Deep Neural Network based Heart Disease Prediction

Bixi Zhang

School of Physics, Nanjing University, Jiangsu Nanjing, 210093, China
652022220082@smail.nju.edu.cn

Abstract. Since the 21st century, cardiovascular disease (CVD), which has a high rate of morbidity and mortality, has become one of the most prevalent and fatal diseases. Therefore, machine-learning and deep-learning-based heart disease prediction models hold substantial research value in the medical field with extensive potential applications and great significance. This article aims to establish a prediction model, utilizing the Deep Neural Network (DNN) algorithm, to cope with latent hazards associated with heart disease. A dataset from the Kaggle database was used for this approach. To assess the performance of the prediction model, evaluation indices including accuracy, recall_0, recall_1, and AUC, etc. were calculated for DNN and five comparative machine-learning algorithms. DNN eventually achieved outstanding performance with an accuracy of 0.76, recall_1 rate of 0.77, and AUC value of 0.84, respectively. The study concluded that DNN showed better average performance compared to other machine learning algorithms. The result could serve as an auxiliary strategy for heart disease diagnoses.

Keywords: Heart Disease, Deep Neural Network, Machine learning, Feature selection, Imbalance learning.

1 Introduction

Since the 21st century, cardiovascular diseases (CVDs) have consistently been one of the deadliest illnesses, leading to massive fatalities worldwide. Heart disease is the critical branch of cardiovascular disease. According to Megan Lindstrom et al., the global all-age mortality rate from heart disease in 2021 is 258.8 per 100,000 people, and there is a prevalence of up to 7852.0 per 100,000 people [1]. With the increase in the modern standard of living, there is a growing emphasis on improving quality of life due to socio-economic and cultural prosperity. Heart disease could pose a serious potential threat to health, resulting in a dramatic decline in quality of life when it occurs to individuals. Therefore, establishing a systematic strategy to predict heart disease would allow earlier treatments for potential patients with medication or surgery, thus reducing its serious consequences.

In recent decades, owing to advancements in computer processing power, AI has made significant progress in disease prediction, particularly in heart disease, skin disease, cancer, etc. [2]. Meanwhile, the open-source databases provided by the Internet facilitate the training of artificial intelligence models. Numerous machine

Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_70

learning and deep learning models for heart disease prediction based on cardiology datasets such as Cleveland and IEEE have already been established and exhibited [3]. These models are relatively well-performing.

Electrocardiograms (ECGs) can provide valuable information when using artificial intelligence to predict heart disease [4]. Y. M. Ayano et al. provided a targeted review of ECG-based heart disease prediction focusing on Interpretable Machine Learning (IML) [5]. E. H. A.-E. Rabie Ahmed et al. overviewed the diagnosis and classification of heart failure. Machine learning strategies comprising supervised, unsupervised, and reinforcement learning were exploited [6]. J. Botros et al. realized a CNN-based classification method for ECGs and applied the model to heart failure prediction, achieving 99.31% accuracy, 99.50% sensitivity, and 99.11% specificity [7]. I. M. El-Hasnony et al. utilized the Cleveland database and implemented five active learning strategies to train and predict heart disease in a manner different from traditional supervised learning. However, the final performance was unsatisfactory, with an accuracy of 57.4% ± 4% and an F-score of 62.2% ± 3.6% [8]. C.-Y. Guo et al. developed a prediction model for heart failure using four machine-learning models. They also applied, for instance, the KNN clustering algorithm to classify the samples and enhance the performance of their models. Despite its high level of professionalism, the dataset from the JHS database, which was applied for machine learning contained an excessive number of variables (over 100), which hindered the practical application of their models [9].

The majority of previous research centered their view on specialized medical data, such as ECG data, resting blood pressure (restbps), plasma cholesterol level (chol), fasting blood sugar (fbs), etc., which can be difficult to obtain, constituting an obstacle to the practical application. Moreover, variables incorporated in the datasets applied by these studies are generally involved, making the algorithm intricate and slow. Overall, it's crucial to develop a prediction model for heart disease that is accessible for daily use.

