Early Detection of Heart Disease using Machine Learning

ASTHA GUPTA, SNEHIL KESHARI, and GURMEHR SOHI, New York University

In this study, we employ machine learning-based methods to detect the presence of a risk for cardiavascular disease in patients. Our study seeks to overcome the issue of dataset size by utilizing five different heart disease datasets. The combined dataset is cleaned and pre-processed and then used for multiple classification algorithms. A comaprison of different methods show that the best performing algorithm was the Random Forest algorithm. We also identified the top features that influence the outcome the most. We identified that cholesterol, Echocardiography slope segment and type of chest pain are the most influential features when it comes to heart disease. The results of our study show that computer-aided methods can be used as screening methods for disease detection.

CCS Concepts: • **Machine Learning**; • **Heart Disease**; • **Supervised Learning**;

Additional Key Words and Phrases: disease detection, early detection, classification

ACM Reference Format:
Astha Gupta, Snehil Keshari, and Gurmehr Sohi. 2022. Early Detection of Heart Disease using Machine Learning. 1, 1 (December 2022), 10 pages.

1 INTRODUCTION

Cardiovascular diseases are a group of diseases which affect the blood flow and lead to heart issues. These diseases account for the highest number of deaths in the world (approximately 17.3 million people die each year owing to cardiovascular diseases) [10]. This number will only grow in the coming times.

Early detection can play a decisive role in reducing the fatality of this disease. If patients are diagnosed at an early stage, it would be possible for heath care practitioners to prescribe treatment options to the patients. Currently, the detection of heart diseases and related conditions are done by invasive methods which come with their own risks. These methods are also expensive and need trained physicians, meaning that they remain inaccessible to the economically marginalized communities. In order to deal with this problem, we need to come up with cost-effective methods to identify patients who are at a high risk of developing cardiovascular issues.

Over the last decade, the advances made in the field of machine learning have changed the landscape of heath care. Machine learning algorithms have been used to detect the presence of many diseases. Neural networks have been employed in the field of radiology to detect cancerous nodes and irregularities. Such methods have shown us that machine learning can be employed successfully in heath care use-cases. Not only do these methods reduce the cost involved in the task, but they also reduce the human effort required for the same.

2 METHODOLOGY

2.1 Dataset

One requirement that is needed to build robust machine learning screening or detection tools is the presence of a large dataset. This is often an issue in the field of medical sciences due to the unavailability of large-scale records of patients.

Authors' address: Astha Gupta, ag7982@nyu.edu; Snehil Keshari, sk9603@nyu.edu; Gurmehr Sohi, gs3541@nyu.edu, New York University .

Table 1. Data Features

Feature	Type
Age	Numerical
Sex	Categorical
Chest Pain Type	Categorical
Resting Blood Pressure	Numerical
Slope of the peak exercise ST segment	Numerical
Cholesterol	Numerical
Fasting Blood Sugar	Numerical
Resting Electrocardiographic Results	Categorical
Heart Rate maximum	Numerical
Exercise-induced ST depression relative to rest	Categorical
Fasting Blood Sugar	Categorical
Angina induced on exercise	Categorical
Heart Disease	Categorical

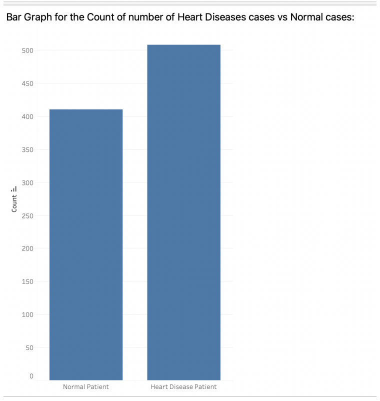

Fig. 1. Count of patients with heart disease vs Count of patients without heart disease.

Our study aims to overcome that challenge by combining different datasets on common features. Our dataset consists of 918 rows, with a total of 11 features like Age, Sex and Cholesterol values (shown in 1). This dataset was created by the University of California, Irvine [2].

