INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

Heart Disease Prediction using Machine Learning Classification Techniques

M.Sampath Kumar, B.Lokeswari, D.Prasannanjali, A.Tanuja, G.Sampujitha

Assistant Professor, Department of CSE, Tirumala Engineering College, NRT, Andhra Pradesh, India

UG Student, Department of CSE, Tirumala Engineering College, NRT, Andhra Pradesh, India

UG Student, Department of CSE, Tirumala Engineering College, NRT, Andhra Pradesh, India

UG Student, Department of CSE, Tirumala Engineering College, NRT, Andhra Pradesh, India

UG Student, Department of CSE, Tirumala Engineering College, NRT, Andhra Pradesh, India

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 describing brief definitions of the most commonly used classification techniques to diagnose heart disease. Then, we review representative research works on using ma- chine learning classification techniques in this field. Also, a detailed tabular comparison of the surveyed papers is presented.

KEYWORDS: heart disease; heart disease diagnosis; heart disease prediction; machine learning; machine learning classification techniques.

I. INTRODUCTION

Artificial intelligence (AI) is a part of computer science that has the task of making computers more intelligent. Since the most basic requirement of intelligence is learning, hence came the sub- field 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 avail- able in modern hospitals and are being used every day, and abun- dant 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 di- agnosed 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 be- came 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 ma- chine 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, encom- passes 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 dis- ease 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 predic- tion 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 dis- ease, 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. Sec- tion 3 contains the literature review of the current proposed re- search 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.

II. BACKGROUND

This section provides descriptions of the related topics of this paper such as machine learning, its techniques with brief descrip- tions, 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 pat- terns 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 con- sists 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 da- taset that is fed into the model is a medical dataset of healthy and unhealthy patients' information, the input data can be a new pa- tient'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 un- healthy 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 tech- nique is classification and regression [5].
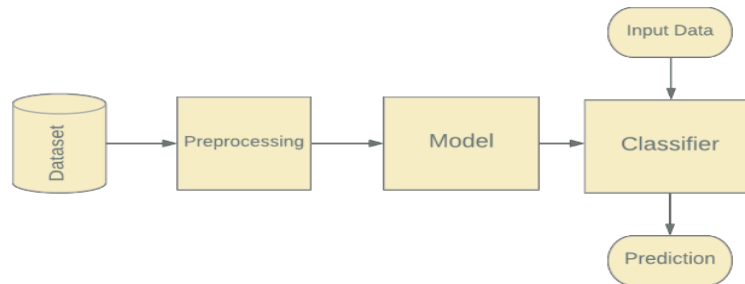
Figure 1: Machine Learning Simplified Representation

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 unsu- pervised 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 per- form 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 classifi- ers based on Naive Bayes theorem. It assumes sturdy independ- ence between the features, and this is the essential part of how this classifier makes predictions. It is easy to build, and it usually per- forms 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 con- ditions [6].

2.3.3. Radial Basis Function (RBF)

This is a type of ANN, and is similar to the Multi-Layer Percep- tron (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 the- ory of evolution. It imitates methods in nature such as mutation, crossover, and natural selection. One of the mostly used ad- vantages 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 predic- tion model.

2.3.8. Ensemble Learning

This method combines multiple classifiers into one model to in- crease 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 pre- vious models results. Stacking is the third type, which means ag- gregating 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 da- taset 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 ana- lyzing 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.

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 data- base 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.

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)

> **Accuracy**: This metric shows the percentage of the accurate results.
> **Precision:** This metric shows how relevant the result is.
> **Recall or Sensitivity:** Measures the returned relevant results.
> **F-Measure:** Combines precision and recall.
> **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].

3.1. Naive Bayes

Vembandasamy *et al.* in [11] used Naive Bayes classifier to diag- nose 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 catego- ries 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 Net- work (ANN) Ensemble method. The Cleveland heart disease da- taset 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 Quanti- zation (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 da- taset 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 accura- cy, which makes them fifteen attributes. The tool used for experi- menting is WEKA tool. The results showed that using the thirteen attributes provided an accuracy of 99.25% whereas using the fif- teen 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 accu- racy achieved was 75.83% using all the features; while the accura- cy is improved to 76.67% using feature selection. Furthermore, when more irrelevant features were removed, the accuracy is im- proved 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 logistic model tree, and random forest. The J48 algorithm outper- formed the rest with an accuracy of 56.76%.

