MACHINE LEARNING FOR MULTIPLE STAGE HEART DISEASE PREDICTION

Khalid Amen[1], Mohamed Zohdy[1] and Mohammed Mahmoud[2]

[1]Department of Electrical and Computer Engineering,
Oakland University, Rochester, MI, USA
[2]Department of Computer Science and Engineering,
Oakland University, Rochester, MI, USA

ABSTRACT

According to the Centers for Disease Control and Prevention (CDC), heart disease is the number one cause of death for men, women, and people of most racial and ethnic groups in the United States. More than one person dies every minute and nearly half a million die each year from it, costing billions of dollars annually. Previous machine learning approaches have been used to predict whether patients have heart disease. The purpose of this work is to predict the five stages of heart disease starting from no disease, stage 1, stage 2, stage 3, and advance condition or severe heart disease. We investigate different potential supervised models that are trained by machine learning algorithms and find out which of these models has better accuracy. In this paper, we describe and investigate five machine learning algorithms (SVM, LR, RF, GTB, ERF) with hyper parameters that maximize classifier performance to show which one is the best to predict the stage at which a person is determined to have heart disease. We found that the LR algorithm performs better compared to the other four algorithms. The experiment results show that LR performs the best with an accuracy of 82%, followed by SVM with an accuracy of 80% when all five classifiers are compared and evaluated for performance based on accuracy, precision, recall, and F measure. This predication can facilitate every step of patient care, reducing the margin of error and contributing to precision medicine. Lastly, this paper aims to improve heart disease prediction accuracy, precision, recall and F measure using UCI heart disease dataset. For this, multiple machine learning approaches were used to understand the data and predict the chances of heart disease in a medical database.

KEYWORDS

machine learning, ml, cnn, dnn, rnn, jupyter, python, cleveland dataset, gradient tree boosting, gtb, random forest, rf, support vector machine, svm, extra random forest, erf, logistic regression, lr.

1. INTRODUCTION

1.1. Machine Learning

Machine learning is the process of teaching a computer system how to make accurate predictions when provided data. It uses algorithms and neural network models to assist computer systems in progressively improving their performance. Machine learning algorithms automatically build a mathematical model using sample data – also known as "training data" – to make decisions without being specifically programmed to make those decisions [1] [2] [6].

Dhinaharan Nagamalai et al. (Eds): CSEIT, WiMoNe, NCS, CIoT, CMLA, DMSE, NLPD - 2020
pp. 205-223, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101118

Those predictions could be answering whether a piece of fruit in a photo is a banana or an apple, spotting people crossing the road in front of a self-driving car, whether the use of the word book in a sentence relates to a paperback or a hotel reservation, if an email is spam, or recognizing speech accurately enough to generate captions for a YouTube video [2] [4].

Machine learning is used across many spheres around the world. The healthcare industry is no exception. Machine learning can play an essential role in predicting presence/absence of locomotor disorders, heart diseases and more. Such information, if predicted well in advance, can provide important insights to doctors who can then adapt their diagnosis and treatment to a per patient basis [3] [4] [5].

Machine learning when applied to health care is capable of early detection of disease which would aid to provide early medical intervention. In heart disease prediction, machine learning techniques have played a significant role. Analysis of disease has become vital in health care sectors. The massive data collected by healthcare sectors are preprocessed and analyzed to discover the underlying information in the data for effective decision making and to provide proper medical intervention. The success of machine learning in the medical industry is its capability in analyzing the huge amount of data gathered by the health sector and its effectiveness in decision-making. Since the medical field involves too many manual processes, it has become necessary to automate these procedures. Remarkable advancements in electronic medical records have made it possible. Diagnosing diseases is an intricate job in the medical field [1] [3] [4] [7].

In order to conduct this prediction, a Jupyter notebook was constructed in Python using the publicly available Cleveland dataset for heart disease, which has over 300 unique instances with 76 total attributes. From these 76 attributes, only 14 of them are commonly used for research to this date. In addition, the hyperparameters used in this prediction come from the recommendations by Dr. Olson, “data-driven advice for applying machine learning to bioinformatics problems” [17]. The libraries and coding packages used in this analysis are: SciPy, Python, NumPy, IPython, Matplotlib, Pandas, Scikit-Learn, and Scikit-Image [18] [23].

1.2. Heart Disease

Heart disease describes a range of conditions that affect the heart. Diseases under the heart disease umbrella include blood vessel diseases such as coronary artery disease, heart rhythm problems, and congenital heart defects, among others [20] [21].

The term "heart disease" is often used interchangeably with the term "cardiovascular disease." Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect the heart's muscle, valves or rhythm, are also considered forms of heart disease [20] [21] [22].

Heart disease causes roughly 735,000 heart attacks each year in the U.S. killing more than 630,000 Americans. According to the American Heart Association, over 7 million have suffered a heart attack in their lifetime [22].

There are several risk factors for heart disease; some are controllable, others are not. Uncontrollable risk factors for heart disease include male, older age, family history of heart disease, being postmenopausal, and race. About half of Americans (47%) have at least one out of three key risk factors for heart disease: high blood pressure, high cholesterol and smoking [21] [26].

Heart disease is the number one killer of both men and women. Heart disease can happen at any age, but the risk increases as people get older. Children of parents with heart disease are more likely to develop heart disease themselves. African-Americans have more severe high blood pressure than Caucasians, and a higher risk of heart disease. Heart disease risk is also higher among Mexican-Americans, American Indians, native Hawaiians and some Asian-Americans. This is partly due to higher rates of obesity and diabetes [20] [21].

