International Journal of Engineering & Technology, 7 (4) (2018) 5373-5379

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Machine learning classification techniques for heart disease prediction: a review

Maryam I. Al-Janabi [1], Mahmoud H. Qutqut [1] [2] *, Mohammad Hijjawi [1]

[1] Faculty of Information Technology, Applied Science Private University, Amman, 11931 Jordan
[2] Telecommunications Research Lab, School of Computing, Queen's University, Kingston, ON K7L 2N8 Canada
**Corresponding author E-mail: qutqut@asu.edu.jo*

Abstract

The most crucial task in the healthcare field is disease diagnosis. If a disease is diagnosed early, many lives can be saved. Machine learning classification techniques can significantly benefit the medical field by providing an accurate and quick diagnosis of diseases. Hence, save time for both doctors and patients. As heart disease is the number one killer in the world today, it becomes one of the most difficult diseases to diagnose. In this paper, we provide a survey of the machine learning classification techniques that have been proposed to help healthcare professionals in diagnosing heart disease. We start by overviewing the machine learning and de-scribing brief definitions of the most commonly used classification techniques to diagnose heart disease. Then, we review represent-able research works on using machine learning classification techniques in this field. Also, a detailed tabular comparison of the sur-veyed papers is presented.

Keywords: *Heart Disease; Heart Disease Diagnosis; Heart Disease Prediction; Machine Learning; Machine Learning Classification Techniques*.

1. Introduction

the task of making computers more intelligent. Since the most basic requirement of intelligence is learning, hence came the subfield of AI that is called machine learning (ML). ML is one of the most rapidly evolving fields of AI which is used in many areas of life, primarily in the healthcare field. ML has a great value in the healthcare field since it is an intelligent tool to analyze data, and the medical field is rich with data. In the past few years, numerous amount of data was collected and stored because of the digital revolution. Monitoring and other data collection devices are available in modern hospitals and are being used every day, and abundant amounts of data are being gathered. It is very hard or even impossible for humans to derive useful information from these massive amounts of data, that is why machine learning is widely used nowadays to analyze these data and diagnose problems in the healthcare field. A simplified explanation of what the machine learning algorithms would do is, it will learn from previously diagnosed cases of patients. The resulting classifier can have many uses such as helping doctors to diagnose new patients with higher speed and efficiency and training students and non-specialists to diagnose patients [1].

Since we have vast amounts of medical datasets, machine learning can help us discover patterns and beneficial information from them. Although it has many uses, machine learning is mostly used for disease prediction in the medical field. Many researchers became interested in using machine learning for diagnosing diseases because it helps to reduce diagnosing time and increases the accuracy and efficiency. Several diseases can be diagnosed using machine learning techniques, but the focus of this paper will be on heart disease diagnosis. Since heart disease is the primary cause of deaths in the world today, and the effective diagnosis of heart disease is immensely useful to save lives [1].

The term heart disease, also called cardiovascular disease, encompasses the diverse diseases that affect the heart. The World Health Organization estimates that 12 million deaths occur worldwide every year due to heart disease. It is the major cause of deaths in many developing countries. For example, in the United States, it kills one person every 34 seconds. It is also the main cause of deaths in India, which proves that heart disease is one of the most dangerous diseases threatening adults lives today [2]. Heart disease diagnosis is one of the most critical and challenging tasks in the healthcare field. It must be diagnosed quickly, efficiently and correctly in order to save lives. It requires the patient to do many tests, and healthcare professionals must carefully examine the results. That is why researchers have been interested in predicting heart disease, and they developed different heart disease prediction systems using various machine learning algorithms [3]. Some of them achieved better results than others. Many used the famous UCI heart disease dataset to train and test their classifier, while others used data obtained from other hospitals accessible to them.

This survey paper provides an overview of the machine learning classification techniques used in the field of diagnosing heart disease, and how previous researchers implemented them. It throws the light on how important is machine learning in the healthcare field and how it can make accurate predictions and help healthcare professionals.

The rest of the paper is organized as follows. Section 2 presents background topics on ML, classification techniques, and the most widely used heart disease dataset by researchers in this field. Section 3 contains the literature review of the current proposed research work in this area. Section 4 presents a tabular comparison between the classification techniques overviewed in section 3 on the basis of their accuracy. Finally, the conclusion is presented in section 5.

2. Background

This section provides descriptions of the related topics of this paper such as machine learning, its techniques with brief descriptions, data preprocessing, performance evaluation metrics and a brief explanation of the most used heart disease dataset.

