DIAGNOSTYKA, 20xx, Vol. xx, No. x

ISSN 1641-6414
e-ISSN 2449-5220
DOI:

EFFICIENT HEART DISEASE DIAGNOSIS BASED ON TWIN SUPPORT VECTOR MACHINE

Youcef BRIK, Mohamed DJERIOUI, and Bilal ATTALLAH
LASS Laboratory, Faculty of technology, University Mohamed Boudiaf of M'sila, Algeria.
{youcef.brik, Mohamed.djerioui, bilal.attallah}@univ-msila.dz

Abstract
Heart disease is the leading cause of death in the world according to the World Health Organization (WHO). Researchers are more interested in using machine learning techniques to help medical staff diagnose or detect heart disease early. In this paper, we propose an efficient medical decision support system based on twin support vector machines (Twin-SVM) for heart disease diagnosing with binary target (i.e. presence or absence of disease). Unlike conventional support vector machines (SVM) that finds only one optimal hyper-plane for separating the data points of first class from those of second class, which causes inaccurate decision, Twin-SVM finds two non-parallel hyper-planes so that each one is closer to the first class and is as far from the second class as possible. Our experiments are conducted on real heart disease dataset and many evaluation metrics have been considered to evaluate the performance of the proposed method. Furthermore, a comparison between the proposed method and several well-known classifiers as well as the state-of-the-art methods has been performed. The obtained results proved that our proposed method based on Twin-SVM technique gives promising performances better than the state-of-the-art. This improvement can seriously reduce time, materials, and labor in healthcare services while increasing the final decision accuracy.

Keywords: Heart diseases, medical data, diagnostic, machine learning, twin support vector machines.

1. INTRODUCTION

Heart disease is one of the main reasons for disability and premature death of people in the world. According to the World Health Organization, about 17.9 million deaths have occurred worldwide due to Heart diseases in 2016 [1]. However, some key factors help us reduce the risk of heart disease, such as controlled blood pressure and lower cholesterol [2]. Therefore, the diagnosing of heart disease is a delicate, risky, and very important factor [3]. If done properly it can be used by the medical staff to save life. This process can be realized by exploring the registered patient data. Usually, the existing healthcare systems use electronic health records to store those data [4]. Advances in computer and information technologies can deal with this routine data to make critical medical decisions [5].

Machine learning (ML), which is part of artificial intelligence, is the research domain of algorithms and statistical techniques that build a mathematical model based on sample data in order to make decisions or diagnosis without explicitly programming them. Actually, many researchers have worked on heart diseases prediction/ diagnosing using ML approaches in order to achieve an accurate diagnosis. In [6], several data mining classifiers such as Naïve Bayes, Decision tree, Rule-based and Artificial Neural Network have been examined with different healthcare data including heart disease prediction. Also, Shouman, et al. [7] have applied a range of Decision Tree techniques for retrieving the better performance in heart disease diagnosing. Chaurasia and Pal [8] have explored WEKA data mining tool for heart disease detection. This tool consists of several machine learning algorithms for mining purpose such as: bagging, Naive Bayes, and J48. Bagging has provided better classification results compared to other techniqus.

The authors in [9] have considered two systems based on Artificial Neural Network (ANN) and Neuro-Fuzzy approaches in order to develop an automatic heart disease diagnosis system. Xiong et al. [10] have realized RhythmNet system for the classification of heart disease from single lead electrocardiogram (ECG). This system used a residual convolutional recurrent neural network as classifier.

Furthermore, the authors in [11] have developed heart sound classification using a combination of convolutional neural network (CNN) and majority voting for cardiovascular disease prediction. According to Amine et al. [12], the prediction accuracy of the cardiovascular disease can be significantly improved by combining different features and classification techniques. The best performing classifier achieved by using vote technique, which combines Naïve Bayes and Logistic Regression techniques. Besides, Padmanabhan and his colleagues [13] have proposed an Auto Machine Learning (AutoML) to evaluate cardiovascular disease diagnosing. The performance evaluation of their system has been conducted by using the Auto-Sklearn library.

