Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/99/2025.17852

# *Research on the Application of Machine Learning in the Field of Heart Disease*

**Hongli Ding[1,a,*], Weilin Li[2]**

[1]*School of Mathematical Science, Beihang University, Beijing, 100191, China*
[2]*School of surveying and mapping engineering, Nanjing University of Posts and Telecommunications, Nanjing, 210003, China*
*a. s285dh@163.com*
**corresponding author*

***Abstract:*** Heart disease has long been a major threat to human health, causing about one-third of all deaths each year. Therefore, there is a need for accurate and effective prediction of cardiac disease using machine learning techniques. This paper uses data from 1,319 patients with heart disease and applies several machine-learning methods to study the relationship between a total of eight factors. The algorithms used in this paper include neural networks, decision trees, random forests, and so on. The best model was established by the random forest method, with an accuracy of 96.97%, precision of 97.03%, recall of 96.97%, and f1-Score of 0.9. It was found that troponin and CK-MB indicators had the highest influence weights, and the sum of the weights of each model accounted for more than 75% of the total weight, which demonstrated their significance in the prediction of heart disease, and the results could be utilized for the future prediction of heart disease. In addition, it also plays an essential role in the prediction of heart disease. To sum up, this is really important for the prediction of future heart disease.

***Keywords:*** Heart disease, prediction, neural network, decision tree, random forest.

## 1. Introduction

China's cardiovascular disease (CVD) prevalence is on the rise. There are estimated to be 330 million CVD sufferers worldwide, and more than 25.7% of them are related to heart disease. The financial toll that heart disease takes on the population and society continues to increase, and the inflexion point for prevention and treatment has not yet been reached [1].

The mechanism of heart disease is complex, with multiple genetic and environmental factors acting together. Scholars at home and abroad have found a correlation between heart disease and smoking and passive smoking, work habits, and family medical history [2-4]. In addition, there is a powerful relationship link cardiac conditions and indicators of some substances in the body, such as cholesterol, triglycerides and so on [5]. With the growing interest in the use of machine-assisted diagnosis trained by large databases in medical research and practice, the effective identification of patients with relevant diseases through machine learning can not only reduce the pressure on physicians but also improve the effectiveness of diagnosis.

Prediction of heart disease has been a global and popular research topic for experts and scholars, and various results have been achieved after decades of efforts [6]. Researchers globally have

DOI: 10.54254/2755-2721/99/2025.17852

employed models such as multilayer deep convolutional neural networks (MLDCNN) optimized by adaptive elephant herd methods (AEHOM), K-nearest neighbor (KNN), Naïve Bayes, and IoT Cloud Technology (IoT), among others [7-9]. The study of Halah et al. shows that the method demonstrates high accuracy and precision in training and testing, as well as improved rates of false positives and real positives, and an increased likelihood of making an accurate prediction [7]. The balanced F-scores indicate that the method preserves equilibrium between recall and precision. However, the study also points to the necessity of additional research and methodological improvement, especially with regard to issues with feature selection, data quality, and optimization methods. Similarly, the research of Albert, Rakesh and K Manoj has shown that Random Forest is a very effective machine learning algorithm that outperforms other methods in heart attack prediction [8]. The algorithm is able to identify high-risk subjects and provide healthcare providers with the data they need to make informed prevention and intervention decisions. The Random Forest model also has outstanding potential for handling high-dimensional data and missing values, which makes it useful in real-world healthcare applications. Additionally, Random Forest provides feature correlation scores that guide healthcare providers toward critical risk factors and prioritize treatments that may significantly reduce CVD mortality. Overall, the use of random forest algorithms enables better assessment and management of cardiovascular health and helps people develop appropriate treatment plans and behavioural changes.

However, most of the literature on the correlation of factors of heart disease is less researched and lacks complete systematic statistics. Therefore, this paper addresses 8 factors (Age, Gender, Heart rate, Systolic blood pressure, Diastolic blood pressure, Blood sugar, CK-MB, and Troponin) to investigate whether they have an effect on heart disease.

To address the shortcomings of a single model, this paper proposes to determine the influence level of each eigenvalue from algorithms such as Decision Tree and Random Forest, aiming at replacing the cumbersome and time-consuming clinical examination to identify heart disease patients with extremely efficient machine learning (ML) algorithms [10,11]. Results acquired by each model are presented in this study.

