Analysis of WEKA data mining algorithms Bayes net, random forest, MLP and SMO for heart disease prediction system: A case study in Iraq

Rana Riad K. AL-Taie, Basma Jumaa Saleh, Ahmed Yousif Falih Saedi, Lamees Abdalhasan Salman
Department of Computer Engineering, Al-Mustansiriyah University, Baghdad, Iraq

Article Info

Article history:

Received Feb 17, 2020
Revised May 29, 2021
Accepted Jun 11, 2021

Keywords:

Classification
Data mining
Dataset
Heart disease
WEKA tool

ABSTRACT

Data mining is defined as a search through large amounts of data for valuable information. The association rules, grouping, clustering, prediction, sequence modeling is some essential and most general strategies for data extraction. The processing of data plays a major role in the healthcare industry's disease detection. A variety of disease evaluations should be required to diagnose the patient. However, using data mining strategies, the number of examinations should be decreased. This decreased examination plays a crucial role in terms of time and results. Heart disease is a death-provoking disorder. In this recent instance, health issues are immense because of the availability of health issues and the grouping of various situations. Today, secret information is important in the healthcare industry to make decisions. For the prediction of cardiovascular problems, (Weka 3.8.3) tools for this analysis are used for the prediction of data extraction algorithms like sequential minimal optimization (SMO), multilayer perceptron (MLP), random forest and Bayes net. The data collected combine the prediction accuracy results, the receiver operating characteristic (ROC) curve, and the PRC value. The performance of Bayes net (94.5%) and random forest (94%) technologies indicates optimum performance rather than the sequential minimal optimization (SMO) and multilayer perceptron (MLP) methods.

Corresponding Author:

Basma Jumaa Saleh
Department of Computer Engineering
Al-Mustansiriyah University
Baghdad, Iraq
Email: eng.basmaj@uomustansiriyah.edu.iq

1. INTRODUCTION

The complications of heart attack can be considered as the main world’s leading causative agent to, to stop attacks in conjunction with early diagnosis. Many of information, usually produced by physicians with rich hidden material, but used inefficiently for forecasting. Hence, utilizing many of data mining strategies helped in turning unused data into a useful data set. Several of signs have not been taken into account, which let to dying people. Professionals of medical should predict heart disease before it happens in any patient [1]. There are many of characteristics that may increase the possibility of heart diseases [2]: i) Smoking: Destroys the lining of the arteries by releasing a fat content, such as atheroma, which decreases the arteries that activate heart attack, ii) High cholesterol: Cholesterol is a waxy material found in the fatty plaques of blood vessels. High cholesterol doesn't really allow sufficient blood to enter the lungs, causing heart disease, iii) Inappropriate diet: Blood pressure and cholesterol are increased by eating so much unhealthy food, that can cause heart disease, iv) Lack of physical activity: An increase in the levels of cholesterol in the muscles, leading to the probability of heart attacks, v) Harmful alcohol intake: A psychoactive series of use that causes damage to health. The purpose may be physiological or behavioural or e.g. Primary to serious drinking depressed episodes. Harmful locations do not always have harmful psychological consequences; however, the social impacts are not adequate to justify a diagnosis of harmful use [3], vi) High sugar levels: Measurements of blood sugar higher than 180 mg/DL or any measurements outside the normal range are abnormal. A blood sugar test of 300 mg/DL or higher may be dangerous. When you have two readings in a sequence of 300 or more, call your doctor. In high insulin cases, the practice of plural, growing hunger and polyphonic [4] is generally observed, vii) Overstress: The unspecific response of the body to any request is a condition generally occurring during its entire lifespan. All individuals in their society and history have felt it. Stress has become one of life's extra characteristics and its essence has been highlighted in order to be explored in all eras of art history and fiction [5], and viii) Blood pressure: A rare disorder in which blood powerful enough for the walls of the artery can ultimately trigger health conditions [6], age, gender and family history of diabetes. A Situation common [7]. Such reasons can be used as lifestyle factors for the prediction of cardiac disease [8]. Many kinds of heart injury conditions are included in the term cardiac illness. Heart disorders are prevalent in:

a. Coronary heart disease: The much more common method of heart disease in the world is coronary heart disease. It is also called heart disease. Statue particles block the coronary channel, causing a reduced flow to the cardiovascular of oxygenated blood.
b. Arrhythmias: It is related to the irregular activity of the heartbeat. It may be a low, fast or irregular heartbeat. Besides irregular heartbeats, there is a defect in the cardiovascular system.
c. Heart failure: It is a condition in which enough blood can not be delivered to the specific body by the heart. Normally, it is pointed to as heart problems.
d. Congenital heart disease: it is often pointed to as a congenital cardiac condition and leads to an abnormal carbonate developmental stage and function. Often, infants with a congenital disorder.
e. Cardiomyopathy: undermining the heart muscle or affecting the musculature due to the improper beating of the heart. Via cardiomyopathy. The most common causes of cardiomyopathy are high blood pressure, alcohol intake, bacterial infections, and genetic abnormalities.
f. Angina pectoralis: is a medical procedure for angina that happens when the heart is not properly supplied with blood; it is a sign of a heart attack. There are several seconds or minutes of chest pain.
g. Myocarditis: It is a cardiac infection usually affecting the heart that is viral, fungal, and bacterial. It is an irritable heartbeat. It is a rare condition that has no direct correlation with pain, arm stiffness or temperature [9].

