Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/46/20241090

Prediction of cardiovascular disease based on machine learning

Jiesi Li

Academy of Cybersecurity in the Cyberspace, Chengdu University of Information Technology, Chengdu, China

moralesvok85230@student.napavalley.edu

Abstract. Cardiovascular disease is one of the deadliest diseases worldwide, causing millions of deaths every year. Major risk factors include hypertension, hyperlipidemia, smoking, unhealthy diet, and lack of physical activity. To achieve a simple and effective prediction of cardiovascular disease, a study comparing the performance of common machine learning algorithms was conducted. The dataset used in this research consists of a population of 70,000 individuals from Kaggle. During the data processing phase, abnormal values within the feature variables were removed, and a BMI feature variable was added to the dataset to visualize the relationships between the data more intuitively. Deep neural networks were used to predict cardiovascular disease and were compared with eight traditional machine learning algorithms with respect to accuracy, F1 score, PR and ROC. The results indicated that the deep neural network (DNN) is the optimal model for predicting cardiovascular disease.

Keywords: Cardiovascular Disease Prediction, Machine Learning Techniques, Deep Neural Networks (DNNs).

1. Introduction

Cardiovascular disease is one of the most prevalent, deadly, and disabling diseases in the world today. According to data from the World Health Organization, millions of people die from cardiovascular disease each year, including heart disease and stroke. Despite significant advancements in medical science in the treatment and prevention of cardiovascular disease, challenges remain. Early prediction and accurate diagnosis of cardiovascular disease are crucial for improving patients' quality of life, reducing healthcare costs, and enhancing preventive measures. Predicting cardiovascular disease requires consideration of multiple risk factors, including age, gender, blood pressure, lipid levels, blood sugar, and family medical history. Early prediction and intervention can assist doctors in identifying high-risk patients earlier and implementing appropriate measures such as medication, lifestyle changes, and regular monitoring. However, traditional statistical methods often struggle to handle large-scale and multidimensional patient data, lacking the required accuracy and personalized predictive capability. This highlights the importance of machine learning in cardiovascular disease prediction.

Machine learning algorithms can leverage extensive clinical data and biomarkers from patients to discover hidden patterns and regularities, facilitating accurate risk prediction for cardiovascular disease. Machine learning-based cardiovascular disease prediction not only provides personalized medical advice and intervention measures but also aids doctors in better assessing patients' risk levels and

DOI: 10.54254/2755-2721/46/20241090

formulating treatment plans. Personalized medical advice can select the most suitable treatment methods and medications based on patients' specific conditions, thus avoiding overtreatment or ineffective treatment and improving treatment effectiveness and patients' quality of life. Additionally, accurate prediction of cardiovascular disease trends can assist public health departments and decision-makers in developing effective prevention and control strategies. By predicting and monitoring the prevalence trends of cardiovascular disease, potential risk factors can be identified and addressed early on, leading to the establishment of relevant public health policies and prevention measures. This helps reduce the incidence and mortality rates of cardiovascular disease while alleviating the burdens on healthcare systems and the economy. The advantages of machine learning lie in its ability to handle large-scale and complex data, mining valuable information from it. Through the analysis of extensive patient data, machine learning algorithms can identify hidden risk factors and anomalous patterns, resulting in more accurate risk prediction for cardiovascular disease. This provides doctors with additional evidence and guidance, enabling them to make more informed decisions, while also offering patients more personalized and precise medical care [1].

Despite the enormous potential of machine learning in cardiovascular disease prediction, there are still challenges and obstacles to overcome. Pooja et al [2]. proposed using the random forest algorithm to construct a model that combines data mining with machine learning models for the purpose of determine the probability of the outcome variable. However, the limitation of this study is the lack of implementation of more advanced and integrated models that could yield higher accuracy than existing models. F. Fahim et al [3]. achieved more accurate prediction of CVD in patients by integrating machine learning methods, including support vector machines. Nonetheless, their study was constrained by the dataset being related to heart sounds, which is not easily obtainable through simple measurements. In their study, LIAQAT ALI et al. [4] introduced a novel framework comprising of two methodologies: the X2 statistics model and the deep neural network. The X2 statistics model is employed for function refinement, while the DNN handles the classification task. The Cleveland dataset, consisting of 303 instances, was employed for evaluation purposes. Remarkably, this proposed framework outperformed previous traditional artificial neural network (ANN) models, leading to a noteworthy classification accuracy of 93.33% using the DNN approach. However, the dataset instances were limited. Therefore, this study adopts a larger and more readily available dataset to establish ensemble learning models and deep neural networks [5]. for predicting cardiovascular disease.

