Prediction of Heart Diseases Using Data Mining Algorithms

Karrar AL-Jammali
Faculty of Pharmacy, University of Kufa, Najaf, Iraq
E-mail: karrara.aljammali@uokufa.edu.iq

Keywords: ANN artificial neural networks, SVM support vector machine, decision tree, heart disease, data mining

Received: October 24, 2022

Data mining has been successfully used in numerous businesses and sectors as a result of its success in great visible areas like e-commerce and marketing. Healthcare is one of the recently identified industries. The healthcare sector continues to be "information-rich." Healthcare systems have access to a multitude of datasets and can use them to find hidden links and trends in data. There aren't enough efficient analysis tools, though. The dataset is analyzed using various machine learning algorithms, i.e., decision trees, neural networks, support vector machines, and algorithms. The experiment makes use of data mining. This study paper aims to present an overview of the most recent methods for knowledge discovery in databases utilizing. Data mining is a technique used in modern medical research, especially to predict heart disease. The primary cause of a significant portion of deaths worldwide is heart disease. Several experiments on the dataset have been done to compare the performance of predictive data mining techniques. The results show that SVM performs better of Other predictive techniques, such as ANN Neural Networks, and the decision tree performs poorly. We are recommending that you test more classifiers, so you may compare the results with other algorithms and improve the system in our earlier work by adding more features. This will help the system predict and diagnose people with heart disease more accurately.

***Povzetek**: Glavni cilj te študije je bil napovedati človeško stanje in ali ima srčno bolezen ali ne.*

1 Introduction

Data mining technology offers a user-focused method for discovering new and hidden patterns in medical data sets. Medical data mining has a lot of potential for clinical diagnosis, and these patterns can be used. However, the readily accessible raw medical data is dispersed, diverse, and substantial. This data must be gathered in a structured manner. A hospital information system can then be created by integrating the acquired data.

According to the World Health Organization, heart disease kills 12 million people every year. Cardiovascular illnesses are responsible for half of all the United States and other affluent nations' fatalities. It is also the main factor in deaths in several countries [1].

The main cause of death in the worldwide is heart disease. In the US, a person dies of heart disease every 34 seconds. There are some types of cardiac disorders, including cardiomyopathy, coronary heart disease, and cardiovascular disease. The term "cardiovascular disease" refers to a broad spectrum of disorders that have an impact on the heart, blood arteries, and how the body pumps and circulates blood. Cardiovascular disease (CVD) causes a variety of ailments, disabilities, and fatalities. Disease diagnosis is a crucial and complex task in medicine [2].

Medical diagnosis is thought of as a significant but challenging duty that must be carried out precisely and effectively. This system would greatly benefit from automation. An automatic medical diagnosis system would likely be incredibly helpful. Clinical tests can be conducted at a lower cost with the help of suitable computer-based information and/or decision support systems. A comparison study of many methodologies available is necessary for the effective and precise implementation of automated systems. In this research, different ways of using predictive and descriptive data mining to diagnose heart disease are looked at [3].

2 Technology for data mining

An artificial neural network (ANN) is a mathematical or computational model that is based on the structural and functional characteristics of biological brain networks. They derived their inspiration from the type of computation carried out by the human brain. ANN is a network of synthetic neurons that uses a connectionist method of computation to analyze input. According to the basic connection principle, mental processes can be modelled as networks of simple, typically uniform units that are interconnected. During the learning phase, ANN frequently acts as an adaptive system, changing its structure in response to external or internal data. In order to find patterns from sets of data, modern neural networks are frequently used to describe complicated interactions between inputs and outputs [4]. ANN is seen as a nonlinear statistical data modelling tool. It is made up of

numerous extremely linked small processing units (artificial neurons). Data is input into ANN using a model of the human brain. An extensive training set is required because ANN is an iterative process. Its unique capability is to extract patterns and directions from complicated data that are too challenging for humans or other computer abilities to identify [5].

In the medical field, medical devices can be monitored by artificial neural networks, which include continuous updating of many requirements, such as heart rate, blood pressure, etc. Neural networks can be trained to learn a classification task and to predict diseases [6].

In the medical field, medical devices can be monitored by artificial neural networks, which include continuous updating of many requirements, such as heart rate, blood pressure, etc. Neural networks can be trained to learn a classification task and to predict diseases [7].