Genetic factors likely play some role in high blood pressure, heart disease, and other related conditions. However, it is also likely that people with a family history of heart disease share common environments and other factors that may increase their risk. Most people with a significant family history of heart disease have one or more other risk factors. Just as you cannot control your age, sex and race, you cannot control your family history; so it's even more important to treat and control any other modifiable risk factors you have [22] [4] [21].The risk for heart disease can increase even more when heredity is combined with unhealthy lifestyle choices, such as smoking cigarettes and eating an unhealthy diet [21] [22] [26].

High blood pressure increases the heart's workload, causing the heart muscle to thicken and become stiffer. This stiffening of the heart muscle is not normal and causes the heart to function abnormally. It also increases risk of stroke, heart attack, kidney failure and congestive heart failure [23].

When high blood pressure is present alongside obesity, smoking, high cholesterol levels or diabetes, the risk of heart attack or stroke increases even more. Some risk factors for heart disease cannot be controlled, like family history, for example. But it's still important to lower the chance of developing heart disease by decreasing the risk factors that can be controlled.

2. RELATED WORK

Many researchers have completed a lot of work on data analysis and survivability analysis through Machine Learning (ML) and Data Mining (DM) approaches. Several studies reported that these techniques are significant for future predictions such as in the field of medical diagnosis. In these studies, the authors applied multiple approaches to specific problems and achieve high classification accuracies e.g. in the healthcare industry, these techniques are used for disease prediction.

In [1], [38] author applied Decision Tree (DT), LL, NB, SVM, KNN, PCA, ICA classifier respectively to analyze the kidney disease data. Early detection and treatment of the diseases prevents it from getting to the worst stage making it not only difficult to cure but also impossible to provide treatment. Breast cancer affects many women, so researchers work on different classifiers such that DT, SMO, BF Tree and IBK help to analyze the breast cancer data and examine the performance of the related techniques in order to accurately predict breast cancer using DT [39] and Weka software [9]. RBF Network, Rep Tree, and Simple Logistic DM techniques are used to predict and resolve the survivability of breast cancer patient [40]. Simple Logistic is used for dimension reduction and proposed RBF Network and Rep Tree model used for fast diagnosis of the other diseases.

The prediction of heart disease and patient survivability has been a critical research problem for a few decades. Globally, heart diseases are one of the major cause of deaths. About 80% of deaths in low and middle-income countries are due due to heart diseases [41]. Researchers use multiple DM techniques to develop a prediction model for the survival of heart disease patients. K-mean, C4.5 techniques are used in [8]. NB, J48 DT and Bagging algorithm, CART, ID3 (Iterative

Dichotomized 3) and Decision Table, Logistics Classification, Multilayer Perception and SMO; these three algorithms are respectively used in [41], [9], [10] to predict heart disease patients and their survivability. However, with the advancement of these technologies, the measurements are not sufficient for the prediction of diseases.

3. BACKGROUND OF CLEVELAND DATASET

Experiments with the Cleveland dataset have concentrated on simply attempting to distinguish the presence of heart disease from absence [15]. The 14 attributes that were used are listed in Table 1.

Table 1: Cleveland dataset attributes

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	0 = female 1 = male
Cp	Discrete	Chest pain: 1 = typical angina, 2 = atypical angina, 3 = non-angina pain, 4 = asymptom
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar>120 mg/dl: 1 = true 0 = false
Restecg	Discrete	Resting Electrocardiograph
Thalach	Continuous	Exercise Max Heart Rate Achieved
Exang	Discrete	Exercise Induced Angina: 1=yes 0=no
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment: 1=up sloping 2=flat 3=down sloping
Ca	Continuous	Number of major vessels colored by fluoroscopy that range between 0 and 3
Tha	Discrete	3=normal 6=fixed defect 7=reversible defect
Class	Discrete	Diagnosis classes: 0=No Presence 1=Least likely to have heart disease 2=>1 3=>2 4=More likely have heart disease

The five stages of our prediction of Heart Disease presented are mapped as follows:

Table 2: Five stages prediction

0	No Heart Disease
1	Stage1
2	Stage2
3	Stage3
4	Heart Disease presented

4. BACKGROUND ON RECOMMENDATION OF MODEL ALGORITHMS

A study conducted by Randal S. Olson provides insightful best practice advice for solving bioinformatics problems with Machine Learning, “Data-driven Advice for Applying Machine Learning to Bioinformatics Problems” [18]. He analyzed 13 state-of-the-art commonly used

Machine Learning algorithms on a set of 165 publicly available classification problems in order to provide data-driven algorithm recommendations to current researchers.

From his findings, he was able to provide a recommendation of five algorithms with hyperparameters that maximize classifier performance across the tested problems, as well as general guidelines for applying machine learning to supervised classification problems. The recommendations are as follows [1]:

Table 3: Five Machine Learning Algorithms

Algorithm	**Parameters**	**Datasets Covered**
GradientBoostingClassifer	Loss= devianceLearning_rate=0.1, n_estimators=500 max_depth=3, max__features = log2	
RandomForestClassifier	n_estimators=500, max__features = 0.25, criterion=entropy	19
SVC	C=0.01, gamma=0.1, degree=3, coef0=10.0	16
ExtraTreesClassifier	n_estimators=1000, max__features = log2, criterion=entropy	12
LogisticRegression	C=1.5, Penalty=L1, Fit_intercept=true	8

The database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by DL researchers to this date. The "num" field in the figure refers to the presence of heart disease in the patient. It is integer valued from zero (no presence) to four. Experiments with the Cleveland database have concentrated on attempting to distinguish presence (values 1,2,3,4) from absence (value 0) [15] [23].

