

Prediction of Heart Failure Occurrence Based on the Categorical Boosting Model

Yinan Gao ^a

International School, Beijing University of Posts and Telecommunications, No.10 Xitucheng Road, Beijing, China

Keywords: Medical Prediction, Machine Learning, Deep Learning, Heart Failure.

Abstract: Cardiovascular disease (CVD) is the world's greatest cause of mortality, taking an estimated 17.9 million lives annually and contributing 31% of all fatalities worldwide. Heart failure is a prevalent CVD-related occurrence and has a five-year survival rate similar to malignant tumors, at around 50%. Therefore, the prevention and advanced Interfere treatment are the key to current research. This study aims to predict heart failure occurrence based on various indicators of patients' physical health by constructing a Categorical Boosting machine learning model. The final trained model achieved a prediction accuracy of 88.13%, which fully validates the feasibility of using this model for practical heart failure prediction. Therefore, the primary focus of this research is to continue optimizing this model in the future, promote clinical validation, and facilitate its practical application. By identifying more potential patients, conducting early diagnosis and treatment, and effectively reducing the incidence of heart failure disease, we aim to realize the actual application of machine learning technology in the medical field.

1 INTRODUCTION


Heart failure (HF) is a frequent occurrence brought on by a number of cardiovascular conditions. If cardiovascular patients develop HF, their five-year survival rate is similar to that of malignant tumors (Zhong, 2007). Even if successfully treated, HF can lead to high rates of readmission and mortality, imposing a heavy burden on patients, their families, and society.

Therefore, preventing and early prediction of HF occurrence are crucial for patients and their families. Early intervention and treatment can greatly increase the five-year survival rate and treatment success rate for patients. By utilizing physiological data and advanced predictive models, it is possible to predict who is more likely to develop HF, enabling early intervention, treatment, and the formulation of better clinical management strategies, thereby reducing the risks borne by patients.

Currently, the prediction and diagnosis of HF mainly rely on medical history, physical examinations, laboratory tests, cardiac imaging examinations, and functional tests (Zhong, 2007). However, these conventional medical diagnoses are

not the optimal approach to addressing such a significant disease as HF. This is because patients often seek medical attention when their bodies have already shown unusual conditions, or even reached an intolerable level. For severe diseases, such manifestations may indicate rapid disease progression, ultimately leading to the development of untreatable conditions due to delayed treatment-seeking. Furthermore, HF is also a potentially acute and sudden-onset disease that often endangers the patient's life and requires immediate rescue. At this time, routine diagnostic testing clearly does not serve the purpose of prediction and prevention.

Therefore, the purpose of this study is to use deep learning and machine learning methods based on the Categorical Boosting model to predict the occurrence of HF based on many routine physiological signs (Vickers J, 2018). With the precise prediction of HF occurrence, potential patients can receive early treatment and care.

^a <https://orcid.org/0009-0002-3073-3400>

2 RELATED WORK

Research on the prediction of HF in the past has mainly focused on analyzing and modeling clinical indicators, biomarkers, imaging features, etc. Additionally, some researchers have investigated the application of statistical and machine learning models, like random forests, logistic regression, and support vector machines, to predict the occurrence and progression of HF (Song, 2023).

Early research on clinical indicators, biomarkers, and imaging features related to HF has been quite mature. Through professional detection indicators, relevant medical history analysis, and imaging examinations, accurate assessment of the patient's current condition can be achieved, leading to a diagnosis of HF. Plasma brain natriuretic peptide (BNP) measurement, radionuclide ventriculography, radionuclide myocardial perfusion imaging, two-dimensional echocardiography, and Doppler ultrasound can all help diagnose HF (Zhong, 2007). However, these detection indicators generally have a strong timeliness and are usually only used for clinical diagnosis. They cannot be directly applied to prediction, and inevitably, doctors' diagnoses may be subjective to some extent, leading to errors in judgment or misdiagnosis. This study aims to achieve objective and accurate judgments through machine models by analyzing relevant features of patients and predicting the occurrence of HF more accurately.

In recent years, research on HF based on machine learning models has been increasing gradually. This trend can be attributed to the rise of machine learning

technologies and the unique advantages of machine learning in model building and data analysis (Song, 2023; Bzdok, 2018). Some studies have applied machine learning in predicting HF, using models such as decision trees, support vector machines for training. By analyzing echocardiograms and related data, the HF incidence in asymptomatic individuals can be obtained (Strom, 2021; Kobayashi, 2021). Although most of these studies can achieve ideal prediction results from the trained models, the features studied are often highly specialized and not easily accessible in normal life, leading to high detection costs. This study aims to predict HF using more universally and easily obtainable physical data, reducing costs and making professional medical diagnosis more generalized and closer to people's daily lives.

