2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE) | 978-1-6654-8474-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/AEMCSE55572.2022.00071

Research of Heart Disease Prediction Based on Machine Learning

Shuge Ouyang*
University of Minnesota Twin City, Minneapolis, MN, USA, 55455;
* Corresponding author: ouyan164@umn.edu

Abstract: **The use of massive clinical data in the medical field for supporting medical decision support is an inevitable development trend. Medical decision support is based on a variety of data sources accumulated and acquired in real-time in the clinic, and various machine learning algorithms are used to achieve classification of patient disease types or prediction of disease risks. This paper assists in performing cardiac disease prediction starting from different heart disease types (coronary heart disease) and data sets, summarizing the currently adopted machine learning diagnosis and prediction methods, highlighting the characteristics and differences of these methods, and analyzing the challenges and future developments. The results show that machine learning techniques have a wide range of applications in cardiac diseases. However, each machine learning method can only be applied to a specific scope due to the non-uniformity of medical data. At the end of the article, the prediction of heart disease is summarized.**

Keywords: heart disease; coronary heart disease; heart disease prediction; machine learning

I. INTRODUCTION

The heart is one of the most essential organs in the human body. The primary function of the heart is to power the flow of blood and run it to all parts of the body. Heart disease, a common circulatory disorder also known as cardiovascular disease, encompasses a variety of heart diseases. Heart disease is today the most severe disease that threatens human life. According to World Health Organization estimates, heart disease kills 12 million people worldwide each year, and on average it takes the life of one person every 34 seconds in the United States alone. Coronary heart disease, arrhythmias, and myocardial infarction are the most common heart diseases, affecting 8 billion people worldwide [1]. Take coronary heart disease, for example, according to recent statistics from the American Heart Association, coronary heart disease accounted for 13 percent of deaths in the United States in 2018 [2].

The sudden onset of heart disease and the short time available for resuscitation lead to very few incidents in which heart attack patients are successfully resuscitated. In many cases, the patients die on the spot. Therefore, the rapid and effective prediction of potential patients with heart disease has become a critical and challenging task in the medical field. Current researchers have been interested in predicting heart disease, and they have developed several prediction methods for different heart diseases. Machine learning is currently one of the most rapidly developing subfields of artificial intelligence. It is an effective intelligence tool for rapid analysis of data and is used in many areas of life, including health care.

Many medical institutions worldwide and hospitals worldwide count data on various characteristics of heart patients, such as the patient's gender, age, heartbeat per minute, blood pressure, and other common data from daily medical examinations. A large amount of patient data is collected. Human observation alone still cannot obtain valid information from these large amounts of data or derive heart patients' characteristics. Machine learning algorithms can learn from existing patient cases and quickly analyze the data for heart disease, which is why machine learning is widely used in the medical field to analyze disease data.

Classification, the most basic type of machine learning, enables quickly dividing the data as required. In predicting heart disease, effective classification can promptly classify the indicated person into two categories by different features, whether they have heart disease or not. For example, the well-known UCI heart disease dataset can be used to train and test classifiers [3].

The accuracy of heart disease prediction has been a concern for researchers because the prognosis of the condition affects the judgment of physicians and the self-protection of potential patients, and incorrect prediction can have incalculable consequences.

The rest of this paper will present heart disease and coronary heart disease predictions based on different datasets and summarize the data results and predictions obtained by researchers through various machine learning methods and compare the performance of different machine learning methods in heart disease prediction.

II. MAIN BODY

A. Background of machine learning in the medical field Margins

Computers were first introduced as an aid to diagnosis by Ledley et al. [1], who introduced mathematical models into clinical science under intelligent diagnosis. The earliest digital systems used symbolic reasoning to aggregate the clinical experience of physicians and existing medical knowledge to create a database of a particular disease. The database contents were called up at the time of diagnosis using symbolic reasoning. However, this approach had substantial limitations and was gradually phased out because the system was too simple to be fully utilized for complex medical data. In its place, machine learning methods are now used by many scholars, and advances in computer and artificial intelligence technology are sufficient to allow intelligent diagnostic models to be constructed. Various common diseases such as cancer, diabetes,

978-1-6654-8474-9/22/$31.00 ©2022 IEEE
DOI 10.1109/AEMCSE55572.2022.00071

heart disease, Alzheimer's disease, etc. [2], can be detected by artificial intelligence diagnostic algorithms.