This paper intends to develop a prediction model for heart disease with accessible variables, to streamline heart disease prediction in daily life. A novel dataset (2022) from Kaggle with 17 variables was employed for this approach. The model was based on Deep Neural Network (DNN) algorithm, and five machine learning algorithms were analyzed for comparison. Feature screening and unbalanced strategies were adopted for performance optimization. The evaluation results were presented separately in section 3.

2 Dataset And Methods

2.1 Dataset

The Dataset in this research was chosen from the Kaggle database. 319795 instances were included in the total. A yes/no approach was used to record whether or not there had been a cardiac heart disease (CHD or MI). Seventeen relevant variables, whose specific meanings and values are shown in Table 1, were concerned. It should be noted that the number of instances counted as "yes" (Potential patient for heart

disease) in the dataset was 27.4k, merely 9% of the total samples. While those recorded as "no" (Not a potential patient for heart disease) were 292k, about 91% of the total samples. The dataset consequently presented a tremendous imbalance essence.

Before training, this paper used the Label Encoder tactic to convert binary and discrete "object" values into "numeric" values. Subsequently, the dataset was split into two segments, with a ratio of 7:3 for training and testing, respectively.

Table 1. Introduction of Variables

Variables	Contents	Type	Values
BMI	Body Mass Index	discrete	23, 24, 25
Smoking	Have you ever smoked at least 100 cigarettes in your entire life	binary	0, 1
AlcoholDrinking	Are you a heavy drinker	binary	0, 1
Stroke	Have you had a stroke or not	binary	0, 1
PhysicalHealth	For how many days during the past 30 do you have physical health problems	discrete	1--30
MentalHealth	For how many days during the past 30 do you have mental health problems	discrete	1--30
DiffWalking	Do you have serious difficulty walking	binary	0, 1
Sex	Your gender	binary	0, 1
AgeCategory	Your age	discrete	1, 2, 3
Race	Your race	discrete	1, 2, 3, 4
Diabetic	You had diabetes or not	discrete	1, 2, 3, 4
PhysicalActivity	Were you doing regular exercise during the past 30 days	binary	0, 1
GenHealth	Your general health condition	discrete	1, 2, 3, 4
SleepTime	Your average sleep hours	discrete	1, 2, 3
Asthma	You had asthma or not	binary	0, 1
KidneyDisease	You had kidney disease or not	binary	0, 1
SkinCancer	You had kidney disease or not	binary	0, 1

2.2 Methods

Deep Neural Network. Fully connected deep neural networks (DNNs) are neural network algorithms based on multi-layer perceptrons (MLPs). They offer higher accuracy and more powerful nonlinear fitting capabilities than traditional machine learning algorithms. DNNs are capable of creating extremely accurate decision boundaries in binary classification problems. While fully connected deep neural networks are prone to overfitting and require a long training period, they could still be a good choice for the prediction algorithm in this study due to simple variables and the small data volume of the selected dataset.

The structure of the neural network is depicted in Fig. 1. This article designed a network with 2 hidden layers containing 64 and 32 neurons, respectively. Unlike traditional DNNs, the neural network in this paper was deliberately designed as "wide and shallow". The structure aimed at striking a balance between comprehensive extraction of features and reducing overfitting. To further reduce overfitting, a ReLU function (refer to Fig. 2) with sparse activation property were chosen as the hidden layer activation function. The Sigmoid function $S(x)$ was used as the output layer function:

$$S(x) = \frac{1}{1 + e^{-x}} \#(1)$$

The loss function chosen for the neural network was the cross-entropy function $J(p, y)$:

$$J(p, y) = -\sum (y \cdot \log p + (1 - y) \log (1 - p)) \#(2)$$

Whereas y represents the real value (in this study, only 1 or 0 was taken, representing diseased and not diseased, respectively), p represents the prediction probability generated from the neural network (a value between 0 and 1). The Adam optimizer was used as a gradient descent rule [10].

