**David Pang Daowei**

HEART DISEASE PREDICTIVE ANALYSIS   
WITH SUPERVISED MACHINE LEARNING

*Abstract*

All around the world, cardiovascular disease is the leading cause of death. Data mining can aid in the retrieval of useful information from health-care data. It aids in the training of a model to forecast patients' health, which is faster than clinical experimentation.

On Cleveland heart datasets, various implementations of machine learning algorithms such as Logistic Regression, K-Nearest Neighbor, Nave Bayes (NB), Support Vector Machine, Decision Tree, and Random Forest will be used, but it will be interesting to see which machine learning model is best for predicting potential heart disease patients. Although the original collection contained 76 variables, all published experiments only used a subset of 14, therefore this study used modeling to discover the relationship between the 14 relevant aspects of Cleveland heart data obtained from the UCI repository.

**The goal is to see how the classifier's performance is affected by attribute dependency. The kNN model generates a graphical representation of the attributes that is both trustworthy and transparent, as well as the capacity to forecast new scenarios. The model's accuracy is 89 percent, and the researchers determined that it had a strong possibility of detecting new patients.**

*Keywords:*Data mining, machine learning, artificial intelligence, k-Nearest Neighbor Classifier

# **INTRODUCTION**

The human heart has been dubbed the most important organ in the body since it is responsible for pumping blood throughout the body. Cardiovascular disorders including coronary artery disease (CAD) are, nonetheless, among the most frequent types of heart disease, impacting millions of people around the world. According to recent data, coronary heart disease is one of the leading causes of death, accounting for 15.6 percent of all deaths resulting from the disease [[1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8147740/#B1)]. Because heart diseases are linked to a number of modifiable risk factors related to lifestyle and action, the timing of diagnosis and diagnostic accuracy are extremely important in the therapeutic care of heart disease patients.

New scientific approaches, like as Machine Learning, have had a substantial impact in the detection and diagnosis of diseases over the last decade [[2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8147740/#B2)]. In general, the Machine Learning approach entails ‘training' an algorithm using a control dataset in which the disease status (disease or no disease) is known, and then applying this trained algorithm to a variable dataset to predict the disease status in patients for whom it is unknown. The Machine Learning system will be better trained as a predictor of disease status when more cohorts of data are included. More accurate illness prediction using machine learning would enable doctors to enhance patient identification, diagnosis, categorization, risk stratification, and management, all while potentially reducing the need for clinical intervention.

The ‘UCI Heart Disease Dataset,' also known as the ‘Cleveland Dataset,' is now freely available on the UCI Machine Learning Repository, making it one of the most widely used clinical datasets in the world [[3](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8147740/#B22)]. Since 2010, there have been over 300 online recordings analysing Machine Learning methods on the dataset, with the majority of studies examining one or more Machine Learning methods at a time for predictive analysis. Nonetheless, because the computer code has not been made publicly available, verifying claims in any of the articles regarding the accuracy of the algorithms will take a lot of time [[4](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8147740/#B23)].

For this paper, a comparative analysis approach was taken by applying six different Machine Learning algorithms using the UCI Cleveland dataset to predict heart disease outcomes from 14 features. As for the six Machine Learning algorithms used in this study, it was found that all six algorithms (logistic regression, decision tree, random forest, support-vector machine, k-nearest neighbour, and naïve bayes) perform well with an accuracy of greater than ∼80%, with the exception of naïve bayes and decision tree classifier. The paper considers accuracy, F1 score and the area under the curve–receiver operating (AUC-ROC) as performance metrics for comparative analysis amongst the Machine Learning models.

It's worth noting that a variety of risk factors have a role in the development of cardiovascular disease, some of which are controllable or adjustable [[5](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8147740/#B10)]. These include high blood pressure, high cholesterol, smoking, diabetes, lack of physical activity, unhealthy diet and stress. Age, sex (gender), family history, and race or ethnicity are all fixed risk factors. Traditional methods of predicting the prognosis of cardiovascular illness rely on these risk factors [[6](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8147740/#B11)]. However, these methods failed to identify a substantial number of people at risk of cardiovascular disease, and some people who aren't at risk were given unnecessary prophylactic treatment [[7](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8147740/#B12)]. As a result, Machine Learning algorithms will be able to summarize the impact of individual factors on the response variable, which will be referred to as "variables of importance" (from the model's perspective), assisting in the construction of correct models [[8](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8147740/#B14)]. In this present study, we also extract the variables of importance in one model in an effort to demonstrate the feasibility in the Machine Learning model.

The dataset and pre-processing section contains information about the data source, data variables, and data pre-processing methods. The next part contains succinct descriptions of the Machine Learning models used in this work, as well as the methodology used to evaluate the model outcomes. In the findings and discussion section, the outcomes of the modelling efforts are shown and debated. The final section explains the current Machine Learning efforts' limits and gives a workflow model framework for expanding our research in the future.

# **dataset preparation**

The dataset used in this study is downloaded from Kaggle.   
This dataset has 14 variables on 303 patients. Table 1 lists all the variables, type of the variables, and a brief description which illustrates the contained information.

Table 1: Attributes of UCI dataset

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| Age | Continuous | Patient age in years |
| Sex | Categorical | Patient gender (1 = male;   0 = female) |
| cp | Categorical | Chest pain (1 = typical angina;   2 = atypical angina;   3 = non-angina pain;   4 = no pain) |
| trestbps | Continuous | Resting blood pressure (in mmHg) on admission to the hospital |
| chol | Continuous | Serum cholesterol in mg/dl |
| fbs | Categorical | Fasting blood sugar higher than 120 mg/dl (1 = true;   0 = false) |
| restecg | Categorical | Resting electro-cardiogram ( 0 = normal;  1 = ST-T wave abnormality;  2 = probable/definite left ventricular hypertrophy) |
| thalach | Continuous | Maximum heart rate achieved (during thallium test) |
| exang | Categorical | Exercise induced angina ( 1 = yes; 0 = no) |
| oldpeak | Continuous | ST depression induced by exercise relative to rest |
| slope | Categorical | Slope of the peak exercise ST segment (1 = up-sloping;   2 = flat;   3 = down-sloping) |
| ca | Categorical | Number of major vessels (0 to 3) colored by fluoroscopy |
| thal | Categorical | Thallium heart scan ( 3 = normal;  6 = fixed defect;  7 = reversible defect) |
| num | Categorical | Diagnosis of heart disease (angiographic disease status)  (0 = absent; 1 to 4 = present) |

The following explains why some variables will be important in predicting heart diseases in patients from a medical perspective:

**Age**: Age is the most important risk factor in developing heart or cardiovascular diseases. It is estimated that 82 percent of people who die of coronary heart disease are 65 and older [[9](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2797320/)].

**Sex**: Men are at greater risk of heart disease than women before the age of 55. After menopause, it has been argued that a woman’s risk is similar to a man’s.

**Angina (cp)**: Angina is chest pain or discomfort caused when the heart muscle doesn’t get enough oxygen-rich blood. It feels like pressure or squeezing in the patient’s chest.

**Resting Blood Pressure (trestbps)**: High blood pressure can damage arteries that feed your heart over a period of time.

**Serum Cholesterol (chol)**: A high level of low-density lipoprotein cholesterol is most likely to narrow arteries, increasing the risk of a heart attack.

**Fasting Blood Sugar** **(fbs)**: Without sufficient insulin or the body not responding to insulin properly will cause the patient body’s blood sugar levels to rise, increasing the risk of a heart attack.

**Resting ECG** **(restecg)**: It has been determined that the potential risks of screening with a resting or exercise ECG equal or exceed the possible benefits for those at low risk of cardiovascular disease.

**Max heart rate achieved (thalach)**: It has been shown that an increase in heart rate by 10 beats per minute was associated with an increase in the risk of cardiac death by at least 20%. [[10](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3908776/)]

**Exercise induced angina (exang)**: The pain or discomfort associated with angina usually feels tight, gripping or squeezing, and can vary from mild to severe.

**Peak exercise ST segment (oldpeak)**: At a lesser workload or heart rate, the presence of horizontal or down-sloping ST-segment depression implies a higher risk of multi-vessel disease.

The correlation heat-map for the 14 variables in the dataset is shown in Figure 1, depicting the Pearson coefficients, or r, corresponding to the association between two variables, which can range from -1 to +1. In this analysis, we considered an absolute value of 0.5 as threshold, that is, if r is greater than 0.5 or less than -0.5, we assume those two variables are correlated. The above correlation matrix shows that only a few variables correlate with coefficients greater than 0.5, demonstrating poor correlation between the variables.

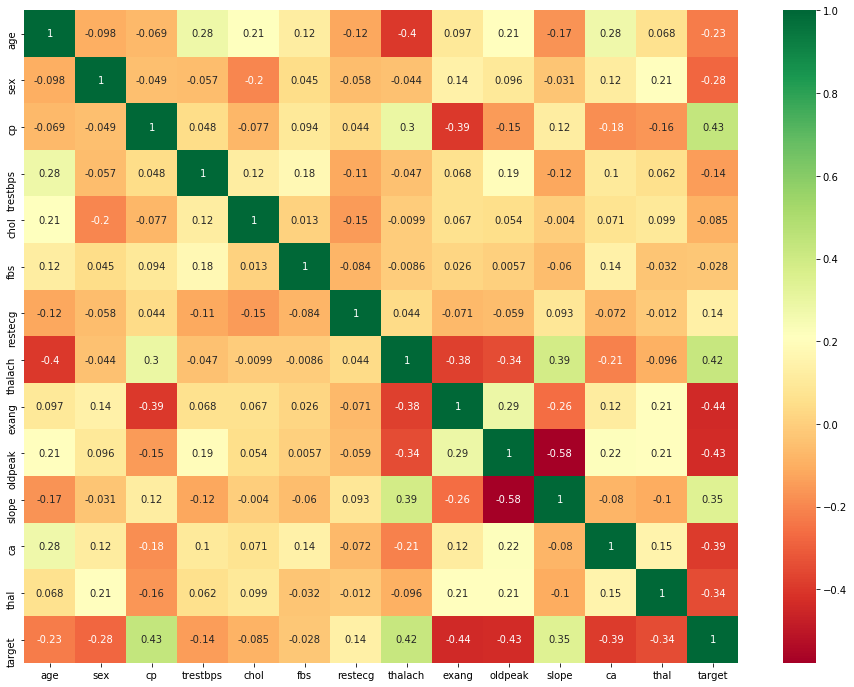


Figure 1: Variables Correlation Heat-map

In this analysis, the label “target” would be set as the dependent variable, in other words, the label refers to whether the patients have heart issues. Regarding the features, age and thalach (maximum heart rate achieved) have slightly strong negative correlations. It may account for that when people get older, their physical condition may undergo some declines along with the heart rate. Moreover, the correlation between thalach (maximum heart rate achieved) and oldpeak (ST depression induced by exercise relative to rest) is -0.34, it indicates that when people start to do more exercise, their heart rate will tend to increase and the ST depression will decrease.

Before proceeding to the modelling methodology, it will be important to encode the categorical data via one hot encoding, so to produce a new column (variable) for each categorical value. It will place a one in the column corresponding to the categorical value present for that observation. Similarly, the continuous data need to be standardized where each feature or variable is scaled to unit variance. It arranges the data in a standard normal distribution which is much more useful in classification problems.

# **MODELLING METHODOLOGY**

In supervised learning label data is given to the machine for prediction. The following will illustrate the general idea on how each of the six supervised machine learning algorithms work on the dataset and any assumptions we make in each case. Before applying the modelling algorithms onto the dataset, each algorithm is first trained (or fitted) with a fraction of the dataset, usually known as the ‘training set’ and then tested on the ‘test set'.

**Logistic regression**

In Logistic regression, instead of fitting a straight line or hyper plane, the logistic regression algorithm uses the logistic function to squeeze the output of a linear equation between 0 and 1. There are 13 independent variables which makes logistic regression good for classification. For the modelling, we left the threshold to the default value, which is 0.5, such that for probability values greater than 0.5, the model predicts the dependent variable to be 1 (the patient has heart issues) and for values less than or equal to 0.5, the model predicts the dependent variable to be 0 (the patient does not have heart issues).

**Decision tree**

A decision tree is a tree-like structure that sorts instances depending on the values of variables to classify them. In a decision tree, each node represents a variable, and each branch represents a possible value for the node. Starting at the root node, instances are categorized and then sorted based on the values of the variables. The root node of the tree is the variable that best divides the dataset. Internal nodes are the decision-making components that make a decision based on several algorithms and then proceed to the next node. The prediction of class label comes from the tree's root. Decision Tree is chosen because they are fast, reliable, easy to interpret and very little data preparation is required.

**Random forest**

Random forest is an ensemble model that resembles a forest in that it consists of many regression trees. Random forest builds a tree out of the data and uses it to generate predictions. It starts by combining several classification trees, training each on a slightly different set of dataset instances and separating nodes in each tree based on a small number of factors. The random forest's final predictions are calculated by averaging the predictions of each individual tree, which improves prediction accuracy for unknown data. Even when substantial sets of record values are missing, the Random Forest technique can yield the same result on large datasets.

**Support vector machine**

In support vector machine, each data point is plotted in an n-dimensional space, with the value of each variable being the value of specific coordinates, and classification is done using the hyperplane that distinguishes the two data classes. Following that, new instance characteristics can be utilized to determine which class a new instance should belong to.

**K-Nearest neighbor**

K-Nearest neighbor (kNN is one of the most basic and nonparametric algorithms available; it makes no assumptions about the underlying data distribution. The approach is based on the Euclidean distance principle, which states that examples in a dataset are often close to other instances with similar features. If each instance has a classification label, the value of an unclassified instance's label can be determined by looking at the class of its closest neighbors.

**Naive bayes**

The Bayes rule is the foundation of the Naive Bayes algorithm. The main and most crucial assumption in establishing a classification is that the dataset's properties are independent. It is simple and quick to predict, and it works best when the assumption of independence is maintained. The Bayes theorem estimates an event's posterior probability given its prior probability.

For each of the algorithms the performance is computed and analysed based on different metrics used such as accuracy, AUC-ROC and F1-measure scores as described further.

**Accuracy:**

It is straight-forward and the most common metric used, which refers to the proportion of total dataset instances that were correctly predicted out of the total instances.

**F1-score:**

F1 is an overall measure of a model’s accuracy that combines precision and recall, hence a good F1 score means low false positives and low false negatives.

**AUC-ROC:**

This is an estimate of the probability that a model ranks a randomly chosen positive instance higher than a randomly chosen negative instance. It measures how well predictions are ranked, rather than their absolute values.

# **RESULTS / DISCUSSION**

As shown in Table 2, 4 out of the 6 models have achieved an accuracy value higher than 0.80 with the exception of the decision tree model and the naïve bayes model. The other performance parameters, F1 score and AUC are also high. A mean value is calculated with all the latter two parameters for each of the models (shown in the last column in Table 2) to determine which model would perform the best. Instead of including the accuracy in calculating the mean, since it is often considered to be a misleading indicator in measuring the performance of models with biomedical datasets [[11](https://www.researchgate.net/publication/222674734_A_systematic_analysis_of_performance_measures_for_classification_tasks)], the analysis have used the accuracy metric as a supporting indicator.

Table 2: Models Performance Metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | F1 Score | AUC-ROC | Mean |
| Logistic  Regression | 0.836 | 0.853 | 0.838 | 0.846 |
| Decision  Tree | 0.705 | 0.719 | 0.706 | 0.713 |
| Random  Forest | 0.869 | 0.875 | 0.870 | 0.873 |
| Support Vector  Machine | 0.836 | 0.853 | 0.838 | 0.846 |
| K-Nearest  Neighbor | 0.885 | 0.892 | 0.887 | 0.890 |
| Naïve  Bayes | 0.508 | 0.500 | 0.500 | 0.500 |

Figure 2 and Figure 3 illustrate the confusion matrixes for the Decision Tree model and k-Nearest Neighbour model respectively. For the Decision Tree model (with poorer performance metrics), it has 23 true positive values and 22 true negative values. The accuracy is around 70% and misclassification is 30%. On the other hand, for the k-Nearest Neighbour model (with the best performance metrics), it has 29 true positive values and 25 true negative values. The accuracy is around 89% and the misclassification is around 11%, which means about 85% of the test data set has been correctly predicted by the model based on the training data set. All the models used tend to produce more false positives than false negatives.

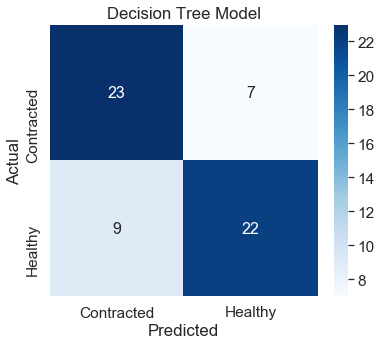


Figure 2: Decision Tree Model Confusion Matrix

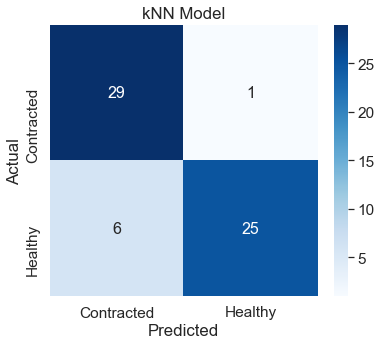


Figure 3: kNN Model Confusion Matrix

Looking at features importance, or in other words the importance of each variable used for modelling, will also illustrate the relative importance of variables and dictate which variables are necessary to be included in the prediction of heart diseases. However, feature importance is not defined for the kNN classification algorithm since kNN algorithm has the best performance metrics out of the 6 models used. The best way to estimate feature importance for kNN model is by taking the sample to predict on, and computing its distance from each of its nearest neighbours for each feature (call these neighb\_dist) and then continue the process repeatedly. Instead the analysis will use the second best model in terms of performance metrics, the Random Forest algorithm, to understand which feature ranks in importance.

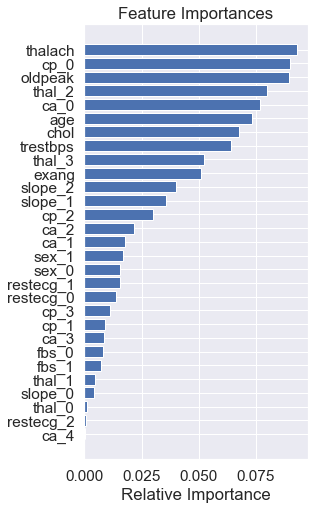


Figure 4: Random Forest Model Features Importance

The top variable playing significant role in predicting the dependent variable “Target” is “trestbps” (resting blood pressure). Usually we use blood pressure readings to analyze and monitor blood pressure. These tests record blood pressure using two measurements: systolic and diastolic blood pressure. Both high systolic and high diastolic blood pressure, according to the study, can contribute to heart attack and stroke [[12](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1124431/)]. The type of chest discomfort and the ST depression generated by exercise compared to rest are the next two important parameters that the Random Forest algorithm uses to forecast the presence of heart disease.

For this analysis, hyper-parameter tuning is also applied for the k-Nearest Neighbor algorithm via GridSearchCV, with the intention to select the best tuning parameters for the model algorithm and obtain the most optimized prediction results. However, the performance metrics are poorer, deriving <80% in terms of accuracy, F1 score and AUC-ROC score. There could be a couple of reasons for this, because GridSearchCV creates subsamples of the data repeatedly. That means the kNN model is trained on 80% of X\_train in each iteration and the results are the mean of predictions on the other 20%. The other cause could be due to conflating the questions of hyper-parameter tuning (selection) and model performance estimation.

# **CONCLUSION**

The study indicates that utilizing a publically available dataset, machine learning algorithms can detect the existence of disease or health conditions with high accuracy and other standard performance criteria. The study also shown that the kNN model outperforms existing Machine Learning-based heart disease detection methods. However, to improve the prediction, there are other models which could be used, such as XGBoost and neural networks. Many studies have been using neural networks for health condition prediction, such as the study by Shadab Hussain [[13](https://arxiv.org/ftp/arxiv/papers/2105/2105.10816.pdf)]. Nonetheless, this research will aid in the inference of heart illnesses, so functioning as a diagnostic tool to aid medical practitioners, and the models can be refined further.

References

*Papers presented:*

[1]

Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015

GBD 2015 Mortality and Causes of Death Collaborators

Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8147740/#B1>

[2]

Machine Learning for Detection and Diagnosis of Disease

Paul Sajda

Available: <https://www.annualreviews.org/doi/abs/10.1146/annurev.bioeng.8.061505.095802>

[3]

Machine learning, medical diagnosis, and biomedical engineering research – commentary

Kenneth R Foster, Robert Koprowski and Joseph D Skufca

Available:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4105825/>

[4]

Machine Learning in Medicine

Rahul C. Deo

Available:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5831252/>

[5]

Predicting the Future — Big Data, Machine Learning, and Clinical Medicine

Ziad Obermeyer and Ezekiel J. Emanuel

Available:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5070532/>

[6]

From Machine Learning to Artificial Intelligence Applications in Cardiac Care: Real-World Examples in Improving Imaging and Patient Access

David Tsay, Cam Patterson

Available:

<https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.118.031734>

[7]

Current applications of big data and machine learning in cardiology

Renato Cuocolo, Teresa Perillo, Eliana De Rosa, Lorenzo Ugga, and Mario Petretta

Available:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6748901/>

[8]

Machine Learning Techniques for Heart Disease Datasets: A Survey

Younas Khan, Usman Qamar, Nazish Yousaf, Aimal Khan

Available:

<https://dl.acm.org/doi/abs/10.1145/3318299.3318343>

[9]

The Burden of Cardiovascular Disease in the Elderly: Morbidity, Mortality, and Costs

Ali Yazdanyar and Anne B. Newman

Available:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2797320/>

[10]

Protective Role of Resting Heart Rate on All-Cause and Cardiovascular Disease Mortality

Arpit Saxena, Dawn Minton, Duck-chul Lee, Xuemei Sui, Raja Fayad, Carl J Lavie and Steven N. Blair

Available:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3908776/>

[11]

A systematic analysis of performance measures for classification tasks

Marina Sokolova

Available:

<https://www.researchgate.net/publication/222674734_A_systematic_analysis_of_performance_measures_for_classification_tasks>

[12]

Systolic blood pressure

Jan N Basile

Available:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1124431/>

[13]

Novel Deep Learning Architecture for Heart Disease Prediction using Convolutional Neural Network

Shadab Hussain, Susmith Barigidad, Shadab Akhtar, Md Suaib

Available:

<https://arxiv.org/ftp/arxiv/papers/2105/2105.10816.pdf>