3 METHOD

3.1 Data Setup

3.1.1 Data Set

"Heart Failure Prediction Dataset," which was chosen from the Kaggle website, is the dataset used in this study. Each record in this dataset represents observations of patients in a hospital setting. The dataset is a combination of five heart datasets from different hospitals, comprising a total of 918 samples across 11 common features (refer to Table 1).

Table 1: Sample Features.

	Feature Names	Measurement Content	Type	Value
1	Age	Age	Discrete type	1, 2, 3
2	Sex	Sex	Category type	M: Male, F: Female
3	ChestPainType	Chest Pain Type	Category type	TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic
4	RestingBP	Resting Blood Pressure	Discrete type	1, 2, 3
5	Cholesterol	Serum Cholesterol	Discrete type	1, 2, 3
6	FastingBS	Fasting Blood Sugar > 120 mg/dl	binary system	0, 1
7	RestingECG	Resting Electrocardiographic Results	Category type	Normal: Normal, ST: ST-T wave abnormality present, LVH: Possible or definite left ventricular hypertrophy

8	MaxHR	Maximum Heart Rate	Discrete type	1, 2, 3
9	ExerciseAngina	Exercise-Induced Angina	Category type	Y: Yes, N: No
10	Oldpeak	ST Segment Depression measured during Resting Electrocardiogram	Discrete type	1, 2, 3
11	ST_Slope	Trend of ST Segment Changes during Peak Exercise	Category type	Up: Upward, Flat: Flat, Down: Downward
12	HeartDisease	Heart Failure	binary system	0, 1

