DOI: 10.54254/2755-2721/82/20241186

Comparative Analysis of Machine Learning, Decision Trees, and K-Nearest Neighbors for Heart Disease Prediction

Xinyi Chang

School of Science and Technology, Hong Kong Metropolitan University, Hong Kong, China

s1339285@live.hkmu.edu.hk

Abstract. This study investigates the efficacy of Machine Learning (ML), Decision Trees, and K-Nearest Neighbors (KNN) techniques in predicting heart disease, aiming to identify their strengths and limitations. ML models are effective in detecting complex patterns and delivering evolving predictions from large datasets but require high-quality data and can be challenging to interpret. Decision Trees provide clear, understandable decision rules, which is beneficial in clinical settings, yet they are susceptible to overfitting and instability. KNN, valued for its simplicity and flexibility, classifies heart disease based on similarity but struggles with high computational costs and sensitivity to noisy data. Experimental results indicate that each model has distinct advantages: ML excels in pattern recognition, Decision Trees offer interpretability, and KNN handles diverse data effectively. However, each also faces challenges that impact performance, such as data quality issues for ML, overfitting for Decision Trees, and computational demands for KNN. The study suggests that balancing these strengths and weaknesses is crucial for optimizing heart disease prediction models. Future research should explore hybrid approaches that combine these models' advantages while addressing their respective limitations to improve predictive accuracy and practical application in real-world scenarios.

Keywords: Heart Disease Prediction, Machine Learning, Decision Trees, K-Nearest Neighbors.

1. Introduction

Over the past decade, heart disease has emerged as the leading cause of death worldwide, underscoring the importance of the heart as a critical organ in the human body. Lifestyle and dietary changes have a profound effect on the health of the heart, significantly influencing its function and overall health. According to the World Health Organization (WHO), cardiovascular disease is responsible for more than 18 million deaths worldwide each year [1]. This staggering statistic highlights the urgent need for improved methods of predicting, diagnosing, and managing heart disease.

Predicting cardiovascular disease is a critical challenge in clinical data analysis [2]. Doctors typically diagnose potential cardiovascular diseases based on previous clinical tests and their experience with patients who have had similar symptoms [3]. In response to the growing mortality rates associated with cardiovascular diseases, researchers have increasingly turned to technological advancements and the analysis of extensive patient data to enhance the effectiveness of heart disease prediction and diagnosis. The availability of large datasets has opened new avenues for extracting valuable insights that can aid healthcare professionals in their efforts to combat this pervasive health issue. This shift towards data-driven approaches is motivated by the increasing global prevalence of heart disease and the need for more accurate and timely diagnostic tools.

Machine learning is central to healthcare. It enables researchers to diagnose, identify and predict a range of diseases [4]. Researchers have been using data mining techniques to help diagnose heart disease, motivated by the global increase in mortality from heart disease and the availability of vast amounts of patient data from which valuable insights can be extracted. Data mining is a process of exploring large datasets for hidden and previously unknown patterns, relationships and knowledge, which traditional statistical techniques make difficult. Thus, it is the process of knowledge extraction from large amounts of data. Its applications serve to improve healthcare policies and prevent hospital errors, detect and prevent diseases early, and reduce avoidable hospital deaths [5]. For example, predicting the risk of heart disease before a heart attack occurs can greatly assist in the treatment of patients. By delving into these large datasets, data mining helps to uncover previously unknown information that can be crucial for diagnosing heart disease more effectively. This process of extracting knowledge from extensive data is essential for developing new strategies to manage and prevent cardiovascular diseases.

In the healthcare industry, data are often abundant but not fully mined for insights necessary to uncover hidden patterns and make effective decisions [6]. The complexity and volume of healthcare datasets can be overwhelming for manual analysis, making machine learning techniques invaluable for extracting useful information. In order to accurately predict the presence or absence of heart disease, these algorithms are becoming increasingly important [7]. The integration of data mining into healthcare marks a substantial advancement in combating cardiovascular diseases. Analyzing and interpreting large volumes of patient data enables more precise and proactive diagnostic and treatment approaches. As research continues to advance these techniques, the focus will continue to be on improving patient outcomes, reducing mortality rates, and improving the quality of care for those affected by heart disease. This paper summarizes recent requests of machine learning (ML) in detecting heart disease and reviews how these algorithms contribute to disease prediction and computational advancements in the medical field.

2. Methodology

2.1. Dataset description and preprocessing

This dataset provides crucial information for predicting heart disease risk by capturing various features related to patients' health profiles [8]. It includes data on demographics, medical history, lifestyle factors and key health metrics including blood pressure, cholesterol and heart rate. Analysis of these characteristics aims to identify patterns and risk factors associated with heart disease. In developed countries, where heart disease remains the leading cause of death, the model could be used to target prevention and intervention. This data comes from the University of California (UC) Irvine Machine [8].