All these conditions are the main causes of death for individuals all over the world. The WHO and CDC has indicated that the major cause of mortality is cardiovascular disease [8] and disease prevention centres. In today's world, data mining in medical treatment is becoming more popular because it offers a great variety of complex knowledge that includes healthcare facilities, medicines, medical devices, patients and disease diagnosis. Such complex data must be processed and evaluated for the retrieval of information, Which, indecisions, is both price-effective and beneficial. In 2011, the World Health Organisation lost 17.5 million patients with heart disease, which is 31 percent of all foreign deaths. Of these, coronary heart disease affected 7.4 million and cerebro-spinal diseases affected 6.7 million. Almost 23.6 million people will die of heart attacks, approximately by the Health Organization in 2030 [10]. Through some hospital system monitoring systems, many clinics electronically storing their patient records. Each day, such devices generate vast quantities of data. Such data can be arranged as servers in infinite text or in picture form. For choice-making criteria, such data can collect useful information. This presumption contributes to the use of knowledge creation in data sets, which converts small-level data into information for heavy-level decision-making. The results can be used and can be further explored in good decisions and analysis. By contacting, data mining is graded, clustered, analyzed and identified [9]. Applications for data mining forecast future developments by knowledge-based decision-making. Cardiovascular disease identification requires an immense amount of data, too difficult and massive for current techniques to be processed and interpreted. A number of techniques of data mining are used by experts. Our aim is to find algorithmically efficient data mining applications. Different algorithms for the analysis of data mining are applied in this paper to health data on heart diseases. This helped determine the best prediction strategy on the collected data set in terms of its precision, Kappa, receiver operating characteristic (ROC) and accuracy.

The rest of the article is arranged accordingly: The problem statement is illustrated in section 2. The review of literature and related works is illustrated in section 3 The method of our experiment is explained in section 4. The results and performance comparisons of our experiment are explained in section 5. Section 6 eventually draws our conclusions.

2. PROBLEM STATEMENT

The implementation of machine learning methods for the classification and prediction of heart disease has been investigated in previous researches. These, however, offer a model for prediction of heart disease for the diagnosis of heart disease incidence. In addition, this analysis aims to determine the best classification method to find the risk of heart disease in a case. This research is justified by a comparative study and observation using four classification techniques, i.e. sequential minimal optimization (SMO), multilayer perceptron (MLP), random forest and Bayes net. The evaluations are used at various levels. While these machine learning methods are widely used, the prediction of heart disease is a critical task requiring the highest possible precision, comparing with [11]-[17]. Therefore, the four algorithms are tested in a number of assessment levels and types. It provides medical researchers and physicians with a greater understanding and helps them find the best way to prevent cardiac disease. WEKA software should be used in the proposed framework. The Weka software tool has been used to evaluate heart disease data. This paper's key contributions are:

a. Classified precision extraction is important for prediction of heart disease.
b. Use the Ibn al-Bitar Hospital Cardiac Surgery and the Baghdad Medical City electronic diagnostic cardiac condition database and the collected actual information database for the training and testing of the program.
c. To achieve the highest level of classification accuracy bayes net (94.5%) and random forest (94%) methods, investigate knowledgeable classification strategies.
d. Evaluation of suggested classifier classification results. And check the performance of the classifiers suggested by comparing them with existing classifiers of other works.
e. Evaluate the best results of the suggested WEKA software classifiers.
f. Comparison of various algorithms for data mining on the dataset for cardiovascular disease.
g. Classification of the best algorithms for prediction of heart disease based on the results.

3. LITERATURE REVIEW AND RELATED WORKS

Dwivedi [18] used six machine learning classification techniques that were applied to the heart disease dataset. In this study, this author used tenfold cross validation for evaluation and eleven performance measures for comparison. Thereafter, a study of Gharehchopogh *et al.* [11]. Researchers used 40 people in their medical records. Blood pressure, gender, age and tobacco use are the conditions used for detection. The model correctly anticipated 85% of cases. Multilayer perceptron (MLP) utilization on the heart disease datasets exceeded accuracy by 80.89% in the WEKA software. Ramotra *et al.* [12] Suggestion of a machine learning model for using the WEKA method for predicting cardiovascular disease. The data contained 303 data and 76 specifications. 297 data with 13 input functions are required for analysis after pretreatment of data and removal of missing values. The authors claim to be 80.89 percent accurate. An efficient heart disease detection system was introduced by Purushottam *et al.* [17] data mining utilization. It can help doctors make parameter-based decisions effectively. The device is formed and tested by a model 10 times, and the precision of 86.3% during the test and 87.3% during the training process is proven. The authors noted that the overall accuracy of the multicoyer perceptron (MLP) classification was 74.85%.

Jothikumar *et al.* [18] Suggestion of a model using a learning method to estimate medical history with 295 samples and 13 characteristics apply to the naive Bayes algorithm in quick producer. Other similar metrics are Kappa 0.499, absolute error 0.247%, RMSE is 0.378, and relative error 24.19%. Sarangam Kodati *et al.* [19] It is suggested that the preceding analysis is 77.9% in Orange and 73.4% in Recall of cardiopathy results. In the WEKA precession, 81.8 percent and recall, 81.9 percent. Comparison between the software Orange and WEKA, Weka is the best reminder and precession.

The sequential minimal optimization (SMO) method was introduced by Platt [20] in 1998 and was the fastest method for optimizing algorithmic programming. Sequential minimal optimization (SMO) is used to prepare the algebraic kernel or RBF kernel vector classification supporters. This replaces all conditional attributes with the null values and transforms them into binary ones. Aung *et al.* Suggests a machine learning approach for predicting heart conditions using the WEKA tool [15], design that utilizes a minimum sequential optimization strategy and a mitigation strategy for lazy classification. The Weka data mining approach has been used to predict heart disease. 66 percent of the data set (training) and 34 percent (testing) for analysis was instructive.

