Heart Failure Prediction with Machine Learning: A Comparative Study

Jing Wang
School of Computer Science, University of Wollongong, NSW 2522, Australia
Corresponding Author’s Email: 1037826627@qq.com

Abstract. Heart failure is a worldwide healthy problem affecting more than 550,000 people every year. A better prediction for this disease is one of the key approaches of decreasing its impact. Both linear and machine learning models are used to predict heart failure based on various data as inputs, e.g., clinical features. In this paper, we give a comparative study of 18 popular machine learning models for heart failure prediction, with z-score or min-max normalization methods and Synthetic Minority Oversampling Technique (SMOTE) for the imbalance class problem which is often seen in this problem. Our results demonstrate the superiority of using z-score normalization and SMOTE for heart failure prediction.

1. Introduction
Heart failure is a serious problem which has a huge impact on people’s life. With the accelerated pace of life, increased portion sizes and inactivity, most people always ignore their health. Moreover, because of the environmental deterioration, those factors can lead to the issue of heart failure which can become more and more common in the future. If people did not pay attention to the issue of heart failure, it would finally cause the death.

In the past years, different researchers used different methods to collect and analyze data with the aim to predict heart failure. These data include electronic health record (EHR) data of patients with heart failure in different hospitals from different countries, Cleveland heart disease dataset, biomedical science datasets from UCI, etc. Based on these data, various methods are being applied, e.g., predicting the survival of patients by utilizing classifiers of machine learning, using supervised deep learning and machine learning algorithms, training a boosted decision tree algorithm, utilizing machine intelligence-based statistical model, random under-sampling method and deep neural network models, using bioinformatic explainable deep neural network (BioExpDNN), etc.

Although there are some work based on machine learning, comprehensive comparisons between different models have not been made, nor have new models such as Deep Forest been attempted. The work of this paper can fill these inadequacies. We use F1 score and accuracy as the evaluation metrics. The F1 score reflects the robustness of the model, and the accuracy reflects the overall accuracy. First, with the min-max normalization method, the results of 18 models are compared. It can be seen from the accuracy that SMOTE (Synthetic Minority Over-sampling Technique) method is higher than that without SMOTE method, and it can be seen from the comparison of F1 score that SMOTE method is higher than that without SMOTE method. By comparing the z-score normalization methods with the 18 models, the accuracy of the SMOTE method is higher than that of the non-SMOTE method, and the F1 score comparison shows that the SMOTE method is higher than that of the non-SMOTE method. Overall, the z-score normalization method is better than the min-max normalization method.
2. Related Work

In this section, we give a short review of recent related works.

In [1], taking advantage of 299 patients who have cardiac failure in 2015. Those data have 13 features for example high blood pressure, sex, and smoking. The authors utilized some different classifiers of Machine Learning to forecast the proportion of survivors, and rank the features corresponding to the most important risk factors. They find serum creatinine and ejection fraction are the most important factor to forecast the proportion of survivors.

In [2], the authors use some data which are from electronic health record (EHR) and use some models about Machine Learning and Deep Learning. The results show novel machine learning models is possible to change the prediction accuracy of model.

In [3], taking advantage of dataset which is from EHR data and has 26,575 patients who have cardiac failure in 2018. Finally, the results show age of patients, creatinine, body mass index, and levels of blood pressure were significant factors in predicting mortality within one year among heart failure patients.

In [4], the authors collect data from 5,822 hospitalized and ambulatory patients with heart failure and it included eight variables, for example, diastolic blood pressure, white blood cell count, platelets, albumin, and red blood cell distribution width. By training a boosted decision tree algorithm, a model was developed to correlate a subset of 5822 inpatient and outpatient heart failure patients with a very high or very low risk of mortality. As a result, using machine learning and off-the-shelf variables, the authors generated and validated a risk score for death in patients with heart failure that was more accurate than other risk scores compared to it.

In [6], the authors use Cleveland heart disease dataset in the research paper and take advantage of machine intelligence -based statistical model, random under-sampling method and DNN model. They found that the given predictive model is helpful for doctors to diagnose heart failure.

In [7], this study used three datasets about biomedical science which are from UCI and utilized bioinformatic explainable deep neural network. They found that the results of research paper showed that the classification accuracies about CDS, CCBRDS, and HFCRDS were 92.59 percent, 100 percent, and 78.9 percent.

In [8], the authors took advantage of data which is from the medical records about cardiac failure patients. The dataset contains 299 cardiac failure patients in medical records and has clinical and lifestyle data. Different machine learning algorithms are used. The result showed that Machine learning algorithms are tools which are useful and effective to classify the records of medicine of patients with cardiac failure.

