Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/45/20241122

Auxiliary prediction of cardiovascular diseases based on deep learning

Zhongzheng Zhang[1], Junnan Zhu[2,3]

[1]School OF Mathematics & Computer Science, Wuhan Polytechnic University, Wuhan, 430040, China.
[2]College Of Information Engineering, Sichuan Agricultural University, Chengdu, 625014, China.

[3]201800890@stu.sicau.edu.cn

Abstract. Assisting in the prediction of cardiovascular diseases is beneficial for both individuals and medical institutions, especially for cardiovascular diseases, which are prevalent worldwide and easily lead to death and health issues. Although many scholars have discussed various datasets from different regions, understanding their internal decision-making process is challenging. This paper aims to discuss and verify data features by collecting data from Kaggle, as well as use the Synthetic Minority Over-sampling Technique (SMOTE) to address data imbalance issues. The research used multiple machine learning, ensemble learning, and deep learning models for analysis, compared the accuracy of these models, and chose the best model to better assist doctors and help individuals in disease prediction. The implemented models in this paper include K Nearest Neighbors (KNN), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting Decision Trees (GBDT), and Deep Neural Network (DNN). The primary focus on XGBoost, GBDT, and DNN yielded AUC scores in the test set of 0.823, 0.825, and 0.832, respectively. The established DNN model demonstrates a good auxiliary prediction capability for cardiovascular diseases.

Keywords: DNN, XGBoost, GBDT, Ensemble Learning, Cardiovascular Disease, Smote.

1. Introduction

According to research from the World Health Organization (WHO), in 2019, approximately one-third of global deaths were related to cardiovascular diseases [1]. Research into the prediction of cardiovascular diseases is of great importance. It helps improve public health, reduce disease burden, and promote the sustainable development of the healthcare system.

Machine learning and artificial intelligence have played a significant role in predicting cardiovascular diseases. Researchers have developed prediction models and clinical datasets, predicting from various perspectives such as a patient's lifestyle, genetic information, and biomarkers. While some studies have achieved commendable success, there are issues related to the overly complex features included in datasets, privacy concerns, and insufficient dataset sizes. These could lead to the risk of overfitting, resulting in models with poor generalization.

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/45/20241122

This paper selects a dataset with a moderate number of features and good distribution for prediction. In addition to traditional machine learning models, we have also employed DNN and ensemble learning models and for comparison. We discuss the dataset's features and, through continuous experimentation and parameter adjustments, aim to complete research with high generalizability and features that are appropriately obtainable.

The sequential steps outlined are as follows: The second section delves into methodology research, starting with the study of the dataset, including discussing features with high correlations within the dataset, as well as discussions on feature transformations, data preprocessing, and data augmentation. This is followed by a discussion on model selection. The third section analyzes experimental results and conclusions, and finally, we summarize the experiment and look to the future.

2. Methodology

In this study, the present paper initially conducted exploratory data analysis on the dataset and performed data augmentation to address the imbalanced segments. Subsequently, the input data underwent preprocessing. Following that, several machine learning models, including KNN, RF, XGBoost, GBDT, and DNN were constructed and trained to acquire the corresponding outcomes for further analysis. The process is illustrated in Figure 1.

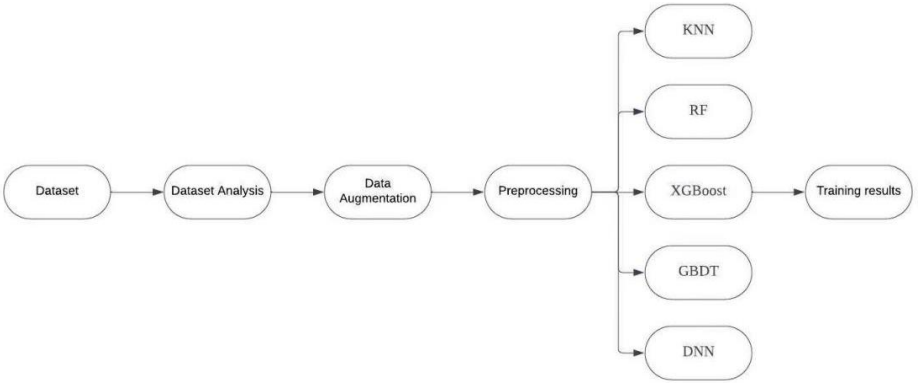


Figure 1. Research Process.

