Acta Scientific MEDICAL SCIENCES (ISSN: 2582-0931)

Volume 7 Issue 10 October 2023

Research Articles

Novel Computational Approach for Early Diagnosis of Cardiovascular Diseases

Saharsh Reddy Peddireddy*

Senior - Commack High School, 1Scholar Lane, Commack, New York, USA

***Corresponding Author:** Saharsh Reddy Peddireddy, Senior - Commack High School, 1Scholar Lane, Commack, New York, USA.

DOI: 10.31080/ASMS.2023.07.1683

Received: September 11, 2023

Published: September 29, 2023

Abstract

Background: Heart disease is the leading cause of death in the US, costing over $200 billion. Current detection methods include invasive angiography, but these have limitations. Machine learning models using computational methods offer precise and trustworthy solutions.

Methodology: The study tested various feature selection techniques on two datasets from the UC Irvine Machine Learning Repository, analyzing their performance using ROC curves, accuracy, and precision metrics. The ET classifier performed best in accuracy, sensitivity, and specificity.

Results: The study tested five machine learning models without feature selection techniques, revealing their high accuracies and sensitivities. The predicted model (RF) came close to the top, while the KNN model had low accuracy. Decision trees and AdaBoost were also tested. The ET classifier, combined with feature selection techniques, performed better than the other models. The Relief method had the highest chi-squared value, suggesting that Relief feature selection can improve accuracies similar to optimal settings.

Discussion: The study showcases the effectiveness of machine learning models, with all five models showing decent accuracy, even with faulty data. The ET classifier, which combines decision trees and ensemble methods, outperformed all other models. It handles numerical and categorical data, handles missing values, and outliers. The LASSO technique, which may have removed essential features, decreased accuracy compared to no feature selection techniques. When combined with feature selection techniques, the Relief method achieved the highest chi-squared value, indicating improved accuracies. This combination has shown promising results in previous studies, making it a suitable choice for this test.

Conclusion: The project aimed to predict heart disease using a computational model and identified the most effective combination of feature selection techniques and classification algorithms. High accuracies were achieved, with the Random Forest algorithm excelling.

Keywords: Computational; Diagnosis; Cardiovascular; Diseases

Introduction

Heart disease is the leading cause of death in the United States, with around 700,000 people killed every year in the US alone.[1] In addition, heart disease is rapidly increasing in both developed and undeveloped countries with the disease costing countries more than 200 billion dollars [1]. Over ten per cent of adults ages 20 andover develop coronary heart disease (CAD) and cardiovascular disease (CVD), the most common types of heart disease.[2] Heart disease is becoming more and morefatal by the day, and this is mostly due to late detection and diagnosis of the disease [2]. Current

methods of heart disease detection are invasive-based techniques such as angiography, but these techniques have their limitations. These include overestimating or underestimating severity, not detecting atheroma's, and sometimes ineffective in predicting myocardial infarction. These are in addition to the technological knowledge, long process due to human-error, and high-end tools required for angiography [3].

Whether you notice it or not, programming and code-related technologies are taking shape in our everyday lives and is one of the largest and most successful markets. Machine learning is a branch of artificial intelligence that creates algorithms to analyze large amounts of data. A model is essentially the "machine" in these types of scenarios and models use human bias, time constraints, and other factors that may affect decisions of a human [4]. Machine learning is a rising technology that is being incorporated all around us to improve society. Data analysis could be used along with machine learning to assist in identifying the most useful data out of a large dataset [5]. This combination between machine learning and data analysis could be really useful and powerful in real life applications, especially in the medicine world. Utilizing feature selection techniques allows for strong, practical data that can be used for computational models [6].

These new technologies can allow for early and accurate detection of heart disease as a new solution.

Expected outcomes and hypothesis

Currently, the bulk of the positive results of cardiovascular disease detection is based on invasive techniques. But as mentioned, these techniques rely heavily on human work and are not entirely accurate. Consequently, it is an absolute necessity for a non-invasive computational program to detect heart disease. As most of the fatalities resulting from CVD and CAD are due to late detection, a computer model will increase efficiency and overcome this obstacle at least by a decent margin. Building a program may be the best solution because this approach is the quickest, easiest to handle, and could be done from any location.