2.2 Exploratory Data Analysis

Before jumping into the development of our classifier, it is important to perform analysis of the data, by means of different graphs and plots.

We use a barplot 1 to analyse the class distribution based on the outcome variable. The plot shows a class imbalance. This is an important conclusion. Going ahead, we cannot rely on classification accuracy as our sole metric as it could be misleading. Thus, to compare the performances of different models, we will need to rely on metrics that take class imbalance into account.

To gain some more insights regarding the class distribution, we use the following graphs. A bar graph is used to visualize the count of healthy and diseased patients based on their chest pain type, the resting echocardiography (ECG)

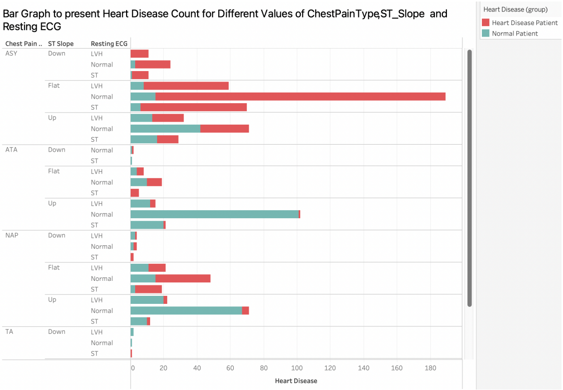

Fig. 2. Heart Disease patients based on chest pain type, ECG results and slope of ST Segments.

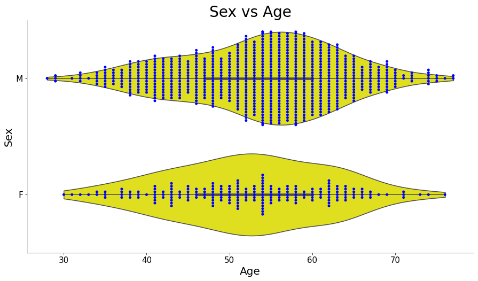

Fig. 3. Violin Plots to see distribution of age and gender.

results and the slope of ST segment during exercise (ST Segments show the results of an ECG) 2. The results to indicate that some conditions are more prevalent in people who have heart disease. For instance, heart disease counts are very high in patients with an asymptomatic chest pain and normal ECG results. We also used violin plots and swarm plots to see the distributions of the age, sex and ECG results amount the patients 4 3. Furthermore, we also use a scatter plot to see the distribution of cholesterol among male and female patients for all ages [5]. The relationship between gender and cholesterol is inconclusive from this plot. Both genders exhibit a diverse range of cholesterol values. Another visualization we use is to see which age groups have the highest number of heart disease. Another visualization we employ is a bar graph to see which age groups have the highest count of heart disease 6. Our results show that in our data, people between ages 50 and 60 have the highest number of heart disease patients.

For detection of outliers, we plotted box plots for our continuous features namely: cholesterol, age, resting blood pressure and maximum heart rate [7]. These figures will help us identify the presence of outliers in our data. We go into detail about their removal in the next section deal. All features except age have some outliers present.

We also use a heat map graph to check the Pearson correlation [8] between different features [8. This figure can help us identify those features which are highly correlated. If such features are detected, then we should remove one of the two correlated features from our data. Details for this filtering are presented in the next section.

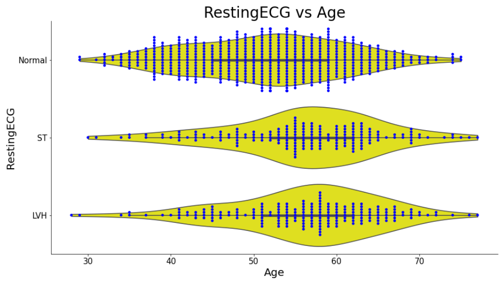

Fig. 4. Violin Plots for Age and Resting ECG results.