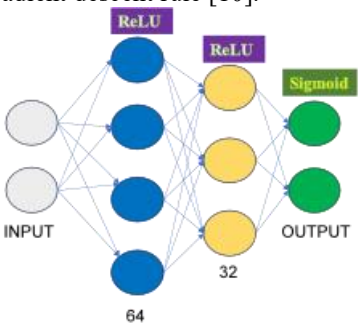

Fig. 1. Deep Neural Network(Photo/Picture credit : Original)

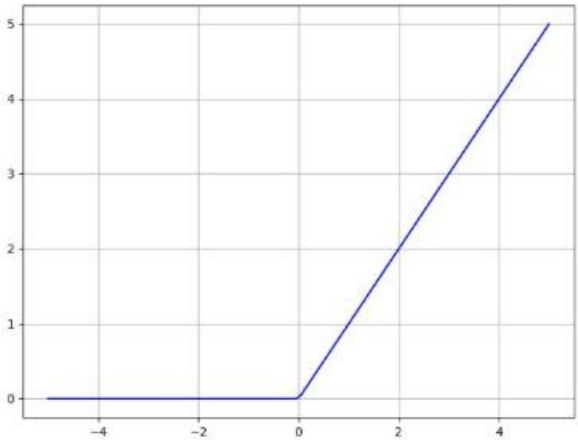

Fig. 2. ReLU Function(Photo/Picture credit : Original)

Comparative Machine Learning Methods. To showcase the effectiveness of fully connected deep neural networks, the study selected five machine learning models to compare their performance with the former algorithm. For those five machine learning models applied, their features and advantages are listed below:

Logistic Regression is fast classification with a small computation volume for quick results. Excellent fit for linear binary classification problems.

Decision Tree is white-box and highly interpretable. Skilled at handling non-linear problems, but susceptible to overfitting.

In contrast to the decision tree model, Random Forest is highly resistant to interference and provides an effective method to mitigate errors when dealing with unbalanced datasets. It significantly outperforms the decision tree in unbalanced datasets.

XGB and GBDT are integrated learning methods with strong robustness and have performed well in AI algorithm competitions in recent years. This study selected them for comparison.

Evaluation Indices. To evaluate the prediction model's performance, mainly four indices were applied, including accuracy, recall_0, recall_1, and AUC. The definitions are given below, respectively:

Accuracy: The percentage of samples that were predicted correctly out of the total number of predictions.

Recall_0: Accuracy among negative samples.

Recall_1: Accuracy among positive samples.

AUC: Area under the ROC curve.

In addition, the Pearson correlation coefficient was used to estimate feature importance. Pearson correlation coefficient is defined as:

$$P(x, y) = \frac{Cov(x, y)}{\quad} \#(3)$$

Cov (*x, y*) represents the covariance between variables *x* and *y*. Variances of the variable *x* and *y* are indicated by σ_x and σ_y, respectively.

3 Result Analysis

3.1 Feature Analysis

Data Visualization. With data visualization, this paper examined the credibility of the dataset. Fig. 3, and Fig. 4 gives some of the visualization results. Fig. 5 shows the linear correlation thermogram as a pre-analysis of the data.

Fig. 3 shows the visualization of the variable "Sleep Time" as an example of analysis. The dataset is evenly and adequately sampled for all age groups. The rate of heart disease increases nearly linearly with age, thus inferring a positive correlation between age and prevalence. The same conclusion can also be deduced from the results shown in Fig.5. The correlation coefficient between age and prevalence is 0.23, which is the highest among all variables.

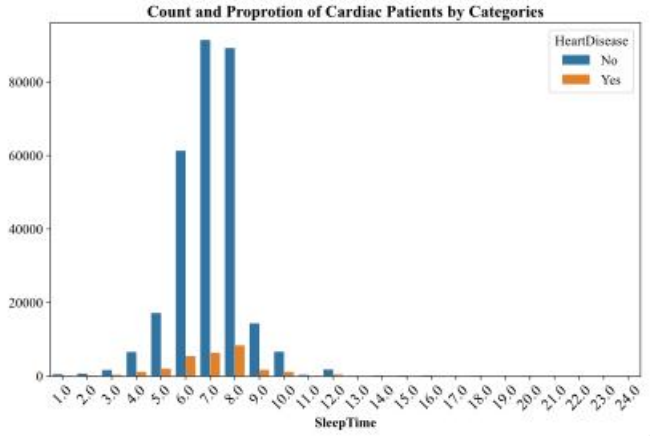

Fig. 3. Count and Proportion of Cardiac Patients by Sleep Time(Photo/Picture credit : Original)

