Improving Heart Disease Prediction Using Random Forest and AdaBoost Algorithms

Halima El Hamdaoui(✉), Saïd Boujraf, Nour El Houda Chaoui, Badreddine Alami, Mustapha Maaroufi
University Sidi Mohamed Ben Abdellah, Fez, Morocco
`halima.elhamdaoui@usmba.ac.ma`

Abstract—heart disease is a major cause of death worldwide. Thus, diagnosis and prediction of heart disease remain mandatory. Clinical decision support systems based on machine learning techniques have become the primary tool to assist clinicians and contribute to the automated diagnosis. This paper aims to predict heart disease using Random Forest algorithm enhanced with the boosting algorithm AdaBoost. The model is trained and tested on University of California Irvine (UCI) Cleveland and Statlog heart disease datasets using the most relevant features 14 attributes. The result shows that Random Forest algorithm combined with AdaBoost algorithm achieved higher accuracy than applying only Radom Forest algorithm, 96.16%, 95.98%, respectively. We compare our suggested model to report machine learning classifiers. Indeed, the obtained result is supporting the efficiency and validity of our model. Besides, the proposed model achieved high accuracy compared to existing studies in the literature that confirmed that a clinical decision support system could be used to predict heart disease based on machine learning algorithms.

Keywords—heart disease, clinical decision systems, machine learning, Random Forest, AdaBoost algorithm, UCI heart disease dataset

1 Introduction

Technological innovations contribute to empowering, enriching, and significantly transforming the health work methods. Indeed, artificial intelligence and machine learning (ML) can analyze, improve diagnosis, predict, and help daily clinical practice [1]–[3]. Hence, healthcare industries are competing to produce machine learning for medical decision support systems for disease prediction. Indeed, a predictive analytic model is used to assist clinicians to make more accurate predictions based on the volume of information gathered through a clinical data; such as data from past treatment and medical research results [4]. These models can be prospectively installed within the clinical settings to investigate whenever patients risk developing diseases. Prediction models using ML have been suggested by various studies to diagnose different diseases such as lung cancer, Liver disease, breast cancer, obesity, Parkinson, Alzheimer, and cardiovascular diseases (CVDs)[5]–[12]. CVDs are the top cause of death

worldwide, with more than 17.9 million people died in 2016 according to the world health organization (WHO), and represented 31% of all global deaths. Accordingly, CVDs in Morocco represented 38% of total causes of deaths in 2016 (Figure 1)[13]. This high mortality rate has attracted significant attention during the last years to improve and automate CVD diagnosis, resulting in numerous approaches.
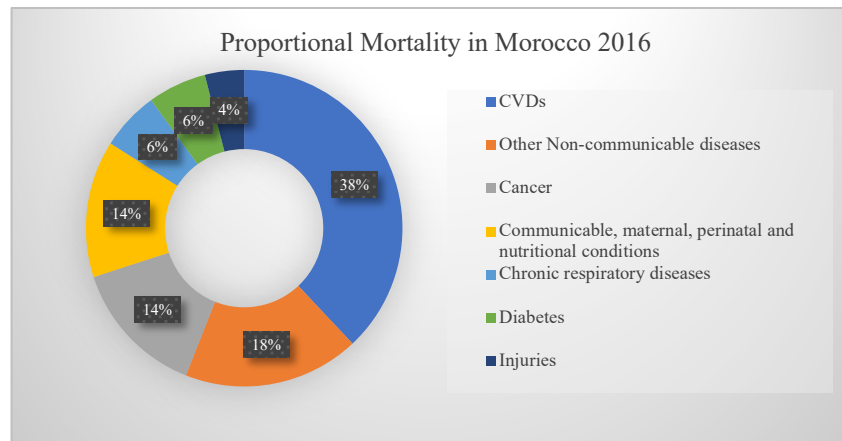

Fig. 1. Proportional diseases mortality in Morocco 2016

2 Clinical relevance

Cardiovascular disease refers to all kinds of diseases that affect the heart or blood vessels; one of these types is heart disease (HD). HD is an umbrella term for various conditions that affect the heart's structure and function. Whereas all heart diseases are cardiovascular diseases, but not all cardiovascular diseases are heart disease. The most familiar type of heart disease is coronary heart disease. It is often referred to simply as heart disease, yet not the only kind of heart disease. Coronary heart disease is mainly affecting the coronary arteries and weakening the heart muscle when plaque (a combination of cholesterol, fat, calcium) occurs in the walls of the artery. The plaque reduces the volume of oxygen-rich blood getting to the heart, which can cause chest pain and block blood flow, leading to the most common cause of a heart attack [14], [15]. Various risk factors expand heart disease incidence; some are uncontrollable risk factors, including sex, age, and heart disease history. Several risk factors can be controlled: high cholesterol, high blood pressure, obesity, physical inactivity, uncontrolled diabetes, stress, depression, anger, smoking, alcohol use, and low diet [16]. Over time, these risk factors might cause changes in the heart and blood vessels yielding heart attacks, heart failure, and stroke [17]. Thus, it is critical to consider risk factors in early life to prevent and predict HD, which remains challenging. In the current clinical routine, angiography is the standard gold technique used to diagnose heart disease. However, as a conventional invasive-based method, angiography has some downsides such as complications and risks (including the dangers of radiation). Also, it is a costly and time-consuming

assessment. To overcome the use of conventional invasive-based methods for the diagnosis of heart disease, researchers attempted to develop different non-invasive automated systems based on predictive machine learning techniques. Machine learning based approaches have been presented by numerous researchers to predict heart disease. Author [18] suggested a predictive model using C4.5 and fast decision tree algorithms applied on the four collected and separated UCI datasets; this model achieved an accuracy of 78.06% and 75.48% for C4.5 and fast decision tree, respectively, using only the Cleveland dataset. Author [19] predicted HD using the meta-algorithm AdaBoost on the Cleveland dataset and suggested reducing the number of attributes from 76 to 28 to provide higher accuracy of 80.14%. A comparative study using four different classifiers, including SVM, KNN, C5.0, and Neural network, was approved by Ref. [20] he achieved a high accuracy of 93.02% by C5.0 algorithm but only using Statlog dataset and validated with train-test split method. Author [21] gained 86.70% in accuracy using decision rules algorithm and cross validation technique.