For various heart diseases, which are clinically chronic and have a high mortality rate, it is difficult to analyze and diagnose these diseases by traditional medical judgment accurately. In recent years, with the involvement of information technology in medical diagnosis, a large number of clinical test reports electronic case and treatment reports have been generated. Various machine learning algorithms can be used to extract and process these medical data effectively to achieve the rapid and correct diagnosis and prediction of heart diseases [3]. In addition, many scholars have proposed various theories and models to improve the detection of heart disease prediction to achieve better diagnostic results and prediction accuracy.

Heart disease is routinely divided into six categories, namely congenital heart disease, pulmonary heart disease, rheumatic heart disease, coronary heart disease, hypertensive heart disease, and cardiomyopathy. There are different datasets for different characteristics of heart diseases. This paper classifies heart diseases according to different etiologies and compare the prediction effect of various machine learning methods for different types of heart diseases.

B. Unclassified heart disease

Many of the available datasets on heart disease do not differentiate according to the type of heart disease. The most commonly used of these datasets, which determine only for whether a patient has the disease, is the heart disease dataset available in the UCI Machine Learning Database, which has 303 samples and 76 features and is a combination of four datasets from Cleveland, Switzerland, Hungary, and Long Beach [4]. Fourteen variables were included in this dataset as influencing factors, with age, blood pressure, cholesterol level, and maximum heart rate as continuous variables and the others as discrete variables (see Table 1 for details).

Table 1. Variable explanation

Variables	Meaning
Age	The person's age in years.
Sex	The person's sex(1=male,0=female).
Cp	The chest pain experienced (Value1: typical angina, Value2: atypical angina, Value3: non-anginal pain, Value4: asymptomatic.
Trestbps	The person's resting blood pressure.
Chal	The person's cholesterol measurement.
Fbs	The person's fasting blood sugar (>120 mg/dl, 1=true; 0=false).
Restecg	Resting electrocardiographic (0=normal, 1=having ST-T wave abnormality, 2=left ventricular hypertrophy by Estes' criteria).
Thalach	The person's maximum heart rate.
Exang	Induced angina (1=yes; 0=no)
Oldpeak	ST depression induced by exercise relative to rest.
Slope	Slope of the peak exercise ST segment.
Ca	The number of major vessels(0-3).
thal	A blood disorder called thalassemia.
Target	Whether have heart disease or not (0=healthy, 1=sick).

1) Decision Trees and Random Forests in heart disease prediction

A decision tree is a supervised learning algorithm, defined by Berry and Linoff as "a structure that can be used to divide many records into multiple parts (see figure 1). Successive smaller sets of records are made by applying a series of simple decision rules. With each successive division, the members of the result set become increasingly similar [5]." In a decision tree, each node represents an object, and the decision point is the final choice, if it is a multi-level decision problem, there can be more than if it is a multi-level decision problem, there can be multiple decision points. The model construction of the decision tree is classified into two parts: feature selection and spanning tree [6].

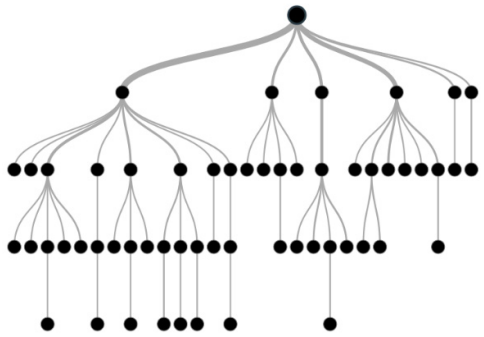

Figure 1. Decision tree diagram

For this dataset, the accuracy is only 77.55% with the decision tree approach [7], and when the decision tree is used in combination with the augmentation technique, it improves the accuracy rate to 82.17%. While in the decision tree for the correctly classified instances only 42.8954% of the percentage performance is very poor. Some scholars have also implemented decision trees using the J48 algorithm and obtained an accuracy improvement to 67.7%, which is a small improvement compared to the former [8].

