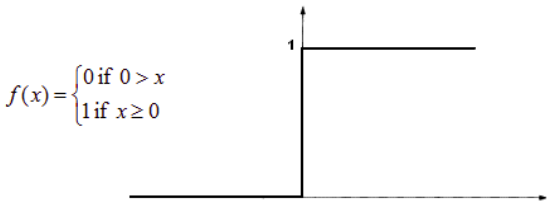

Figure 22 [23]

Activation function A = “activated” if Y > threshold else not or A=1 if y >threshold 0 otherwise.
The problem with this function is for creating a binary classifier (1 or 0), but if you want multiple such neurons to be connected to bring in more classes, Class1, Class2, Class3, etc. In this case, all neurons will give 1, so we cannot decide [23].

- **Sigmoid Activation Function — (Logistic function)**
  A Sigmoid function is a mathematical function having a characteristic "S"-shaped curve or sigmoid curve which ranges between 0 and 1, therefore it is used for models where we need to predict the probability as an output[23].

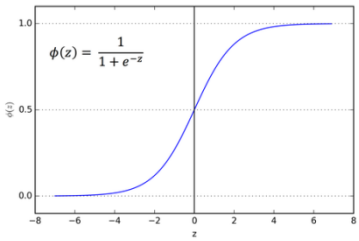

Figure 23 [23]

The Sigmoid function is differentiable, means we can find the slope of the curve at any 2 points. The drawback of the Sigmoid activation function is that it can cause the neural network to get stuck at training time if strong negative input is provided.

- **Rectified Linear Units — (ReLu)**

ReLu is the most used activation function in CNN and ANN which ranges from zero to infinity. [0, ∞) [23]

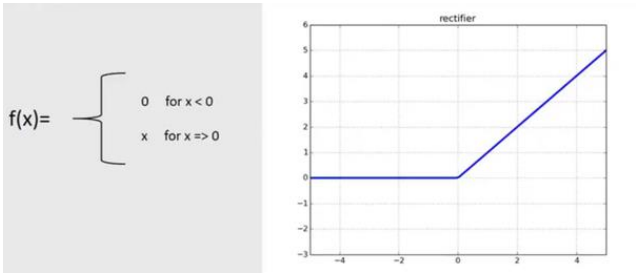

Figure 24 ReLu [23]

It gives an output 'x' if x is positive and 0 otherwise. It is like having the identical problem of linear function as it is linear in the positive axis. Relu is non-linear and a combination of ReLu is also non-linear. it is a good approximate and any function can be approximated with a combination of Relu, ReLu is 6 times improved over hyperbolic tangent function
It should only be applied to hidden layers of a neural network. So, for the output layer use softmax function for the classification problem, and regression problem use a Linear function.
Here one problem is some gradients are fragile during training and can die. It causes a weight update which will make it never activate on any data point again. ReLu could result in dead neurons. To fix the problem of dying neurons, Leaky ReLu was introduced. So, Leaky ReLu introduces a small slope to keep the updates alive. Leaky ReLu ranges from -∞ to +∞ [23].

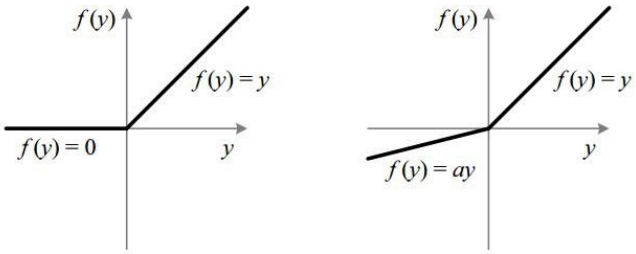

Figure 25 ReLu vs Leaky ReLu

Leak helps to increase the range of the ReLu function, usually, the value of a = 0.01 or so. When a is not 0.01, then it is called Randomized ReLu [23].

4. EXPERIMENTAL WORK

We used classification techniques because the data set labeled into 2 classes, it's a two-step process such as:

- **Learning Step (Training Phase)**
  Different Algorithms are accustomed to building a classifier by making the model learn using the training set available. The model must be trained for the prediction of accurate results.

- **Classification Step**
  Models are used to predict the class for the instances. The experiment was conducted by the python language for developing and testing the various ML models.