2. Methods of diagnosis

2.1. Database

This experiment utilizes a dataset curated by Kuzak Dempsey on the perils of cardiovascular diseases in adults. The dataset comprises a total of 70,000 instances, with 45,530 males and 24,470 females. It encompasses twelve feature variables, namely: label, identification number, age, gender, height, weight, diastolic pressure, systolic pressure, glucose level, cholesterol level, alcohol consumption status, smoking status, and physical activity level (Table 1).

Table 1. Database feature variables.

Field	Variable name abbreviation	variable	Specific measurement content	type
1	age	Age	Person's age	Int
2	gender	Gender	Person's gender	Str
3	height	Height	Person's height	Int
4	weight	Weight	Person's weight	Int
5	ap_hi	Systolic Pressure	The systolic pressure reading	Int
6	ap_lo	Diastolic Pressure	The diastolic pressure reading	Int

DOI: 10.54254/2755-2721/46/20241090

Table 1. (continued).

7	cholesterol	Cholesterol	One's individual cholesterol levels.	Int
8	gluc	Glucose	One's individual glucose level	Int
9	smoke	Smoke	Personal smoking situation	Boolean
10	alco	Drinking	Personal drinking consumption status.	Boolean
11	active	Active activities	Individual's level of physical activity	Boolean
12	BMI	Body Mass Index	Personal body mass index	Int
13	cardio	Cardiovascular disease		Boolean

2.2. Data preprocessing

In order to enhance the accuracy of the model's predictions, it is advised to remove data rows corresponding to diastolic blood pressure below 40 or above 199, as well as systolic blood pressure below 60 or above 299, as outliers. Additionally, considering relevant studies have indicated a clear correlation between body mass index (BMI) and cardiovascular diseases [6], this experiment augments the dataset with BMI as a new feature variable, leveraging existing data on height and weight and incorporating the following formula:

$$BMI = weight / (height)^2 \quad (1)$$

Herein, weight is measured in kilograms and height in meters.

By evaluating the relationship between the feature variables and cardiovascular diseases, it is observed that age, cholesterol, body mass index, diastolic blood pressure, systolic blood pressure, and glucose content hold significant importance in predicting cardiovascular diseases. This suggests that these features exert a substantial influence on the prediction of such diseases (Figure 1). In feature selection and engineering, incorporating these important variables into the model can significantly enhance predictive performance [7].

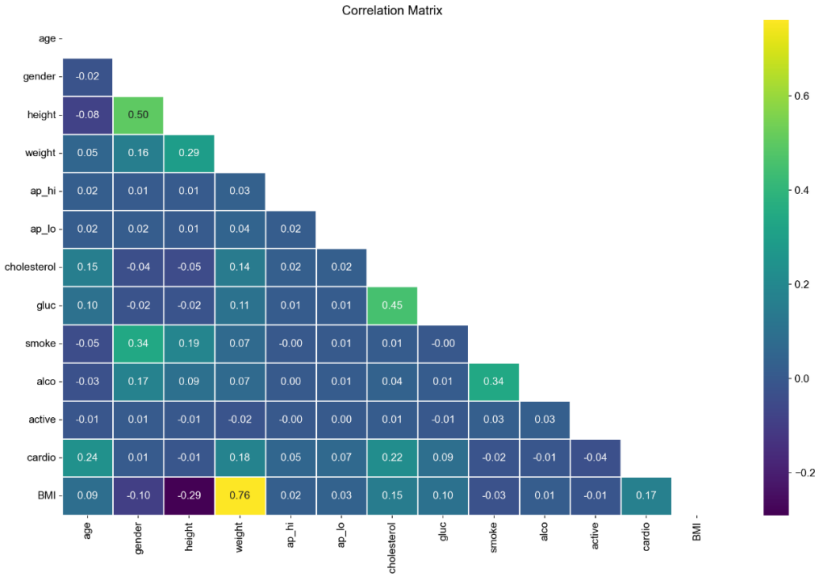

Figure 1. Correlation Matrix.