Decision trees, although low in complexity and high in efficiency, are prone to overfitting, leading to limitations [9]. Random forest is an algorithm to improve the accuracy of models to address the limitations of single decision trees. The algorithm was proposed by Breiman [10] in the 1990s mainly for classification and regression, and is now widely used in biological processing, face recognition, etc.

Using the random forest algorithm, the importance of the features in the dataset is analyzed and it found that "trestbps", which is a person's resting blood pressure, is the feature with the highest importance because it has the greatest influence on whether a person has heart disease [6]. ROC (Receiver Operating Characteristic), whose main analysis tool is a curve drawn on a two-dimensional plane with the horizontal coordinate of the false positive rate (FPR) and the vertical coordinate of the true positive rate (TPR). The value of AUC is the size of the area below the ROC curve. Usually, the value of AUC is between 0.5 and 1.0, and a larger AUC means better performance. From the plotted ROC curve (see figure 2), for this dataset, the random forest algorithm has the largest AUC value i.e., the largest area under the ROC curve with a value of 0.965, and the accuracy of the algorithm is improved by 10.8% compared to the decision tree. For this dataset, the random forest algorithm obtained better results, but the data volume of

this dataset is small and it is not possible to prove the effectiveness of the algorithm in large datasets [6].

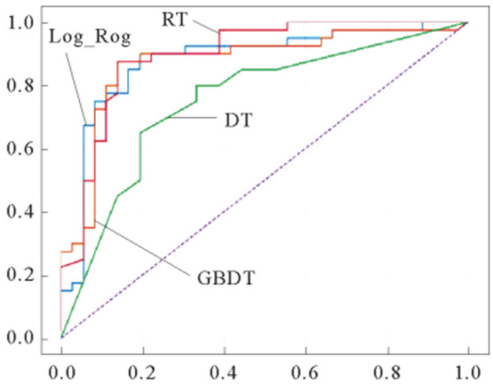

Figure 2. Comparison of ROC curves between different methods

2) SVM in heart disease prevention

Support vector machine (SVM) can be used as both predictor and classifier, making it a very popular supervised learning method. In the classification process, a hyperplane space is found in the feature space, and regions are classified. The training data points are represented by the SVM model as points in the feature space and then mapped in such a way that the points belonging to different classes are separated by the widest possible edges. Eventually, the test data points are mapped to the same space and classified by determining on which side of the boundary the data points fall.

Parthiban and Srivatsa [12] used a dataset of 500 patients collected at the Institute of Research, Chennai, India. The plain Bayesian algorithm and SVM algorithm was implemented using WEKA to diagnose heart disease in diabetic patients. Weka (Waikato Environment for Knowledge Analysis) is a java-based machine learning and data mining software. As an open data mining platform, WEKA integrates many machine learning algorithms that can undertake data mining tasks, including data preprocessing, classification, regression, clustering, association rules, and visualization in a new interactive interface. The results showed that the accuracy of the plain Bayesian algorithm was 74% and the accuracy of SVM was up to 94.60%.

In contrast to most researchers who use a single machine learning method to classify and evaluate the results, Tan et al. [13] proposed a hybrid technique that combines two machine learning algorithms, genetic algorithm (GA) and SVM, using a wrapper approach, and implemented and analyzed on LIBSVM and WEKA. The experimental results show that the diagnostic accuracy of 84.07% for the heart disease dataset using a hybrid approach of genetic algorithm and support vector machine.