In order to evaluate heart disease, Mirmozaffari *et al.* [16] proposed a method for the classification of various data mining methods. It has developed a particular model of different filters and methods of analysis. For multi-layer pre-process filtering, the superior approach and the more precise clinical resolution assistance systems for the diagnosis of diseases are used, as are varying.

The UCI system information is routinely viewed in a database or in a report. This work uses the Waikato framework for knowledge evaluation. The data sets must be in the attribute-relation file format (ARFF), to use this data for the WEKA method. In pre-processing the dataset, the WEKA method is used. Just major attributes, i.e. 13 in this case, are taken into account when evaluating all these 13 attributes, which provide better and clearer results. After all, unimportant attributes are discarded. The 13th is essentially an expected class feature. Through analyzing the various decision tree algorithms inside WEKA tools extensively and making the choices it makes, the device will help predict the probable existence of cardiac diseases in a patient and definitely help diagnose cardiac diseases well in preparation and cure them in good time. Some of the standard machine learning of data mining challenges is in the following areas:

a. Extraction of valuable information and development of scientific decision-making capability for disease treatment and diagnosis.
b. Classification of the developments of effective medical treatments for various ailments.
c. Too many attributes available for decision-making so must determine which the best prediction of heart disease.
d. With the assistance of computerisation, voluminous real data (text, graphs, and images) are now being processed, but it is still more difficult to collect.
e. Handling noisy (containing errors or outliers), confusing (containing code or name discrepancies) and lack of attributes to be pre-processed for medical data problems.
f. Determine the best tools and algorithms for analysis the datasets by using WEKA tools, and for future work trying to use MATLAB program for developing the work.

4. RESEARCH METHOD

The purpose of this study is to successfully predict possible heart attacks from the compilation of medical data. Using prediction algorithms to evaluate the characteristics of cardiac disease by certain attributes, a model have been developed. Data mining is used in this work to create class predictive models based on features selected. The Waikato environment for knowledge research (WEKA) has been used for prediction because of its ability to discover, study, and forecast trends. It is typically possible to divide the entire process into 6 stages:

4.1. Description of the algorithms

Heart disease is a word used to describe a large variety of health circumstances associated with the heart. These medical conditions specifically describe the pathological diseases of the heart as well as all parts of it. A substantial health concern is heart disease. Over the years, the number of people who have heart disease has increased [20]. Several studies focused on the management of heart disease have been conducted. Various techniques for diagnostic data mining have been applied and various probabilities have been obtained. Many studies are being conducted to assess the inefficiency of MLP, Bayes net, SMO and random forest algorithms. There are several possible strategies to treat heart disease [21]:

a. MLP: The perceptron multi-layer algorithms help the problems of regression and classification. It is also called, for short, artificial neural networks or just neural networks. Neural networks are a challenging algorithm to be used for predictive modeling since there are so many parameters of configuration that can be effectively tuned only by observation and a number of trial and error [20].
b. Random forest: An ensemble of random decision tree classifiers is a random forest that makes predictions by combining the individual trees' predictions. In the decision tree construction process, various methods are possible to incorporate randomness. To make forecasts about classification or characteristics, a random forest can be used. One of the best predictive analytics is random forests [22].
c. Sequential minimal optimization (SMO): is an algorithm for solving the quadratic programming (QP) problem that arises during the training of support-vector machines (SVM). It is commonly used for machine learning training, support and is introduced by the common LIBSVM tool. In the SVM community, the publishing of the SMO algorithm in 1998 created a lot of anticipation, as previously available techniques for SVM training were much more complicated and costly third-party QP solvers were required [23].
d. Bayes net: The Bayesian network is a combination of probability and graphic models. It is widely applicable in machine learning, data mining, and diagnostics. because it has a solid evidentiary-based conclusion that is familiar to human intuition [24].

4.2. Tools and data source

The Waikato Environment for Knowledge Analysis 2018 (WEKA 2018, version 3.8.3) had been adopted as the standard interface to compare different data mining techniques and determine the best methods [25]. The standard data kit had been getting from the Iraqi hospitals under the oversight of the National Minister of Health includes 200 samples. To detect heart disease with a high degree of accuracy, a large range of relevant inputs must be considered. The physician relies on all the recorded symptoms, patiently answering questions, medical testing and laboratory performances. Overall, the data had been collected from the Ibn al-Bitar Hospital and the Baghdad Medical city based on these medical factors to provide appropriate medical criteria for the detection of heart disease. There has been considerable difficulty in collecting these factors by some medical variables, such as (Maximum cardiac rate, ST depression, fairly restful exercise, the slope of the ST highest exercise section, and Number of key fluoroscopy-colored vessels) Such variables are therefore substituted for medical causes by cardiologists (heart rate, family history, smoking, hyperkinesia Echo, and an earlier angina assault) [26].

The adapted medical variables, consider the causal factor, the family medical history, besides the observed echo the probability of prior angina to get adequate medical causes; these data would include four classes of heart disease besides normal classes. Table 1 shows the availability of the five classes of cardiac diseases and includes 13 medical features required for cardiovascular treatment. To create a diagnostics system, these factors are turned into a numerical simplification [27].

