Heart Disease Prediction Using Classification

Jeffrey Davis
Department of industrial and Systems Engineering
Mississippi State, Mississippi

Abstract

This is research a paper about heart disease prediction using classical machine learning classification methods. The objective of this research is to find the most effective classification model to predict heart disease from various clinical records of patients who have, and do not have heart disease. The models that are applied in this research are random forest classification, support vector machine classification, Multilayer perception classification, and gradient boost Classification. To evaluate and compare these classification models, a dataset from 1988 that consists of four databases of clinical heart disease attributes with clinical records from Cleveland, Hungary, Switzerland, and Long Beach V is used. Training and Testing accuracy for these models will be derived from this dataset and will be compared to determine which model is best fit for this application. This research can help physicians determine if their patients are at risk for heart disease based on other clinical factors.

1. Keywords
Heart Disease Prediction, machine learning, Classification

Heart Disease is the leading cause of death in the United States and worldwide, according to the center for disease control. 1 in every 5 deaths can be attributed to heart disease. According to the CDC, around 47% of people in the United States have at least one risk factor for heart disease such as high blood pressure, or low daily physical activity [5]. Between 2019 and 2020 heart disease cost the U.S 252.2 billion due to, the cost of care, medicine, and lost productivity [1]. Heart disease is encompassing of several different ailments such as arrhythmia, atherosclerosis, cardiomyopathy, congenital heart defects, coronary artery disease and heart infections.

Heart disease can be prevented by living a healthy lifestyle for most, but there are some risk factors that cannot be controlled such as family history, ethnicity, and age [5]. This makes early detection and prevention through early screenings essential. At these screenings, vitals such as cholesterol, blood pressure, and blood sugar levels can be measured [2]. Due to the advancement in the field of machine learning, many top biomedical companies, hospitals, and researchers such as the Mayo Clinic have found that building machine learning models using large databases of patient vitals can help to better predict whether someone is susceptible to heart disease, especially if they appear asymptomatic.

This paper contains various vitals which can be indicators of heart disease. The focus is on the efficiency and accuracy of three classification models, Random Forest, Gradient Booster, and Multilayer perception. To evaluate these models, one dataset is used. The data set contains 13 features with 1025 data points each and is primarily focused on heart disease clinical attributes that could lead to heart disease [13]. To prove the viability of a model compared to another test data accuracy and train data accuracy will be compared amongst all models with the dataset.

This paper is structured into five sections. The first section is an overview of heart disease and a general overview of this paper. Similar research about using machine learning classification to determine heart disease is presented in section 2. Section 3 gives a brief overview of each classification method used in this research, multilayer perception, and random forest and gradient boosting. Section 4 contains the experimental process and dataset description and will contain test results and comparisons. Section 5 will contain a conclusion and insights and lessons learned that may be useful in future research or real-life application.

2. Literature Review

Classification is a predicative modeling process in which machine learning models use algorithms to analyze data from a training data set to predict a correct label for input data [4]. Models are trained used training data which tends to 70 to 75% of a data set. Once the model is trained, its performance can be evaluated against the test data which is the remaining 25 to 30% of the data set. If performance is acceptable then the model can be used to make real-world predictions. This paper focuses on applying several common classification algorithms to a heart disease data set to predict whether a patient is at risk for heart disease or not.

2.1 Classification methods used for heart diseases diagnosis.

There are currently many tests available to be used to diagnose heart disease. These tests can be separated into two categories, invasive and non-invasive. Invasive tests consist of cardiac catheterization, coronary angiography, and electrophysiology. Non-invasive tests include electrocardiogram (EKG), echocardiogram, stress tests, carotid ultrasound, Holter monitor, tilt-table test, CT scan and hear MRI [5]. With the advancement in the field of machine learning, studies have concluded that with biomarkers obtained from the test's previously mentioned, and the help of a qualified medical practitioner, classification models such as Decision Tree, Support Vector Machines, Random Forests, and Bagging and Boosting can achieve a high accuracy in predicting heart disease [6].