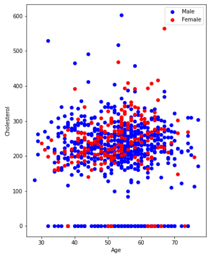

Fig. 5. Distribution of cholesterol values based on age and gender.

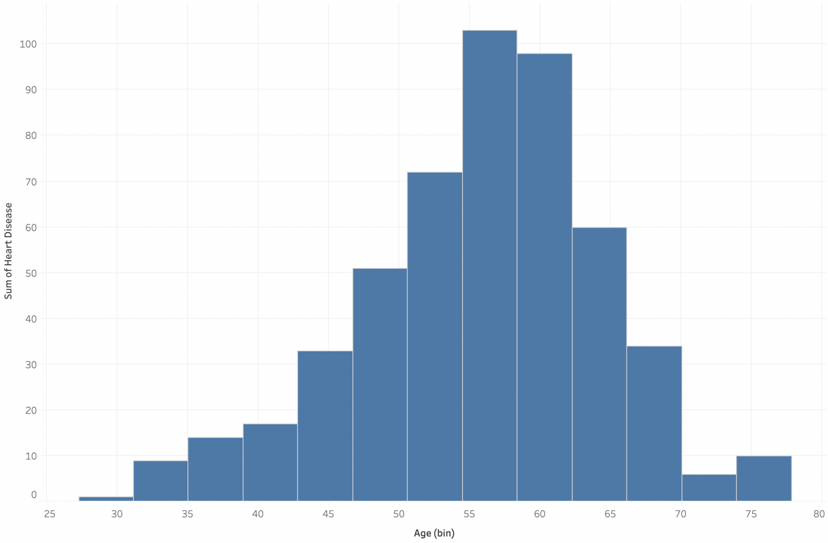

Fig. 6. Count of people with heart disease in different age groups.

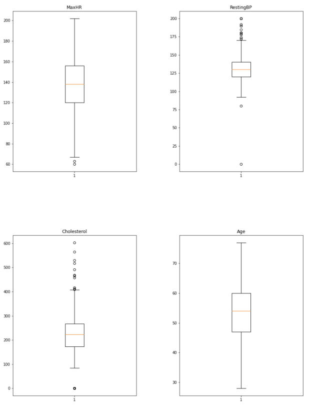

Fig. 7. Boxplots for continuous features for outlier detection.

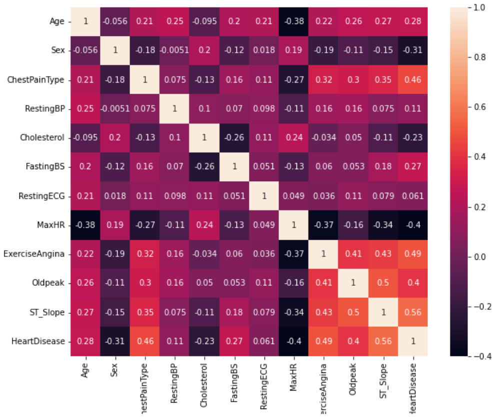

Fig. 8. Pearson correlation between data features.

2.3 Data Wrangling

Data cleaning is an important step in any machine learning application. A clean data set ideally leads to a better results as there is no missing data, outliers or correlated data features present.

For our study, we first analyze the data to check the presence for any missing values. Our results indicate that no null values are present in our data. Thus, we proceed with the next step which is feature encoding. Machine learning algorithms usually cannot work with categorical features. Thus, it is imperative that we convert any categorical features to numerical ones. In our data, features of sex, exercise angina, chest pain type, resting ECG and ST Slope are categorical. We assign each class a numerical value starting from 0.

Once the data does not have any null values and any categorical features, we can proceed to determine the Pearson correlation among different features. An absolute correlation value above 0.7 indicates that the two features are highly correlated [1]. Thus, we perform a check for all correlation values and if any two features have a high correlation value, we decide to drop that feature. Based on our outcome, we did not need to drop any columns from our data.