The authors in [14] have presented a one dimensional deep CNN to classify multiple heart diseases where a modified ECG signal has been considered as an input signal. In [15], an automated diagnostic system for the prediction of heart disease has been proposed. This system used a statistical model for features refinement and Deep Neural Network (DNN) for classification. Sellami et al. [16], have presented a deep CNN based on state-of-the-art deep learning techniques for accurate heartbeat classification using ECG signals.

However, the Deep learning techniques are often a time-consuming and costly procedure in term of parameter optimization. Furthermore, these techniques require a lot of structured data [5] [11]. This disadvantage has been lifted by the SVM technique which is chosen due to, first, its speed in learning phase and its performance [17]. Second, thanks to the structural risk minimization theory, SVM has effectively solved the local minimum problems and high dimensionality. Third, SVM is very powerful tool for solving binary problems [18].

In fact, SVM can perfectly classify binary data by finding the optimal hyper-plane that separates the data points of first class from those of second class. In clinical decision support systems, SVM has attracted many attentions, especially in heart disease diagnosing. In Tan et al. [19], SVM has been joined with Genetic Algorithm using wrapper approach to classify five heart disease data sets. In [20] SVM with many machine learning techniques such as: Baysian Network, Decision tree, Artificial Neural Network, and Fuzzy pattern tree have been used to classify the Cleavland heart disease data set using 10-fold-cross validation. SVM achieved the highest prediction accuracy compared to other classifiers. Otoom et al. [21] have presented a system for Coronary artery disease detection and monitoring where three machine learning techniques are performed such as: Bayes Net, SVM, and Functional Trees. The authors have used WEKA tool for feature selection and detection. SVM has provided the best accuracy with 85.1%.

However, enormous difficulties have been presented in dealing with complex data that is nonlinearly inseparable and unstructured where single hyper-plane cannot efficiently maximize the margin between the classes [22]. Furthermore, SVM is very sensitive to noisy data which makes it predisposed to over-fitting [23], [24]. In order to remedy these drawbacks, Khemchandani and Chandra [25] have proposed a new SVM variant called Twin Support Vector Machine (Twin-SVM) for the binary classification. Unlike traditional SVM, Twin-SVM would find two non-parallel hyper-planes, such that each one is closer to the first class and is as far as possible from the second class. Therefore, Twin-SVM provides lower computational complexity and better generalization ability compared to conventional SVM [26]. All of these advantages make Twin-SVM very adequate for heart disease diagnosing system that contains data of patient records with binary target, i.e. referring to the presence or absence of heart disease. Furthermore, SVM can be learned efficiently for heart disease diagnosing without optimizing a large amount of hyper-parameters [18], [27].

In this paper, we propose a Twin-SVM for a heart disease purpose for butter diagnosis that can be obtained by using two non-parallel hyper-planes. Our work focuses on the following points:

- Set up a system architecture for Heart diseases based on Twin-SVM in order to make an adapted decision to the Heart diseases diagnosing;
- A comparative study between Twin-SVM and other SVM variants;
- A comparison of Twin-SVM results against those of many well-known classifier methods such as Multilayer Perceptron (MLP), Logistic Regression, Decision Trees, Random Forest, and K-Nearest Neighbors (KNN).
- Furthermore, we compare our proposed method with the state-of-the-art methods that used the same datasets, the same experimental protocol, and the same performance measurements.

The remainder of this paper is organized as follows: Section 2 presents the Heart diseases diagnosing system and gives an overview of the conventional SVM and Twin-SVM. In Section 3, we describe the dataset used, the evaluation metrics and discuss the obtained results. In Section 4, we make a meaningful comparison between the proposed method and some well-known classifiers in the diagnosing purpose. Section 5 reports the comparison of the proposed method with the state-of-the-art techniques that used the same heart disease dataset in evaluation. Finally, the conclusions drawn from this work are presented in Section 6.

2. PROPOSED METHOD

The present work was conducted on real heart disease dataset using machine learning techniques. The flowchart given in Fig. 1 presents the proposed heart diseases diagnosing system. We focus in the next on the theoretical background of SVM and Twin-SVM.

