RESEARCH ARTICLE | MARCH 07 2024

Heart disease prediction using ensemble of k-nearest neighbour, random forest and logistic regression method

Mohd Syafiq Asyraf Suhaimi ✉; Nor Azuana Ramli; Noryanti Muhammad

AIP Conf. Proc. 2895, 040009 (2024)
https://doi.org/10.1063/5.0192203

Heart Disease Prediction using Ensemble of k-Nearest Neighbour, Random Forest and Logistic Regression Method

Mohd Syafiq Asyraf Suhaimi [1, a)], Nor Azuana Ramli [1,b)] and Noryanti Muhammad [1,c)]

[1] *Centre for Mathematical Sciences, College of Computing & Applied Sciences, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Pahang, Malaysia.*

[a)]Corresponding author: syaraf95@gmail.com
[b)]azuana@ump.edu.my
[c)]noryanti@ump.edu.my

Abstract. Coronary heart disease has been ranked as the number one leading cause of death in Malaysia. Based on the recent data published by WHO in 2018, death caused by this disease has reached 34,766 which brought up to 24.69 of the total deaths and places the Malaysian population 64th in the world. Medical researchers all around the world believe that there are multiple circumstances for this disease which include health problems, unhealthy personal habits, genetics, and family history. It is not an easy task to predict heart disease since the study needs a broad range of expertise from many disciplines. Recently, machine learning had been applied as one of the methods to predict heart disease. To test the accuracy of different machine learning methods, this study is conducted by applying the data extracted from the machine learning repository. The proposed predictive modelling in this study was developed using the ensemble method. The ensemble technique used was stacking where logistic regression was used as the meta-level classifier while Random Forest and k-nearest neighbour method were applied as the meta-level classifiers. Results obtained from this study show that the proposed method outperforms other single methods with 82.42 accuracies. Although the accuracy and RMSE of the ensemble method are similar to Random Forest, the proposed method is still the best method since it has a 0.903 value for the area under the ROC and 0.843 value for F1 score. This proposed predictive model will be applied by using smartwatch datasets for future study.

INTRODUCTION

As modern society evolves into an aging society, the health problems caused by chronic diseases intensify. In Malaysia, ischemic heart disease remained the leading cause of death, numbering 15.0% of the 109,164 physician-certified deaths in 2019 followed by pneumonia (12.2%), cerebrovascular disease (8.0%), transport accidents (3.8%) and malignancies and finally tumours of the trachea, bronchi, and lungs (2.4%) [1].

Even in the absence of risk factors, absolute cardiovascular disease risk levels gradually increase with age [2], but risk factors for cardiovascular disease differ by age group. The prevalence of cardiovascular disease is lower in young people than in other age groups because physiological risk factors including lipid metabolism and vascular status have no discernible effect on the condition. The characteristics of cardiovascular disease risk factors in middle-aged people were gradually affected by physiological risk factors compared to young people, however, it took some time for things to get significantly worse. As a result, middle-aged people are only considered a potential risk group for cardiovascular disease, and they do not receive the same level of attention as the true risk group, the elderly [2]. In other words, if the current risk group of middle-aged people continues to live in poor health, adolescent lifestyles are more likely to become a high-risk group for heart problems [3]. It is critical to predict and warn a person at risk of developing cardiovascular disease in advance, and a tool for predicting the onset of cardiovascular disease will be extremely useful in this regard.

With the help of many information retrieval techniques, it is much more viable to detect coronary artery disease at an early stage. The purpose of information extraction is to extract statistics from a specific data set and reshape the statistics into an understandable form for similar use. To provoke with the paints, it is best to start by collecting

3rd International Conference on Applied & Industrial Mathematics and Statistics 2022 (ICoAIMS2022)
AIP Conf. Proc. 2895, 040009-1–040009-9; https://doi.org/10.1063/5.0192203
Published by AIP Publishing. 978-0-7354-4820-9/$30.00