1. #3 (age)
2. #4 (sex)
3. #9 (cp)
4. #10 (trestbps)
5. #12 (chol)
6. #16 (fbs)
7. #19 (restecg)
8. #32 (thalach)
9. #38 (exang)
10. #40 (oldpeak)
11. #41 (slope)
12. #44 (ca)
13. #51 (thal)
14. #58 (num) (the predicted attribute)

5. APPROACH

The Data is split into 80% training (237 people) and 20% testing (60 people) after we dropped six from missing values. Several different models are evaluated through k-fold Cross-Validation with k-fold = 10 using GridSearchCV, which iterates on different algorithm's hyperparameters:

1. Gradient Tree Boosting (GradientBoostingClassifier).
2. Random Forest (RandomForestClassifier).
3. Support Vector Machine (SVC).
4. Extra Random Forest (ExtraTreesClassifier).
5. Logistic Regression (LogisticRegression).

5.1. Gradient Tree Boosting

Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error. The target outcome for each case in the data depends on how much changing that case's prediction impacts the overall prediction error [24]:

- If a small change in the prediction for a case causes a large drop in error, the next target outcome of the case is a high value. Predictions from the new model that are close to its targets will reduce the error.
- If a small change in the prediction for a case causes no change in error, the next target outcome of the case is zero. Changing this prediction does not decrease the error.

The name gradient boosting arises because target outcomes for each case are set based on the gradient of the error with respect to the prediction. Each new model takes a step in the direction that minimizes prediction error, in the space of possible predictions for each training case [25].

This algorithm builds an ensemble of trees in a serial approach, where a weak model, e.g., a tree with only a few splits, is trained first and consecutively improves its performance by maintaining to generate new trees [34]. Each new tree in the sequence is responsible for repairing the previous prediction error [30]. Based on the *grid* search, we set the learning parameters as follows:

Table 4: GTB Parameters

Parameter	Meaning	Value
Learning rate	Impact of each tree on the final outcome	0.1
N_estimators	Number of sequential tree to modeled	500
Max_depth	Max depth of a tree	3
Max_features	Number of features	Log2

5.2. Random Forest

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction.

The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science speak, the reason that the random forest model works so well is a large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models [36]. The low correlation between models is the key. Just like how investments with low correlations come together to form a portfolio that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. The reason for this effect is the trees protect each other from their individual errors. While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction.

This classifier takes bagging of decision tree procedure to evoke a large collection of trees to improve performance. Compared to other similar ensembles, Random Forest (RF) that requires less hyperparameter tuning [27]. Original bagging decision tree yields tree-mutuality, which suffers from the effect of high variance. Hence, RF offers a variance reduction by introducing more randomness into the tree-generation procedure. Based on grid search, we set the learning parameters as follows:

Table 5: RF Parameters

Parameter	Meaning	Value
N_estimators	Number of trees in the forest	500
Max_features	Number of features to consider	1
Criterion	Function to measure the quality of a split	Entropy

5.3. Support Vector Machine (SVM)

Support Vector Machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. In 1960s, SVMs were first introduced but later they were refined in 1990. SVMs have their unique way of implementation as compared to other machine learning algorithms. Lately, they are extremely popular because of their ability to handle multiple continuous and categorical variables [29].

An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).

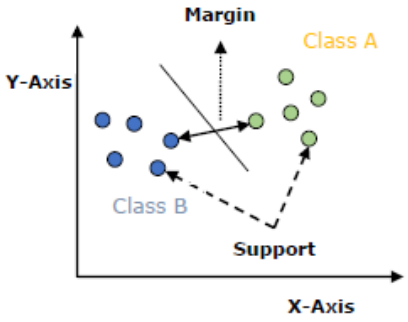

Figure 1: SVM Classes

The import concepts in SVM are:

- Support Vectors – Datapoints that are closest to the hyperplane are called support vectors. Separating line will be defined with the help of those data points.
- Hyperplane – As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes.
- Margin – It may be defined as the gap between two lines on the closet data points of different classes. It can be calculated as the perpendicular distance from the line to the support vectors.
- Large margin is considered a good margin and small margin is considered a bad margin.

The main goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH) and it can be done in the following two steps

- First, SVM will generate hyperplanes iteratively that segregate the classes in the best way.
- Then, it will choose the hyperplane that separates the classes correctly.

In practice, SVM algorithm is implemented with kernel that transforms an input data space into the required form. SVM uses a technique called the kernel trick in which kernel takes a low dimensional input space and transforms it into a higher dimensional space. In simple words, kernel converts non-separable problems into separable problems by adding more dimensions to it. It makes SVM more powerful, flexible and accurate. The following are some of the types of kernels used by SVM [28] [29] [30].

- Linear Kernel – It can be used as a dot product between any two observations. The formula of linear kernel is:

  $$K(x, y) = sum(x * y) \quad (1)$$

  From the above formula, the product between two vectors (x) and (y) is the sum of the multiplication of each pair of input values.
- Polynomial Kernel – It is more generalized form of linear kernel and distinguish curved or nonlinear input space. The formula of Polynomial kernel is:

  $$K(x, y) = 1 + sum(x * y)\text{^}d \quad (2)$$

  where d is the degree of polynomial which it be specified manually in the learning algorithm.
- Radial Basis Function (RBF) kernel – it is mostly used in SVM classification. It maps input space in indefinite dimensional space. The formula if RBF is:

  $$K(x, y) = exp\big(gamma * sum(x - y^2)\big) \quad (3)$$

  Gamma ranges from 0 and 1 and it needs to be specified manually in the learning algorithm.

