IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Prediction of Heart Failure Using Random Forest and XG Boost

Aidan Gao
The Westminster Schools, 1424 W Paces Ferry Rd NW, Atlanta, GA 30327, USA, aidanlgao@gmail.com

ABSTRACT: Heart Failure (HF), a type of cardiovascular disease (CVD), is a prevalent illness that can lead to life-threatening complications. Each year, approximately 17.9 million patients around the world die from this disease. It is challenging for heart specialists and surgeons to predict heart failure accurately and on time. Fortunately, available classification and prediction models can assist the medical field in efficiently using medical data. This study aims to enhance the accuracy of heart failure 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: Robotics and intelligent machines, machine learning, heart failure, diagnosis, prediction modeling, binary classification, forest, XGBoost, cardiovascular disease.

Introduction

The primary cause of heart stroke is the obstruction of arteries, also known as cardiovascular disease or arterial hypertension.[1] Heart disease affects approximately 26 million people worldwide, and this number is expected to rise rapidly if adequate measures are not taken.[2] Dietary indiscretions, tobacco use, excessive sugar consumption, and obesity commonly contribute to heart disease.[3] Pain in the arms and chest commonly manifests as 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. Adequate analysis of this data type can enhance the diagnostic process and assist cardiac surgeons. Previous research used techniques like Random Forest, Support Vector Machine, and other AI classification models.[4] This study aims to surpass previous studies' random forest model accuracy to predict heart failure better before it manifests.

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).[4]

Table 1: Performance Comparison (The following table is an accuracy comparison of the most recent modeling done on this dataset).

Model	Alotaibi[4]	UCI, 2019[5]	UCI, 2017[6]	UCI, 2017[6]
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 employed in this study, named 'Heart Failure Prediction Dataset,' is sourced from Kaggle.[7] The dataset combines five datasets with over 11 common attributes. These five datasets combine data from Cleveland, Hungary, Switzerland, and 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 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:

Prior to putting the data into the model, data preprocessing methods are applied to make it useful for modeling.[8] Data points, such as sex and ExerciseAngina, were converted to binary values of 1 and 0. One hot encoding was employed in

DOI: 10.36838/v6i5.1

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 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. To further increase accuracy, HyperOPT[9] and RandomizedSearch10 were used to finetune the hyperparameters for XGBoost and Random Forest, respectively.[9,10]

Random Forest:

The Random Forest algorithm addresses classification challenges. 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, each working on a subset of data, and the average is calculated to improve prediction accuracy.[11] Rather than relying on the prediction of a single tree, the Random Forest algorithm combines the trees using an estimated outcome and voting procedure.[12] 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 used for various tasks such as regression, classification, and ranking. XGBoost creates a model that boosts an ensemble of weak classification trees by gradient descent, which optimizes the loss factor.[13] 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

In binary classification problems, four primary metrics are considered: true positives, true negatives, false positives, and false negatives. 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. This research aimed to reduce the number of false negatives and false positives to improve accuracy in predicting heart disease. Using these models and methods, the result was 91.56% accuracy after cross-validation for XGBoost, with nine false negative cases out of 181 cases (see Figure 1). In Random Forest, there was 92.90% accuracy after cross-validation with six false negatives out of 181 cases (see Figure 2).

Compared 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 needed to be higher to implement machine learning approaches. Alotaibi artificially increased the size of the data 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 suitable for the artificial patient, thus creating useless data for the model. This causes much of the data to carry either a random target value or no target value, which would cause the model's accuracy to decrease[a].

Another metric used in binary classification problems is the ROC curve, the area under the receiver operating characteristic, a standard metric for evaluating binary classification models. A model with a higher AUC is thought of as a better model.[13] 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.

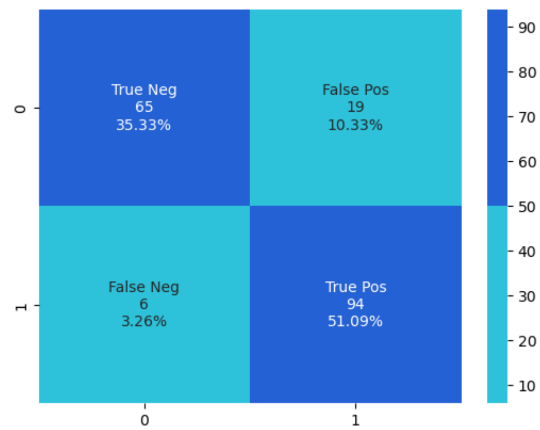

Figure 1: Confusion Matrix of XGBoost Mode
(The top left square shows the number of negative results from the model that were actually negative; false positives show the number of predictions that the model decided were positive but were actually false. False negatives show the number of predictions that the model decided were negative but were actually true, and true positives show the number that were actually positives.)[1]

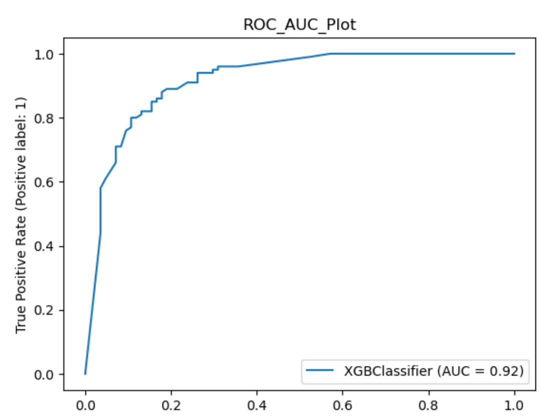

Figure 2: ROC_AUC Plot of XGBoost
(AUC ROC stands for the area under the receiver operating characteristic curve. This curve measures the ability of a binary classification model to distinguish between classes.)