The final data wrangling step we undertook was the removal of outliers from continuous features. For this step, we used a Z-score [14] to determine which rows to drop. For each feature with outliers, we filtered the data such that only values which have a Z-score value between -3 to 3 are kept in the data. Any values with scores lower than -3 or higher than 3 were dropped.

Our final data set had a total of 906 rows. Then, we performed a train-test split on the data. The training set comprised of 80% of the data, and the test set had the remaining 20%.

2.4 Training

For our study, we wanted to compare the performance of multiple algorithms to see which one gave the best performance. The first model we used was Logistic Regression. Logistic Regression is very widely used for binary classification tasks such as ours [7]. The algorithm uses a sigmoid function on top of a linear regressor to predict the probability that the given example belongs to the positive class. Next, we utilized the Naive Bayes classifier [15]. This algorithm uses the conditional probability to predict the output class of an example. The algorithm assumes that all variables are independent from one another.

Another popular algorithm is the Support vector algorithm [6]. This method tries to find the separating boundary plane between the two classes. The best boundary is the one that maximizes the distance between the two closest points of the two classes.

Finally, we also use tree-based decision algorithms like Random Forest [3], eXtreme Gradient Boosting (XGBoost) [5] and Gradient Boosting [11]. These methods perform splits on the data based on the values of different features to come up with decision trees which decide the output variable. Gradient Boosting is a more evolved form of Random Forest where each tree in the forest of trees is built successively (one after the other).

For each model, we printed the classification accuracy, precision. recall, F1 score and the Area Under the Receiver Operating Curve (AUROC) scores [4]. We also used the feature importance methods provided by the Random Forest, Gradient Boosting and XGBoost algorithms to determine which features affect the outcome of our problem the most.

3 RESULTS

Since we are dealing with a slightly imbalanced dataset, classification accuracy is not the best metric to use to compare model performances. For this reason, we used the AUROC values to compare models. This metric uses the True Positive Rate and the False Positive Rate to adjudge the outcome. A high value of the AUROC score implies a better model.

The figures [9 - 14] show the Receiver Operating Curves for all the models we trained. We can see that XGBoost was the best one among all, with an AUROC score of approx. 0.93 and the best Precision and Recall Scores.

Table [2] compares the precision, recall, accuracy score and roc scores for all models.

Looking at the feature importance bar plots, we can say that ST_Slope (Slope of ST_Segment), Chest Pain Type and Cholesterol were the common features in the top five feature importance values among all the three decision tree based methods.

4 DISCUSSIONS

The aim of our study was to build a classifier for the detection of heart disease using supervised learning methods. Our results show that we were able to leverage machine learning models for our task. All algorithms we used give us a

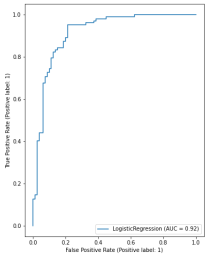

Fig. 9. Logistic Regression ROC Curve.

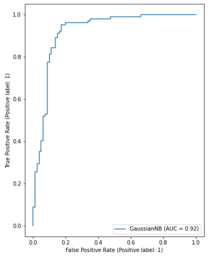

Fig. 10. Naive Bayes ROC Curve.

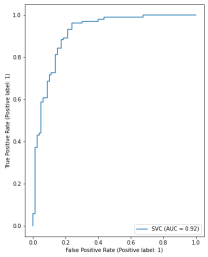

Fig. 11. Support Vector ROC Curve.

Table 2. Comparsion of Different Classification Methods

Model	Precision	Recall	F1-Score	Area under the ROC Curve
Logistic Regression	0.84	0.84	0.84	0.92
Naive Bayes	0.88	0.87	0.88	0.92
Support Vector	0.85	0.85	0.85	0.92
Random Forest	0.85	0.85	0.85	0.92
Gradient Boosting	0.85	0.85	0.85	0.93
XGBoost	0.87	0.87	0.87	0.93

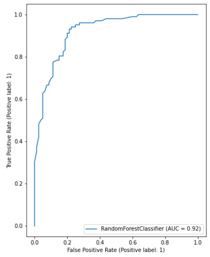

Fig. 12. Random Forest ROC Curve.