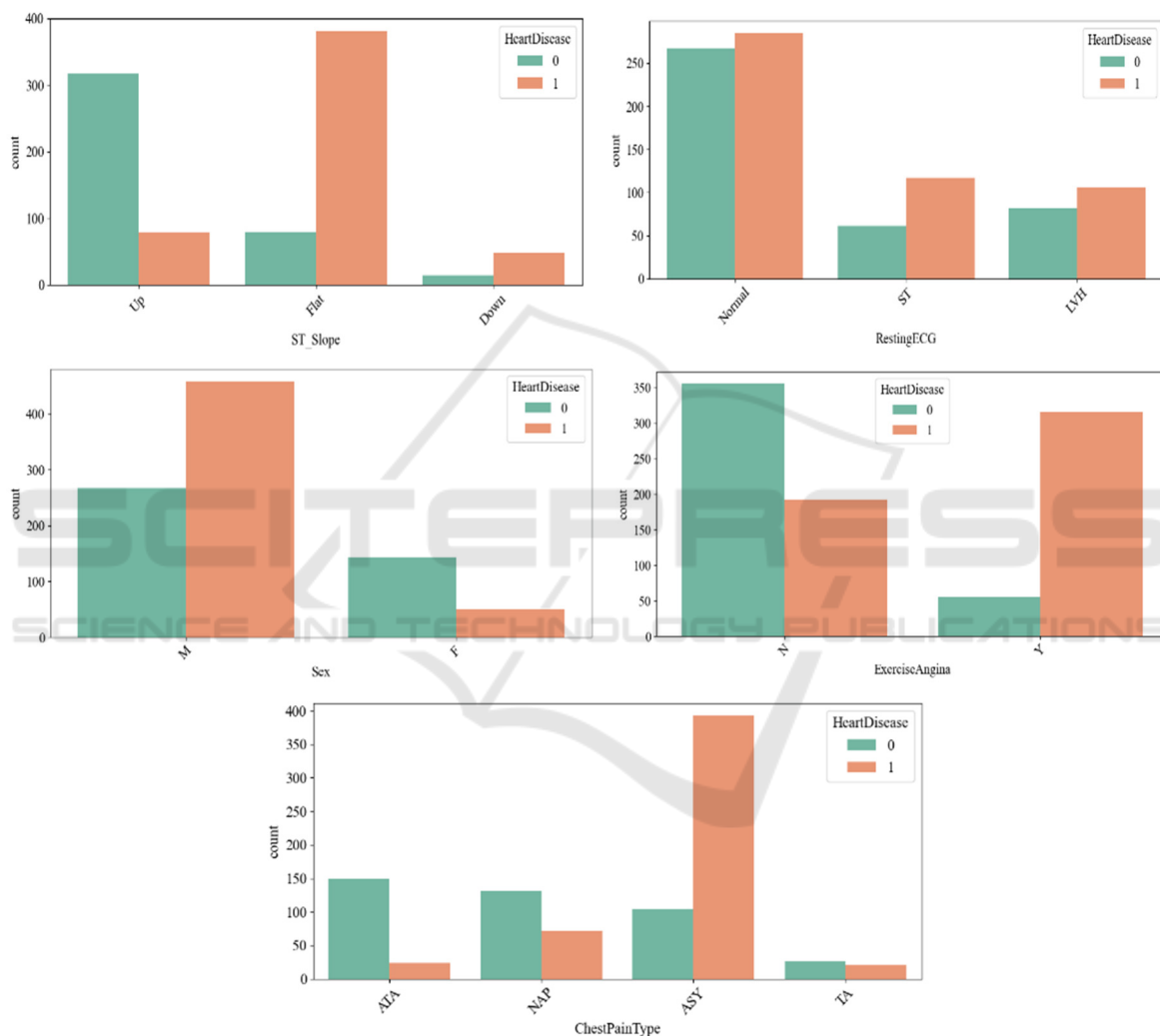


Figure 1: Distribution plots of non-continuous columns (Photo/Picture credit: Original).

3.1.2 Data Processing and Analysis

The dataset is clean and complete, with no invalid data or missing values, so there is no need for deletion or filling. The categorical data in the label-encoded dataset will be transformed into discrete data. Subsequently, the dataset will be split into 70% training set and 30% test set.

Next, distribution plots will be drawn for each feature to analyze the distribution of the data and its correlation with the target feature (whether the patient has HF) (refer to Figure 1) (refer to Figure 2).

From the distribution plots, it can be observed that the data in this dataset is fairly evenly distributed without any significant imbalance. Additionally, there is no clear linear relationship between the features and

