A Proposal of Clinical Decision Support System Using Ensemble Learning for Coronary Artery Disease Diagnosis

Rawia Sammout[1(✉)], Kais Ben Salah[2], Khaled Ghedira[3], Rania Abdelhedi[4], and Najla Kharrat[4]

[1] National School of Computer Sciences, Manouba, Tunisia
rawia.sammout@ensi-uma.tn
[2] Computing and Information Technology Faculty of Computing and Information Technology, Jeddah, South Africa
Kbensalah@uj.edu.sa
[3] SSOIE COSMOS National School of Computer Sciences, Manouba, Tunisia
khaledghedira3@gmail.com
[4] Laboratory of Molecular and Cellular Screening Processes Centre of Biotechnology of Sfax, Sfax, Tunisia
rania.abdelhedi@gmail.com, najla.kharrat@gmail.com

Abstract. Coronary Artery heart Disease (CAD) is the leading cause of mortality in the world. It is a complex and multifactorial disease resulting in several acute coronary syndromes and lead to death. In healthcare, an accurate clinical decision support system (CDSS) for CAD prediction has become increasingly important for making granted decisions at premature stage. Intensive research has been conducted on improving classification performance using machine learning techniques and metaheuristics algorithms. But most of these studies introduced the "classic risk factors" for CAD diagnosis i.e., demographic and clinical data. In this study, we present a novel CDSS based on ensemble learning for CAD prediction and we emphasize on adding other medical markers i.e., therapy data, some genetic polymorphisms along with classical factors. The new framework exploits the potential of three base classifiers including Support Vector Machines, Naïve Bayes and Decision Tree C4.5 to improve the prediction performance. Six experimental data used to build the proposed framework: the first one is collected from a Tunisian biotechnology center and the five other datasets from the University of California at Irvine repository. The analysis of the results shows that the proposed CDSS has the highest rate on classification accuracy, precision, recall and F1-measure when compared with CSGA Bagging and Adaptive boosting on the different datasets and proves that some medications and genetic polymorphisms such as Antivitamin K, Dose Beta Blocker, Proton pump inhibitor, CYP2C19*17, Clopidogrel active metabolite have an impact in CAD diagnosis.

Keywords: Coronary artery heart disease · Genetic factors · Medications · Ensemble learning

J. Ye et al. (Eds.): MobiHealth 2020, LNICST 362, pp. 300–314, 2021.
https://doi.org/10.1007/978-3-030-70569-5_19

1 Introduction

According to the World health organization report of 2017 [1], Coronary heart disease (CHD) represents the highest death rate among non-infectious diseases in the world. Various forms of cardiovascular disease exist such as stroke, rheumatic fever/rheumatic heart disease, high blood pressure, valvular heart disease and coronary heart disease on which our paper is focused. A blood clot resulting in a heart attack is typically the main cause of a sudden blockage of a coronary artery which leads to the reduction of blood and oxygen supply to the heart and to the coronary artery disease (CAD) [2]. Moreover, the atherosclerotic plaque growth model combines information from genetic and biological data of the patients. Therefore, it is essential to study the effect of certain genetic polymorphisms in the genes of patients with biological markers for CAD diagnosis. To the best of our knowledge, previous studies have used mainly different factors to diagnose CAD such as demographic, clinical, Electrocardiogram (ECG), symptoms and physical examination features [3, 4, 5]. Only a few of studies utilized some genetic polymorphisms in CAD diagnosis. Hence, it is still an active research in finding indicators for CAD diagnosis. However, Various techniques are used in CAD diagnosis such as ECG, Echocardiogram, Stress test, Cardiac catheterization and angiogram, Heart scan [3, 5] etc. But unfortunately, all these methods are expensive, protracted, and invasive. Moreover, the treatment cost for CAD is very expensive (estimated to US $ 14 billion per year) in the USA [6]. Therefore, new alternatives based on data mining (artificial neural networks, boosting, SVM) and soft computing (fuzzy logic, genetic algorithms) have been proposed to overcome time complexity, high diagnosis and treatment costs and adverse effects issues. Y. Niranjana Devi and S. Auto [7] used the decision tree algorithm to select significant attributes and then extract crisp if-then rules to constitute the fuzzy rule base for the fuzzy system. Finally, they applied genetic algorithm GA to optimize the fuzzy membership function. The results showed the performance of the system was significantly better than other systems. Next, Wiga Maulana Baihaqi et al. [8] examined the combination of datamining techniques (C4.5, CART, and RIPPER) and the fuzzy expert system to generate fuzzy rules to diagnose CAD. As a result, C4.5 and the fuzzy expert system outperforms studied classifiers with an accuracy of 81.82%. A recent research carried out by Kathleen H, Miao et al. [9] proposed for CAD diagnosis. An advanced ensemble machine learning model based on adaptative boosting (AdaBoost) algorithm was applied on four cardiac open datasets. The results indicated that the proposed ensemble achieves accuracy of 80.14% for Cleveland data, 89.12% for Hungarian data, 77.78% for Long beach data, and 96.72% for Switzerland data and outperforms existing models. Further, A new diagnosis model for CAD was introduced by N. Samadiani and S. Moameri [10]. The studied factors are extracted from SPECT heart disease images. Then a feature selection step was performed using Cuckoo Search CS and Genetic Algorithm GA to find the most significant features for CAD diagnosis. Then, the results are classified using the bagging classifier. The results of the proposed model (77,19%) are significantly better than GA or CS with a bagging classifier. Additionally, Kai Lei et al. [11] applied a weighted Naïve Bayes model on attribute relevancy for CAD diagnosis. The studied risk factors incorporated in this study are CAD symptoms. The improved Naïve Bayes model outperforms standard Naïve Bayes because of the studying of attributes relevance. While most of previous research yielded successful results for