In [9], the authors used 299 patients who have cardiac failure in EHR and they are recorded at the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad, from April to December in 2015. Oversampling was used by authors to solve the imbalance problem by adding information in a few classes. Random forest (RF) was also utilized by the authors to classify. Finally, the SMOTE is the most suitable way for oversampling EHR data to predict the proportion of death in patients with cardiac failure.

In [10], the authors took advantage of these data of patients with cardiac failure from the Faisalabad Institute of Cardiology and the Faisalabad United Hospital. Using the multilayer perceptron neural network to analyze. The results showed that the heart failure dataset, with 88 percent accuracy, outperformed other studies in predicting heart failure.

3. Dataset Description

We utilize a common dataset in public: https://www.kaggle.com/andrewmvd/heart-failure-clinical-data. This data contains the following input characteristics and prediction targets. The target we want to predict is the Death Event. The features have Age, anaemia, high blood pressure, creatinine phosphokinase, diabetes, ejection fraction, platelets, sex, serum creatinine, smoking, and time. Since the number of positive and negative samples is different, I will try SMOTE method later to obtain a uniform sample distribution.
In figure 1, it introduces the distribution of death events in patients.

Figure 1. The distribution of death events in patients.

All input features are numeric data. The correlation between different input characteristics and Death Event is shown in Figure 2.

Figure 2. The correlation between different columns.
4. Models and Experiments

4.1. Models
I want to introduce all machine learning models I use in my research paper. With the development of big data and artificial intelligence, machine learning has been successful in a series of problems. For heart failure prediction, there are also some relevant studies as we introduced in the related work part. However, there lacks a comprehensive comparison between different machine learning models.

Logistic Regression (LR): Logistic regression which is a classification model, and it is often used in dichotomy. Logistic Regression is one of algorithms of ML to solving binary (0 or 1) problems, which is used to estimate the likelihood of some things.

K Nearest Neighbour (kNN): KNN is one of easiest algorithms in data analysis and it is a nonparametric statistical method for classification and regression.

Naives Bayes (NB): Naive Bayes classifier is a series of simple probability classifier based on Bayes theorem under strong (naive) independence between assumed features. For example, Bayes' theorem is to solve the problem which often encountered in the real life.

Decision Tree (DT): Decision Tree is a tree structure which is binary or non-binary.
Each non-leaf node represents a test on a feature attribute, each branch represents the output of this feature attribute on a range of values, and each leaf node stores a category.

Linear SVM (SVM-Linear) and RBF SVM (SVM-RBF): SVM is one of supervised learning methods, can be widely used in statistical classification and regression analysis. SVM belongs to generalized linear classifiers, which are characterized by their ability to minimize empirical error and maximize geometric edge region at the same time. Therefore, another name of SVM is maximum edge region classifier.

Multi-Level Perceptron (MLP): MLP is a kind of feed-forward neural network (ANN). The term MLP is vague, sometimes loosely applied to any feed-forward ANN, and sometimes strictly refers to a network consisting of multiple layers of perceptron.

Ridge Classifier (RC): In the least squares of ordinary linear regression, it is only valid when the matrix is nonsingular matrix, but in most cases the matrix is not rank, and ridge regression can be used to solve this problem.

Random Forest (RF): Random forest is a classifier which includes a lot decision trees, and its class of output is decided by mode of class of outputs of individual trees.

Quadratic Disc. Analysis (QDA): Quadratic Discriminant Analysis is improved on Linear Discriminant Analysis.

AdaBoost (AdaBoost): The Adaboost model is an iterative algorithm, whose core idea is to train different weak classifiers for the same training set, and then assemble these weak classifiers to form a strong classifier. Adaboost can handle classification and regression problems Gradient Boosting (GB).

Linear Disc. Analysis (LDA): LDA is a dimension reduction technique of supervised learning, that is to say, each sample of its data set is output by category.

Extra Trees (ET): Extra Trees is one of the most useful algorithms in Machine Learning. It is similar with Random Trees and it is generated by a lot decision trees.

Extreme Gradient Boosting (XGBoost): XGBoost is an open-source library, and it can provide an efficient and effective implementation of gradient enhancement algorithms.

Light Gradient Boosting (LGB): Light Gradient Boosting Machine is a framework to implement GBDT algorithm, which supports efficient parallel training, faster training speed, lower memory consumption, better accuracy, distributed support, and fast processing of massive data.

CatBoost Classifier (CatBoost): CatBoost is an open-source gradient enhancement library, and it is based on decision trees that provides off-the-shelf classification feature to support Python and R.