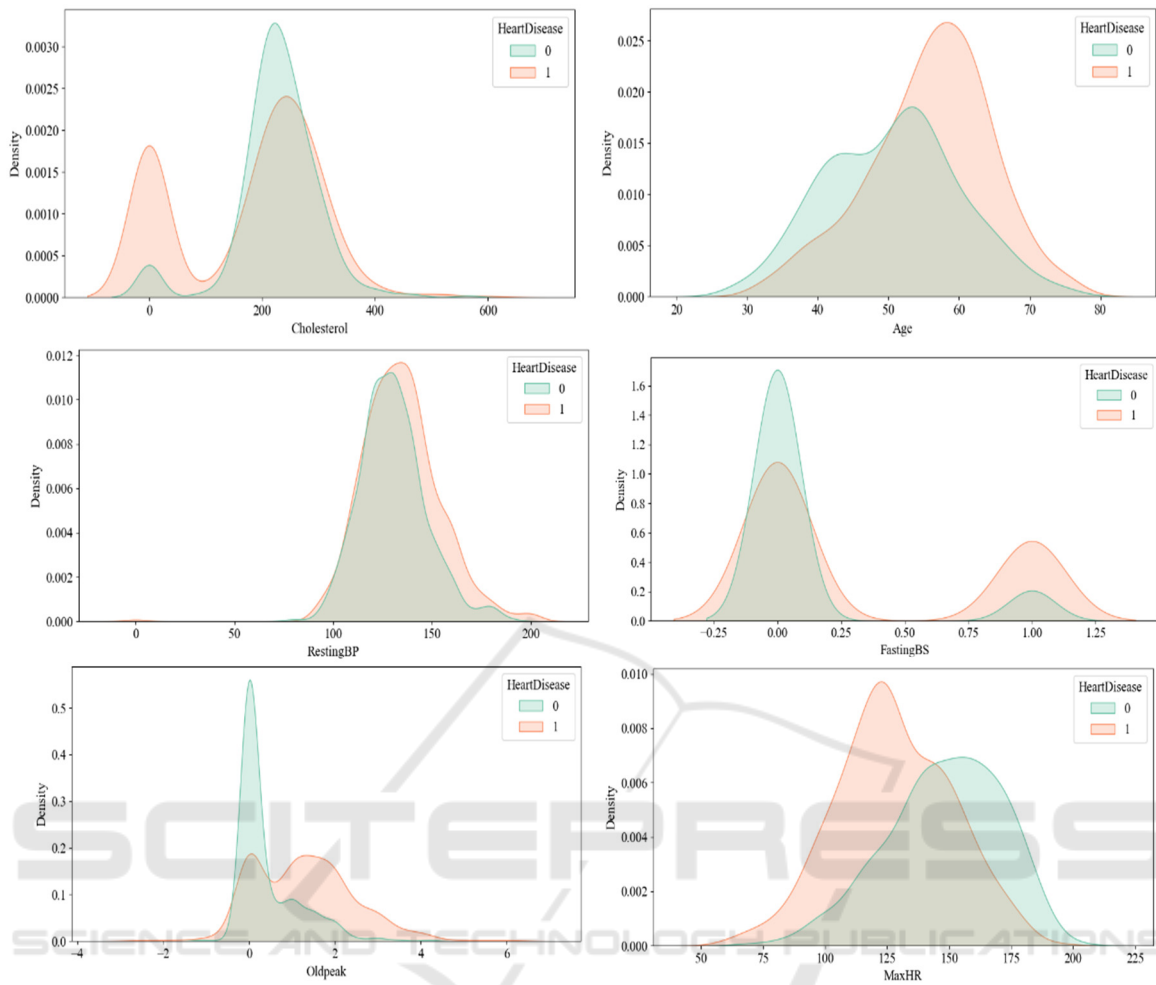


Figure 2: Distribution plots of continuous columns (Photo/Picture credit:Original).

the target feature. However, each feature shows a strong correlation with the target feature. Notably, features such as MaxHR, Oldpeak, and ST_Slope exhibit significant correlations with the likelihood of HF. This suggests that electrocardiogram readings can effectively reflect the condition of the heart and aid in predicting the occurrence of HF (Lyon, 2018). Furthermore, a heatmap displaying the correlation between features (refer to Figure 3) will be generated to visually represent the relationships among the different features.

From the figure 3, it is observed that, in terms of the correlation with the presence of HeartDisease, all features except Resting ECG show relatively ideal values, with Age, Sex, and 10 other features exhibiting strong correlations. This indicates that most features in the dataset have good correlations, either positive or negative, with the target feature. The strong correlations between the features and the target feature, along with the even distribution of the

data, demonstrate the usefulness of this dataset. Consequently, the following step would be to train the model with the training set and then evaluate its performance by testing it on the test set.

3.2 Model

3.2.1 Model Introduction

Categorical Boosting (CatBoost) is a specialized gradient boosting framework tailored for machine learning models that deal with categorical features in supervised learning (Kuan, 2019). This GBDT (Gradient Boosting Decision Tree) framework is founded on oblivious trees as base learners, recognized for its minimal complexity, provision for categorical variables, and exceptional accuracy. It effectively tackles the challenge of efficiently managing categorical features (Duan, 2019).

