第 49 卷 第 2 期
2022 年 2 月

湖南大学学报（自然科学版）
Journal of Hunan University (Natural Sciences)

Vol. 49 No. 2
February 2022

Open Access Article https://doi.org/10.55463/issn.1674-2974.49.2.4

An Evaluation of Artificial Neural Networks and Random Forests for Heart Disease Prediction

Wan Aezwani Wan Abu Bakar*[1], Nur Laila Najwa B. Josdi*[1], Mustafa B. Man*[2], Yaya Sudarya Triana[3]

[1] Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Besut Campus, 22200 Besut, Terengganu, Malaysia

[2] Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, 21300 Kuala Nerus, Terengganu, Malaysia

[3] Faculty of Computer Science, Universitas Mercu Buana, Jakarta, Indonesia

Abstract: Heart diseases are serious problem in many countries worldwide. In Malaysia, it has been a major killer since 1980. Many health conditions are closely related to heart disease. However, a large amount of data that medical centers have collected each year is not well-mined to find connections between them that can aid in the prognosis of heart disease. Therefore, the purpose of this study is to propose a predictive model of heart disease based on machine learning for prognosis to help individuals with symptoms to seek early advice and treatment. By following the Knowledge Discovery in Database (KDD) methodology that includes data selection, data pre-processing, data transformation, data mining, and interpretation or evaluation of acquired knowledge, this study has tested a dataset taken from UCI Machine Learning Repository. The classification of Artificial Neural Network and Random Forest was used. They were selected based on their adequacy in the medical field, particularly in the aspect of prognosis and diagnosis. The accuracy results obtained by the relevant works from previous authors are also high and reliable. This study uses a few ways to determine the maximum accuracy achieved by both algorithms: dataset splitting and K-Fold Cross-Validation. The results of the study on the test set that has been subdivided into several subsets showed that Artificial Neural Network and Random Forest produced stable accuracies by reaching 67.9% and 64.6%, respectively. The accuracy shown by the Artificial Neural Network is more stable for both subsets, training, and testing sets. In conclusion, Artificial Neural Network has been selected as the algorithm capable of working well with the Heart Disease Prediction Model, referring to the accuracy of the test set, which is slightly better than Random Forest.

Keywords: artificial neural network, accuracy, data mining, heart disease prediction, random forest.

人工神经网络和随机森林预测心脏病的评估

摘要：心脏病是全世界许多国家的严重问题。在马来西亚，自1980年以来，它一直是主要杀手。许多健康状况与心脏病密切相关。然而，医疗中心每年收集的大量数据并没有很好地挖掘它们之间的联系，从而有助于心脏病的预后。因此，本研究的目的是提出一种基于机器学习的心脏病预测模型进行预后，以帮助有症状的个体寻求早期建议和治疗。通过遵循数据库中的知识发现方法，包括数据选择、数据预处理、数据转换、数据挖掘以及对获得的知识的解释或评估，本研究测试了取自加州大学尔湾分校机器学习存储库的数据集。使用了人工神经网络和随机森林的分类。他们的选择是基于他们在医学领域的充分性，特别是在预后和诊断方面。前人相关著作所获得的准确度结果也较高且可靠。本研究使用几种方法来确定

Received: November 9, 2021 / Revised: December 19, 2021 / Accepted: January 10, 2022 / Published: February 28, 2022

About the authors: Wan Aezwani Wan Abu Bakar, Nur Laila Najwa B. Josdi, Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Besut Campus, Besut, Malaysia; Mustafa B. Man, Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, Kuala Nerus, Malaysia; Yaya Sudarya Triana, Faculty of Computer Science, Universitas Mercu Buana, Jakarta, Indonesia

Corresponding authors Wan Aezwani Wan Abu Bakar, wanaezwani@unisza.edu.my; Nur Laila Najwa B. Josdi, laila455659@gmail.com; Mustafa B. Man, mustafaman@umt.edu.my

这两种算法实现的最大准确度：数据集拆分和K折交叉验证。对已细分为几个子集的测试集的研究结果表明，人工神经网络和随机森林的准确率分别达到了67.9%和64.6%。人工神经网络显示的准确性对于子集、训练集和测试集都更加稳定。综上所述，人工神经网络被选为能够与心脏病预测模型配合良好的算法，参考测试集的准确性，略优于随机森林。

关键词：人工神经网络，准确性，数据挖掘，心脏病预测，随机森林。

1. Introduction

Data mining has enormous potential for the healthcare industry to systematically assist data and analytics in health systems to identify the slack and the most efficient method to increase care and cut costs. Many professionals consider that an opportunity to boost the care and reduce fare simultaneously can contribute up to 30% of total health care expenses, which results in a great bargain. However, due to some of the intricacies in healthcare and delays in the adoption of technology, our industry is relatively backward compared to other countries' industries in implementing effective data analysis and data mining strategies [1].

Many industries have successfully implemented data mining where it has helped the retail sector model customer response. For instance, it helps banks forecast their customer profits; for example, it helps banks predict customer profits and any similar use cases in telecom, manufacturing, and more. As for healthcare, data mining can assist surgeons in analyzing massive datasets and acquiring apt comprehension to perform an efficient and smooth surgery.

An artificial neural network (ANN) is an algorithm inspired by brain structures created to help machines and computers resemble humans' brain and how it works [2]. Artificial neural networks have become the highlight of nowadays trends. Following the human brain's functions, they are a computing system of mutually connected nodes categorized in large raw datasets. After finding the pattern or relationship among data, they can then perform several tasks such as solving difficult problems, classifying inputs, or generating complex predictions. Artificial neural networks facilitate discovering new patterns and information related to heart disease. In this case, Multilayer Perceptron (MLP) is used on behalf of ANN due to the limited availability in WEKA. MLP is a straightforward neural network used in deep learning. Even so, the techniques introduced by Multilayer Perceptron have paved the way for more advanced neural networks [3]. It is used for various tasks, including stock analysis, diagnosis prediction, and image identification.

