A Clinical support system for Prediction of Heart Disease using Machine Learning Techniques

Halima EL HAMDAOUI[1, 2], Saïd BOUJRAF[1], Nour El Houda CHAOUI[2], Mustapha MAAROUFI[1]

[1]Clinical Neuroscience Laboratory, Dept. Biophysics & Clinical MRI Methods, Faculty of Medicine,
University of Sidi Mohamed Ben Abdellah, Fez, Morocco
[2]Transmission and Treatment of Information Laboratory, University of Sidi Mohamed Ben Abdellah
Fez, Morocco

halima.elhamdaoui@usmba.ac.ma

Abstract— **Heart disease is a leading cause of death worldwide. However, it remains difficult for clinicians to predict heart disease as it is a complex and costly task. Hence, we proposed a clinical support system for predicting heart disease to help clinicians with diagnostic and make better decisions. Machine learning algorithms such as Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, Random Forest, and Decision Tree are applied in this study for predicting Heart Disease using risk factors data retrieved from medical files. Several experiments have been conducted to predict HD using the UCI data set, and the outcome reveals that Naïve Bayes outperforms using both cross-validation and train-test split techniques with an accuracy of 82.17%, 84.28%, respectively. The second conclusion is that the accuracy of all algorithm decrease after applying the cross-validation technique. Finally, we suggested multi validation techniques in prospectively collected data towards the approval of the proposed approach.**

Keywords—Heart disease, clinical decision systems, machine learning, UCI heart disease data set.

I. INTRODUCTION

According to the world health organization (WHO), cardiovascular diseases (CVDs) are the first cause of death worldwide, with more than 17.9 million people died in 2016[1]. CVDs are a group of syndromes affecting blood vessels and heart; they include heart disease (HD), which is often expressed as coronary heart disease [2]. However, HD can be prevented by observing a healthy lifestyle and avoiding risk factors. Thus, understanding what is contributing to these alarming factors may help for the prevention and prediction of HD. Typical, angiography is the primary diagnosis method; it is used to determine the localization of heart vessels' stenosis. Being costly, time-consuming, and invasive had motivated researchers to develop automatic systems based on information gathered through a set of medical data, such as data from past treatment outcomes as well as the latest medical research results and databases [3]. Nowadays, machine learning techniques are used to assist clinicians in making more accurate predictions of HD based on medical data; these data can be demographic, symptom and examination, ECG, and laboratory. Several studies were carried on diagnosing and predicting heart disease using ML techniques [4].

Most researches have used the UCI heart disease data set due to its availability [5]. This data set contains four sub data set and 76 attributes; the number of selected attributes and common features used in each study is ranging from 76 to 8, including the class attribute. Generally, the studies that used many attributes have applied feature selection to improve relevance [6, 7]. Hence, most studies perform only 14 attributes, including (Age, Gender, Chest pain, blood pressure ...) that are relevant for the risk factors of HD diagnosis values [8-11]. Various prediction models were built using well-known ML techniques. The author [8] suggested a predictive model using C4.5 and fast decision tree algorithms applied on the four collected and separated UCI data sets; this model achieved an accuracy of 78.06% and 75.48% for C4.5 and fast decision tree respectively using only Cleveland data set. The author [7] Combined Infinit Latent feature selection method with SVM classifier and achieved an accuracy of 89.93% using three data sets, including Cleveland, Hungarian, and Switzerland, with 58 attributes. The author [6] predicted HD using the meta-algorithm Adaboost on Cleveland data set and suggested reducing the number of attributes from 76 to 28 to provide higher accuracy of 80.14%. The author [12] used Alizadeh Sani data set to develop a hybrid method by enhancing the performance of Neural network using Genetic algorithm and yielded an accuracy of 93%. A comparative study using four different classifiers including SVM, KNN, C5.0, and Neural network, was approved by the author [9], he achieved a high accuracy of 93.02% by C5.0 algorithm using 14 of attributes but different data sets. Despite a substantial research output, no gold-standard model is available to predict HD. Hence, there is

978-1-7281-7513-3/20/$31.00 ©2020 IEEE

still a need for improvement. Also, many parameters impact the construction of the HD prediction model; these include the data set of choice, the number of attributes and the output class, and the algorithm used.

This paper aims to build a clinical decision support system allowing predicting the risk level of HD using UCI Cleveland data set. A classification model is proposed to detect patterns in existing HD patient's data. In the next section, our methodology is described with a brief detail of the data set used. Section 3 presents the experiments and the different representations of outcomes. Finally, conclusions are given in section 4.

