

Using Machine Learning Algorithms for Assisting Heart Disease Diagnosis

Jiaxi Miao

USA Veritas prep school, Beijing, China

miaojiaxi@me.com

Abstract. Heart disease continues to be one of the top causes of death around the world, which highlights the importance of catching it early to ensure timely treatment. This study explores the use of various body health information, such as Age, Gender etc. and physiological responses such as chest pain, shortness of breath, Fatigue, Dizziness etc. The meaning of this study is to find out how does different body responses correspond to if the patient catches heart disease and which kind of method of diagnosis would be more accurate and efficient. This research investigates applying various types of classifier models finding the better performance in heart disease diagnosis. Kaggle data set is used in this research. To evaluate the performance of different clarification models to choose the most validate model and analyze or determine important features to help doctors' diagnosis. At the end, the study found out that models including SVC and Random Forest model are the accurate ones while studying the data.

Keywords: Data mining; machine learning; heart disease.

1. Introduction

Cardiovascular disease (CVD) has always been one of the leading causes of death worldwide for years. The heart is the body's most important organ for transporting blood through it. Many people die suddenly due to this unexpected disease annually. The World Health Organization reported that in 2016, cardiovascular diseases (CVDs) were responsible for approximately 17.9 million deaths, accounting for 31% of all deaths worldwide. Among these, heart attacks and strokes made up about 85% of the total [1]. As many people have not noticed, many factors including lifestyle, genetics, and health conditions all affect the health of one's heart. According to Columbia Heart Surgery, approximately 25% of all heart attacks go unrecognized, especially among women, highlighting the importance of regular health check-ups [2]. Therefore, analyzing various predictive models based on patients' body reactions and evaluating their performance in detecting heart disease is crucial.

Previous and older diagnostic methods mostly rely on medical tests and measurements such as electrocardiograms and blood tests. Doctors and researchers analyzed aspects such as heart rate, blood pressure and cholesterol level. Although those methods are useful, but they have not captured the relationship between some implicit but also important variables. Traditional methods for predicting heart disease involve multiple steps that require the doctors to examine the results of several tests closely, making the process highly time-consuming. On the other hand, heart disease prediction heavily relies on the knowledge and experience of medical professionals. However, a same result could be interpreted differently while showing it to different doctors. In this case, applying Artificial Intelligence (AI) technology could successfully avoid all those implicit problems. Nowadays, technology progressed rapidly and researchers have made many new achievements. For instance, Deep Mind collaborated with Moor fields Eye Hospital in London to develop an AI model that predicts cardiovascular health risks through retinal scans [3], showing AI's potential to help patients predict the risk. As a result, it is essential to utilize automated technologies in heart disease diagnosis to help doctors identify earlier with high accuracy. It's not feasible to let the patients do all tests which are really time consuming; AI should be used to predict and calculate which combination of tests would be the fastest and the most accurate one.

The research separated the data into three parts, using different models to see which combination could help doctors to determine whether the patient will catch heart disease. The data includes the body reaction of 18 different body indexes from patients. The categories not only include basic information for medical examination but also include patient's basic health information such as age and family history and diagnosis information. The research applies models including the random forest model, decision tree model and support vector classifier to prove which combination could be more effective and cost less. This research is important for bridging the gap between traditional medical assessments and modern AI techniques. Furthermore, this study contributes to the improvement and process of AI-driven healthcare solutions.

2. Method

2.1. Data Set Preparation

2.1.1. Data set description and settings.

To prevent over-fitting during the modeling process, cross-validation should be applied. In this study, the dataset was split into two parts: 80% was used as the training set to build the model, while the remaining 20% served as the test set to evaluate the model's performance on unseen data. The data set was sourced from Kaggle. It contains 18 attributes, the total samples are 70K. 69K data set is used in the model, 1K is reserved to verify the accuracy manually. In this project, a dataset provided by Kaggle was used [4]. The dataset contains 70,000 patients with their age, family history, chest pain, shortness of breath, fatigue, palpitation, dizziness, swelling, pain arms jaw_back, cold_sweats nausea, high Blood Pressure (BP), high cholesterol, diabetes, smoking, obesity, sedentary lifestyle, chronic stress, gender and whether they caught heart disease. Then the study will analyze those data and the result of whether they will get heart disease to see if they have some traits in common. The data mining process typically begins with extracting the data, followed by transforming it and loading it into the system. After these steps, the data is stored in a warehouse for further analysis. Then manage the data into a multidimensional data base system. With the help of high technique software, abstract the useful data for the researchers to use them in the following. There are different types of analysis including data visualization, rule induction, nearest neighbor method, clustering, generic algorithm, decision tree model etc. In the data base, the users already knew the results. Also, Data mining can discover unexpected patterns and results that haven't been considered in the beginning.