Based on grid search, we set the learning parameters as follows:

Table 6: SVM Parameters

Parameter	Meaning	Value
C	Penalty or Regularization Parameter	0.01
Gamma	Gamma coefficient	0.1
Kernel	Kernel Type	rbf
Degree	Degree of the polynomial function	3
Coef0	Independent term in kernel function	10.0

5.4. Extra Random Forest (Extra Tree Classifier)

Extra Trees Classifier is an ensemble machine learning algorithm. In other words, it is an ensemble of decision trees and is related to other ensembles of decision trees algorithms such as bootstrap aggregation (bagging) and random forest [31]. The Extra Trees algorithm works by creating a large number of unpruned decision trees from the training dataset. Predictions are made by averaging the prediction of the decision trees in the case of regression or using majority voting in the case of classification [32]. Unlike bagging and random forest that develop each decision tree from a bootstrap sample of the training dataset, the Extra Trees algorithm fits each decision tree on the whole training dataset [31] [32] [37].

Like random forest, the Extra Trees algorithm will randomly sample the features at each split point of a decision tree and select a split point at random.

Based on grid search, we set the learning parameters as follows:

Table 7: ERF Parameters

Parameter	Meaning	Value
N_estimators	The number of trees in the forest	1000
Max_features	The number of features to consider	Log2
Criterion	The function to measure the quality of a split	Entropy

5.5. Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of the target or dependent variable is dichotomous, which means there will be only two possible classes [33]. In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, diabetes prediction, cancer detection etc. [17].

5.5.1.Type of Logistics Regressions

Logistic Regression means binary logistic regression having binary target variables, but there can be two more categories of target variables that can be predicted by it. Based on those number of categories, Logistic regression can be divided into following types [38]:

- Binary or Binomial – In such a classification, a dependent variable will have only two possible types either 1 and 0. For example, these variables may represent success or failure, yes or no, win or loss etc.
- Multinomial – In such a classification, dependent variable can have three or more possible unordered types or the types having no quantitative significance. For example, these variables may represent “Type A” or “Type B” or “Type C”.
- Ordinal – In such a classification, dependent variable can have three or more possible ordered types or the types having a quantitative significance. For example, these variables may represent “poor”, “good”, “very good” or “excellent” and each category can have scores such as 0, 1, 2 or 3.

Based on grid search, we set the learning parameters as follows:

Table 8: LR Parameters

Parameter	Meaning	Value
C	Inverse of regularization strength	1.5
Penalty	Specify the norm used in the penalization	L2
Fit_intercept	Specifies if a constant should be added to decision function	True

6. METHODOLOGY

The proposed methodology using fiveclassification techniques; Gradient Tree Boosting, Random Forest, Support Vector Machine (SVM), Extra Random Forest, and Logistic Regression to predict heart disease as the proposed methodology shown in Fig 4. These classifiers are used to improve the prediction. We applied the classifiers in Fig 5to heart disease data that comes from the Cleveland dataset to predictin which of five stages a patient has heart problems. The performance of these classifiers are to evaluate on the bases of accuracy, precision recall, and F measure.

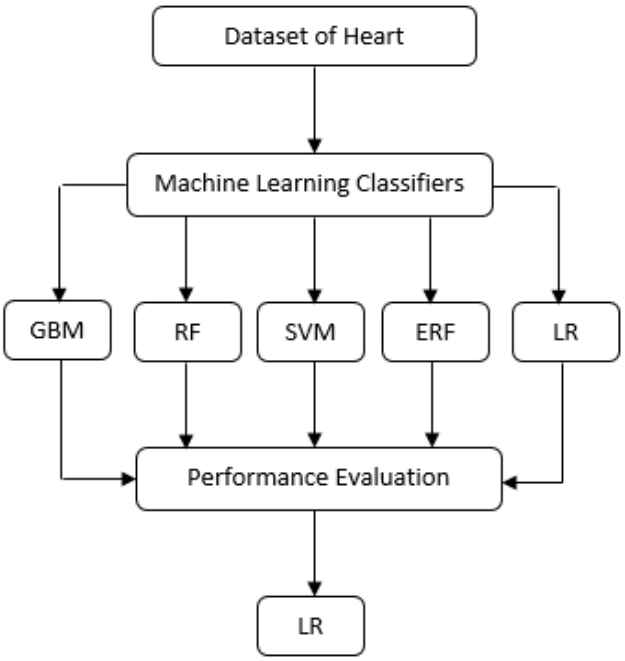

Figure 2: Proposed Methodology

The dataset of heart is taken from UCI repository [23], the classifier taking it as input for disease prediction. These classifiers are implemented in Python language. Python is a powerful interpreter language and a reliable platform for research [19]. The accuracy of prediction increased by comparing the results of these five classifiers using evaluation parameters. The experimental result describes which classifier is best between them.

A. Evaluation Parameters

Some evaluation parameters in data mining are accuracy, precision, recall, and F measure. Where TPTrue Positive, TN- True Negative, FP- False Positive and FN- False Negative [19].

- Accuracy is defined as the number of accurately classified instances divided by the total number of instances in the dataset as in (4).

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

- Precision is the average probability of relevant retrieval as described in (5).

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

- The recall is defined as the average probability of complete retrieval as defined in (6).