the diagnosis of CAD using single classifiers, ensemble classifiers also showed expected excellent results in CAD classification [9]. Therefore, research on using ensemble model for CAD diagnosis is still active. Even though several CDSS have been introduced for CAD diagnosis, most of them have incorporated specific risk factors with the studied population such as American, Indian, Indonesian, Chinese etc. But environmental factors, lifestyle, diet habits aren't the same. On the other hand, it might be other factors that may help to assess CAD disease in another community. The existing CDSSs are not able to incorporate new risk factors. These limitations are handled in this research by taking into consideration more heterogeneous factors (72 biomarkers) including four genetic features such as CYP2C19*2, CYP2C19*17, CYP2C9*2 (rs1799853) and CYP2C9*3 (rs1057910) polymorphisms and some medications plus demographic and clinical features to build a new CDSS for CAD diagnosis. The proposed framework aims to improve the prediction accuracy. This paper is organized as follows: Sect. 2 introduces the techniques used to build the proposed framework. Section 3 covers the experiment datasets, finding and a discussion of the results. Finally, Sect. 4 concludes this paper.

2 Materials and Methods

2.1 Design of the Proposed CDSS

The proposed CDSS for CAD diagnosis is presented in the following flowchart given in Fig. 1. It consists of three main phases detailed below: preprocessing, classification and prediction and evaluation.

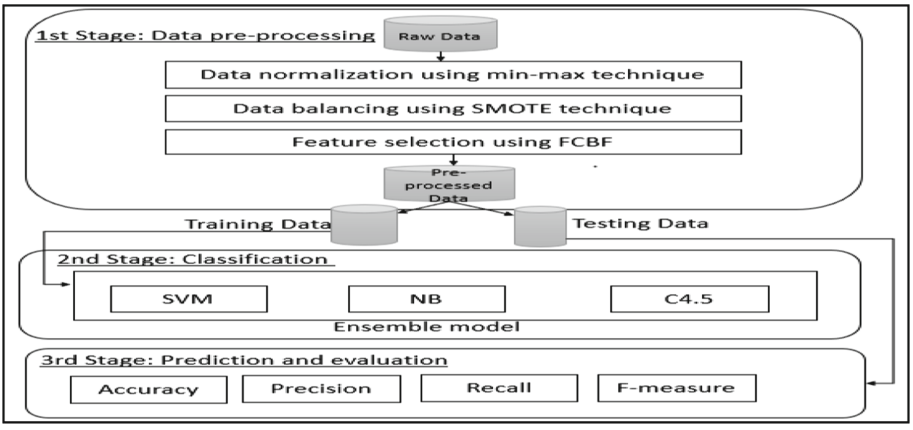

Fig. 1. General design of the proposed CDSS

2.2 Data Pre-processing Phase

A data preprocessing phase consists of three main steps: scale normalization, sampling, and feature selection, detailed below:

Normalization Using Min-Max Technique

Using data with different measurement units may have effect on the analysis and leads to different results. For example, using meters to measure the height instead of inches will lead to giving greater importance to the attributes with greater weight [12]. Therefore, normalization represents an essential step in preprocessing in order to give all attributes equal importance (weights). It aims to transform an original range of data to a new range. Also, it may be helpful to maintain the large variation in prediction or forecasting [13]. Min-max technique is widely used in the literature and known as a very simple method that provides a linear mapping of data from an unstructured range to new values of data. It also insures keeping relationship among original data values [14, 15]. Normalization is calculated using the following formula:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} * (new_{max} - new_{min}) + new_{min}, \tag{1}$$

where $X^{'}$ *is the new value*, X_{min} is the minimum value and X_{max} is the maximum value in the attribute.

In the present study, the original data are mapping in the range [0, 1] (where $new_{min} = 0$ and $new_{max} = 1$) and the simplified following formula is used:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}, \tag{2}$$

Sampling Using Smote Technique

Class imbalance ratio is high specifically in genomic dataset where the number of instances from one class is higher than the other class. The class having the higher number is called majority class, while the other one is known as minority class. Generally, classifiers are more sensitive to select majority class and less sensitive to detect minority class. Therefore, it may lead to a biased classification output. Hence, a combination between a classification algorithm and a sampling technique becomes mandatory. In this study, an oversampling technique known by synthetic minority oversampling technique (SMOTE) [16] is selected to handle this issue while the studied datasets are small. It has an ability to generate synthetically observations from the minority class samples to over-sample the minority distribution by joining any/all the k minority class nearest neighbors [17, 18]. It aims to balance a dataset with a binary target variable. Figure 2 below explains the process of this technique.