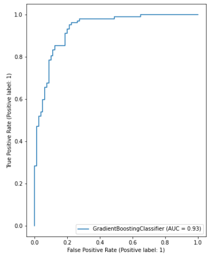

Fig. 13. Gradient Boosting ROC Curve.

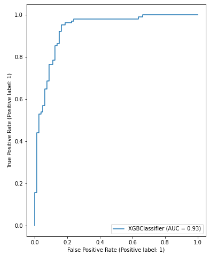

Fig. 14. XG Boost ROC Curve.

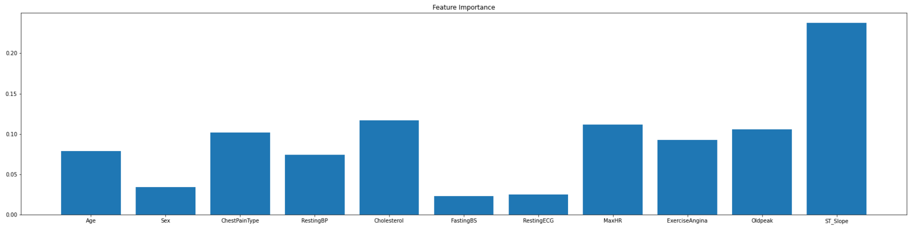

Fig. 15. Random Forest Feature Importance

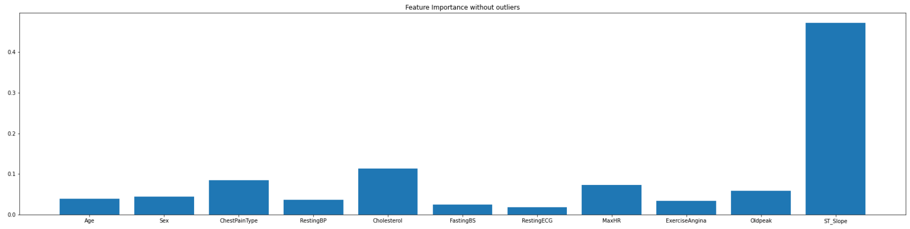

Fig. 16. Gradient Boosting Feature Importance

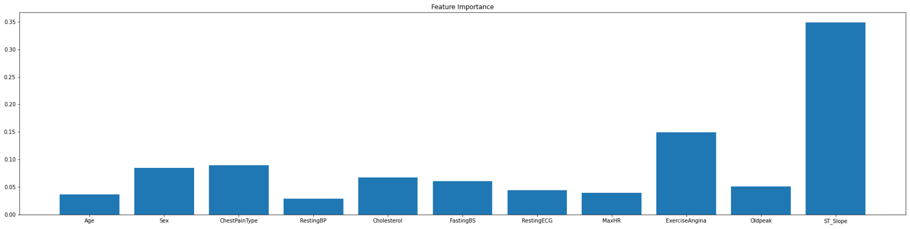

Fig. 17. XG Boost Feature Importance

good performance and the best model found is consistent with the literature that exists. Tree-based methods have been known to outperform their counterparts in classification tasks.

The visualizations we performed during the data analysis paved the way for the data wrangling part of our study. It was important to study the distribution of data with respect to the features. Correlation study helped us ensure that the features were independent of each other. Similarly, the outlier detection enabled us to clean our data further.

As for the feature importance values achieved, it does match our expectations for heart disease. Cholesterol measures determine the level of fatty deposits in our arteries which can lead to blockage of blood flow, impacting the heart directly [9]. Chest pain is often associated with cardiovascular health and can be associated to the risk of a heart disease [13]. Slope of ST_Segment relate to the results of echocardiography, and thus can be a direct indication of underlying cardiovascular issues [12].

