Prediction of Heart Failure Using Random Forest and XG Boost

By Aidan Gao

AUTHOR BIO

Aidan Gao is a sophomore attending The Westminster Schools in Atlanta, Georgia. He is interested in the uses of computer science and machine learning in the medical realm. He hopes to study computational medicine in the future.

ABSTRACT

Heart Failure (HF), a type of Cardiovascular Disease (CVD), is a prevalent illness that can lead to hazardous situations. Each year, approximately 17.9 million patients globally die of this disease. It is challenging for heart specialists and surgeons to predict heart failure accurately and on time. Fortunately, there are classification and prediction models available that can assist the medical field in efficiently using medical data. The objective of this study is to enhance the accuracy of heart failure prediction by prediction modeling a Kaggle dataset composed of five sets of data over 11 patient attributes. Multiple machine learning approaches were utilized to understand the data and forecast the likelihood of heart failure in a medical database. The results and comparisons show a definite increase in the accuracy score of predicting heart failure. Integrating this model into medical systems would prove beneficial for aiding doctors predictions of heart disease in patients

Keywords: machine learning, heart failure, diagnosis, prediction modeling, binary classification, random forest, XGBoost, cardiovascular disease

INTRODUCTION

The primary cause of heart stroke is the obstruction of arteries, also known as cardiovascular disease or arterial hypertension World Health Organization (n.d.). Heart disease affects approximately 26 million people worldwide, and this number is expected to rise rapidly if effective measures are not taken (Savarese & Lund, 2017). Unhealthy food, tobacco, excessive sugar, and obesity are common contributors to heart disease (Benjamin et al., 2019). Pain in the arms and chest are common symptoms, but the disease often presents with different symptoms based on sex and age. In addition to maintaining a healthy lifestyle and diet, timely diagnosis and comprehensive analysis are critical factors in identifying heart disease. However, many patients undergo multiple tests that can be physically and financially burdensome. Proper analysis of this type of data can improve the diagnosis process and assist heart surgeons. Previous research has used various techniques such as Random Forest, Support Vector Machine, and other AI classification models (Alotaibi, 2019). This study aims to surpass previous studies' random forest model accuracy in order to better predict heart failure before it manifests.

LITERATURE REVIEW

Previous work has utilized a subset of the dataset used in this paper to predict heart failure. The University of California Irvine (UCI) used Decision Tree, Logistic Regression, Random Forest, Naïve Bayes, and SVM reaching results around the ~85% accuracy mark. Through the use of tenfold cross validation and an enlarged dataset, previous studies have enhanced the accuracy of the UCI models (see Table. 1) (Alotaibi, 2019).

Table 1. Performance Comparison

Model	Alotaibi (Alotaibi, 2019)	UCI, 2019 (Bashir et al., 2019)	UCI, 2017 (Ekiz and Erdogmus, 2017)	UCI, 2017 (Ekiz and Erdogmus, 2017)
Decision Tree	93.19%	82.22%	60.9%	67.7%
Logistic Regression	87.36%	82.56%	65.3%	67.3%
Random Forest	89.14%	84.17%	X	X
Naïve Bayes	87.27%	84.24%	X	X
SVM	92.30%	84.85%	67%	63.9%

DATA OVERVIEW

The dataset utilized in this paper is collected from Kaggle under the name "Heart Failure Prediction Dataset" (Ortega, 2021). The dataset combines five datasets with over 11 common attributes. These five datasets combine data from Cleveland, Hungarian, Switzerland, Long Beach VA, and Stalog datasets. In total, the dataset contains 918 rows. Row definitions are provided below:

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or

depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

METHODS

Data Preprocessing

Before putting the data into the model, data preprocessing methods are applied in order to make it useful for modeling (Al-Mudimigh et al., 2009). Binary values, such as sex, were converted to binary numbers (1 and 0). ExerciseAngina was converted to binary as well. One hot encoding was employed in ChestPainType. The ordinal encoder from SKLearn was employed in the ordinal variables, ST_Slope and RestingECG.

