ISSN 2959-6157

Application of Machine Learning in the field of Heart Disease Prediction and its Accuracy Study

Qiantong Gao

Guangzhou Tianxing Experimental School, Guangzhou, Guangdong province, China
*Corresponding author: 1807040124@stu.hrbust.edu.cn

Abstract:

The application of machine learning in the healthcare sector is a prominent area of research presently. Among these, there is significant potential in employing machine learning for predicting and treating heart disease. Hence, this study aims to explore the accuracy and application of machine learning models in predictive diagnosis and treatment of heart disease. This paper first presents commonly used machine learning algorithms within relevant fields, followed by gathering data and instances showcasing the use of machine learning in predicting and treating heart disease. Through data analysis, this study summarizes current cutting-edge development directions, trends, as well as the integration of machine learning algorithms into real medical processes. Finally, it concludes by summarizing the existing prospects and challenges faced by machine learning in predicting and treating heart disease. This paper can serve as a valuable source of inspiration and reference for researchers involved in related fields concerning heart disease prediction and treatment.

Keywords: Machine learning, Heart disease prediction and diagnoses, Prediction algorithms.

1. Introduction

With the advent of industrialization, urbanization, population aging, and shifts in people's lifestyles and living conditions, cardiovascular disease (CVD) has emerged as a leading global cause of mortality, exerting significant impacts on both human health and socioeconomics. Over 500 million individuals worldwide continue to be affected by heart disease, resulting in approximately 20.5 million deaths in 2021 - accounting for nearly one-third of all global fatalities. This represents a notable increase from the previously estimated figure of 121 million CVD-related deaths [1]. Therefore, it is crucial to accurately predict heart disease in a timely manner and ensure prompt diagnosis and medical intervention to effectively reduce mortality rates.

With medical technology's continuous advancement and data science development, machine learning techniques show great potential in heart disease prediction and diagnosis. Compared to conventional medical diagnostic methods, the utilization of machine learning and data processing techniques can effectively address the issue of limited availability of high-quality medical resources. This includes enhancing the productivity of doctors in larger hospitals and enabling primary care physicians to provide intermediate-level diagnosis and treatment services, thereby facilitating the implementation of a "hierarchical diagnosis and treatment" approach. Standard machine learning algorithms for heart disease prediction and diagnosis include, but are not limited to, the following: logistic regression, decision trees, random forests, K-nearest neighbors, support vector machines, Extreme Gradient Boost, neural networks, and so on. application: K-Nearest Neighbor, Logistic Regression, Random Forest, SVM and Extreme Gradient Boost [2]. Shadman Nashif et al. applied and tested Naïve Bayes, Artificial Neural Network, SVM, Logistic Regression and Random Forest five standard models and practiced them in real-time cardiovascular health monitoring system [3]. Yingying Du proposed a non-parametric improved logistic regression model based on Nadaraya-Watson estimation and compared the predicted ROC curves of the generalized logistic regression model with those of the generalized logistic regression model [4].

However, in practical applications, utilizing real-world data for disease prediction and medical treatment still faces some technical challenges, such as poor data quality, high data dimensionality, and data imbalance. Therefore, this thesis aims to explore the current status and future

trends of the application of machine learning techniques in heart disease prediction. By studying and analyzing the practical application examples of machine learning algorithms in cardiovascular disease prediction, this paper will focus on the following aspects of the discussion. First text will introduce the basic principles of machine learning in the medical field and several common machine learning algorithms in heart disease medical diagnosis, analyze their advantages and disadvantages as well as their specific applications in heart disease prediction. Secondly, it will analyze the advantages and challenges of machine learning in heart disease healthcare from the aspects of prediction, diagnosis, and treatment, respectively. Machine learning techniques can handle large-scale medical data, uncover potential features and patterns, and improve the accuracy and precision of prediction. However, issues such as data privacy, model specification, and changes in clinical practice are topics faced by machine learning in healthcare. Finally, the future trends of machine learning in the field of heart disease prediction are envisioned. With the continuous improvement of data collection technology and algorithm optimization, the application of machine learning in heart disease prediction will be more promising. At the same time, we will strengthen the cooperation of the academic system to promote the combination of machine learning technology and medical practice, and promote the development of heart disease prediction refinement and personalization. This paper aims to promote the prediction and treatment of heart disease to provide new ways of thinking and methods, and make positive contributions to improving human health.