2.1. Dataset

The dataset is sourced from the Kaggle website, comprised of data classified into four categories: Objective (directly obtainable objective data), Examination (medical examination data), Subjective (subjective data), and Target (target data). It includes a total of 70,000 data entries, with 11 features and 1 target value, as detailed in Table 1.

Table 1. Data Feature.

Name	Attribution	Type
Age	Objective	Int(day)
Height	Objective	Int(cm)
Weight	Objective	Float
Gender	Objective	Category
Systolic blood pressure	Examination	Int
Diastolic blood pressure	Examination	Int
Cholesterol	Examination	Category code
Glucose	Examination	Category code
Smoking	Subjective	Binary
Alcohol	Subjective	Binary
Physical activity	Subjective	Binary
Cardiovascular disease	Target	Binary

DOI: 10.54254/2755-2721/45/20241122

2.2. Data Discussion

2.2.1. Correlation Analysis and Feature Processing. Firstly, this paper visualizes and analyzes each feature using a heatmap. As can be seen from the figure 2, the correlation between age, weight, and cholesterol is the most noteworthy. Cholesterol is a type of lipid molecule that plays a vital physiological role in the human body. However, Elevated levels of cholesterol, particularly an increase in low-density lipoprotein cholesterol, are associated with a higher susceptibility to cardiovascular diseases [2]. But, as certain professional medical indicator, it will not be discussed in detail.

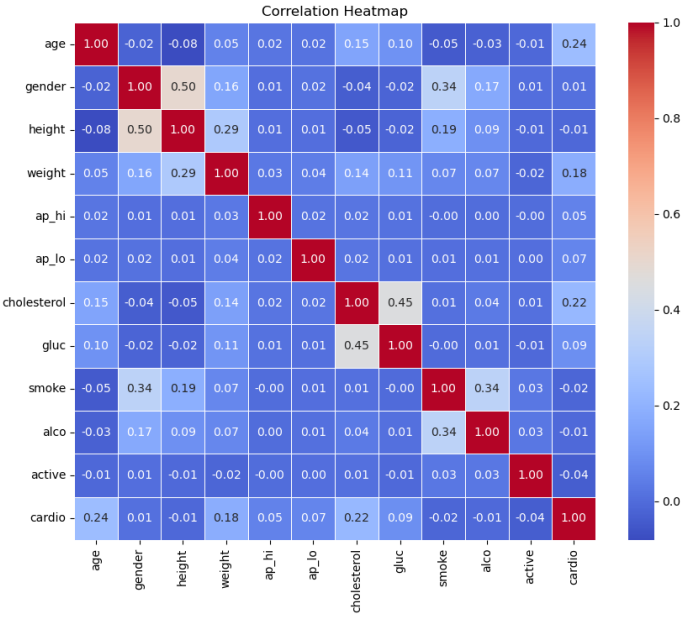


Figure 2. Feature Heatmap.

In terms of age, there is a close relationship between age and cardiovascular diseases. As age increases, the risk of an individual developing cardiovascular diseases also significantly rises. With the progression of age, arteries gradually lose their elasticity, and the risk of atherosclerosis (accumulation of plaques on the arterial walls) increases. This can lead to the onset of hypertension, coronary heart disease, heart attacks, strokes, and other cardiovascular diseases [3]. The initial dataset expressed age in terms of days. To improve visualization, days were converted to years using the following formula:

$$Age_{years} = \frac{Age_{days}}{365.25} \tag{1}$$

After obtaining the year, and observing the age results obtained, we first used a histogram with a Kernel Density Estimate (KDE) curve. As shown in the figure 3. The age range is between 35 and 65, which is also the age group with a high incidence of cardiovascular diseases. The age bracket falls between 35 and 65, encompassing the age range where cardiovascular diseases are prevalent. The utilization of box plots indicates that the age range most susceptible to cardiovascular diseases is 50 to 60.