Table 1. Collected dataset (CD)

Age	Real (0-76)			
Sex	Male "0"		Female "1"	
CP	typ_angina "1"	Asympt "2"	non_anginal "3"	atyp_angina "4"
BP	Real			
Col	Normal "0"			Abnormal "1"
Fobs value	>120 mg/del "1" true			<120 mg/del "0" false
Rest ECG	Normal "0"	Abnormality "1"		left_vent_hyper "2"
Thalach Value	Real			
Exam	No "0"		Yes "1"	
FH	No "0"		Yes "1"	
SM	No "0"		Yes "1"	
HYP	No "0"		Yes "1"	
PERANGINA	No (Negative) "0"		"1"	Yes (Positive)

4.3. Attribute description

a. Age: represents in years the numeric value of age.
b. Sex: which will be represented in binary (0=male, 1=female).
c. Cup Type: the abbreviation of Chest pain types, which will be introduced as follows:
   Value 1: Typical Burning Sensation in heart.
   Value 2: Acute stabbing (such as pain).
   Value 3: Burning Sensation.
   Value 4: Acute Crushing Pain in heart.
d. Col: Cholesterol level in patient where 1 = Abnormal, 0 = Normal
e. Fobs: fasting blood sugar level where 1 = true (>120 mg/del), 0 = false (<120 mg/del).
f. Rest ECG: the abbreviation of Rested Electrocardio Graphic, the indicated values of the report are:
   Value 0: normal
   Value 1: ST-T wave Abnormality
   Value 2: Ventricular Abnormality.
g. Thalach Value: shows the achieved maximum heart rate.
h. Exam: which implies the engine that caused by exercising (1 = yes, 0 = no).
i. FH: family history can be as strong of a marker for heart disease (1 = yes, 0 = no).
j. SM: Smoking increases the risk of developing cardiovascular diseases (1 = yes, 0 = no).
k. HYP-Echo finding for hypo Kinesis (1 = yes, 0 = no).
l. PERANGINA: previous attack of angina (1 = yes, 0 = no).

m. Class: Class of Patients with heart disease. Value 0: Coronary Heart disease, Value 1: Angina pectoris, Value 2: Congestive heart failure, Value 3: Arrhythmias, Value 4: Normal.

4.4. Performance metrics

The metrics used in the analysis will be defined in detail throughout this section [25]:

4.4.1. Precision

Precision1: the part between the accumulated instances of major cases. The precision equation is:

$$\text{Precision1} = \text{TP1}/\ (\text{TP1}+\text{FP1}) \tag{1}$$

4.4.2. Recall

The small subset of the required instances in the overall number of particular instances. The recall equation is:

$$\text{Recall1} = \text{TP1}/\ (\text{TP1} + \text{FN1}) \tag{2}$$

4.4.3. F-Measure

The f-measure is examined based on the 2-fold precision reminder period separated by the sum of accuracy and reminder [28]. The F-Measure equation is provided in (3).

$$F = \frac{TP1 * TN1 - FP1 * FN1}{\sqrt{(TP1+FP1)(TP1+FN1)(TN1+FP1)(TN1+FN1)}} \tag{3}$$

4.4.4. Area of ROC

ROC equations are commonly used as visuals about any cutoff, including clinical sensitivity and accuracy, for an assessment or a variety of tests, relationships, and trade-off.

4.4.5. Area of PRC

The number of lower grades of patients without a diagnosis are not affected by curves for correct recall. It is particularly important to use precision recording formulas to supplement the ROC formulas to obtain the complete spectrum during analysis and selection. The classification model product [26], as shown in Table 2.

Table 2. Various effects of a two-class model

Specific Class	Class Predicted	
	YES	No
YES	Positive True (TP1)	Negative False (FN1)
No	Positive False (FP1)	Negative True (TN1)

a. Positive true (TP1): It was fairly expected that patients were positive (Patients are likely to require heart failure and heart cauterisation.).
b. Positive false (FP1): If TP1 and TN1 are approximately 100 percent, the model is ideally predicted to be negative, because they are not supposed to have a cardiac catheterization.
c. TN1 is a negative true: Healthy people are properly classified as healthy.
d. FN1 is negative false: Classified incorrectly as healthy [28] heart disease patients.
e. Correct classified cases (CCC): This represents the proportion of patients who need and not need heart surgery and are diagnosed correctly. Accuracy [29] is also known as elk (4).

$$\text{Accuracy}= \frac{TP1+TN1}{TP1+TN1+FP1+FN1} \tag{4}$$

f. Mean absolute error (MAE): A test of predictors. The calculation of 1- ACC is probable. A strong system has a very high absolute mean error [30].
g. Kappa: Prediction identification with a correct class is checked by Kappa. The statistical effect of a kappa is a score in the 0-1 range. A value greater than 0 means it is better than average for the classifier [31].
h. Root mean squared error (RMSE1): The difference between the value predicted and the value observed [32] is the root mean squared error.

$$\text{RMSE1} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\frac{\rho_{i1,j1}-\tau_j}{\tau_j}\right)^2} \quad (5)$$

$\rho_{i1,j1}$ = Value predicted.
τ_j= Max value of fitness applicability of j.

4.5. Proposed strategies

This research aims to predict the possibility of heart disease occurrence by early automatic diagnosis within short time. In addition, that will help healthcare professionals to treat their patients early based on accurate decision-making. In addition, the proposal has a crucial role in healthcare Organization especially for experts with having less knowledge and skills. The accurate results considered as the major limitation of existing methodology. The proposal used both data mining techniques and machine learning algorithms SMO, MLP, random forest and Bayes net, with k-fold cross-validation to predict the occurrence of heart disease. Many of medical attributes had been used to identify if the patient either has heart disease or not, such as blood pressure, cholesterol, age, blood sugar, sex, and heart rate. The data set had been analyzed and computed using the WEKA software. WEKA is open-source software that includes a set of machine learning algorithms for the data mining tasks. WEKA had been implemented with Java code. WEKA contains several tools, which are important in data mining tasks: preprocessing data, regression, clustering, classification, association and visualization. The analysis of WEKA methodology as shown in Figure 1.