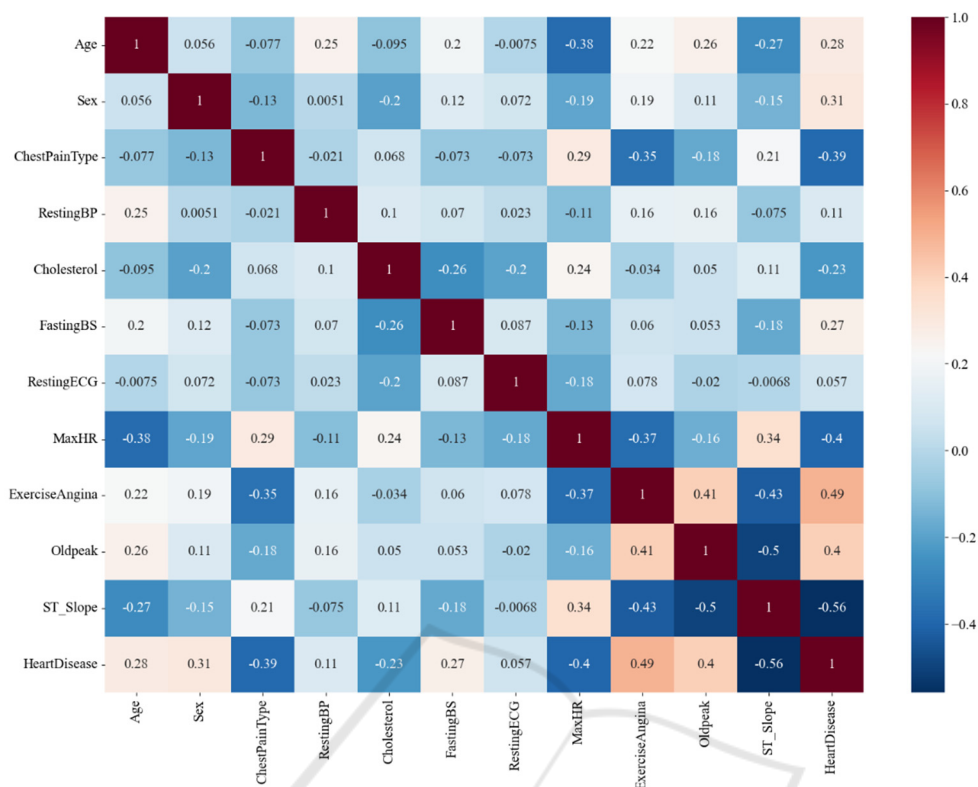


Figure 3: Heatmap (Photo/Picture credit: Original).

The methodology involves initially randomizing the order of all samples. For a specific value within a categorical feature, the conversion of this feature to a numerical representation for each sample is based on the average of the category labels preceding the current sample. Additionally, CatBoost incorporates priorities and weight coefficients for these priorities. The parameter "countInClass" denotes the count of samples with a label value of 1 corresponding to the current categorical feature value. The parameter "prior" signifies the initial value of the numerator, calculated using initial parameters. The parameter "totalCount" indicates the total count of samples, including the current sample, sharing the same categorical feature value across the entire dataset.

$$avg_target = \frac{countInClass + prior}{totalCount + 1} \quad (1)$$

CatBoost has the capability to handle categorical features and missing values. Therefore, within its

internal logic, it first processes categorical features, including ordered feature splitting algorithms and feature importance evaluation. CatBoost employs a gradient boosting tree-based ensemble learning method, where it initially trains a base learner (i.e., a decision tree) and then trains the next tree based on the residuals. These base learners are combined to reduce errors.

Furthermore, the CatBoost model utilizes a method of dynamically growing trees and employs an approach to finding the optimal solution for tree depth and leaf structure selection. CatBoost also balances tree depth using symmetric trees to mitigate the risk of overfitting. During the training process, CatBoost evaluates the model through techniques such as cross-validation and optimizes the model based on the evaluation results, including parameter tuning, feature engineering, and other optimizations (Figure 4).

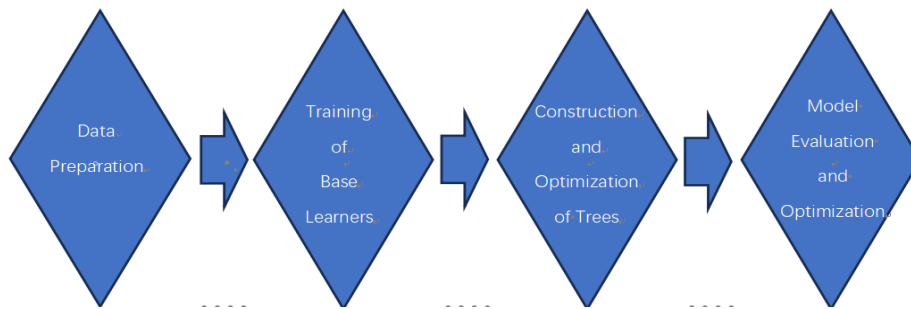


Figure 4: The training process of the CatBoost mode (Photo/Picture credit :Original).

3.2.2 Model Advantages

CatBoost processes categorical features during the training process, rather than handling them in the feature preprocessing stage. This is particularly beneficial when dealing with datasets containing a large number of categorical features, as it eliminates the need to encode these features using techniques such as label encoding or one-hot encoding, which can lead to loss of correlation information.

When constructing the tree structure, CatBoost employs an algorithm known as Symmetric Tree Growth to calculate leaf nodes. This method improves the model's robustness against overfitting, particularly in scenarios where the dataset is not large, and overfitting is a potential issue.

Furthermore, CatBoost has the capability to automatically adjust the learning rate during the training process. This adaptive learning rate adjustment ensures that the learning rate dynamically adapts based on the model's performance, facilitating faster convergence to the optimal state.