Fast Correlation-Based Feature Selection

A multivariate subset search technique called fast correlation-based filter selection (FCBF) [19] is used to select the subset of the most relevant and irredundant features among the full set of features. The attributes are ranked using an evaluation criterion called symmetric uncertainty (SU) [20]. Then, a threshold value of this latter is fixed and the attributes with values above this threshold (have highest dependency on the output variable) are selected to construct the model and the rest of attributes with values below the threshold (have low dependency) are removed. However, this technique has the ability of capturing non-linear correlation between features and modeling feature

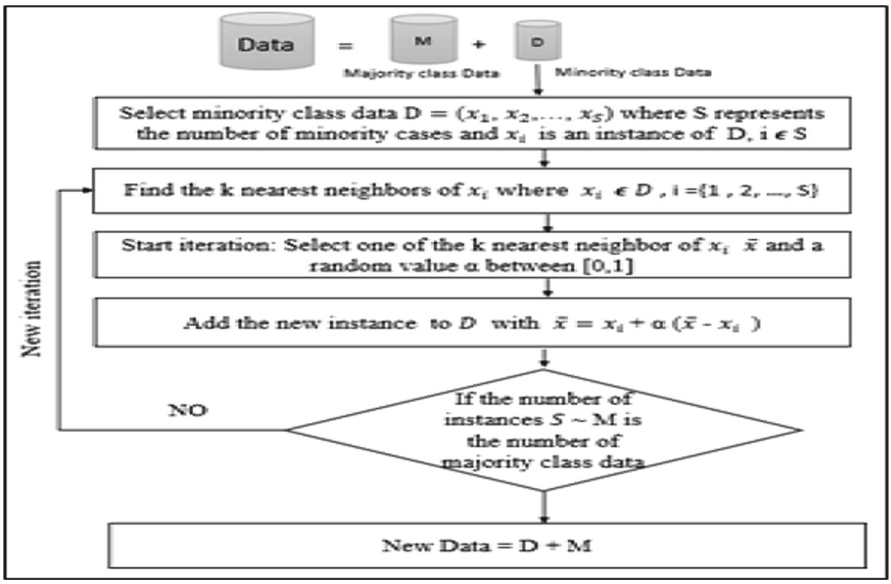

Fig. 2. Oversampling process using SMOTE technique

dependencies. Besides, it helps to reduce overfitting problem and time complexity and to improve the learner's performance [21]. The formula for calculating the SU measure is given below:

$$SU(X|Y) = 2\left[\frac{IG(X|Y)}{H(X)+H(Y)}\right] \tag{3}$$

Where $IG(X|Y)$ [22] is the information gain and represents the amount of the decrease of entropy of X provided as additional information by Y and calculating by the formula as follows:

$$IG(X|Y) = H(X) - H(X|Y) \tag{4}$$

With H(X) represents the uncertainty of a random variable X known by the entropy and is defined as:

$$H(X) = -\sum_i P(x_i) log_2(P(x_i)) \tag{5}$$

With $P(xi)$ is the probability of xi and H(X|Y) is the entropy of X after seeing values of Y and is calculated using (6) given by:

$$H(X|Y) = -\sum_j P(y_j) \sum_i P(x_i|y_j) log_2(P(x_i|y_j)) \tag{6}$$

With $P(yj)$ is the probability of yj and $P(xi|yj)$ is the conditional probability of xj given that yj has occurred.

$$P(x_i|y_j) = \frac{P(x_i \cap y_j)}{P(y_j)} \tag{7}$$

2.3 Classification Phase: Proposed Ensemble Learning Model

Ensemble Learning

Ensemble learning is a new concept that combines more than one model to predict a target output with more efficiency and accurate decisions than single model [23]. Thus, it leads to excellent classification results superior to those of a single classifier in many fields including cardiac arrhythmia [24], DNA microarray classification [25], and different heart diseases [26]. Diversity of ensemble members and different classification properties are required in ensemble learning in order to achieve high classification performance [27] with a good management of bias-variance errors [28]. A good ensemble strategy is ensured by the complementarity between its classifiers where the diversity between classifiers could be ensured by establishing sample techniques or training the classifiers by different training sets [27]. In this work, three techniques SVM, NB and DT are selected to build the ensemble learning, they will be discussed in the following subsections. The results of the analysis carried out by the discussed techniques are combined using a combination technique that will be explained below:

Support Vector Machines

As the sample studied in this review is a small dataset, support vector machines (SVM) is selected as a base classifier to be used in this study. It is recommended in the literature as an efficient classification technique for small-sample data [29]. Moreover, this classification model has been widely used to classify genomic datasets and yielded to excellent results. SVM is a supervised learning algorithm introduced by Vapnik (1998). It is a two-class classifier. It aims to design a N dimensional hyperplane that classify all training vectors (target variable/class label and feature variables) into two classes and leaves the maximum margin from both classes [30]. To maximize the margin, we have to solve a quadratic (nonlinear) optimization problem in order to maximum margin hyperplane as illustrated in Fig. 3 below:

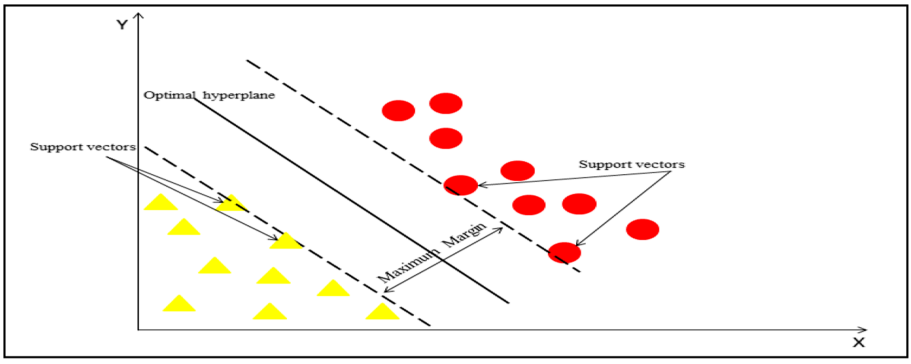

Fig. 3. Support Vector Machines

Naïve Bayes
Naive Bayes (NB) is a probability-based classification technique. It applies Bayes 'theorem with considering the independence assumption between all features [31]. The NB classifier calculates the probability that a given instance X belongs to a class label y. Given an instance X, characterized by a set of attributes (x_1, $x_{2,\ldots,}x_n$), and a class output y, the Bayes theorem consists of calculating the posterior probability $P(y/X)$ using the following formula:

$$P(y/X) = \frac{P(y)P(X/y)}{P(X)} \tag{8}$$

Moreover, NB classifier yields generally to excellent classification results and surprisingly outperforms more sophisticated algorithms in classification even without considering the independence assumption [32].

Decision Tree C4.5
Decision trees have become one of the most powerful and popular classification approaches used in the literature. They have many advantages, such as being comprehensible, easy and they require low computational effort [33]. In this paper, we emphasize the study on C4.5 decision tree algorithm as it is one of the most popular algorithms which is widely used for genomic dataset analysis [34, 35]. C4.5 is a top-down tree growth algorithm proposed by Ross Quinlan in 1993 [36], and its algorithm starts by calculating entropy and equivalent information gain to measure the importance of the attributes. Feature with the highest information gain tends to be selected as the most influential attribute in the classification process. The set of examples will be splitted according to the possible values of the selected feature. This process will be repeated iteratively until the decision tree learns from the set of the training examples. The formula for measuring information gain IG and entropy H are described above in Eqs. (4) and (5) respectively.

Weighted Majority Voting
Weighted Majority Voting represents one of the simplest methods for combining several classifiers. Let f is is the decision function of the i^{th} model where $i \in n$ and n represents the number of classifiers in the ensemble models and C is the class label $C_j = \{j = 1, 2, \ldots, C)$. However, the final decision fem(x) of the ensemble models is calculated as follow:

$$f_{em}(x) = \text{argmax}_C \sum_i w_i \delta(C,\ f_i(x)) \tag{9}$$

With w_i is the weight for the prediction model and $\delta(C, fi(x))$ is the probability for each instance of the class C label according to the classifier i.

2.4 Performance Evaluation Measurement

To evaluate the proposed CDSS performance with other models, we used the basic metrics such as precision, recall, classification accuracy and F-measure [37]. The main formulations of these metrics are:

$$Precision = \frac{T_p}{T_p + F_p} \tag{10}$$

$$Recall = \frac{T_p}{T_p + F_N} \tag{11}$$

$$Accuracy = \frac{T_N + T_p}{T_p + T_N + F_N + F_p} \tag{12}$$

$$F_{measure} = 2 * \frac{Precision * Recall}{Precision + Recall}, \tag{13}$$

With TP = True Positive, TN = True Negative, FN = False Negative, FP = False Positive. Indeed, accuracy represents the percentage of a correct CAD prediction (test is true) and a non-CAD prediction (test is false). Recall (sensitivity) is the true positive rate of CAD while precision is the positive predicted value of CAD. F-measure represents the weighted harmonic mean of precision and recall. In addition, a ten-fold cross validation (CV) has been successfully used for evaluating the performance of a machine learning algorithm(s) as it offers reliable approximates for the classification accuracy on each classification task [38]. Moreover, it is able to reduce the variability but increases the selection bias in case of feature selection or model parameters 'tuning. Thus, an external cross validation [39] is needed by holdout a testing set (30% of the sample) and applied 10-fold CV on the training (70%) and then evaluate model accuracy using the hold out testing set. This technique helps to reduce the selection bias and therefore guarantee the tradeoff between the bias and the variance.