Deep Forest (DF): Deep Forest is an integration of the breadth and depth of the traditional forest. The purpose of integration in depth is to improve the classification capability. Purpose of integration in breadth is to reflect the difference of input data.
4.2. Experiments
In this part, we describe the experiment details. Firstly, I divided the data set, and divided 10% of the dataset, namely 30 sample points, as the test dataset. 90% of the entire dataset as the training dataset. Moreover, I compare the following situations: min-max normalization without SMOTE, min-max normalization with SMOTE, z-score normalization without SMOTE, z-score normalization with SMOTE. The accuracy of different machine learning models shows in Table 1. The F1. score of different machine learning models shows in Table 2.

Table 1. The accuracy of different machine learning models.

| Model   | Min-max Normalization | Z-score Normalization |
|---------|-----------------------|-----------------------|
|         | No SMOTE | SMOTE     | No SMOTE | SMOTE   |
| LR      | 0.8333   | 0.7667    | 0.8667   | 0.7667  |
| kNN     | 0.7000   | 0.8000    | 0.7333   | 0.8000  |
| NB      | 0.6333   | 0.8000    | 0.6333   | 0.7667  |
| DT      | 0.8333   | 0.8333    | 0.8333   | 0.8333  |
| SVM (linear) | 0.8333 | 0.8667 | 0.8667 | 0.8000 |
| SVM (RBF) | 0.8667 | 0.8000 | 0.8333 | 0.8333 |
| MLP     | 0.7667   | 0.7667    | 0.8667   | 0.7667  |
| RC      | 0.7333   | 0.8000    | 0.8000   | 0.8333  |
| RF      | 0.8667   | 0.8333    | 0.8667   | 0.8000  |
| QDA     | 0.6987   | 0.8667    | 0.7333   | 0.8667  |
| Adaboost| 0.7000   | 0.8333    | 0.7000   | 0.7333  |
| GB      | 0.7667   | 0.8000    | 0.7333   | 0.8000  |
| LDA     | 0.8000   | 0.8000    | 0.8000   | 0.8000  |
| ETC     | 0.8333   | 0.8667    | 0.8333   | 0.8667  |
| EGB     | 0.8667   | 0.8000    | 0.7667   | 0.8333  |
| LGB     | 0.8667   | 0.8000    | 0.7667   | 0.8667  |
| CC      | 0.8333   | 0.8000    | 0.8333   | 0.8333  |
| DF      | 0.8333   | 0.8000    | 0.8333   | 0.8000  |

Table 2. The F1 score of different machine learning models.

| Model   | Min-max Normalization | Z-score Normalization |
|---------|-----------------------|-----------------------|
|         | No SMOTE | SMOTE     | No SMOTE | SMOTE   |
| LR      | 0.6667   | 0.6316    | 0.7500   | 0.6316  |
| kNN     | 0.1818   | 0.6667    | 0.5333   | 0.6250  |
| NB      | 0.0000   | 0.6250    | 0.0000   | 0.5333  |
| DT      | 0.6667   | 0.6667    | 0.6667   | 0.6667  |
| SVM (linear) | 0.6667 | 0.8182 | 0.7500 | 0.6667 |
| SVM (RBF) | 0.7500 | 0.7000 | 0.7059 | 0.7059 |
| MLP     | 0.4615   | 0.6316    | 0.7500   | 0.6316  |
| RC      | 0.3333   | 0.6667    | 0.5714   | 0.7059  |
| RF      | 0.7500   | 0.7368    | 0.7500   | 0.6667  |
| QDA     | 0.0000   | 0.7500    | 0.4286   | 0.7778  |
| Adaboost| 0.4000   | 0.7059    | 0.4000   | 0.5000  |
| GB      | 0.5882   | 0.6667    | 0.5000   | 0.6667  |
| LDA     | 0.6250   | 0.6667    | 0.6250   | 0.6667  |
| ETC     | 0.6667   | 0.7778    | 0.6667   | 0.7778  |
| EGB     | 0.6316   | 0.7000    | 0.6316   | 0.7619  |
| LGB     | 0.5333   | 0.6667    | 0.5333   | 0.7778  |
| CC      | 0.6667   | 0.7000    | 0.6667   | 0.7368  |
| DF      | 0.7059   | 0.6667    | 0.7059   | 0.6667  |
We choose F1 score and accuracy as our evaluation metrics. The F1 score reflects the robustness of the model, and the accuracy reflects the overall accuracy. The conclusion is made by comparing the accuracy as well as the F1 score. First, the min-max normalization method, which uses 18 models, is compared. It can be easy to see from the accuracy that SMOTE method is higher than that without SMOTE method, and it can be seen from the comparison of F1 score that SMOTE method is higher than that without SMOTE method. By comparing the z-score normalization method with the 18 models, the accuracy of the SMOTE method is higher than that of the non-SMOTE method, and the F1 value comparison shows that the SMOTE method is higher than that of the non-SMOTE method. The z-score normalization method is better than the min-max normalization method.