Random Forest is a versatile and intuitive algorithm capable of producing great classification and prediction results even by using hyper-default parameters, making it a very compelling algorithm nowadays. The process involves the operation of various Decision Trees to obtain optimal results by selecting most of them as the best value [4].

Heart disease is a medical condition where the flow of blood vessels is congested with fat or cholesterol [5]. The higher reported disease cases led to the major contributor of mortality or death cases in Malaysia. Latest statistics by the Department of Statistic Malaysia Official Portal showed that ischemic heart diseases remained as the principal cause of death, 15.0 percent of the 109,164 medically certified deaths in 2019. That was followed by pneumonia with 12.2%, cerebrovascular diseases with 8.0%, accidents by transportations with 3.8%, and lung, malignant neoplasm of trachea and bronchus with 2.4% [6]. Few risk factors have been detected as the causes of heart disease. Both factors like age and heredity are beyond our control [7]. According to the National Heart, Lung, And Blood Institute (NHLBI), it is reported that these 2 factors would increase when the age reaches 55 and 45 for women and men, respectively [8]. It could be more severe when it comes to the genetic history of close family members.

However, other factors such as unhealthy diet, obesity, and diabetes are human-controlled factors [9]. Despite genetic factors, the major contributor comes from an unhealthy lifestyle. If these factors are taken care of properly, it can reduce the risk of heart disease [10]. Heart disease is easier to treat when detected early [11]. Early diagnosis is highly recommended as it can help an individual keep track of how the individual's body is doing. As it has been the leading cause of death worldwide, heart disease should be a major public health concern. Therefore, the development of this model is crucial to many existing treatment guidelines. Predicting cardiac disease helps practitioners make accurate and correct decisions on the next actions to be taken prior to the current patient's health status. Thus, machine learning plays a crucial role in reducing and understanding the symptoms related to heart disease [12]. It also can motivate people to change their lifestyle and behavior and adhere to medications.

2. Related Works

ANN has explored a wide range of highly beneficial

applications in the medical field, especially in disease diagnosis and monitoring. The most successful ANN applications are found in challenging medical situations. Random Forest is widely used in medicine, especially in diagnosing cardiovascular disease, diabetes, and cancers [13]. The disease will be identified by analyzing the patient's medical history record. Random Forest also can be used to identify the correct combination of components in medicine as medicine has a complex mixture of certain chemicals [14].

There have been lots of related work regarding the use of machine learning in medicine. Fat liver disease diagnosis through machine learning algorithms is the effort of [15]. They use algorithms such as Naïve Bayes (NB), Random Forest (RF), Logistic Regression (LR), and Artificial Neural Networks (ANN). The result of the RF model showed accuracy predicted with 87.48% higher than other classification models. The extensive experiments conducted by [16] via several datasets of diabetics, breast cancer, heart spectrum disease, thyroid, and liver disorders in testing the effectiveness of the machine learning algorithms. Dermatology datasets that have used Random Forest have achieved a maximum accuracy of 97.26%. On average, Random Forest has provided accurate results for each dataset considered. The research in [17] proposed a computational method using protein information sequences and classifies feature vectors based on random forests to identify Alzheimer's disease genes. The protein information is extracted from k-skip-n-gram, where the accuracy attained from the model is up to 85.5 percent in the UniProt dataset. The RF model was best suited for infectious diarrhea. It also integrated the autocorrelation and seasonality of the morbidity. Filtering should be done in RF to increase the accuracy [18]. Then, a new model was enhanced in Random Forest with and named RFRF-ILM. It detects the factors of cardio disease [19]. The SVM is utilized and shows that diabetes relates to coronary artery disease with an accuracy percentage of 96.6. This model saves cost and diagnostic time while improving treatment accuracy.

The authors in [20] proposed a hybrid algorithm based on K-Means and ANN in predicting heart disease to improve prediction accuracy. The implementation result shows a higher accuracy rate of 97% of disease detection. The model by [21] is based on weights and variable attributes determined during the training process of diagnostic. A multi-level model has been proposed to predict Cardiovascular Heart Disease (CHD) using highly unbalanced clinical data, with qualitative and quantitative attributes [22]. The research confirms that the proposed Convolutional Neural Network (CNN) architecture achieves the strength of correctly classifying CHD attendance of 77%. In comparison, 81.8% is achieved in precisely classifying the absence of CHD cases on test data, making up 85.70% of the entire dataset despite the high-class imbalance in the NHANES dataset. The proposed CNN model also predicts negative (non-CBC) cases with higher accuracy and a better balance accuracy (79.5%) than other machine learning methods, such as SVM and Random Forest.

Then, work in [23] proposed a deep learning-based Heterogeneous Modified Artificial Neural Network (HMANN) method for the segmentation and early diagnosis of chronic renal disease on the Internet of Medical Things (IoMT) platform. It is classified as a Support Vector Machine and Multilayer Perceptron (MLP) with a Backpropagation (BP) algorithm. The proposed method has been shown to reduce noise and help identify the location of kidney stones clearly by segmenting the kidney picture. In kidney segmentation, the proposed HMANN method achieves an accuracy of 97.50% and reduces the time to draw contours impressively. Finally, a paper by [24] highlighted a neural network classification model to estimate the association among gender, race, BMI, age, smoking, kidney disease, and diabetes in hypertensive patients. Artificial neural network techniques applied to large clinical data sets can provide a meaningful data-driven approach to categorize patients for population health management and support in managing and detecting hypertensive patients, which are part of a critical factor for heart disease. The results from an imbalanced dataset of 24,434 with (69.71%) non-hypertensive patients and (30.29%) hypertensive patients indicate a sensitivity of 40%, a specificity of 87%, precision of 57.8%, and a measured AUC of 0.77 (95% CI [75.01–79.01]). This paper shows results that are more effective than previous studies performed by the authors using a statistical model with similar input features that presents a calculated AUC of 0.73.