DOI: 10.54254/2755-2721/46/20241090

2.3. Building a neural network of profound depth

Compared to traditional machine learning models [8], fully connected neural networks can learn and capture complex nonlinear relationships between input features. They can extract more useful information through appropriate feature combinations and representation learning, automatically learning different levels of feature representation without the need for manual feature definition and selection. In this experimental study of CDV, the hidden layers of the fully connected neural network are set to two layers, with each layer consisting of 14 neurons. The activation functions *f* used are ReLU function [9]. and Sigmoid function [10]. The figure 2 is the fully connected neural network structure is presented below:

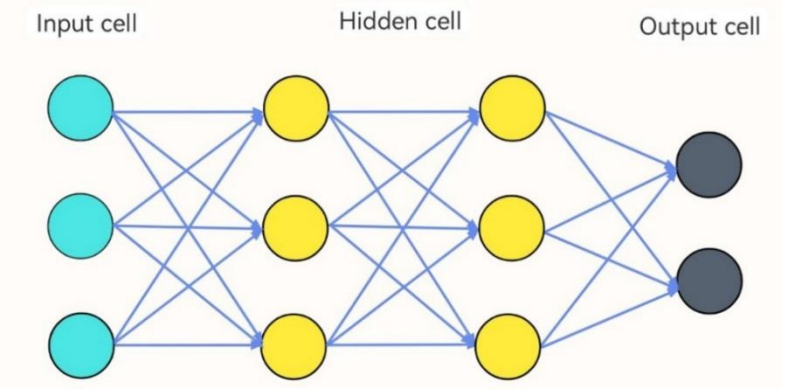

Figure 2. The schematic diagram of the fully connected neural network.

In the realm of deep neural networks, the activation functions play a pivotal role within both the hidden and output layers. In the context of predicting cardiovascular diseases, the output layer serves as a binary classification task. This study aims to map the network's output to a probability value between 0 and 1, indicating the likelihood of a given sample belonging to the positive category (indicating the presence of cardiovascular disease). To achieve this objective, this experiment has chosen to employ the Sigmoid function as the activation function for the output layer. The Sigmoid function adeptly maps real numbers to the interval of (0, 1), thus interpreting the output as a probability value.

Within the hidden layers, it is significant to consider the non-linear expressive power and the avoidance of the vanishing gradient problem. To address this, the experiment has opted for the Rectified Linear Unit as the activation function for the hidden layers. The ReLU function exhibits linearity for inputs greater than 0, while non-linearity is introduced for inputs less than or equal to 0. This design not only imbues the network with non-linear expressive capabilities but also mitigates the risk of overfitting by setting negative input values to zero. Given the potential complexity involved in extracting and learning features for cardiovascular disease prediction, appropriate non-linear capabilities are of utmost importance to the model's performance. The linearity of the ReLU function for positive inputs aids in the propagation of information between different layers and facilitates the learning of high-level features. Additionally, the simplicity and efficiency of the ReLU function's computation contribute to increased training speed of the network.

For the entire hidden layer or output layer, the linear transformation can be represented in matrix form as follows:

$$Z = W * x^T + b \tag{2}$$

Here, Z represents the result of the linear transformation, W is the weight matrix, X is the input feature matrix, and b is the bias vector.

Additionally, eight machine learning models have been established, namely Gaussian Naive Bayes, Logistic Regression, Decision Trees, Random Forests, Extremely Randomized Trees, Xgboost, LightGBM, and Support Vector Classifier, and compared them with deep neural networks.

DOI: 10.54254/2755-2721/46/20241090

3. The outcomes and analysis of the experiment

3.1. Metrics for experimental evaluation

In this study, we are confronted with a binary classification problem, aiming to predict whether an individual is afflicted with cardiovascular disease (CVD). To assess the performance of the classification model we have devised, it is imperative to employ several essential evaluation metrics.

Accuracy is a pivotal measure, denoting the proportion of correctly classified samples to the total number of predictions made by the model. Higher accuracy implies greater precision in model classification.

Recall, another significant performance evaluation metric, informs us of the model's ability to correctly identify positive instances out of all true positives. Since missing cases may incur greater consequences than false positives in disease prediction, recall is a crucial consideration.

The F1 score, a commonly used comprehensive evaluation metric, incorporates both precision and recall. It aids us in holistically evaluating the model's accuracy and coverage. When there is a need to weigh these two factors together, the F1 score proves to be a valuable measure.

When dealing with imbalanced datasets, where the ratio of positive to negative samples is severely skewed, the use of metrics such as accuracy and recall may be limited. In such circumstances, we can employ Average Precision (AP) to evaluate the model's performance. AP more effectively reflects the model's performance on imbalanced datasets.

Furthermore, when we need to compare the performance of multiple models, the Area Under the ROC Curve (AUC) is a commonly used metric. AUC provides an intuitive measure to assess the accuracy and discriminative ability of the model. By comparing AUC values, we can swiftly determine which model is more effective for a given task and dataset.

