Prediction of Heart Diseases (PHDs) based on Multi-Classifiers

Amirah Al Shammari[1]
Department of Computer Science
College of Computer, Al jouf University
Al jouf, Saudi Arabia

Haneen al Hadeaf[2], Hedia Zardi[3]
Department of Computer Science
College of Computer, Qassim University
Buraydah, Saudi Arabia

***Abstract*—At present, the number of articles on Heart Disease Detection (HDD) based on classification searched by Google Scholar search engine exceeds 17,000. The medical sector is one of the most important fields that benefit from ML. Heart diseases (HDs) are considered to be the leading cause of death worldwide, as it is difficult for doctors to predict them earlier. Therefore, the HDD is highly required. Today, the health sector contains huge data that has hidden information where this information can be considered as essential to make diagnostic decisions. In this paper, a new diagnostic model for the detection of HDs is on a multi-classifier applied to the heart disease dataset, which consists of 270 instances and 13 attributes. Our multi-classifier is composed of Artificial Neural Network (ANN), Naïve Bays (NB), J48, and REPTree classifiers, which select the most accurate of them. In addition, the most effective feature on prediction is determined by applying feature selection using the "GainRatioAttributeEval" technique and "Ranker" method based on the full tainting set. Experimental results show that the NB classifier is the best, and our model yields over 85% accuracy using the WEKA tool.**

***Keywords*—Classification; diseases; heart-attack; multi-classifier; heart disease detection**

I. Introduction

Pumping blood to the whole body is the most critical task in human bodies. Therefore, the heart is the most important organ for humans. All of the Heart Diseases (HDs) concern to categorize this kind of cardiovascular diseases like coronary heart disease, Angina pectoris, coronary heart collapse, Cardiomyopathy, coronary cardiovascular illness, Arrhythmias, along with Myocarditis [1]. HDs are still the main cause of death worldwide. HDs are the leading cause of death in the United Kingdom, the United States, Canada, and Australia. According to the Centers for Disease Control (CDC), about 610,000 people die of HDs in the United States every year. It is estimated that 25% of deaths in the United States occur as a result of HDs. The possibility of detection at an early stage will help prevent the attacks. HD is defined as a variety of diseases, conditions, and disorders that affect the heart and the blood vessels. People die having experienced symptoms that were not taken into consideration. In addition, the quality of services in health care centers implies diagnosing disease correctly and delivers effective handlings for patients. On the other hand, poor diagnosis can lead to disastrous consequences, which are unacceptable.

Moreover, there is a need for medical practitioners to predict HDs before they occur in their patients. In these cases, many studies have been done on predicting heart disease by applying different data mining techniques to predict the accuracy of heart disease from related data sets. So, this study is presented.

Today, most health organizations around the world exploit information systems to manage their healthcare, including data about patients. Usually, these systems store significant amounts of data (numbers, text, charts, and images), especially about patients. This data is a golden (vital) resource to support clinical decision making because it is a rich source of hidden information that is largely unexploited. This is a great reason that motivates researchers to generate datasets from these unused data. Then, the data can be analyzed by employing a variety of data mining techniques in order to detect many diseases, especially HDs.

Knowledge Discovery in Databases (KDD) is the process of determining useful knowledge from a collection of data [2] using data mining and Machine Learning (ML) techniques. KDD includes data integration, data cleansing, data selection, and incorporating prior knowledge on datasets and interpreting accurate solutions from the observed results [3]. Classification is one of the most popular data mining tasks that assigns new, unknown items in a collection to target predefined categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks [4].

In this paper, a model for the prediction of HDs (PHDs) using multi-classifiers is proposed to detect the existence of HD (sick or normal), as it is shown in section IV. The research question of this paper is: "Which is the most efficient technique for the prediction of Heart diseases by considering the factors of accuracy and speed?"

The remainder of this paper is organized as follows. Section II reviews the background of the used ML techniques (J48, Naïve Bays, ANNs, and REPTree). Many HDD based on ML is discussed in Section III, followed by a full description of the proposed PHDS model in Section IV. Next, in Section V, the experiments and results are discussed. Finally, conclusions and suggested future work are given in Section VI, and VII, respectively.

II. BACKGROUND

PHDs model differentiates between four ML techniques to choose the most accurate of them by applying them using the WEKA tool. In this section, a simple background is introduced WEKA and about each of the used techniques.