Table 1 and Table 2 depict the summary of all the related works of Random Forest and ANN used in medical research.

Table 1 Summary of related work in Random Forest (RF)

Year	RF	K-Slip-N-Gram-RF	RF + IOT	RF+NB+ANN+Logistic Regression (LR)	RF Ensemble +Feature Selection
2019 [15]				√	
2019 [16]			√		
2019 [17]		√			
2020 [18]	√				
2020 [19]					√

Table 2 Summary of related work in Artificial Neural Network (ANN)

Year	Artificial Neural Network (ANN)	K-Means + ANN	Support Vector Machine + (MLP) + Back Propagation (BP)	Convolutional Neural Network (CNN)
2017 [20]		√		
2019 [21]	√			

Continuation of Table 2

2020 [22]			√
2020 [23]		√	
2020 [24]	√		

2.1. Random Forest

Random Forest (RF) [4] is a supervised learning algorithm, and it often trains the dataset using a "bagging" method. The term "random forest" indicates that this algorithm creates a forest with multiple decision trees randomly to produce a precise and reliable prediction. Interdependence between forest volume and the result can be manipulated by increasing the number of trees in the forest — the bigger the forest, the more precise the result. RF looks for the best features among a random subgroup rather than the influential features when splitting the nodes, leading to a broad variance that will positively affect the models. The formula shown in Equation 1 is used for Random Forest calculation. Mean squared error (MSE) is used to solve regression problems. This formula calculates each node from the predicted actual value while determining the best branch for the forest.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{n} (\text{fi-yi})^2 \quad (1)$$

N is the number of data points involved, while yi is the actual value for the tested data point at a certain node i. fi is the value returned by the decision tree.

2.2. Artificial Neural Network

Artificial neural network (ANN) fits for diagnosing diseases using scans as it does not require a particular algorithm to identify the disease. Without having to take into account the quantity, a set of examples that represent all the variations of disease is vital for a neural network to work. The details to recognize the disease are not needed since neural networks learn from examples, which must be selected thoroughly to guarantee a reliable and efficient system. ANN has explored a wide range of highly beneficial applications in the medical field, especially in disease diagnosis and monitoring. The most successful ANN applications are found in challenging medical situations. Equation 2 shows the formula used for Artificial Neural Network.

$$F\left(b + \sum_{i=1}^{n} x_i w_i\right) \quad (2)$$

The b refers to bias or maybe similar to weight. It satisfies unexpected or invisible factors. Therefore, it attaches to all neurons not present at the input layer. That carries a value equal to the weight and helps control the value at which the activation function will be triggered.

The x indicates the input to the neuron, while w is the corresponding weight. i is the counter from 1 to n, and n on the upper part of the sigmoid is the number of inputs from the incoming layer.

2.2.1. Multilayer Perceptron

A multilayer perceptron (MLP) [23] is a class of feedforward artificial neural networks (ANN). Multilayer perceptron (MLP) is arranged in several layers and works with a few additional perceptrons to solve complex problems. For example, a three-layer MLP requires each perceptron in the first layer (input layer) to send output to all perceptrons in the second layer (hidden layer), followed by the transmission of output from all perceptrons in the second layer to the final layer (output layer). Each perceptron will send various signals where each signal will lead to each perceptron in the next layer, carrying a different weight and output. Each layer can have many perceptron and layers for maximizing the complexity of the multilayer perceptron system.

Typically, Multilayer Perceptron is used for supervised learning problems. The data is trained on a set of input-output pairs and learns to show the interdependence between inputs and outputs. The training process includes adjusting the parameters or weights of the model to reduce errors. Backpropagation is used to adjust those weights and parameters relative to the error. The error itself can be measured in various ways, including by root mean squared error (RMSE). MLP is designed to approach continuous function and solve problems that cannot be separated linearly, especially in pattern classification, recognition, prediction, and approximation.

3. Research Methodology

This study is based on the Knowledge Discovery in Database (KDD) approach [24] and revised with the methodology depicted in Figure 1. Once problems and objectives of the study are determined, the first step is selecting the data to be used for the study. The dataset used is the Statlog (Heart) dataset taken from UCI Machine Learning Repository [25]. It consists of 13 attributes and 270 instances. This data is downloaded in a DAT file before being converted to a CSV file ensuring its compatibility in WEKA.

The next step is data pre-processing, where data removal and data integration take place. However, since Statlog (Heart) dataset did not have any missing or repetitive values, there is no data removal happening in this stage. Each data is unique and meaningful to each attribute. The dataset is then integrated into WEKA before it is transformed.

The next step, transformation, transformed this dataset whereas it is split into three subsets that consist of both training and test data. Data must be partitioned to develop an accurate and relevant model for data to be trained and collected in the future. By partitioning this data into training and test datasets, the effectiveness and accuracy of this model will be able to be maximized. The data is split into 65% training set and 35% test set, 70% training set and 30% test set and 80% training set and 20% test set.

The next phase required the study to choose the data mining to use. The latest pattern and rule mining effort is made by [26-28], focusing on Equivalence Class Transformation (Eclat), where the dataset is organized in vertical database format. Meanwhile, solving database integration issues among multiple database formats is handled in [29] via the implementation of JSON format, especially in big data storage. For this study, ANN and Random Forest have been chosen for experimentation since their involvement in medicine has proven to help hospitals work more efficiently and effectively. According to the availability of WEKA, Multilayer Perceptron is chosen on behalf of Artificial Neural Network. The subsets are tested on a few layers and depth to find the most effective accuracy. The dataset is also tested once by a few K-Fold Cross-Validation.