In summary, accuracy, recall, F1 score, AP, and AUC are crucial metrics for evaluating the performance of classification models. When conducting model evaluation and comparison, it is essential to consider these measures comprehensively in order to select the most suitable model for our specific problem.

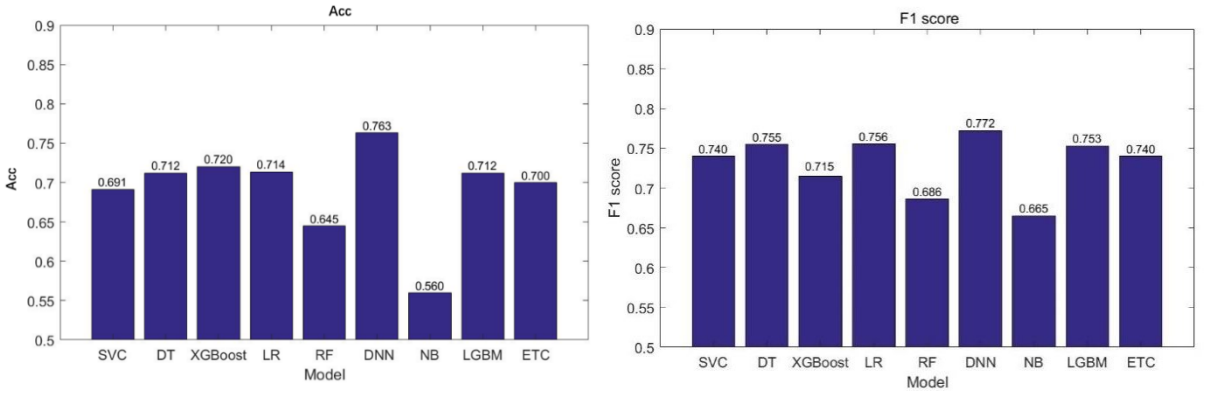

Figure 3. The results of Acc and F1score.

According to Figure 3, it is evident that in terms of the accuracy metric, the Deep Neural Network (DNN) surpasses the other eight machine models significantly. As for the F1 score, the Deep Neural Network outperforms the rest, ranking first with a substantial advantage of 0.016 over the second-ranked Logistic Regression model.

DOI: 10.54254/2755-2721/46/20241090

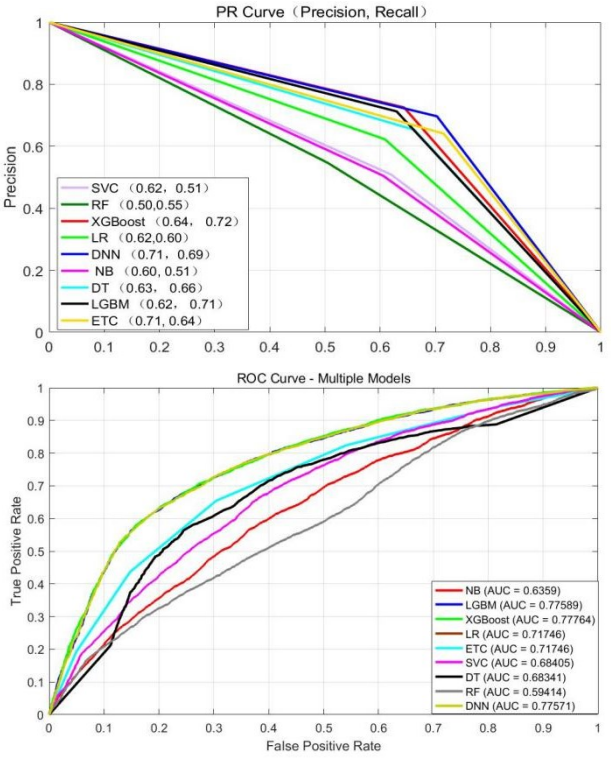

Figure 4. PR and Roc curve.

Based on Figure 4, it is evident that the deep neural network representation model performs well with regard to accuracy and recall in the PR metric. This implies that the model can accurately identify individuals with cardiovascular diseases and cover the actual affected population more comprehensively. In the ROC metric, the deep neural network, ranked third with a marginally higher value of 0.00193 than XGBoost, showcases strong discriminatory ability between individuals with cardiovascular diseases and non-affected individuals.