There have been many other studies done that have explored the use of classification models with using patient clinicals as input data. In previous research done conducted using. In a study conducted by Hosam El-Sofany, Belgacem Bouallegue & Yasser M. Abd El-Latif , the machine learning classification models logistic regression, random Forest, K-nearest neighbor, decision tree, bagging, adaboost, XGboost, voting, support vector machines, and naïve bayes classier were used on two separate data sets to determine the accuracy of each algorithm based on two separate data sets containing clinical attributes associated with heart disease. When compared to the other algorithms, XGBoost performed with a high accuracy, sensitivity, specificity, precision, F1 score, and AUC. Because of how well this this model performed, the authors are now in the process of implementing it into the creation of a phone app that will allow for users to enter symptoms and known vitals to predict if they are at risk for heart disease [10].

Through their study, several limitations were also found that could affect attempts to replicate or expand of this research. Data set quality and availability, the performance and reliability of ML models depend on the quality and availability of testing and training datasets. Another limitation is missing data. In the real world in real-world healthcare, missing data is widespread and can dramatically impair predictive models. It is crucial to have to address missing data to have the most accurate model [10]. These are just two limitations stated, but there are many more one must be aware of.

3. Methodology

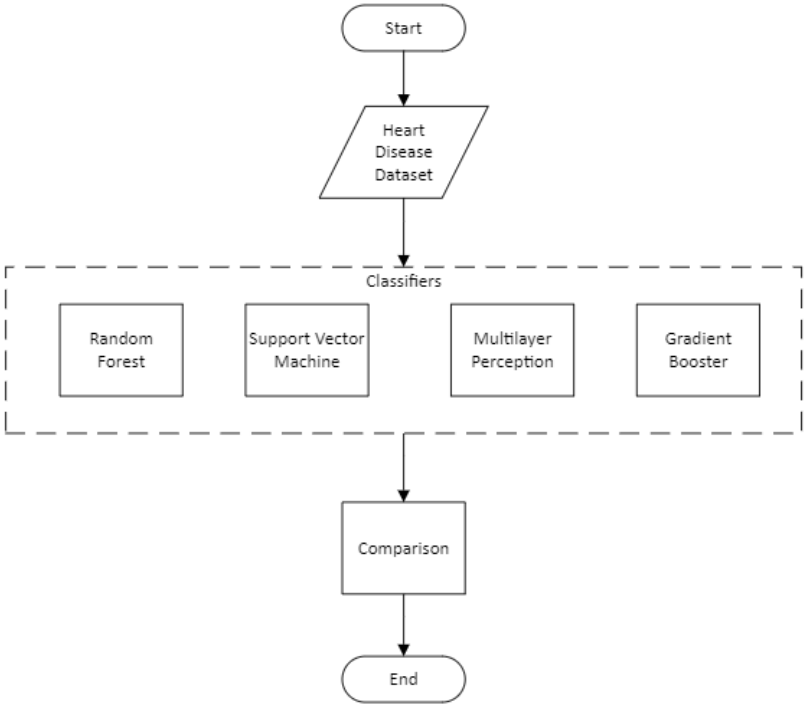


Fig 1. Project Process Flow

The goal of classification is to create a model that can make a prediction from a set of pre-determined outcomes based on the input given. This is achieved by taking a dataset and splitting it into training and testing data. Training data is fitted to a selected model. Testing data is then used to validate the model. Validation in this case is how close the models predictions are. This accuracy is given as a percentage. Random Forest, Multilayer Perception, and Gradient Booster are used in this paper. These models were selected they have tunable parameters which can allow data to be better fit, and they are extremely popular choices. Fig 1 shows the process flow of this project.

3.1 Support Vector Machine Classification

Support Vector Machines were developed in the 1990's by Vladimir N. Vapnik and his colleagues. SVM's are commonly used with classification problems. They distinguish between two classes by finding the optimal hyperplane that maximizes the margin between the closet data points of the opposite class. The number of input features determine if the hyper plane will be in a 2-d space or n-dimensional space. [12]