2.1. Machine learning

Machine learning (ML) is a domain of artificial intelligence that involves constructing algorithms that can learn from experience. The way that ML algorithms work is that they detect hidden patterns in the input dataset and build models. Then, they can make accurate predictions for new datasets that are entirely new for the algorithms. This way the machine became more intelligent through learning; so it can identify patterns that are very hard or impossible for humans to detect by themselves. ML algorithms and techniques can operate with large datasets and make decisions and predictions [4]. Figure 1 represents a simplified representation of how machine learning works. In this figure, the dataset, which in our case can be a patient database, is preprocessed first. The preprocessing phase is crucial as it cleans the dataset and prepares it to be used by the machine learning algorithm. The model consists of a single algorithm, or it can contain multiple algorithms working together in a hybrid approach. The output of the model is a classifier; this is where the intelligence is, and this is what will make the prediction. If the classifier receives input data, it can predict without any human interruption. For example, if the dataset that is fed into the model is a medical dataset of healthy and unhealthy patients' information, the input data can be a new patient's information. This input data is entirely new to the classifier and has never been seen before. The classifier will receive this data and will predict whether this new patient is healthy or unhealthy based on past data.

2.2. Machine learning techniques

The main ML techniques can be classified as follows:

2.2.1. Supervised learning

In this technique, a dataset exists with examples and their response (the output). The algorithm can learn from the dataset through a training process; then it can respond to any new input based on what it has learned. An example of the supervised learning technique is classification and regression [5].

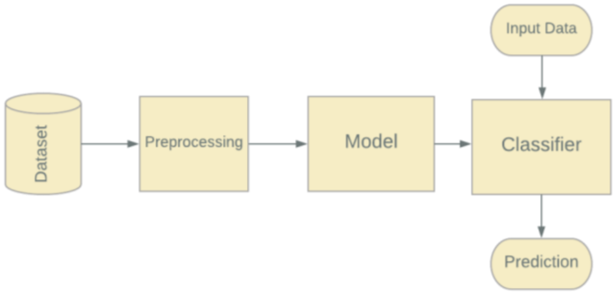

Fig. 1: Machine Learning Simplified Representation.

2.2.2. Unsupervised learning

The dataset does not contain the responses in this technique. So, the algorithm tries to recognize the similarities between input values and categorizes them based on their similarities. The unsupervised learning technique contains the clustering method [5].

2.2.3. Reinforcement learning

This technique is in the middle of supervised and unsupervised learning, where the model improves its performance as it interacts with the environment. Hence, learn how to correct its mistakes. It ought to get the correct result through examination and trying out different possibilities [5].

The most common type of learning is the supervised learning technique; especially the classification technique that is widely used for prediction. In this paper, we mainly focus on the papers that used classification algorithms to diagnose heart disease.

2.3. Classification machine learning techniques

Classification, which is a type of supervised ML techniques perform predictions for future cases based on a previous dataset. In this section, we present a brief definition of the most widely used classification techniques for heart disease prediction.

2.3.1. Naive bayes (NB)

Naive Bayes classifier belongs to a family of probabilistic classifiers based on Naive Bayes theorem. It assumes sturdy independence between the features, and this is the essential part of how this classifier makes predictions. It is easy to build, and it usually performs well which makes it suitable for the medical science field and diagnosing diseases [6].

2.3.2. Artificial neural network (ANN)

This algorithm was developed to imitate the neurons in the human brain. It consists of some nodes or neurons that are connected, and the output of one node is the input of another. Each node receives multiple inputs, but the output is only one value. The Multi-Layer Perceptron (MLP) is a widely used type of ANN, and it consists of an input layer, hidden layers, and an output layer. A different number of neurons are assigned to each layer under different conditions [6].

2.3.3. Radial basis function (RBF)

This is a type of ANN, and is similar to the Multi-Layer Perceptron (MLP) Neural Network but has a different number of hidden layers, approximation technique, number of parameters, and other factors [6].

2.3.4. Decision tree (DT)

This algorithm has a tree-like structure or flowchart-like structure. It consists of branches, leaves, nodes and a root node. The internal nodes contain the attributes while the branches represent the result of each test on each node. DT is widely used for classification purposes because it does not need much knowledge in the field or setting the parameters for it to work [6].

2.3.5. K-nearest neighbor (KNN)

This algorithm predicts the class of a new instance based on the most votes by its closest neighbors. It uses Euclidean distance to calculate the distance of an attribute from its neighbours [6].

2.3.6. Support vector machine (SVM)

This algorithm has a useful classification accuracy. It is defined as a finite-dimensional vector space which consists of a dimension for every feature/attribute of an object [6].