DOI: 10.54254/2755-2721/45/20241122

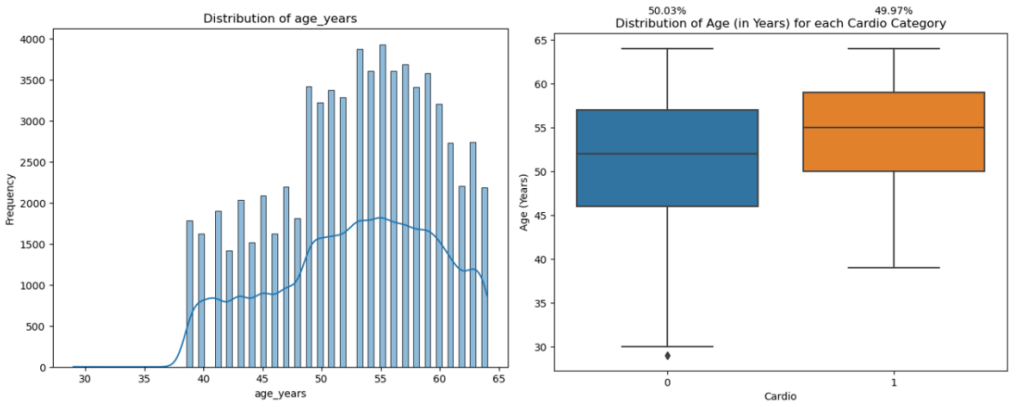


Figure 3. KDE Histogram and Box Plot of Age.

Meanwhile, in the research by Sun W, Zhang P, they directly used the 'year' feature [4] and did not process using days. Therefore, this paper also follows this approach, adding the feature of the year.

In research related to weight, it has indicated that obesity stands as a notable risk element for coronary heart disease. It can cause chronic inflammation, affect vascular function, and increase the risk of atherosclerosis (accumulation of plaques on the arterial walls), leading to coronary artery disease. At the same time, it might result in lipid abnormalities like high cholesterol and high triglycerides, thereby increasing the risk of cardiovascular diseases [5]. However, one cannot directly determine if a person is obese based solely on weight. Factors such as height and age also need to be considered.

In the paper by Farmer R E, Mathur R, a reference value for BMI is suggested [6]. Therefore, this paper adds the BMI feature for subsequent model training. The formula is as follows:

$$BMI = \frac{Weight(kg)}{[Height(m)]^2} \tag{2}$$

After adding the BMI and Year features, we did not directly delete the original Height, Weight, and Day features. Instead, during subsequent training, we combined them to see whether the new features or the original features yield better results.

2.2.2. Data Augmentation

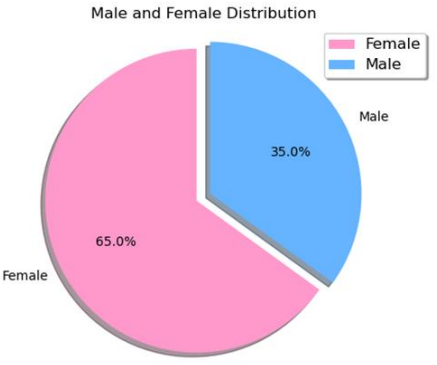


Figure 4. Distribution of Gender.

Considering that the dataset only has 70,000 entries, and noting the uneven distribution between males and females as shown in Figure 4, according to a survey by Nielsen M W, Stefanick M L [7], the proportion of males who smoke and drink is higher than that of females. The smaller number of males

DOI: 10.54254/2755-2721/45/20241122

in the dataset compared to females might lead to an inability to truly capture the relationship between smoking or drinking and the eventual onset of disease. As a result, we considered addressing this using the SMOTE method. SMOTE generates synthetic samples by interpolating between the less prevalent category, thus increasing the number of less prevalent category samples and making the proportions between different classes more balanced [8]. This helps the model to better capture patterns of the less prevalent category.

The formula is as follows:

Assume there's a less prevalent category sample X_i, and its Knearest neighbor minority class samples. For each nearest neighbor sample X_{zi},we can generate a new synthetic sample X_{new},using the following formula:

$$X_{new} = X_i + \alpha \times (X_{zi} - X_i) \tag{3}$$

α is a random value between [0,1] used to control the position between X_i and X_{zi}.

In this paper, when using the method, the sampling_strategy parameter is set to 'auto'. It will automatically balance the sample quantity among different classes, making the sample count for each category relatively even. As a result, the total number of entries in the dataset increased to 90,000.