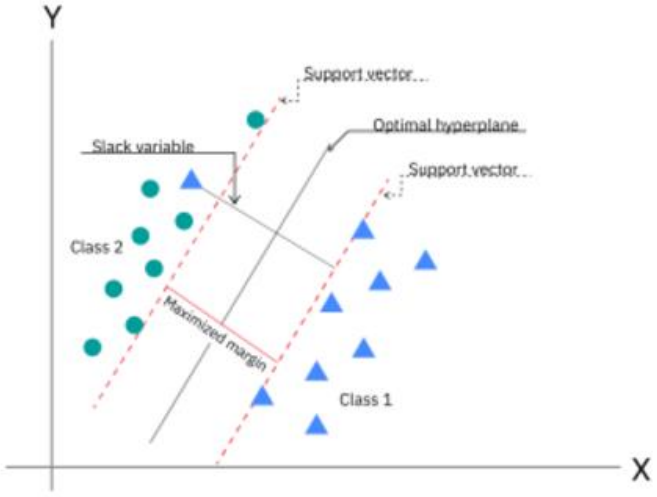


Fig 3: SVM Classification Diagram [12]

3.1 Random Forest Classifier

The concept of random decision forests was first introduced by Salzberg and heath in 1993. They developed a method the used a randomized decision tree algorithm to generate multiple distinct trees and combine their outputs using majority voting [7]. Random forests are a part of the ensemble family, and use bootstrapping (bagging) with replacement, which is taking random samples of the original dataset to produce a new tree.

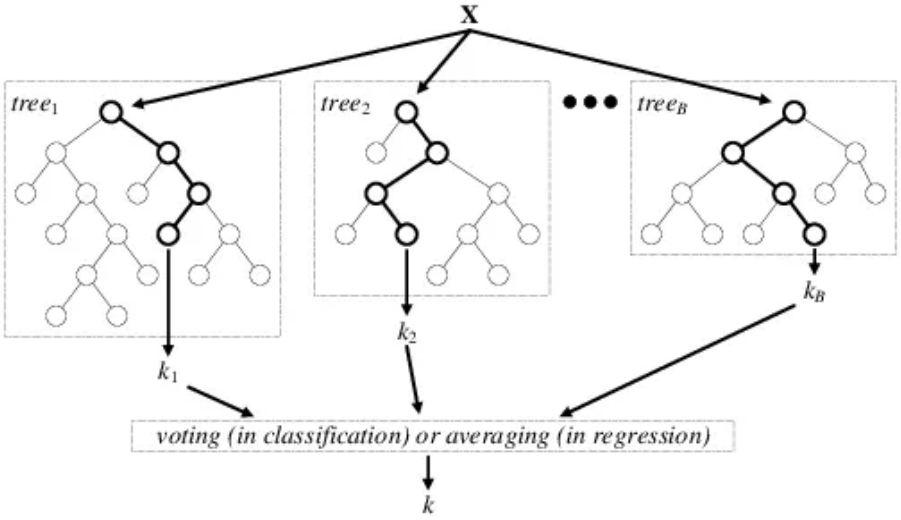


Fig 2: Random forest diagram

3.2 Gradient Boosting

Gradient boosting works by building simpler prediction models sequentially and each models tries to improve upon the error of the previous model [9]. This can be thought of as a simulation of the human learning process where we try and improve upon mistakes we made previously to improve. An example of a gradient boosting algorithm with three trees is as follows:

$$y = A_1 + (B_1 * X) + e_1$$

y is the output model when only fit to one decision tree, e_1 is the residual from the decision tree, and X is input variables. When gradient boosting is applied to the initial tree, the consecutive decision trees are fit to the residual of the previous one [9].

$$e_1 = A_2 + (B_2 * X) + e_2$$

and

$$e_2 = A_3 + (B_3 * X) + e_3$$

The final model of the gradient boosted decision tree is as follows:

$$y = A_1 + A_2 + A_3 + (B_1 * X) + (B_2 * X) + (B_3 * X)$$

3.3 Multilayer Perception Classifier

A Multilayer Perception Classifier is a type of neural network that consists of an input layer where data is firsts received, a hidden layer where computation happens and an output layer where a prediction is made [10]. A neural network receives its namesake because it is derived from the way a human neuron receives and processes information.