4.1 Data collection and Understanding

We import Python libraries and load the dataset, describe the dataset for numerical and categorical features, then we plotted data using Histogram to understanding Categorical data, we have found that the incidence of heart disease in males is higher than females, that those with a blood sugar level greater than 120 are the most affected by heart disease, that patients without angina are the most affected with heart disease, patients who have had wave abnormality with beat heart is most affected by heart disease, patients who have had the slope of the peak exercise ST-segment to down is most affected by heart disease, most patients of those with heart disease do not suffer from angina pectoris at the time of exercise and patients with the fixed defect have a significantly higher incidence of heart disease. (See, Figure 26), then we show the correlation between attributes by Heatmaps, we found the chest pain type (cp) is the highest correlation to target, Heatmaps is a great way of representing correlation visually (See, Figure 27).

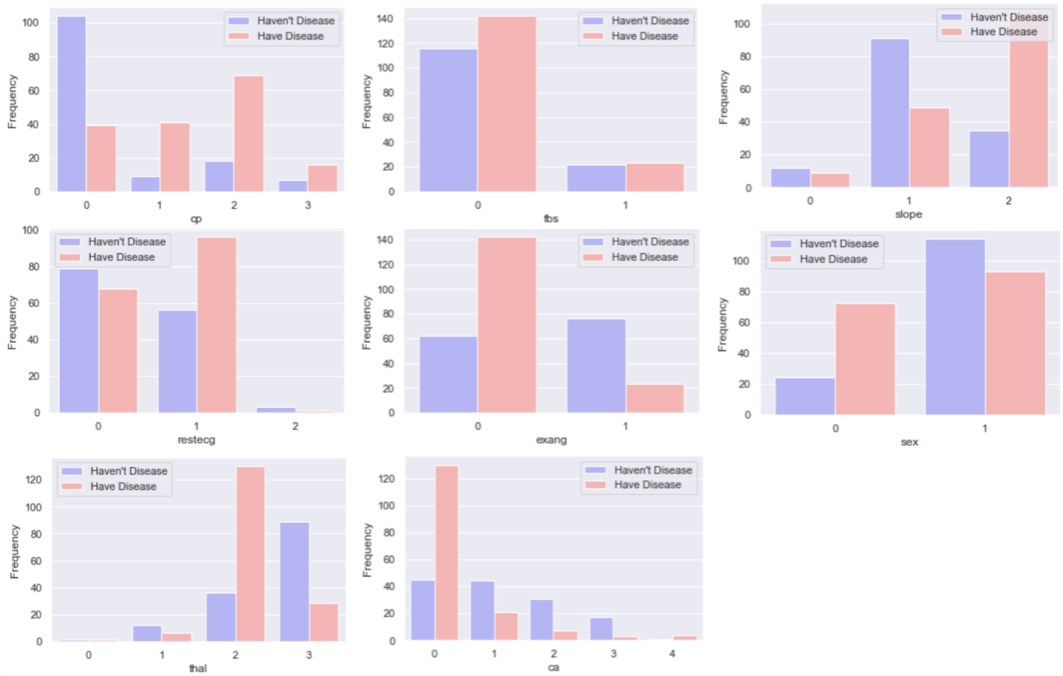

Figure 26 Histogram

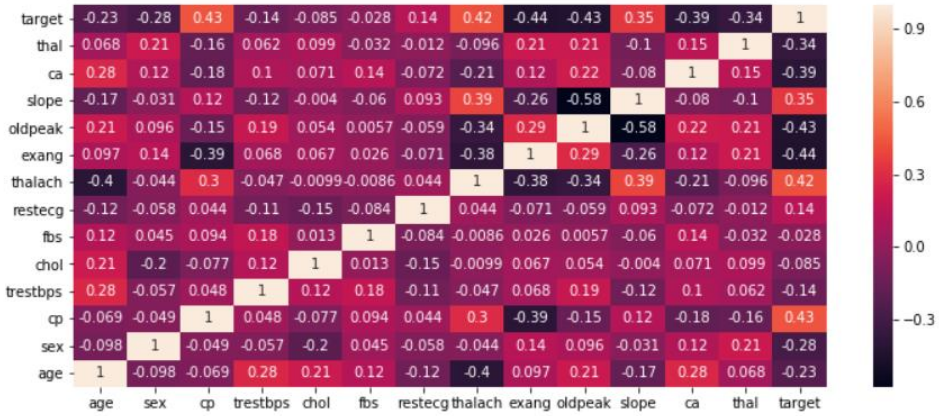

Figure 27 Heatmaps

4.2 Data selection and preprocessing

In any Machine Learning process, Data Preprocessing is that step during which the data gets transformed, or Encoded, to bring it to such a state that now the machine can easily parse it and to show the result with high accuracy. In other words, the features of the data set can be easily interpreted by the algorithm when it is preprocessed.
Because data is usually taken from multiple sources which are normally not too reliable which too in numerous formats, quite half our time is consumed in addressing data quality issues when acting on a machine learning problem. It is simply unrealistic to expect that the data is perfect. There are also problems because of human error, limitations of measuring devices, or flaws within the data collection process.