Table 2. Heart disease diagnosis summary table

Researchers	Algorithm	Dataset	Accuracy
Chaurasia, Pal [14]	Naïve Bayes J48 Bagging	UCI Machine Learning Database (seen Table1)	82.31% 84.35% 85.03%
Vembandasamy et al. [15]	Naïve Bayes	Institute of Research Chennai	86.42%
Parthiban, Srivatsa [12]	SVM Naïve Bayes	Institute of Research Chennai	94.6% 74%
Tan et al. [13]	GA+SVM	UCI Machine Learning Database (seen Table 1)	84.07%

From table 2, it can be visualized that for the diagnosis of heart disease, for both the UCI and Chennai datasets, SVM achieved the highest accuracy rates of 85.1% and 94.60%, respectively. The advantage of SVM over other algorithms is that it can correctly achieve classification when faced with limited samples, is not prone to over the advantage of SVM is that it can correctly classify a limited number of samples without overfitting and has better robustness. The disadvantage is that SVM is computationally cost is high, so it runs slowly and is not suitable for large training The disadvantage is that SVM is computationally expensive and therefore slower, and is not suitable for large training samples, and requires a combination of models for multiple classification tasks.

C. Coronary heart disease prediction

Coronary heart disease, also known as ischemic heart disease, is a heart condition caused by coronary atherosclerosis leading to myocardial ischemia and hypoxia. As the most common cardiovascular disease in the world, coronary heart disease (CHD) is one of the leading causes of human mortality today. According to statistics published by the World Health Organization, in 2015, about 17.7 million people died of coronary heart disease, accounting for 31% of the world's deaths from the disease, and more than three-quarters of them occurred in low- and middle-income countries [16]. The diagnosis and treatment of coronary heart disease are more complex, especially in developing countries, where the scarcity of diagnostic instruments and the lack of medical personnel affect the early detection and further treatment of patients with coronary heart disease, leading to serious consequences (Table 3).

In recent years, many scholars have been trying to apply machine learning and neural network algorithms to the field of coronary artery disease diagnosis, which is of great practical significance to reduce the cost and difficulty of diagnosis, improve the efficiency of diagnosis, and the accuracy of prediction. Alizadehsani et al. proposed a support vector machine (SVM)-based model for coronary heart disease prediction on the Z-Alizadeh Sani dataset with a model accuracy of 96.4% [17]. Liu et al. used clinical data and predicted coronary heart disease based on a logistic regression algorithm with a model accuracy of 93.6% [18]. Palaniappan et al. selected 15 features in the Cleveland dataset and used Naive Bayes for the prediction of coronary heart disease with a model accuracy of 86.12% and discussed the important factors

affecting coronary heart disease [19]. Xiaoli Wang et al. used clinical data to assess the risk of coronary heart disease in the elderly based on the XGBoost algorithm, with a model accuracy of 82% [20]. On ECG signal data, neural network-based algorithms performed better, including Hui Meng et al. who used IPSO-BP neural network to model and predict clinical ECG signals with an overall model accuracy of 87% [21]. Wang et al. used ECG signal data based on CWT-CNN neural network algorithm and achieved a diagnostic accuracy of 98.7% [22]. Based on the imbalance in the NHANES coronary heart disease survey dataset, Dutta et al. used a randomized secondary sampling technique and selected 38 features to make predictions for coronary heart disease using a convolutional neural network (CNN), with a final model accuracy of 81.8% [23].

Table 3. Comprehensive comparison of existing related studies.