This research focuses on two models to predict heart disease: Random Forest and XG Boost. Random Forest was implemented with the goal to surpass previous research accuracy with its cross-validation score. XGBoost was also implemented as a popular model among many Kaggle dataset winners. Both models were cross-validated ten times across an 80-20 split of training and test data, respectively. In order to further increase accuracy, HyperOPT (Komer et al., 2019) and RandomizedSearch (Agrawal, 2021) were used to finetune the hyperparameters for XGBoost and Random Forest, respectively.

Random Forest

The Random Forest algorithm is utilized to address classification issues. Its approach is based on ensemble learning, which combines multiple classifiers to enhance the algorithm's performance. The algorithm is composed of several Decision Trees classifiers to create a forest (Donges, 2018), each working on a subset of data, and the average is calculated to improve prediction accuracy. Rather than relying on the prediction of a single tree, the Random Forest algorithm combines the trees using an estimated outcome and voting procedure (Bashar et al., 2019). The model then considers predictions from each tree and determines the outcome based on majority voting.

XGBoost

XGBoost, which stands for "Extreme Gradient Boosting," is a popular machine learning model that has been used for a variety of tasks such as regression, classification, and ranking. XGBoost creates a model in the form of boosting an ensemble of weak classification trees by gradient descent which provides optimization to the loss factor (Cui et al., 2017). It is an ensemble learning algorithm that combines multiple decision tree models to improve the accuracy and robustness of predictions. XGBoost has gained popularity due to its speed, scalability, and performance. It uses a gradient-boosting framework and can handle missing values, regularization, and parallel processing. Additionally, it has various hyperparameters that can be tuned to achieve better performance. Overall, XGBoost is a powerful machine learning model that is widely used in industry and academia.

RESULTS

When looking at binary classification problems, there are four relevant metrics: true positives, true negatives, false positives, and false negatives. Out of these four metrics, the most harmful to prediction would be false negatives, or results that report no risk of heart failure despite the patient being at risk. The goal of this research was to reduce the number of false negatives and false positives in order to improve accuracy in predicting heart disease. Using these models and methods, the result was 91.56% accuracy after cross validation for XGBoost, with 9 false negative cases out of 181 cases (see Fig. 1). In Random Forest, there was 92.90% accuracy after cross validation with 6 false negatives out of 181 cases (see Fig. 2). In comparison to Alotaibi (2019), Random Forest performed considerably better, increasing from 89.14 to 92.90% accuracy. The accuracy increase may be due to the hyperparameter tuning for Random Forest or the cross-validation method. Also taken into consideration is the artificial addition of rows. In Alotaibi (2019), the size of the Cleveland data was too low to implement machine learning approaches. Alotaibi increased the size of the data artificially by randomizing values between minimums and maximums. The issue with randomizing these values is that there is no way to tell whether the target value is right for the artificial patient, thus creating noisy data that is useless for the model. This causes much of the data to carry either a random target value or no target value, which would cause the accuracy of the model to go down[1].

Another metric used in binary classification problems is the ROC curve, or the area under the receiver operating characteristic, a common metric for evaluating binary classification models. A model with a higher AUC is thought of as a better model (Javeed et al., 2019). This value was 0.92 for XGBoost and Random Forest. Both models performed well in terms of accuracy and the area under the ROC curve, indicating their effectiveness in predicting heart disease.