2.3.7. Genetic algorithm

It is an evolutionary algorithm that is built based on Darwin's theory of evolution. It imitates methods in nature such as mutation, crossover, and natural selection. One of the mostly used advantages of the genetic algorithm is its usage to initialize weights of the neural network model [8]. That is why its use alongside ANN is witnessed in many researches to produce a hybrid prediction model.

2.3.8. Ensemble learning

This method combines multiple classifiers into one model to increase the accuracy. There are three types of Ensemble learning method. The first type is Bagging, which is aggregating classifiers of the similar kind by voting technique. Boosting is the second type, which is like bagging, yet the new model is affected by previous models results. Stacking is the third type, which means aggregating machine learning classifiers for various kinds to produce one model [6].

2.4. Data preprocessing

The performance and accuracy of the predictive model is not only affected by the algorithms used, but also by the quality of the dataset and the preprocessing techniques. Preprocessing refers to the steps applied to the dataset before applying the machine learning algorithms to the dataset. The preprocessing stage is very important because it prepares the dataset and puts it in a form that the algorithm understands.

Datasets can have errors, missing data, redundancies, noise, and many other problems which cause the data to be unsuitable to be used by the machine learning algorithm directly. Another factor is the size of the dataset. Some datasets have many attributes that make it harder for the algorithm to analyze it, discover patterns, or make accurate predictions. Such problems can be solved by analyzing the dataset and using the suitable data preprocessing techniques. Data preprocessing steps includes: data cleaning, data transformation, missing values imputation, data normalization, feature selection, and other steps depending on the nature of the dataset [9].

2.5. Performance evaluation metrics

The metrics mentioned below are used by researchers to evaluate prediction models and show their performance results. We provide a short definition for each method without delving into the deep details and mathematical equations.

1) Accuracy: This metric shows the percentage of the accurate results.
2) Precision: This metric shows how relevant the result is.
3) Recall or Sensitivity: Measures the returned relevant results.
4) F-Measure: Combines precision and recall.
5) Receiver Operation Characteristic (ROC): Is a graph for visualizing the classifier's performance. It shows the correctly classified cases as well as the incorrectly classified ones [6].

The most widely used performance evaluation metric is accuracy, which is used in all research papers discussed in our article. Hence, the focus of this overview article is on categorizing, comparing and reviewing previous work based on the accuracy.

2.6. Heart disease dataset

The dataset that is used in the majority of research papers is the heart disease dataset obtained from the UCI (University of California, Irvine C.A) Center for machine learning and intelligent systems. It contains four databases from four hospitals. Each database has the same number of features, which is 14, but different numbers of records. The Cleveland dataset is the most used dataset by machine learning researchers, due to containing less missing attributes than the other datasets and having more records. The "num" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. The Cleveland dataset contains 303 instances [10]. Table 1 shows the 14 attributes/features as they exist in the dataset alongside the description of each attribute.

3. Current classification techniques for predicting heart disease

There are various classification techniques used for predicting heart disease by many researchers. In this section, we provide a summary of the surveyed papers in this area. We grouped the papers based on the algorithms that were used in their prediction model. Most researchers combined multiple algorithms in their research work or provided a comparison between them; this can be found in the last section, called the "Hybrid Approach" section.

Table 1: Dataset attributes

Number	Attribute	Description
1	Age	Age in years
2	Gender	Male or Female
3	cp	Chest pain type
4	trestbps	Resting blood pressure in mmHg
5	chol	Serum cholesterol in mg/dl
6	fbs	Fasting blood sugar
7	restecg	Resting electrocardiographic results
8	thalach	Maximum heart rate achieved
9	exang	Exercise induced angina
10	oldpeak	ST depression induced by exercise relative to rest
11	slope	The slope of the peak exercise ST segment
12	ca	Number of major vessels (0-3) colored by flourosopy
13	thal	Thallium heart scan
14	num	Diagnosis of heart disease (angiographic disease status)

3.1. Naive bayes

Vembandasamy et al. in [11] used Naive Bayes classifier to diagnose either the presence or absence of heart disease. The dataset used in the research is obtained from one of the leading diabetic research institutes in Chennai and contained records of about 500 patients and had 11 attributes (including the diagnosis). Waikato Environment for Knowledge Analysis (WEKA) tool, which is a collection of ML algorithms, is used to apply Naive Bayes classifier. The accuracy of their research work was 86.4198%.

Medhekar et al. in [12] proposed a system that categorized the data into five categories using Naive Bayes classifier. The categories are no, low, average, high and very high. The system predicts the possibility of heart disease in the input data. The dataset used for training and testing is the UCI heart disease dataset shown in table 1. The system showed an accuracy of 88.96%.