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

- F- Measure is the calculated by using both precision and recall as shown in (7).

$$\text{F Measure} = \frac{2 * (Precision * Recall)}{Precision + Recall} \quad (7)$$

Where all evaluation parameters accuracy, precision, recall, and F measure are calculated from dataset when splitting the dataset into training data and test data.The Pseudo codes for the evaluation parameters are as follow:

```
Def evaluationParameters(X_train, y_train, X_test, y_test):
        X_train ← fit_transform(X_train)
        Classifier ←sklearn()
        y_pred ← classifier.predict(X_test)
        cm_test ← confusion_matrix(y_pred, y_test)
        y_pred_train ← classifier.predict(X_train)
        cm_train ← confusion_matrix(y_pred_train, y_train)
        training_accuracy = (cm_train[0][0] + cm_train[1][1])/len(y_train)
        test_accuracy = (cm_test[0][0] + cm_test[1][1])/len(y_test)
        training_percision = cm_train[0][0]/(cm_train[0][0] + cm_train[1][0])
        test_percision = cm_test[0][0]/(cm_test[0][0] + cm_test[1][0])
        training_recall = cm_train[0][0]/(cm_train[0][0] + cm_train[0][1])
        test_recall = cm_test[0][0]/(cm_test[0][0] + cm_test[0][1])
                training_f_measure ← (2 * training_percision *
        training_recall)/(training_percision + training_recall))
        test_f_measure ← (2 * test_percision * test_recall)/(test_percision + test_recall))

return (training_accuracy, test_accuracy, training_percision, test_percision, training_recall,
test_recall,          training_f_measure,  training_f_measure)
```

B. Dataset

To perform the research, heart disease datasetis used. This heart disease dataset contains 14 attributes and 303 instances. This dataset is taken from UCL repository. It's an online repository that contains 412 diverse datasets. UCI provides data for ML to perform analysis in a different direction. The UCI database is known for its extensiveness in data, its completeness, and accuracy [23].

C. Machine Learning Classifiers:

In this research, five classification methods are implemented in python using the pandas and keras libraries. These models are used to improve prediction. These classifiers are compared to find out which of the five stages best predictsthe chance of heartdisease in patients. In the next section, we briefly describe these classification techniques/ classifiers.

1) Logistics Regression (LR): is the predictive analysis to conduct on discrete (binary) values based on a specified set of independent variables. LR describes the data and clarifies the relationship between one (binary) dependent variable and independent variables. It predicts the event occurrence probability by fitting the data into a logit function. Therefore, it is also called logit regression [32].

Input values x are linearly combined using coefficient values b, to calculate an output value p. The output values as predictable lies between 0 and 1. Input data associated with coefficient b (constant value) learned from training data. Where p is the output, b0 is an intercept term and b1 is the coefficient of input value x as shown in (8).

$$P = \frac{e^{(b0+b1+x)}}{1 + e^{(b0+b1+x)}} \quad (8)$$

2) Support Vectors Machine (SVM): is a classification and regression algorithm. In SVM, every data item is plotted in n-dimensional space, a number of dimensions are equivalent to the number of features or attributes. Where n represents the number of attributes. The value of each attribute being the value of certain coordinates. Once plotting all the data items then performed classification by drawing a line or by finding the optimal hyperplane that separates two classes completely. For example, if we have two features of individual like hair and height length. First, we plot these two features in two-dimensional space where every point has two coordinates (these co-ordinates are also known as Support Vectors) [29].

3) Random Forests (RF): are ensemble learning technique for regression, regression, classification, and for other tasks. That operate by making a multitude of Decision Tree (DT) at training stint and outputting that is the mean prediction (regression) or mode of classes (classification) of the distinct trees.

Every tree in the forest contributes for a classification. To classify new case based on its attributes. We identify the tree "votes" for that class so the forest indicates the classification of the case that is taking the most votes [37].

Every tree is planted and grown as follows:

- If the number of objects N in training set, the sample of N objects is taken randomly with replacement. This sample act as a training set for growing tree.
- If there is an input variable N,numbern < N is stated that at each node, randomly selection of n variable out of input variable N. So, the best splitting on n is used to split the node. The value of m (node splitter) is constant during gowning the forest.
- Each tree is growing up to the largest magnitude possible so there is no trimming.

4) Gradient Tree Boosting: is a machine learning technique for regression and classification problems, which produces a predication model in the form of an ensemble of weak predication models. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function [35].

It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error. How are the targets calculated? The target outcome for each case in the data depends on how much changing that case's prediction impacts the overall prediction error.

5) Extra Random Forest (ERF): is very similar to Random Forests (RF) but there are two main differences:

- ERF does not resample observations when building a tree. They do not perform bagging.
- ERF does not use the best split. Like RF, ERF selects a random subset of predictors for each split. Instead of the best split for predictors, ERF makes a small number of randomly chosen splits-points for each of the selected predictors. ERF then selects the best split from this small number of choices.

In ERF, the features and splits are selected at random. Since splits are chosen at random for each feature, it is less computationally expensive than RF [31].