II. PROPOSED METHOD

The proposed architecture of the model to predict the presence of HD is shown in Fig. 1. As soon as the preprocessing was carried out to handle missing values in the initial step of the process, machine learning algorithms are used to predict whether a patient has HD or not. Finally, the performance measures based on the confusion matrix considered to evaluate the algorithms.

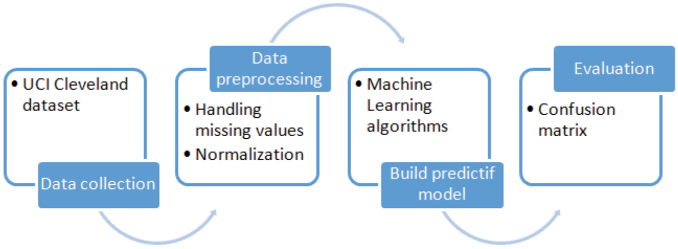

Fig. 1. Flow Chart of our model for prediction of HD

A. *Medical dataset*

In this paper, the Cleveland heart disease data set is retrieved from the UCI machine learning repository [5]. Although there are a total of 303 records of 76 attributes, we used only 14 most relevant attributes. The percentage of the presence of HD in the used data set is 45.54 %. The used 14 attributes with their description are shown in Table 1.

TABLE I. DESCRIPTION OF SELECTED ATTRIBUTES AND THEIR VALUES

Attributes	**Description**
Age	Age in years [29-77]
Sex	Male =1 , Female=0
Fbs	Fasting blood sugar >120 mg/dl, Value 1=yes, Value 0=no
Cp	Chest pain types Value 1= typical angina, Value 2= atypical angina Value 3= non-angina, Value 4= asymptomatic
Trestbps	Resting blood pressure in mm Hg [94-200]
Chol	Serum cholesterol in mg/dl [126--564]
Restecg	Resting electrocardiographic , Value 0=normal, Value1= having ST-T wave abnormality Value2= left ventricular hypertrophy by Estes' criteria
Thalach	Maximum heart rate achieved, [71-202]
Exang	Exercise induced angina value 0= no, 1= yes
Oldpeak	Measure of ST depression induced by exercise relative to rest [0--6.2]
Slope	Measure of slope for peak exercise ST segment, Value 1= up sloping , Value 2= flat, Value 3= down sloping
Ca	Number of major vessels colored by flourosopy [0-3]
Thal	Thallium stress test, Value 3= normal, Value 6= fixed defect, Value 7= reversible defect
Num	Value 1 = presence of HD Value 0= absence of HD

B. *Data preprocessing*

Data preparation is the most critical first step in any predictive model; it helps to transform data into an understandable format to enhance model efficiency. Medical data are generally incomplete, lacking attribute values, and noisy since containing outliers or irrelevant data [13].

The UCI Cleveland data set used in this study contains six missing values, including four missing values for the number of major vessels (Ca) attribute and two missing values for Heart rate (Thal) attribute. The handle these missing values, we used the "Mode" imputation method that replaced missing values by the most frequently occurring value since all missing values are categorical [14].

The predicted attribute (num) of the original data set contained five values; a value 0 indicated the absence of HD and values between 1 and 4 reported different levels of HD, respectively. In this study, we are interested in the presence or absence of HD without interest in the exact disease classification. Hence, the class attribute is reclassified into a binary value of 0 or 1, indicating the absence or presence of HD in the patients, respectively.

C. *Data classification*

Predictive modeling is an approach to building a model able to make predictions. This model includes a machine learning algorithm that enables data-driven models to learn specific information from observed data in a training data set to make those predictions [4]. In this study, we have a particular target used to predict output for new use cases, which determines either heart disease is present or not. Hence a supervised learning classification algorithm would be ultimate to train the data. The following ML algorithms (NB, KNN, SVM, RF, and DT), are used to build our proposed model.

Naïve Bayes (NB) is a probabilistic classifier based on applying Bayes' Law and naïve conditional independence assumptions. In other words, Naïve Bayes Classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Hence, the status of a specific feature in a class does not affect the status of another feature. Since it is based on conditional probability, it is considered as a robust algorithm employed for classification purposes [15].

K-Nearest Neighbor (KNN), a supervised learning model as well, is used to classify the test data using the training samples directly. It. is a method for classifying objects based on closest training data in the feature space. Otherwise, the class of a new sample is predicted based on some distance metrics where the distance metric can be a simple Euclidean distance. In the practical steps, KNN first calculates k (Number of the nearest neighbors), and it finds the distance between the training data and then sorts the distance. Subsequently, a class label will be assigned to the test data based on the majority voting [16].

