*e-ISBN: 978-967-25608-8-3*

*Publication Date: 04-Apr-2023*

# EARLY DETECTION OF HEART DISEASE USING RANDOM FOREST ALGORITHM

*Muhammad Iqbal Bin Suhaidin[1], Syarifah Adilah Mohamed Yusoff[2], Elly Johana Johan[3], Azlina Mydin[4] and Wan Anisha Wan Mohamad[5]
**2021886804@student.uitm.edu.my[1], syarifah.adilah@uitm.edu.my[2],* ellyjohana@*uitm.edu.my[3],*
*azlin143@ uitm.edu.my[4],* wannan122@ *uitm.edu.my[5]*

[1]Kolej Pengajian Komputer, Informatik dan Media,
Universiti Teknologi MARA Cawangan Kuala Terengganu, Malaysia

[2,3,4,5]Jabatan Sains Komputer & Matematik (JSKM),
Kolej Pengajian Pengkomputeran, Informatik dan Media,
Universiti Teknologi MARA Cawangan Pulau Pinang, Malaysia

**Corresponding Author*

## ABSTRACT

*The prediction of heart diseases is a crucial aspect of healthcare, as it helps medical professionals to diagnose and treat the condition at an early stage. This is a preliminary study that aims to investigate the Random Forest Algorithm (RFA) that accurately predicts the presence of heart diseases, enabling healthcare providers to take proactive measures to prevent severe health complications and improve patient outcomes. RFA as a machine learning classification model has the potential to provide more accurate predictions than traditional methods. This potential has been investigated by thoroughly compared with several other studies across implementation of different types of dataset and algorithms. Furthermore, additional prototypes could be used in clinical settings, providing valuable insights to healthcare providers and contributing to the advancement of medical research.*

***Keywords: heart disease, random forest algorithm, classification algorithm, prediction analysis***

## Introduction

The cardiovascular system in the human body consists of the heart and blood vessels. Endocarditis, rheumatic heart disease, and conduction system abnormalities are just a few of the problems that can arise in the cardiovascular system. Cardiovascular diseases (CVDs), also known as heart diseases are group of disorders of the heart and blood vessels refers to the several conditions; 1) Coronary Artery Disease (CAD); 2) Cerebrovascular Disease (CVD); 3) Peripheral Artery Disease (PAD); and 4) Aortic Atherosclerosis (Atherosclerosis of the Aorta).

The current consumerism and technology-driven culture, which is associated with longer work hours, longer commutes, and less leisure time for recreational activities, may explain the significant and steady increase in CVD rates over the last few decades. According to WHO in 2019, an estimated 17.9 million people (representing 32% of all global) died from CVDs. Physical inactivity, a high-calorie diet, saturated fats, and sugars are specifically linked to the development of atherosclerosis and other metabolic disturbances such as metabolic syndrome, diabetes mellitus, and hypertension, all of which are common in people with CVDs (Curry et al., 2018). Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, poor diet and obesity, physical inactivity, and excessive alcohol consumption. Early detection as possible is crucial in order to arrange treatment such as counselling, self-care support and medication (Riegel et al., 2017).

According to Xie et al. (2022) heart disease risk soars after COVID even with a mild case, where people who had recovered showed stark increase in 20 CVDs problems. For example, 50% more likely to have stroke and heart attack than those in contemporary control group. The major challenge in heart disease is its detection and managing it are very extensive depending on clinical situation. Early detection of heart diseases can decrease the mortality rate and overall complications. Furthermore, early detection is critical for providing strong education on the importance of second prevention through risk factor and lifestyle modification (Lopez et al., 2022). There are tools that can predict heart disease, but they are either expensive or ineffective at predicting the possibility of heart disease in humans. However, because it takes more insights, time, and expertise, it is impossible to accurately monitor patients every day in all circumstances and consult with patients for 24 hours when there are no doctors available. In the modern world, we have a lot of data, so we can use a variety of machine-learning algorithms to examine the data and look for hidden patterns. In medical data, the hidden patterns can be used for health diagnosis.