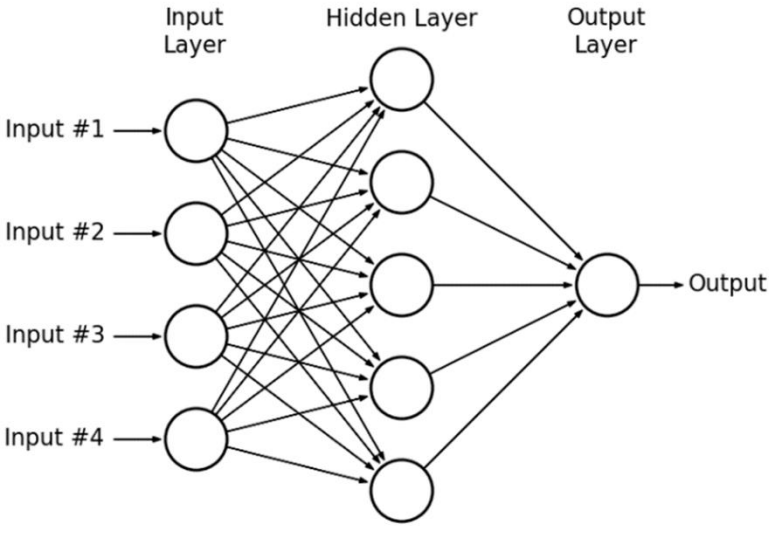


Fig 3. Basic Multilayer Perception

3.4 Stratified K-Fold Cross-Validation

Cross Validation is a generalized way of assessing model performance. Data is split into equal K groups or folds. The model will be evaluated K times where 1-fold will act as the test set, and the remainder will be used as a training set. A new test set and training sets are selected for each evaluation. The final performance of the model is the mean of K evaluations. Using stratified K-Fold ensures that the data is split in such a way that the proportions between classes are the same as they are in the entire dataset.