3.4. K-Nearest Neighbor (KNN)

Shouman *et al.* in [17] applied K-Nearest Neighbor (KNN) to predict heart disease using the Cleveland dataset. The paper com- pared 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 var- ious 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 in- cluded 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 pre- processed using a min-max scaler. The next stage was training the algorithm on the dataset which was done using the SVM algo- rithms 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 dif- ferent algorithms or made a comparison between several algo- rithms. 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 di- vided 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 dis- ease 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 in- creased 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), Sin- gle Conjunctive Rule Learner (SCRL), Radial Basis Function (RBF) and Support Vector Machine (SVM). The paper also in- cluded 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 per- formance 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, sev- eral neuron numbers were experimented to arrive at the best com- bination which is thirteen, seven and two for the input, hidden and

output layers respectively. The research was divided into two ex- periments: the first one included comparing the different classifi- ers 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 con- ducted by the American Heart Association

and contains thirteen attributes. Data analysis involved preprocessing the data to re- move 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 peo- ple 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 nor- malization. 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 car- ried 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 us- ing .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 tech- niques 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 com- pared which are: Naive Bayes, Decision Tree, Discriminant, Ran- dom 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 clas- sification. 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 fea- tures 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 re- sults 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 algo- rithms 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 dum- my values to protect the privacy of patients), gender, cardiogram, age, chest pain, blood pressure level, heart rate, cholesterol, smok- ing, 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.

IV. COMPARISON OF ML CLASSIFICATION TECH- NIQUES FOR HEART DISEASE PREDICTION

This section provides a tabular comparison between all the re- search 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 algo- rithms. 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 in- put 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 Technique/s	Classification Found	Best Technique	Tool	Dataset	Accuracy
dasamy et al. [11]	NB	*n/a	WEKA	A diabetic research institute in Chennai	86.4198%
Medhekar et al. [12] 88.96%			n/a	Not mentioned	Cleveland (UCI)
Das et al. [7]	ANN Ensemble miner 5.2	n/a	SAS enterprise	Cleveland (UCI)	89.01%
Chen et al. [13]	ANN LVQ	n/a	C and C#	Cleveland (UCI)	80%
Cleveland and	Nearly 100% Statlog (UCI)				
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] 97.4%	KNN	n/a	Not mentioned	Cleveland (UCI)	
Wiharto et al. [18] 61.86%	SVM	BT SVM	Not mentioned	Cleveland (UCI)	
Khateeb and Usman [3] 79.20% bagging technique	NB, KNN, DT and	KNN	WEKA	Cleveland (UCI)	
NB, DT, MLP, KNN, Pouriyeh et al. [6] 84.81% bagging, boosting and	SCRL, RBF, SVM, SVM	Boosting with	Not mentioned	Cleveland (UCI)	
American Heart	89% Association dataset				
ANN and Genetic Waghulde and Patil [8] 98% system	Algorithm hybrid	n/a	American Heart Association dataset MATLAB		
Amma [20]		n/a	Not mentioned	Cleveland (UCI)	94.17%
enkatalakshmi and 85.03% Shivsankar [21]	NB and DT	NB	WEKA	UCI	
Palaniappan and Awang [22]	DT, NB and ANN	NB	DMX	Cleveland (UCI)	86.53%
Cleveland and	Nearly 100% Statlog (UCI)				
NB, DT, Discriminant, Zriqat et al. [24] and SVM	Random Forest, and	DT	MATLAB	99.01% for Cleveland and 98.15% for Statlog (UCI)	Cleveland and Stat...
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%

nd Radial Basis records and 19	SVM	Not mentioned	dataset of 214	Indian patients 86.42% Function attributes
Masethe and Masethe [27] Net	J48, NB, REPTREE, Simple Cart, and Bayes	Simple Cart	WEKA 11 attributes	South African dataset containing 99.0741%

V. CONCLUSION AND FINAL REMARKS

This paper overviews the literature of machine learning classifica- tion 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 ma- chine 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 pro- fessionals 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 reposito- ry. Since the quality of the dataset is an essential factor in the pre- diction'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

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.

[2]   J. Soni et al., "Intelligent and effective heart disease prediction sys- tem 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 sys- tem using k-nearest neighbor classification technique," in Proceed- ings of the International Conference on Big Data and Internet of Thing (BDIOT), New York, NY, USA: ACM, 2017, pp. 21–26.

[4]   H. Almarabeh and E. Amer, "A study of data mining techniques accuracy for healthcare," International Journal of Computer Appli- cations, 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.

[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 Communica- tions (ISCC). Heraklion, Greece: IEEE, July 2017, pp. 204–207.

INTERNATIONAL JOURNAL
OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

📱 9940 572 462 🟢 6381 907 438 ✉ ijircce@gmail.com

Scan to save the contact details