2.1.2. Proposed machine learning approaches.

In this study, it used Decision Trees, Random Forests, and support vector classifier. A decision tree is a basic, yet powerful supervised learning algorithm commonly used for both classification and regression. It works by following a tree-like structure, where decisions are made through a series of questions or conditions. At each internal node, a feature is tested; each branch shows a possible outcome of that test, and the leaf nodes provide the final prediction — either a category for classification or a number for regression. For each study subject, a final classification is obtained by a majority vote across the classification trees grown in the different bootstrap samples [5]. Decision trees have wide applications in many fields. In medical diagnosis, they can assist doctors in judging diseases based on various symptoms and examination results. Random forest is another important supervised machine learning algorithm. A random forest is composed of multiple decision trees. It uses the bagging (bootstrap aggregating) technique to build different decision trees on bootstrap samples of the training data. Each tree in the forest is trained independently then makes its own predictions. The final prediction of the random forest is usually determined by aggregating the predictions of all individual trees, often using methods like majority voting for classification tasks or averaging for regression tasks [4]. The reason for keeping the tree size small is based on the principle that simpler explanations are generally favored when accounting for a set of phenomena. Also, small trees produce decisions faster than large trees, and they are much easier to look at and understand [6, 7]. It can face complicated data with a large number of features effectively and can automatically

select important features. The Support Vector Classifier (SVC) model is a significant supervised learning algorithm in the machine learning domain, primarily used for classification tasks [8-10]. SVC is for finding an optimal hyperplane in a sophisticated space that divides different classes of data. The decision boundary separates the two classes is chosen such that the margin between the closest data points of different classes (support vectors) is maximized. SVC has a wide range of applications in various fields such as image classification, text classification, bioinformatics (e.g., gene expression data classification), and medical diagnosis (e.g., disease prediction). It can handle both small and large datasets, although for very large datasets, computational resources and time may be a cancer.

3. Results and Discussion

When comparing model accuracy, all three models achieved over 98% accuracy on the test dataset. As shown in Table 1, the differences among them are minimal and not statistically significant.

Table 1. Models' accuracy

Name	Percentage
Decision Trees	98.2%
Random Forests	99.2%
SVC	99.1%

Decision trees and random forests have different feature importance scores, and the importance of different features is also ranked differently. However, from the two models, the first 9 features are quite consistent Table 2. Ages show the high importance score in both Decision Trees model and Random Forests Model, given other features only have two options of yes/no, age has a much wider distribution. Further analysis was carried out to examine the relationship between heart disease and age distribution. As shown in Fig. 1, individuals aged 45 to 68 appear to be more affected by or concerned about heart disease compared to other age groups. One possible reason is that many middle-aged individuals are so focused on their jobs that they tend to overlook regular health check-ups. In today's fast-paced and competitive world, the pressure of supporting a family and meeting work demands often leaves little time or energy for personal health. As a result, they may only seek medical attention when symptoms start to appear, which could explain the higher prevalence rate among this age group.

Table 2. Feature importance

Feature Name	Decision trees score	Random forests score
cold sweats Nausea	0.3644	0.1175
Age	0.2874	0.1540
Chest Pain	0.0752	0.0988
Palpitations	0.0589	0.0814
Shortness of Breath	0.0573	0.0824
Fatigue	0.402	0.0902
Pain Arms Jaw Back	0.0290	0.1000
Dizziness	0.0255	0.0865
Swelling	0.0232	0.1014
Sedentary Lifestyle	0.0060	0.0118
Smoking	0.0049	0.0090
High Cholesterol	0.0048	0.0104
Family History	0.0046	0.0126
High BP	0.0044	0.0117
Diabetes	0.0043	0.0091
Chronic Stress	0.0039	0.0094
Gender	0.0034	0.0045
Obesity	0.0025	0.0093

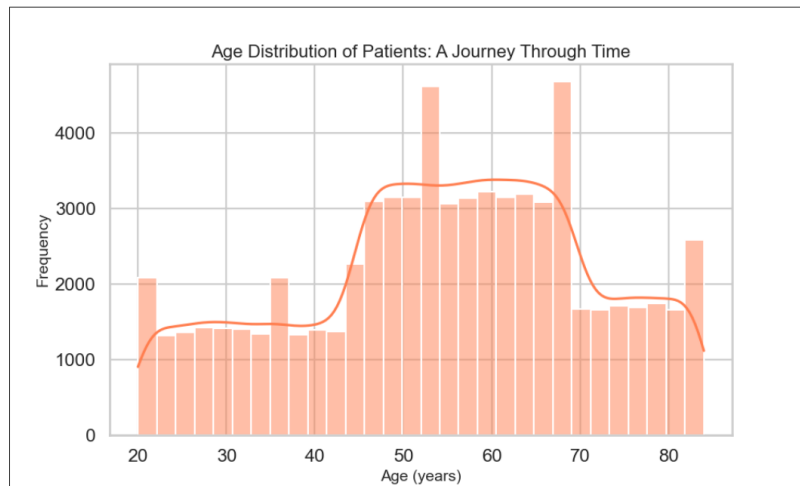


Figure 1. Age distribution of patients (Picture credit: Original)