WEKA is used as a platform for machine learning as it has a collection of artificial intelligence algorithms in a domain for data mining tasks. It includes specific tools for data preprocessing and preparation, classification, clustering, regression, association rules, and visualization. WEKA gives us the ability to build models in order to detect hidden patterns in data and make a prediction without human interruption. Moreover, it contains a collection of predefined methods to evaluate the results of the techniques.

A. Artificial Neural Network (ANN)

It consists of an interconnected group of artificial neurons. ANN processes information using a connectionist approach. In most cases, an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Modern neural networks are usually used to model complex relationships between inputs and outputs or to find patterns in data [5]. ANN has the main three key advantages that make it more appropriate for ML problems: It can learn and model non-linear, complicated relationships, generalize, and does not enforce any restrictions on the distribution of input variables.

B. Naïve Base (NB)

A Naive Bayes (NB) classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. The classifier assumes that the presence (or absence) of a particular feature of a class (attribute) is unrelated to the presence (or absence) of any other feature. Even if these features depend on each other or upon the existence of the other features, the classifier considers all of these properties to contribute to that probability independently. The NB classifier performs reasonably well, even if the underlying assumption is not true [6]. NB is selected in this paper based on many advantages points as it requires less training data, fast to predict the class of the test data set, performs well in the multi-class prediction, and handle either continuous or discrete data.

C. J48 Decision Tree

It is the implementation of algorithm Iterative Dichotomiser 3 (ID3). J48 is developed by the WEKA project team. J48 classifier is a straightforward C4.5 decision tree for classification, which creates a binary tree. It is the most useful decision tree approach for classification problems [7]. This technique constructs a tree to model the classification process.

In general, decision tree algorithms are [8] robust to errors, handle missing values by observing the data into other attributes, and generate understandable rules. Also, the learning and classification processes are uncomplicated and quick, with accuracy superior to the others.

D. Reduced Error Pruning Tree (REPTree)

REPTree algorithm uses the regression tree logic. It generates multiple trees in different iterations. Afterward, it chooses the best of them as the representative tree [8]. In pruning the tree, it uses the mean square error on the predictions made by the tree. Fundamentally, REPTree is a fast decision tree learner, which builds a decision/regression tree using information gain as the splitting criterion and prunes it using reduced error. This ML can be effectively exploited in experimental comparisons to find the smallest optimally pruned tree with respect to the test set. Additionally, the property of this method is its linear computational complexity since each node is visited only once to evaluate the opportunity of pruning it.

III. RELATED WORK

Before research on HDD focused their efforts on applying different data mining techniques, many data mining techniques for the diagnosis of heart disease were implemented following different approaches, such as Decision Tree, NB, ANNs, which give different levels of accuracies [9].

Patel et al. [9] reported the results of the comparison between three different algorithms based on the decision tree looking for the best performance in HDD using WEKA. The tested algorithms were the J48 algorithm, Logistic model tree algorithm, and Random Forest algorithm. They concluded, after experiments, that the winning algorithm for best performance was J48.

Sudhakar and Manimekalai [10] proposed a model that generates a class of data based on association rules from a training data set. Their model classifies the test data set into predefined class labels using the three different data mining classification techniques: ANNs, Decision Tree, and NB. Their overall objective was to study the different data mining techniques available for the prediction of HDs and to compare them in order to identify the best HDD prediction method.

A hybrid algorithm with the ANN (backpropagation) approach for HDs prediction was proposed by Dewan and Sharma [11]. This hybrid algorithm extracts unknown patterns and relations related to heart diseases from a past heart disease database record.

Masethe and Masethe [12] compared the performance of J48, Bayes Net, NB, Simple Cart, and REPTree in the prediction of possible HDs attacks to determine which model gives the highest percentage of correct prediction. They concluded that the most accurate classification techniques were J48 followed by REPTree and Simple Cart algorithms, while Bayes Net and NB algorithms had less accuracy rat.

Kim, Lee, and & Lee [13] proposed a predictive model for coronary heart disease (CHD) based on data collected for Disease Control and Prevention. This model incorporates fuzzy logic and CART-based rule induction to support the prediction of CHD. Rule induction was conducted to generate the rules. The fuzzy logic was used in the prediction model as an inference model. The experimental results showed that the accuracy and receiver operating characteristic curve values of the proposed systems were 69.51% and 0.594.