2.1. Support Vector Machine

The SVM method proposed by Vapnik has been studied extensively for classification and regression [17], [18]. The SVM algorithm was developed for prediction by using an ε–insensitive loss function.

The goal of SVM is to identify a function $f(x)$ that for all training patterns x has a maximum deviation ε from the target values y and has a maximum margin [24]. The estimating function f is taken in the form:

$$f(x) = w\phi(x) + b, \tag{1}$$

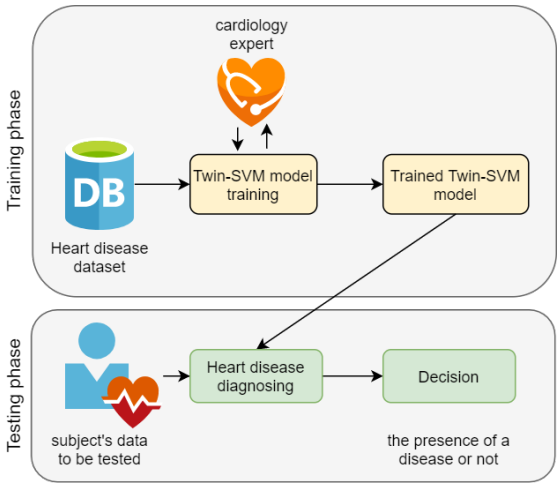


Fig. 1. Heart diseases diagnosing system.

where w and $b$are the coefficients that have to be estimated from data. $\phi(x)$ is the non-linear function in feature space. The objective is to find the values of w and b such that $f(x)$ can be determined by minimizing the following cost function:

$$R(C) = \frac{1}{2}\|w\|^2 + C\frac{1}{N}\sum_{i=1}^{k} L_\varepsilon\,(d_i, y_i), \tag{2}$$

where L_ε is the extension of ε-insensitive loss function defined as:

$$L_\varepsilon(d,y) = \begin{cases} |d-y| - \varepsilon, & |d-y| \geq \varepsilon \\ 0 & otherwise \end{cases} \tag{3}$$

After introducing slack variables, the risk function can be expressed in the following constrained form:

$$\text{Minimise} \quad R(w,\xi^*) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{k}(\xi_i + \xi^*{}_i) \tag{4}$$

With subject to:

$$d_i - w\phi(x_i) - b_i \leq \varepsilon + \xi_i \tag{5}$$

$$w\phi(x) + b - d_i \leq \varepsilon + \xi^*{}_i, \tag{6}$$

where $\xi_i \; and \; \xi^*{}_i \geq 0$.

Solution of the above problem (4) using primal dual method leads to the following dual problem that can be expressed as:

$$Q(\alpha_i, \alpha_i^*) = \sum_{i=1}^{k} d_i\,(\alpha_i - \alpha_i^*) - \varepsilon\sum_{i=1}^{k}(\alpha_i + \alpha_i^*) - \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}(\alpha_i - \alpha_i^*)\,(\alpha_j - \alpha_j^*)K(X_i, X_j) \tag{7}$$

Subject to

$$\sum_{i=1}^{k}(\alpha_i - \alpha_i^*) = 0 \tag{8}$$

$$0 \leq a_i, \alpha_i^* \leq C, i = 1,2,\ldots.n, \tag{9}$$

where a_i and α_i^* are the Lagrange multipliers that act as forces pushing the predictions towards the target value d. The computation in input space can be performed using kernel function in feature space as follows:

$$K(X_i, X_j) = \emptyset(X_i)\emptyset(X_j) \tag{10}$$

Note that any function that satisfies Mercer's theorem [28] can be used as a kernel function. The kernel parameters are user's defined where C controls the smoothness of approximating function and ε determines the margin within which the error is tolerated. Finally, the estimating function can be expressed as:

$$f(x) = \sum_{i=1}^{nsv}(\alpha_i - \alpha_i^*)K(X, X_i) + b, \tag{11}$$

where nsv is the number of support vectors.

We present in Tab. I five well-known types of SVM kernels that we use in this work including linear, polynomial, radial basis function (RBF), sigmoid (hyperbolic tangent), and Laplace kernels.