From feature importance score, it can be identified that Category 1 and Category 3 run a very important role in the model accuracy. Below table shows how's accuracy with different category features Table 3.

Table 3. Feature category accuracy

Category#	Decision Trees	Random Forests	SVC
Category1	89.78%	89.80%	89.11%
Category1\$2	90.84%	91.33%	91.84%
Category1&3	98.04%	98.85%	98.99%

Decision Trees (DT) perform worse than the other two models because it's over-fitting, decision trees tend to over-fit the training data, capturing noise instead of generalizable patterns. Second is because of the instability, a small change in the dataset can lead to a completely different tree structure, making DT highly sensitive to variations. Last is because of locally optimal splitting, since DT uses a greedy algorithm for selecting split points, it may not always find the globally best solution. On the other hand, Random Forest (RF) performs better because it consists of many decision trees so there won't be a single tree overfitting the whole. It could balance variances of single decision trees, leading to a more well done model. Some features are not as important as other features such as age. Although women could have a higher possibility of catching a heart disease, since the body structure of women and men have little difference. However, this feature is relatively low in impact while comparing with other features.

4. Conclusion

This study compared the performance of three common classification models under different features. All three methods showed high accuracy and there was not much difference between them, but Random Forests and SVC shows slightly better performance if not all features are available. This result is of great significance for medical workers to better screen outpatients with heart diseases. For large-scale screening, potential patients only need to provide the most basic health data. By using the models, it is possible to quickly identify the group of patients who are likely to have heart diseases, with a recognition rate of nearly 98%. At the same time, this result also proves that the diagnosis of heart diseases does not require medical instrument examination, thus reducing costs and accelerating the diagnostic speed. However, while inputting too much data inside the models, it requires longer time to process the data and also after getting the results, it fails to give professional explanations to explain the reason for the complex relationship between them. In the future, the study would increase models that could explain the reasoning of the results of current present models such as SHAP or LIME, to better understand the relationships between features and predictions. On the other hand,

cooperating with professionals' doctors will help validate the model in settings and guide it into real use in real life. Finally, the goal is to develop an efficient, explainable and accessible diagnostic system for early heart disease detection

References

- [1] World Health Organization. Cardiovascular diseases (CVDs). WHO, 2020. [https://www.who.int/zh/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/zh/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)).
- [2] Columbia University Department of Surgery. 10 facts you may not have known about heart attacks. Columbia Surgery, 2014. <https://columbiasurgery.org/news/2014/07/28/10-facts-you-may-not-have-known-about-heart-attacks>.
- [3] Wagner SK, Chopra R, Ledsam JR, Askham H, Blackwell S, Faes L, Balaskas K, Back T, Keane PA. Diagnostic accuracy and interobserver variability of macular disease evaluation using optical coherence tomography. *Investigative Ophthalmology & Visual Science*. 2019 Jul 22; 60 (9): 1849-1849.
- [4] Kaggle. HeartRiskAI: Predicting heart disease risk, 2025, <https://www.kaggle.com/code/mahatiratusher/heartriskai-predicting-heart-disease-risk/notebook>.
- [5] Suzuki H, Kawakami T, Ito T, et al. Explainable and local correction of classification models using decision trees. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, 36 (8).
- [6] Haddouchi M, Berrado A. A survey of methods and tools used for interpreting random forest. In: 2019 1st International Conference on Smart Systems and Data Science (ICSSD). IEEE, 2019.
- [7] Zhu T. Analysis on the applicability of the random forest. *Journal of Physics: Conference Series*, 2020, 1607 (1).
- [8] Noble W S. What is a support vector machine?. *Nature Biotechnology*, 2006, 24 (12): 1565-1567.
- [9] Kecman V. Support vector machines – an introduction. In: *Support Vector Machines: Theory and Applications*. Berlin, Heidelberg: Springer, 2005: 1-47.
- [10] Guido R, Ferrisi S, Lofaro D, Conforti D. An overview on the advancements of support vector machine models in healthcare applications: a review. *Information*. 2024 Apr 19; 15 (4): 235.