### Random Forest Algorithm for Prediction Analysis

Random forests are a combination of machine learning algorithms. These are combined with a subset of tree classifiers, each of which cast a unit vote for the most popular class, the final sort result is then obtained by combining these results. High classification accuracy, good noise and outlier tolerance, and no overfitting are all characteristics of random forests. In both data mining and biological fields, random forests have been one of the most widely used research techniques (Liu et al., 2012).

Random Forest is a popular ensemble machine-learning algorithm that is used for both classification and regression tasks. It is based on the concept of decision trees, which are simple yet powerful models that can be used to make predictions by recursively partitioning the data into subsets based on the values of the input features. The key idea behind Random Forest is to train multiple decision trees on different subsets of the data and then average or vote on their predictions to improve the overall performance of the model.

### What is Decision Tree?

The Random Forest model grows a "forest" out of several decision trees. One more algorithm used to categorize data is a decision tree. It starts at a single point and branches off into two or more directions, with each branch of the decision tree offering a different possible outcome. To put it simply, think of it as a flowchart that clearly illustrates a pathway to a decision or outcome (Meltzer, 2021).

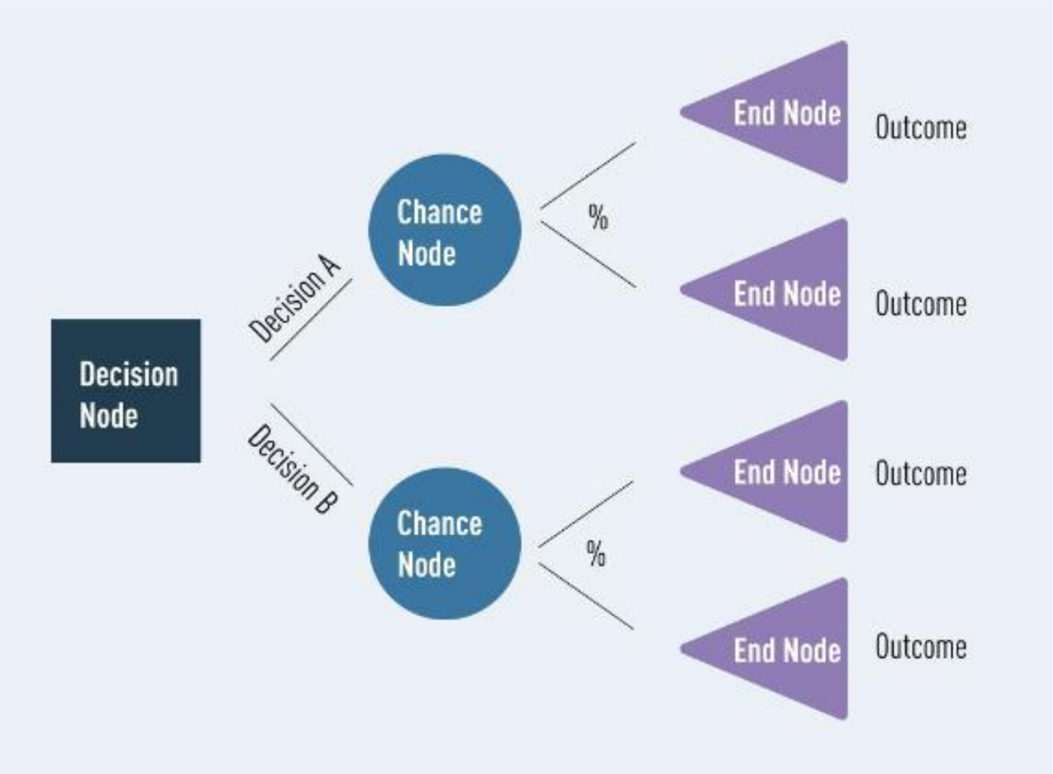

Figure1: Flow of Decision Tree *(Source: Hillier, 2021)*

For a more precise prediction, Random Forest grows multiple decision trees that are then combined. The Random Forest model is based on the idea that multiple uncorrelated models (the various decision trees) work much better together than they do separately. Each tree provides a classification or a "vote" when using Random Forest for classification. The classification with the most "votes" is chosen by the forest. When performing regression using Random Forest, the forest selects the average of all tree outputs (Meltzer, 2021).