2.2.3. Reprocessing. To enhance the model's generalization capability, prevent weight biases among features, and improve model convergence speed, normalization was applied to ensure all features are within a similar scale range. This also facilitates a more intuitive comparison of relationships between features. The Min-Max normalization method was utilized. This method is suitable for various data distributions. Regardless of whether the data follows a normal distribution, it can scale the data value range to [0, 1]. It retains the original data distribution relationship and does not alter the data's relative order. This is essential for the performance and interpretability of some models, as it doesn't change the inherent characteristics of the data [9]. The formula is as follows:

$$X_{normalization} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{4}$$

2.3. Model

In this study, when selecting the model, we compared three types of algorithms: traditional machine learning (KNN), ensemble machine learning (RF, XGBoost, and GBDT), and deep learning (DNN).

2.3.1. KNN. The KNN algorithm is an instance-based supervised learning algorithm. Once the KNN model has been constructed, adding new training samples is very straightforward, eliminating the need to retrain the entire model [10]. KNN is a non-parametric algorithm, indicating it doesn't make assumptions about the distribution of the data. This gives it an advantage when dealing with complex, non-linear problems. The algorithmic process is as follows

Distance metrics are used to calculate the distance between the sample to be predicted and each sample in the training data. In this study, the Euclidean distance is employed for these calculations. Suppose there are two sample points $X_i$and X_j, with their feature vectors being $V_i$and $V_j$respectively. The calculation is then performed based on the following formula:

$$D(V_i, V_j) = \sqrt{\sum_{k=1}^{n} (v_{ik} - v_{jk})^2} \tag{5}$$

Where n represents the number of features, v_{ik} and v_{jk} are the values of the two samples for the feature.

Following that, we enter the neighbor selection phase. For the sample to be predicted, the training samples closest in distance to it are identified. These training samples are its k nearest neighbors. Finally, for classification tasks, the occurrence of each category within these nearest neighbors is counted. The sample to be classified is then assigned to the category with the highest occurrence.

DOI: 10.54254/2755-2721/45/20241122

2.3.2. RF. RF is an ensemble learning algorithm employed in tasks encompassing both classification and regression [11]. It's an ensemble method based on decision trees. By constructing multiple decision trees and integrating their prediction results, the performance and robustness of the model can be improved. Each decision tree's classification result is combined through a voting mechanism. For classification problems, the final classification will be the result of the majority vote from the decision trees.

The flow of how RF (Random Forest) makes predictions is illustrated in Figure 5.

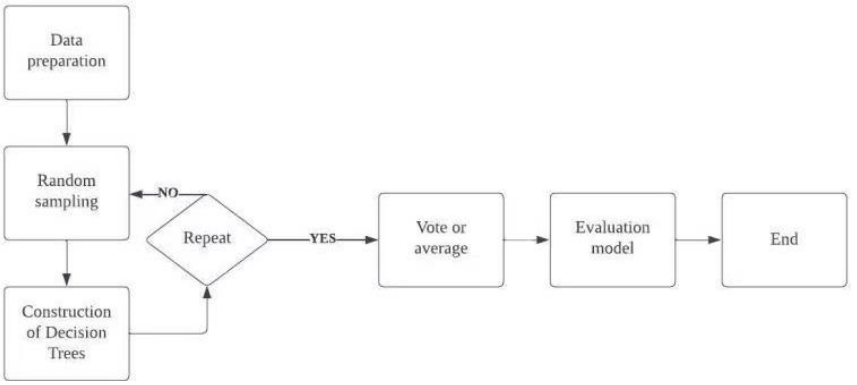


Figure 5. Workflow of RF.

2.3.3. GBDT. GBDT primarily utilizes decision trees as its base learner [12]. A decision tree is a type of non-parametric learner that can flexibly fit complex data relationships. The core idea is to progressively improve the model by fitting the residuals (the variance between the projected and observed values) of the previous round's model. During each iteration, a novel decision tree is trained to effectively capture the residuals of the preceding model. Each decision tree has a weight, and by ensemble-weighting the predictions of multiple trees, the final prediction is obtained. The process is illustrated in Figure 6.