3 Decision tree

Data mining software is essential to the process of discovering knowledge, It uncovers important hidden information. To create fresh target patterns, vast data collection can also be processed.

Decision trees are used in many fields, including machine learning, information extraction, applications in biomedicine, and categorization research in science. Systems that produce classifiers are one of the most widely used data mining techniques. Data classification algorithms in data mining can process a large volume of data or knowledge [8]. It can be applied to infer conclusions about category class names, to categorize information according to training materials and class descriptions, and to categorize newly available machine learning techniques for data classification containing multiple algorithms, and this work used the general decision tree algorithm[9].

The decision tree can process nominal and numerical data simultaneously, can be visually explained, visually analyzed, and easily extract rules. When the data set is tested, the decision tree's size is independent of the database size, its running speed is relatively quick, and it can be extended to large databases.

The decision tree does not require more expertise in the subject. Fast and simple to understand. Decision trees can handle a variety of data types, including binary, real, ordinal, and nominal values [10].

4 Support vector machine

Support vector machines are a supervised machine learning method It functions both as a predictor and a classifier; it locates a hyper-plane in the feature space for categorization that distinguishes between classes[11]. After that, the test data points are mapped in the same area and are categorized according to either side of a wide margin [12].

5 Heart disease data

We pick a dataset from UCI Machine Learning and download it [13]. We present 13 attributes in this database were extracted from a larger set of 75. The dataset, which includes 13 variables related to heart disease, was created using data from 270 individuals, some of whom were diagnosed with heart disease. While others were not. It is thought that 14 characteristics are a class. Data analysis aims to determine whether or not there is heart disease (1 is none and 2 is present). Three classifiers were utilized in the procedures to identify the new suspect patient's condition.

6 Results

Apply classification models to the following steps that have been taken with the Rapid Miner Framework: split validation for training and test data.

90% of the data is used for training and 10% is used for testing in the ANN classifier The model is then optimized for maximum performance, and the ANN's class detection accuracy is improved by using a confusion matrix. With the first step's default configuration to get the accuracy.

We have some steps to configuration. The first step is to add only one hidden layer and increase its neuron count. The second step is to check shuffle data. We normalize values, and we do some steps in other models, like SVM and Decision Tree, Table 1 shows Confusion Matrix for ANN.

Table 1: Confusion Matrix for ANN

	Sick	Normal	Class Precision
Prediction. Sick	30	2.5	80.56%
Prediction. Normal	34	3.0	84.44%
Class Recall	80.86%	84.44%	

Table 2 shows the Confusion Matrix for Support vector machine.

Table 2: Confusion Matrix for SVM

	Sick	Normal	Class Precision
Prediction. Sick	10	1	90.91%
Prediction. Normal	2	14	88.25%
Class Recall	83.33%	93.33%	

Table 3 shows the Confusion Matrix of the Decision Tree.

Table 3: Confusion Matrix of the Decision Tree

	Sick	Normal	Class Precision
Prediction. Sick	9	2	81.82%
Prediction. Normal	3	13	81.25%
Class Recall	75.00%	86.67%	

Figure 1 shows Confusion Matrix of ANN, SVM, and Decision Tree

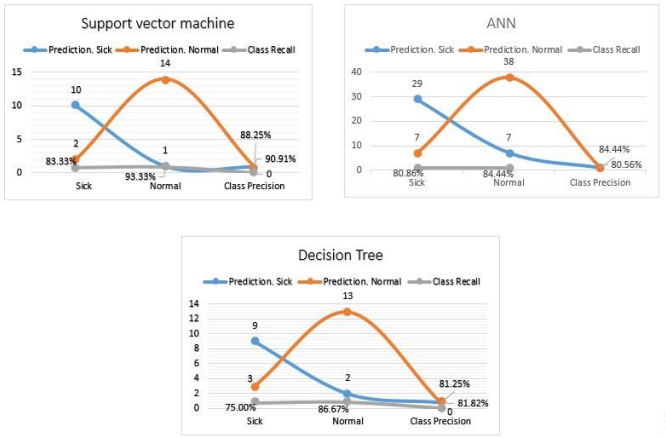

Figure 1: Confusion Matrix of ANN, SVM, and Decision Tree.