Nevertheless, traditional ML algorithms present some limitations in term of accuracy improvement. Thus, several other researchers have employed different approaches including the hybridization and ensemble learning to improve the performance of the prediction of HD. Author [22] Combined Infinit Latent feature selection method with SVM classifier and achieved an accuracy of 89.93% using three datasets, including Cleveland, Hungarian, and Switzerland, with 58 attributes. Author [23] used 'Z-Alizahed sani' dataset to develop a hybrid method by enhancing Neural network's performance using Genetic algorithm, and it yielded an accuracy of 93%. Author [24] developed a clinical decision support system for the accurate prediction of HF, using a hybrid approach based on both ANN and Fuzzy AHP. The result achieved an average prediction accuracy of 91.10%, which is 4.40% higher in comparison to that of the conventional ANN method. and author [25] proposed optimized stacked SVM algorithms. Author [26] implemented a homogeneous ensemble learning method that involves randomly partitioning the dataset into smaller subsets using a mean based splitting method, classification and regression tree (CART) was applied to model each partition. A homogeneous ensemble was then created using an accuracy based weighted aging classifier ensemble. The proposed method achieved an accuracy of 93% on the Cleveland dataset and an accuracy of 91% Framingham test set. yet the validation was carried out using only the train-test split method, which may cause overfitting. Author [27] trained and tested the optimized XG Boost approach only on the Cleveland dataset and achieved 91.80% accuracy using the cross-validation method. Author [28] contributed to a clinical decision support system for the improvement of the prediction of HF. The proposed model is introduced with different combinations of features and several known classification techniques. They combined random forest with a linear model (HRFLM) to enhanced performance level with an accuracy level of 88.7%.

The current work's contribution is to introduce a clinical decision support system to predict the risk level of HD using the patient's data set. A predictive model is proposed to detect patterns in existing HD patients' data, the classification approach (RFAB) combining Random Forest (RF) and AdaBoost algorithms has been proposed to predict HD. The effectiveness of the RFAB method is proved evidently by comparison with other studies and other algorithms including Naïve Bayes (NB), Decision Tree C4.5

(DT), Support Vector Machine (SVM), and RF to demonstrate the performance of the selected classifications algorithms to classify the HD data best. In the next section, our methodology is described with a brief detail of datasets. Section 4 presents the experimental results achieved. The different representations of outcomes are discussed in section 5. Finally, conclusions are given in section 6.

3   Proposed model

The studied model was based on available online datasets. And the proposed architecture of our system to predict the presence of HD is shown in Figure 2. As soon as the preprocessing was carried out in the initial step of the process, the Random Forest algorithm is used as a classifier to predict whether a patient has HD or not, followed by the AdaBoost boosting algorithm to improve the efficiency of the RF classifier. To evaluate the performance and validate the proposed approach, we used the confusion matrix that measures the accuracy. Finally, we implement different comparisons with other studies that existed in the literature.
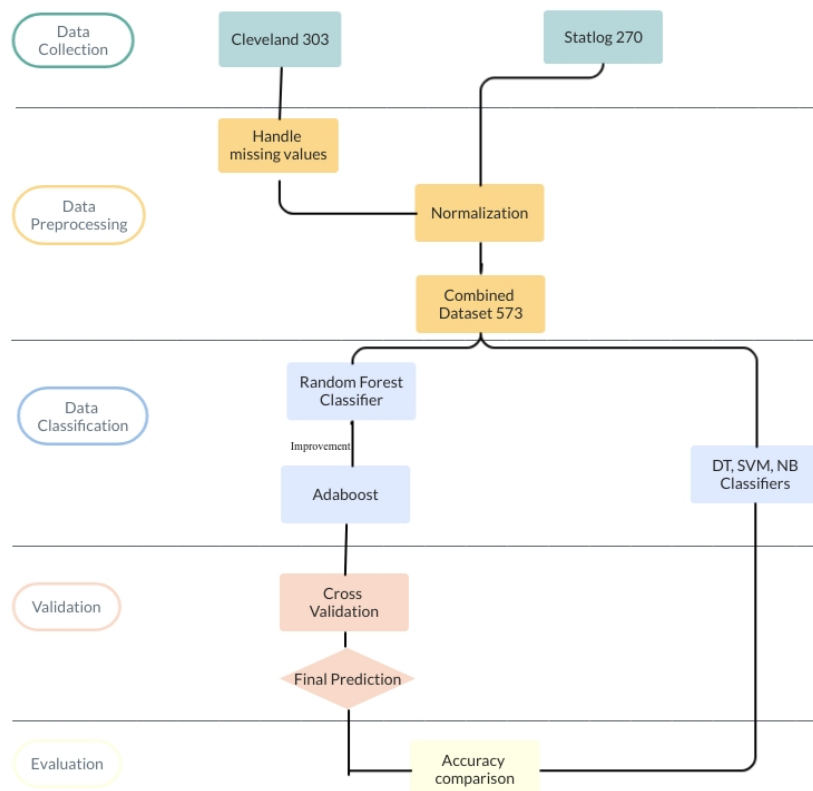

**Fig. 2.** Flow Chart of the proposed model to predict HD

3.1 Data collection

The Heart disease datasets that have been used for the experiments are Cleveland and Statlog datasets [29], [30]. Both datasets were retrieved from the machine learning repository database of the University of California, Irvine (UCI), Cleveland was collected from V.A. Medical Center, Long Beach, and Cleveland Clinic Foundation. The principal responsible for data collection was 'Roberto Detrano'.