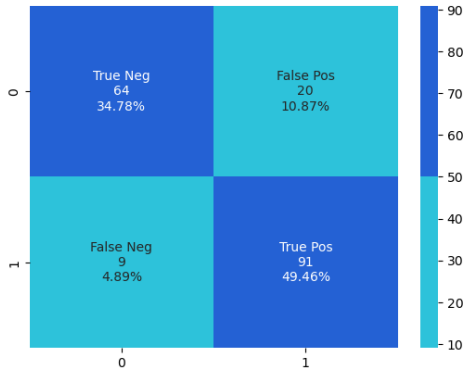

Fig. 1. Confusion Matrix of XGBoost Model

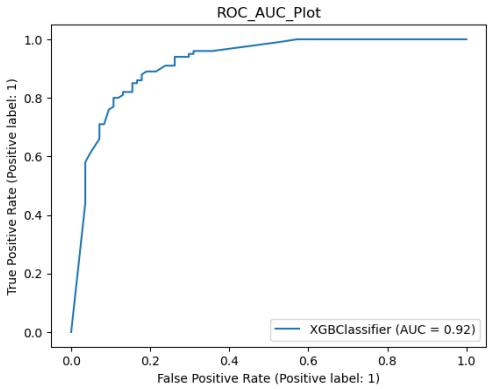

Fig. 2. ROC_AUC Plot of XGBoost

[1] This may not be the case, but there is no evidence in the paper or the references to explain the choice made by Alotaibi (2019).

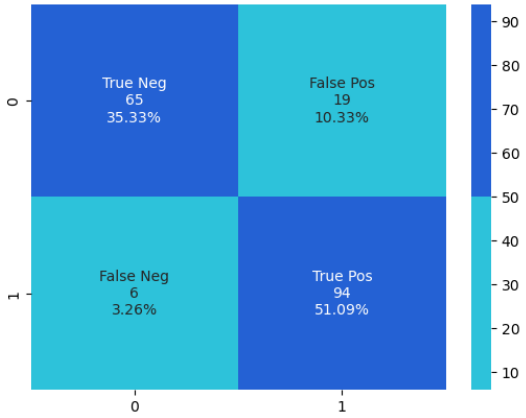

Fig. 3. Confusion Matrix of Random Forest Model

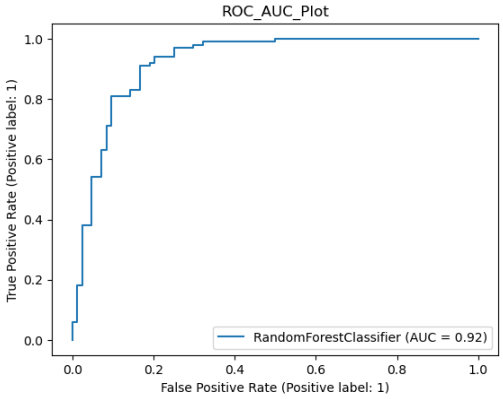

Fig. 4. ROC_AUC Plot of Random Forest Model

Table 2. Performance Comparison

Model	This Study	Alotaibi (Alotaibi, 2019)	UCI, 2019 (Bashir et al., 2019)	UCI, 2017 (Ekiz and Erdogmus, 2017)	UCI, 2017 (Ekiz and Erdogmus, 2017)
Decision Tree	X	93.19%	82.22%	60.9%	67.7%
Logistic Regression	X	87.36%	82.56%	65.3%	67.3%
Random Forest	92.90 %	89.14%	84.17%	X	X
Naïve Bayes	X	87.27%	84.24%	X	X
SVM	X	92.30%	84.85%	67%	63.9%
XGBoost	91.56 %	X	X	X	X

DISCUSSION

The results of this research reveal the potential use of machine learning in the landscape of cardiovascular healthcare. The degree of accuracy achieved by the current models opens up a pathway for research that may dramatically impact the diagnosis of heart disease. If integrated into medical information systems, these models could facilitate the collection and analysis of live data from patients, allowing doctors to make predictions in real-time. Hypothetically, machine learning models could form a network. Accessible to healthcare providers across the world, such a system could flag patients the models decide show significant risk of heart disease, prompting immediate intervention and follow-up. This would have a profound effect on the field of heart disease management, further shifting it from a reactive to a proactive field. Moreover, the system could be used in more rural, underserviced, areas where traditional diagnostic resources are unavailable.