Table I. Kernel functions used in SVM training.

Kernel name	Mathematical function
Linear	$K(x_i, x_j) = x(i)x(j)$
Polynomial	$K(x_i, x_j) = (\gamma x(i)x(j) + L)^D$
RBF	$K(x_i, x_j) = exp(-\gamma\|x(i) - x(j)\|^2)$
Sigmoid	$K(x_i, x_j) = \tanh(\gamma x(i)x(j) + L)$
Laplace	$K(x_i, x_j) = exp(-\gamma\|x(i) - x(j)\|)$

2.2. Twin Support Vector Machine

Twin-SVM is one of the new emerging machine learning approaches suitable for both classification and regression problems [25]. The target of Twin-SVM is to generate the above two non-parallel hyper-planes in the n-dimensional real space R^n, such that each plane is closer to one of the two classes and is as far as possible from the other [26]. For linear case, the two nonparallel hyper-planes can be formulated as:

$$f_1(x) = (w_1.x) + b_1 = 0 \tag{12}$$

and

$$f_2(x) = (w_2.x) + b_2 = 0, \tag{13}$$

where $w_1, w_2 \in R^n$ are normal vectors and $b_1, b_2 \in R$ are bias terms. The linear classifiers are obtained by solving the following optimization problems.

$$\min_{w_1,b_1,\xi} \frac{1}{2}\|Aw_1 + e_1b_1\|^2 + c_1e_2^T\xi \tag{14}$$

Subject to

$$-(Bw_1 + e_2b_1) + \xi \geq e_2 \tag{15}$$

$$\min_{w_2,b_2,\eta} \frac{1}{2}\|Bw_2 + e_2 b_2\|^2 + c_2 e_1^T \eta \quad (16)$$

Subject to

$$-(Aw_2 + e_1 b_2) + \eta \geq e_1, \quad (17)$$

where c_1 and c_2 are penalty parameters, ξ and η are slack positive factors, e_1 and e_2 are vectors of ones of appropriate dimensions.

By introducing the Lagrangian multipliers, the dual quadratic programming problems (QPPs) of (14) and (16) can be represented as followings

$$\max_{\alpha} e_2^T \alpha - \frac{1}{2}\alpha^T G(H^T H)^{-1} G^T \alpha \quad (18)$$

Subject to

$$0 \leq \alpha \leq c_1 e_2, \quad (19)$$

and

$$\max_{\beta} e_1^T \beta - \frac{1}{2}\beta^T H(G^T G)^{-1} H^T \beta \quad (20)$$

Subject to

$$0 \leq \beta \leq c_2 e_1, \quad (21)$$

where $H = [A\ e_1], G = [B\ e_2]$.

After solving the dual problems (18) and (20), the two nonparallel hyper-planes can be produced by

$$\begin{bmatrix} w_1 \\ b_1 \end{bmatrix} = -(H^T H)^{-1} G^T \alpha, \begin{bmatrix} w_2 \\ b_2 \end{bmatrix} = (G^T G)^{-1} H^T \beta \quad (22)$$

Twin-SVM then can easily assign a label "+1" or "-1" to a testing instance x by

$$Class\ i = \underset{k=1,2}{\operatorname{argmin}} |w_k x^T + b_k|, \quad (23)$$

where $|.|$ is the absolute value.

In order to make Twin-SVM non-linear, the kernel functions reported in Tab.1 can be used to map the original data samples into a new non-linear feature space where the decision function of equation (23) becomes

$$Class\ i = \underset{k=1,2}{\operatorname{argmin}} \left|w_k K(x_i, x_j) + b_k\right| \quad (23)$$

For a new input data, its distance is measured from both kernel surfaces and is assigned to the class from which its distance is smaller.

3. EXPERIMENTAL RESULTS

In this section, we describe the dataset used in this study as well as the different evaluation metrics involved in the performance assessment. Furthermore, we present the result of our proposed method against other SVM variants applied on Heart UCI dataset. Then, we compare our work to the state-of-the-art methods. It should be noted that the evaluation are performed using MATLAB environment on 1.9 GHz CPU processor with 8 GB RAM memory.