3.2. Artificial neural network (ANN)

Das et al. in [7] proposed a system using Artificial Neural Network (ANN) Ensemble method. The Cleveland heart disease dataset is shown in table 1 was used. The ensemble model provided increased generalization by combining a number of models trained on the same task. The tool used to implement the experiment was SAS enterprise miner 5.2, and the results showed that the model predicted heart disease with an accuracy of 89.01%.

Chen et al. in [13] developed a heart disease prediction system (HDPS) using Artificial Neural Network. Learning Vector Quantization (LVQ), which is a type of ANN was used in this research. The ANN model in this paper used thirteen neurons for the input layer, six neurons for the hidden layer and two neurons for the output layer. The dataset used in the paper is the Cleveland dataset in table 1. The developed system has a user-friendly interface and requires users to fill in the thirteen medical attributes to be able to make predictions. The output displays the result of the prediction, either healthy or unhealthy alongside the ROC curve, the accuracy, sensitivity, specificity and the running time it took to display the result. The tool used to develop the system is C programming language and C# for making the user interface. The results showed that the model obtained an accuracy, sensitivity, and specificity of approximately 80%, 85%, and 70% respectively.

Dangare and Apte in [14] used ANN to develop a Heart Disease

Prediction system (HDPS) to predict the presence or absence of heart disease in patients. It used the Cleveland heart disease dataset shown in table 1 for training the algorithm, and the Statlog dataset for testing; both obtained from the UCI repository and contain thirteen medical attributes. Additional two attributes which are smoking and obesity were added to increase the accuracy, which makes them fifteen attributes. The tool used for experimenting is WEKA tool. The results showed that using the thirteen attributes provided an accuracy of 99.25% whereas using the fifteen attributes provided an accuracy of nearly 100% for predicting the disease.

3.3. Decision tree (DT)

Sabarinathan and Sugumaran in [15] used the Decision Tree J48 algorithm for feature selection and for predicting heart disease. The dataset used contains thirteen medical attributes/features, and 240 records were used for training and 120 for testing. The accuracy achieved was 75.83% using all the features; while the accuracy is improved to 76.67% using feature selection. Furthermore, when more irrelevant features were removed, the accuracy is improved to 85%. The paper claims that the J48 algorithm enables selecting minimum features to enhance prediction accuracy.

Patel et al. in [16] compared several decision tree algorithms using WEKA tool on the UCI dataset to determine the presence or absence of heart disease. The different algorithms tested were J48, logistic model tree, and random forest. The J48 algorithm outperformed the rest with an accuracy of 56.76%.

3.4. K-nearest neighbour (KNN)

Shouman et al. in [17] applied K-Nearest Neighbor (KNN) to predict heart disease using the Cleveland dataset. The paper compared the results of applying KNN only and applying KNN with the voting technique. Voting is the method of dividing the data into subsets and applying the classifier to each subset. Evaluation is done using 10-fold cross-validation. The results showed that without voting, the accuracy ranged from 94% to 97.4% with various values for K. When K=7, the accuracy was the highest at 97.4%. Using the voting technique, however, did not improve the accuracy. The results showed that at K=7, the accuracy decreased to 92.7%.

3.5. Support vector machine (SVM)

Wiharto et al. in [18] studied the accuracy of SVM algorithm types on the UCI dataset to diagnose heart disease. The study included various SVM types such as Binary Tree Support Vector Machine (BTSVM), One-Against-One (OAO), One-Against-All (OAA), Decision Direct Acyclic Graph (DDAG) and Exhaustive Output Error Correction Code (ECOC). The dataset was first preprocessed using a min-max scaler. The next stage was training the algorithm on the dataset which was done using the SVM algorithms mentioned above. In the performance evaluation, BTSVM performed better than the other algorithms with 61.86% overall accuracy.

3.6. Hybrid approach

This section contains research work that built a model using different algorithms or made a comparison between several algorithms.

Khateeb and Usman in [3] experimented with various classification algorithms such as Naive Bayes, KNN, decision tree and bagging technique on the UCI Cleveland dataset. The work was divided into six cases, and the accuracy is calculated for every case by every classifier. In case 1, all the classifiers were applied to the dataset without feature reduction. In case 2, feature reduction was used where instead of using all the 14 attributes in the dataset, only seven attributes, which are the most important for heart disease diagnosis, were selected. In case 3, only the most generic features were removed such as age, sex and resting blood sugar. In case 4, the dataset was resampled by WEKA tool and only the seven most essential attributes were used. The resampling increased the accuracy of each classifier. In case 5, resampling was applied to all the 14 attributes. Finally, in case 6, the Synthetic Minority Over-sampling Technique (SMOTE) was applied in WEKA tool. The best result achieved was using KNN on case 5, which yielded 79.20% accuracy.