The last step involved in evaluating or interpreting knowledge where it evaluates and interprets the results, rules, and reliability of the objectives is identified in the first step. By the end of the experiment, the accuracy obtained by the test set will be compared to use for the model development because this set allows a data sample that provides an impartial assessment of the appropriate final model on the training dataset. This test set serves as a proxy for new data. In this study, accuracy is used to determine how predictions can be made.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

In WEKA, TP (True Positive) + TN (True Negative) is used for Correctly Classified Instances, whereas FP (False Positive) + FN (False Negative) is used for Incorrectly Classified Instances. The % of Correctly Classified Instances in WEKA gives the model's accuracy.

Fig. 1 shows the phase of KDD used in this study.

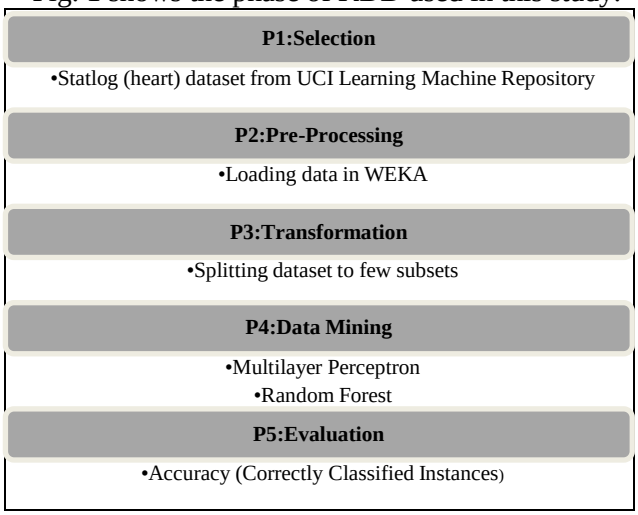

Fig. 1 Research Methodology

4. Result & Analysis

Statlog (Heart) dataset is tested using Artificial Neural Network and Random Forest, involving various layers and depths according to WEKA availability. This study used two types of testing to obtain the results: dataset percentage splitting and K-Folds Cross-Validation. For dataset percentage splitting, the original Statlog (Heart) dataset is split into two new subsets, training and testing dataset. There are three parts of subsets involved. The first subset is 65% (training set) and 35% (test set), the second subset is 70% (training set) and 30% (test set) and the last subset consist of 80% (training set) and 20% (test set). For this study, K-Fold Cross-Validation uses 2, 5, and 10 folds for the dataset. These numbers are randomly selected to see the distribution of results and how they can affect the accuracy.

Artificial Neural Network (Multilayer Perceptron) involves 2 types of layers, namely Single Layer (no hidden layer) and 3-Layer (a, 6, 8). In contrast, Random Forest involves one level of depth, a maximum depth of 0 (unlimited). These parameters are used to compare and observe how the number of layers involved can affect the dataset model's range of results and accuracy. For the Random Forest classifier, the subsets are tested on Max Depth 0, with an unlimited number of trees used to generate the result. This study chose not to do another Random Forest parameter on the dataset as the default depth, 0, offers an unlimited depth that will traverse as many trees in this dataset as possible. Table 3 shows the result obtained by testing on an Artificial Neural Network and Random Forest for a split dataset.

Table 3 Summary of Split Percentage Dataset Result

			Training Set			***Testing Set***		
			65%	70%	80%	35%	30%	20%
Acc	MLP	Single Layer	99.4 %	98.9 %	99.5 %	64.2 %	66.7 %	64.6 %
		3-Layer	98.3 %	97.8 %	97.2 %	57.8 %	*67.9 %*	64.8 %
	RF	Max Depth 0	100.0 %	100.0 %	100.0 %	58.9 %	59.2 %	*64.6 %*

The results obtained show a very high level of accuracy for all training sets in this layer. As expected, the accuracy results dropped after the test set was applied to the dataset. The accuracy varies according to the percentage of samples used, and the number of layers included. Adding more layers can increase the number of hidden layers and the number of neurons in each of those layers. That, in turn, allows the model to adapt to more complex and difficult functions. For Artificial Neural Network (Multilayer Perceptron), the subsets are tested with 500 Epochs, Learning Rate: 0.3 and under Momentum = 0:2.

The results obtained by the training set were not much different from the results obtained at Single-Layer, where it was at a very high level. Although it is not significantly different from the Single-Layer accuracy results, the test set for the 35% dataset showed slightly less accuracy than the other data. The overall fixed results can be categorized as good and can

be used for predictive models.

The accuracy obtained by the test set is relevant because its accuracy will usually not exceed the accuracy of the training set as the model optimizes training data most of the time. Reaching a high accuracy for the test set should be a sign that the set has leaked into the training set, or there may also be an error during the process of dividing parts of the dataset. The difference also means this model is suitable for training data but unfortunately performs poorly on invisible data.

For Random Forest, the accuracy results obtained reached 100% for all the training datasets involved. The accuracy of the test set, on the other hand, showed very similar results to the test set in Artificial Neural Network (Multilayer Perceptron), with a minimum difference where it ranged from 58.9% to 64.6%.

The accuracy results from each test applied to the subset using Artificial Neural Networks (Multilayer Perceptron), and Random Forest showed consistent and reliable accuracy results. It can guarantee that this dataset can produce strong and reliable predictions.

Table 4 summarized the accuracy obtained by Statlog (Heart) dataset after being tested as a whole set on a few K-Fold Cross-Validation.