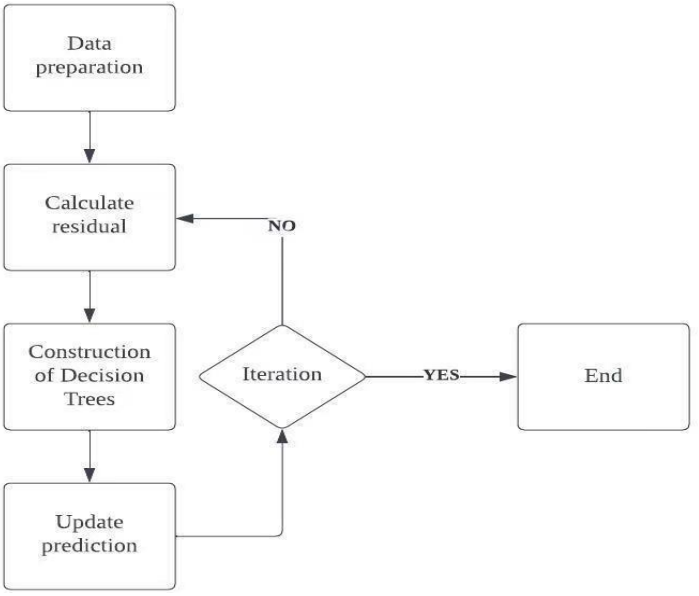


Figure 6. Workflow of GBDT.

2.3.4. XGBoost. XGBoost's main feature is that it iteratively trains a series of decision trees and gradually integrates them to improve the model's performance [13]. It introduces regularization terms,

DOI: 10.54254/2755-2721/45/20241122

including L1 (Lasso) and L2 (Ridge) regularization, to control the model's complexity and reduce the risk of overfitting. XGBoost can calculate the importance of each feature to the model and has a built-in cross-validation function, which helps to conveniently select the best hyperparameters. XGBoost was created to support the performance of machine learning models and to improve computing speed. XGBoost is a highly accurate and scalable GBDT implementation. The process is illustrated in Figure 7.

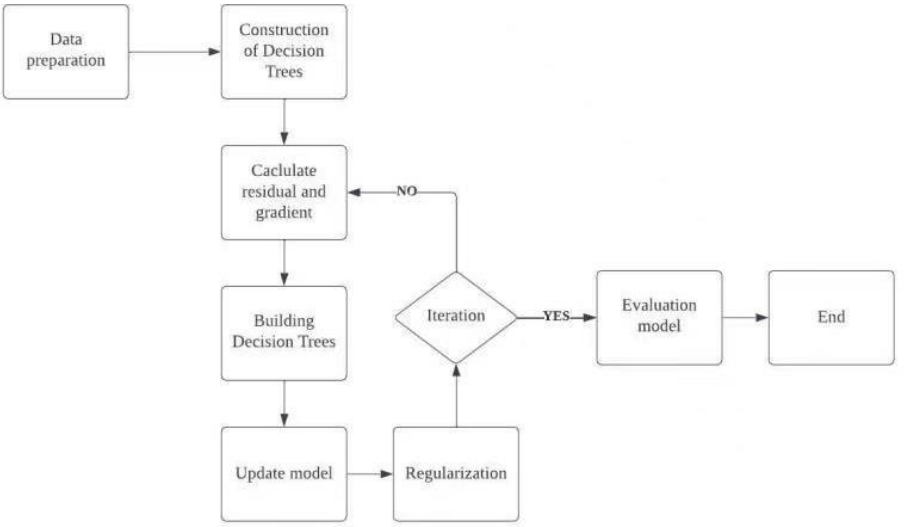


Figure 7. Workflow of XGBoost.

2.3.5. DNN. A DNN (Deep Neural Network) is an extension of neural networks, characterized by its multi-layer hidden layer structure, and is used to address complex machine learning and artificial intelligence tasks [14]. DNN uses multi-layer non-linear transformations to learn representations of features from raw data, allowing for the extraction of higher-level abstract features layer by layer.

It mainly comprises three structural layers:

Input Layer: Receives raw data features as input. Each neuron corresponds to one feature, and the amount of those in the input layer is equal to the amount of features.

Hidden Layers: One or additional layers of neurons positioned inside the input and output layers, tasked with gradually extracting abstract features. Each hidden layer comprises numerous neurons.

Output Layer: Produces the final prediction result. Each neuron in the network receives input signals from the previous layer's neurons through weighted connections, comparing the weighted sum of these received signals with a threshold. An activation function then generates the output signal. The network is trained using the error backpropagation (BP) algorithm by adjusting weights to minimize prediction errors. Figure 8 describes an example of a network.