3 Experimental Results and Discussion

3.1 Datasets

Based on a recent study of the National Public Health Institute 2018, heart diseases are the primary risk of death in Tunisia rather than infectious diseases. The studied population (see Table 1) consists of 213 patients from the south of Tunisia. The patients were admitted in the biotechnological Center in Sfax Tunisia for coronary artery disease diagnosis. The period of the study extends from January 1, 2010 to April 30, 2013. The dataset contains 72 categorical and numerical features considered for the prediction. The diagnosis result as the target variable. The studied features (see Table 2) represent clinical characteristics, genetic polymorphisms, and some medications for example. The target variable has a binary CAD diagnosis (1: diseased, 0: healthy). To ensure efficiency of the proposed CDSS, four majors most widely used cardiac databases from UCI repository are studied. They are Cleveland, Hungarian, Switzerland, and Long Beach [39]. These datasets consist of 76 attributes, but 14 of them are the mainly used. Furthermore, a cardiac Single Proton Emission Computed Tomography (SPECT) images dataset is studied for a comparison purpose. It is composed of 267 patient SPECT image records and 23 extracted binary features.

Table 1. Description of the experiment datasets

Dataset	Number of attributes	Number of classes	Number of positive cases	Number of negative cases	Total number of cases
Studied population	72	2	150	63	213
Cleveland	14	2	139	164	303
Hungarian	14	2	106	188	294
Switzerland	14	2	155	8	123
Long Beach	14	2	149	51	200
SPECT	23	2	212	55	267

Table 2. Description of the studied features in the Tunisian dataset

Variables	Description
Genetic polymorphisms	CYP2C19*2, CYP2C19*17, CYP2C9*2 (rs1799853), YP2C9*3 (rs1057910)
Biomarkers	Time of collection (Hours), Number of dilated arteries, Systolic blood pressure, Dyastolic blood pressure, Glycemia, Creatinine, Urea, CPK (creatine phosphokinase), Triglyceride, Cholesterol total, Na (sodium), CL (chlorine), K (potassium), Leukocytes, Hemoglobin, Platelets, Number of stents, Coronarography results, Event time (month), Event, Diagnosis (angina effort, SCA ST−, SCA ST+), INDICATION (TTT, PAC, ATL), Type of artery 1, Age, Sex, Non-insulinodependant diabetes, Insulin-dependent diabetes, Smoking, Dyslipidemia, HyperCT, HyperTG, Mixed dyslipidemia, Family history of CAD, Renal failure, Previous MI, Previous PCI, Previous CABG, Previous stroke, Alcohol
Medications	Clopidogrel loading dose, Clopidogrel maintenance dose, Clopidogrel treatment duration, Clopidogrel carboxylic acid (ng/ml), Clopidogrel (pg/ml), Clopidogrel acyl glucuronide (ng/ml), Clopidogrel active metabolite, Statins, Dose statins, Aspirin, Aspirin loading dose, AVK (Antivitamin K), ACE inhibitor, DOSE IEC, Angiotensin II receptor antagonist, Beta blockers, DOSE BB, Calcium channel blocker, Diuretic, DIURETIQ ARAII, proton pump inhibitor, Dose ipp, Nitrated derivatives, AGRASTAT, Reopro

3.2 Hyperparameters Setting

Hyperparameters represent parameters of the classifier that must be tuned before training to guarantee good classification results. In our case, SVM has two main parameters to optimize i.e., gamma, the coefficient C and the kernel, while DT has other parameters to tune such as number of features in each split, the minimum number of samples that

must be in the leaf node, the minimum number of samples required in an internal node. Grid search algorithm (GS) is used in this study. It is a heuristic technique that aims to find the optimal parameters of a model among a given subset of hyperparameters space [40]. This algorithm is the most widely used algorithm because of its simplicity. The principle of this algorithm is to minimize a loss function using a combination of a tuple of parameters among the defined space. However, the grid search results must be evaluated using cross validation/boosting or hold-out test on the performance metrics to estimate the generalization performance. In this work, a ten-fold cross validation technique is used with grid search algorithm.

3.3 Results and Discussion

We applied the proposed CDSS to a Tunisian population dataset and four benchmark cardiac datasets to prove its Dataset Number of attributes Number of classes Number of positive cases Number of negative cases Total number of cases Studied population 72 2 150 63 213 Cleveland 14 2 139 164 303 Hungarian 14 2 106 188 294 Switzerland 14 2 155 8 123 Long Beach 14 2 149 51 200 SPECT 23 2 212 55 267 Variables Description Genetic polymorphisms CYP2C19*2, CYP2C19*17, CYP2C9*2 (rs1799853), CYP2C9*3 (rs1057910) Biomarkers Time of collection (Hours), Number of dilated arteries, Systolic blood pressure, Dyastolic blood pressure, Glycemia, Creatinine, Urea, CPK (creatine phosphokinase), Triglyceride, Cholesterol total, Na (sodium), CL (chlorine), K (potassium), Leukocytes, Hemoglobin, Platelets, Number of stents, Coronarography results, Event time (month), Event, Diagnosis (angina effort, SCA ST−, SCA ST+), INDICATION (TTT, PAC, ATL), Type of artery 1, Age, Sex, Non-insulinodependant diabetes, Insulin-dependent diabetes, Smoking, Dyslipidemia, HyperCT, HyperTG, Mixed dyslipidemia, Family history of CAD, Renal failure, Previous MI, Previous PCI, Previous CABG, Previous stroke, Alcohol Medications Clopidogrel loading dose, Clopidogrel maintenance dose, Clopidogrel treatment duration, Clopidogrel carboxylic acid (ng/ml), Clopidogrel (pg/ml), Clopidogrel acyl glucuronide (ng/ml), Clopidogrel active metabolite, Statins, Dose statins, Aspirin, Aspirin loading dose, AVK (vitamin K), ACE inhibitor, DOSE IEC, Angiotensin II receptor antagonist, Beta blockers, DOSE BB, Calcium channel blocker, Diuretic, DIURETIQ ARAII, proton pump inhibitor, Dose ipp, Nitrated derivatives, AGRASTAT, Reopro efficacity. The experiments conducted for evaluating the performance of the proposed ensemble learners and all the studied classifiers are performed using 10-fold CV strategy to alleviate the insufficiency of small studied samples. The proposed CDSS is implemented using 70% of a training set and testing splitting on 10-fold CV and a validation set of 30% and running on 100 different seeds to validate the results with the mean accuracy value. As described in Table 1 below, the skewed Tunisian dataset is composed of 163 CAD patients (as majority class) and 63 non-CAD patients as minority class. After applying SMOTE technique, a balanced dataset (BD) is generated with equal class sizes. Indeed, Table 3 compares results of DT classifier using 10-fold cross validation and grid search techniques before and after oversampling the data using different performance evaluation metrics. The results obtained when the data is imbalanced show that the positive class CAD has effective prediction results with high rates in precision 81%, recall 94% and F1-measure 87% while the negative class No CAD has low rates in precision 9%,