## 2. Methods and Data

### 2.1. Data

The heart attack dataset utilized in this study was gathered from January to May 2019 at Zheen Hospital in Erbil, Iraq. Its attributions include blood sugar, CK-MB, age, gender, heart rate, systolic and diastolic blood pressure, and troponin with positive or negative results.

The data used in this article contains 1319 instances and 8 variables without any missing values. Creating an algorithm to categorize individuals with and without heart disease by using all the available clinical characteristics including age, gender, heart rate, systolic and diastolic blood pressure, blood pressure, CK-MB and Troponin. All 8 variables are represented in the Table 1.

Table 1: Different types of variables

| Term | Type | Range |
|---|---|---|
| Age | Numeric | 14-103 |
| Gender | Categorical | 0-Female, 1-Male |
| Heart rate | Numeric | 20-135 |
| Systolic blood pressure | Numeric | 42-223 |
| Diastolic blood pressure | Numeric | 38-154 |
| Blood sugar | Numeric | 35-541 |
| CK-MB | Numeric | 0.321-300 |
| Troponin | Numeric | 0.01-10.3 |

DOI: 10.54254/2755-2721/99/2025.17852

## 2.2. Methods

### 2.2.1. Neural Network

The Neural network algorithms, usually referred to as backpropagation neural networks (BP neural networks) consist of two main processes: transmission of the signal forward and the error backwards. In forward propagation, the input signal is nonlinearly transformed by the hidden layer to generate a signal for output; if the expected and actual outputs are different, it will enter the error's backpropagation phase. The error is back propagated from the output layer to the input layer and is used to adjust the weights and thresholds of each layer to reduce the error. The network parameters are optimized through repeated training to achieve minimum error. Neuron, as The fundamental component of a neural network, are responsible for receiving multiple input signals from other neurons, calculating them through specific activation functions, generating output signals, and transferring these signals to other neurons in the network to realize the transmission and processing of information. The neuronal composition formula can be written as:

$$a = g\left(\sum_{i=1}^{n} w_i\, x_i + b\right) \tag{1}$$

According to the formula, a represents the neuron's output, $g(\cdot)$ indicates the function of activation, $w_i$ denotes the mass of the $i$ signal input, $x_i$ denotes the $i$ input signal, and b denotes the bias.

The neuron receives a plurality of input signals, $x_i$ each of which is multiplied by the corresponding weight $w_i$ with bias b. Their weighted sums are then fed into the activation function $g(\cdot)$ for a nonlinear transformation to generate the output of the neuron a. The output of the neuron can be passed on to other neurons or be used in the output of the neural network.

### 2.2.2. Decision tree

Decision trees are commonly used classification algorithms that consist of several internal nodes, leaf nodes, and a root node. The basic features include: Leaf nodes represent classification results and other nodes are used for attribute testing. The collection of samples in a node is assigned to child nodes according to the attribute test's findings. Overfitting can be dealt with by techniques such as pruning.

Decision trees are usually sensitive to changes in the data which may result in modifications to the tree's structure, although its simplicity is effective, the Random Forest algorithm can improve stability and performance by integrating multiple decision trees.

A schematic representation of the decision tree algorithm is displayed in Figure 1.

### 2.2.3. Random Forest

A classifier with several decision trees, Random Forest is an integrated method, which gives better performance and prevents overfitting in contrast to one decision tree. In prediction, every decision tree gives a classification outcome, and the category with the most votes is finally selected as the model's forecast. Random forest performs better than single decision trees and effectively prevents overfitting. The Random Forest algorithm schematic diagram is displayed in Figure 1:

DOI: 10.54254/2755-2721/99/2025.17852

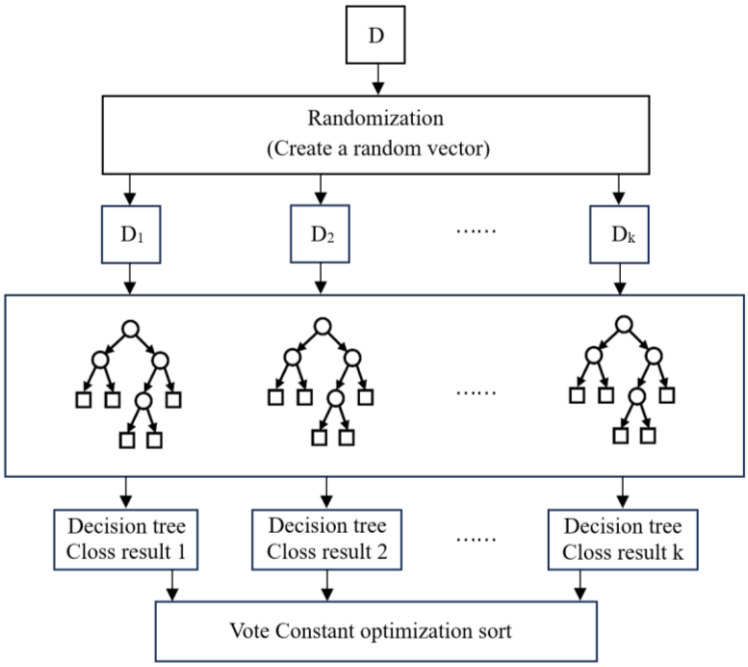


Figure 1: The schematic diagram of the Random Forest

In this paper, three main techniques for machine learning mentioned are to analyze the data. In the data processing analysis, independent variables in Table 1 are selected, while the result is selected as the variable that is reliant on neural network modelling, decision tree modelling and random forest modelling respectively.

## 3. Result And Discussion

### 3.1. Neural Network

Figure 2 below shows the performance of the model on both the training set and the test set, respectively. Positive indicates the subjects have heart disease, while negative shows they do not.

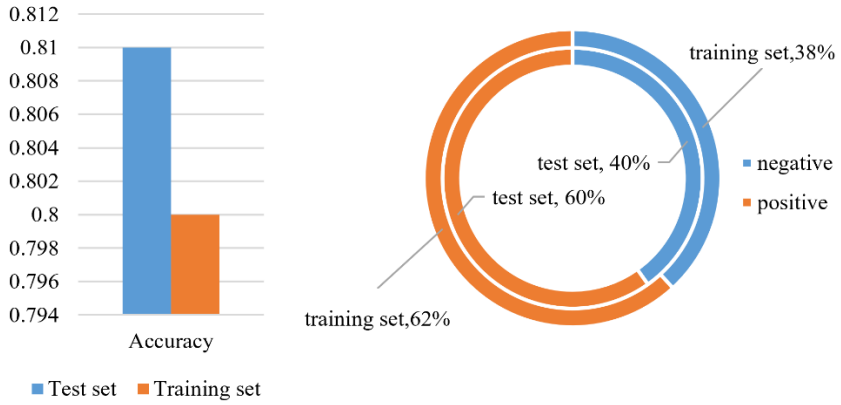


Figure 2: Neural network training effect

It can be seen from Figure 2 that the proportion of heart attack patients in the training set is 61.8%, and the proportion in the test set is 59.8%.

DOI: 10.54254/2755-2721/99/2025.17852

Table 2: Model summary table (Neural network)

| Term | Parameter Name | Value |
|---|---|---|
| Model parameter setting | Data preprocessing | mms |
| | Training set proportion | 0.8 |
| | Hidden layer neuron setting | (100) |
| | Activation function | relu |
| | Weight optimization method | adam |
| | L2 regularization coefficient | 1.0E-4 |
| | Initial learning rate | 0.001 |
| | Optimization method | constant |
| | Minibatch size | auto |
| | Maximum number of iterations | 200 |
| | Optimization tolerance | 1.0E-4 |
| Model evaluation effect | Accuracy | 80.303% |
| | Precision (comprehensive) | 80.371% |
| | Recall rate (comprehensive) | 80.303% |
| | f1-score | 0.803 |

According to Table 2, the accuracy of the final model on the test set is 80.30%, the precision(comprehensive) is 80.37%, the recall rate (comprehensive) is 80.30%, and the f1-score is 0.80. The model effect is acceptable.