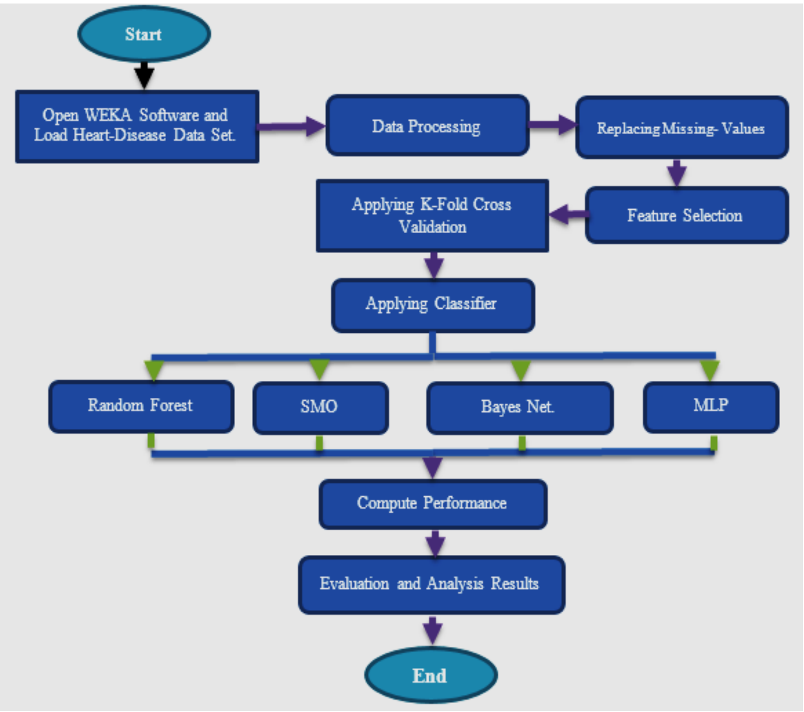

Start
Open WEKA Software and Load Heart-Disease Data Set
Data Processing
Replacing Missing- Values
Feature Selection
Applying K-Fold Cross Validation
Applying Classifier
Random Forest
SMO
Bayes Net
MLP
Compute Performance
Evaluation and Analysis Results
End