The way Random Forest algorithm works is as follows:

1. Bootstrapping: The first step in training a Random Forest is to create multiple subsets of the data, called bootstrap samples, by randomly selecting data points with replacements. Each decision tree in the forest is trained on a different bootstrap sample, which helps to reduce overfitting by introducing randomness into the training process.

2. Feature Randomness: When training each decision tree, the Random Forest algorithm also randomly selects a subset of the features to use for splitting each node. This process, known as feature randomness, helps to decorrelate the trees and further increase performance.

3. Training: Once all the decision trees are trained, the Random Forest algorithm combines their predictions by averaging them (for regression tasks) or voting on them (for classification tasks).

4. Prediction: Finally, the Random Forest returns the combined prediction as the final output. Figure 2 is depicted the concept of how the random forest algorithm mechanism for prediction analysis.

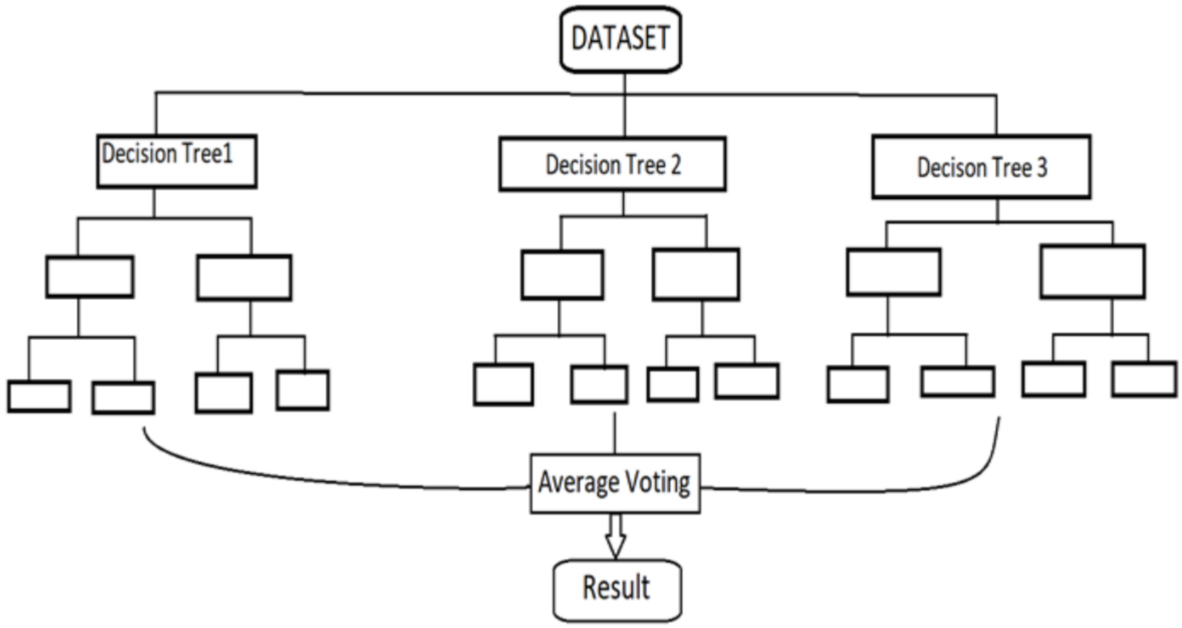

Figure 2: Mechanism of prediction using Random Forest Algorithm

The beauty of Random Forest is that, it can handle large data sets with higher dimensionality by creating multiple decision trees (forest) and making predictions by taking the average or majority of the predictions. It also helps to reduce overfitting which is a common problem with Decision Trees.