DOI: 10.54254/2755-2721/45/20241122

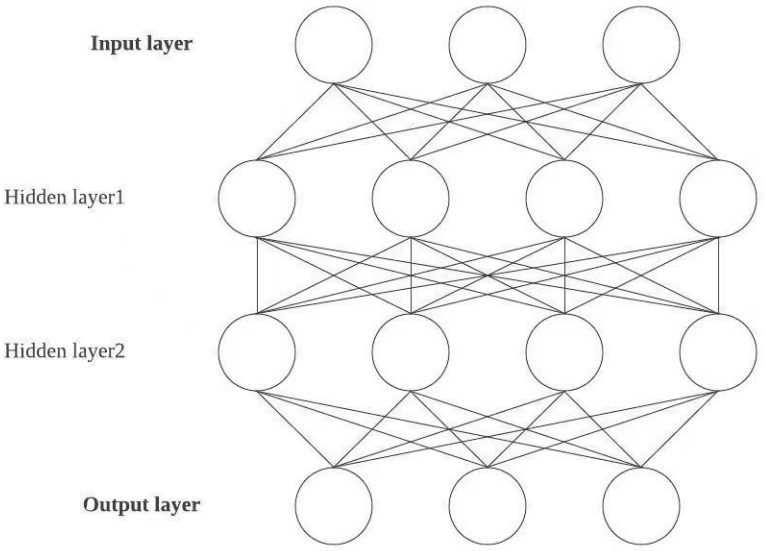


Figure 8. Example of DNN.

3. Research results and analysis

In this chapter, we determined the performance evaluation function, then conducted data experiments, and then made a second selection of the model based on the experimental findings, and then used the data augmented data set for comparison.

3.1. Model selection and construction

For the final model evaluation, this paper adopted five aspects for assessment, including Precision, Accuracy, F1-score, Recall, and Area Under Curve (AUC), which are shown in table 2.

Beforehand, it's essential to understand several concepts:

True Positive (TP): The count of positive instances correctly predicted as positive.

True Negative (TN): The count of negative instances correctly predicted as negative.

False Positive (FP): The count of negative instances incorrectly predicted as positive.

False Negative (FN): The count of positive instances incorrectly predicted as negative.

Table 2. Evaluation Function.

$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ ……(6)	Accuracy denotes the ratio of correctly predicted samples by the model to the total sample count.
$Precision = \frac{TP}{TP+FP}$……(7)	Precision signifies the ratio of genuinely positive samples among those predicted as positive by the model.
$Recall = \frac{TP}{TP+FN}$……(8)	Recall indicates the ratio of true positives that the model accurately identifies as positives
$F1-score = \frac{2}{\frac{1}{Recall}+\frac{1}{Precision}}$……(9)	F1-score takes into account both the precision and recall of the model

DOI: 10.54254/2755-2721/45/20241122

3.2. Feature selection experiment

3.2.1. Experimental result. In the dataset discussion, we attempted to use BMI as a replacement for Height and Weight, and Year in place of Day. Four combinations were explored: "Day + Height/Weight", "Day + BMI", "Year + Height/Weight", and "Year + BMI". By comparing these combinations under different models, the best current result was obtained with the "Day + Height/Weight" combination, exemplified using DNN (Figure 9).