3.2. Analysis of Experimental Results

This experiment is based on a dataset provided by Kaggle. During the data processing phase, feature variables with outliers were removed, and a BMI feature was added to better illustrate the correlations between the data. The performance of a cardiovascular disease detection model, constructed using a deep neural network (DNN) and eight traditional machine learning algorithms, was evaluated using four metrics: accuracy, F1 score, PR and ROC

The results show that the deep neural network (DNN) is the optimal model for predicting cardiovascular diseases. Through a comparative analysis of performance metrics, we found that the DNN model excels in terms of accuracy, F1 score, PR and ROC. This finding holds significant implications for cardiovascular disease prediction. Future research can further investigate and optimize the application potential of deep neural networks in the healthcare field.

Accurate and convenient prediction of CVD has always been a challenging focal point. Despite the significant achievements of deep neural networks (DNNs) in predicting cardiovascular diseases, there is room for further improving model performance. Future research can explore more complex neural network structures, incorporate more feature variables, or utilize data from other domains to enhance the accuracy and stability of the model. Additionally, besides traditional clinical data, the prediction of cardiovascular diseases can also consider the fusion of multi-modal data, such as genomic data and imaging data (e.g. cardiac CT scan images). Future research can investigate how to integrate these different types of data to improve the overall capability of prediction models.

DOI: 10.54254/2755-2721/46/20241090

4. Conclusion

This experiment is based on a dataset provided by Kaggle. During the data processing phase, feature variables with outliers were removed, and a BMI feature was added to better demonstrate the correlations between the data. A comparison was made between a DNN and eight traditional machine learning algorithms in predicting cardiovascular diseases. The performance of the models was evaluated using four metrics: accuracy, F1 score, AP and AUC.

The results showed that the DNN performed the best among all the cardiovascular disease prediction models. Through the analysis of performance metrics, it was found that the DNN model exhibited outstanding performance regarding accuracy, F1 score, AP and AUC. This finding holds great significance for cardiovascular disease prediction. Future research can further explore and optimize the potential application of deep neural networks in the field of health. Accurate and convenient prediction of CVD has always been a challenging issue. Despite the significant achievements made by deep neural networks (DNNs) in predicting cardiovascular diseases, there is still room for improving the performance of the model. Future research can explore more complex neural network structures, incorporate more feature variables, or utilize data from other domains to improve the accuracy and stability of the model. In addition to traditional clinical data, the prediction of cardiovascular diseases can also consider the integration of multimodal data, such as genomic data and image data (e.g., cardiac CT scan images). Future research can explore how to combine these different types of data to improve the comprehensive predictive ability of the model.

References

[1] Desai, F., Chowdhury, D., Kaur, R., Peeters, M., Arya, R. C., Wander, G. S., Gill, S. S., Buyya, R. (2022). HealthCloud: A system for monitoring health status of heart patients using machine learning and cloud computing. *Internet of Things*, **17**, 100485.

[2] Anbuselvan, P., 2020. Heart disease prediction using machine learning techniques. *Int. J. Eng. Res. Technol,* **9**,515-518.

[3] F. Fahim, M. T. Ahmed, M. N. M. Shuvo and M. R. Islam, A Comparison between Different Kernels of Support Vector Machine to Predict Cardiovascular Diseases Using Phonocardiogram Signal, *2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies,* Bhilai, India, 2022, 1-4.

[4] L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed and J. A. Khan, An Automated Diagnostic System for Heart Disease Prediction Based on $\chi 2$ Statistical Model and Optimally Configured Deep Neural Network, in *IEEE Access*, **7**, 34938-34945, 2019,

[5] Jürgen Schmidhuber (2015). Deep learning in neural networks: An overview. *Neural Networks*, **61**, 85-117.

[6] Susanna C Larsson and others, Body mass index and body composition in relation to 14 cardiovascular conditions in UK Biobank: a Mendelian randomization study, *European Heart Journal*, **41, 2**, 221–226.

[7] A. M.A. and P. A. Thomas, Comparative Review of Feature Selection and Classification modeling, International *Conference on Advances in Computing, Communication and Control*, Mumbai, India, 2019, 1-9.

[8] A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, Prediction of Heart Disease Using Machine Learning, *2018 Second International Conference on Electronics, Communication and Aerospace Technology*, 2018, 1275-1278,

[9] J. Si, S. L. Harris and E. Yfantis, A Dynamic ReLU on Neural Network, *2018 IEEE 13th Dallas Circuits and Systems Conference*, Dallas, 2018, 1-6.

[10] R. Pogiri, S. Ari and K. K. Mahapatra, Design and FPGA Implementation of the LUT based Sigmoid Function for DNN Applications, *2022 IEEE International Symposium on Smart Electronic Systems*, Warangal, India, 2022, 410-413