## 3.2. Decision Tree

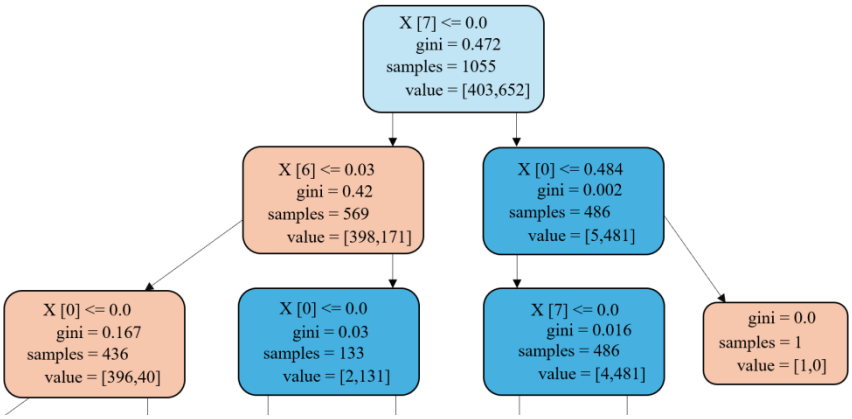


Figure 3: Part of the decision tree structure diagram

Figure 3 above shows part of the decision tree. The first line in each block is the name of the property used to split the node and the indicator. The number after *X* represents the index of the analysis item placed in *X* The second line, gini/entropy, represents an indicator of purity. The value indicates the number of samples of different categories.

Table 3: Feature weight value table (Decision tree)

| Term | Weighted value |
|---|---|
| Age | 0.009 |
| Gender | 0.028 |
| Heart rate | 0.009 |

DOI: 10.54254/2755-2721/99/2025.17852

Table 3: (continued).

| | |
|---|---|
| Systolic blood pressure | 0.006 |
| Diastolic blood pressure | 0.021 |
| Blood sugar | 0.037 |
| CK-MB | 0.354 |
| Troponin | 0.537 |

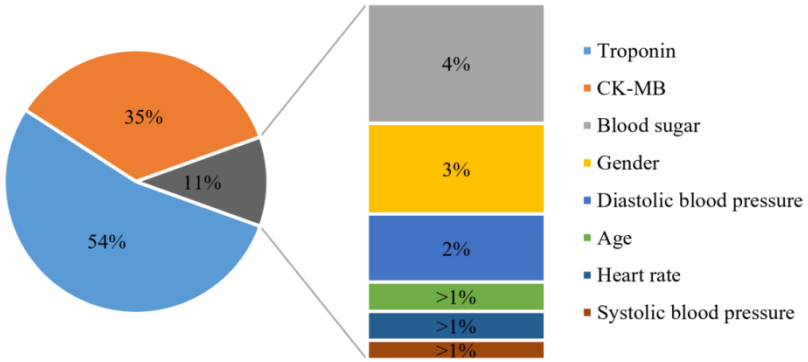


Figure 4: Feature weight value table (Decision tree)

Table 3 and Figure 4 show each term's contribution to the model. From the above table, it can be seen that Troponin accounts for 53.66%, has the highest weight and plays a crucial role in the pattern construction: CK-MB accounts for 35.36%, being the second importance, while other features have little influence on the decision making.

Table 4: Model summary table (Decision tree)

| Term | Parameter Name | Value |
|---|---|---|
| Model parameter setting | Data preprocessing | norm |
| | Training set Proportion | 0.8 |
| | Node splitting Standard | gini |
| | Node partitioning method | best |
| | Node splitting Minimum sample number | 2 |
| | Leaf node Minimum sample number | 1 |
| | Tree Maximum depth | unlimited |
| Model evaluation effect | Accuracy | 95.076% |
| | Precision (comprehensive) | 95.071% |
| | Recall rate (comprehensive) | 95.076% |
| | f1-score | 0.951 |

From Table 4, the accuracy of the final model on the test set is 95.08%, the precision (comprehensive) is 95.07%, the recall rate (comprehensive) is 95.08%, and the f1-score is 0.95, which means the model effect is better.

DOI: 10.54254/2755-2721/99/2025.17852

### 3.3. Random forest

Table 5: Feature weight value table (Random Forest)

| Term | Weighted value |
|---|---|
| Age | 0.048 |
| Gender | 0.029 |
| Heart rate | 0.038 |
| Systolic blood pressure | 0.033 |
| Diastolic blood pressure | 0.033 |
| Blood sugar | 0.036 |
| CK-MB | 0.251 |
| Troponin | 0.532 |