2. Machine learning and common machine learning models:

2.1 Machine Learning

Machine Learning (ML) is an important branch of Artificial Intelligence (AI) that enables computers to learn from data and make predictions or decisions. The principles and methodologies of Machine Learning span a number of subject areas, including but not limited to statistics, probability theory, computational complexity theory, cybernetics, information theory, philosophy, physiology, and neurobiology. This interdisciplinary knowledge provides a solid foundation for the development of machine learning. From an engineering perspective, the core of machine learning revolves around three main steps: the formulation of hypotheses (models), the collection of data to test the hypotheses (validation of the model), and the refinement (iteration) of the hypotheses (models) [5]. This process involves a variety of algorithms, of which gradient descent is an important iterative method used to optimize the model parameters [5] continuously. There are various approaches to machine learning, and different choices lead to a range of machine learning methods that can be combined and adapted to the needs of specific problems in practical applications [5].

The fundamentals of machine learning involve the ability to learn and recognize patterns from data automatically, and this ability is based on a series of trained algorithms that can automatically learn from data and make predictions. Machine learning can be categorized into various types such as supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning, each of which has its own specific application scenarios and algorithm design.

The application of machine learning in disease prediction has become an important research direction in the field of healthcare. In particular, for the prediction of common diseases such as cardiovascular disease, diabetes, and kidney disease, machine learning algorithms such as decision trees, random forests, and support vector machines have been shown to have good performance. These algorithms can discover useful patterns from large amounts of medical data and help doctors diagnose diseases earlier, thus improving the survival rate of patients. Developing deep learning techniques has opened up new possibilities for disease prediction. The application of deep learning models in disease risk prediction has shown that such models are able to handle complex nonlinear relationships and improve the accuracy of disease prediction.

2.2 Common Machine Learning Models

The common machine learning algorithms applied to predicting and diagnosing heart disease are Logistic Regression, Support Vector Machine, K- Nearest Neighbor, Random Forest and Neural Network.

2.2.1 Logistic Regression

Logistic Regression (Logistic Regression) is a widely used statistical method for classification problems, especially prominent in binary classification problems. It predicts the likelihood of an event by converting the output of a linear regression model to a probability value, which is usually realized by a Sigmoid function. Logistic regression centers on modeling the relationship between one or more independent variables (predictor variables) and a dependent variable (outcome variable) to estimate the probability of an event occurring given the independent variables.

A key advantage of logistic regression is its simplicity and ease of interpretation, which makes it the classification

tool of choice in many situations. However, logistic regression also suffers from some limitations, such as being prone to overfitting and being sensitive to outliers, and it may not be effective enough in dealing with nonlinear relationships. To overcome these limitations, researchers have proposed a variety of improvement methods, including introducing penalty terms for regularization, using different loss functions, employing integrated learning methods, and combining other machine learning algorithms.

2.2.2 Support Vector Machine

Support Vector Machine (SVM) is a machine learning method developed on the basis of statistical learning theory to solve classification and regression problems. It achieves effective classification or regression of data by finding an optimal hyperplane, which is especially suitable for dealing with small-sample learning problems in high-dimensional spaces. The core idea of SVM is to maximize the edges, i.e., to find a hyperplane such that that hyperplane separates data of different categories and the distance between the two edges is as large as possible.

The basic principle of SVM consists of mapping the original data into a higher dimensional space to find a valid hyperplane for classification in this new space. This process involves the application of the kernel trick, which allows SVMs to handle nonlinear problems. The kernel trick simplifies the computational process and improves the algorithm's efficiency by calculating the inner product between data points to avoid direct mapping to higher dimensional spaces.

Although SVM has many advantages, such as globally optimal solution, simple structure, and strong generalization ability. However, it also has some limitations, such as inefficient processing of large-scale datasets, particularly sensitive to noise and isolated points. Therefore, researchers have been exploring new methods and techniques to overcome these limitations and further improve SVM's performance and application scope.

2.2.3 K-Nearest Neighbors

The KNN (K-Nearest Neighbor) algorithm is a classification and regression method widely used in the field of data mining and machine learning. The basic idea is that for a given test sample, K nearest neighbors is calculated based on their distance from all samples in the training set, and then the categories of the test sample are predicted based on the category information of these K neighbors. The advantages of the KNN algorithm lie in its simplicity and ease of implementation, the fact that it does not require assumptions on the data distribution and that it is insensitive to outliers. However, the KNN algorithm also has some problems and challenges, such as high computational complexity, inapplicability to large datasets, and sensitivity to parameter selection (especially the selection of K-values).