Data preprocessing techniques:

- **Data Cleaning**
  The data set may have many irrelevant and missing parts. To handle this part, data cleaning is performing. It involves handling of missing data, Outliers, etc. Our data set is already cleaned so we skipped this step.

- **Scaling**
  feature scaling is performed when the numeric attributes have a different scale. in the heart disease data, the oldpeak ranges 0-6.2 while the chol ranges 126-564. It is not much difference between them, but the models perform better after feature scaling. The goal of scaling is to change the values of every numerical feature within the data set to a standard scale. Scaling can be done with min-max scaling or standardization.

- **Dimensionality Reduction**
  In classification problems, different factors influence the final classification. One of those factors called features. The more these features of the data set, the more difficult it will be to train these data on the model. Here comes the importance of dimensionality reduction algorithms.
  Dimensionality reduction is a method to decrease the number of random variables and obtaining the principal variables. There are two techniques for Dimensionality reduction. Feature selection and feature extraction, we skipped this step because we have 14 features only!

4.3 Apply machine learning techniques

The next step after we have prepared the data is training it with different classification algorithms. ML classification algorithms were applied in this work are SVM, Knn, Random forest, Logistic regression, Ensemble learning (voting), ANN (Artificial neural network), all These algorithms have been explained in section II. All features have been to train the models to classify data into (0 = no heart disease, 1 = yes have heart disease). This paper does not apply any of the data preprocessing except for feature scaling. Experimental results showed that the Ensemble learning (voting) results was very good compare to the other algorithms.

4.4 Performance evaluation

Evaluating your machine learning algorithm is a vital part of any project. In this work we used to evaluate the model:
Cross-Validation Score splits the dataset into K equal groups. Each group is referred to as a fold. Some of the folds are used for training and the remaining for testing the model.

Accuracy Score Function, computes subset accuracy in a multi label classification dataset and is equal to the Jaccard Score function in binary and multiclass classification.
F1 Score is the weighted average of Precision and Recall.
Mean Square Error is the average of the square of the difference between the observed and predicted values of a variable.
Confusion matrix shows the corrected and wrong predictions, in comparison with the actual labels. It shows the model's ability to correctly predict or separate the classes
True Positive – model predicted positive class correctly to be a positive class
False Positive – model predicted negative class incorrectly to be a positive class
False Negative – model predicted positive class incorrectly to be the negative class
True Negative – model predicted negative class correctly to be the negative class

Our result of Confusion matrix:
Confusion matrix about Logistic Regression Classifier: array [[19, 9], [2, 31]] (See, Figure 28).
Confusion matrix about Support vector machines: array [[20, 8], [2, 31]] (See, Figure 29).
Confusion matrix k-nearest neighbors Classifier: array [[19, 9], [3, 30]] (See, Figure 30).
Confusion matrix about Random Forest Classifier: array [[20, 8], [3, 30]] (See, Figure 31).
Confusion matrix about Voting Classifier: array [[22, 6], [2, 31]] (See, Figure 32).
Confusion matrix about Artificial neural network: array [[19, 9], [2, 31]] (See, Figure 33).

Table 3 Summary of machine learning modules result

model	Accuracy	Precision	Recall	F1 score	Cross-Validation Score	mean error
Random Forest Classifier	0.81967213	0.7894	0.9090	0.8450	79.750%	0.4155
Logistic Regression Classifier	0.81967213	0.775	0.9393	0.8493	83.517%	0.3964
Support vector machines	0.83606557	0.7948	0.9393	0.8611	81.817%	0.4163
k-nearest neighbors	0.80327868	0.7692	0.9090	0.8333	82.650%	0.4027
Artificial neural network	0.81967213	0.775	0.9393	0.8493	----	----
Voting Classifier (we use Logistic Regression , SVM, KNN and Random Forest model)	0.86885245	0.8378	0.9393	0.8857	84.283%	0.3757