Pouriyeh et al. in [6] conducted a comprehensive comparison of different classification techniques on the Cleveland heart disease dataset to determine which classifier outperforms the rest. The classifiers included were Decision Tree (DT), Naive Bayes (NB), Multi-layer Perceptron (MLP), K-Nearest Neighbor (KNN), Single Conjunctive Rule Learner (SCRL), Radial Basis Function (RBF) and Support Vector Machine (SVM). The paper also included comparing ensemble techniques as bagging, boosting and stacking. The authors used the K-Fold Cross Validation technique to estimate the accuracy of classifiers. For each classifier, the performance evaluation metrics were accuracy, precision, recall, F-measure and ROC curve. For the KNN classifier, different values of K were tried, resulting in K=9 as the best value. For ANN, several neuron numbers were experimented to arrive at the best combination which is thirteen, seven and two for the input, hidden and output layers respectively. The research was divided into two experiments: the first one included comparing the different classifiers mentioned above, while the second one involved applying the ensemble techniques. The results showed that SVM outperformed the other classifiers in the first experiment at an accuracy of 84.15%. In the second experiment, using the boosting technique with SVM also proved to be the most efficient with an accuracy of 84.81%.

Amin et al. in [19] proposed a hybrid system for predicting heart disease using ANN and Genetic algorithm. The dataset used in this research was collected from 50 people through a survey conducted by the American Heart Association and contains thirteen attributes. Data analysis involved preprocessing the data to remove missing or incorrect values. The dataset was divided into 70% of the data for training and 15% for testing and validation. The system was implemented using MATLAB R2012a through Global Optimization Toolbox and the Neural Network Toolbox. The results showed an accuracy of 89% for predicting whether a person has heart disease or not.

Waghulde and Patil in [8] developed a heart disease prediction system using ANN and Genetic algorithm. The method used a genetic algorithm to initialize the weights in the Neural Network. The experiment was done using MATLAB on a dataset of 50 people collected by the American Health Association and included thirteen attributes. The results generated an accuracy of 98% and 84% when carried out using six hidden nodes and ten hidden nodes respectively.

Amma in [20] presented a system for heart disease diagnosis by combining ANN and Genetic algorithm. The dataset used was the Cleveland dataset. Preprocessing the dataset consisted of filling out missing values and normalizing the data using Min-Max normalization. The weights of the neural network were determined using the genetic algorithm. The accuracy obtained was 94.17%.

Venkatalakshmi and Shivsankar in [21] included a comparison between Naive Bayes and Decision Tree to determine which one has the highest accuracy for heart disease prediction. The dataset used was the UCI heart disease dataset. The experiment was carried out using WEKA tool and showed an accuracy of 85.03% and 84.01% for Naive Bayes and Decision Tree respectively. The paper suggested using a genetic algorithm in MATLAB to reduce the number of features before feeding the dataset into the WEKA tool for future work.

Palaniappan and Awang in [22] proposed an Intelligent Heart Disease Prediction System (IHDPS) using multiple classification techniques which are Decision Tree, Naive Bayes and Neural Network. The system is web-based and was implemented using .NET framework. The data source consisted of 909 records with fifteen attributes obtained from the Cleveland Heart Disease database. Data Mining Extension (DMX) query language was used to create the model. The results showed that Naive Bayes proved to be the most efficient with 86.53% correct predictions followed by Neural Network with only 1% difference.

Dangare and Apte in [23] developed a model for predicting heart disease. The dataset used is the Cleveland database consisting of 303 records alongside the Statlog database comprising of 270 records. Instead of using only the thirteen attributes present in the dataset, they added two attributes: obesity and smoking. WEKA tool used for preprocessing the dataset. The classification techniques used for analyzing the dataset were Decision Tree, Naive Bayes, and ANN. According to the results, ANN gave an accuracy of 100%, Decision Tree 99.62%, and Naive Bayes 90.74% which proves that Artificial Neural Network is the highest performing algorithm.

Zriqat et al. in [24] developed an effective intelligent medical decision support system. Five classification algorithms were compared which are: Naive Bayes, Decision Tree, Discriminant, Random Forest, and Support Vector Machine. The analysis was done using MATLAB on two datasets, the Cleveland Heart Disease and the Statlog Heart Disease. The results showed that Decision Tree performed the highest accuracy for both datasets at 99.01% and 98.15% for the Cleveland and Statelog datasets respectively.