Table 4 Summary of K-Fold Cross-Validation Result

			Folds		
			2	5	10
Accuracy	MLP	Single-layer	65.1%	62.2%	58.5%
		3-layer	33.3%	64.4%	*67.4%*
	RF	Max Depth 0	60.3%	*64.8%*	64.4%

Based on the results shown above, it can be seen that the results vary according to the number of folds involved. The results obtained from this K-Fold Cross-Validation test show only a slight difference from the accuracy achieved by the test on dataset splitting. For Artificial Neural Network (Multilayer Perceptron), the results range from 58.5% to 65.1% for Single-Layer. However, it decreased significantly in the 3-Layer test for 2-Fold Cross-Validation with an accuracy of only 33.3%. For Random Forest, the result shows nearly the same accuracy for each fold, starting from 60.3% to 64.8%.

The study compared all the accuracy achieved by the test set from the dataset splitting and K-Fold Cross-Validation to determine the most appropriate algorithm to use. This study proposed developing a Heart Disease Prediction Model using an Artificial Neural Network (Multilayer Perceptron) based on several aspects such as consistency and complexity. The accuracy obtained by Multilayer Perceptron was not much different from each other, may it be in dataset splitting or K-Fold Cross-Validation. Its accuracy varies according to the layers involved, where this flexibility can bring more choices on experiments and results simply by manipulating the layers to be used. However, the results obtained by Artificial Neural Network (Multilayer Perceptron) were more applicable as both training, and test sets showed a consistent accuracy. The result was not too far apart from each other.

Fig. 2 indicates the suggested model to build Heart Disease Prediction Model.

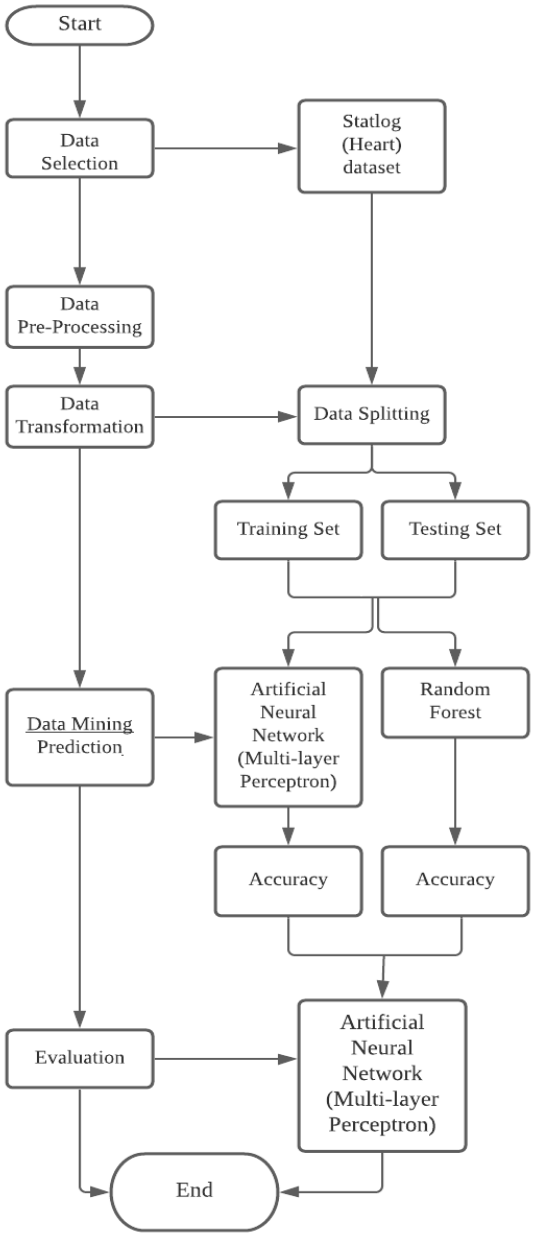

Fig. 2 Figure Heart Disease Prediction Model

5. Conclusions

This study aims to propose a predictive model for heart disease using machine learning. The study proposed a Heart Disease Prediction Model Using Artificial Neural Network (Multilayer Perceptron) based on the obtained result. Initially, the study was conducted by testing the dataset with Artificial Neural Network (Multilayer Perceptron) and Random Forest on a few subsets of training and test sets. The study results concluded that the Artificial Neural Network

(Multilayer Perceptron) showed slightly better performance than Random Forest. The accuracy achieved by both subsets is at a satisfactory level, reaching 99.5% for the training set and 67.9% for the test set. The high accuracy achieved by the training data suggested that the current model configuration successfully captures the complexity of the data set. In addition, the results obtained also did not show any signs of overfitting in the model.

While Random Forest achieved 100% for training set accuracy for all sets, the test set accuracy, unfortunately, portrayed a slightly substandard performance than Artificial Neural Network (Multilayer Perceptron).

The study also performed several tests using k-fold cross-validation, and the accuracy obtained varied, depending on the number of folds involved. The smallest folds gave the least accuracy, and the accuracy increases as the number of folds are built on. The highest accuracy gains by Artificial Neural Network (Multilayer Perceptron) is 67.4% meanwhile Random Forest reach 64.8% for its highest peak.

This study chose to use both methods, dataset splitting, and K-Folds Cross-Validations, to determine which algorithms could achieve maximum accuracy. That is to test the extent of the difference when testing a dataset by dividing it into several subsets and testing the entire dataset at once. There is a possibility that some of the split data may not fit in its subset, so by using K-Fold Cross-Validation, all instances in the dataset will be tested fairly and thoroughly. Meanwhile, the use of dataset splitting also has several reasons. One of the few is to get the well-fitted model while preventing overfitting. Overfitting could happen when the dataset is doing much better on a training set rather than a test set. If the model is still experiencing overfitting even after using the dataset splitting technique, one way that can be used is to increase the volume of data for the training subset. Upon completing the training phase of a model, it should supposedly be tested using new inputs to see if the model is ideal and to expect the model's success rate. Herewith, the study was able to observe and compare the accuracy obtained. The three types of subsets and the three types of folds used to produce accuracy variations vary according to capability, and each has its description.