4 ANALYSIS OF EXPERIMENTAL RESULTS

4.1 Experimental Details

This study primarily adopts the approach of implementing model invocation and training using Python language on the PyCharm platform for experimentation. Following model training, the trained model is evaluated using a 30% test set partitioned during the data preprocessing stage. The experiment leverages the PyCaret library to automatically tune hyperparameters, aiming to obtain the optimal model.

For the final evaluation metrics, this experiment selects Accuracy, AUC (Area Under the Curve)

curve, Recall, Precision, and F1 score as the five key evaluation metrics (refer to Figure 5) (Khaled , 2022).

Based solely on these metrics, with values around 0.90 and not falling below 0.88, it can be concluded that the CatBoost model performs well regarding accuracy, AUC, recall, precision, and F1 score. Therefore, the trained model demonstrates precise predictions and exhibits strong performance.

4.2 Performance Comparison

In subsequent experiments, to validate the superiority of the CatBoost model in this study, we utilized the PyCaret library to invoke and train multiple models, obtaining relevant evaluation metrics (refer to Table 2).

It is evident from the table that the model ranks first in all five selected evaluation metrics. For medical disease prediction, it is imperative to focus on the five key metrics chosen for this experiment: Accuracy, Recall, Precision, and F1 score. A higher F1 score indicates precise model predictions and overall lower error rates, while a high AUC value signifies the reliability and practical value of the detection method in this experiment. Notably, the CatBoost model achieved rare results with scores surpassing those of other machine learning models across multiple evaluation metrics in this experiment. The following analysis can be performed to determine the causes behind the generally poor performance of popular models like RT and LR in this study.

Regarding the LR model, it is important to note that LR is a linear model with limited capability to model complex nonlinear relationships. Given that the features in this study do not exhibit clear linear relationships with the target feature, and that cardiovascular disease diagnosis in real-life scenarios cannot be solely based on simple linear relationships, CatBoost, as a gradient boosting tree model, possesses strong nonlinear modeling capabilities to construct accurate and robust predictive models based on complex features.

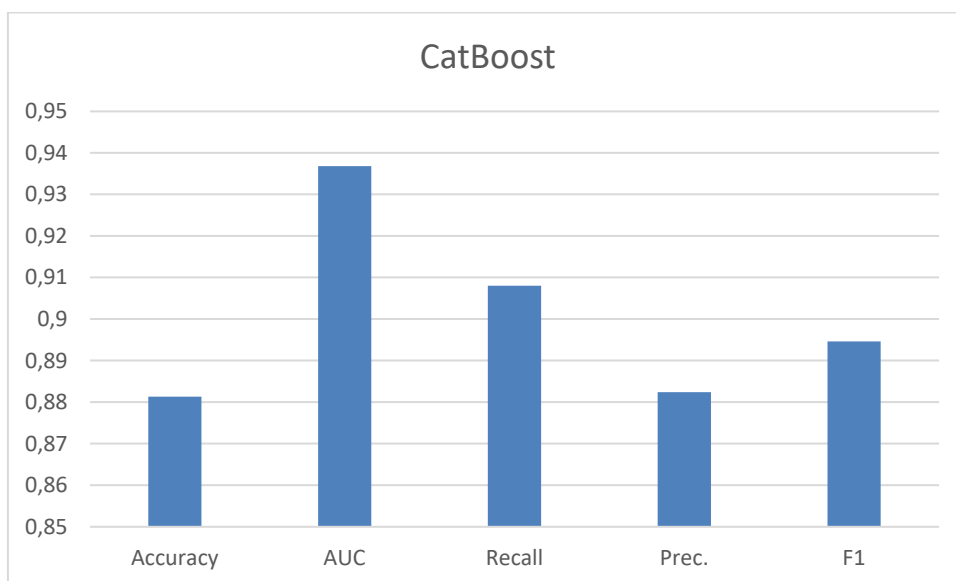


Figure 5: The commonly used evaluation metrics for the CatBoost model (Photo/Picture credit :Original).

Table 2: Comparison of Evaluation Metrics.