2.2. Proposed approach

In recent years, ML has made significant strides in detecting heart disease. This paper reviews how these techniques are applied for disease prediction and computational diagnosis, focusing on their effectiveness in enhancing diagnostic accuracy. It will explore key components of ML models, including data preprocessing, feature extraction, model training, and validation. A detailed pipeline diagram will illustrate the end-to-end process from data collection to clinical application. This review aims to help understand how modern ML methods are revolutionizing cardiovascular diagnostics and highlight both their benefits and challenges in real-world scenarios, as shown in the Figure 1.

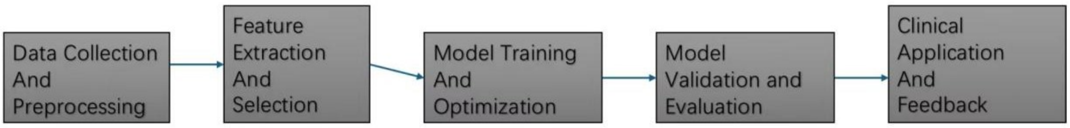
Data Collection And Preprocessing
Feature Extraction And Selection
Model Training And Optimization
Model Validation and Evaluation
Clinical Application And Feedback

Figure 1. Machine learning techniques.

2.2.1. ML. ML is a part of Artificial Intelligence (AI) that focuses more on researching algorithms that continuously learn from experience. Unlike traditional programming, it is better at identifying hidden content in data sets and effectively using that content to build new models. These newly built models are then used to analyze and predict future data that has never been encountered before. This learning process allows ML systems to improve their performance over time, as they become better at detecting complex patterns that may be difficult or even impossible for humans to discern independently [9].

In essence, ML enables computers to automatically adapt and refine their predictive capabilities by processing large volumes of data. In fields such as healthcare, finance and marketing, where complex and subtle patterns in data can provide critical insights, this adaptability is particularly valuable. By leveraging ML, organizations can enhance decision-making processes, optimize operations, and uncover valuable information that would otherwise remain hidden. As ML techniques continue to evolve, they promise to further advance our ability to analyze and interpret data in increasingly sophisticated ways.

2.2.2. Decision Tree (DT). A DT is a graphical representation used to model decisions and their potential outcomes. It is structured as a tree where each node represents a decision point or attribute. Each branch shows the possible choices. or outcomes of that decision. The root node signifies the initial decision, internal nodes represent subsequent decisions based on specific criteria, and leaf nodes denote the final outcomes or classifications [10]. In a decision tree, attributes are used to split the data into subsets, with branches illustrating the decision rules applied at each node. This hierarchical structure allows for straightforward interpretation and visualization of the decision-making process.

For example, Sabarinathan and Sugumaran used the J48 DT algorithm to select features and predict heart disease. Their study used a dataset with medical features, with some records used for training and testing. Notably, by applying feature selection, they improved accuracy to 76.67% [11]. This demonstrates the decision tree's effectiveness in handling complex classification tasks and emphasizes the importance of feature selection in improving model performance. The structure is shown in the Figure 2.

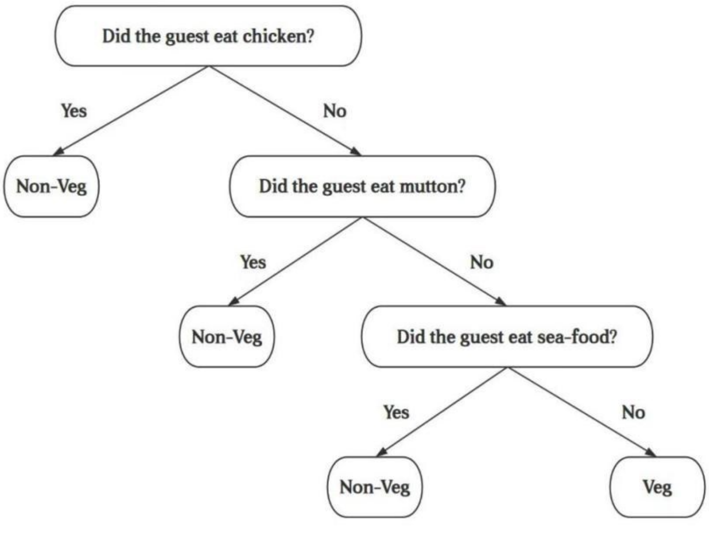
Did the guest eat chicken?
Yes
No
Non-Veg
Did the guest eat mutton?
Yes
No
Non-Veg
Did the guest eat sea-food?
Yes
No
Non-Veg
Veg

Figure 2. Decision Tree.

2.2.3. K-Nearest neighbor (KNN). KNN is a machine learning tool used for classification and regression problems. It works on the same principle that similar things are close together. Shouman et al's study used KNN to predict heart conditions using database [12].

The algorithm works by taking a point and looking at the 'K' nearest data points. The data point is then classified on the basis of the majority of the data points. For example, in this research the highest accuracy was achieved when K=7, meaning the classification was based on the 7 nearest neighbours.

The study also explored the use of a voting technique in conjunction with KNN. This involved dividing the data into subsets and applying the KNN classifier to each subset. However, the use of this technique did not improve the accuracy of the predictions. In fact, it was found that the accuracy decreased when the voting technique was used.

Overall, KNN is a simple yet powerful algorithm that can be very effective in making predictions if the appropriate value of 'K' is used.