Based on a previous personal study with HIV/AIDS and other studies in the field of disease detection prediction, the Random Forest (RF) classification model will best benefit the program in making predictions. Random Forest is a supervised algorithm and uses ensemble learning for regression. RF regression is a strong choice because it combines predictions of multiple different machine learning models instead of one, more specifically, the mean of these models. Since outliers will be eliminated with the feature selection techniques, utilising the means is a promising approach. Additionally, one of the most vital aspects of the RF model is that it reduces overfitting in decision trees with extra features, and this leads to comparatively high accuracies. Random Forest is not usually incorporated into a multi-step program as it follows a rule-based approach and can automate missing values, but it is a powerful tool and was interesting to see how it would perform within the realm of heart disease detection.

Various feature selection techniques were also incorporated to "clean up" the data and make the data better for the models to run on. The Least Absolute Shrinkage and Selection Operator (LASSO) feature selection tool is a capable option that will benefit the program. LASSO Regularization (L1) is an embedded method commonly used in combination with RF Models and eliminates the disadvantages of Random Forest regression. Even more so, LASSO generally results in higher accuracies and favours data with less collinearity and a large number of features. Going off of these points, LASSO (L1) Regularization seemed as the best choice for the model.

Methodology

Data extraction

Having a robust and well-organized dataset reflects the performance of the model. The University of California Irvine Machine Learning Repository contains an abundant amount of datasets, and the repository contains four datasets on heart disease. The datasets contain 76 attributes, but studies only use 14 of them, including age, sex, chest pain, BP, cholesterol, blood sugar, and more for over 1000 instances. The Cleveland database is the one with the most published studies so we selected that one, and we also selected the one from the University Hospital of Basel, Switzerland. These two datasets felt like a good base for this model.

Feature selection techniques to preprocess data and regression models

The model will be tested without any feature selection techniques and then the techniques will be implemented to compare and evaluate the effect on the model. The feature selection tools

used were Fast correlation-based filter (FCBF), Variance Threshold, Forward Feature Selection, Least Absolute Shrinkage and Selection Operator (LASSO), and Relief. These techniques are all different in some ways and are different categories of feature selection, such as filter methods, wrapper methods, embedded methods, and hybrid methods. The raw data was downloaded as a comma-separated value (CSV) format file, and Python was the preferred language used to build the model. The model was built with the help of some libraries such as pandas and sci-kit-learn and others that helped construct the regression models. Additionally, libraries were used for visualisation, as it is vital for good understanding, for example, matplotlib for creating ROC curve line graphs (specificity vs. sensitivity). Several different individual pieces of code were developed to test each combination of feature selection and regression, and the results were analysed.

Accuracy of the model and analysis

Previously, a decision tree was used in another project, which was relatively easier to analyze and visualize. So, another method was needed to analyze and compare the most effective methods. A confusion matrix was developed to keep track of four outcomes: TruePositive, True Negative, False Positive, False Negative. Based on these four values it was possible to retrieve the accuracy, precision, Matthews correlation coefficient, and ultimately for visualization, a ROC curve. In addition to the graph, the accuracies for the different combinations were displayed in data tables, for readability purposes of the reader. These processes were automated for simplicity and less work.

Results

Model	Accuracy	Sensitivity	Specificity
Random Forest	92.42	92.46	92.08
K-Nearest Neighbors(k = 5)	88.34	89.12	87.25
Extra-Tree	96.44	95.96	97.32
Decision Tree	85.78	89.43	85.11
AdaBoost	92.30	92.18	91.79

Table 1: No Feature Selection Techniques.

With no feature selection techniques implemented, for all five models the accuracies were prettydecent with high sensitivities and specificities as well. This strongly demonstrates the power of these models and the capabilities of machine learning as even with some faulty data present techniques are used by the models to redirect through them. The predicted model (RF) did not perform the best in this scenario, but still came close to the top and regarding the KNN model, only k = 5 was tested, so the low accuracy was reflected off of that choice. Similarly, a decision tree theoretically doesn't make the most sense to use in this scenario, but it is a good mark to compare against other models. Finally, AdaBoost is a technique that is comparatively newer thanthe others and came out close to the top, but the ET classifier was a clear winner.

Model	Accuracy	Sensitivity	Specificity
Random Forest	86.48	88.05	84.44
K-Nearest Neighbors (k = 5)	83.98	82.76	85.20
Extra-Tree	89.41	88.93	90.34
Decision Tree	84.36	83.66	85.02
AdaBoost	85.97	86.41	85.28

Table 2: LASSO Feature Selection Technique.