**Discussion and Conclusion**

The aim is to see the effectiveness of Random Forest in making predictions and how accurately it can perform. Thus, comparison has been made with several difference studies in order to investigate the usefulness and performance of the RFA across different implementation. Table 1, summaries several the performance of RFA across different studies.

Table1: RFA implementation for predictive analysis and performance across several studies

| No | Title | Problem | Algorithm | Result | Reference |
|---|---|---|---|---|---|
| 1 | An Ensemble Random Forest Algorithm for Insurance Big Data Analysis | Algorithms for classifying data, such as logistic regression and support vector machines, are challenging to use when modelling insurance business data (SVM) | • Random forest<br>• SVM | The ensemble random forest algorithm outperformed SVM and other classification algorithms in both performance and accuracy within the imbalanced data | (Lin et al., 2017) |
| 2 | Early Diagnosis of Breast Cancer Prediction using Random Forest Classifier | Late prediction of Breast Cancer may greatly reduce survival chances and machine learning algorithm for earlier prediction method is necessary and some algorithm was explored such as Random Forest. | Random forest | The patients are classified and a final accuracy of 98% is obtained. | (Anisha et al., 2021) |
| 3 | Random Forest Classifier for Remote Sensing Classification | Compare accuracy between Random Forest and SVM | - Random forest<br>- SVM | The result given by RF is 88.37% while SVM is 87.9%. | (Pal, 2005) |

According to the analysis in Table 1, the RFA is one of the outstanding predictive algorithms in area of medical, insurance and remote sensing data. This information suggested that the RFA is preferable to be implemented in any robust and difficult data analysis.

## References:

Anisha, P. R., Kishor Kumar Reddy, C., Apoorva, K., & Meghana Mangipudi, C. (2021). Early Diagnosis of Breast Cancer Prediction using Random Forest Classifier. IOP Conference Series: Materials Science and Engineering, 1116(1), 012187. https://doi.org/10.1088/1757-899x/1116/1/012187

Curry, S. J., Krist, A. H., Owens, D. K., Barry, M. J., Caughey, A. B., Davidson, K. W., Doubeni, C. A., Epling, J. W., Kemper, A. R., Kubik, M., Landefeld, C. S., Mangione, C. M., Silverstein, M., Simon, M. A., Tseng, C. W., & Wong, J. B. (2018). Risk assessment for cardiovascular disease with nontraditional risk factors: US preventive services task force recommendation statement. JAMA - Journal of the American Medical Association, 320(3), 272–280. https://doi.org/10.1001/jama.2018.8359

Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. (2017). An ensemble random forest algorithm for insurance big data analysis. IEEE Access, 5, 16568–16575. https://doi.org/10.1109/ACCESS.2017.2738069

Liu, Y., Wang, Y., & Zhang, J. (2012). New Machine Learning Algorithm: Random Forest. In LNCS (Vol. 7473).

Lopez, E. O., Ballard, B. D., & Jan, A. (2022). Cardiovascular disease. In StatPearls [Internet]. StatPearls Publishing. [PubMed PMID: 30571040]

MELTZER. (2021, July 15). What is Random Forest? CareerFoundry. Retrieved January 14, 2023, from https://careerfoundry.com/en/blog/data-analytics/what-is-random-forest/

Pal, M. (2005). Random forest classifier for remote sensing classification. International Journal of Remote Sensing, 26(1), 217–222.

Riegel, B., Moser, D. K., Buck, H. G., Dickson, V. V., Dunbar, S. B., Lee, C. S., Lennie, T. A., Lindenfeld, J., Mitchell, J. E., Webber, D. E., & Research, O. (2017). Self-Care for the Prevention and Management of Cardiovascular Disease and Stroke: A Scientific Statement for Healthcare Professionals from the American Heart Association. Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease, 6(9). https://doi.org/10.1161/JAHA.117.006997

Xie, Y., Xu, E., Bowe, B., & Al-Aly, Z. (2022). Long-term cardiovascular outcomes of COVID-19. Nature Medicine, 28(3), 583–590. https://doi.org/10.1038/s41591-022-01689-3