Random Forest (RF) refers to ensembles of simple tree predictors, and each tree produce and outcome. For the regression model, the tree outcome is an estimate of the dependent value given the predictors. On the other hand, for the classification model, the tree outcome takes the form of a class membership that classifies a set of independent values given the predictors with one of the categories present in the dependent value [18].

Support Vector Machine (SVM) is one of the standard set of supervised learning model employed in classification. It is defined as the finite-dimensional vector spaces where each dimension characterizes a feature of a particular sample. A support vector machine aims to find the best highest-margin separating hyperplane between the two classes. Due to its computational competence on big data sets SVM has been proved as an effective method in high-dimensional space problems. [17].

Decision Tree (DT), currently, is one of the most potent and popular classifications and prediction algorithms used in machine learning. It has been widely used to examine data and make decisions by many researchers as classifiers in the healthcare domain. DT creates a model that predicts the value of a target variable by learning simple decision rules inferred from the data features and splitting data into branch-like segments. The input values can be continuous or discrete. The leaf nodes return class labels or probability scores. In theory, the tree can be converted into decision rules. These rules of classification can easily be presented visually [19, 20].

D. Performance measurement

In order to evaluate the validity of the predictive model, various measurements can be calculated suchlike sensitivity, specificity, accuracy, and precision, by using the confusion matrix (Table 2). Specificity measures the proportion of negatives which are correctly identified, and sensitivity measures the percentage of real positives that are correctly identified [21]. These measures can be mathematically represented by the following formulas. Where TP, TN, FP, and FN signify True Positive (number of positive data that were correctly labeled by the classifier), True Negative (number of negative data that were correctly labeled by the classifier), False Positive (number of negative data that were incorrectly labeled as positive), and False Negative (number of positive data that were mislabeled as negative), respectively.

TABLE II. CONFUSION MATRIX

		Actual values	
		positive	negative
Predicted values	positive	TP	FP
	negative	FN	TN

$$Specificity = \frac{TN}{TN + FP} \qquad (1)$$

$$Sensitivity = \frac{TP}{TP + FN} \qquad (2)$$

$$Precision = \frac{TP}{FP + TP} \qquad (3)$$

$$Accuracy = \frac{(TN + TP)}{(TN + FP + FN + TP)} \qquad (4)$$

III. RESULTS

Different experiments are conducted to evaluate the performance and the validation of the developed model using the Cleveland training data set with 14 different attributes. Results are estimated using confusion matrix measurements and to compare the accuracy using different algorithms.

For the first experiment, we used the train-test split validation technique where our data-set is divided into two parts, and we made several tests with different percentages, the best splitting we achieved is the 70% of the data for training and 30% for testing. Fig. 2 shows the results obtained by applying NB, KNN, SVM, RF, and DT algorithms. Based on the experimental results shown Fig. 2, it is clear that the classification accuracy of the NB algorithm is the highest, followed by DT compared to other algorithms.

However, the train-test split validation technique usually causes overfitting since the evaluation may depend mainly on which data is used in the training set and which is used in the test set. Hence, the evaluation may be significantly different depending on how the split is made. Thus, we proposed the cross-validation technique to handle this problem. The cross-validation is a robust preventative measure against overfitting, it uses the initial training data to generate multiple mini train-test splits to tune the model. In our second experiment we used 10-fold cross-validation, where the original dataset is splited into 10 same size subsamples, and the accuracy is averaged over all 10 trials to get the total effectiveness of our model. As reflected, each data point gets to be in a test set once and set k-1 times in the training. This significantly reduces bias and variance, since we used most of the data for fitting and in the test set.

Table 3 gives the accuracy obtained using a 10-fold cross-validation technique. It can be observed from the table that the NB still worked better, building the model with an accuracy of 82.17%, and SVM was second with an accuracy of 79.20%.

We can conclude that NB performs better and gets better accuracy compared to other algorithms. Also, all the results' accuracy is decreased after using the cross-validation technique.

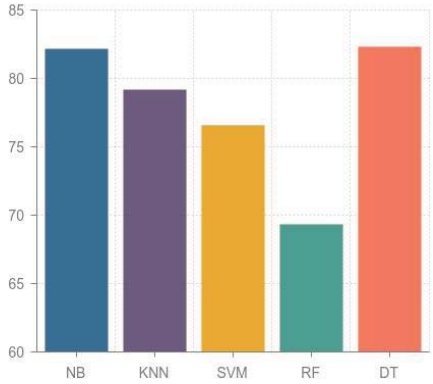

Fig. 2. Comparison of the accuracy of ML algorithms using train/test split technique