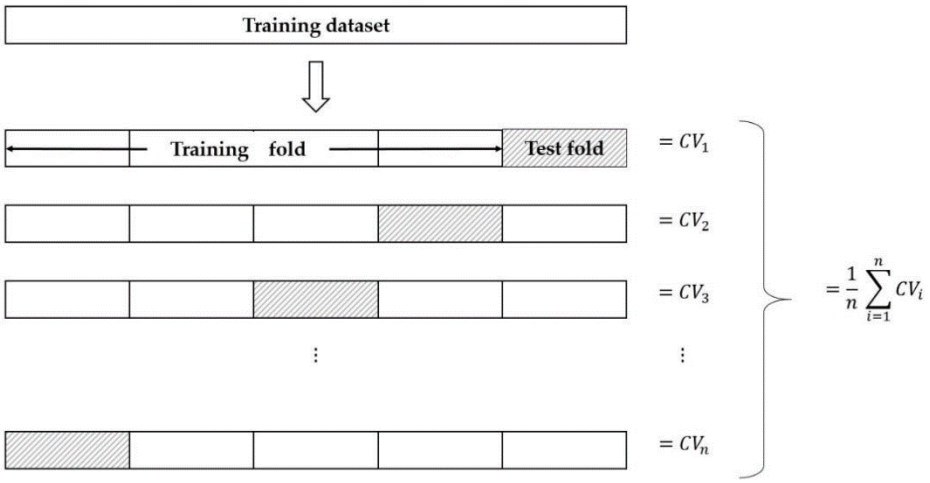


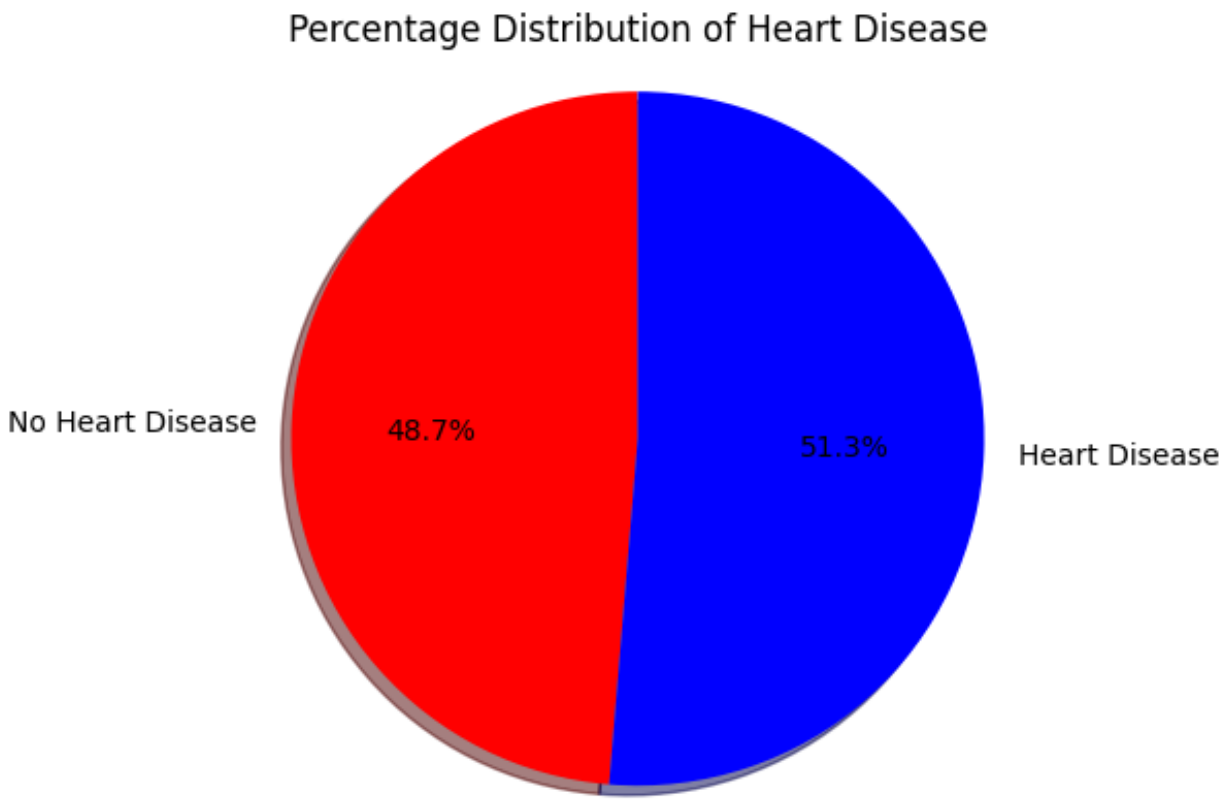
Figure 4. K-Fold Cross Validation [11]

4. Data Description

In this research, the data set applied a heart disease dataset pulled from Kaggle. This data set has been used in many studies before with varied outcomes. There are originally 76 attributes that was reduced to the 14 most important attributes for research purposes. Data is collected from the clinical records' of 1025 adults of various ages and states of health in Cleveland, Hungary, Switzerland, and Long Beach V.

Table 1: Summary of Heart Disease Dataset

Attribute Number	Attribute Description	Range	Mean	Standard Deviation
1	Age	29-77	54.43	9.07
2	Sex (M - 1/ F - 0)	0-1	0.70	0.46
3	Chest Pain Type	0-4	0.94	1.03
4	Resting Blood Pressure (mm/Hg)	94-200	131.61	17.52
5	Serum Cholesterol (mg/dl)	126-564	246.00	51.6
6	Fasting Blood Sugar > 120 mg/dl	0-1	0.15	0.36
7	Resting Electrocrardiographic results	0-2	0.53	0.53
8	Maximum Heart Rate	71-202	149.11	23.01
9	Exercise induced Angina	0-1	0.34	0.47
10	ST depression induced by exercise relative to rest	0-6.2	1.07	1.17
11	Slope of Peak Exercise ST	0-2	1.39	0.62
12	Majors Vessels Detected by Flourosopy	0-4	0.75	1.03
13	Thalassemia (0 - Normal, 1 - Fixed, 3 - reversible)	0-3	2.32	0.62