recall 3% and F1 measure 4. However, after balancing the dataset we can see an increasingly prediction improvement for the negative class with 78% precision, 72% recall and 75% F1-measure. The sampling process is repeated for the four benchmark datasets while they have also imbalanced class distribution. Next, the numerical attributes of the balanced data are normalized using min_max normalization technique to avoid large variation in the prediction results and improve the prediction accuracy. Using the same classifier (DT) on the same data shows an improvement from 75,72% to 76,58% on accuracy rate and from 76% to 77% for other metrics.

Table 3. Performance evaluation before and after balancing the Tunisian dataset

	Imbalanced data		Balanced data	
Metrics/Class	CAD	No CAD	Metrics/Class	CAD
Precision	81%	9%	Precision	81%
Recall	94%	3%	Recall	94%
F1-measure	87%	4%	F1-measure	87%
Accuracy	77%		75,72%	

Then, this study has investigated the determination of CAD factors and emphasized on studying the impact of some genetic polymorphisms and medications which may help in the diagnosis of CAD. However, we performed a feature selection process using FCBF model to select the most significant attributes independently of the classifier. Then, we applied C4.5 algorithm to test the select features subset on prediction accuracy improvement as DT is simple and widely used in biology. Based on the results obtained in Table 4, we consider that the best subset of medical markers is sufficient to predict CAD with a high accuracy and provides less computational time than using all the features set. For example, the eight selected significant features from the Tunisian dataset represent one genetic feature (CYP2C19*17) among the four studied ones and five drugs (Antivitamins K (AVK), Dose Beta blockers, Proton pump inhibitor, Clopidogrel active metabolite) among all the studied medications and three other clinical markers (Event time/month, Previous stroke, Obesity). Hence, these results prove that the selected genetic factors and drugs are important indicators to diagnose CAD.

Furthermore, a classification stage is performed using the novel ensemble learners based on a weighted majority voting technique to aggregate the prediction results. The model weights are estimated according to their prediction accuracy (the model with the highest accuracy rate has the highest weight and so on). The proposed CDSS is examined on five different populations to prove its generalization ability and it yielded successful results. Table 5 lists the existing ensemble models including adaptive boosting (AdaBoost) [9] and CSGA Boosting [10] in the comparison. The results (Table 5) show that the new system achieved the best classification accuracies when comparing with the two existing ensembles. Indeed, comparing with AdaBoost on the five studied data, the new system yielded the highest prediction accuracies on the five studied populations i.e., Tunisian, Cleveland, Hungarian, Switzerland, and Long Beach respectively

Table 4. Performance evaluation before and after feature selection

Data	Attributes Number	Selected Features	DT Accuracy	Execution Time (s)
Tunisian	72	All the features	75,72%	3.73
Tunisian	8	Event time, AVK, Dose Beta Blockers, Proton pump inhibitor, Previous stroke, CYP2C19*17, Clopidogrel active metabolite, Obesity	78,85%	3.30
Cleveland	14	All	74,08%	3.44
Cleveland	8	Sex, cp, restecg, thalach, exang, old peak, ca, thal	78,66%	3.11
Hungarian	14	All	77,13%	3.61
Hungarian	8	Sex, cp, chol, fbs, exang, oldpeak, slope, thal	81,14%	3.04
Switzerland	14	All	86,09%	3.35
Switzerland	4	Sex, cp, fbs, exang	95,22%	3,12
Long beach	14	All	73,82%	3.541
Long beach	5	Age, sex, cp, exang, oldpeak	77.52%	3.28

with 79.41%, 82.27%, 89.48% and 97.45% compared with AdaBoost 71.15%, 80.14%, 89.12% and 96.72% and with CSGA Bagging 69.18%, 81,11%, 88,78%, 93,4%, 76,13%. Furthermore, Table 5 shows that the new framework achieved the highest accuracy rate of 79.72% on SPECT dataset while AdaBoost 76.41% and CSGA+ Bagging 77.19%. In conclusion, the proposed framework contributes efficiently to the prediction performance improvement due to its complementarity and diversity. However, the complementarity is ensured between the three used classifiers (SVM, DT and NB) by complementing the weaknesses between them and maximally improving the classification accuracy of the ensemble. Whereas the diversity is ensured by their different natures like probabilistic nature of NB and the complex nature of SVM and the tree-based nature of DT.