As show in Figure 2, the diagnosis model has four stages, the first stage pre-processing that the SVM will use to make a diagnosis. The second step is to set up the SVM algorithm so that it gives the best results. The third step is to use the SVM algorithm to figure out what's wrong with a new case, once you put in the details of a new case, SVM will use of training data to figure out how to handle the new case. In the last step, a medical expert checks the results to make sure they are correct. The new case data is added to the training data to make a better accurate model. By adding the new case's results to the training set, our model will get better. After some time, the amount of training data we have will grow. After many more steps, there will be two types of records in the training data. The original data collect before but wasn't check it by a doctor, and the other records have been checked one after one.
More verified data records will make the model more accurate, and the training data will also continue to grow. We can make less mistakes in training data if we add new patient data that has been checked. SVM and doctors classify this information about patients.

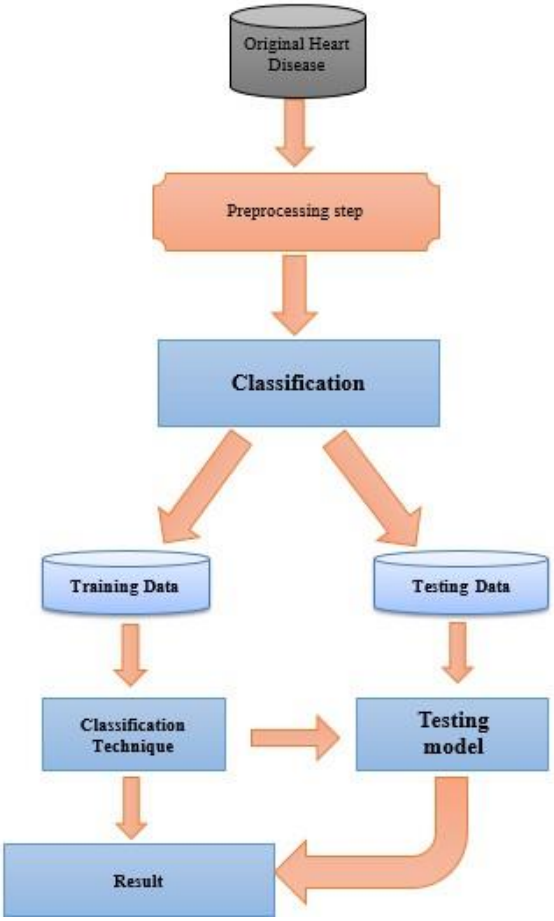

Figure 2: Diagnoses Model

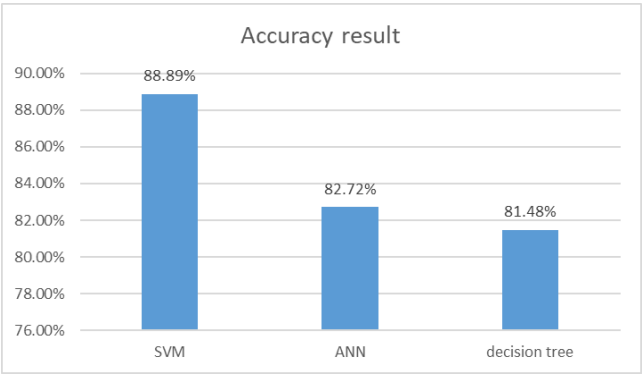

Figure 3: Accuracy of three model

7 Analyze the results

The results of the SVM classification algorithm are much better than those of ANN classification. There is a clear difference in how accurate they are. When we go beyond the training data, the following Table 3 shows that SVM is better at diagnosing heart disease than ANN learning and Decision Tree. This happens because the ANN model was trained with some examples. This shows that the ANN model should be used in deep learning with large datasets to get a better result. Regarding accuracy reflex, the results of the two model-based training methods were the same, and the result of decision tree model is less than SVM in accuracy reflex that the model-based the technique in the training.

On the other hand, the experiment in Table 3 shows that the SVM classifier gives the best results. By comparing the tables, the result of SVM is a better model than other algorithms for classifying heart disease shown in Table 4. so, we have done some tests to see how well and how practical different classification algorithms are for making predictions about Heart Patients shows in Figure 3. And in the medical field, getting things accurate is very important. Figure 2 shows the Diagnoses Model.

8 Compare the results

We describe three key categorization models, decision trees, artificial neural networks, support vector machines and using overfitting and hyper parameters to forecast and identify disease. SVM classifier gives the best result that mean SVM was more accurate than ANN Artificial Neural Networks and Decision Tree by 88.89% and we obtain ANN with the greatest accuracy of 82.72%, and the decision tree is (81.48%).