Another observation is that age did not play a very important role in the risk of disease. This is not surprising given that with changing diets and lifestyle, young people are also prone to develop issues related to heart diseases.

5 CONCLUSION

Through our study, we were able to leverage a combined heart disease dataset to build a robust classifier for detecting heart disease. Our best model was the eXtreme Gradient Boosting algorithm which achieved an AUROC score of 0.93. Similarly, we identified the top three features that impact heart disease. The results we achieved indiacte that machine learning methods can be successfully employed for screening tools. Over time, researchers can refine this study and use an even larger dataset to build applications that can be deployed for use in areas where people may not have access to large hospitals. In the future, computer-aided technologies can significantly reduce the cost of diagnostics and act in tandem with professionals to add another layer of screening.

REFERENCES

[1] Haldun Akoglu. “User’s guide to correlation coefficients”. In: *Turkish Journal of Emergency Medicine* 18.3 (2018), pp. 91–93. ISSN: 2452-2473. DOI: https://doi.org/10.1016/j.tjem.2018.08.001. URL: https://www.sciencedirect.com/science/article/pii/S2452247318302164.

[2] Matthias Pfisterer Andras Janosi William Steinbrunn and Robert Detrano. *UCI Machine Learning Repository.* 2021. URL: https://archive.ics.uci.edu/ml/datasets/heart+disease.

[3] Leo Breiman. “Machine Learning, Volume 45, Number 1 - SpringerLink”. In: *Machine Learning* 45 (Oct. 2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[4] Gürol Canbek et al. “Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights”. In: *2017 International Conference on Computer Science and Engineering (UBMK).* 2017, pp. 821–826. DOI: 10.1109/UBMK.2017.8093539.

[5] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* KDD ’16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145/2939672.2939785.

[6] C. Cortes and V. Vapnik. “Support Vector Networks”. In: *Machine Learning* 20 (1995), pp. 273–297.

[7] Jan Cramer. *The Origins of Logistic Regression.* Tinbergen Institute Discussion Papers 02-119/4. Tinbergen Institute, 2002. URL: https://EconPapers.repec.org/RePEc:tin:wpaper:20020119.

[8] David Freedman, Robert Pisani, and Roger Purves. “Statistics (international student edition)”. In: *Pisani, R. Purves, 4th edn. WW Norton & Company, New York* (2007).

[9] S. M. Jeong et al. “Effect of Change in Total Cholesterol Levels on Cardiovascular Disease Among Young Adults”. In: *J Am Heart Assoc* 7.12 (June 2018).

[10] Lawrence J. Laslett et al. “The Worldwide Environment of Cardiovascular Disease: Prevalence, Diagnosis, Therapy, and Policy Issues”. In: *Journal of the American College of Cardiology* 60.25_Supplement (2012), S1–S49. DOI: 10.1016/j.jacc.2012.11.002. eprint: https://www.jacc.org/doi/pdf/10.1016/j.jacc.2012.11.002. URL: https://www.jacc.org/doi/abs/10.1016/j.jacc.2012.11.002.

[11] Alexey Natekin and Alois Knoll. “Gradient Boosting Machines, A Tutorial”. In: *Frontiers in neurorobotics* 7 (Dec. 2013), p. 21. DOI: 10.3389/fnbot.2013.00021.

[12] J. H. O’Keefe et al. “ST-segment elevation: defined by the company it keeps”. In: *Mayo Clin Proc* 87.7 (July 2012), pp. 610–613.

[13] J. Robson et al. “Clinical value of chest pain presentation and prodromes on the assessment of cardiovascular disease: a cohort study”. In: *BMJ Open* 5.4 (Apr. 2015), e007251.

[14] Songwon Seo. “A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets”. In: 2006.

[15] Harry Zhang. “The Optimality of Naive Bayes”. In: *The Florida AI Research Society.* 2004.