Krishnaiah, Narsimha, and Chandra [14] built a model to predict heart disease patients based on a fuzzy approach. In this model, the diagnosis was based on historical data. To remove the uncertainty of the data, the Fuzzy K-NN classifier employed, and the results showed the capability to remove the redundancy of the data and the better accuracy of the system.

Choi, Schuetz, Stewart, and Sun [15] explored whether the use of deep learning to model temporal relations between events in electronic health records would improve model performance in predicting the initial diagnosis of heart failure compared to conventional methods that ignore temporality. They used Recurrent neural network models with gated recurrent units to detect relations among time-stamped events. Based on the experiments, deep learning models appear to improve the performance of models for the detection of incident heart failure.

Comparison between the performance of ANN, NB, J48, and REPTree classifiers to the best of our knowledge has not been reported. In this study, we investigate and report such a comparison with our PHDs model, as explained in the following sections.

IV. PHDs Model

The PHDs model is a classification model based on supervised learning of classifiers and testing. Four classifiers in PHDs learned to select the most accurate of them based on our dataset. After training, the most accurate classifier is considered as the classifier of the PHDs model. After that, the considered classifier is used to detect any new unknown instance. The PHDs model is composed of two stages containing four steps.

A. The Learning Stage is Composed of Three Steps

In this stage, the model is built based on the learning of four ML classifiers using one dataset with known instants "classified instants" to choose the most accurate among them.

B. The Classification Stage is Composed of a Single Compound Step (Last Step)

In this stage, any new unknown instant can be prepared and then classified to normal or sick using the most accurate classifier.

These two stages are consisting of four steps divided into two stages, as is shown in Fig. 1 and as discussed follows:

1) First stage: Learning and selecting the classifier:

This stage is composed of three steps:

a) Preprocessing and Preparation: As the quality of data is a key issue with data mining, so data preprocessing and preparation is a required step for serious, effective data mining. The results of data mining tasks as classification are affected by the quality of the dataset. So, in order to increase the accuracy of the mining, data preprocessing has to be performed. This stage includes handling missing values by using the average of attribute values from the same class. It should be noted that there are no noise data or inconsistencies. The other important tasks of preprocessing data, such as normalizing the attributes before conducting the ANN technique. In addition, the class attribute is transformed into binomial. The selected dataset prepared as follows:

- Convert data text format into comma .csv format for WEKA.
- Name columns (attribute values) by title.
- Choose the class attribute and convert it into binomial some class values were digits which not allowed for ANN classifiers.
- Open *.csv file from WEKA.
- Select class attribute as a label with importing wizard (important for classifications).
- Handle missing values using the average of instants of each class.
- Normalize the input data (1,0, -1) to process by ANN classifier.

b) Classifiers' Learning and Testing: In this step, each of the four classifiers learned and tested using dataset. The accuracy results of each classifier in recorded to be used later in the next step.

2) Classifier selection: In this step, the most accurate classifier is selected as the classifier of the model. The selected classifier is used to detect HDs for any new instance.

3) Second stage: Classification: This stage is composed of a single compound step. In this stage, any new unknown instant will be prepared and then classified to either normal or sick using the most accurate classifier selected from the *first stage.*