3. Result and Discussion

In heart disease prediction, ML, Decision Trees, and KNN each offer distinct advantages and face specific challenges. ML excels in uncovering complex patterns in large datasets, leading to accurate and evolving predictions. However, it relies heavily on high-quality data and can be difficult to interpret, which may hinder trust in its diagnostic decisions. Decision Trees are valued for their clear, comprehensible structure, allowing for a straightforward understanding of how various features influence predictions. This can be particularly useful in clinical settings. Nonetheless, they are prone to model overtraining and can be unstable, as minor changes in the data may lead to different tree structures and results. KNN offers simplicity and flexibility, handling diverse data types effectively, and making it useful for classifying heart disease based on similarities to known cases. Yet, KNN can be computationally intensive with large datasets and sensitive to noisy data, potentially impacting its reliability. Balancing these strengths and weaknesses is essential for optimizing heart disease prediction models, ensuring they are both accurate and practical in real-world applications.

ML, Decision Trees, and KNN each encounter unique challenges. ML models, while powerful in detecting complex patterns, are highly dependent on large, high-quality datasets. Poor data can lead to inaccurate predictions, and the "black box" nature of many ML models makes them difficult to interpret. In contrast, Decision Trees offer greater interpretability with their clear, visual decision rules, but they are prone to overfitting, especially with complex trees that capture noise rather than general trends. This overfitting issue can also contribute to instability, as slight data changes can significantly alter tree structures and results. On the other hand, KNN, known for its simplicity and flexibility, struggles with high computational costs due to the need to calculate distances between data points for large datasets. Additionally, KNN’s performance can be adversely affected by noisy or irrelevant features and is highly sensitive to the choice of the parameter K, which can greatly influence classification accuracy. While ML provides advanced predictive capabilities, it shares the challenge of requiring substantial computational resources with KNN. Meanwhile, Decision Trees’ interpretability comes at the cost of potential overfitting and instability. Addressing these issues involves balancing data quality, computational efficiency, and model robustness to improve heart disease prediction outcomes.

4. Conclusion

This study explores the effectiveness of various predictive heart disease diagnostic models, specifically focusing on ML, Decision Trees, and KNN techniques. The primary objective is to assess and enhance diagnostic accuracy by analyzing the strengths and limitations of these models in the context of heart disease prediction. The research involved extensive experimentation to evaluate the proposed models. The findings reveal that while ML models are proficient at detecting complex patterns and improving accuracy over time, they require high-quality data and can be difficult to interpret. Decision Trees offer clear decision rules but are susceptible to overfitting and instability. KNN, though straightforward and effective, faces challenges such as high computational costs and sensitivity to noisy data. Each model presents distinct advantages and limitations, indicating that combining these approaches or improving their robustness could lead to enhanced predictive performance. Future studies will focus on developing a mixed model that integrates the strengths of ML, Decision Trees, and KNN. The goal is to improve diagnostic accuracy while addressing each model's weaknesses. This will involve creating advanced techniques to handle data variability, enhance model interpretability, and optimize computational efficiency, ultimately aiming for more accurate heart disease prediction.

References

[1] Ali F El-Sappagh S Islam S M R et al. 2020 A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion Information Fusion vol 63 pp 208-222

[2] Kiran P Swathi A Sindhu M et al. 2022 Effective heart disease prediction using hybrid machine learning technique South Asian Journal of Engineering and Technology vol 12 no 3 pp 123-130

[3] Yazdani A Varathan K D Chiam Y K et al. 2021 A novel approach for heart disease prediction using strength scores with significant predictors BMC medical informatics and decision making vol 21 no 1 p 194

[4] Bhatt C M Patel P Ghetia T et al. 2023 Effective heart disease prediction using machine learning techniques Algorithms vol 16 no 2 p 88

[5] Patel J TejalUpadhyay D Patel S 2015 Heart disease prediction using machine learning and data mining technique Heart Disease vol 7 no 1 pp 129-137

[6] Dangare C S Apte S S 2012 Improved study of heart disease prediction system using data mining classification techniques International Journal of Computer Applications vol 47 no 10 pp 44-48

[7] Ramalingam V V Dandapath A Raja M K 2018 Heart disease prediction using machine learning techniques: a survey International Journal of Engineering & Technology vol 7 no 2 pp 684-687

[8] Kamil P 2022 Indicators of Heart Disease Retrieved on 2024, Retrieved from: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[9] Aljanabi M Qutqut M H Hijjawi M 2018 Machine learning classification techniques for heart disease prediction: a review International Journal of Engineering & Technology vol 7 no 4 pp 5373-5379

[10] Mahesh B 2020 Machine learning algorithms-a review International Journal of Science and Research vol 9 no 1 pp 381-386

[11] Sabarinathan V and Sugumaran V 2014 Diagnosis of heart disease using the decision tree International Journal of Research in Computer Applications & Information Technology vol 2 no 6 pp 74-79

[12] Shouman M Turner T and Stocker R 2012 Applying k-nearest neighbor in diagnosing heart disease patients International Journal of Information and Education Technology vol 2 no 3 pp 220-223