TABLE III. COMPARISON OF PERFORMANCE OF ML ALGORITHMS USING BOTH SPLIT AND CROSS-VALIDATION TECHNIQUES

ML algorithms	Accuracy (CrossValidation) %	Accuracy (split data)%
NB	82.17	84.28
KNN	76.56	81.31
SVM	79.20	81.42
RF	69.30	77.14
DT	75.57	82.28

IV. CONCLUSION

This study has been conducted to help clinicians to produce an accurate and efficient predictive system; the model validation is conducted with both cross-validation and the train-test split of data. The results showed that NB achieved the highest accuracy compared to other algorithms using both validation techniques. Despite the accuracy is decreased when we applied the cross-validation, we believe that this technique is the best in our model since the used data-set is not large so the process didn't take a long time, at the same time we solved the problem of overfitting.

The results were significant, and we believe that the achieved results using our predictive model based on ML algorithms could improve the knowledge on the prediction of heart disease risk through better diagnoses and interpretation; therefore, appropriate clinical decisions.

REFERENCES

[1] URL:http://who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] URL: http://nhlbi.nih.gov. National heart, lung, and blood institute.

[3] N. Mishra and S. Silakari "Predictive Analytics: A Survey, Trends, Application, Opportunities and Challenges," International Journal of computer science and information technologies, vol 3(3), pp. 4434-4438, 2012.

[4] H. Alharti. "Healthcare predictive analytics: An overview with a focus on Saudi Arabia," Journal of Infection and Public Health, vol 11(6), pp. 749-756, 2018.

[5] R. Detran Heart Disease Dataset. "Retrieved from: http://archive.ics.edu/ml/machine-learning-databases/heart-disease/cleveland.data" 1988.

[6] K. H., Miao, J. H. Miao & G. Miao. "Diagnosing Coronary Heart Disease Using Ensemble Machine Learning," International Journal of Advanced Computer Science and Applications,vol 7(10), 2016.

[7] L. M. Hung, D. T. Toan, & V. T. Lang. "Automatic Heart Disease Prediction Using Feature Selection and Data Mining Technique,". Journal of Computer Science and Cybernetics,vol 34(1), pp. 33-47, 2018.

[8] R. El-Bialy, M. A. Salamay, O. H.Karam, & M.E. Khalifa. "Feature Analysis of Coronary Artery Heart Disease Data Sets". International Conference on Communication, Management and Information Technology. Procedia Computer Science, vol 65, pp. 459-468, 2015.

[9] M. Abdar, R.Sharareh, N. Kalhori, T. Sutikno, I.M. I. Subroto & G . Arji. "Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases," International Journal of Electrical and Computer Engineering, vol 5(6), pp. 1569-1576, 2015.

[10] A. K. Paul, P. C. Shill, R. I. Rabin, & M. A. H. Akhand. "Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease,". IEEE 5th International Conference on Informatics, Electronics and Vision. 2016.

[11] Purushottam, K.Saxena, & R.Sharma (2016). Efficient Heart Disease Prediction System. Procedia Computer Science, 85, 962-969.

[12] Z. Arabasadi, R. Alizadehsani , M. Roshanzamir , H. Moosaei , & A. A. Yarifard. "Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm," Computer Methods and Programs in Biomedicine, vol 41, pp. 19-26, 2017.

[13] H. M. Zolbanin, D. Delen, & A. H. Zadeh. "Predicting overall survivability in comorbidity of cancers: A data mining approach," Decision Support Systems, vol 74, pp. 150-161, 2015.

[14] Z. Zhang. "Missing data imputation: focusing on single imputation," Ann Transl Med, vol 4(1), pp. 9, 2016.

[15] I. Kononenko; " Inductive and Bayesian learning in medical diagnosis," Applied Artificial Intelligence, vol 7(4), pp. 317-337, 1993.

[16] N. S. Altman. “An introduction to kernel and nearest-neighbor nonparametric regression,” The American Statistician, vol 46(3), pp. 175–185, 1992.

[17] C. Cortes & V. Vapnik. ”Support-vector networks,” Machine Learning, vol 20(3), pp. 273–297, 1995.

[18] T.K. Ho. “Random Decision Forests“ Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, pp. 278–282, 1995.

[19] L. Breiman, J. H. Friedman, R. A. Olshen & C. J. Stone “Classification and Regression Trees,” Chapman & Hall/CRC, 1984.

[20] J. R. Quinlan. “Induction of decision trees”. Machine Learning, vol 1(1), pp. 81-106, 1986.

[21] S. V Stehman. “Selecting and interpreting measures of thematic classification accuracy,” Remote Sensing of Environment, vol 62(1), 77-89, 1997.