Cleveland dataset comprised 303 instances and contained 76 raw attributes. We used 14 reliable features, including the predicted attribute that was assigned 0 for the absence of HD and 1 to 4 for the ascending risk levels of HD. Patient's sensitive information like name and SSN were all removed for confidentiality purposes. Statlog project heart disease dataset also consists of 14 features and 270 of total instances, the distribution and percentage of absence and presence of HD in both datasets is shown in Table 1. In Table 2, we present 14 attributes applied for modeling with their description and their values. Also, the histogram of all features is shown in Figure 3.

Table 1. Datasets distribution in term of absence and presence of HD

Dataset	Total of instances	Absence of HD	Presence of HD
Cleveland	303	164 (54.12%)	139 (45.87%)
Statlog	270	150 (55.55%)	120 (44.44%)
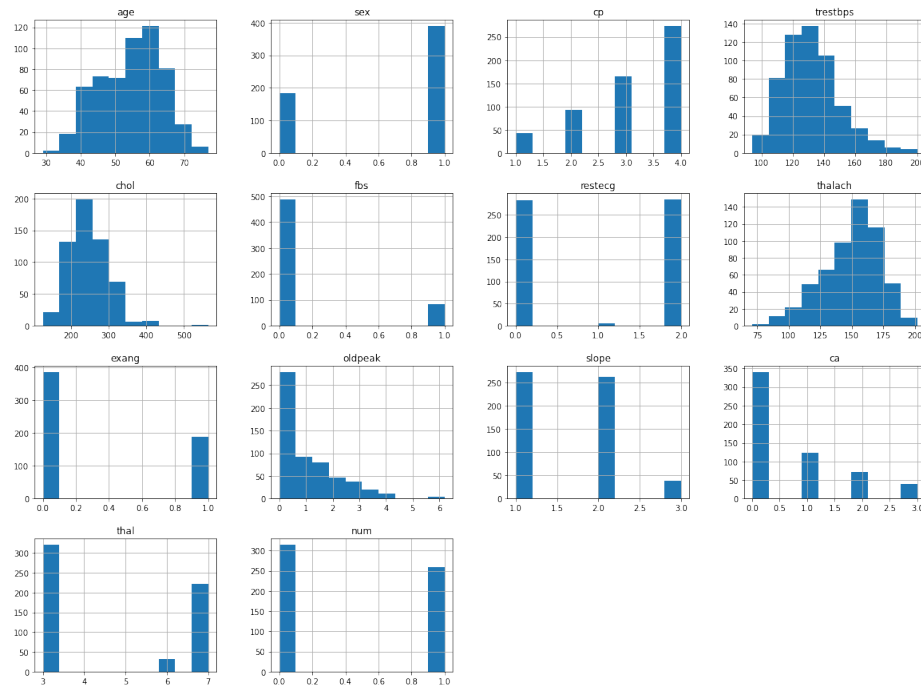
Fig. 3. Histogram of 14 attributes

Table 2. Description of selected attributes and their values

N	Name of Attribute	Description	Value
1	Age	Age in years	[29--77]
2	Sex	Male or female represented in binary	Male =1 Female =0
3	Cp	Chest pain types	Value 1= typical angina Value 2= atypical angina Value 3= non-angina Value 4= asymptomatic
4	Trestbps	Resting blood pressure	Represented in mm Hg [94--200]
5	Chol	Serum cholesterol	Represented in mg/dl [126--564]
6	Restecg	Resting electrocardiographic results	Value 0=normal Value1= having ST-T wave abnormality (Twave inversions and/or ST elevation or depression of >0.05 mV) Value2= showing probable or definite left ventricular hypertrophy by Estes' criteria
7	Fbs	Fasting blood sugar >120 mg/dl	Value 1=yes Value 0=no
8	Thalach	Maximum heart rate achieved	[71--202]
9	Exang	Exercise induced angina value	1= yes 0= no
10	Oldpeak	Measure of ST depression induced by exercise relative to rest	[0--6.2]
11	Slope	Measure of slope for peak exercise ST segment	Value 1= up sloping Value 2= flat Value 3= down sloping
12	Ca	Number of major vessels colored by flourosopy	Values range between [0--3]
13	Thal	Thallium stress test result	Value 3= normal Value 6= fixed defect Value 7= reversible defect
14	Num	Absence or presence of HD	Value 0= absence of HD Value 1 = presence of HD

3.2   Data preprocessing

Data preprocessing involves transforming data to an accurate and understandable format to improve the model efficiency and accuracy. Medical data are usually fuzz, incomplete, deficient attribute values, and contain irrelevant data[31]. In our research, the Statlog dataset has no missing values. However, the Cleveland dataset is characterized with six missing values, including four missing values for Number of major vessels (Ca) attribute and two missing values for Heart rate (Thal) attribute, even though both attributes had less than 5% missing values. We believe that imputing values would improve the data rather than removing them. To impute them, we use the "Mode" imputation method that replaced missing values with the most frequently occurring value since missing values are categorical [32].