Similarly, it can be deduced through Fig. 4 “Count and Proportion of Cardiac Patients by Gender” that the percentage of males having suffered from heart disease is significantly greater than that of females, which accords with the findings of Walli-Attaei M et al. [11]. In Fig. 5 “Correlation Heatmap”, the variables that strongly correlate with the prevalence of heart disease were identified with red circles. A history of skin cancer, kidney disease, stroke, etc. could increase the chances of heart disease, the habit of smoking could lead to an increase in heart disease, and obesity is also a possible factor that increases its prevalence. Meanwhile, the black circles in Fig. 5 mark the variables with strong inner correlations, which lays the foundation for

dimensionality reduction methods such as feature screening.

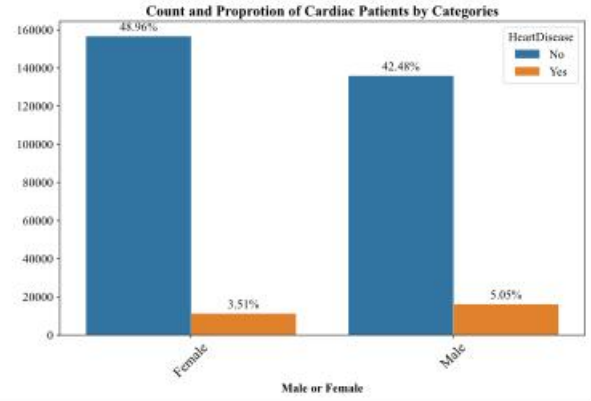

Fig. 4. Count and Proportion of Cardiac Patients by Gender(Photo/Picture credit : Original)

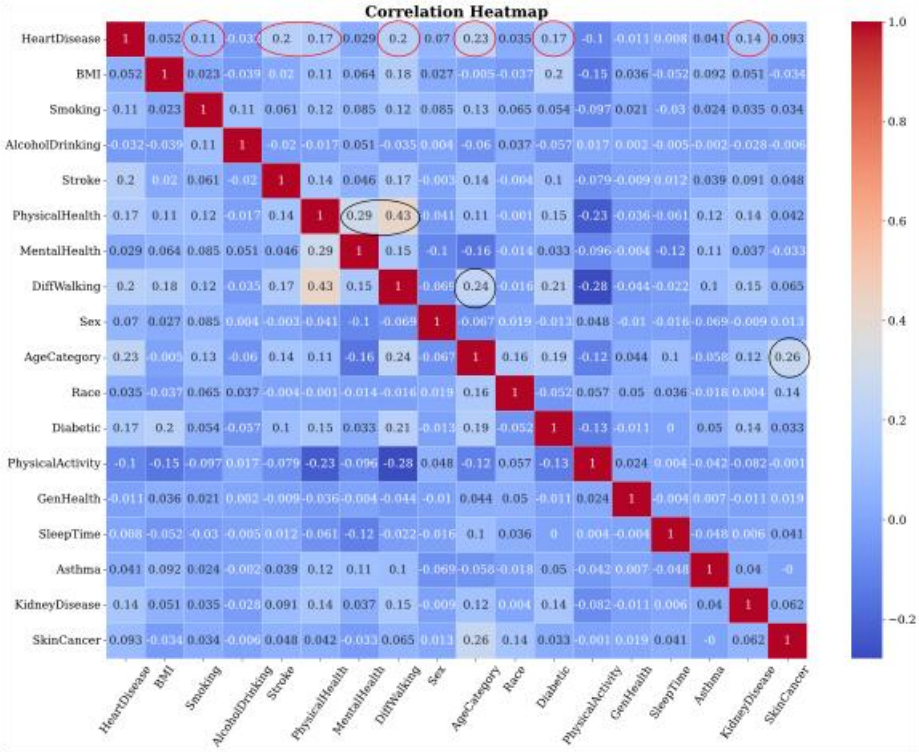

Fig. 5. Correlation Heatmap(Photo/Picture credit : Original)