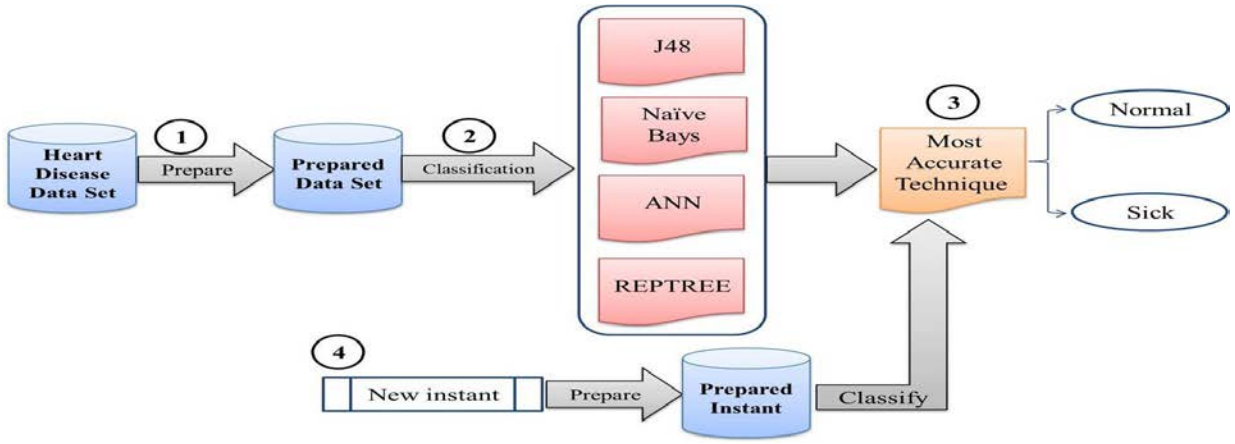

Fig. 1. PHDs Model.

V. Experiments and Discussion

Experiments were conducted with the WEKA tool using the dataset of heart disease from the UCI Machine Learning Repository [16]. WEKA is chosen to conduct experiments for many reasons, which are: it has many visualization tools and algorithms for data analysis and predictive modeling, which give easy access and use. WEKA supports graphical user interfaces either for process data or to illustrate the results, which makes it easy to understand. Also, it has an extensive collection of data preprocessing and modeling techniques.

On the other hand, the heart disease dataset preferred because this dataset collected and designed for a classification task, has a suitable number of records, and Only 3 of its attributes have missing value. The data set specifications listed in Table I.

This dataset is taken from 270 individuals; the diagnosis of some of them was definite for having heart disease, has 14 attributes. The last attribute is a special one for "class", either the presence or the absence of HDD. These attributes are represented in Table II. The purpose of analyzing the dataset was to detect for the presence of HDs (normal is none and sick is present).

The four algorithms applied to the data set using the percentage of data for learning and the remainder for testing in order to assess the performance of the classification technique for predicting a class. Many experiments with variants of parameters for training and testing data and evaluation options (percentage split, cross-validation) conducted. The best results are shown.

A. Percentage Split

The experiment's results of NB and ANN algorithms with test mode: 80.0% training and 20.0% testing are illustrated in Tables III and IV, respectively. While results of the REPTree algorithm with test mode: 85.0% training and 15.0% testing are shown in Table V. Finally, results of the J48 algorithm with test mode: 90.0% training and 10.0% testing are represented in Tables VI. All the accuracy and time results are illustrated in Fig. 2 and Fig. 3, respectively.

Using the percentage split, ANNs then J48 algorithm are achieved higher accuracy while NB then REPTree algorithm is less. So overall confusion matrices and Fig. 2 and 3, it is concluded that ANN is the most accurate, and the J48 is the fastest algorithm.

TABLE I. Dataset Specifications

Data Set Characteristics	Multivariate
Attribute Characteristics	Categorical, Integer, Real
Associated Tasks	Classification
Number of Instances	270
Number of Attributes	13
Missing Values?	Yes

TABLE II. Dataset Attributes

Symbol	Attribute
A	Age
B	Sex
C	Chest pain type (4 values)
D	Resting blood pressure
E	Serum cholesterols in mg/dl
F	Fasting blood sugar > 120 mg/dl
G	Resting electrocardiographic results (values 0,1,2)
H	Maximum heart rate achieved
I	Exercise induced angina
J	Old peak = ST depression
K	The slope of the peak exercise ST segment
L	Number of major vessels (0-3) colored by fluoroscopy
M	Thallium:3=normal; 6=fixed defect; 7=reversible defect
Class	Class, Absence (Normal) or presence (Sick)

TABLE III. Confusion Matrix of NB Algorithm

	Predicted Sick	**Predicted Normal**
Actual Sick	TP (22)	FP (4)
Actual Normal	FN (4)	TP (24)

TABLE IV. CONFUSION MATRIX OF ANN ALGORITHM

	Predicted Sick	Predicted Normal
Actual Sick	TP (24)	FP (2)
Actual Normal	FN (3)	TP (19)

TABLE V. CONFUSION MATRIX OF REPTREE ALGORITHM

	Predicted Sick	Predicted Normal
Actual Sick	TP (14)	FP (4)
Actual Normal	FN (4)	TP (24)

TABLE VI. CONFUSION MATRIX OF J48 ALGORITHM

	Predicted Sick	Predicted Normal
Actual Sick	TP (8)	FP (4)
Actual Normal	FN (0)	TN (15)