The original datasets contain five output values; a value 0 indicated the absence of HD, and values between 1 and 4 showed different HD levels, respectively. In this study, we are interested in the presence or absence of HD. Thus, the output attribute is reclassified into a binary value of 0 or 1, indicating HD's absence or presence in the patients, respectively.
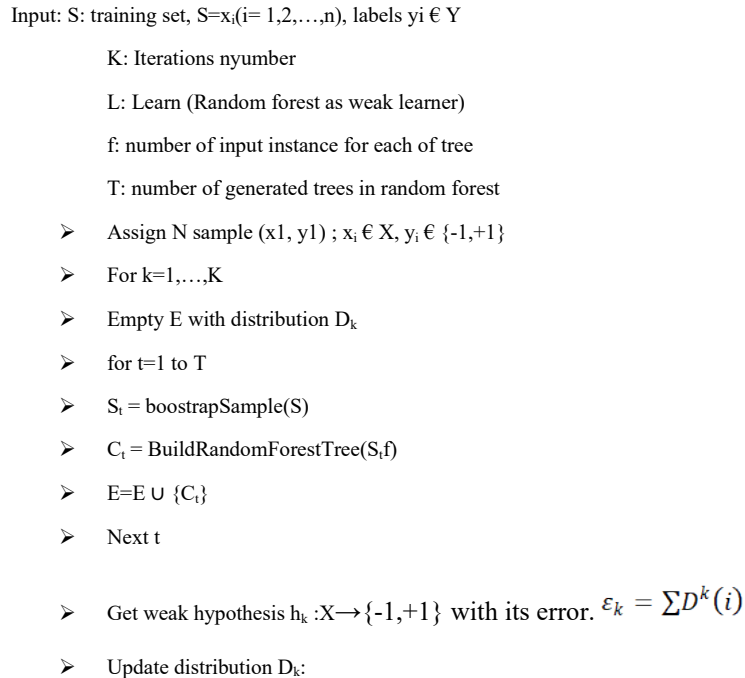
3.3 Classification method

Predictive analytic and machine learning go side by side, as predictive models generally embrace a machine learning algorithm as a machine learning algorithm that enables data-driven models to learn information from observed data in a training dataset to make the intended predictions [33]. There are two subtypes of predictive models: Regression and classification. Regression models analyze the relationships between variables to make predictions about continuous variables, while the classification assigns discrete class labels to particular observations as outcomes of a prediction[33]. Classification tasks can be organized into two main sub-categories: Supervised and unsupervised learning[4]. Our study has a specific target used to predict output for new use cases, which determines either the heart disease is present or not. Hence a supervised learning classification algorithm would be ultimate to train our data using Random Forest and AdaBoost algorithms.

Random Forest is a fusion of tree predictors that constructs multiple decision trees at training time and produces the class by voting on individual trees. It is similar to the decision tree algorithms concept, yet the algorithm builds a forest of decision trees with locations of attributes chosen at random. Its main advantage is improving the prediction accuracy without increasing the computational cost [34].

Boosting is a common approach for learning classifiers using an optimally weighted majority vote of weak classifiers to generate a robust classifier [35]. The most known boosting algorithm is AdaBoost introduced by Yoav Freund and Robert Schapire [36] as a meta greedy algorithm that builds up a powerful classifier by optimizing the weights and adding one weak classifier at a time.

The AdaBoost algorithm aims to maintain a set of weights over the training set. The weight on training example i on round k is denoted $S_k(i)$. Initially, all weights are initialized relatively, but on each round, the weights of misclassified samples are increased so that the weak learner is forced to focus on the hard examples in the trading set [37].

We propose a combination of Adaptive Boosting (AdaBoost) with RF as a base decision tree. As mentioned above, AdaBoost is a meta-algorithm that can be used in conjunction with many other learning algorithms to improve their performance and flexibility. It works by repeatedly running a given weak or base learning algorithm on diverse groups on the training data, afterward combining the weak learner's classifiers into a single composite classifier [38]. We will refer to the proposed learning method as RFAB. Steps of the hybrid AdaBoost and random forests algorithm are given in Figure 4.

Input: S: training set, S=x$_i$(i= 1,2,…,n), labels yi € Y

         K: Iterations nyumber

         L: Learn (Random forest as weak learner)

         f: number of input instance for each of tree

         T: number of generated trees in random forest

➤   Assign N sample (x1, y1) ; x$_i$ € X, y$_i$ € {-1,+1}

➤   For k=1,…,K

➤   Empty E with distribution D$_k$

➤   for t=1 to T

➤   S$_t$ = boostrapSample(S)

➤   C$_t$ = BuildRandomForestTree(S$_t$f)

➤   E=E ∪ {C$_t$}

➤   Next t

➤   Get weak hypothesis h$_k$:X→{-1,+1} with its error. $\varepsilon_k = \sum D^k(i)$

➤   Update distribution D$_k$:

Fig. 4. Illustration of RFAB process

3.4 The performance measurement

In order to measure the performance of the proposed classification algorithms, various evaluation measures have been implemented, including sensitivity, specificity, accuracy, recall (proportion of instances classified as a given class divided by the actual total in that class), MCC (a measure of the quality of binary classifications) and ROC curve. All these measures are calculated based on the confusion matrix described in Table 3.

Table 3. Confusion matrix

Confusion Matrix		Observed class	
		yes	*no*
Predicted class	yes	TP	FP
	no	FN	TN

The confusion matrix, True Negative (TN) signifies that the model correctly classifies a healthy person. True Positive (TP) represents that person having heart disease is correctly classified by the model. False Positive (FP) shows that the model misclassifies a healthy person. False Negative (FN) notifies that patient having heart disease is incorrectly classified by the model.