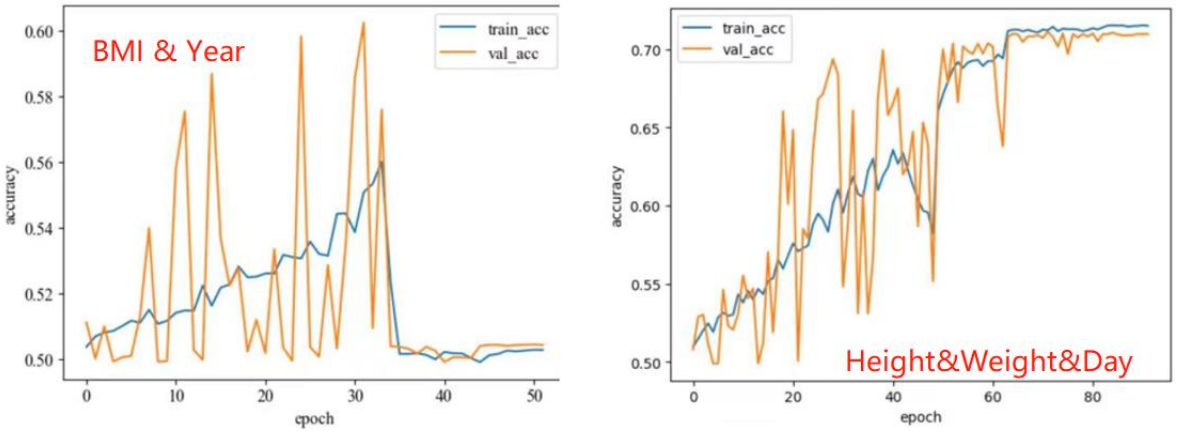


Figure 9. Features Comparison.

3.2.2. Experimental analysis. Analyzing the experimental results, the accuracy of Height, Weight, and day is above 70%, while the accuracy of BMI and year eventually converges to around 50%. This indicates that it is unreasonable in some articles to directly use them without discussion.

The reduction in features first leads to information loss. Height and Weight, as independent features, provide more information. Replacing them with BMI could result in information loss since BMI is derived from the combination of height and weight. Height and Weight are, to some extent, independent features, while BMI is essentially a function of them, potentially introducing feature redundancy. The model might be more sensitive to the relationship between height and weight, overlooking the information from BMI as a derivative feature. If the model assumes a linear relationship between features, using BMI might be inappropriate. The computation of BMI is fundamentally linear, but the relationship between height and weight might be nonlinear.

Additionally, replacing day with year also results in decreased performance. Considering granularity, which refers to the level of detail or refinement of data or information. In terms of time features, granularity denotes the size of the time intervals or the intervals between data points. A finer granularity means smaller time intervals, enabling capturing more detailed temporal changes; while a coarser granularity implies larger time intervals, offering a broader temporal overview [15].

3.3. Data experiment

3.3.1. Experimental results. In the experiments on the dataset before data augmentation, the hyperparameters of the DNN were adjusted. A three-layer structure was employed: the input layer had 128 neurons, one hidden layer with 32 neurons, and the output layer consisted of a single neuron (Table 3-4).

DOI: 10.54254/2755-2721/45/20241122

Table 3. Results before data augmentation.

Model Name	Accuracy	Precision	Recall	Fl-score	AUC
KNN	72.0	71.2	66.2	68.6	76
RF	73.0	73.3	72.2	72.7	79
XGBoost	74.5	74.5	72.5	73.4	82
GBDT	74.0	74.1	71.3	72.6	81
DNN	75.0	75.0	72.4	73.6	81

Subsequently, model comparisons were conducted on the dataset after data augmentation. Meanwhile, the DNN network was optimized by adding an additional hidden layer with 64 neurons.

Table 4. Results after data augmentation.

Model Name	Accuracy	Precision	Recall	Fl-score	AUC
KNN	80.1%	79.8%	82.2%	82.1%	79.9%
RF	83.5%	82.1%	83.3%	83.2%	81.2%
XGBoost	83.0%	81.8%	82.2%	81.9%	82.3%
GBDT	84.5%	82.3%	83.3%	82.7%	82.5%
DNN	85.0%	83.1%	84.1%	83.5%	83.2%

3.3.2. Experimental analysis. From the results, it's evident that ensemble machine learning models have an advantage over traditional machine learning models. Ensemble machine learning is capable of achieving satisfactory performance on medium and small scale data in a relatively short time. In this task, the performance difference between ensemble machine learning and deep learning is not significant, with DNN having a slight edge overall. As the volume of data increases, DNN displays superior performance. DNNs have a vast number of parameters and, through the continual addition of hidden layers, they can learn and represent complex functions, processing intricate, high-dimensional data to better capture inherent patterns. Conventional machine learning approaches frequently necessitate manual feature engineering, whereas DNNs can automatically learn high-level and abstract features from raw data. This automatic feature extraction makes them especially remarkable on large datasets. DNNs allow for end-to-end training, which means direct learning from raw input to the final output, eliminating the need for intermediate steps. This approach enables the model to learn the optimal feature representations directly from the data.