Liu et al. in [25] proposed a hybrid model for diagnosing heart disease. The dataset used was the Statlog heart disease dataset from the UCI repository. The model developed with MATLAB consisted of two subsystems which are: feature selection and classification. The feature selection subsystem uses the Relief method to estimate the weight of features then used the feature selection approach Rough Set method (RFRS) to remove unnecessary features and improve the accuracy of the model. The classification subsystem used Ensemble classifier with the C4.5 algorithm (which is used to generate a Decision Tree) as the base. The results showed 92.59% classification accuracy.

Ghumbre et al. in [26] compared Support Vector Machine and Radial Basis Function (RBF), which is a type of ANN. The algorithms were applied to a patient dataset in India consisting of 214 records and 19 attributes and predicting whether a person has heart disease or not. The performance of the algorithms was evaluated using the overall average through training and testing the dataset, 5-fold cross-validation, and 10-fold cross-validation. The overall average performance yielded 86.42% and 80.81% accuracy for SVM and RBF respectively. Their results showed that SVM provided a better accuracy.

Masethe and Masethe in [27] applied several algorithms namely: J48, Naive Bayes, REPTREE, Simple Cart (Classification and Regression Tree) which is a type of Decision Tree, and Bayes Net to diagnose heart disease. The dataset used for this work has been obtained from South African physicians containing eleven attributes which are: patient identification number (replaced with dummy values to protect the privacy of patients), gender, cardiogram, age, chest pain, blood pressure level, heart rate, cholesterol, smoking, alcohol consumption and blood sugar level. The tool used in the experiment was the WEKA tool. The performance evaluation was done using 10-fold cross-validation to assess the efficiency of the built model. The results showed an accuracy of 99.0471% for J48, 99.0471% for REPTREE, 97.222% Naive Bayes, 98.1481% for Bayes Net, and 99.0741% for the simple cart, showing that simple cart outperformed the rest.

4. Comparison of ML classification techniques for heart disease prediction

This section provides a tabular comparison between all the research papers described above.

The comparison is made on the basis of accuracy and can be seen in table 2. The table has six elements which are as follow:

1) Author: This shows the author/s of the paper and the reference number.
2) Classification Technique/s: This represents the classification algorithm used in the research; whether it was a single algorithm, a comparison or a hybrid model.
3) Best Technique Found: This column is only applicable to papers having a comparison between multiple algorithms. It represents the best algorithm found in the research work, which is the algorithm with the highest accuracy.
4) Tool: The framework or programming language used to build the model is shown in this column. It is what the researcher used to pre-process the input dataset, create the predictive model and test it.
5) Dataset: This shows the dataset that was used as an input for the classification algorithm.
6) Accuracy: This represents the accuracy of the results of the proposed model. If the paper contained a comparison, this column only shows the accuracy of the best technique found by the author.

Table 2: Comparison of Classification Techniques for Heart Disease Prediction

Author	Classification Technique/s	Best Technique Found	Tool	Dataset	Accuracy
Vembandasamy et al. [11]	NB	*n/a	WEKA	A diabetic research institute in Chennai	86.4198 %
Medhekar et al. [12]		n/a	Not mentioned	Cleveland (UCI)	88.96%
Das et al. [7]	ANN Ensemble	n/a	SAS enterprise miner 5.2	Cleveland (UCI)	89.01%
Chen et al. [13]	ANN LVQ	n/a	C and C#	Cleveland (UCI)	80%
Dangre and Apte [14]	ANN	n/a	WEKA	Cleveland and Statlog (UCI)	Nearly 100%
Sabarinathan and Sugumaran [15]	DT	J48 with feature selection	Not mentioned	A dataset with 240 records for testing and 120 for training	85%
Patel et al. [16]		J48	WEKA	Cleveland (UCI)	56.76%
Shouman et al. [17]	KNN	n/a	Not mentioned	Cleveland (UCI)	97.4%
Wiharto et al. [18]	SVM	BT SVM	Not mentioned	Cleveland (UCI)	61.86%
Khateeb and Usman [3]	NB, KNN, DT and bagging technique	KNN	WEKA	Cleveland (UCI)	79.20%
Pouriyeh et al. [6]	NB, DT, MLP, KNN, SCRL, RBF, SVM, bagging, boosting and stacking	Boosting with SVM	Not mentioned	Cleveland (UCI)	84.81%
Amin et al. [19]		n/a	MATLAB	American	89%