Researchers	Model Algorithm	Data	Accuracy	Features and shortcomings
Alizadehsani et al. [17]	SVM	Sani dataset	96.4%	Suitable for small-sample nonlinear and generalizable learning, lacking interpretability analysis of the model.
Liu et al. [18]	Logistic regression	clinical data	93.6%	Suitable for small and specific datasets, the importance of each feature is discussed
Palaniappan et al. [19]	Naïve Bayes	Cleveland dataset	86.1%	Suitable for small-scale data, insensitive to missing data, and discusses the importance of model features
Xiaoli Wang et al. [20]	XGBoost	clinical data	82%	Regularization is applied to the loss function to prevent overfitting of the model, but the model lacks interpretability
Hui Meng et al. [21]	IPSO-BP	ECG	87%	Requires large amounts of data for training and poor model interpretation
Jahmunah et al. [24]	GaborCNN	ECG	98.7%	Good noise handling ability, more accurate subdivision of coronary artery disease types, but interpretable poorly interpretable
Shu et al. [25]	CNN-LSTM	ECG	98.5%	The learning process is slow, requires large amounts of data for training, and the model is not interpretable
Feng et al. [26]	CNN-LSTM	ECG	95.54%	Friendly to non-linear data, but the model is more complex, the operation time is long, and the model does not have interpretability
Sowmiya et al. [22]	CWT-CNN	ECG	98.7%	Allow samples to have large deficiencies and distortions, run fast, and the model has certain interpretability
Dutta et al. [23]	CNN	NHANES clinical data	81.8%	Good fault tolerance and parallel processing capability, random secondary sampling for unbalanced samples, no interpretable analysis of the model

As can be seen from table 3, machine learning methods such as random forests and neural networks are mainly used for the prediction of coronary heart disease, which have high performance and accuracy, but the lack of interpretability of the models is not conducive to assisting physicians in making final medical decisions, and most of the research data come from ECG data and clinical data, which are more difficult to obtain.

III. CONCLUSION

This paper presents several common approaches based on machine learning for medical decision support in the prediction of heart disease and coronary heart disease in medicine, and reviews and analyzes the advantages and disadvantages of the involved machine learning methods applied in different disease models. Based on the above analysis machine learning methods have a great range of applications in predicting heart-related diseases. However, by comparing the prediction results of different methods it can be concluded that each of the above methods will perform well in some cases, but in other cases, their performance will become worse. For example, decision trees perform very well when using principal component analysis, but in other cases, they perform very poorly due to overfitting. Random forests solve the overfitting problem by using multiple algorithms and decision trees. Support vector machines have a very good performance in most heart disease predictions. Although various machine learning techniques have been able to achieve very good accuracy in predicting heart-related diseases, A great deal of research is still needed in solving overfitting problems or dealing with large data sets. There are still many efficient algorithms to be developed for specific types of data. The choice of network models for different diseases, how to train and optimize the model parameters, how to improve the efficiency of learning methods, and many other scientific questions need to be addressed.

In summary, it is found that today's data set for the collection of medical data has not been perfected, and the lack of overall unified standards for data in the industry has led to data incompatibility and structural complexity, and the quality of data is greatly affected. The abnormal data structure in machine learning technology processing will inevitably lead to a waste of resources due to the consumption of model computing power. Therefore, standardization and unification of data is a major bottleneck on the road to intelligence in healthcare.

REFERENCES

[1] Ledley R S, Lusted L B. “Reasoning foundations of medical diagnosis,” Science, 130:9-21, (2009).

[2] Kumar G. “A survey on machine learning techniques in health care industry,” International Journal of Recent Research Aspects, 3(2) pp.128-132, (2016).

[3] Tiezheng Sun, Zehao Yu, “A study of heart disease case classification prediction based on machine learning, 17, (26), pp. 439-448, (2021).

[4] Shutong Liang, Maozu Guo, Lingling Zhao. “Survey on medical decision support systems based on machine learning. Computer Engineering and Applications, 55(19), pp.1-11, (2019).

[5] Thenmozhi K., Deepika P., “Heart Disease Prediction Using Classification with Different Decision Tree Techniques, International Journal of Engineering Research and General Science 2(6), (2014).

[6] Jingchao Zhao, Yi Li, “Research on heart disease prediction algorithm based on optimized random forest,” Journal of Qingdao University of Science and Technol, pp. 491-500, (2021).

[7] Seyedamin Pouriyeh, et. al. “A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease”, 22nd IEEE Symposium on Computers and Communication (ISCC 2017): Workshops - ICTS4eHealth (2017).