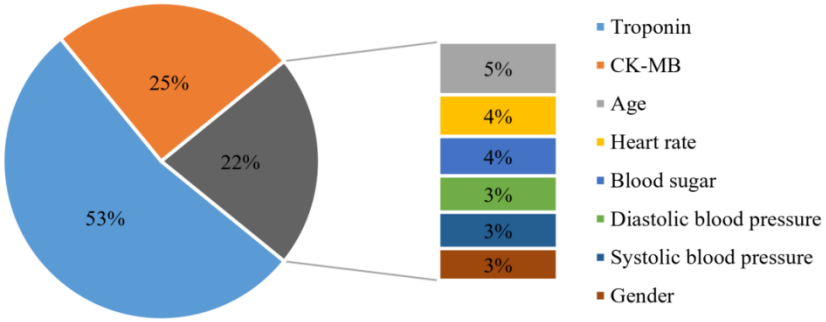


Figure 5: Feature weight value table (Random Forest)

From the above Table 5 and Figure 5, it can be seen that Troponin accounts for 53.21%, and CK-MB accounts for 25.06%. Other characteristics have less influence on decision making.

Table 6: Model summary table (Random Forest)

| Term | Parameter Name | Value |
|---|---|---|
| Model parameter setting | Data preprocessing | norm |
| | Training set Proportion | 0.8 |
| | Decision tree quantity | 100 |
| | Node splitting Standard | gini |
| | Node splitting Minimum sample number | 2 |
| | Leaf node Minimum sample number | 1 |
| | Tree Maximum depth | unlimited |
| | Limit of Maximum number of features | auto |
| | put back sampling | Yes |
| | perform out-of-pocket data testing | Yes |
| Model evaluation effect | Accuracy | 96.970% |
| | Precision (comprehensive) | 97.033% |
| | Recall rate (comprehensive) | 96.970% |
| | f1-score | 0.970 |

In Table 6, the accuracy of the final model on the test set is 96.97%, the precision (comprehensive) is 97.03%, the recall rate (comprehensive) is 96.97%, and the f1-Score is 0.97.

DOI: 10.54254/2755-2721/99/2025.17852

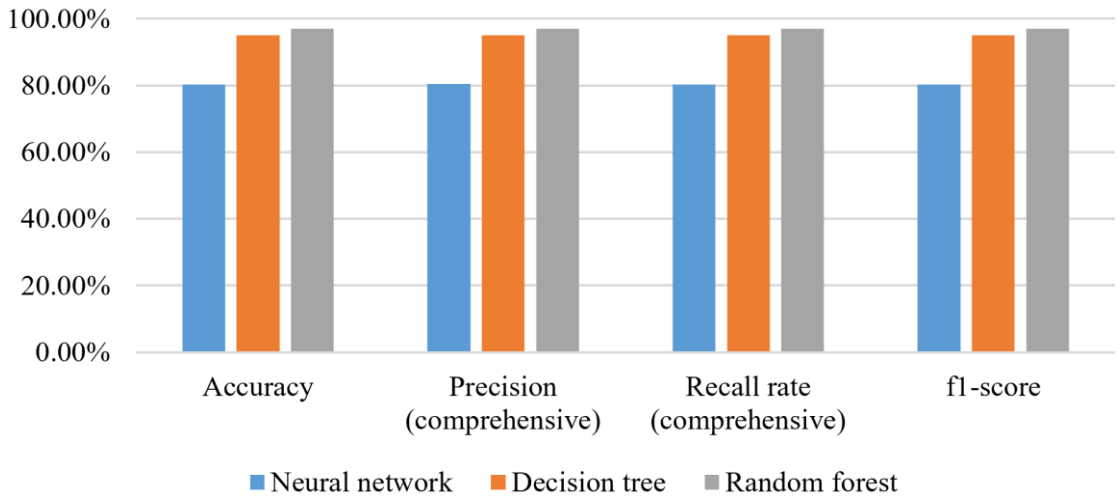


Figure 6: Comparison of different methods

The results above show that approaches based on tree models are typically better than neural networks as the comparison shown in Figure 6. While the model based on random forest performs the best.

All the methods essentially solve the problem by breaking it down step by step. By gradually segmenting the feature space along various characteristics, tree-based methods maximize information acquisition. Neural networks, on the other hand, enable each neuron to monitor particular input and feature space regions (with a variety of overlaps), after which specific neurons are triggered. Tree-based models take a deterministic approach, whereas neural networks take a probabilistic approach to this piece-by-piece model fitting.