Figure 28 Logistic Confusion matrix

Figure 29 SVM Confusion matrix

Figure 30 KNN Confusion matrix

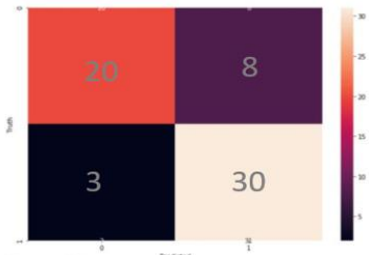

Figure 31 Random forest Confusion matrix

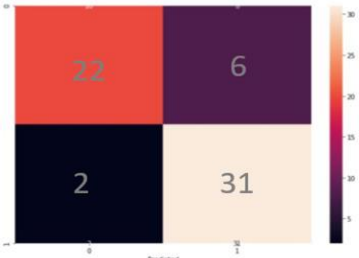

Figure 32 Voting Confusion matrix

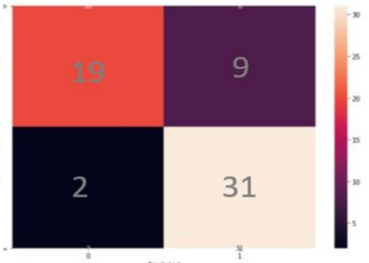

Figure 33 ANN Confusion matrix

5. RESULTS

After training the model and evaluating it with different techniques, a comparison has been made between them, as shown in the table in the Performance evaluation section, and as Figure in drawings 34-37. After observing the table, for accuracy, the voting model is better than the rest, with a difference of 0.03 from the best of them, and its accuracy rate is 0.868, which is very good. In fact, voting model obtained the best performance in terms of Precision (0.83), Cross-Validation Score (0.84), F1 score (0.88), Mean squared error (0.37), not just the accuracy! As for the recall, it was equal to some other models with percentage of 0.939. This result is like some of the related work such as Bruno Aldo Lunardi and Vitalii Mokin they Both concluded after the experiment that the voting model is the best

The results of some models improved well before and after the feature scaling for example, the accuracy for KNN was 0.64 and now it is 0.80 and the recall was 0.67 and now it's 0.90! So, the feature scaling has a good effect to the performance unlike the PCA, which negatively affected and reduces the accuracy of most models.

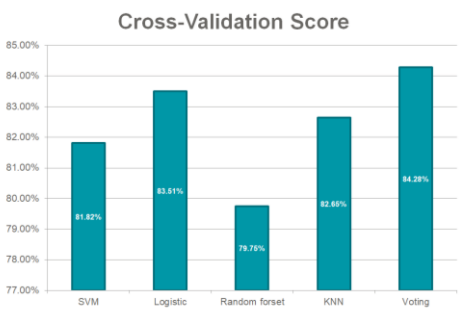

Figure 28

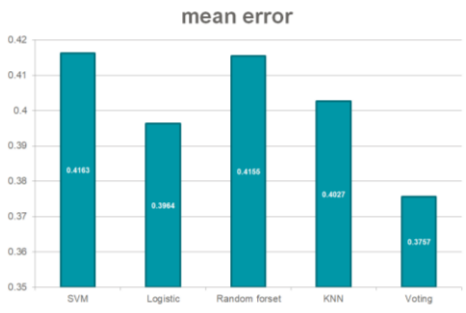

Figure 29

Figure 30

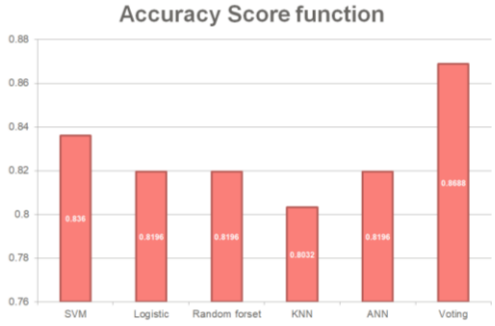

Figure 31

6. Conclusion

This paper analyzed the heart disease dataset using feature scaling technique and six popular ML algorithms to classify if they have heart disease or not.
This research work reveals that the feature scaling can help improve the diagnosis of know early heart disease using machine learning techniques.
Future work can be directed towards developing the chosen approach into a potential practical method for aiding doctors with a quick second opinion in heart disease. Future work can also consider comparing more ML algorithms used for heart disease. More disease options can also be considered in future works.