In conclusion, this study aims to evaluate the performance of the two machine learning algorithms. Our research perspective is to present an ideal one to help develop an efficient and effective heart disease prediction model. However, the limitation is towards the dataset size where we have segregated the original data into 2 sets, i.e., training and testing sets. Artificial Neural Network (Multilayer Perceptron) is expected to help develop this predictive model with the accurate results obtained. That would make it easier for the health practitioners to provide an accurate prognosis and diagnosis while helping to reduce the number of patients with serious heart disease.

Several suggestions can be put forward to ensure the effectiveness of this study for future work. Future studies could explore this issue further by increasing the number of datasets and algorithms used. This study can provide more reliable and accurate correlation coefficient results and develop a system capable of assisting the early prognosis of heart disease.

5.1. Acknowledgments

We wish to thank the Center for Research Excellence and Incubation Management (CREIM), UniSZA and Research Management and Innovation Center (RMIC), UMT for providing financial support for this study. Thanks to the corresponding authors (Dr. Wan Aezwani Wan Abu Bakar as the project leader, Miss Nur Laila Najwa Bt Josdi as the MIT postgraduate student of UniSZA and Assoc. Prof. Dr. Mustafa Man as the RMIC collaborator of UMT) and Dr. Yaya Sudarya Triana as the international and network linkages. Finally, our gratitude goes to all collaborators from UniSZA and UMT for morale and technical support in checking for spelling errors and synchronization inconsistencies as the overall proofreading processes.

References

[1] BARKLEY S., STARFIELD B., SHI L., and MACINKO J. The contribution of primary care to health systems and health. In: *Family medicine: The classic papers*. CRC Press, Boca Raton, 2016, 191-239. https://doi.org/10.1201/9781315365305

[2] ABIODUN O. I., JANTAN A., OMOLARA A. E., DADA K. V., MOHAMED N. A., and ARSHAD H. State-of-the-art in artificial neural network applications: A survey. *Heliyon*, 2018, 4(11), e00938. https://doi.org/10.1016/j.heliyon.2018.e00938

[3] EDUCATIVE. *What is a multi-layered perceptron?* 2021. https://www.educative.io/edpresso/what-is-a-multi-layered-perceptron

[4] KUMAR A. Random Forest for prediction. *Towards Data Science*, 2020. https://towardsdatascience.com/random-forest-ca80e56224c1

[5] BERRÍOS-TORRES S. I., UMSCHEID C. A., BRATZLER D. W., LEAS B., STONE E. C., KELZ R. R., REINKE C. E., MORGAN S., SOLOMKIN J. S., MAZUSKI J. E., and DELLINGER E. P. Centers for disease control and prevention guideline for the prevention of surgical site infection, 2017. *Journal of the American Medical Association Surgery*, 2017, 152(8), 784-791. https://doi.org/10.1001/jamasurg.2017.0904

[6] DEPARTMENT OF STATISTICS MALAYSIA OFFICIAL PORTAL. *Statistics on Causes of Death, Malaysia*. 2020. https://www.dosm.gov.my.

[7] BENJAMIN E. J., MUNTNER P., ALONSO A., BITTENCOURT M. S., CALLAWAY C. W., CARSON A. P., CHAMBERLAIN A. M., CHANG A. R., CHENG S., DAS S. R., and DELLING F. N. Heart disease and stroke statistics — 2019 update: a report from the American Heart Association. *Circulation*, 2019, 139(10), e56-528.