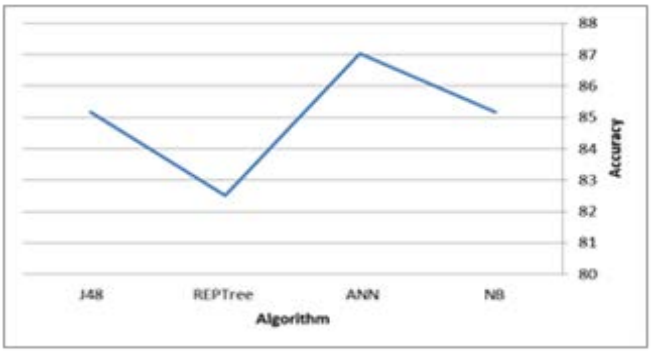

Fig. 2. Accuracy of Models using Percentage Split.

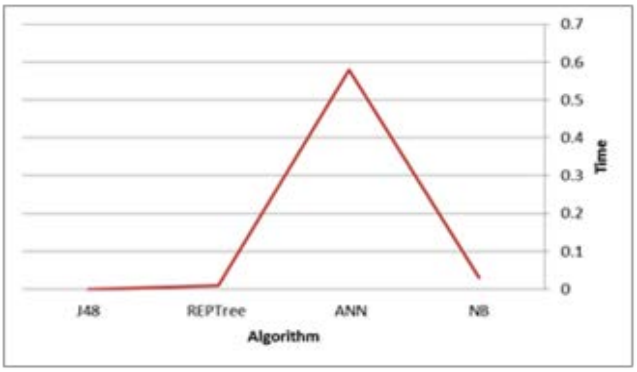

Fig. 3. Conducting Time using Percentage Split (by Seconds) for the Best Result of each Algorithm.

B. Cross-Validation

The experiment's result of the NB algorithm with 15 folds shown in Tables VII. On the other hand, the ANN and J48 algorithms with test mode: 21 folds are illustrated in Tables VIII and X, respectively. Finally, the results of the REPTree algorithm with 24 folds presented in Tables IX.

All the accuracy and time results based on cross-validation are illustrated in Fig. 4 and Fig. 5, respectively. Using cross-validation, NB, then the REPTree algorithm becomes the most accurate while ANN then J48 algorithm is less.

TABLE VII. CONFUSION MATRIX OF NB ALGORITHM

	Predicted Sick	Predicted Normal
Actual Sick	TP (97)	FP (23)
Actual Normal	FN (20)	TN (130)

TABLE VIII. CONFUSION MATRIX OF ANN ALGORITHM

	Predicted Sick	Predicted Normal
Actual Sick	TP (92)	FP (28)
Actual Normal	FN (24)	TP (126)

TABLE IX. CONFUSION MATRIX OF REPTREE ALGORITHM

	Predicted Sick	Predicted Normal
Actual Sick	TP (89)	FP (31)
Actual Normal	FN (18)	TP (132)

TABLE X. CONFUSION MATRIX OF J48 ALGORITHM

	Predicted Sick	Predicted Normal
Actual Sick	TP (93)	FP (27)
Actual Normal	FN (29)	TP (121)

TABLE XI. ACCURACY OF ALGORITHMS

Algorithm	Percentage-Split	Cross-Validation
NB	85.18	84.07
ANN	89.58	81.65
REPTree	80.60	81.55
J48	85.182	79.26

So, we conclude that the NB algorithm is the best one due to its high-performance using a cross-validation method. In addition, it is the fastest.

The comparison between the accuracies of all algorithms based on percentage-split and cross-validation is shown in Table XI and Fig. 6. So, we conclude that the NB algorithm is the best one due to its high-performance using a cross-validation method. Also, it is the fastest.

Predefined instances of the heart disease dataset, 18 new unlabeled instances are supposed with random values to apply the second stage of the PHDs model. After inputting them to PHDs, they are labeled, as shown in Fig. 7.

Like Comparing our results with other research works, our results are inconsistent with them, that is in [9] and [12], J48 is the most accurate, while in [13] and [15] is the ANN. In our opinion, this inconsistency due to many factors as a dataset and an applied tool. Moreover, none of the research work applied the algorithms together on the same dataset and using the same tool. We believe that these results can give a chance to develop a new effective diagnostic tool to help doctors and HDs patients.