Table 5. Comparison of performance between the proposed CDSS and existing models

Dataset	Proposed ensemble	Adaptive boosting [9]	CSGA + Bagging [10]
Tunisian	79.41%	71.15%	69.18%
Cleveland	82.27%	80.14%	81,11%
Hungarian	89.48%	89.12%	88.78%
Switzerland	97.45%	96.72%	93.4%
Long beach	79.91%	77.78%	76.13%
SPECT	79.72%	76.41%	77.19%

4 Conclusion and Perspectives

In this study, we propose a new ensemble learning system based on three base classifiers SVM, NB and C4.5 DT in order to improve the prediction performance for CAD as a classification problem. The performance of the proposed CDSS is tested with 10-fold

cross validation on different cardiac datasets from different populations such as Tunisian, Hungarian, Switzerland, etc. The original datasets have an uneven distribution which may affect the classification performance and lead to an overfitting. Hence, a SMOTE technique has been applied to balance the class distribution. Then, we applied a feature selection technique called FCBF in order to determine the most effective features needed in the diagnosis of CAD and reduce the classification time complexity. Further, it may eventually help to reduce the cost of CAD diagnosis by limiting clinical markers needed and administrate some specific medications for CAD. However, the results of this process prove that some medications and genetic polymorphisms such as Antivitamin K, Dose Beta Blockers, Proton pump inhibitor, CYP2C19*17, Clopidogrel active metabolite have an impact in CAD diagnosis. Finally, the reduced data are classified using the new proposed ensemble learning model and, as a result, we found that the proposed CDSS has the highest prediction rates comparing with the two existing ensemble models CSGA+ Bagging and Adaptive boosting on the different datasets. These results demonstrate the effectiveness of ensemble learning models in improving classification performance. For future work, several directions must be considered. First, we will examine the significance of the studied variables by using other feature selection techniques. Then, a fuzzification approach may be introduced to envisage information vagueness and decision-making uncertainty in engineering problems. Finally, we will focus to find way to reduce the computation time problem of the proposed system.

References

1. AHA Statistical Update.: Heart Disease and Stroke Statistics 2010 Update: Summary, A Report from the American Heart Association (2010)
2. Rajkumar, A., Reena, G.S.: Diagnosis of heart disease using datamining algorithm. Global J. Comput. Sci. Technol. **10**(10), 38 (2010)
3. Genders, T.S., Steyerberg, E.W., Alkadhi, H., et al.: A clinical prediction rule for the diagnosis of coronary artery disease: validation, updating, and extension. Eur. Heart J. **32**(11), 1316–1330 (2011)
4. Xu, H., Duan, Z., Miao, C., Geng, S., Jin, Y.: Development of a diagnosis model for coronary artery disease. Indian Heart J. **69**(5), 634–639 (2017)
5. AlHosani, A., AlShizawi, S., AlAli, S., Saleh, H., Assaf, T., Stouraitis, T.: Automatic detection of coronary artery disease (CAD) in an ECG signal. In: 24th IEEE International Conference on Electronics, Circuits and Systems (ICECS) (2017)
6. Martono, G.H., Adji, T.B.: Penggunaan Principal Component Analysis dan Pohon Keputusan untuk Mendeteksi Penyakit Jantung Koroner, Unpublished thesis, Dept. Elect. Eng., Universitas Gadjah Mada, Yogyakarta (2012)
7. Niranjana Devi, Y., Anto, S.: An evolutionary-fuzzy expert system for the diagnosis of coronary artery disease. Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET) **3**4) (2014)
8. Baihaqi, W.M., Setiawan, N.A., Ardiyanto, I.: Rule extraction for fuzzy expert system to diagnose coronary artery disease. In: 1st International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE),Yogyakarta, Indonesia (2016)
9. Miao, K.H., Miao, J.H., Miao, G.J.: Diagnosing coronary heart disease using ensemble machine learning. Int. J. Adv. Comput. Sci. Appl. **7**(10), 30–39 (2016)
10. Samadiani, N., Moameri, S.: Diagnosis of coronary artery disease using cuckoo search and genetic algorithm in single photon emision computed tomography images. In: 7th International Conference on Computer and Knowledge Engineering (ICCKE 2017), 26–27 October 2017