Again, the ET classifier performed considerably better although contradictory to the hypothesis, the LASSO technique brought down accuracy compared to no feature selection techniques at all.

Feature Selection Method	P-value	Chi-squared value
LASSO	0.0000078186	84.9204
FCBF	0.0000089352	110.2975
Relief	0.0000056396	114.9927

Table 3: Conclusion of Benefit of Feature Selection Techniques.

The ET classifier was used in combination with these feature selection techniques for this test asthe ET classifier performed far better than the other models in the other tests. The P-value for all the methods was quite small, but the difference occurred in the chi-squared values, where the Relief method resulted in the highest value. With the ET classifier, the Relief feature selection will improve accuracies similar to the optimal feature settings.

Discussion

The study demonstrates the power of machine learning models, with all five models showing decent accuracy with high sensitivities and specificities even with faulty data. The predicted model (RF) came close to the top, while the KNN model had low accuracy due to its choice of k=5. AdaBoost and the ET classifier were the clear winners.

The ET classifier outperformed all other models in this scenario. Its high accuracy and performance make it the ideal choice for this particular problem. The ET classifier combines the strengths of decision trees and ensemble methods, making it a powerful tool for classification tasks. With its ability to handle numerical and categorical data, it is well-suited for the dataset. Additionally, the ET classifier's ability to handle missing values and outliers further enhances its performance. Overall, the ET classifier proves to be the most effective model for this scenario.

Again, the ET classifier performed considerably better, although contradictory to the hypothesis, the LASSO technique brought down accuracy compared to no feature selection techniques. This unexpected result suggests that the LASSO technique may have removed essential features from the dataset, leading to decreased accuracy. Further investigation is needed to understand the underlying reasons for this outcome. Additionally, it is worth noting that the performance of the ET classifier indicates that the selected features could capture relevant information for accurate classification.

When combined with feature selection techniques, the ET classifier outperformed other models in this test, with the Relief method achieving the highest chi-squared value, indicating improved accuracies.

Additionally, the Relief feature selection method offers the advantage of reducing the dimensionality of the dataset, which can be particularly beneficial in cases where computational resources are limited. By selecting the most relevant features, the Relief method helps to focus the classifier's attention on the most informative attributes, thereby improving the model's overall accuracy. This combination of the ET classifier and Relief feature selection has shown promising results in previous studies, making it a suitable choice for our current test.

Conclusion

The purpose of this project was to accurately make predictions in detecting heart disease using a computational model and which combinations of feature selection techniques and classification algorithms would be the most successful. The first goal was definitely met as pretty high accuracies were achieved with the computer model and this method is much quicker than invasive techniques like angiography. The Random Forest algorithm as predicted performed comparatively well but was beat out by Extra-Tree classification in every scenario. Additionally, surprisingly the LASSO feature selection affected the model negatively as accuracies went down when no feature selection was applied. One suggestion that should have been made was testing different values for k for the K-Nearest Neighbor algorithm. Maybe higher accuracies could havebeen achieved, swaying the results a bit. Since the ET algorithm performed the best, testing was done to see with what combination of feature selection it performed best in, the answer being theRelief method. With the Relief feature selection the accuracy of ET went up 2% to 3% putting the overall accuracy just under 95%. Overall, compared to previous studies the results achieved in this project are quite amazing and with more and more data available to train the models it is expected that these computational algorithms will become more reliable by the day. For future work, other feature selection tools and classification algorithms can be tested, maybe newly developed ones and the effect on the model can be investigated. Furthermore, an application or other user interface can be developed for consumers to use the model.

Bibliography

1. "American heart month toolkits 2023". Centers for Disease Control and Prevention (2023).
2. Chauhan NS. "Model Evaluation Metrics in Machine Learning". *KD nuggets* (2020).
3. Dagli Y. "Feature selection using relief algorithms with python example".
4. Gupta A. "Feature selection techniques in machine learning". *Analytics Vidhya* (2020).
5. Jordan MI, Mitchell TM. "Machine learning: Trends, perspectives, and prospects". *Science* 349.6245 (2015): 255-260.
6. Muhammad Y., *et al*. "Early and accurate detection and diagnosis of heart disease using intelligent computational model". *Scientific Reports* 10.1 (2020): 19747.