2.2.4 Random Forests

Random Forests (RF) is an integrated learning method that improves the accuracy and generalization of a model by constructing multiple decision trees and combining their predictions. Each decision tree is trained on randomly selected samples from the original dataset, which reduces the correlation between models and thus improves the stability and accuracy of the overall model. The basic idea of Random Forest is to use the self-help resampling (bootstrapping) method to draw multiple samples from the original samples, model the decision tree for each self-help sample, and then combine the predictions of multiple decision trees to arrive at the final prediction by voting or other means. This method can handle not only classification problems but also regression problems and is very tolerant to outliers and noise and is not prone to overfitting.

Although it performs well in many cases, just like any machine learning model, it has its limitations and drawbacks, such as reduced processing efficiency for large datasets, inability to adapt to dynamically changing data, and sensitivity to the balance of the dataset.

2.2.5 Neural Networks

Neural network algorithm is a kind of intelligent computing model that imitates the behavioral characteristics of biological neural networks, and it solves various complex problems through parallel processing and distributed information processing. The principle of a neural network algorithm is based on simulating the basic characteristics of the human brain or natural neural network, and it realizes the processing and storage of information by constructing a network composed of multiple nodes (or called "neurons"). By connecting and interacting with each other, these nodes are able to analyze, learn, and make decisions about input data. The learning process of a neural network typically involves adjusting the weights of the connections between these nodes to minimize the difference between the predicted output and the actual output, a process known as back-propagation (BP) learning.

Some advantages of neural networks over other machine learning algorithms are their high expressiveness, parallel processing, and adaptive learning capabilities. However, neural network algorithms also have some limitations and defects, such as: overfitting problem, long training time and high demand for computational resources, BP algo-

rithm and other common training algorithms such as slow convergence and easy to fall into the local optimum problem.

3. Application of Machine Learning Models for Medical Diagnosis of Heart Disease

3.1 Heart disease prediction through machine learning algorithms

Machine learning algorithms play a key role in recent advances in heart disease prediction. Deep learning techniques, especially neural network-based methods, have been shown to outperform traditional statistical methods in predicting the natural prognosis of cardiovascular disease and the safety of interventions. In addition, machine learning models, such as artificial neural networks, random forests, and support vector machines, have shown high diagnostic accuracy in early cardiac arrest prediction. Integrated learning methods, such as combining different classifiers such as Random Forest, XGBoost and Extra trees classifier, have been shown to improve prediction accuracy. Feature selection, especially ECG parameters, significantly impacts model performance, emphasizing its importance in heart disease prediction models.

In exploring machine learning algorithms to achieve heart disease prediction, this paper analyzes existing research results from several perspectives.

First, there are various machine learning algorithms that have been used in the research of heart disease prediction. For example, the XGBoost algorithm, which is a machine learning algorithm based on Gradient Boosting Tree (GBT) with a high degree of flexibility and accuracy, is used in the analysis of real-time electrocardiogram (ECG) data to detect different types of cardiac signals, including normal, atrial fibrillation (AF), tachycardia, bradycardia, and arrhythmia [6]. In addition, MLbPM model utilizes a combination of data scaling methods, split ratios, optimal parameters, and machine learning algorithms to predict heart disease. Pierre Claver Bizimana et al. achieved up to 96.7% prediction accuracy by combining data scaling methods, split ratios, optimal parameters, and machine learning algorithms [7]. These studies show that the accuracy of heart disease prediction can be significantly improved by reasonably selecting and optimizing machine learning models and their parameters.

On the other hand, different machine learning algorithms perform differently in heart disease prediction. For example, models such as Random Forest (RF), Decision Tree Classifier (DT), Multilayer Perceptron (MP), and XGBoost (XGB) were used for cardiovascular disease prediction in a study and hyperparameter tuning was performed to optimize the results by grid search CV [8]. This study found that the multilayer perceptron had the highest accuracy of 87.28% under cross-validation [8]. This suggests that it is important to consider the algorithm's and dataset's characteristics when selecting a suitable machine learning algorithm.

Some studies have also explored applying specific machine learning techniques in heart disease prediction. For example, a deep learning-based cardiovascular disease risk prediction model utilizes recurrent neural networks to learn representations of a patient's historical electronic medical record data, which effectively captures temporal features in the data and enhances the model's fitting ability and interpretability through an attention mechanism [9]. This model has high recall, F1 value, and AUC value compared to other methods [9].