The best model that can be used for achieving the results is support vector machines SVM.

Table 4: Model accuracy.

Model	Accuracy result
SVM	88.89%
ANN	82.72%
Decision Tree	81.48%

9 Conclusion

This project included research about one of the most well-known data mining tasks. The main objective of this study was to assess if a person has heart disease or not by comparing three classification algorithms. Since more informative models produce more accurate results, we use SVM, which is more accurate than ANN or decision trees.

We describe three key categorization models, decision trees, artificial neural networks, support vector machines and using overfitting and hyper parameters to forecast and identify disease. Overfitting conditions can result from tedious configuration operations, such as setting arguments. Additionally, our experimental results showed that train sets and test sets of data determined model performance and accuracy to evaluate the model's correctness, we employ a confusion matrix. Therefore, the same factors can be utilized to diagnose a state.

270 instances of dataset are used in this study's experiments, which are carried out using RapidMiner and validated using split validation techniques. We conclude from our experiments that, when used to solve the classification issue for the heart disease data analysis task, the SVM classification model performs more accurately than ANN and decision trees, which use sequential minimal optimization. We conducted these tests to make predictions about human health and whether or not he has heart disease. According to the computer's learning theory, the system may forecast new unclassified circumstances after learning from previously classified data.

References

[1] Ramalingam, V. V., Ayantan Dandapath, and M. Karthik Raja. "*Heart disease prediction using machine learning techniques: a survey." International Journal of Engineering & Technology* 7.2.8 (2018): 684-687. [Online].Available:https://doi.org/10.14419/ijet.v7i2.8.10557

[2] Palaniappan, Sellappan, and Rafiah Awang. "Intelligent heart disease prediction system using data mining techniques." 2008 IEEE/ACS international conference on computer systems and applications. IEEE,2008.[Online].Available:https://doi.org/10.1109/aiccsa.2008.4493524

[3] Dangare, Chaitrali S., and Sulabha S. Apte. "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques." *International Journal of Computer Applications*, 47.10 (2012), 44-48.[Online]. Available: https://doi.org/10.5120/7228-0076

[4] D.J. Montana and L. Davis. Training Feedforward Neural Networks Using Genetic Algorithms. *IJCAI*, 1989.[Online].Available:https://www.ijcai.org/Proceedings/89-1/Papers/122.pdf

[5] Krizhevsky, Alex. Sutskever and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks "(2012).[Online].Available:https://doi.org/01145/3065386

[6] J. Schmidhuber, An Overview of Deep Learning in Neural Networks. *Neural networks,* 61: 80-115, 2015. [Online].Available:https://doi.org/10.1016/j.neunet.2014.09.003

[7] Katarya, Rahul, and Sunit Kumar Meena. "Machine learning techniques for heart disease prediction: a comparative study and analysis."87-97. [Online].

Available: https://doi.org/10.1007/s12553-020-00505-7

[8] Sabarinathan, V., and V. Sugumaran. "Diagnosis of heart disease using decision tree." International Journal of Research in Computer Applications & Information Technology 2.6 (2014): 74-79. [Online]. Available: https://www.researchgate.net/publication/298181341

[9] Najjar, Noor. "Analyzing Data Mining Statistical Models of Bio Medical." (2018). [Online]. Available:https://edit.elte.hu/xmlui/handle/10831/41034?key=Noor

[10] Singla, Anshu, Swarnajyoti Patra, and Lorenzo Bruzzone. "A novel classification technique based on progressive transductive SVM learning." *Pattern Recognition Letters* 42 (2014): 101-106. [Online]. Available:https://doi.org/10.1016/j.patrec.2014.02.03

[11] Zhang, Ying, et al. "Sample-specific svm learning for person re-identification." *Proceedings of the IEEE conference on computer vision and pattern recognition*.2016.[Online].Available:https://doi.org/0.1109/cvpr.2016.143

[12] Wang, Shengzheng, Dacheng Tao, and Jie Yang. "Relative attribute SVM+learning for age estimation." *IEEE transactions on cybernetics* 46.3 (2015)p.825-835.[Online].Available: https://doi.org/10.1109/tcyb.2015.2416321

[13] M. Lichtman, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], Irvine, University of California, Irvine, School of Information and Computer Sciences (2013). [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Heart+Disease