Specificity measures the ratio of negatives that are correctly classified, sensitivity measures the ratio of real positives that are correctly identified, and the accuracy of the classification model shows the overall performance of the model and can be calculated[39]. The following formulas can mathematically represent these measures:

$$Specificity = \frac{TN}{TN+FP} \tag{1}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{2}$$

$$Precision = \frac{TP}{FP+TP} \tag{3}$$

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

$$Specificity = \frac{TN}{TN+FP} \tag{5}$$

4   Results

The model is also validated using the K-Fold cross-validation (CV) method, which is the most convenient method to avoid overfitting and get more accurate results in the testing set. In K-Fold CV, the whole dataset is split into k equal parts. The (k-1) parts are utilized for training, and the rest is used for the testing at each iteration. This process continues for k-iteration; in this study, k = 10 is used for experimental work since it produces significant results.

The results are estimated using confusion matrix measurements to evaluate the proposed model's performance; Table 4 shows results achieved with both the RF algorithm and the hybrid model RFAB combining RF with AdaBoost algorithm, using a combined dataset, and illustrated in Figure 5. RF outperforms with 95.98% accuracy. However, the implementation of the hybrid method RFAB has enhanced the accuracy by 0.18 to achieve 96.16%. We are also confirming the proposed model's efficiency by illustrating the AUC of the ROC chart in Figure 6.

Table 4. Our proposed method performance on the combined dataset

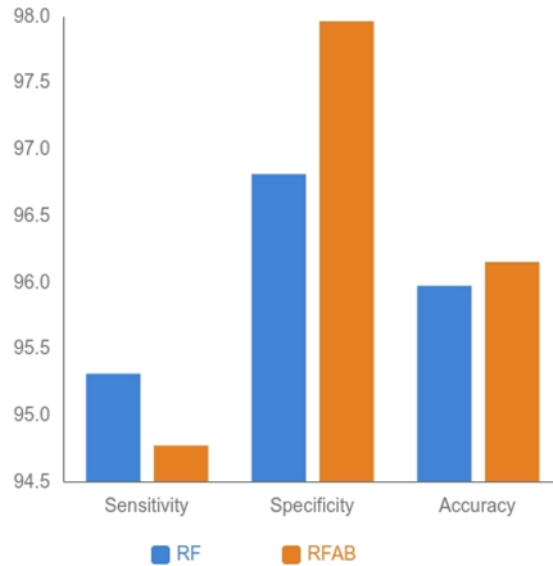| Methods	Recall	MCC	Precision	Specificity%	Sensitivity%	Accuracy %
RF	0.960	0.919	0.960	96.82	95.32	95.98
RFAB	0.962	0.923	0.962	97.97	94.78	96.16

Fig. 5. Measures performance comparison between simple RF and proposed RFAB
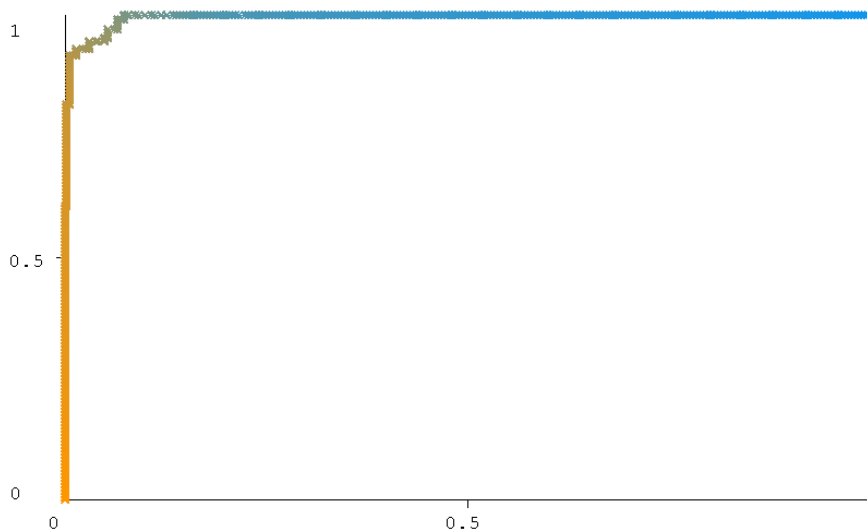

Fig. 6. ROC chart of the proposed model (ROC curve, AUC= 0.99, where: X: False positive rate, Y: True positive rate)

The comparison of performances of the proposed method with well-known ML algorithms including Naïve Bayes (NB) [40], C4.5 decision tree (DT) [41], [42], Support Vector Machine (SVM) [43], and Random forest (RF) [44]using both separated and combined datasets are listed in Table 5. NB achieved high accuracy of 82.17% on the

Cleveland dataset, while on Statlog dataset, both NB and DT achieved significant accuracy of 79.25%. All the algorithms listed, including the proposed method, achieved better results on the combined dataset comparing to each individual dataset. Figure 7 demonstrates the accuracy comparison; the proposed RFAB achieves the best result performance with 96.16.34% accuracy comparing to other classifiers such as NB: 87.60%, DT: 77.13%, and SVM: 94.58%.