Figure 1. The steps followed in the methodology

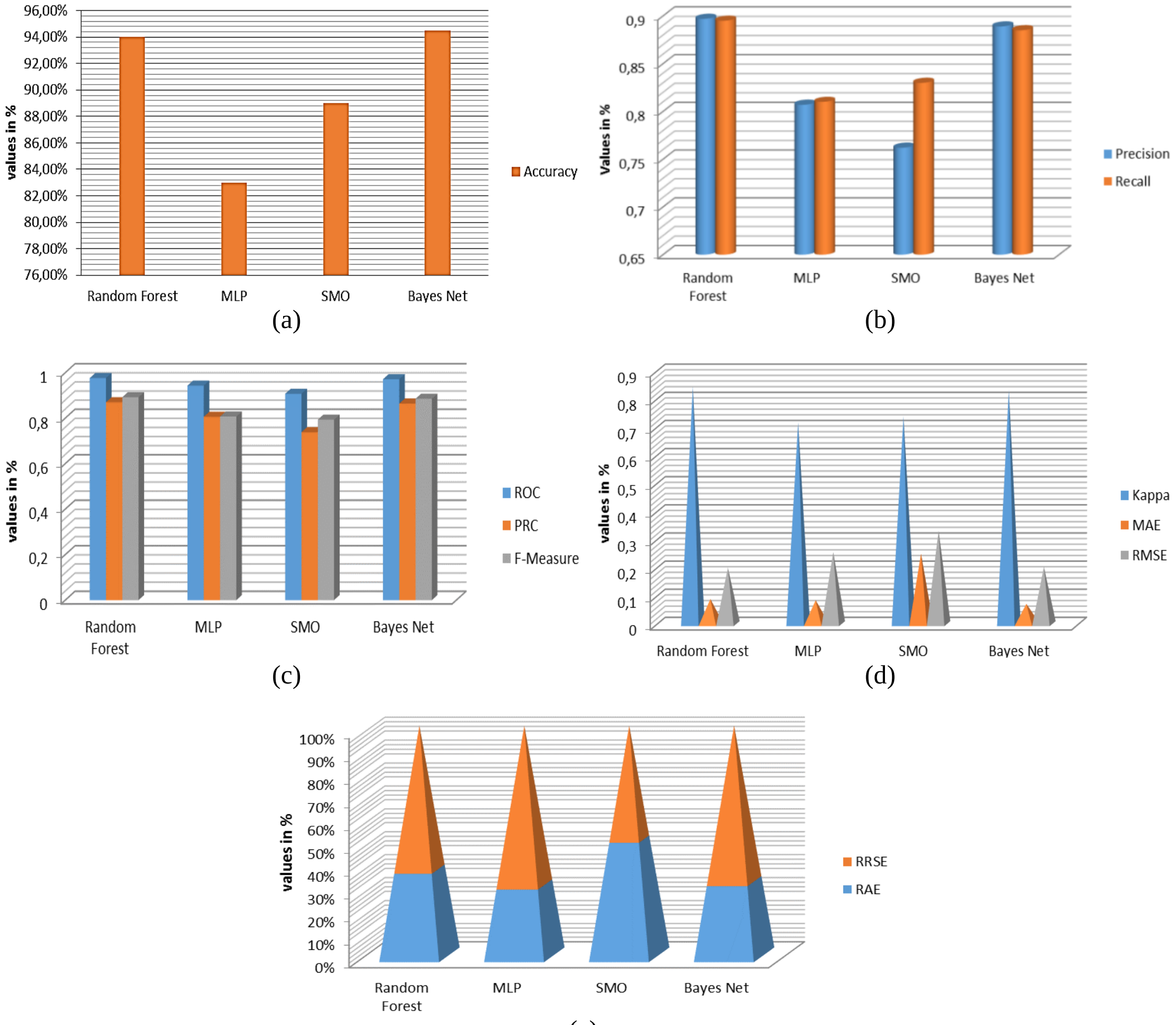
5. RESULTS AND PERFORMANCE COMPARISONS

The Weka data mining tool has been used for clinical forecasts, Part of the dataset is used for learning and the remainder of testing the classification results are shown in the Table 3. The results of the classification diagram are shown in Figure 2. To conclude, the estimate results separated by a different machine learning techniques. The platform was built with 200 units. Several evaluation metrics were compared and shown in Figure 2: Classification precision, REM, Kappa, MAE, RMSE, RAE, RRSE, F-measure, PRC and ROC values. Based on the accuracy of classifications which is calculated by (4), Figure 2(a), The highest accuracy is given by Bayes Net, about 94.50% and The worst (83%) achieved by multilayer perceptron (MLP) in estimating heart disease cases. From Figure 2(b), it can be concluded that Random Forest has the highest precision measure (0.897), and recall (0.895). Mostly on base of Figure 2(b).

It can also be shown that in the recall metrics, multilayer perceptron (MLP) and SMO are almost identical, but distinct in precision. Figure 2(c) demonstrates that Bayes Net performs the best, providing maximum F-measure values (0.885), ROC (0.971) and PRC (0.864). Figure 2(d) demonstrates that, while random forest, MAE (0.088) highly smaller than sequential minimal optimization (SMO) (0.2492) but RMSE of random forest (0.1992) is the least and Kappa of random forest (0.8431) is the greatest. Figure 2(e) shows RRSE of Bayes Net (54.84%), which contributes to better prediction results.

Table 3. The comparison result of the four classifiers

Algorithm Parameter	Random Forest	Multilayer Perceptron (MLP)	Sequential Minimal Optimization (SMO)	Bayes Net
Correctly classified instances	89.5%	81%	83%	88.5%
Incorrectly classified instances	10.5%	19%	17%	11.5%
Kappa statistic	0.8431	0.7156	0.7374	0.8301
Mean absolute error	0.088	0.0835	0.2492	0.0707
Root mean squared error	0.1992	0.255	0.3303	0.202
Relative absolute error	32.2994%	30.6718%	91.4867%	25.9664%
Root relative squared error	54.0999%	69.2363%	89.7072%	54.8438%
Accuracy	94%	83%	89%	94.5%
Construction Time	0.14 Sec	0.56 Sec	0.13 Sec	0.01 Sec



(a) (b) (c) (d) (e)

Figure 2. Evaluation of performance metrics using percentage split of, (a) Accuracy, (b) Precision and recall metrics, (c) ROC, PRC, and F-measure metrics, (d) Kappa, MAE and RMSE metrics, (e) RAE and RRSE metrics

To evaluate the efficiency of classification strategies for class prediction and determination accuracy, the algorithm is employed in the data set through stratified 10-fold testing. The resulting uncertainty matrix calculates the measurements for accuracy, sensitivity, and specificity. The matrix applies to samples labeled as true, others as false and others as wrong. Confusion matrix estimation reveals that sequential minimal optimization (SMO), multilayer perceptron (MLP), random forest and Bayes net show 200 instances with the positive causal factor for a heart attack. Predictions show a predictive model. The methods strongly advise that techniques for data mining can predict a diagnostic class. The matrix of uncertainty specifically classifies the functional accuracy. The matrix confirms the model's performance.

We checked the formulas for the classification of heart diseases mentioned in the work experience section of the suggested classifiers. It contrasts the suggested WEKA method classification method with other research findings in Table 4 and Figure 3, in comparison with the current methods and experimental tests, we consider that our proposed system is better than the other model in prediction and diagnosis of heart disease. Therefore, the precision of the classification of current models is improved.

Table 4. Comparison between heart disease prediction system using different techniques and other works

Model of other works	Techniques used	Accuracy	Proposed Model
Gharehchopogh *et al.* [11] (2011)	MLP	80.89%	83%
A. K. Ramotra *et al.* [12] (2020)	MLP	80.89%	83%
Purushottam *et al.* [13] (2016)	MLP	74.85	83%
Aung Nway Oo *et al.* [14] (2019)	SMO	82.6%	89%
Mirpouya Mirmozaffari *et al.* [15] (2017)	Bayes Net	80.83%	94.5%
R. Jothikumar *et al.* [16] (2016)	Random Forest	78.24%	94%
Sarangam Kodati *et al.* [17] (2018)	Random Forest	81.9%	94%

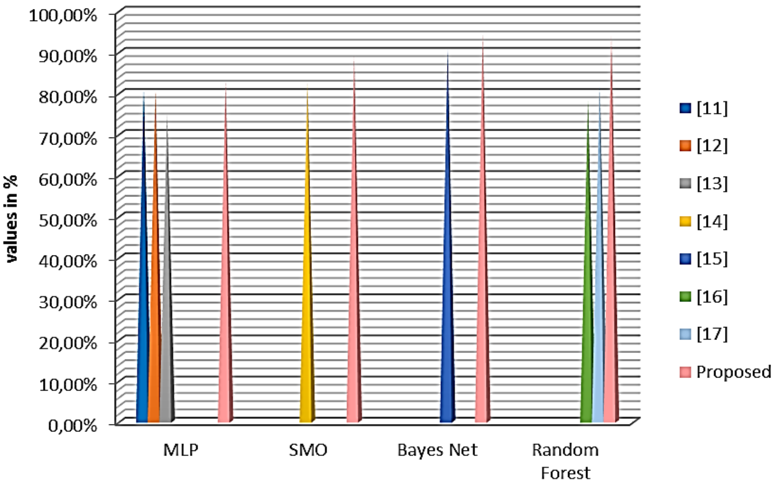


Figure 3. Comparison between heart disease prediction system using different techniques and other works