D. Scaling the Data:

To accomplish the five stages output prediction for a patient to be diagnosed with one of five stages, it is important to scale the data so the machine learning algorithms do not overfit to the wrong features. Using the MinMaxScaler() method on Python, the values are scaled per features based on the minimum and maximum between 0 and 1. This keeps the information from being lost but allows the machine learning algorithms to correctly train with the data. The training data and test data are scaled between 0 and 1 and the output data is scaled between 0 and 1 as well. Then, the scaled output value is mapped as follows:

Table 9: Five Stages

Output Value	Stage
0	No disease presented
0 < and <= 0.25	Stage1
0.25 < and <= 0.5	Stage2
0.5 < and <= 0.75	Stage3
0.75 < and <= 1	Advance disease presented

7. EXPERIMENTAL RESULT

The experiment is conducted for the prediction of heart disease stages by applying various machine learning classifiers. From the experiment results, we identify that Logistic Regression performs better as compared to the other four ML classifiers in the prediction of these diseases. In this experiment, we use multiple stages of heart disease prediction to forecast the stage at which a person is determined to have heart disease. In previous works [42] [43] [44] [45], the study used two outcome predications, either a person has the disease or not; that is represented by (0 ,1) or (true, false).The Pseudocodes for the experiment are as follow:

```
data_frame ← read_CSV_file
X ← data_frame [column: 0 - 12]
y ← data_frame [column: 13]
target ← preprocessing.scale(y)
data ← preprocessing.scale(X)

for k ← 0 to data - 1
        if data[k] = 0 then
        data[k] ← 'no disease'
        if data[k] > 0 && data[k] <= 0.25 then
        data[k] ← 'stage1'
        if data[k] > 0.25 && data[k] <= 0.5 then
        data[k] ← 'stage2'
        if data[k] > 0.5 && data[k] <= 0.75 then
        data[k] ← 'stage3'
        else
        data[k] ← 'disease presented'

X_train, X_test, y_train, y_test ← train_test_split(X, y, test_size=0.2, random_state=0)
svm(X_train, y_train, X_test, y_test)
lr(X_train, y_train, X_test, y_test)
rf(X_train, y_train, X_test, y_test)
gtb(X_train, y_train, X_test, y_test)
erf(X_train, y_train, X_test, y_test)
```

The below Figures show the performance of various evaluation parameters in the prediction of heart disease. The experimental results show the comparison of LR, ERF, GTB, SVM and RF classifiers and evaluate the performance on the bases of accuracy, precision, recall and F measure. In all classifiers, LR performs the best with an accuracy of 82%, followed by SVM with an accuracy of 80%.

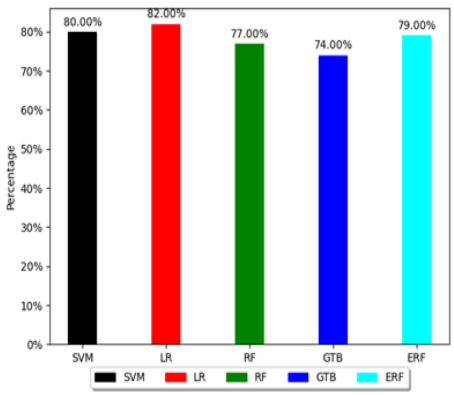

Figure 3: Heart Disease Accuracy

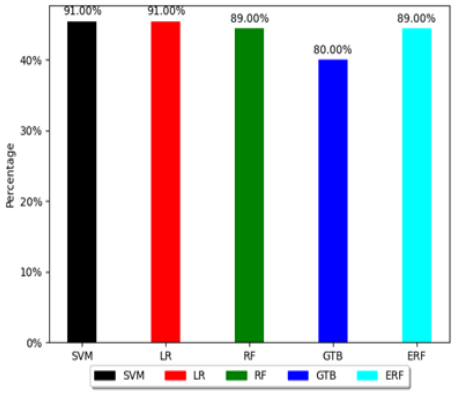

Figure 4: Heart Disease Precision

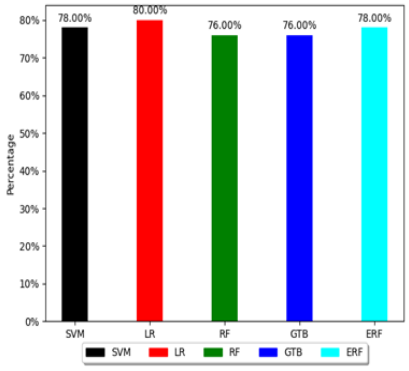

Figure 5: Heart Disease Recall

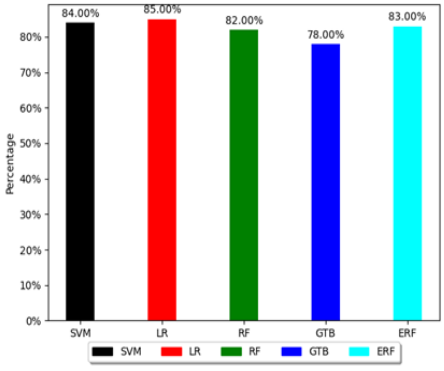

Figure 6: Heart Disease F Measure

Table 10: ML Algorithms Comparison

Algorithm	Accuracy	Precision	Recall	F Measure
SVM	80%	91%	78%	84%
LR	82%	91%	80%	85%
RF	77%	89%	76%	82%
GTB	74%	80%	76%	78%
ERF	79%	89%	78%	83%

8. CONCLUSIONS

The importance of extracting the valuable information from raw data has very good consequences in many fields of life such as the medical area, business area, and more. In this study, we proposed a multiple stage detection model of heart disease based on five algorithms to compare which one performs better. The proposed method was built by the stacking of five different ensemble learners, such as the Random Forest, Gradient Boosting Machine, Extreme Random Forest, Logic Regression, and Support Vector Machine. The proposed detection model was tested on well-known Cleveland dataset in order to provide a fair benchmark against existing studies. Based on the experimental results, our proposed model was able to outperform heart disease detection methods with respect to accuracy, precision, recall and F measure. The result reflected

the highest result obtained showed that Logic Regression has a better result comparing to the other four methods or algorithms. The performance was further enhanced using feature selection techniques. The feature selection techniques helped to improve the accuracy, precision, recall, and F measure of the ensemble algorithms.The experiment results show that LR performs the best with an accuracy of 82%, followed by SVM with an accuracy of 80% when all five classifiers are compared and evaluated for performance based on accuracy, precision, recall, and F measure

9. LIMITATIONS

The Cleveland heart dataset from UCI machine learning repository was utilized for training and testing purposes. The ML algorithms SVM, LR, RF, GTB, and ERF were employed for experiments. As far as the dataset are concerned, they need to be amplified as the main limitation in this work is the small size of the dataset. If the dataset has bad data and is not caught, then this would generate bad or inaccurate predictions. The dataset has a limited number of patient records; therefore, the dataset was augmented using appropriate techniques.