https://doi.org/10.1161/CIR.0000000000000659
[8] SUBCZYNSKI W. K., PASENKIEWICZ-GIERULA M., WIDOMSKA J., MAINALI L., and RAGUZ M. High cholesterol/low cholesterol: effects in biological membranes: a review. *Cell Biochemistry and Biophysics*, 2017, 75(3), 369-385. https://doi.org/10.1007/s12013-017-0792-7
[9] FLORA G. D., & NAYAK M. K. A brief review of cardiovascular diseases, associated risk factors, and current treatment regimes. *Current Pharmaceutical Design*, 2019, 25(38), 4063-4084. https://doi.org/10.2174/1381612825666190925163827
[10] BALLA C., PAVASINI R., and FERRARI R. Treatment of angina: where are we? *Cardiology*, 2018, 140(1), 52-67. https://doi.org/10.1159/000487936
[11] BOWDEN J., & SINATRA S. T. *The Great Cholesterol Myth, Revised and Expanded: Why Lowering Your Cholesterol Won't Prevent Heart Disease - and the Statin-Free Plan that Will*. Fair Winds Press, Beverly, 2020.
[12] HEMANTH D. J. Data mining technique based critical disease prediction in medical field. In: *Intelligent Systems and Computer Technology*. IOS Press, Amsterdam, 2020.
[13] SHARMA S., & OSEI-BRYSON K. M. Toward an integrated knowledge discovery and data mining process model. *The Knowledge Engineering Review*, 2010, 25(1), 49-67. https://doi.org/10.1017/S0269888909990361
[14] ALAM M. Z., RAHMAN M. S., and RAHMAN M. S. A Random Forest based predictor for medical data classification using feature ranking. *Informatics in Medicine Unlocked*, 2019, 15, 100180. https://doi.org/10.1016/j.imu.2019.100180
[15] WU C. C., YEH W. C., HSU W. D., ISLAM M. M., NGUYEN P. A., POLY T. N., WANG Y. C., YANG H. C., and LI Y. C. Prediction of fatty liver disease using machine learning algorithms. *Computer Methods and Programs in Biomedicine*, 2019, 170, 23-29. https://doi.org/10.1016/j.cmpb.2018.12.032
[16] KAUR P., KUMAR R., and KUMAR M. A healthcare monitoring system using random forest and internet of things (IoT). *Multimedia Tools and Applications*, 2019, 78(14), 19905-19916. https://doi.org/10.1007/s11042-019-7327-8
[17] XU L., LIANG G., LIAO C., CHEN G. D., and CHANG C. C. K-skip-n-gram-RF: a random Forest based method for Alzheimer's disease protein identification. *Frontiers in Genetics*, 2019, 10, 33. https://doi.org/10.3389/fgene.2019.00033
[18] SAADOON Y. A., & ABDULAMIR R. H. Improved Random Forest Algorithm Performance for Big Data. *Journal of Physics: Conference Series*, 2021, 1897(1), 012071. https://doi.org/10.1088/1742-6596/1897/1/012071
[19] GUO C., ZHANG J., LIU Y., XIE Y., HAN Z., and YU J. Recursion enhanced random forest with an improved linear model (rerf-ilm) for heart disease detection on the internet of medical things platform. *Institute of Electrical and Electronics Engineers Access*, 2020, 8, 59247-59256. https://doi.org/10.1109/ACCESS.2020.2981159
[20] MALAV A., KADAM K., and KAMAT P. Prediction of heart disease using k-means and artificial neural network as hybrid approach to improve accuracy. *International Journal of Engineering and Technology*, 2017, 9(4), 3081-3085. http://dx.doi.org/10.21817/ijet/2017/v9i4/170904101
[21] COSTA W. L., FIGUEIREDO L. S., and ALVES E. T. Application of an Artificial Neural Network for Heart Disease Diagnosis. In: *XXVI Brazilian Congress on Biomedical Engineering*. Springer, Singapore, 2019, 753-758. http://dx.doi.org/10.1007/978-981-13-2517-5_115
[22] DUTTA A., BATABYAL T., BASU M., and ACTON S. T. An efficient convolutional neural network for coronary heart disease prediction. *Expert Systems with Applications*, 2020, 159, 113408. https://doi.org/10.1016/j.eswa.2020.113408
[23] MA F., SUN T., LIU L., and JING H. Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network. *Future Generation Computer Systems*, 2020, 111: 17-26. https://doi.org/10.1016/j.future.2020.04.036
[24] TECHOPEDIA. *Knowledge Discovery in Databases (KDD)*. 2017. https://www.techopedia.com/definition/25827/knowledge-discovery-in-databases-kdd
[25] UNIVERSITY OF CALIFORNIA. *Machine Learning Repository*. 2019. https://archive.ics.uci.edu/ml/index.php
[26] ABU BAKAR W. A. W., MAN M., MAN M, and ABDULLAH Z. I-Eclat: Performance enhancement of Eclat via incremental approach in frequent itemset mining. *Telkomnika*, 2020, 18(1), 562-570. http://dx.doi.org/10.12928/telkomnika.v18i1.13497
[27] ABU BAKAR W. A. W., JALIL M. A., MAN M., ABDULLAH Z., and MOHD F. Postdiffset: an Eclat-like algorithm for frequent itemset mining. *International Journal of Engineering & Technology*, 2018, 2(28), 197-199. http://dx.doi.org/10.14419/ijet.v7i2.28.12911
[28] JUSOH J. A., & MAN M. Modifying iEclat Algorithm for Infrequent Patterns Mining. *Advanced Science Letters*, 2018, 24(3), 1876-1880. https://doi.org/10.1166/asl.2018.11180
[29] YUSOF M. K., & MAN M. Efficiency of JSON for data retrieval in big data. *Indonesian Journal of Electrical Engineering and Computer Science*, 2017, 1, 250-262. http://dx.doi.org/10.11591/ijeecs.v7.i1.pp250-262

参考文:

[1] BARKLEY S., STARFIELD B., SHI L., 和 MACINKO J. 初级保健对卫生系统和健康的贡献。在：家庭医学：经典论文。化学橡胶公司出版社，博卡拉托, 2016, 191-239. https://doi.org/10.1201/9781315365305
[2] ABIODUN O. I., JANTAN A., OMOLARA A. E., DADA K. V., MOHAMED N. A., 和 ARSHAD H. 人工神经网络应用的最新技术：一项调查。赫利昂, 2018, 4(11), e00938. https://doi.org/10.1016/j.heliyon.2018.e00938
[3] EDUCATIVE. 什么是多层感知器？ 2021. https://www.educative.io/edpresso/what-is-a-multi-layered-perceptron
[4] KUMAR A. 用于预测的随机森林。迈向数据科学, 2020. https://towardsdatascience.com/random-forest-ca80e56224c1
[5] BERRÍOS-TORRES S. I., UMSCHEID C. A., BRATZLER D. W., LEAS B., STONE E. C., KELZ R. R., REINKE C. E., MORGAN S., SOLOMKIN J. S., MAZUSKI J. E., 和 DELLINGER E. P. 疾病控制和预防中心预防手术部位感染指南，2017 年。美国医学会外科杂志, 2017, 152(8), 784-791. https://doi.org/10.1001/jamasurg.2017.0904
[6] DEPARTMENT OF STATISTICS MALAYSIA