4. Conclusion

This paper integrates machine learning and deep learning to find a precise method to assist in predicting cardiovascular diseases. The models used include KNN, RF, XGBoost, GBDT, and DNN. Among all these models, the DNN ensemble learning method showed the best auxiliary prediction effect for cardiovascular diseases. During the construction of the DNN model, two layers of neurons were employed. After training the model, data visualization yielded satisfactory results: the DNN's Accuracy is 85% and AUC is 83.2%. This shows a clear advantage over traditional machine learning. It can also be observed that XGBoost's Accuracy is 83%, with an AUC of 82.3%, indicating that ensemble learning algorithms obviously have a better accuracy rate. Another advantage of ensemble learning methods is their ability to evaluate the relative importance of each attribute during the training process, which leads to more accurate predictions. However, issues like data imbalance and the risk of overfitting remain. In the experimental part, the SMOTE method was used for data augmentation. With the enhancement of the dataset and continuous parameter adjustments, the DNN demonstrated increasingly improved performance.

Although the DNN ensemble learning displayed promising results, the AUC still stands at 83.2%, indicating room for improvement. It was also observed that features without strong correlation posed

DOI: 10.54254/2755-2721/45/20241122

concerns related to granularity and information loss due to reduced feature numbers, leading to decreased prediction performance. In future research, a more relevant dataset can be sourced. Moreover, a better DNN model can be constructed for further optimization, aiming for a more precise auxiliary prediction of cardiovascular diseases

Authors Contribution

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

References

[1] Lippi G, Henry B M, Sanchis-Gomar F. Physical inactivity and cardiovascular disease at the time of coronavirus disease 2019 (COVID-19). *European journal of preventive cardiology*, 2020, 27(9): 906-908.

[2] Packard C, Chapman M J, Sibartie M, et al. Intensive low-density lipoprotein cholesterol lowering in cardiovascular disease prevention: opportunities and challenges. *Heart*, 2021, 107(17): 1369-1375.

[3] Bae S A, Kim S R, Kim M N, et al. Impact of cardiovascular disease and risk factors on fatal outcomes in patients with COVID-19 according to age: a systematic review and meta-analysis. *Heart*, 2021, 107(5): 373-380.

[4] Sun W, Zhang P, Wang Z, et al. Prediction of cardiovascular diseases based on machine learning. *ASP Transactions on Internet of Things*, 2021, 1(1): 30-35.

[5] Nielsen R E, Banner J, Jensen S E. Cardiovascular disease in patients with severe mental illness. *Nature Reviews Cardiology*, 2021, 18(2): 136-145.

[6] Farmer R E, Mathur R, Schmidt A F, et al. Associations between measures of sarcopenic obesity and risk of cardiovascular disease and mortality: a cohort study and Mendelian randomization analysis using the UK Biobank. *Journal of the American Heart Association*, 2019, 8(13): e011638.

[7] Nielsen M W, Stefanick M L, Peragine D, et al. Gender-related variables for health research. *Biology of Sex Differences*, 2021, 12(1): 1-16.

[8] Ishaq A, Sadiq S, Umer M, et al. Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. *IEEE access*, 2021, 9: 39707-39716.

[9] Singh D, Singh B. Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, 2020, 97: 105524.

[10] Nawaz M S, Shoaib B, Ashraf M A. Intelligent cardiovascular disease prediction empowered with gradient descent optimization. *Heliyon*, 2021, 7(5).

[11] Yang L, Wu H, Jin X, et al. Study of cardiovascular disease prediction model based on random forest in eastern China. *Scientific reports*, 2020, 10(1): 5245.

[12] Mary M M A, Beena T L A. Heart disease prediction using machine learning techniques: A survey. *Int. J. Res. Appl. Sci. Eng. Technol*, 2020, 8(10): 441-447.

[13] Srinivas P, Katarya R. hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost. *Biomedical Signal Processing and Control*, 2022, 73: 103456.

[14] Manaa A, Brahimi F, Chouiref Z, et al. Cardiovascular Diseases Prediction Based on Dense-DNN and Feature Selection Techniques, International Symposium on Modelling and Implementation of Complex Systems. *Cham: Springer International Publishing*, 2022: 333-347.

[15] Cirillo D, Núñez-Carpintero I, Valencia A. Artificial intelligence in cancer research: learning at different levels of data granularity. *Molecular Oncology*, 2021, 15(4): 817-829.