Table 5. Results comparison between ML algorithms and our proposed method on all datasets

Dataset	ML techniques	Precision%	Specificity%	Sensitivity%	Accuracy %
Cleveland	SVM	76.60	78.88	73.94	76.56
	RF	69.61	71.29	68.20	69.30
	NB	82.31	83.46	81.25	82.17
	DT	75.81	79.18	76.56	75.57
	RFAB	68.81	69.64	68.64	68.64
Statlog	SVM	77.41	74.78	79.47	77.40
	RF	69.91	69.30	70.41	70.00
	NB	79.23	78.07	80.12	79.25
	DT	79.26	79.09	79.37	79.25
	RFAB	68.62	68.81	68.36	68.51
Combined Dataset	SVM	94.62	94.18	94.92	94.58
	RF	96.01	96.82	95.32	95.98
	NB	87.61	87.30	87.85	87.60
	DT	77.20	77.58	76.83	77.13
	RFAB	96.20	97.97	94.78	96.16
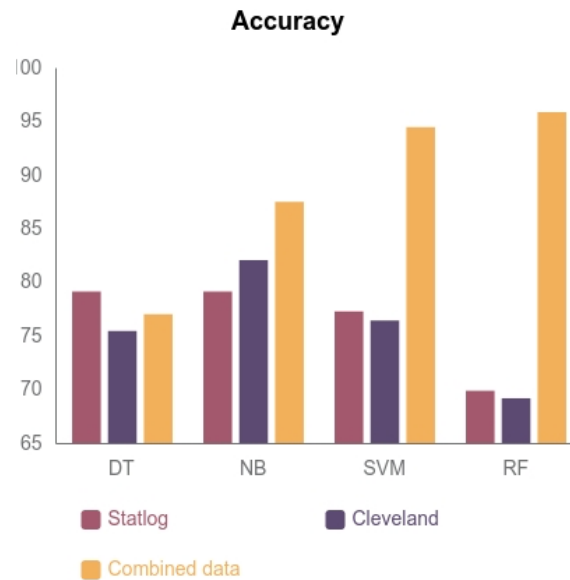
Fig. 7. Comparison of our proposed method with known ML techniques

5   Discussion

Several studies were carried on diagnosing and predicting heart disease using ML techniques are listed in Table 6. Most researches have used four datasets, namely Cleveland[29], Hungarian, Long Beach, and Switzerland[45]. These data were all collected from the UCI machine learning repository. Further studies have used the Cleveland dataset only since not lacking values[11], [13], [16]–[19]. In contrast, other datasets showed more than 90% of some attributes' missing values which might compromise the accuracy and the quality of results, e.g "thal"and "ca" attributes that shown to have high correlation with the output attribute. The number of selected attributes and standard features was ranging from 76 to 8, including the class attribute. Generally, the studies that used many attributes have applied feature selection to improve the relevance[19], [22]. Hence, we perform only 14 attributes, including (Age, Gender, Chest pain, blood pressure, etc.) that are relevant for the risk factors of HD diagnosis. Various prediction models were built using ML techniques. Nowadays, despite a substantial research output, no gold-standard model is available to predict HD. Hence, there is still a need for improvement. Our model applied a combination of Random Forest and AdaBoost algorithm trained and tested in both Cleveland and Statlog datasets with 14 attributes obtained evidencing results with high accuracy of 96.16% using cross validation technique. The summary of obtained comparative results with various classifiers is shown in Figure 10. Our proposed method reached the highest accuracy compared to NB, DT, SVM, and RF. This result demonstrated an enhanced efficiency of the proposed hybrid method. Indeed, many parameters impact the construction of the HD prediction model; these included the dataset of choice, the number of attributes and the output class, the algorithm used, and the validation method. Therefore, comparing our proposed method with other studies should consider all mentioned parameters to have a measurable comparison. Nevertheless, the proposed method outperforms in terms of accuracy and validation technique compared to other studies.

Table 6.   A comparison study with other recent studies

Ref	Datasets	Instances	Attributes	ML techniques	Validation	Accuracy %
[18]	Cleveland	303	14	C4.5 Fastdecision tree	10- cross-validation	78.54 77.55
	Statlog	270	14	C4.5 Fastdecision tree		76.60 76.60
[20]	Statlog	270	14	KNN SVM C5.0 NN	70-30 train-test	88.37 86.05 93.02 80.23
[28]	Cleveland	297	14	HRFLM approach	--	88.70
[46]	Cleveland	303	14	Decision rules	10-cross-validation	86.70
[19]	Cleveland	303	29	AdaptiveBoosting Algorithm	--	80.14
[23]	Z-alizahed sani	303	54	Genetic NN	10-cross-validation	93.85
	Cleveland	303	14			89 .4

Ref	Datasets	Instances	Attributes	ML techniques	Validation	Accuracy %
[24]	Cleveland	297	14	ANN and fuzzy AHP	Train-test-validation/Split	91.10
[22]	Cleveland+ Hungarian+ Switzerland	699	58	Linear SVM NB Logistic regression	10-cross-validation	89.93 77.70 85.61
[25]	Cleveland	297	14	Optimized Stacked SVM	70-30 train-test	92.22
[27]	Cleveland	303	14	Optimized XGBoost	10 cross-validation	91.80
[26]	Cleveland	303	14	Randomized decision tree ensemble	70–30 train-test	93.00
RFAB	Cleveland +Statlog	573	14	Improved AdaBoost with Random forest (FRAB)	10-cross-validation	96.16

6 Conclusion

In conclusion, we developed a clinical support system based on machine learning; we used a helpful clinical dataset for clinicians to produce an accurate and efficient diagnosis system. Indeed, the achieved results emphasize our model's validity with a high level of accuracy of 96.16% and we demonstrated the effectiveness of the proposed method using 10-fold cross-validation. Random forest along with AdaBoost had given a very promising result in predicting heart disease. As perspectives, we look forward for training our model on large datasets. Finally, a comparison with ensemble learning methods would be of interest.