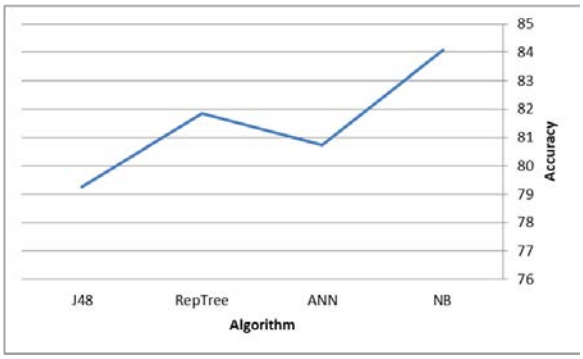

Fig. 4. Accuracy of Models using Cross-Validation.

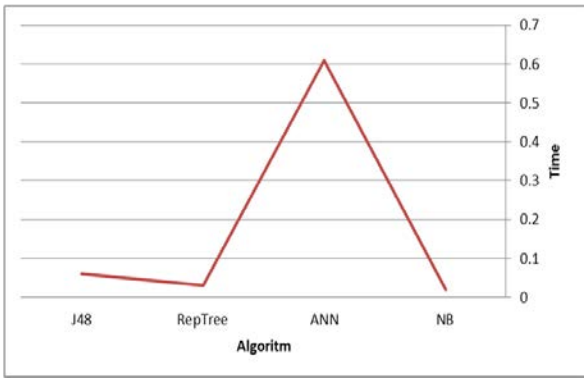

Fig. 5. Conducting Time using Cross-Validation (by Seconds) for the Best Result of each Algorithm.

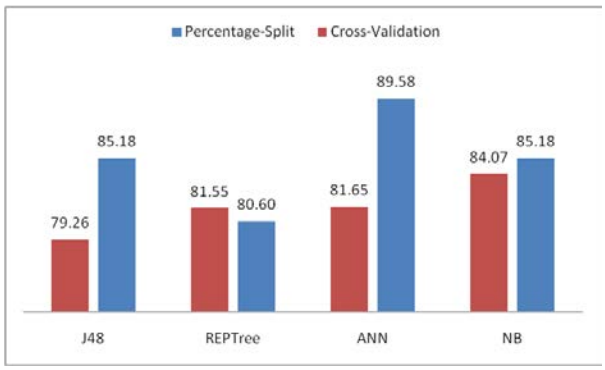

Fig. 6. Comparison of the Accuracy of the Algorithm.

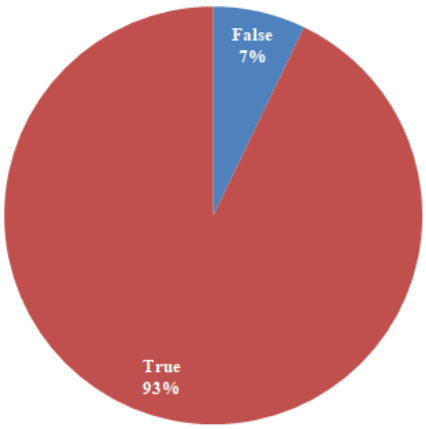

Fig. 7. Predict of New Instances using NB.

```
Ranked attributes:
 0.2054  13 M
 0.1906   3 C
 0.1701  12 L
 0.148   10 J
 0.1421   9 I
 0.1231   8 H
 0.11    11 K
 0.0738   2 B
 0.0567   1 A
 0.0241   7 G
 0        5 E
 0        4 D
 0        6 F

Selected attributes: 13,3,12,10,9,8,11,2,1,7,5,4,6 : 13
```

Fig. 8. Results of Feature Selection Technique.

To determine the most effective feature on the prediction, feature selection is applied using the "GainRatioAttributeEval" technique and "Ranker" method based on the full tainting set. As it is shown in Fig. 8, all the (13) attributes have effects on the prediction, but the most three attributes in order are (M,C,L) ,which are: Thalassemia, Chest pain type, Number of major vessels; respectively.