[a] This may not be the case, but there is no evidence in the paper or the references to explain the choice made by Alotaibi[4]

DOI: 10.36838/v6i5.1

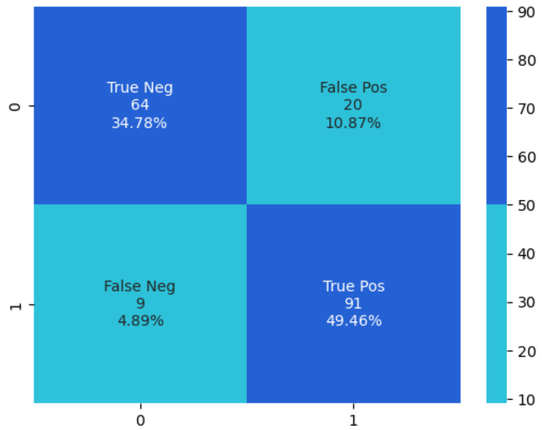

Figure 3: Confusion Matrix of Random Forest Model (Refer to Figure 1 explanation)

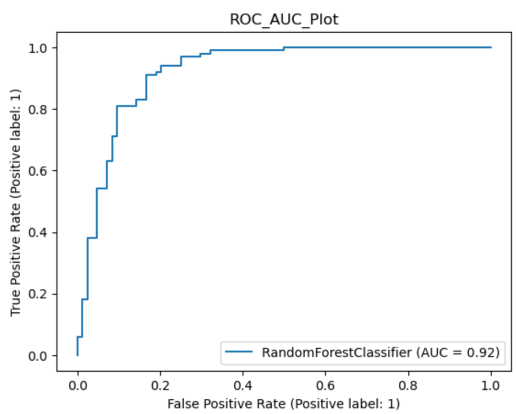

Figure 4: ROC_AUC Plot of Random Forest Model (Refer to Figure 2 explanation)

Table 2: Performance Comparison (The following table shows the performance of this study's models compared to the models from previous studies using the same dataset.)

Model	This Study	Alotaibi[4]	UCI, 2019[5]	UCI, 2017[6]	UCI, 2017[6]
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

This research demonstrates the potential application of machine learning within 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 patient data, allowing doctors to make predictions in real-time. Hypothetically, machine learning models could form a network. Accessible to healthcare providers worldwide, such a system could flag patients the models decide to show as being at significant risk of heart disease, prompting immediate intervention and follow-up. This would profoundly affect 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, underserved 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 the 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 need help to mirror large populations accurately, which is 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 expand the dataset to 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 incorporating machine learning with other predictive tools. With further development, these models could bring a new era of predictive healthcare in cardiology.

Conclusion

Heart failure represents a significant global health issue, affecting millions worldwide. Machine learning models, specifically Random Forest and XGBoost, can 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.

Acknowledgments

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

References

1. World Health Organization. Cardiovascular diseases. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (accessed March 15, 2023).
2. Savarese, G.; Lund, L. Global public health burden of heart failure. Cardiac Failure Review 2017, 3(1), 7-11.
3. Benjamin, E. J.; Muntner, P.; Alonso, A.; Bittencourt, M. S.; Callaway, C. W.; Carson, A. P.; et al. Heart disease and stroke statistics-2019 update: A report from the American Heart Association. Circulation 2019, 139(10), e56-e528.
4. Alotaibi, F. S. Implementation of machine learning model to predict heart failure disease. Int. J. Adv. Comput. Sci. Appl. 2019, 10(6), 261-268. DOI: 10.14569/IJACSA.2019.0100637.
5. Bashir, S.; Khan, Z. S.; Khan, F. H.; Anjum, A.; Bashir, K. Improving heart disease prediction using feature selection approaches. In 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST); 2019; pp 619-623.
6. Ekiz, S.; Erdogmus, P. Comparative study of heart disease classification. In 2017 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, EBBT 2017; 2017; pp 1-4.

DOI: 10.36838/v6i5.1

7. Ortega, F. Heart failure prediction dataset. Kaggle. https://www.kaggle.com/fedesoriano/heart-failure-prediction (accessed February 25, 2023).
8. Al-Mudimigh, F.; Ullah, Z. S. A.; Saleem, A. A framework of an automated data mining systems using ERP model. Int. J. Comput., Electr., Autom., Control Inf. Eng. 2009, 1(5), 598-605.
9. Komer, B.; Bergstra, J.; Eliasmith, C. Hyperopt-sklearn. In Automated Machine Learning: Methods, Systems, Challenges; Springer: 2019; pp 97-111.
10. Agrawal, T. Hyperparameter optimization using scikit-learn. In Hyperparameter Optimization in Machine Learning; Apress: 2021; pp 21-41.
11. Donges, N. The random forest algorithm. Towards Data Science. https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd (accessed March 15, 2023).
12. Bashar, S. S.; Miah, M. S.; Karim, A. H. M. Z.; Al Mahmud, A.; Hasan, Z. 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); 2019; pp 1-5.
13. Cui, Y.; Cai, M.; Stanley, H. E. Comparative analysis and classification of cassette exons and constitutive exons. BioMed Res. Int. 2017, 2017, 1-8.
14. Javeed, A.; Zhou, S.; Yongjian, L.; Qasim, I.; Noor, A.; Nour, R. An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection. IEEE Access 2019, 7, 180235-180243. DOI: 10.1109/ACCESS.2019.2950861.

■ Author

Aidan Gao is a junior 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.