7. REFERENCES

[1] Wikipedia, About Machine learning. [online] Available at: https://en.wikipedia.org/wiki/Machine_learning [Accessed 14 Apr 2020]

[2] Medical News Today, About Heart disease: Types, causes, and treatments. [online] Available at: https://www.medicalnewstoday.com/articles/237191#symptoms [Accessed 14 Apr 2020]

[3] Kaggel, About Heart Disease UCI. [online] Available at: https://www.kaggle.com/ronitf/heart-disease-uci [Accessed 14 Apr 2020]

[4] Healthline, About Serum Cholesterol: Understanding Your Levels. [online] Available at: https://www.healthline.com/health/serum-cholesterol [Accessed 14 Apr 2020]

[5] Medical News Today, About What to know about fasting blood sugar?. [online] Available at: https://www.medicalnewstoday.com/articles/317466#fasting-blood-sugar-levels [Accessed 15 Apr 2020]

[6] Wikipedia, About Electrocardiography. [online] Available at: https://en.wikipedia.org/wiki/Electrocardiography [Accessed 15 Apr 2020]

[7] Mayo Clinic, About 2 easy, accurate ways to measure your heart rate. [online] Available at: https://www.mayoclinic.org/healthy-lifestyle/fitness/expert-answers/heart-rate/faq-20057979 [Accessed 15 Apr 2020]

[8] WebMD, About Angina (Ischemic Chest Pain). [online] Available at: https://www.webmd.com/heart-disease/heart-disease-angina#1 [Accessed 16 Apr 2020]

[9] Wikipedia, About ST depression. [online] Available at: https://en.wikipedia.org/wiki/ST_depression [Accessed 16 Apr 2020]

[10] Mayo Clinic, About 2 easy, accurate ways to measure your heart rate. [online] Available at: https://www.mayoclinic.org/diseases-conditions/heart-disease/diagnosis-treatment/drc-20353124 [Accessed 16 Apr 2020]

[11] Vitalii Mokin, Professor at Vinnytsia National Technical University, Vinnytsia, Vinnytsia Oblast, Ukraine,Joined 2 years ago to kaggel, http://mmss.vntu.edu.ua/index.php/en/staff-en

[12] Rajesh kumar jha, student at BNMIT,Bengaluru, Karnataka, India, Joined 2 years ago to kaggel, https://github.com/rajeshjnv

[13] Bruno Aldo Lunardi, Data Scientist Student,Santo André, State of São Paulo, Brazil, Joined a year ago to kaggel . https://www.kaggle.com/akumaldo

[14] Toward data science, About support vector machines , [online] Available at: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 [Accessed 16 Apr 2020]

[15] Lipo Wang , E.D (2010) , Book Title Support vector machines: theory and applications ,Book Publisher Springer , [online] Available at: https://books.google.com.sa/books?id=uTzMPJjVjsMC&printsec=frontcover&dq=what+is+support+vector+machines&hl=ar&sa=X&ved=0ahUKEwi-zIiw2_ToAhXLUBUIHaFpCrQQ6AEIKTAA#v=onepage&q=what%20is%20support%20vector%20machines&f=false page 11- 22 [Accessed 16 Apr 2020]

[16] java point, About K-Nearest Neighbor , [online] Available at: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning [Accessed 16 Apr 2020]

[17] Toward data science, About Random forest, [online] Available at: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 [Accessed 17 Apr 2020]

[18] Yu.l. Pavlov , P.O (2019) , Book Title Random Forests, Book Publisher De Gruyter , [online] Available at: https://books.google.com.sa/books?id=07gpKU3npYUC&printsec=frontcover&dq=what+is+random+forest&hl=ar&sa=X&ved=0ahUKEwjz6v-

lk_XoAhVRQEEAHT_TCQ8Q6AEIMTAB#v=onepage&q&f=false [Accessed 17 Apr 2020]

[19] Statistic Solution, About logistic regression, [online] Available at: https://www.statisticssolutions.com/what-is-logistic-regression/ [Accessed 17 Apr 2020]

[20] saedsayad , About logistic regression , [online] Available at: https://saedsayad.com/logistic_regression.htm [Accessed 17 Apr 2020]

[21] Toward data science, About Ensemble learning [online] Available at: https://towardsdatascience.com/simple-guide-for-ensemble-learning-methods-d87cc68705a2 [Accessed 17 Apr 2020]

[22] Digital Trends About artificial neural network, [online] Available at: https://www.digitaltrends.com/cool-tech/what-is-an-artificial-neural-network/ [Accessed 17 Apr 2020]

[23] Toward data science, about artificial neural network [online] Available at: https://towardsdatascience.com/introduction-to-artificial-neural-networks-ann-1aea15775ef9 [Accessed 17 Apr 2020]