6. CONCLUSION

In this analysis, we have submitted an effective prediction method for heart disease with data extraction and test the accuracy of heart disease prediction with a group of classifiers. The collected heart database for training and testing purposes was used from the hospital of Ibn al-Bitar and Baghdad medical city. This program will assist physicians inaccuracy, parameter-specific decisions. The research has been successfully performed in several techniques for the classification of data mining (SMO, MLP, Bayes Net and Random Forest) with a diagonal output of tenfold, and it is found that the Bayes Net algorithm gives greater accuracy than the other data set supplied (94.5%). It can also be used with many classification techniques.

ACKNOWLEDGEMENTS

The authors would like to thank Mustansiriyah University (www.uomustansiriyah.edu.iq) Baghdad-Iraq for its support in the present work.

REFERENCES

[1] A. K. Sen, S. B. Patel, and D. Shukla, “A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level,” *International Journal of Engineering and Computer Science (IJECS),* vol. 2, no. 9, pp. 1663–1671, 2013.

[2] K. Divya and K. Navpreet, “Review On Prediction System For Heart Diagnosis Using Data Mining Techniques,” *International Journal of Latest Research in Engineering and Technology (IJLRET),* vol. 1, no. 5, pp. 9–14, 2015.

[3] J. Adamson, N. Heather, V. Morton, and D. Raistrick, “Initial preference for drinking goal in the treatment of alcohol problems: II. Treatment outcomes,” *Alcohol and Alcoholism,* vol. 45, no. 2, pp. 136–142, 2010, doi: 10.1093/alcalc/agq005.

[4] S. Lal, “Diabetes: Causes, Symptoms And Treatments,” *Public Health Environment and Social Issues in India. Edition: 1. Chapter: 5. India: Serials Publications*, pp. 55–67, 2016.

[5] A. M. Shahsavarani, E. A. M. Abadi, and M. H. Kalkhoran, “Stress: Facts and Theories through Literature Review,” *International Journal of Medical Reviews*, vol. 2, no. 2, pp. 230–241, 2015.

[6] Blood Pressure, “What Is High Blood Pressure,” *The South Carolina State Library Digital Collections*, 2017. [Online]. Available: https://dc.statelibrary.sc.gov/handle/10827/25131.

[7] F. Charles, M. Catherine, F. Julian, W. DeWayne, and B. Andrew, “Sex and family history of cardiovascular disease influence heart rate variability during stress among healthy adults,” *J. Psychosom Res.*, vol. 110, pp. 54–60, 2018, doi: 10.1016/j.jpsychores.2018.04.011.

[8] M. Sultana, A. Haider, and M. S. Uddin, “Analysis of data mining techniques for heart disease prediction,” *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT).* Dhaka, Bangladesh, 2016, pp. 1–5, doi: 10.1109/CEEICT.2016.7873142.

[9] B. J. Saleh, A. Y. F. Saedi, A. T. Q. Al-Aqbi, and L. A. Salman, “A Review Paper: Analysis of Weka Data Mining Techniques For Heart Disease Prediction System,” *Library Philosophy and Practice*, pp. 1–17, 2020, doi: 10.30491/ijmr.2020.221474.1078.

[10] S. Chaitrali and S. Apte, “A Data Mining Approach for Prediction of Heart Disease Using Neural Networks,” *International Journal of Computer Engineering and Technology (IJCET)*, vol. 3, no. 3, pp. 30–40, 2012.

[11] S. Charehchopogh and Z. Khalifelu, “Neural network application in diagnosis of patient: a case study,” *2011 International Conference on Computer Networks and Information Technology*, Abbottabad, 2011.

[12] K. Ramotra, A. Mahajan, R. Kumar, and V. Mansotra, “Comparative Analysis of Data Mining Classification Techniques for Prediction of Heart Disease Using the Weka and SPSS Modeler Tools,” *Smart Trends in Computing and Communications. Smart Innovation, Systems and Technologies*, vol. 165, pp. 89–97, 2020, doi: 10.1007/978-981-15-0077-0_10.

[13] R. Jothikumar and V. Sivabalan, “Analysis of Classification Algorithms for Heart Disease Prediction and its Accuracies,” *Middle-East Journal of Scientific Research*, pp. 200–206, 2016.

[14] Heart Disease, “General Info and Peer reviewed studies,” [Online]. Available: http://www.aristoloft.com.

[15] N. Aung and T. H. Naing, “SMO and Lazy Classifiers for Heart Disease Prediction,” *International Journal of Advance Research and Innovative Ideas in Education (IJARIIE),* vol. 5, no. 2, pp. 2395–4396, 2019.

[16] M. Mirpouya, A. Alinezhad, and A. Gilanpou, “Data Mining Apriori Algorithm for Heart Disease Prediction,” *International Journal of Computing, Communication and Instrumentation Engineering*, vol. 4, no. 1, pp. 20–23, 2017, doi: 10.15242/IJCCIE.DIR1116010.

[17] Purushottama, K. Saxenab, and R. Sharma, “Efficient Heart Disease Prediction System,” *Procedia Computer Science,* vol. 85, pp. 962–969, 2016, doi: 10.1016/j.procs.2016.05.288.