Model	Accuracy	AUC	Recall	Prec.	F1
CatBoost	0.8813	0.9368	0.908	0.8824	0.8946
Gradient Boosting	0.8643	0.9298	0.8795	0.8762	0.8771
Light Gradient Boosting Machine	0.8606	0.9233	0.8903	0.8639	0.8763
Random Forest	0.8583	0.9276	0.8861	0.8641	0.8739
Extra Trees	0.8582	0.9232	0.8884	0.864	0.8741
Linear Discriminant Analysis	0.857	0.9255	0.884	0.8635	0.8728
Logistic Regression	0.8521	0.9249	0.8816	0.8579	0.8689
Extreme Gradient Boosting	0.851	0.922	0.8772	0.8582	0.8671
Naive Bayes	0.8473	0.914	0.86	0.8659	0.8618
Ada Boost	0.8473	0.9109	0.8707	0.8574	0.8633
K Neighbors	0.8316	0.89	0.8644	0.8415	0.8512
Decision Tree	0.7832	0.7804	0.8072	0.8028	0.8041
Quadratic Discriminant Analysis	0.7698	0.8237	0.7837	0.8103	0.785

As for the RT model, firstly, RT models have several hyperparameters that can be adjusted, such as the number of trees, maximum depth, and learning rate. It is possible that the hyperparameters used in this study may not have been the most suitable for this dataset, leading to suboptimal performance. Secondly, RT models have strong fitting capabilities, which can sometimes result in overfitting issues. In contrast, the CatBoost model effectively mitigates the

risk of overfitting through techniques such as regularization and adaptive learning rates.

In conclusion, the application of the CatBoost model for HF prediction not only demonstrated excellent performance in this experiment but also holds promising research value and prospects for practical applications.

5 CONCLUSION

This study selected a suitable heart dataset and conducted machine learning model training on the PyCharm platform. The PyCaret library was introduced to facilitate the calling and training of multiple models. Ultimately, through comparative analysis with numerous models, the CatBoost model demonstrated the best performance. The study also analyzed the characteristics and training principles of this model, as well as briefly explained the reasons for its superior performance. Further comparisons revealed that its evaluation metrics (Accuracy, AUC, F1, Precision, Recall, etc.) surpassed those of other models, indicating a significant advantage of the CatBoost model in predicting HF.

The sample size of the dataset chosen for this study was relatively small, which may limit the generalizability of the resulting models. Future work could involve collecting additional relevant datasets, combining them to increase the sample size, and then conducting further model training and comparisons.

Future endeavors could focus on collaborating with clinical practitioners to validate and implement the model in clinical settings, in order to assess its reliability and practical utility. Simultaneously, exploring the application of the model in mobile healthcare devices and remote monitoring systems could assist healthcare professionals and enable personalized health management for patients. These efforts have the potential to enhance the efficiency and quality of the healthcare industry, offering new possibilities for disease prevention and management.

REFERENCES

- Bzdok, D., Altman, N., Krzywinski, M. 2018. Statistics versus machine learning. *Nature Methods*, 15(4): 233-4.
- Duan B., S, J. 2021. Analysis and classification of heart rate using CatBoost feature ranking model. *Biomedical Signal Processing and Control*, 68.
- Kobayashi, M., Huttin, O., Magnusson, M., et al. 2021. Machine Learning-Derived Echocardiographic Phenotypes Predict Heart Failure Incidence in Asymptomatic Individuals. *JACC. Cardiovascular Imaging*.
- Kuan, H. K., C, V. R., et al. 2019. Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods. *AJR. American journal of roentgenology*, 212(1): 38-43.
- Khaled, R., Raymond, B., Dewar, F., et al. 2022. Machine learning and the electrocardiogram over two decades: Time series and meta-analysis of the algorithms, evaluation metrics and applications. *Artificial Intelligence in Medicine*, 132: 102381-102381.
- Lyon, A., Mincholé, A., Martínez, J. P., et al. 2018. Computational techniques for ECG analysis and interpretation in light of their contribution to medical advances. *Journal of the Royal Society Interface*.
- Song, Y. X., Ma, X. Y., Wang, S. C., et al. 2023. Research progress on the application of machine learning in predicting heart failure. *Chinese Journal of Evidence-Based Cardiovascular Medicine*, 15(01): 118-119+126.
- Strom, J. B., Sengupta, P. P. 2021. Predicting Preclinical Heart Failure Progression: The Rise of Machine-Learning for Population Health. *JACC. Cardiovascular Imaging*.
- Vickers, J. 2018. What's the difference between machine learning and deep learning? *Vision Systems Design*, 23(9).
- Zhong H. 2007 Guidelines for the Diagnosis and Treatment of Chronic Heart Failure. *Chinese Journal of Cardiovascular Disease*, 35(12): 1076-1095.