Feature Selection. Variable screening using a correlation coefficient (Pearson correlation coefficient) larger than 0.05 resulted in eleven variables with high correlation rates. Fig. 6 shows the comparative performance of the five machine learning models before and after feature selection, where AUC is employed as the evaluation index. Except for the logistic regression model, which has a slight increase in AUC (marked with a red circle in Fig. 6), the rest of the models appear to have a different magnitude of decline after feature screening. This might have resulted from the linear essence of the logistic regression model, corresponding to the Pearson correlation coefficient. Screening out redundant variables with low linear correlation could help to improve the predictive performance of linear classification models. In contrast, the Decision Tree and Boosting integrated learning algorithms with a nonlinear nature performed poorly, as linear feature screening could have instead caused their main information to be discarded, resulting in performance degradation.

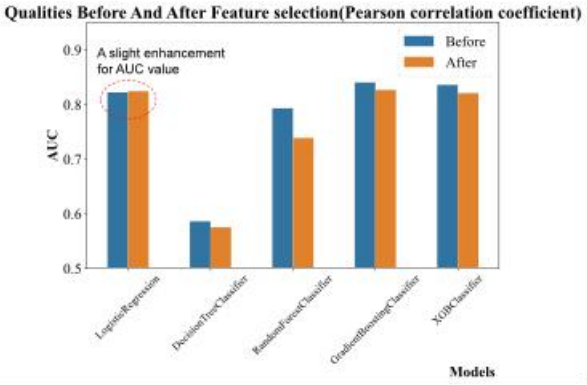

Fig. 6. Performance Before and After Feature Selection(Photo/Picture credit : Original)

3.2 Initial Results

Fig. 7 presents the assessment results described by the four evaluation indices for DNN and five machine learning models. All models exhibited high accuracy and recall_0 rate, with a relatively low recall_1 rate (marked with a black rectangle in Fig.7). For DNN model, Recall_1 rate was merely 0.06.

Since the total number of category 1 (sick) samples was only 1/10th of the category 0 (healthy) samples, all models tended to give higher weight to the category 0 results. Thus, the imbalanced nature of the dataset could have affected the Recall_1 rate tremendously. However, Recall_1 rate is a crucial metric for disease prediction. Hence, imbalanced strategies must be utilized to enhance the performance of predictive models.

In Fig. 7, the decision tree model exhibited an exclusively lower AUC value compared to the others. A special analysis was conducted to gain better insight into the problem. Fig. 8 shows the cross-validation results for the decision tree model. The portion of the depth above 10 produced large splits in the training and test set results;

presumably the severe overfitting phenomenon was the cause of its low performance.

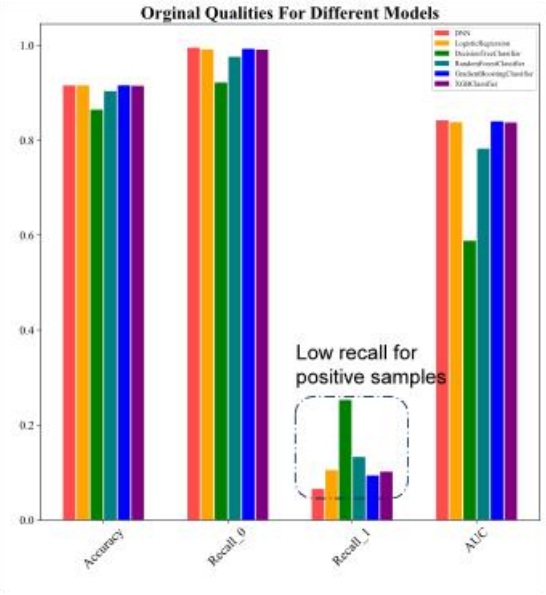

Fig. 7. Original Model Performance(Photo/Picture credit : Original)