10.FUTURE WORK

There are many possible improvements that could be explored to improve accuracy, precision, recall, and F measure of this prediction system. Due to time limitations, the following research/ work needs to be performed in the future:

- There is a need of more ML algorithms for comparison.
- Large dataset to be trained
- Build a system or a framework for automation of heart disease predication.
- Real data from health care organizations and agencies needs to be collected and all the available techniques will be compared for the optimum accuracy.
- Use deep learning to structure algorithms in layers to create Artificial Neural Network (ANN) or Convoluted Neural Network (CNN) that can learn and make intelligent decision.

11.ACKNOWLEDGEMENTS

This paper and the research behind it would not have been possible without the grace, the bounty, and the blessing of almighty Allah (God) first and foremostand the exceptional support of my professors, Mohamed Zohdy and Mohammed Mahmoud. Their enthusiasm, knowledge and exacting attention to detail have been an inspiration and kept my work on track from my first encounter with machine learning research to the final draft of this paper.

REFERENCES

[1] R. Garg, "A Comparative Study of Different Classification Algorithms on Kidney Disease Prediction," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 6, no. 2, pp. 741–746, Feb. 2018, doi: 10.22214/ijraset.2018.2132.0

[2] Nick Health, "What is machine learning? Everything you need to know", https://www.zdnet.com/article/what-is-machine-learning-everything-you-need-to-know/, Sep. 2018

[3] G. Kaur and A. Sharma, "Predict chronic kidney disease using data mining algorithms in hadoop," in *2017 International Conference on Inventive Computing and Informatics (ICICI)*, Nov. 2017, pp. 973–979, doi: 10.1109/ICICI.2017.8365283.

[4] P. K. Sahoo, S. K. Mohapatra, and S.-L. Wu, "Analyzing Healthcare Big Data With Prediction for Future Health Condition," *IEEE Access*, vol. 4, pp. 9786–9799, 2016, doi: 10.1109/ACCESS.2016.2647619.

[5] Karan Bhanot, "Predicting presense of Heart Diseases using Machine Learning", https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c , Feb. 2019

[6] Keith Foote, "A Brief History of Machine Learning", https://www.dataversity.net/a-brief-history-of-machine-learning/, Mar. 2019.

[7] N. A. R. and D. R. Vincent, "Heart Disease Prediction System Using Ensemble of Machine Learning Algorithms," *Recent Patents Eng.*, vol. 13, Mar. 2019, doi: 10.2174/1872212113666190328220514.

[8] A.Rairikar, V. Kulkarni, V. Sabale, H. Kale, and A. Lamgunde, "Heart disease prediction using data mining techniques," in *2017 International Conference on Intelligent Computing and Control (I2C2)*, Jun. 2017, pp. 1–8, doi: 10.1109/I2C2.2017.8321771.

[9] V. Chaurasia, "Early Prediction of Heart Diseases Using Data Mining," *Caribb. J. Sci. Technol.*, vol. 1, pp. 208–217, 2013, doi: 10.1377/hlthaff.2014.0041.

[10] S. Vijiyarani, S. Sudha, and M. P. Research Scholar, "An Efficient Classification Tree Technique for Heart Disease Prediction," 2013

[11] A. Chaudhary and P. Garg, "Detecting and Diagnosing a Disease by Patient Monitoring System," *Int. J. Mech. Eng. Inf. Technol.*, vol. 2, no. 6, pp. 493–499, 2014.

[12] P. R. V. A. M Archana Bakare, "Prediction of Diseases using Big Data Analysis," *Int. J. Innov. Res. Comput. Commun. Eng. (An ISO Certif. Organ.*, vol. 3297, no. 6, pp. 11449–11455, 2016, doi: 10.15680/IJIRCCE.2016.

[13] A.V. Solanki, "Data Mining Techniques Using WEKA classification for Sickle Cell Disease," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 4, pp. 5857–5860, 2014, [Online]. Available: http://www.ijcsit.com/docs/Volume 5/vol5issue04/ijcsit20140504222.pdf.

[14] V. S and D. S, "Data Mining Classification Algorithms for Kidney Disease Prediction," *Int. J. Cybern. Informatics*, vol. 4, no. 4, pp. 13–25, 2015, doi: 10.5121/ijci.2015.4402.

[15] Heart Disease in Cleveland, https://www.rpubs.com/aepoetry/log_reg_heart

[16] V. Kunwar, K. Chandel, A. S. Sabitha, and A. Bansal, "Chronic Kidney Disease analysis using data mining classification techniques," in *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, Jan. 2016, pp. 300–305, doi: 10.1109/CONFLUENCE.2016.7508132.

[17] G. Caocci, R. Baccoli, R. Littera, S. Orru, C. Carcassi, and G. La, "Comparison Between an Artificial Neural Network and Logistic Regression in Predicting Long Term Kidney Transplantation Outcome," in *Artificial Neural Networks - Architectures and Applications*, InTech, 2013.