7 References

[1] F. Jiang *et al.*, "Artificial intelligence in healthcare: Past, present and future," *Stroke Vasc. Neurol.*, vol. 2, no. 4, pp. 230–243, 2017, https://doi.org/10.1136/svn-2017-000101

[2] I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," *Artif. Intell. Med.*, vol. 23, no. 1, pp. 89–109, 2001, https://doi.org/10.1016/S0933-3657(01)00077-X

[3] A. Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?," *Brain Informatics*, vol. 3, no. 2, pp. 119–131, 2016, https://doi.org/10.1007/s40708-016-0042-6

[4] N. Mishra and S. Silakari, "Predictive Analytics : A Survey , Trends , Applications ," *Int. J. Comput. Sci. Inf. Technol.*, vol. 3, no. 3, pp. 4434–4438, 2012,[Online].Available: http://0ieeexplore.ieee.org.oasis.unisa.ac.za/stamp/stamp.jsp?tp=&arnumber=7426629

[5] Q. Chen *et al.*, "An automatic system to identify heart disease risk factors in clinical texts over time," *J. Biomed. Inform.*, vol. 58, pp. S158–S163, 2015, https://doi.org/10.1016/j.jbi.2015.09.002

[6] M. R. Vinutha and J. Chandrika, "Prediction of Liver Disease using Regression Tree," *Int. J. online Biomed. Eng.*, vol. 17, no. 2, pp. 164–172, 2021, https://doi.org/10.3991/ijoe.v17i02.19287

[7] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," *Procedia Comput. Sci.*, vol. 83, no. Fams, pp. 1064–1069, 2016, https://doi.org/10.1016/j.procs.2016.04.224

[8] K. Almohammadi, "Conceptual framework based on type-2 fuzzy logic theory for predicting childhood obesity risk," *Int. J. online Biomed. Eng.*, vol. 16, no. 3, pp. 95–106, 2020, https://doi.org/10.3991/ijoe.v16i03.12701

[9] M. Nilashi, O. Ibrahim, H. Ahmadi, L. Shahmoradi, and M. Farahmand, "A hybrid intelligent system for the prediction of Parkinson's Disease progression using machine learning techniques," *Biocybern. Biomed. Eng.*, vol. 38, no. 1, pp. 1–15, 2018, https://doi.org/10.1016/j.bbe.2017.09.002

[10] A. Khan and M. Usman, "Alzheimer's disease prediction model using demographics and categorical data," *Int. J. online Biomed. Eng.*, vol. 15, no. 15, pp. 96–109, 2019, https://doi.org/10.3991/ijoe.v15i15.11472

[11] H. El Hamdaoui, S. Boujraf, N. E. H. Chaoui, and M. Maaroufi, "A Clinical support system for Prediction of Heart Disease using Machine Learning Techniques," *2020 Int. Conf. Adv. Technol. Signal Image Process. ATSIP 2020*, 2020, https://doi.org/10.1109/atsip49331.2020.9231760

[12] N. A. Nayan *et al.*, "Cardiovascular disease prediction from electrocardiogram by using machine learning," *Int. J. online Biomed. Eng.*, vol. 16, no. 7, pp. 34–48, 2020, https://doi.org/10.3991/ijoe.v16i07.13569

[13] "Cardiovascular diseases (CVDs)." https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed Feb. 08, 2021).

[14] "National Heart, Lung, and Blood Institute (NHLBI) | National Institutes of Health (NIH)." https://www.nih.gov/about-nih/what-we-do/nih-almanac/national-heart-lung-blood-institute-nhlbi (accessed Feb. 08, 2021).

[15] "Heart Disease Symptoms & Signs of Other Heart Problems." https://www.webmd.com/heart-disease/guide/heart-disease-symptoms#1 (accessed Feb. 08, 2021).

[16] "Risk Factors for Heart Disease." https://www.webmd.com/heart-disease/risk-factors-for-heart-disease#1 (accessed Feb. 08, 2021).

[17] "Heart Disease and Stroke | Healthy People 2020." https://www.healthypeople.gov/2020/topics-objectives/topic/heart-disease-and-stroke (accessed Feb. 08, 2021).

[18] R. El-Bialy, M. A. Salamay, O. H. Karam, and M. E. Khalifa, "Feature Analysis of Coronary Artery Heart Disease Data Sets," *Procedia Comput. Sci.*, vol. 65, no. Iccmit, pp. 459–468, 2015, https://doi.org/10.1016/j.procs.2015.09.132

[19] K. H., J. H., and G. J., "Diagnosing Coronary Heart Disease using Ensemble Machine Learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 10, pp. 30–39, 2016, https://doi.org/10.14569/ijacsa.2016.071004

[20] M. Abdar, S. R. N. Kalhori, T. Sutikno, I. M. I. Subroto, and G. Arji, "Comparing performance of data mining algorithms in prediction heart diseses," *Int. J. Electr. Comput. Eng.*, vol. 5, no. 6, pp. 1569–1576, 2015, https://doi.org/10.11591/ijece.v5i6.pp1569-1576

[21] Purushottam, K. Saxena, and R. Sharma, "Efficient Heart Disease Prediction System," *Procedia Comput. Sci.*, vol. 85, pp. 962–969, 2016, https://doi.org/10.1016/j.procs.2016.05.288

[22] H. M. LE, T. D. TRAN, and L. VAN TRAN, "Automatic Heart Disease Prediction Using Feature Selection and Data Mining Technique," *J. Comput. Sci. Cybern.*, vol. 34, no. 1, pp. 33–48, 2018, https://doi.org/10.15625/1813-9663/34/1/12665

[23] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A. A. Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network-

Genetic algorithm," *Comput. Methods Programs Biomed.*, vol. 141, pp. 19–26, 2017, https://doi.org/10.1016/j.cmpb.2017.01.004

[24] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction," *Expert Syst. Appl.*, vol. 68, pp. 163–172, 2017, https://doi.org/10.1016/j.eswa.2016.10.020

[25] L. Ali *et al.*, "An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure," *IEEE Access*, vol. 7, pp. 54007–54014, 2019, https://doi.org/10.1109/access.2019.2909969

[26] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Informatics Med. Unlocked*, vol. 20, p. 100402, 2020, https://doi.org/10.1016/j.imu.2020.100402

[27] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized XGBoost based diagnostic system for effective prediction of heart disease," *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, 2020, https://doi.org/10.1016/j.jksuci.2020.10.013

[28] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, https://doi.org/10.1109/access.2019.2923707

[29] "UCI Machine Learning Repository: Heart Disease Data Set." https://archive.ics.uci.edu/ml/datasets/heart+disease (accessed Feb. 08, 2021).