Additionally, neural networks indirectly guide the activation of subsequent neurons by transforming inputs through fitted parameters. Decision trees, on the other hand, explicitly fit parameters to direct the flow of information. The findings in this paper on whether there is a risk of heart disease are clear, with only two outcomes: negative or positive. Each set of data contains multiple basic features, requiring minimal probability calculations. The issue is not overly complex, and the data volume is not very large, so the tree model approach outperforms neural networks in the data collected for this paper.

Comparing decision trees and random forests, random forest is a method based on multiple decision trees that divides the data into separate subsets and builds a decision tree on each subset to enhance the model's accuracy and stability. It can be seen as an integrated learning method based on multiple decision trees, and for this reason, random forests outperform decision trees in assessing the risk of heart disease.

## 4. Conclusion

This paper focuses on a machine learning method that can effectively predict heart disease and analyses the factors that influence the largest percentage of heart disease. In the research process, the most effective machine learning method was obtained by comparing the accuracy results of different machine learning methods, and the factors affecting heart disease were also analyzed as a percentage of the factors affecting heart disease, so as to arrive at the factors affecting heart disease the most.

Throughout the course of the research project, by examining data from the Random Forest machine learning model, this paper found that heart disease risk was most closely associated with troponin and creatine kinase-MB. These substances only become apparent after heart muscle damage and cell rupture, and creatine kinase is released from heart muscle cells into the bloodstream, which can only

DOI: 10.54254/2755-2721/99/2025.17852

be detected by elevated creatine kinase isoenzymes in blood tests. Regular medical check-ups are therefore essential to check the levels of these substances and to determine if there is a risk of heart disease.

This paper found that the accuracy rate of the decision tree model on the test set is 95.08%, the precision (comprehensive) is 95.07%, the recall rate (comprehensive) is 95.08%, and the f1-score is 0.95. The best model was established by the random forest method, with an accuracy of 96.97%, precision of 97.03%, recall rate of 96.97%, and f1-Score of 0.97.

The domain of predicting the risk of the heart is an essential direction of research in machine learning these years. In terms of classification of methods, machine learning algorithms such as neural networks, decision trees, and random forests have been widely used in heart disease risk prediction studies with good accuracy results. The research methodology explored in this paper for the application of machine learning in the field of heart disease risk prediction is just a new idea, and it is expected that in the future, technical methods for heart disease risk prediction can be researched with higher accuracy and better results.

## Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

## References

[1] *Liu, M., He, X., Yang, X., et al. (2024). Interpretation of the key points of the China Cardiovascular Health and Disease Report 2023. Chinese Family Medicine, 1–19. Retrieved August 30, 2024.*

[2] *Dunstan, D. W., Thorp, A. A., & Healy, G. N. (2011). Prolonged sitting: Is it a distinct coronary heart disease risk factor? Current Opinion in Cardiology, 26, 412–419.*

[3] *Pohjola-Sintonen, S., Rissanen, A., Liskola, P., & Luomanmäki, K. (1998). Family history as a risk factor of coronary heart disease in patients under 60 years of age. European Heart Journal, 19, 235–239.*

[4] *Doi, T., Langsted, A., & Nordestgaard, B. G. (2022). Elevated remnant cholesterol reclassifies risk of ischemic heart disease and myocardial infarction. Journal of the American College of Cardiology, 79(24).*

[5] *Yang, J., & Luo, Y. (2024). Research on the application of machine learning in heart disease risk prediction. Fujian Computer, 40(08), 12–16.*

[6] *Alshaikh, H. A., et al. (2024). Comprehensive evaluation and performance analysis of machine learning in heart disease prediction. Scientific Reports, 14(1), 7819–7819.*

[7] *Stonier, A. A., Gorantla, R. K., & Manoj, K. (2023). Cardiac disease risk prediction using machine learning algorithms. Healthcare Technology Letters, 11(4), 213–217.*

[8] *Patro, S. P., & Padhy, N. (2023). A secure IoT-cloud-based remote health monitoring for heart disease prediction using machine learning and deep learning techniques. Engineering Proceedings, 56(1), 241.*

[9] *Mineni, H. (2020). Second heart attack prediction using machine learning and deep learning (Doctoral dissertation). Dublin Business School.*

[10] *Yang, Y., & Li, R. (2024). Influence factors of heart disease based on decision tree-logistic regression model. Industrial Control Computer, 37(08), 114–116.*

[11] *Singh, A., et al. (2024). Heart disease detection using machine learning models. Procedia Computer Science, 235, 937–947.*