VI. Conclusion

This paper intends to present a new diagnostic model for the detection of the HDs using the most efficient classifier based on accuracy and time using the WEKA tool. Experiments on PHDs model are conducted using Artificial Neural Network (ANN), Naïve Bays (NB), J48, and REPTree classifiers and are tested based on percentage split and cross-validation. The overall results of experimental results show that the NB classifier is the best, and our model yields over 85% accuracy using it based on the WEKA tool. It can conclude from the analysis of the experimental results that the NB technique turned out to be the most accurate classifier for the HDD. Also, results showed that this technique was the fastest of all. These results are incompatible with many previous works, which conclude that the NB did not give the best accuracy. These results are significant in the field of detecting the HDs also to decrease the reasons for death. Moreover, the results show that the three most effective attributes in order are: Thalassemia, Chest pain type, Number of major vessels.

VII. Future Work

In the future, more experiments using many variants of data sets can be conducted to prove the current results or explore new conclusions. In addition, Text mining can be employed to predict and diagnose the HDs.

References

[1] Purusothaman, G., & Krishnakumari, P. (2015). A survey of data mining techniques on risk prediction: Heart disease. Indian Journal of Science and Technology, 8(12), 1.

[2] Ho, T. B. (2016). Knowledge Discovery. In Knowledge Science, pp. 70-93. CRC Press.

[3] Fayyad, U., Piatetsky-Shapiro, G., and Smyth P. “From data mining to knowledge discovery in databases”, AI magazine, 17 (3), (1996): pp. 37-54.

[4] Bharathi, A., and E. Deepankumar. “Survey on classification techniques in data mining”, International Journal on Recent and Innovation Trends in Computing and Communication 2.7 (2014): 1983-1986.

[5] Gurney, K., An introduction to neural networks. CRC press, 2014.

[6] Lewis, David D. "Naïve (Bayes) at forty: The independence assumption in information retrieval", European conference on machine learning. Springer, Berlin, Heidelberg, 1998.

[7] Ramesh, D., Pasha, S. N. and Roopa, G. (2017). "A Comparative Analysis of Classification Algorithms on Weather Dataset Using Data Mining Tool", Oriental Journal of Computer Science and Technology, 10 (4).

[8] Kalmegh, S. (2015). "Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and Random Tree for Classification of Indian News". International Journal of Innovative Science, Engineering & Technology (IJISET), 2(2), pp. 438-446.

[9] Patel, J., Tejal Upadhyay, D. and Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. Heart Disease, 7(1), 129-137.

[10] Sudhakar, K. and Manimekalai, D. M. (2014). Study of heart disease prediction using data mining. International journal of advanced research in computer science and software engineering, 4(1).

[11] Dewan, A., and Sharma, M. (2015, March). Prediction of heart disease using a hybrid technique in data mining classification. In 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 704-706). IEEE.

[12] Masethe, H. D. and Masethe, M. A. (2014, October). Prediction of heart disease using classification algorithms. In Proceedings of the world Congress on Engineering and computer Science (Vol. 2, pp. 22-24).

[13] Kim, J., Lee, J., & Lee, Y. (2015). Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree. Healthcare informatics research, 21(3), 167-174.

[14] Krishnaiah, V., Narsimha, G., & Chandra, N. S. (2015). Heart disease prediction system using data mining technique by fuzzy K-NN approach. In Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) Volume 1 (pp. 371-384). Springer, Cham. [16] Blake, C. L. and Merz, C. J. UCI Repository of machine learning databases. Irvine, CA, 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html

[15] Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association, 24(2), 361-370.

[16] Blake, C. L. and Merz, C. J. UCI Repository of machine learning databases. Irvine, CA, 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html

AUTHORS' PROFILE

Amirah Al Shammari is a Teaching Assistant in the Computer Science department at Aljouf University, Saudi Arabia and is currently completing my Master's degree in the Informatics department at the University of Qassim, Buraydah, Saudi Arabia.

Haneen Al Hadeaf is a Teaching Assistant in the Computer Science department at Qassim University, Saudi Arabia. Currently, her studying a master's degree in the Informatics department at Qassim University, Buraydah, Saudi Arabia. She received her B.Sc. in Computer Science from Imam Mohammad Ibn Saud Islamic University in Riyadh, Saudi Arabia. And her area of research in computer vision, machine learning, and software engineering.

Hedia Zardi an assistant professor in the computer science department at Qassim University, Saudi Arabia. She received the PhD degree in 2016 from Sorbone University, Paris, France and Manouba University, Tunisia.