[30] "UCI Machine Learning Repository: Statlog (Heart) Data Set." https://archive.ics.uci.edu/ml/datasets/statlog+(heart) (accessed Feb. 08, 2021).

[31] H. M. Zolbanin, D. Delen, and A. Hassan Zadeh, "Predicting overall survivability in comorbidity of cancers: A data mining approach," *Decis. Support Syst.*, vol. 74, pp. 150–161, 2015, https://doi.org/10.1016/j.dss.2015.04.003

[32] Z. Zhang, "Missing data imputation: Focusing on single imputation," *Ann. Transl. Med.*, vol. 4, no. 1, 2016, https://doi.org/10.3978/j.issn.2305-5839.2015.12.38

[33] H. Alharthi, "Healthcare predictive analytics: An overview with a focus on Saudi Arabia," *J. Infect. Public Health*, vol. 11, no. 6, pp. 749–756, 2018, https://doi.org/10.1016/j.jiph.2018.02.005

[34] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, https://doi.org/10.1023/A:1010933404324

[35] N. Noorbakhsh-Sabet, R. Zand, Y. Zhang, and V. Abedi, "Artificial Intelligence Transforms the Future of Health Care," *Am. J. Med.*, vol. 132, no. 7, pp. 795–801, 2019, https://doi.org/10.1016/j.amjmed.2019.01.017

[36] Y. Freund and R. E. Schapire, "Experiments with a New Boosting Algorithm," *Proc. 13th Int. Conf. Mach. Learn.*, pp. 148–156, 1996, http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.133.1040

[37] R. E. Schapire, "A brief introduction to boosting," *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2, no. 5, pp. 1401–1406, 1999.

[38] A. J. Wyner, M. Olson, J. Bleich, and D. Mease, "Explaining the success of AdaBoost and random forests as interpolating classifiers," *J. Mach. Learn. Res.*, vol. 18, pp. 1–33, 2017.

[39] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote Sens. Environ.*, vol. 62, no. 1, pp. 77–89, Oct. 1997, https://doi.org/10.1016/s0034-4257(97)00083-7

[40] I. Kononenko, "Inductive and bayesian learning in medical diagnosis," *Appl. Artif. Intell.*, vol. 7, no. 4, pp. 317–337, 1993, https://doi.org/10.1080/08839519308949993

[41] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986, https://doi.org/10.1007/bf00116251

[42] J. R. Quinlan, "Book Review : C4 . 5 : Programs for Machine Learning," *Mach. Learn.*, vol. 6, pp. 235–240, 1994.

[43] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, https://doi.org/10.1007/bf00994018

[44] Tin Kam Ho, "Random Decision Forests Tin Kam Ho Perceptron training," *Proc. 3rd Int. Conf. Doc. Anal. Recognit.*, pp. 278–282, 1995, [Online]. Available: https://doi.org/10.1109/icdar.1995.598994

[45] "Index of /ml/machine-learning-databases/heart-disease." https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/ (accessed Feb. 08, 2021).

[46] Purushottam, K. Saxena, and R. Sharma, "Efficient Heart Disease Prediction System," *Procedia Comput. Sci.*, vol. 85, pp. 962–969, 2016, https://doi.org/10.1016/j.procs.2016.05.288

[47] A. K. Paul, P. C. Shill, M. R. I. Rabin, and M. A. H. Akhand, "Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease," *2016 5th Int. Conf. Informatics, Electron. Vision, ICIEV 2016*, pp. 145–150, 2016, https://doi.org/10.1109/iciev.2016.7759984

Authors

Halima El Hamdaoui is a full time Ph.D. student at Artificial Intelligence, Data Sciences and Emerging Systems Laboratory, Ecole Nationale des Sciences Appliquées, University of Sidi Mohamed Ben Abdellah. Her research area includes Ma-chine learning and deep learning for Biomedical Engineering and intelligent systems.

Said Boujraf is head of Clinical Neurosciences Laboratory, Dept. of Biophysics and Clinical MRI Methods, Faculty of Medicine and Pharmacy, University of Sidi Moha-med Ben Abdellah, Fez, Morocco. His main research interest is biomedical engineering as well as neuroscience.

Nour El Houda Chaoui is senior lecturer at the Dept. Electrical and Computer Engineering, Ecole Nationale des Sciences Appliquées, University of Sidi Mohamed Ben Abdellah, Fez, Morocco. Her fields of interests include big data, Artificial intel-ligence, and decision systems.

Badreddine Alami is a Radiologist and lecturer at Faculty of Medicine and Pharmacy, University of Sidi Mohamed Ben Abdellah, Fez, Morocco. His main research interest is medical imaging and neuroscience.

Mustapha Maaroufi is head of Dept. Radiology and senior lecturer at Faculty of Medicine and Pharmacy, University of Sidi Mohamed Ben Abdellah, Fez, Morocco. His main research interest is radiology and medical imaging.