4.1 Data Visualization

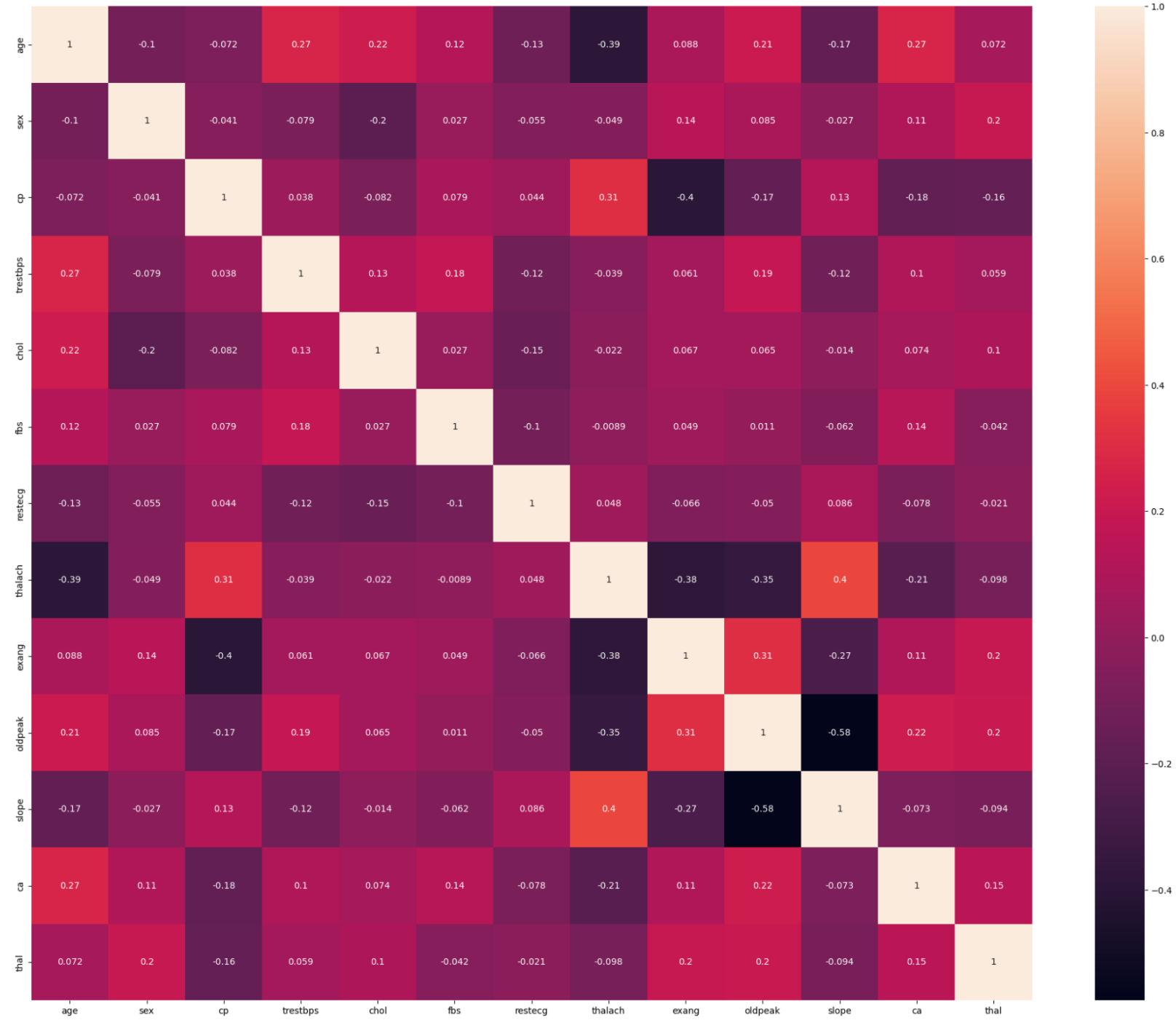


Fig 3. Indicators Of Heart Disease Correlation Heat Map

4.2 Results and comparison

To evaluate the performance of these models, accuracy, precision, recall, roc-auc, and f1 scoring are used on the test results. These measures can be given by the following equations, where TP is True Positive, TN is true negative, FP is false positive, and FN is false negative.

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

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$Roc\ Auc = \frac{\text{Total Positive Rate (recall)}}{\text{False Positive Rate}}$$

$$F1 = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

10-fold cross validation is used to estimate the test error in each iteration, 10 iterations are performed for each model test. Test results are given in Table 2.