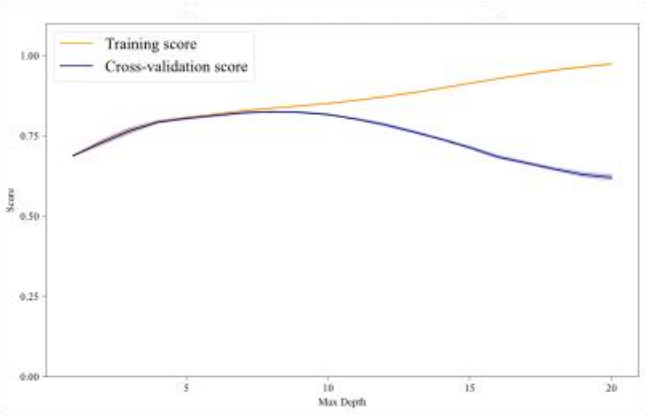

Fig. 8. Cross-validation Curve of Decision Tree Classifier(Photo/Picture credit : Original)

3.3 Imbalance Learning Strategy

This research employed random oversampling, random undersampling, and SMOTE techniques to handle imbalanced datasets in five machine learning models. For random oversampling, class 1 samples were increased to five times, hence the ratio of

class 0 to class 1 samples in the training set was about 2:1 after processing. For random undersampling, the class 0 samples were reduced by 2/3, so the ratio of class 0 to class 1 samples was eventually about 3:1. For the SMOTE strategy, there was no control over the exact ratio of oversampling and undersampling.

The performance of the processed model is shown in Fig. 9, 10, and 11. All three strategies reduced the accuracy of the five models. After implementing imbalanced strategies, the AUC values of four models apart from the decision tree model, declined. By contrast, Recall_1 rates were improved to varying degrees for all five models. Notably, Recall_1 rate of the logistic regression model saw the most significant increase by 82.8% after the SMOTE strategy. It can be concluded that the imbalanced strategies improved the effectiveness of class 1 predictions by sacrificing the accuracy of class 0 predictions.

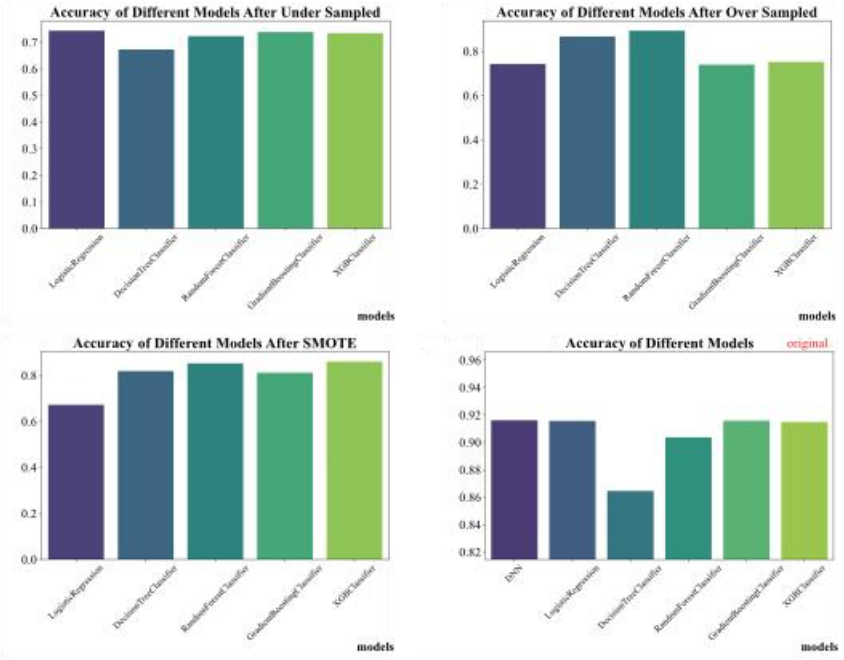

Fig. 9. Accuracy Comparison Before and After Imbalance Learning(Photo/Picture credit : Original)