[18] A. Dwivedi, “Performance evaluation of different machine learning techniques for prediction of heart disease,” *Neural Comput and Applic,* vol. 29, no. 10, pp. 685–693, 2018, doi: 10.1007/s00521-016-2604-1.

[19] K. Sarangam and R. Vivekanandam, “Analysis of Heart Disease using in Data Mining Tools Orange and Weka,” *Double Blind Peer Reviewed International Research Journal*, vol. 18, no. 1, pp. 16–22, 2018.

[20] J. Platt, “Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines,” *Technical Report MSR-TR*, pp. 98–14, 1998.

[21] J. Brownlee, “How To Use Classification Machine Learning Algorithms in Weka,” *Machine learning mastery*. [Online]. Available: machinelearningmastery.com.

[22] C. Vens, “Random Forest,” *Encyclopedia of Systems Biology, Springer*, 2013.

[23] H. Naveed, G. Khan, A. U. Khan, A.Siddiqi, and M. U. G. Khan, “Human activity recognition using mixture of heterogeneous features and sequential minimal optimization,” *Int. J. Mach. Learn. and Cyber*, vol. 10, pp. 2329–2340, 2019, doi: 10.1007/s13042-018-0870-1.

[24] L. Nguyen, “Overview of Bayesian Network,” *Science Journal of Mathematics and Statistics*, vol. 2013, pp. 1–99, 2013.

[25] M. Kumar, K. Nikhil, S. Koushik, and K. Deepak, “Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools,” *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 3, no. 3, pp. 887–898, 2018.

[26] T. T. Hasan, “Design and Implementation of Intelligent Algorithm for the Diagnosis of Heart Disease Using FPGA,” MSc. Thesis, University of Technology, Iraq, 2017.

[27] H. Benjamin, F. David, and S. Belcy, “Heart Disease Prediction Using Data Mining Techniques,” *ICTACT Journal On Soft Computing*, vol. 9, no. 1, pp. 1824–1830, 2018.

[28] I. Cvitić, D. Perakovic, M. Perisa, and B. Gupta, “Ensemble machine learning approach for classification of IoT devices in smart home,” *Int. J. Mach. Learn. and Cyber*, 2021, doi: 10.1007/s13042-020-01241-0.

[29] S. Sakr *et al.*, "Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford ExercIse Testing (FIT) Project," *PLoS ONE*, vol. 13, no. 4, pp. 1–18, 2018, doi: 10.1371/journal.pone.0195344.

[30] A. Masrani, M. Shukla, and K. Makadiya, "Empirical Analysis of Classification Algorithms in Data Stream Mining," *Advances in Intelligent Systems and Computing*, pp. 657–669, 2020, doi: 10.1007/978-981-15-5113-0_53.

[31] S. Ghanem and H. Elgazzar, "Predicting the behavior of reinforced concrete columns confined by fiber reinforced polymers using data mining techniques," *SN Applied Sciences*, vol. 3, no. 2, 2021, doi: 10.1007/s42452-020-04136-5.

[32] H. V. Sánchez and H. S. Angulo, "Use of Data Mining for Root Cause Analysis of Traffic Accidents in Colombia," *Cross Reality and Data Science in Engineering*, pp. 674–688, 2020, doi: 10.1007/978-3-030-52575-0_56.

BIOGRAPHIES OF AUTHORS

Rana Riad K. AL-Taie was born on January 02 1984. M.Sc., Information and Communication Engineering at Al Nahrain Uni. 2014, B. Sc., Computer Engineering dept. at Al-Mustansiriyah Uni. 2005. Academic staff member in Computer Engineering department at Al-Mustansiriyah University (www.uomustansiriyah.edu.iq). Networks, Web Application, Data Security, Robotic controller, Data mining and image processing. Email: Ranaal_taie@uomustansiriyah.edu.iq. Scopus Author ID: 57221643523. ORCID: 0000-0003-2570-3901.

Basma J. Saleh was born on Feb. 13, 1988. M.Sc., Electrical Engineering dept. at Baghdad Uni. 2016, B. Sc., Computer Engineering department at Al-Mustansiriyah Uni. 2010. Academic staff member in Computer Engineering department at Al-Mustansiriyah University (www.uomustansiriyah.edu.iq). Interested area: Artificial Neural Networks, intelligent algorithms, Optimization Methods, Robotic controller, Data mining and image processing. Email: eng.basmaj@uomustansiriyah.edu.iq Scopus Author ID: 57217582038. Web of Science: ResearcherID Y-6577-2019. ORCID: 0000-0002-0364-6005.

Ahmed Y. Saedi was born on Feb. 18, 1986. M.Sc., Ulyanovsk State Technical Uni. at Russia 2017, B. Sc., Computer Engineering department at Al-Mustansiriyah Uni. 2007. Academic staff member in Computer Engineering department at Al-Mustansiriyah University (www.uomustansiriyah.edu.iq). Interested area: Artificial Neural Networks, intelligent algorithms, Optimization Methods, Robotic controller, Data mining and image processing. Email: ahmed.yousif@uomustansiriyah.edu.iq. Scopus Author ID: 57204188050. ORCID: 0000-0003-2049-8371.

Lamees A. Salman was born in Des. 24, 1985. B. Sc. Uni. Of Technology, Iraq 2007. Academic staff member in Computer Engineering department at Al-Mustansiriyah University. Interested area: Artificial Neural Networks, intelligent Optimization algorithms. Email: lameesiteng@uomustansiriyah.edu.iq. Scopus Author ID: 57221395951. ORCID: 0000-0001-7904-3334.