[18] Dheeru Due, UC Irvine Machine Learning Repositoty, https://archive.ics.uci.edu/ml, Sep. 2018

[19] Simon Tavasoli, Machine Learning Algorithms,https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article, Jul. 2020.

[20] WebMD, Risk Factors for Heart Disease, https://www.webmd.com/heart-disease/risk-factors-heart-disease

[21] CDC, Know Your Risk for Heart Disease, https://www.cdc.gov/heartdisease/risk_factors.htm, Dec. 2019

[22] American Heart Association, Understand Your Risks to Prevent a Heart Disease, https://www.heart.org/en/health-topics/heart-attack/understand-your-risks-to-prevent-a-heart-attack#:~:text=An%20inactive%20lifestyle%20is%20a,blood%20pressure%20in%20some%20people , Jun. 2016

[23] Steven Smiley, Diagnositc for Heart Disease with Machine Learning (ML), https://github.com/stevensmiley1989. Feb. 2011.

[24] Jake Hoare, Gradient Boosting Explained - The Coolest Kid on The Machine Learning Block, https://www.displayr.com/gradient-boosting-the-coolest-kid-on-the-machine-learning-block, DisplayR

[25] Wikipedia, Gradient Boosting, https://en.wikipedia.org/wiki/Gradient_boosting, Aug. 2020.

[26] CDC, Heart Disease Facts, https://www.cdc.gov/heartdisease/facts.htm, Jun. 2020.

[27] Bayu Adhi Tama, Sun Im, Seungchul Lee, "Improving an Intellegent Detection System for Coronary Heart Disease Using A Two-Tier Classifer Ensemble", vol 2020, article ID 9816142, April 2020.

[28] KD Nuggets, What is a Support Vector Machine and Why Would Use it?, https://www.kdnuggets.com/2017/02/yhat-support-vector-machine.html, Feb. 2017.

[29] Savan Patel, Chapter 2: SVM (Support Vector Machine) Theory, https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72, May 2017.

[30] Tutorialspoint, ML – Support Vector Machine, https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_support_vector_machine.htm.

[31] Geek for Geeks, ML – Extra Tree Classifer for Feature Selection, https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-

selection/#:~:text=Extremely%20Randomized%20Trees%20Classifier(Extra,to%20output%20it's%20classification%20result, Jan. 2020

[32] Json Brownlee, How to Develop an Extra Trees Ensemble with Python, https://machinelearningmastery.com/extra-trees-ensemble-with-python/, Apr. 2020

[33] Tutorialspoint, Machine Learning – Logistic Regression, https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_logistic_regression.htm.

[34] Pablo Diez, Smart Wheelchairs and Brain-Computer Interfaces, https://www.sciencedirect.com/topics/engineering/confusion-matrix, 2018

[35] Dan Nelson, Gradient Boosting Classifiers in Python with Scikit-Learn, https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn/.

[36] Steven Smiley, Diagnostic for Heart Disease with Machine Learning, https://towardsdatascience.com/diagnostic-for-heart-disease-with-machine-learning-81b064a3c1dd, Jan. 2011.

[37] Frank Ceballos, An Intuitive Explanation of Random Forest and Extra Trees Classifiers, https://towardsdatascience.com/an-intuitive-explanation-of-random-forest-and-extra-trees-classifiers-8507ac21d54b, Jul. 2019.

[38] Sunil Ray, Commonly used Machine Learning Algorithms (with Python and R Codes), https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/,Sep. 2017.

[39] S. A. Kaur Guneet, "Predict Chronic Kidney Disease using Data Mining Algorithms in Hadoop," international J. Adv. Comput. Eng. Netw. , vol. 5, no. 6, pp. 1–5, 2017.

[40] J. Joshi, R. Doshi, and J. Patel, "Diagnosis and Prognosis Breast Cancer Using Classification Rules," *Int. J. Eng. Res. Gen. Sci.*, vol. 2, no. 6, pp. 315–323, 2014, [Online]. Available: www.ijergs.org.

[41] V. Chaurasia and S. Pal, "Data mining techniques: To predict and resolve breast cancer survivability," *Int. J. Comput. Sci. Mob. Comput. IJCSMC*, vol. 3, p. 15, 2017.

[42] A. Vikas Chaurasia and I. Saurabh Pal, "Data Mining Approach to Detect Heart Dieses," *Int. J. Adv. Comput. Sci. Inf. Technol.*, vol. 2, no. 4, pp. 2296–1739, 2013, [Online]. Available: http://ssrn.com/abstract=2376653.

[43] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[44] D. S. P. Jaymin Patel, Prof.TejalUpadhyay, "Heart Disease Prediction Using Machine learning and Data Mining Technique," *IJCSC*, vol. 7, no. March, pp. 129–137, 2016, doi: 10.090592/IJCSC.2016.018.

[45] L. Parthiban and R. Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm," *Int. J. Biol. Med. Sci.*, vol. 3, no. 3, pp. 157–160, 2008.

[46] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction," *Int. J. Comput. Appl.*, vol. 17, no. 8, pp. 43–48, Mar. 2011, doi: 10.5120/2237-2860.

AUTHORS

Khalid Amen is a System Engineering and Computer Science PhD student in the Electrical and Computer Engineering department, Oakland University, Rochester, MI, USA.

Dr. Mohammed Zohdy is a professor in the Electrical and Computer Engineering department, Oakland University, Rochester, MI, USA.

Dr. Mohammed Mahmoud is a professor in the Computer Science and Engineering department, Oakland University, Rochester, MI, USA.