As can be seen from the accuracy value, SMOTE has improved the accuracy of 10 of the 18 models in the min-max normalization method, and 13 of the 18 models in F1 score. SMOTE has improved accuracy of 9 of the 17 models in the z-score normalization method, and 11 of the 17 models in F1 score. Thus, the SMOTE approach is helpful for most models used in this study.

I further evaluate the feature importance of these machine learning models, especially the tree-based models. The feature importance graphs of the Random Forest, AdaBoost, Light Gradient Boosting, Extra Trees, XGBoost and CatBoost models are shown in Figure 3-8. As we can see feature importance of those models from these figures. Feature ‘time’ is the most important input features, which is different from other models. Feature ‘serum_creatinine’ is the second most important input features in the entire models. However, different models may assign different feature importance scores for different features.

![Feature importance of Random Forest.](image1)

![Feature importance of AdaBoost.](image2)

![Feature importance of Light Gradient Boosting.](image3)

![Feature importance of Extra Trees.](image4)
5. Conclusion
In this paper, I analyzed and compared the performance of 18 different machine learning models for heart failure prediction based on 12 clinical features. I utilize F1 score and accuracy as the evaluation metrics. The F1 score reflects the robustness of the machine learning model, and the accuracy reflects the ratio of correctly classified cases. Our results demonstrate that z-score normalization is better than min-max normalization and using SMOTE for class imbalance is helpful. At the same time, we also show the feature importance of different models. Input feature ‘time’ is the most significant, which is different from other models in different models. Heart failure is a serious problem which have a negative effect on people’s live and more than 550,000 people every year are affected by the disease. I need to predict for this disease and utilize some approaches to reduce its impact in the body of human.

In the future, I will continue to search for other approaches which can be optimal for the performance of prediction of heart failure. For example, combining different models of machine learning.

References
[1] Chicco, Davide, and Giuseppe Jurman. "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone." BMC medical informatics and decision making 20.1 (2020): 16.
[2] Guo, Aixia, et al. "Heart Failure Diagnosis, Readmission, and Mortality Prediction Using Machine Learning and Artificial Intelligence Models." Current Epidemiology Reports (2020): 1-8.
[3] Guo, A., et al. "The Use of Synthetic Electronic Health Record Data and Deep Learning to Improve Timing of High-Risk Heart Failure Surgical Intervention by Predicting Proximity to Catastrophic Decompensation. Front. Digit." Front. Digit. Health 2 (2020): 576945.
[4] Olsen, Cameron R., et al. "Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure." American Heart Journal (2020).
[5] Adler, Eric D., et al. "Improving risk prediction in heart failure using machine learning." European journal of heart failure 22.1 (2020): 139-147.
[6] Ali, L., and S. A. C. Bukhari. "An approach based on mutually informed neural networks to optimize the generalization capabilities of decision support systems developed for heart failure prediction." IRBM (2020).
[7] Fang, Hao, Cheng Shi, and Chi-Hua Chen. "BioExpDNN: Bioinformatic Explainable Deep Neural Network." 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2020.
[8] Eletter, Shorouq, et al. "Building an Intelligent Telemonitoring System for Heart Failure: The Use of the Internet of Things, Big Data, and Machine Learning." 2020 21st International Arab Conference on Information Technology (ACIT). IEEE, 2020.

[9] Kim, Young-Tak, et al. "A Comparison of Oversampling Methods for Constructing a Prognostic Model in the Patient with Heart Failure." 2020 International Conference on Information and Communication Technology Convergence (ICTC). IEEE, 2020.

[10] Le, Minh Tuan, et al. "Predicting heart failure using deep neural network." 2020 International Conference on Advanced Technologies for Communications (ATC). IEEE, 2020.

[11] Jiang W, Zhang L. Geospatial data to images: A deep-learning framework for traffic forecasting[J]. Tsinghua Science and Technology, 2018, 24(1): 52-64.

[12] Jiang W. Applications of deep learning in stock market prediction: recent progress[J]. arXiv preprint arXiv:2003.01859, 2020.

[13] Jiang W, Zhang L. Edge-siamnet and edge-triplenet: new deep learning models for handwritten numeral recognition[J]. IEICE Transactions on Information and Systems, 2020, 103(3): 720-723.

[14] Jiang W. Time series classification: nearest neighbor versus deep learning models[J]. SN Applied Sciences, 2020, 2(4): 1-17.

[15] Jiang W, Luo J. Graph Neural Network for Traffic Forecasting: A Survey[J]. arXiv preprint arXiv:2101.11174, 2021.

[16] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[17] Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features[J]. arXiv preprint arXiv:1706.09516, 2017.

[18] Zhou Z H, Feng J. Deep forest[J]. arXiv preprint arXiv:1702.08835, 2017.