3.2 Realization of Heart Disease Diagnosis through Machine Learning Algorithms

Research on the realization of cardiac diagnosis through machine learning algorithms has shown significant progress in this field. The application of machine learning techniques in cardiac diagnosis focuses on ECG analysis, coronary heart disease detection, and data mining based on patient medical records.

Electrocardiogram (ECG) is an important tool for the diagnosis of cardiovascular diseases, and machine learning algorithms can effectively classify and detect abnormalities in ECG signals. For example, a study based on the MIT-BIH data file extracted the feature information of ECG signals through wavelet transform and designed and implemented a classification algorithm based on softmax regression and neural network, and the experimental results showed that the correct rate of classification and identification was stable at more than 90% [10].

In addition, the study of clinical ECG classification algorithm based on deep learning also shows that the accuracy and recall of ECG classification can be effectively improved by the one-dimensional convolutional Res Net network structure with multi-lead two-dimensional structure [11].

Coronary heart disease, as a kind of cardiovascular disease, its early noninvasive and nondestructive detection is of great significance for the prevention and treatment of the disease. Studies have shown that the method based on integrated deep learning of bimodal signals can effectively improve the accuracy of coronary heart disease detection, in which the classification accuracy, sensitivity, and spec-

ificity of the dual-input neural network architecture are 95.62%, 98.48%, and 89.17%, respectively, which is better than that of the unimodal signals [12].

In addition to ECG and coronary heart disease detection, machine-learning techniques have been widely used in the prediction of heart disease based on patients' medical records. For example, using machine learning models such as random forest and logistic regression, the probability of heart disease can be predicted based on a patient's medical history text and other relevant risk factors. These studies show that machine learning techniques can effectively process and analyze large amounts of medical data, providing strong support for early diagnosis and risk assessment of heart disease.

3.3 Application of machine learning algorithms in heart disease treatment

The application of machine learning algorithms in heart disease treatment mainly focuses on the optimization of treatment plans. This paper summarizes several key application areas of machine learning in heart disease treatment: personalizing treatment plans, risk assessment and early warning, and improving treatment effects.

The use of machine learning techniques to improve the personalization of heart disease treatment plans first requires an understanding of the current status and potential of the application of machine learning in the diagnosis and treatment of heart disease. Current research is categorized into the following areas:

First, machine learning models with explanatory properties are the cornerstone of personalized cardiology treatment: a study by De Rong Loh et al. proposes the use of the SHAP (SHapley Additive exPlanations) method to provide interpretability of machine learning predictions, which is essential to support personalized cardiology strategies [13]. Physicians can better develop personalized treatment plans by understanding how specific features affect a patient's cardiac structure.

Second, research has shown that a variety of machine learning algorithms, such as decision trees, plain Bayes, logistic regression, support vector machines, and random forests, have applications in predicting heart disease. The accuracy and efficiency of prediction is improved by comparing the performance of these algorithms and selecting the one that best suits the particular dataset and problem.

Finally, through the examination of the patient's medical characteristics, data-oriented models can be established using the patient's medical attributes. This will enable the development of various machine learning techniques to aid in the timely identification of cardiovascular disease. These approaches have the potential to support healthcare professionals in devising personalized treatment strategies based on individual patient circumstances.

In conclusion, the application of machine learning in heart disease risk assessment and early warning covers a wide range of aspects from data collection and processing to disease detection and prediction, to risk assessment model construction and assisted diagnosis. By utilizing the powerful data processing and analysis capabilities of machine learning, doctors are able to more accurately assess the risk of heart disease and provide timely warning and intervention, thus improving the survival rate and quality of life of heart disease patients.

4. Challenges and Prospects of Machine Learning Applications in Healthcare

The application of machine learning in healthcare faces multiple challenges, while showing great potential and outlook. Current challenges of machine learning in healthcare: on the one hand, current applications of machine learning in healthcare suffer from data quality and labeling problems Disease labels in electronic health records (EHRs) are inaccurate, conditions may contain multiple potential subtypes, and healthy individuals are underrepresented in the data [14]. In addition, the application of deep learning methods in medical image analysis faces data quality challenges.

Next is the need for real-time prediction. For certain heart disease prediction tasks, such as real-time prediction of paroxysmal atrial fibrillation, models are required to process a large amount of data and give prediction results in a short period. This requires the model to have high accuracy and low-latency processing capability.