OFFICIAL PORTAL. 马来西亚死因统计. 2020. https://www.dosm.gov.my.
[7] BENJAMIN E. J., MUNTNER P., ALONSO A., BITTENCOURT M. S., CALLAWAY C. W., CARSON A. P., CHAMBERLAIN A. M., CHANG A. R., CHENG S., DAS S. R., 和 DELLING F. N. 心脏病和中风统计——2019 年更新：美国心脏协会的报告。循环, 2019, 139(10), e56-528. https://doi.org/10.1161/CIR.0000000000000659
[8] SUBCZYNSKI W. K., PASENKIEWICZ-GIERULA M., WIDOMSKA J., MAINALI L., 和 RAGUZ M. 高胆固醇/低胆固醇：对生物膜的影响：综述。细胞生物化学和生物物理学, 2017, 75(3), 369-385. https://doi.org/10.1007/s12013-017-0792-7
[9] FLORA G. D., 和 NAYAK M. K. 心血管疾病、相关危险因素和当前治疗方案的简要回顾。当前的药物设计, 2019, 25(38), 4063-4084. https://doi.org/10.2174/1381612825666190925163827
[10] BALLA C., PAVASINI R., 和 FERRARI R. 心绞痛的治疗：我们在哪里？心脏病学, 2018, 140(1), 52-67. https://doi.org/10.1159/000487936
[11] BOWDEN J., 和 SINATRA S. T. 伟大的胆固醇神话，修订和扩展：为什么降低胆固醇不会预防心脏病 - 以及可以预防他汀类药物的计划。顺风出版社, 贝弗利, 2020.
[12] HEMANTH D. J. 基于数据挖掘技术的医学领域危重疾病预测[J].在：智能系统和计算机技术。IOS出版社，阿姆斯特丹, 2020.
[13] SHARMA S., 和 OSEI-BRYSON K. M. 迈向集成的知识发现和数据挖掘过程模型。知识工程评论, 2010, 25(1), 49-67. https://doi.org/10.1017/S0269888909990361
[14] ALAM M. Z., RAHMAN M. S., 和 RAHMAN M. S. 一种基于随机森林的预测器，用于使用特征排序的医学数据分类。医学信息学解锁, 2019, 15, 100180. https://doi.org/10.1016/j.imu.2019.100180
[15] WU C. C., YEH W. C., HSU W. D., ISLAM M. M., NGUYEN P. A., POLY T. N., WANG Y. C., YANG H. C., 和 LI Y. C. 使用机器学习算法预测脂肪肝疾病。生物医学中的计算机方法和程序, 2019, 170, 23-29. https://doi.org/10.1016/j.cmpb.2018.12.032
[16] KAUR P., KUMAR R., 和 KUMAR M. 使用随机森林和物联网的医疗保健监控系统。多媒体工具和应用程序, 2019, 78(14), 19905-19916. https://doi.org/10.1007/s11042-019-7327-8
[17] XU L., LIANG G., LIAO C., CHEN G. D., 和 CHANG C. C. 一种基于随机森林的阿尔茨海默病蛋白质鉴定方法。遗传学前沿, 2019, 10, 33. https://doi.org/10.3389/fgene.2019.00033
[18] SAADOON Y. A., 和 ABDULAMIR R. H. 改进了大数据的随机森林算法性能。物理学杂志：系列会议, 2021, 1897(1), 012071. https://doi.org/10.1088/1742-6596/1897/1/012071
[19] GUO C., ZHANG J., LIU Y., XIE Y., HAN Z., 和 YU J. 基于改进线性模型的递归增强随机森林在医疗物联网平台上用于心脏病检测。电气和电子工程师协会访问, 2020, 8, 59247-59256. https://doi.org/10.1109/ACCESS.2020.2981159
[20] MALAV A., KADAM K., 和 KAMAT P. 使用 k 均值和人工神经网络作为提高准确性的混合方法预测心脏病。国际工程技术杂志, 2017, 9(4), 3081-3085. http://dx.doi.org/10.21817/ijet/2017/v9i4/170904101
[21] COSTA W. L., FIGUEIREDO L. S., 和 ALVES E. T. 人工神经网络在心脏病诊断中的应用。在：XXVI 巴西生物医学工程大会。新加坡施普林格, 2019, 753-758. http://dx.doi.org/10.1007/978-981-13-2517-5_115
[22] DUTTA A., BATABYAL T., BASU M., 和 ACTON S. T. 一种用于冠心病预测的高效卷积神经网络。具有应用程序的专家系统, 2020, 159, 113408. https://doi.org/10.1016/j.eswa.2020.113408
[23] MA F., SUN T., LIU L., 和 JING H. 使用基于深度学习的异构改进人工神经网络检测和诊断慢性肾病。下一代计算机系统, 2020, 111: 17-26. https://doi.org/10.1016/j.future.2020.04.036
[24] TECHOPEDIA. 数据库中的知识发现. 2017. https://www.techopedia.com/definition/25827/knowledge-discovery-in-databases-kdd
[25] UNIVERSITY OF CALIFORNIA. 机器学习库. 2019. https://archive.ics.uci.edu/ml/index.php
[26] ABU BAKAR W. A. W., MAN M., MAN M, 和 ABDULLAH Z. 工具等价类聚类和自下而上的格遍历：通过频繁项集挖掘中的增量方法来增强 Eclat 的性能。泰尔科姆尼卡, 2020, 18(1), 562-570. http://dx.doi.org/10.12928/telkomnika.v18i1.13497
[27] ABU BAKAR W. A. W., JALIL M. A., MAN M., ABDULLAH Z., 和 MOHD F. 后差异集：一种用于频繁项集挖掘的等价类聚类和自下而上的格遍历算法。国际工程技术杂志, 2018, 2(28), 197-199. http://dx.doi.org/10.14419/ijet.v7i2.28.12911
[28] JUSOH J. A., 和 MAN M. 修改仪器等价类聚类和自下而上的格遍历算法以进行不频繁模式挖掘。高级科学快报, 2018, 24(3), 1876-1880. https://doi.org/10.1166/asl.2018.11180
[29] YUSOF M. K., 和 MAN M. 爪哇对象表示法在大数据中的数据检索效率。印度尼西亚电气工程与计算机科学杂志, 2017, 1, 250-262. http://dx.doi.org/10.11591/ijeecs.v7.i1.pp250-262