Waghulde and Patil [8]	ANN and Genetic Algorithm hybrid system	n/a	MATLAB	Heart Association dataset American Heart Association dataset	98%
Amma [20]		n/a	Not mentioned	Cleveland (UCI)	94.17%
Venkatalakshmi and Shivsankar [21]	NB and DT	NB	WEKA	UCI	85.03%
Palaniappan and Awang [22]	DT, NB and ANN	NB	DMX	Cleveland (UCI)	86.53%
Dangare and Apte [23]		ANN	WEKA	Cleveland and Statlog (UCI)	Nearly 100%
Zriqat et al. [24]	NB, DT, Discriminant, Random Forest, and SVM	DT	MATLAB	Cleveland and Statlog (UCI)	99.01% for Cleveland and 98.15% for Statlog
Liu et al. [25]	ReliefF and Rough Set (RFRS) for feature reduction, Ensemble using C4.5 for classification	n/a	MATLAB	Statlog (UCI)	92.59%
Ghumbre et al. [26]	SVM and Radial Basis Function	SVM	Not mentioned	Indian patients dataset of 214 records and 19 attributes	86.42%
Masethe and Masethe [27]	J48, NB, REPTREE, Simple Cart, and Bayes Net	Simple Cart	WEKA	South African dataset containing 11 attributes	99.0741 %

*n/a: not applicable.

5. Conclusion and final remarks

This paper overviews the literature of machine learning classification methods for diagnosing heart disease. Many representational papers on using machine learning classification techniques were surveyed and categorized. The accuracy of the proposed models vary depending on the tool used, the dataset used, the number of attributes and records in the dataset, the preprocessing techniques, as well as the classifier implemented in the model. It depends on whether it is a hybrid model or not and whether the model uses feature selection or not. From table 2, we can conclude that the researchers who produced the highest accuracy were Dangare and Apte using Artificial Neural Network (ANN), WEKA tool and a combination of the Cleveland and Statlog heart disease datasets.

We conclude that to build an accurate heart disease prediction model, a dataset with sufficient samples and correct data must be used. The dataset must be preprocessed accordingly because it is the most critical part to prepare the dataset to be used by the machine learning algorithm and get good results. Also, a suitable algorithm must be used when developing a prediction model. We can notice that Artificial Neural Network (ANN) performed well in most models for predicting heart disease as well as Decision Tree (DT).

Finally, the field of using machine learning for diagnosing heart disease is an important field, and it can help both healthcare professionals and patients. It is still a growing field, and despite the massive availability of patient data in hospitals or clinics, not much of it is published. As observed in table 2, most researchers got their datasets from the same source which is the UCI repository. Since the quality of the dataset is an essential factor in the prediction's accuracy, more hospitals should be encouraged to publish high-quality datasets (while protecting the privacy of patients) so that researchers can have a good source to help them develop their models and obtain good results.

Acknowledgement

This work was made possible by the financial support from the Applied Science Private University in Amman, Jordan.

References

[1] I. Kononenko, “Machine learning for medical diagnosis: History, state of the art and perspective,” *Artificial Intelligence in Medicine*, vol. 23, no. 1, pp. 89–109, 2001. https://doi.org/10.1016/S0933-3657(01)00077-X.

[2] J. Soni *et al.*, “Intelligent and effective heart disease prediction system using weighted associative classifiers,” *International Journal on Computer Science and Engineering*, vol. 3, no. 6, pp. 2385–2392, 2011.

[3] N. Khateeb and M. Usman, “Efficient heart disease prediction system using k-nearest neighbor classification technique,” *in Proceedings of the International Conference on Big Data and Internet of Thing (BDIOT)*, New York, NY, USA: ACM, 2017, pp. 21–26. https://doi.org/10.1145/3175684.3175703.

[4] H. Almarabeh and E. Amer, “A study of data mining techniques accuracy for healthcare,” *International Journal of Computer Applications*, vol. 168, no. 3, pp. 12–17, Jun 2017.

[5] M. Fatima and M. Pasha, “Survey of machine learning algorithms for disease diagnostic,” *Journal of Intelligent Learning Systems and Applications*, vol. 9, no. 01, pp. 1–16, 2017. https://doi.org/10.4236/jilsa.2017.91001.

[6] S. Pouriyeh *et al.*, “A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease,” *in Proceedings of IEEE Symposium on Computers and Communications (ISCC).* Heraklion, Greece: IEEE, July 2017, pp. 204–207. https://doi.org/10.1109/ISCC.2017.8024530.

[7] R. Das, I. Turkoglu, and A. Sengur, “Effective diagnosis of heart disease through neural networks ensembles,” *Expert systems with applications*, vol. 36, no. 4, pp. 7675–7680, 2009. https://doi.org/10.1016/j.eswa.2008.09.013.