Although machine learning algorithms can process large amounts of complex data, their decision-making process is unfamiliar to many healthcare professionals. This leads to distrust of the algorithm's results, limiting its use in clinical practice.

While there are numerous challenges in applying machine learning to healthcare, the potential for enhancing diagnostic accuracy and advancing personalized medicine is gradually being explored and realized as technology progresses and interdisciplinary collaboration strengthens. In the future, machine learning is anticipated to assume a more prominent role in the medical domain by addressing concerns such as data quality, algorithms interpretability, data silos, and ethical considerations surrounding privacy.

5. Conclusion

With the advancement of society and changes in people's lifestyles, machine learning and artificial intelligence have become increasingly important in predicting, diagnosing, and treating heart disease. This paper aims to provide an overview of the fundamental principles and commonly used algorithms of machine learning in medical diagnosis for cardiovascular diseases. Additionally, it examines their strengths, weaknesses, and specific applications through practical examples. In addition, the advantages and challenges of machine learning in heart disease medical treatment are analyzed from three aspects: prediction, diagnosis and treatment, and the future trends of machine learning in the field of heart disease prediction are looked forward to. The aim of this paper is to provide new ideas and references for heart disease prediction and treatment, as well as directions for related researchers. In summary, although machine learning still has some limitations and challenges in heart disease prediction, diagnosis and treatment, it still has great potential and application prospects in heart disease prediction. Through further research and innovation, machine learning technology is expected to become an important tool to improve the accuracy and precision of heart disease prediction and positively contribute to improving human health.

6. References

[1] World Heart Report 2023: Confronting the World's Number One Killer. Geneva, Switzerland. World Heart Federation. 2023.

[2] Allah, E.M.A., El-Matary, D.E., Eid, E.M. and El Dien, A.S.T. (2022) Performance Comparison of Various Machine Learning Approaches to Identify the Best One in Predicting Heart Disease. Journal of Computer and Communications, 10, 1-18.

[3] Nashif, S., Raihan, Md.R., Islam, Md.R. and Imam, M.H. (2018) Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System. world Journal of Engineering and Technology, 6, 854-873.

[4] Yingying Du. Analysis of heart disease prediction effect of different classification models[J]. Modeling and Simulation, 2023, 12(6): 5600-5607.

[5] Jung, A. (2018). Machine Learning: basic principles. arXiv: Learning.

[6] D. Bertsimas, L. Mingardi and B. Stellato, "Machine Learning for Real-Time Heart Disease Prediction," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 9, pp. 3627-3637, Sept. 2021, doi: 10.1109/JBHI.2021.3066347.

[7] Pierre Claver Bizimana, Zuping Zhang, Muhammad Asim, Ahmed A. Abd El-Latif, "[Retracted] An Effective Machine Learning-Based Model for an Early Heart Disease Prediction", BioMed Research International, vol. 2023, Article ID 3531420, 11 pages, 2023. https://doi.org/10.1155/2023/3531420

[8] Bhatt, C.M.; Patel, P.; Ghetia, T.; Mazzeo, P.L. Effective Heart Disease Prediction Using Machine Learning Techniques. algorithms 2023, 16, 88. https://doi.org/10.3390/a16020088

[9] An, Y., Huang, N.J., Yang, R., et al. Deep learning-based cardiovascular disease risk prediction model[J]. Chinese Journal of Medical Physics,2019,36(09):1103-1112.

[10] LIU Teng, TANG Hong, ZHANG Shibing. Research on arrhythmia signal classification algorithm based on machine learning[J]. Computer Application Research,2020,37(03):940-943.DOI:10.19734/j.issn.1001-3695.2018.07.0545.

[11] Liu Shouhua, Wang Xiaosong, Liu Yu. Deep learning based classification algorithm for clinical ECG[J]. Computer and Modernization,2021(08):52-57.

[12] Li Han. Research on coronary heart disease detection based on deep learning of bimodal signal integration[D]. Shandong University,2021.DOI:10.27272/d.cnki.gshdu.2020.003894.

[13] De Rong Loh, Si Yong Yeo, Ru San Tan, Fei Gao, Angela S Koh, Explainable machine learning predictions to support personalized cardiology strategies. European Heart Journal - Digital Health, Volume 3, Issue 1, March 2022, Pages 49-55, https://doi.org/10.1093/ehjdh/ztab096

[14] Ghassemi, M., Naumann, T., Schulam, P.F., Beam, A., & Ranganath, R. (2018). Opportunities in Machine Learning for Healthcare. arXiv, abs/1806.00388.