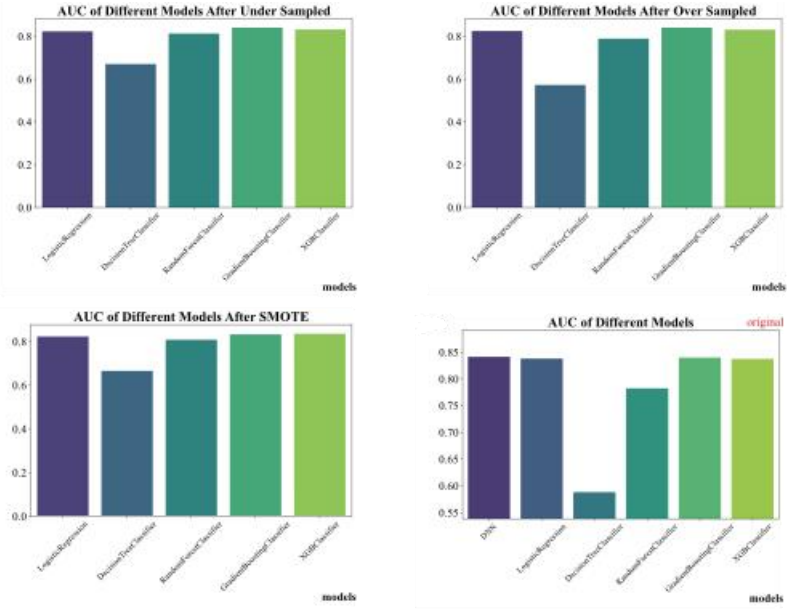

Fig. 10. AUC Comparison Before and After Imbalance Learning(Photo/Picture credit : Original)

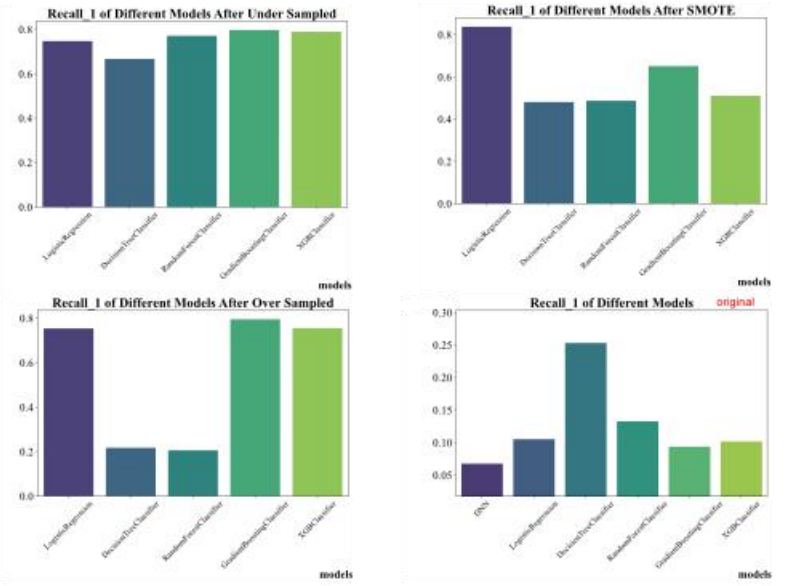

Fig. 11. Recall_1 Comparison Before and After Imbalance Learning(Photo/Picture credit : Original)

Besides, this study employed imbalanced strategies for DNN model and compared the results with an integrated learning strategy. For DNN model, two methods were utilized. On the one hand, F1_Score, AUC, and Recall metrics were monitored through training. On the other hand, the decision threshold was reduced from 0.5 to 0.1 to increase the weight of class 1 samples. Fig. 12 shows the results of the integrated learning strategy and DNN model after imbalanced strategies. Two models were compared under each evaluation metric. Fig. 12 shows that DNN resulted in slightly higher evaluation metrics than the integrated learning strategy except for Recall_1 and Precision_0. The DNN model achieved an accuracy of 0.76, an AUC value of 0.84, and a Recall_1 rate of 0.77 after applying the imbalanced strategies, owning the best overall performance. Whereas machine learning models such as logistic regression models performed worse in terms of accuracy (see Fig. 9), and tree models generally had poor Recall_1 performance (see Fig. 11). It is hypothesized that linear models are prone to overfitting class 1 samples after imbalanced strategies, and tree models are consistently more prone to overfitting class 0 samples. Both have shortcomings in some aspects.