This study helps to revolutionize early detection methods for heart disease, which in turn, significantly impacts patient outcomes. When applied on a larger scale, this approach could reduce mortality rates and enhance quality of life in patients diagnosed with this condition.

LIMITATIONS AND FURTHER WORK

These models are trained on historical data and their predictions are based on patient data patterns exhibited in past cases. However, upon application on a larger scale, much more patient data could be collected, training the model to keep up with modern times. One limitation of this study is that the dataset utilized

to model was relatively small. Therefore, the dataset may not be able to accurately mirror large populations, a significant limitation for a model meant to be used at the population level. To increase the accuracy and reliability of these models, future research should aim to expand the dataset in not just rows but also columns. With the addition of more lifestyle variable columns, the model's accuracy is likely to rise. Future studies should explore handling outlier cases that do not align with patterns in historical data. This could involve using hybrid models that incorporate machine learning with other predictive tools. With further development, these models could potentially bring the rise of a new era of predictive healthcare in cardiology.

CONCLUSION

Heart failure is a significant health issue that affects millions of people worldwide. Machine learning models, specifically Random Forest and XGBoost, can be used to accurately predict heart failure based on medical data. These models can be integrated with medical information systems to improve the accuracy of predictions and assist healthcare providers in detecting heart disease early.

ACKNOWLEDGEMENTS

I would like to acknowledge Ms. Avi-Yona Israel for her amazing support and guidance in the writing process of this paper

REFERENCE

World Health Organization. (n.d.). Cardiovascular diseases. Retrieved March 15, 2023, from https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

Savarese, G., & Lund, L. (2017). Global public health burden of heart failure. Cardiac Failure Review, 3(1), 7-11.

Benjamin, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C. W., Carson, A. P., ... & Virani, S. S. (2019). Heart disease and stroke statistics-2019 update: A report from the American Heart Association. Circulation, 139(10), e56-e528.

Alotaibi, F. S. (2019). Implementation of machine learning model to predict heart failure disease. International Journal of Advanced Computer Science and Applications, 10(6), 261-268. doi: 10.14569/IJACSA.2019.0100637

Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., & Bashir, K. (2019). Improving heart disease prediction using feature selection approaches. In 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST) (pp. 619-623).

Ekiz, S., & Erdogmus, P. (2017). Comparative study of heart disease classification. In 2017 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, EBBT 2017 (pp. 1-4).

Ortega, F. (2021). Heart failure prediction dataset. Kaggle. Retrieved February 25, 2023, from https://www.kaggle.com/fedesoriano/heart-failure-prediction.

Al-Mudimigh, F., Ullah, Z. S. A., & Saleem, A. (2009). A framework of an automated data mining systems using ERP model. International Journal of Computer, Electrical, Automation, Control and Information Engineering, 1(5), 598-605.

Komer, B., Bergstra, J., & Eliasmith, C. (2019). Hyperopt-sklearn. In Automated Machine Learning: Methods, Systems, Challenges (pp. 97-111). Springer.

Agrawal, T. (2021). Hyperparameter optimization using scikit-learn. In Hyperparameter Optimization in Machine Learning (pp. 21-41). Apress.

Donges, N. (2018). The random forest algorithm. Towards Data Science. Retrieved March 15, 2023, from https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

Bashar, S. S., Miah, M. S., Karim, A. H. M. Z., Al Mahmud, A., & Hasan, Z. (2019). A machine learning approach for heart rate estimation from PPG signal using random forest regression algorithm. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-5).

Cui, Y., Cai, M., & Stanley, H. E. (2017). Comparative analysis and classification of cassette exons and constitutive exons. BioMed Research International, 2017, 1-8.

Javeed, A., Zhou, S., Yongjian, L., Qasim, I., Noor, A., & Nour, R. (2019). An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection. IEEE Access, 7, 180235-180243. doi: 10.1109/ACCESS.2019.2950861