[8] N. Waghulde and N. Patil, “Genetic neural approach for heart disease prediction,” *International Journal of Advanced Computer Research*, vol. 4, no. 3, pp. 778, 2014.

[9] S. Garcia *et al.*, “Big data preprocessing: methods and prospects,” *Big Data Analytics*, vol. 1, no. 1, p. 9, Nov 2016. https://doi.org/10.1186/s41044-016-0014-0.

[10] A. Janosi *et al.*, “Heart disease data set,” Jul 1988. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/heart Disease.

[11] K. Vembandasamy, R. Sasipriya, and E. Deepa, “Heart diseases detection using naive bayes algorithm,” *International Journal of Innovative Science, Engineering & Technology*, vol. 2, no. 9, pp. 441–444, 2015.

[12] D. Medhekar, M. Bote, and S. Deshmukh, “Heart disease prediction system using naive bayes,” *International Journal of Enhanced Research In Science Technology & Engineering*, vol. 2, no. 3, pp. 1–5, 2013.

[13] A. Chen *et al.*, “HDPS: Heart disease prediction system,” in *Computting in Cardiology*, Hangzhou, China: IEEE, 2011, pp. 557–560.

[14] C. Dangare and S. Apte, “A data mining approach for prediction of heart disease using neural networks,” *International Journal of Computer Engineering & Technology*, vol. 3, no. 3, pp. 30–40, 2012.

[15] V. Sabarinathan and V. Sugumaran, “Diagnosis of heart disease using decision tree,” *International Journal of Research in Computer Applications & Information Technology*, vol. 2, no. 6, pp. 74–79, 2014.

[16] J. Patel *et al.*, “Heart disease prediction using machine learning and data mining technique,” *Heart Disease*, vol. 7, no. 1, pp. 129–137, 2015.

[17] M. Shouman, T. Turner, and R. Stocker, “Applying k-nearest neighbour in diagnosing heart disease patients,” *International Journal of Information and Education Technology*, vol. 2, no. 3, pp. 220, 2012. https://doi.org/10.7763/IJIET.2012.V2.114.

[18] W. Wiharto, H. Kusnanto, and H. Herianto, “Performance analysis of multiclass support vector machine classification for diagnosis of coronary heart diseases,” *International Journal on Computational Science & Applications*, vol. 5, no. 5, pp. 27–37, 2015. https://doi.org/10.5121/ijcsa.2015.5503.

[19] S. Amin, K. Agarwal, and R. Beg, “Genetic neural network based data mining in prediction of heart disease using risk factors,” in *IEEE Conference on Information Communication Technologies*. Thuckalay, Tamil Nadu, India, April 2013, pp. 1227–1231. https://doi.org/10.1109/CICT.2013.6558288.

[20] N. Amma, “Cardiovascular disease prediction system using genetic algorithm and neural network,” in *International Conference on Computing, Communication and Applications*. Dindigul, Tamilnadu, India: IEEE, Feb 2012, pp. 1–5. https://doi.org/10.1109/ICCCA.2012.6179185.

[21] B. Venkatalakshmi and M. Shivsankar, “Heart disease diagnosis using predictive data mining,” *International Journal of Innovative Research in Science*, Engineering and Technology, vol. 3, no. 3, pp. 1873–1877, 2014.

[22] S. Palaniappan and R. Awang, “Intelligent heart disease prediction system using data mining techniques,” in *IEEE/ACS International Conference on Computer Systems and Applications*. Doha, Qatar, March 2008, pp. 108–115. https://doi.org/10.1109/AICCSA.2008.4493524.

[23] C. Dangare and S. Apte, “Improved study of heart disease prediction system using data mining classification techniques,” *International Journal of Computer Applications*, vol. 47, no. 10, pp. 44–48, 2012.

[24] I. Zriqat, A. Altamimi, and M. Azzeh, “A comparative study for predicting heart diseases using data mining classification methods,” *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 14, no. 12, pp. 868–879, 2017.

[25] X. Liu *et al.*, “A hybrid classification system for heart disease diagnosis,” *Computational and Mathematical Methods in Medicine*, vol. 2017, pp. 1-11, 2017. https://doi.org/10.1155/2017/8272091.

[26] S. Ghumbre, C. Patil, and A. Ghatol, “Heart disease diagnosis using support vector machine,” in *International conference on computer science and information technology*. Pattaya, Thailand: Planetary Scientific Research Centre, 2011, pp. 84–88.

[27] H. Masethe and M. Masethe, “Prediction of heart disease using classification algorithms,” in *Proceedings of the world congress on Engineering and Computer Science*, San Francisco, USA: International Association of Engineers (IAENG), 2014, pp. 22–24.