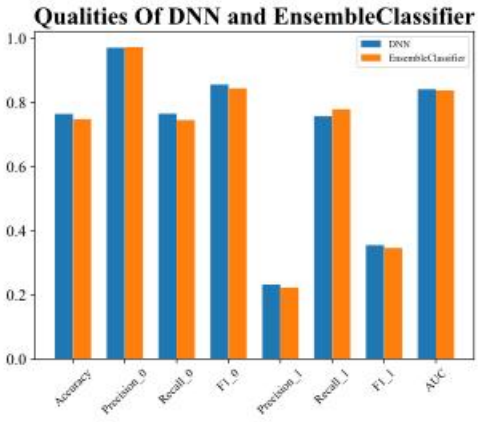

Fig. 12. Performance Comparison Between EnsembleClassifier and DNN After Imbalance Learning(Photo/Picture credit : Original)

4 Conclusion

In this article, heart disease prediction was performed by deep neural network DNN using the dataset from Kaggle. The results were compared with five machine learning models. The DNN eventually reached a prediction accuracy of 0.76, an AUC of 0.84, and a Recall_1 rate of 0.77 utilizing feature filtering and imbalanced strategies. The Recall_1 rate had been significantly improved compared to the initial value (0.06), achieving an ideal predictive result for heart disease. In contrast, machine learning models behaved lower individual evaluation metrics due to overfitting problems (e.g., logistic regression models tended to overfit class 1 samples after applying imbalanced

strategies, and tree models consistently tended to overfit class 0 samples). The results of this paper can assist people in detecting heart disease. However, due to the black-box characteristics of machine learning and deep neural network algorithms, this paper did not conduct an in-depth theoretical study. The enhancement introduced by the imbalanced strategies can be systematically and theoretically analyzed in the future.

Reference

1. Lindstrom M., DeCleene N., Dorsey H., Fuster V., Johnson C.O., LeGrand K.E., Mensah G.A., Razo C., Stark B., Varieur Turco J., Roth G.A.: Global Burden of Cardiovascular Diseases and Risks Collaboration, 1990-2021. J Am Coll Cardiol. (2022).
2. Kumar Y., Koul A., Singla R., Ijaz M.F.: Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J Ambient Intell Humaniz Comput. (2023).
3. Chandrasekhar N., Peddakrishna S.: Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization. Processes. (2023).
4. Somani S., Russak A.J., Richter F., Zhao S., Vaid A., Chaudhry F., De Freitas J.K., Naik N., Miotto R., Nadkarni G.N., Narula J., Argulian E., Glicksberg B.S.: Deep learning and the electrocardiogram: review of the current state-of-the-art. Europace. (2021).
5. Ayano Y.M., Schwenker F., Dufera B.D., Debelee T.G.: Interpretable Machine Learning Techniques in ECG-Based Heart Disease Classification: A Systematic Review. Diagnostics (Basel). (2022).
6. Eman H.A., Rabie A.: Clinical Applications of Machine Learning in the Diagnosis, Classification and Prediction of Heart Failure. Journal of Electrical Systems. (2024).
7. Botros J., Mourad-Chehade F., Laplanche D.: CNN and SVM-Based Models for the Detection of Heart Failure Using Electrocardiogram Signals. Sensors (Basel). (2022).
8. El-Hasnony I.M., Elzeki O.M., Alshehri A., Salem H.: Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction. Sensors (Basel). (2022).
9. Guo C.Y., Wu M.Y., Cheng H.M.: The Comprehensive Machine Learning Analytics for Heart Failure. Int J Environ Res Public Health. (2021).
10. Kingma D., Ba J.: Adam: A Method for Stochastic Optimization. Computer Science. (2014).
11. Walli-Attaei M., Rosengren A., Rangarajan S., Breet Y., Abdul-Razak S., Sharief W.A., Alhabib K.F., Avezum A., Chifamba J., Diaz R., Gupta R., Hu B., Iqbal R., Ismail R., Kelishadi R., Khatib R., Lang X., Li S., Lopez-Jaramillo P., Mohan V., Oguz A., Palileo-Villanueva L.M., Poltyn-Zaradna K., R S.P., Pinnaka L.V.M., Serón P., Teo K., Verghese S.T., Wielgosz A., Yeates K., Yusuf R., Anand S.S., Yusuf S.: Metabolic, behavioural, and psychosocial risk factors and cardiovascular disease in women compared with men in 21 high-income, middle-income, and low-income countries: an analysis of the PURE study. Lancet. (2022).