Table 2: Test Results - Heart Disease Dataset

No.	Model	Accuracy Mean(Std.)	Precision Mean(Std.)	Recall Mean(Std.)	Roc_Auc Mean(Std.)	F1 Mean(Std.)
1	SVM	80.96%(0.06%)	76.37%(0.06%)	91.81%(0.05%)	89.52%(0.04%)	83.27%(0.05%)
2	**Gradient Booster**	**94.43%(0.03%)**	**93.14%(0.04%)**	**96.40%(0.03%)**	**98.27%(0.02%)**	**94.70%(0.03%)**
3	Random Forest	93.07%(0.03%)	90.95%(0.04%)	96.20%(0.04%)	97.92%(0.02%)	93.44%(0.03%)
4	Multi-Layer Perception	74.10%(0.08%)	79.78%(0.10%)	73.74%(0.24%)	91.00%(0.04%)	69.00%(0.14%)

Mean model fit times are given in table 3.

Table 3: Heart Disease Dataset - Model Run Time

No.	Model	Mean Time(s)(Std.)
1	SVM	0.04(0.03)
2	Gradient Booster	0.12(0.05)
3	Random Forest	0.16(0.04)
4	**Multi-Layer Perception**	**6.73(2.85)**

5. Conclusion and Future Studies

The results of this study show that data mining methods can be successfully applied to the medical world to help screen for heart disease in patients. In this study four machine learning models were used on one data set. The model that performed the best for this dataset was the gradient boosting classifier. Since we Know that the gradient booster works well, XGboost may work even better. Although this model performed well, it cannot be the sole tool used when diagnosing heart disease. Future studies will have to be done on different datasets relating to heart disease detection based to determine what model may be best based on the dimensionality and number of records in an input dataset. Future work on this topic base off this research is determining if tree-based models work as well on other type of disease prediction.

References

[1] Centers for Disease Control and Prevention. (2024, May 15). Heart disease facts. Centers for Disease Control and Prevention. https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html

[2] Cardiovascular Institute of the South. (2024, April 23). *Early screening: The key to heart disease prevention*. https://www.cardio.com/blog/early-screening-the-key-to-heart-disease-prevention/

[3] Mayo Foundation for Medical Education and Research. (2024, July 12). *Science Saturday: Artificial Intelligence won't replace your doctors, but it could make them better - mayo clinic news network*. Mayo Clinic. https://newsnetwork.mayoclinic.org/discussion/science-saturday-artificial-intelligence-wont-replace-your-doctors-but-it-could-make-them-better/

[4] Belcic, I., & Stryker, C. (2024, October 10). What is classification in machine learning?. IBM. https://www.ibm.com/think/topics/classification-machine-learning

[5] Donovan, R. (2023, July 27). Heart disease: Risk factors, prevention, and more. Healthline. https://www.healthline.com/health/heart-disease#diagnosis

[6] Ahmed, I. (2022, December). A study of heart disease diagnosis using ... https://scholarworks.lib.csusb.edu/cgi/viewcontent.cgi?article=2748&context=etd

[7] Cuello, F. (2024, August 5). Random Forest. TrendSpider Learning Center. https://trendspider.com/learning-center/random-forest/

[8] Dash, S. (2022, April 20). Gradient boosting - a concise introduction from scratch. Machine Learning Plus. https://www.machinelearningplus.com/machine-learning/gradient-boosting/

[9] GeeksforGeeks. (2023, October 12). *Multi-layer perceptron a supervised neural network model using Sklearn*. https://www.geeksforgeeks.org/multi-layer-perceptron-a-supervised-neural-network-model-using-sklearn/

[10] El-Sofany, H., Bouallegue, B., & El-Latif, Y. M. A. (2024, October 7). A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. Nature News. https://www.nature.com/articles/s41598-024-74656-2

[11] Lee, D.-H., Woo, S.-E., Jung, M.-W., & Heo, T.-Y. (2022, March 9). Evaluation of odor prediction model performance and variable importance according to various missing imputation methods. MDPI. https://www.mdpi.com/2076-3417/12/6/2826#

[12] Ibm. (2024a, August 13). What is support vector machine?. IBM. https://www.ibm.com/topics/support-vector-machine

[13] Lapp, D. (2019, June 6). Heart disease dataset. Kaggle. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset