Predicting Heart Disease Using Multiple Supervised Learning Methods

Hangcen Xie

School of Science, Hong Kong University of Science and Technology, Hong Kong, New Territories, Clear Water Bay, China
hxieam@connect.ust.hk

Abstract. Cardiovascular diseases (CVDs) are responsible for a significant number of deaths worldwide and are one of the leading causes of mortality. The main types of CVDs include coronary heart disease, rheumatic heart disease, and congenital heart disease. This article employs logistic regression to investigate whether individual features can predict heart disease. Following model training, the accuracy of each individual feature's prediction results is documented. The analysis reveals that most features alone cannot effectively predict heart disease. However, clinical features such as chest pain type, ST slope, and oldpeak (measured by ST depression) demonstrate relatively high accuracy. To explore how the combination of all features can predict heart disease, decision tree, random forest, XGBoost, and neural network models are utilized in this study. After model training, their performance is assessed and compared using various evaluation metrics including precision, recall, confusion matrices displayed in heat maps, ROC curves, and loss curves. The results indicate that the random forest model outperforms the others across all evaluation metrics, establishing it as the most effective model developed in this study for predicting heart disease.

Keywords: Heart Disease, Supervised Learning, Binary Classification, Model Performance Evaluation

1. Introduction

Cardiovascular diseases (CVDs) are responsible to a significant number or deaths worldwide and are one of the leading causes of mortality [1]. The coronary heart disease, rheumatic heart disease, congenital heart disease are main types of CVDs. Heart attacks and strokes cause more than 80% of CVD deaths of which one third are premature deaths for people under 70, according to the World Health Organization (WHO). There are a few leading behavioral risk factors of heart diseases and stroke that promote heart complications, including sedentary lifestyle, overuse of tobacco, excessive consumption of alcohols, and poor diet. Because of the behavioral risk factors, abnormal health indicators such as high blood pressure, high blood glucose, and high blood lipids that indicate high risks of CVDs may be observed. Premature deaths can be significantly reduced when those risks are identified and proper treatment [2].

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_23

The focus of this article will be one type of the CVDs, heart diseases. The heart diseases are causing large number of deaths around the globe, and it is causing serious social issues in less-developed countries. It is founded that approximately 26 million people globally are affected by heart diseases [3]. Besides, the global death rate is found to be around 32% by WHO [2]. Heart diseases claim the life of an individual in the United State every 34 seconds [4]; In India, $237 billion deaths from 2005 to 2015 is due to heart diseases as estimated by WHO [2]. Furthermore, it is pointed out by WHO that more than 75% of the deaths because of CVDs are from nations having low and middle income [2]. It is getting more and more difficult for less-developed countries like India and Bangladesh to provide affordable diagnoses and treatment because of the fast-growing population and heart disease cases. Because of the financial crisis of the government, it remains a question that whether most patient can receive equitable, appropriate, and adequate treatment [5]. To make things worse, a great many people in these less-developed countries cannot afford to have the opportunity for diagnosis of heart disease.

As early treatment can significantly reduce premature death if high risks are identified, early diagnosis of heart diseases can really save peoples' life and resolve social issues. Thus, predicting whether or not a person will have heart diseases using multiple health features is of great importance. However, making accurate predictions by only experience is difficult given the great number of risk factor. Consequently, supervised learning methods in machine learning can be of great help.

2. Related Work

With an increasing number of individuals focusing on the application of machine learning methods in the medical field, more of them identify the use of machine learning as an efficient and accurate way for early prediction for heart diseases [6]. The following are a series of studies that apply conventional machine learning methods in predicting heart diseases. Bashir et al. employed a few machine learning models including Naïve Bayes (NB), Decision Tree (DT), and Support Vector Machines (SVM) to construct an ensemble-based model that helps analyze and predict heart patients, achieving an accuracy of 87.37% [7]; A machine-learning-based myocardial infarction prediction based on J48 algorithms developed by Daraei and Hamidi achieve 82.57% accuracy[8]; Tomor and Agarwal acquired an accuracy of 85.59% using a least-squares twin support vector machine (LSTSVM) based on feature selection [9].

Aside from traditional machine learning models and their derivatives, neural networks are introduced to the medical domain because it can normally have better performance, model non-linear relationship more easily and handle larger and more complex dataset more efficiently. Convolutional Neural Network (CNN) is used by Dutta to help diagnose coronary heart disease, reaching 79.5% balanced accuracy [10]; An average accuracy of 89.25% is achieved by Li et al who created the CraftNet based on Deep Neural Network (DNN)[11]; In the work of Paul et al., maximum relevance minimum redundancy back propagation neural network is used and reached accuracy of 97.85%. In all, most of the examples of previous work fail to exceed 90% accuracy and the accuracy values crowd around 85%. Higher accuracy may be achieved by using a bigger dataset and better choice of machine learning models.

In the prediction task of this article, a well-documented dataset which is a combination of multiple existing datasets from Kaggle is used [12]. Given the target label, the heart disease label, is binary, the following machine learning methods are used. First, logistic regression is used to predict the heart disease using single features. Then, more complicated models that use various features are applied, including some tree-based models like random forest, and neural network. Finally, some visualizations like ROC curves and confusion matrices are used for displaying the performance of the trained models.

3. Methods

3.1. Logistic Regression

Although linear regression is really a popular regression model, logistic regression is used here as the result is binary. The equation behind logistic regression is.

$$\ln\left(\frac{\hat{Y}}{1-\hat{Y}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{1}$$

Xi represents the series of independent variables related to the outcome. Ŷ represents the probability of predicted outcome and lies in the range of 0 to 1 according to the definition of ln function. β0 is the intercept. The rest of the βi are the weights of each independent random variables Xi, indicating the slope of the regression line or say the change in Ŷ resulted by one unit change of Xi. In the algorithm, maximum likelihood estimation is used where the linear combination of independent random variables Xi with the best likelihood of correct prediction is found iteratively.

3.2. Decision Tree

The decision tree model is a popular tree-based supervised learning model for data classification tasks. A series of nodes make up the decision tree structure, branch nodes which is also known as internal nodes, and leaf nodes. The system of nodes is constructed to categorize the dataset input into various homogeneous subgroups. The original dataset will be collected into a root node where it is partitioned into a series of relatively more homogeneous subgroups named internal nodes according to certain thresholds. The partitioning process ends when it reaches the leaf node. The threshold of each node is chosen normally based on the information gain measured by entropy decrease of connected nodes. The threshold is chosen to maximize the information gain. The entropy is calculated according to the following equation [13].

$$Entropy(S) = -\sum_{c\in C} p(c)\log p(c) \tag{2}$$

The unit of entropy will be bit. S represents the dataset whose entropy is calculated. C stands for all classes in S. p(c) measures the ratio of data of class c in the dataset S.

Entropy can have value in range of 0 to 1. If all data belong to one class, the entropy will be 0.

3.3. Random Forest

Random forest is one of the most popular models for prediction and multi-classification. It is a kind of supervised ensemble learning algorithm that solve some issues of decision tree, including bias and over-fitting, leading to higher prediction accuracy. The bagging ensemble learning, and the random subspace method are the methods that helps achieve better performance for random forest model. The first step is to randomly create subsets according to the two methods: The process of bagging ensemble learning is essentially randomly picking certain number of features of the dataset to be trained in a decision tree model; The process of the random subspace method is essentially using bootstrapping to create a dataset of same size as the original dataset. The second step is to feed each new dataset generated by the previous random process to series of decision tree models. The third step is to yield the final result of the random forest model according to the results of individual decision trees. It is the two methods that make random forest algorithm less vulnerable to multicollinearity, and make it more stable dealing with missing and noisy data. The number of decision tree models and the number of features selected for training in each decision tree will influence the overall performance of a random forest model. Small number of decision tree models may cause low accuracy issue while large number of decision tree models will lead to long training time [14].

3.4. Extreme Gradient Boosting (XGBoost)

XGBoost is a supervised learning algorithm that is recognized for its flexibility and adaptability in addressing a wide range of regression and classification problems. To acquire higher accuracy and better generalization, it integrates several machine learning models. The classification and regression tree (CART) is used as the base classifier and several inter-related decision tree models are combined to make the final decision. As for the inter-related relationship among the decision tree models, the data used as the input in one decision tree model is the prediction result of the previous decision tree. Because of the complexity of the model, details will not be discussed here. Details can be found in the article published by Chen and Guestrin who originally proposed this great algorithm [15].

3.5. Neural Network

The neural network is a supervised learning model that imitates the collaborative function of neurons in the human brain to arrive at decisions. The neural network consists of several layers of neurons, comprising an input layer, a few hidden layers, and an output layer. The input layer has the same number of neurons as the input data's features, while the output layer has neurons equivalent to the number of potential output outcomes. Each neuron is assigned with certain weight that indicate the importance of its output and some threshold that determine whether the output will be activated and passed on to the next neuron. Activation functions are defined between layers of neurons to transform the previous layer's output into the current layer's input. One type of the activation functions, the sigmoid activation function which is used in this article will be introduced here [16]:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{3}$$

The value x denotes the output of previous layer and $\sigma(x)$ represents the input to the neurons in the current layer after transformation. The sigmoid activation function will generate an input in the range of 0 to 1 which makes it a prevalent for binary classification tasks.

After the input data pass through the network, the loss value will be calculated according to the loss function. The mean square error, binary cross-entropy loss, categorical cross-entropy loss, hinge loss, and Huber loss are the mainstream loss functions. Only the binary cross-entropy loss will be discussed here as the model in this article uses cross-entropy loss as its activation function [17].

$$\mathcal{L}_{BCE}(y, \hat{y}) = -(ylog(\hat{y}) + (1-y)\log(1-\hat{y})) \tag{4}$$

In the given function, y represents the actual value of the test set, while $\hat{y}$ represents the predicted value of the model. The binary cross-entropy loss calculates the disparity between the actual and predicted labels, which is used to impose a penalty on the network for inaccurate predictions for better performance.

The neural network model is trained for lots of iterations called epochs to improve its performance gradually. After one iteration of training, back propagation algorithm will be used to adjust the weights of neurons to improve model performance. Backpropagation functions by computing the gradient descent of the loss function, which is essentially the direction in which the loss function descends the most according to the neuron weights. This enables the model to adjust the neuron weights repeatedly during each iteration after each backpropagation, allowing it to enhance its performance.

4. Result

In the experiment part of this article, a well-documented dataset which is a combination of multiple existing datasets from Kaggle is used [12].

As for the evaluation metrics to be used, in the first part of the experiment, only accuracy is used as the evaluation metric. The second part involves using various evaluation metrics such as precision, recall, confusion matrix displayed in heat maps, ROC curve, and a loss value curve specifically for neural networks.

4.1. Single Feature Prediction using Logistic Regression

All the clinical features are used to predict the heart disease individually using logistic regression. The accuracy rate of each feature is listed in descending order in table 1.

Table 1. The accuracy rate of each feature

Feature	Accuracy
Chest Pain Type	0.804347826
ST Slope	0.798913043
Oldpeak	0.684782609
Exercise Angina	0.657608696
Max Heart Rate	0.652173913
Sex	0.641304348
Age	0.597826087
Fasting Blood Sugar	0.586956522
Resting ECG	0.581521739
Resting Blood Pressure	0.467391304
Serum Cholesterol	0.461956522

Regarding the fundamental settings during model training, 20% of the dataset is allocated to be the testing set and the random state is established to be 42.

It can be observed that some direct indicators of heart disease turn out to have the highest accuracy including chest pain type, ST slope, old peak which is measured by ST depression, exercise angina, and max heart rate. Possible reason may be their close relationship with heart activity. Other clinical indicators such as fasting blood sugar, resting blood pressure and serum cholesterol have poor performance in predicting heart disease. Possible explanation is that these clinical features can be affected by other diseases, thus, it cannot predict heart diseases individually, they may perform better when making prediction in combination.

Low accuracy for resting electrocardiogram (ECG) seems to contradict previous assumption as it is obviously indicator closely related to heart disease. However, it is reported that abnormality of resting ECGs may be caused by various factors such as lung diseases, thyroid disorder, and low levels of blood potassium and calcium [18].

4.2. Multiple Features Prediction using Multiple Algorithms

This section employs various models, including decision tree, random forest, XGBoost, and neural network, to predict heart disease by utilizing a combination of all 11 clinical features.

Some settings for training are as follows:

(1) Because all the models used does not expect categorical features as input, encoding is required before training the model. The encoding method used here is label encoding.

(2) For decision tree, random forest, and XGBoost, because of the complex non-linear relationship between the 11 clinical features, polynomial and interaction features are generated from the original features to improve model performance with the help of "PolynomialFeatures" function and the degree is set to be 2; However, because only shallow neural network is used here, to prevent issues of overfitting, original combination of features is used.

(3) 20% of the dataset is allocated to be the testing set, and the rest is allocated to be the training set.

(4) As for the random state settings after some optimization:

The random state of decision tree is set to be 43. Significant difference in model performance is observed for different random state as the order of data and the order of features being evaluated will affect the performance of decision trees.

However, although both random forest and XGBoost are tree-based algorithm, minor difference in model performance is observed for different random state. It is because of the bagging method of random forest used so that the random state will only affect the training set for each decision tree and will not have significant influence on the model as a whole. Similarly, for the XGBoost, the stochastic gradient boosting used will only affect the subset of dataset and the subset of features used in each iteration but will not affect the overall performance of the model. But for reproducibility, the random state is set to be 42 in both cases.

As for neural network, because of the larger number of hyperparameters of neural network, the settings will be discussed together.

4.3. The hyperparameter setting for neural network

The neural network constructed in this article comprises of two hidden layers. The input layer consists of 11 neurons. After optimization, the first hidden layer is set to contain 88 neurons, and the second hidden layer is set to have 85 neurons. The output layer of the neural network comprises of two neurons, representing the binary categorical label.

As for the activation function, sigmoid activation function are used for all neuron outputs because it is a binary classification task.

As for the learning rate, it is set to 0.01 after optimization. It determined the extent the weight of the network change according to the result of error during back propagation. Both large and small value of learning rate will lead to poor performance. A high learning rate can lead to the model converging too rapidly, causing the overshooting of the optimal outcome. Conversely, a low learning rate can cause the model to converge too slowly, resulting in extended training time and the model becoming stuck at a particular suboptimal point.

As for the epochs, the number of complete iterations of training the models, it is set to be 700 after optimization. Large value of epochs may cause overfitting issues while if the value of epochs is too small, the model cannot learn the pattern well which will cause underfitting issues.

As for the type of loss function used, cross-entropy loss is used which calculates the disparity between the probability distribution of the actual values and the probability distribution of the predicted values of the target label.

4.4. Result and discussion

After introducing the important settings, the result of the trained models are shown as follows:

Table 2 is the values that measures the accuracy of the model prediction:

Table 2. the accuracy of the model prediction

Model\Parameter	Accuracy	Precision	Recall	F1
Decision Tree	0.842391304	0.875	0.85046729	0.862559242
Random Forest	0.89673913	0.923076923	0.897196262	0.909952607
XGBoost	0.864130435	0.91	0.85046729	0.879227053
Neural Network	0.85326087	0.862745098	0.871287129	0.866995074

In the case of predicting heart disease, the models need to reduce the number of cases which mistakenly predict negative cases to be positive so that fewer people will receive unnecessary medical treatment and reduce the number of cases which mistakenly predict positive cases to be negative so that fewer people with potential heart disease will miss the opportunity for early treatment so that they can have longer time span. Thus, the important evaluation metrics for this case are precision and recall.

It can be observed that the value of precision is generally higher than the value of recall except for neural network. Thus, it can be concluded that for decision tree, random forest, and XGBoost, it is less likely for them to mistakenly predict people without heart disease to have heart disease than to predict people with heart disease to be healthy. And for the neural network, the case is opposite.

As for the value of precision, in descending order, there are random forest, XGBoost, decision tree, and neural network. As for the recall value, in descending order, there are random forest, neural network, and surprisingly decision tree and XGBoost having the same value.

In conclusion, the random forest model are best at both reducing number of people without heart diseases being predicted to have heart disease and the number of people with heart diseases being predicted to be healthy.

To further assist understanding of the performance of the four models, the confusion matrices plotted using heat maps of the four models are shown as in figure 1.

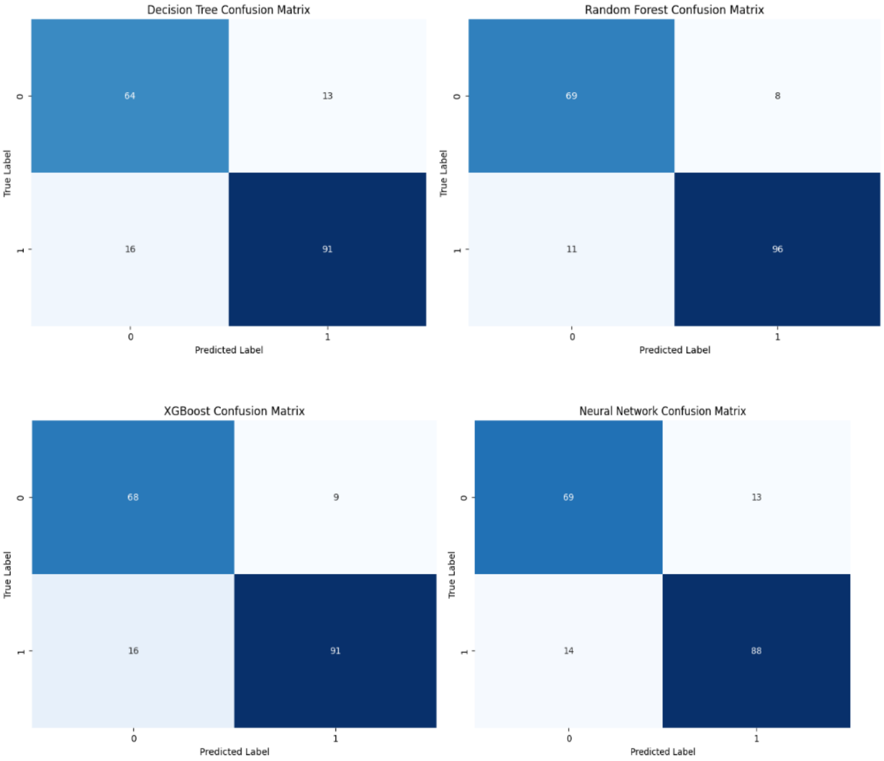

Figure 1 The confusion matrix of different methods

In addition to utilizing evaluation metrics such as precision and recall values, the ROC curves are plotted on the same graph and the area under the ROC curves (AUC) is then calculated, which are both frequently employed for evaluating the performance of binary classification tasks.

The ROC curves display the trade-off between the true positive rate and the false negative rate for different classification threshold which is a value set to determine the class label of a binary classification task.

The AUC values have a range of 0 to 1. A value of 1 signifies perfect performance; If the value is 0.5, it indicated the model is essentially normal guessing; The bigger the value indicates the better the performance.

The graph that displays all the ROC curves and all the AUC values of the four models is shown in figure 2.

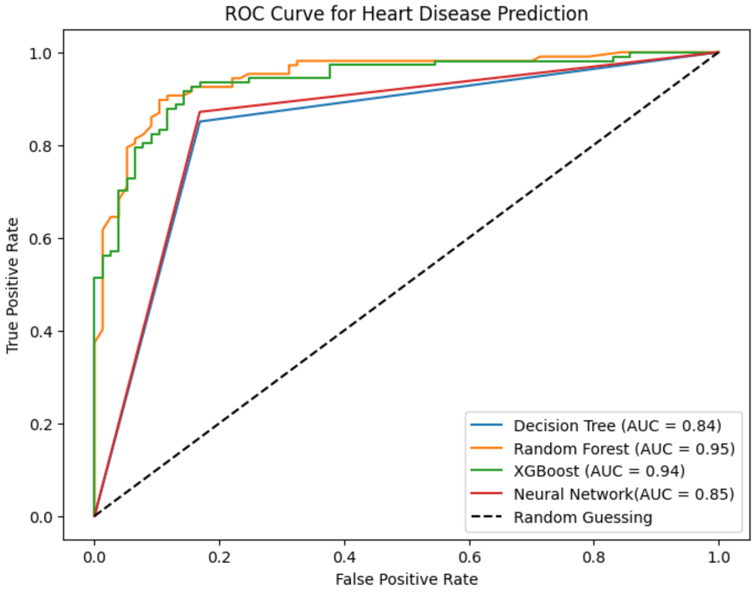

Figure 2 The result of ROC

It is noticeable that the random forest model still performs the best among the four models based on this evaluation metric. For most of the classification threshold, the ROC curve for random forest is above all of the rest of the three models. Furthermore, the AUC value of the random forest is also the highest among the three (Figure 2).

The behavior of ROC curves of decision tree and neural network which are like two joint straight lines and the relatively small value of AUC values indicate that the performance of the two models are relatively poor.

Because of the special learning pattern of neural network, another evaluation metric is introduced especially for neural network, the loss value curve. It is the curve that displays the loss value across the training process, and it is displayed in figure 3.

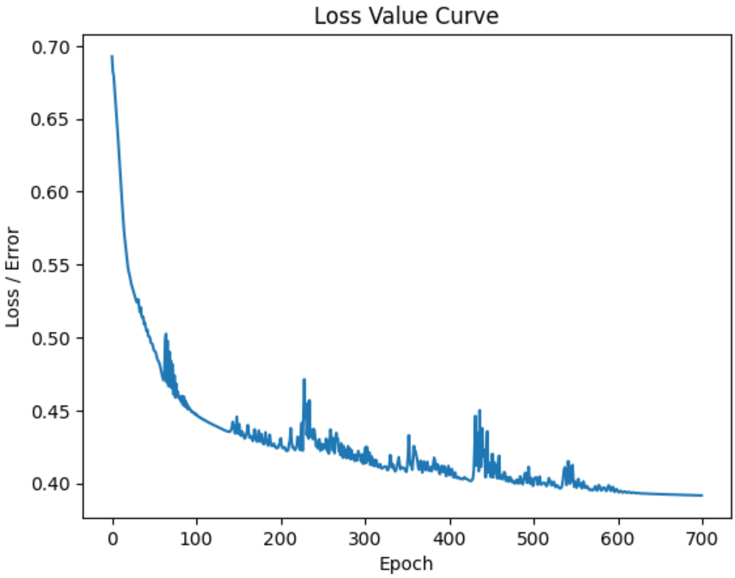

Figure 3 Loss curve of NN

It can be observed that the neural network finally converges as indicated by the loss values. The optimal value of learning rate is found by trial and error. Larger value of learning rate will lead to significant oscillation in the loss value curve indicating failure to converge, Smaller value of learning rate lead to lower accuracy, precision and recall. Simplicity of the neural network is one of the causes of the relatively large loss value after optimization. More complicated network should be used in the future to achieve better performance.

5. Conclusion

To study if single feature can predict heart disease, logistic regression is used in this article. After training the model, the accuracy of each single feature prediction result is listed. It turns out that most features cannot predict the heart disease well individually. However, clinical features including chest pain type, ST slope, and oldpeak which is measured by ST depression achieve relatively high accuracy.

To study how the combination of all the features can predict heart disease, decision tree, random forest, XGBoost, and neural network are used in this article. After the training of the models, their performance is measured and compared using a series of evaluation metrics such as precision, recall, confusion matrix displayed in heat maps, ROC curve, and loss value curve. The results indicate that the random forest model outperforms the others in terms of all evaluation metrics, signifying that it is the most effective model created in this article for predicting heart disease.

There are some limitations for the work done in this article so that future study can work on:

(1) A larger dataset should be used or constructed to achieve better performance for the models.

(2) To better compare the performance of single feature predicting heart disease using logistic regression, more evaluation metrics should be used. As mentioned before, in the task of predicting heart disease, the value of precision and recall is more of concern.

(3) The tuning method of neural network should also be improved for better optimization. The tuning process in this article uses the trial-and-error method. However, more advanced methods such as grid search and random search can also be employed.

References

1. Ahsan, M.M., Mahmud, M.A., Saha, P.K., Gupta, K.D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. Technologies, 9(3), 52.
2. World Health Organization. (2022, March). Cardiovascular diseases. Retrieved from https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1
3. Savarese, G., & Lund, L. H. (2017). Global public health burden of heart failure. Cardiac Failure Review, 3(1), 7-11.
4. Heidenreich, P. A., Trogdon, J. G., Khavjou, O. A., et al. (2011). Forecasting the future of cardiovascular disease in the United States. Circulation, 123(8), 933-944.
5. Islam, A. K. M. M., & Majumder, A. A. S. (2013). Coronary artery disease in Bangladesh: a review. Indian Heart Journal, 65(4), 424-435
6. Cao, J., et al. (2012). An intelligent scoring system and its application to cardiac arrest prediction. IEEE IEEE Transactions on Information Technology in Biomedicine, 16(6), 1324-1331.
7. Bashir, S., et al. (2016). A multicriteria weighted vote-based classifier ensemble for heart disease prediction. Computational Intelligence, 32(4), 615-645.
8. Daraei, A., & Hamidi, H. (2017). An efficient predictive model for myocardial infarction using cost-sensitive j48 model. Iranian Journal of Public Health, 46(5), 682.
9. Tomar, D., & Agarwal, S. (2014). Feature selection based least square twin support vector machine for diagnosis of heart disease. International Journal of Bio-Science and Bio-Technology, 6(2), 69-82.
10. Acton, S. T., et al. (2020). An efficient convolutional neural network for coronary heart disease prediction. Expert Systems with Applications, 159, 113408.
11. Gao, Y., et al. (2020). Craftnet: a deep learning ensemble to diagnose cardiovascular diseases. Biomedical Signal Processing and Control, 62, 102091.
12. Kaggle. (n.d.) Heart Failure Prediction Dataset. Retrieved from https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
13. Dastani, M., et al. (2024). Causal Entropy and Information Gain for Measuring Causal Control. Communications in Computer and Information Science, vol 1947.
14. Zhang, M., & Wang, M. (2009). Search for the smallest random forest. SI, 2(3), 381-381.
15. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

16. Chen, C., et al. (2021). A Review of Microfluidic Droplet-Based Assays for High-Throughput Screening of Single Cells. Micromachines, 12(10), 1183.
17. Rundo, L., et al. (2022). Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation. CMIG, 95, 102026.
18. Mayo Clinic. (2021, March 5). Electrocardiogram (ECG or EKG). https://www.mayoclinic.org/tests-procedures/ekg/about/pac-20384983.