[8] Ramalingam V.V., Ayantan Dandapath, M Karthik Raja, “Heart disease prediction using machine learning techniques: a survey, International Journal of Engineering & Technology, 7 (2.8) pp. 684-687, (2018)

[9] Hongyan Yu, Qian Feng, “Research review of random forest algorithm,” Journal of Hebei Academy of Sciences, (2019).

[10] Coffey, Sean, et al. “Global epidemiology of valvular heart disease,” Nature Reviews Cardiology 18.12 pp. 853-864, (2021).

[11] Rani, Pooja, et al. "A decision support system for heart disease prediction based upon machine learning." Journal of Reliable Intelligent Environments 7.3 pp. 263-275, (2021).

[12] KSasipriya V. R, Deepa E. “Heart diseases detection using naive Bayes algorithm,” International Journal of Innovative Science, 2, pp.441-444, (2015).

[13] Tan Teoh, Yu Q. et al. “A hybrid evolutionary algorithm for attribute selection in data mining,” Expert Systems with Applications, 36, pp.8616-8630, (2009).

[14] Chaurasia V, Pal S. “Data mining approach to detect heart disease,” International Journal of Advanced Computer Science and Information Technology, 2(4) pp. 56-66, (2013).

[15] Katarya, Rahul, and Sunit Kumar Meena. “Machine learning techniques for heart disease prediction: a comparative study and analysis,” Health and Technology 11.1 pp.87-97, (2021).

[16] Baudet, Daugareil, Laulom, et al. “Cardiovascular diseases,” Annales cardiology et angiology, 68 (1), pp.49-52, (2018).

[17] Alizadehsani R, Hosseini M J, Khosravi A, et al. “Non-invasive detection of coronary artery disease in high-risk patie nts based on the stenosis prediction of separate coronary arteries,” Computer Methods and Programs in Biomedicine, 162 (1), pp. 119-127, (2018).

[18] Jiang Liu. “Prediction of all-cause mortality in coronary artery disease patients with atrial fibrillation based on machine learning models,” BMC Cardiovasc Disord. 21 (1) pp. 499-507, (2021).

[19] Palaniappan S, Awang R. “Intelligent heart disease prediction system using data mining techniques,” ACS International Conference on Computer Systems and Applications. Piscataway, pp.108-115, (2008).

[20] Xiaoli Wang, Tianxing Shi, Derong Peng, et al. “Comparative study on the effectiveness of two machine learning algorithms in building risk assessment model of coronary heart disease in the elderly,” Chinese Journal of General Practice, 19 (04), pp. 523-527, (2021).

[21] Hui Meng, Jiahong Zhang, Min Li, et al. “Prediction and classification of coronary heart disease based on IPSO-BP neural network and BCG signal,” Journal of sensing technology, 33 (10), pp.1379-1385, (2020).

[22] Sowmiya, C., and P. Sumitra. "A hybrid approach for mortality prediction for heart patients using ACO-HKNN." Journal of Ambient Intelligence and Humanized Computing 12.5 pp. 5405-5412, (2021).

[23] Dutta A, Batabyal T, Basu M, et al. “An Efficient Convolutional Neural Network for Coronary Heart Disease Prediction,” Expert Systems with Applications, 159 (2) pp. 113408-113408, (2020).

[24] Jahmunah V, Ng E, San T R, et al. “Automated detection of coronary artery disease, myocardial infarction and congestive heart failure using GaborCNN model with ECG signals,” Computers in Biology and Medicine, 134: 104457, (2021).

[25] Shu L O, Vicnesh J, Ru S T, et al. “Comprehensive electrocardiographic diagnosis based on deep learning,” Artificial intelligence in medicine, 103: 101789, (2020).

[26] Feng K, Pi X, Liu H, et al. “Myocardial infarction classification based on convolutional neural network and recurrent neural network,” Applied Sciences (MDPI), 9 (9) pp. 1–12, (2019).