11. Lei, K., Zhang, L., Shen, Y., Huang, X., Wu, J.: Syndromes diagnostic model for coronary artery disease (CAD): an improved naïve bayesian classification model based on attribute relevancy. In: IEEE 2nd International Conference on Big Data Analysis (ICBDA) (2017)
12. Han, J., Kamber, M., Pei, J.: Data Mining Concepts and Techniques. Morgan Kaufmann, Burlington (2011)
13. Shalabi, L.A., Shaaban, Z., Kasasbeh, B.: Data mining: a preprocessing engine. J. Comput. Sci. **2**(9), 735–739 (2006)
14. Gopal Krishna Patro, S., Parimita Sahoo, P., Panda, I., Sahu, K.K.: Technical analysis on financial forecasting. Int. J. Comput. Sci. Eng. **03**(01), 1–6. E-ISSN 2347-2693 (2015)
15. Panda, S.K., Nag, S., Jana, P.K.: A smoothing based task scheduling algorithm for heterogeneous multi-cloud environment. In: 3rd IEEE International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat. IEEE (2014)
16. Kotsiantis, S.B., Pintelas, P.E., Kanellopoulus, D.: Handling imbalanced datasets: a review. In: GESTS International Transactions on Computer Science and Engineering **30** (2006)
17. Wang, S., Yao, X.: Multiclass imbalance problems: analysis and potential solutions. IEEE Trans. Syst. Man Cybern. Part B: Cybern. **42**(4), 1119 (2012)
18. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
19. Wang, L., Ni, M., Zhu, L.: Correlation coefficient of dual hesitant fuzzy sets and its applications. Appl. Math. Model. **38**, 12 (2013).
20. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res. **10**(5), 1205–1224 (2004)
21. Saeys, Y., Inza, I., Larranaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics Advance Access, 24 August 2007
22. Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T.: Numerical Recipes in C. Cambridge University Press, Cambridge (1988)
23. Pradhan, D., Padhy, S., Sahoo, B.: Enzyme classification using multiclass support vector machine and feature subset selection. Comput. Biol. Chem. **70**, 211–219 (2017)
24. Bolón-Canedon, V., Sánchez-Maroño, N., Alonso-Betanzos, A.: Data classification using an ensemble of filters. Neurocomputing **135**, 13–20 (2014)
25. Bashir, S., Qamar, U., Khan, F.H.: IntelliHealth: a medical decision support application using a novel weighted multi-layer classifier ensemble framework. J. Biomed. Informatics **59**, 185–200 (2016)
26. Zhou, L., Lai, K.K., Yu, L.: Least squares support vector machines ensemble models for credit scoring. Expert Syst. Appl. **37**, 127–133 (2010)
27. Yu, L., Lai, K.K., Wang, S., Huang, W.: A bias-variance-complexity trade-off framework for complex system modeling. In: Gavrilova, M. (ed.) ICCSA 2006. LNCS, vol. 3980, pp. 518–527. Springer, Heidelberg (2006). https://doi.org/10.1007/11751540_55
28. Vapnik, V.N.: The Nature of Statistical Learning Theory, Springer, New York (2000). https://doi.org/10.1007/978-1-4757-3264-1
29. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines. Cambridge University Press, Cambridge (2000)
30. Gunn, S.: Support Vector Machines for classification and Regression, Technical report, University of Southampton (1998)
31. Mitchell, T.M.: Machine Learning, 1st edn. McGrawHill, New York (1997)
32. Rokach, L.: Decision forest twenty years of research. Inf. Fusion **27**, 111–125 (2016)
33. Verma, L., Srivastava, S., Negi, P.C.: An intelligent noninvasive model for coronary artery disease detection. Complex Intell. Syst. **4**(1), 11–18 (2018)
34. Sharma, P., Saxena, K., Sharma, R.: Heart disease prediction system evaluation using C4.5 rules and partial tree. In: Behera, H.S., Mohapatra, D.P. (eds.) Computational Intelligence in

Data Mining—Volume 2. AISC, vol. 411, pp. 285–294. Springer, New Delhi (2016). https://doi.org/10.1007/978-81-322-2731-1_26
35. Kinaci, A.C., Yucebas, S.C.: Cost reduction in thyroid diagnosis: a hybrid model with SOM and C4.5 decision trees. In: Arik, S., Huang, T., Lai, W.K., Liu, Q. (eds.) ICONIP 2015. LNCS, vol. 9490, pp. 440–448. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-26535-3_50
36. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., Burlington (1993)
37. Özçift, A.: Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. Comput. Biol. Med. **41**(5), 265–271 (2011)
38. Azar, A.T., Elshazly, H.I., Hassanien, A.E., Elkorany, A.M.: A random forest classifier for lymph diseases: Comput. Meth. Programs Biomed. **113**(2), 465–473 (2014)
39. Ambroise, C., McLachlan, G.J.: Selection bias in gene extraction on the basis of microarray gene-expression data. In: Proceedings of the National Academy of Sciences of the United States of America, vol. 99, no. 10, pp. 6562–6566 (2002)
40. Blake, B.K.-S.C., Merz, C.J.: UCI repository of machine learning databases: Dep. Inf. Comput. Sci. Univ. California, Irvine, CA (1998)
41. Cheung, B.K., Ng, A.C.: An efficient and reliable algorithm for non-smooth nonlinear optimization. Neural Parallel FJ Sci. Comput. **3**, 115–128 (1995)