3.1. Dataset Description

The Heart UCI dataset has been collected from UCI machine learning repository [29]. This dataset contains in total 303 patient records with 76 attributes for each one, but only 14 of them are used for our evaluation to make our scores comparable to previous works. In particular, the Cleveland dataset is the only one that has been used by ML researchers to this date [6], [7], [12], [13], [22], [23], [30-33]. Tab. II provides a brief description about the selected attributes and their proprieties. The last attribute serves as the prediction target that indicates the absence or presence of heart disease in a patient (0 or 1 value, respectively). Of the 303 records, 138

Table II. Heart disease dataset description.

No.	Attribute	Type	Description	Range of values
1	Age	Continuous	Age in years	29 to 79
2	Sex	Discrete	Gender of the person	0, 1
3	Cp	Discrete	Chest pain type	1, 2, 3, 4
4	Trestbps	Continuous	Resting blood pressure (in mm Hg)	94 to 200
5	Chol	Continuous	Serum cholesterol (in mg/dL)	126 to 564
6	Fbs	Discrete	Fasting blood sugar in mg/dL	0, 1
7	Restecg	Discrete	Resting Electrocardiographic Results	0, 1,2
8	Thalach	Continuous	Maximum Heart Rate Achieved	71 to 202
9	Exang	Discrete	Exercise induced angina	0, 1
10	OldPeak	Continuous	ST depression induced by exercise relative to rest	1 to 3
11	Slope	Discrete	The slope of the peak Exercise ST segment	1, 2, 3
12	Ca	Discrete	Number of major vessels colored by fluoroscopy	0 to 3
13	Thal	Discrete	Nature of defect	3, 6, 7
14	Target	Discrete	Presence or absence of heart disease	0,1

ones are that of patients with target 0 and 165 with target 1.

3.2. Evaluation Metrics

Usually, the accuracy rate is the most performance metric used to evaluate the classifiers such as the proposed model. However, due to the imbalanced nature of our dataset, typical measures such as accuracy or error rates are heavily biased and do not reflect the real performance of the system. For this reason, metrics that are insensitive to the imbalanced set are involved based on the confusion matrix (see Fig. 2).

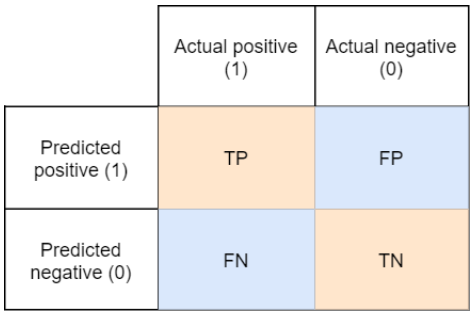


Fig.2. A Generic confusion matrix: TP denotes number of true positives, FP denotes number of false positives, TN denotes number of true negatives and FN denotes number of false negatives.

In our work, we considered other four metrics to properly asses the model performance, such as Sensitivity, Specificity, Matthews Correlation Coefficient, and Balanced accuracy. The formulation of Accuracy metric as well as the other four ones are given as follows:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{24}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{25}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \tag{26}$$

$$\text{Matthews Correlation Coefficient} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{27}$$

$$\text{Balanced accuracy} = (\text{Sensitivity} + \text{Specificity})/2 \tag{28}$$

3.3. Results and discussion

We carried out many experiments to demonstrate the effectiveness of the proposed method. We recall that 5-fold cross validation has been employed to evaluate the performance of our system. The performance of Twin-SVM for heart disease diagnosing was evaluated with the kernel functions mentioned in Tab. I. The obtained results for training and testing phases of Twin-SVM with different kernel functions are reported in Tab. III. It should be noted that all the parameters of the kernel functions are defined empirically according to the loss minimization. Because our data has different ranges, we normalize each column values to fit [0-1] scale without distorting the differences in the ranges of values.

From Tab. III, we observe that the linear kernel-based Twin-SVM outperforms the other Twin-SVM variants in both training and testing accuracies. Furthermore, Twin-SVM with linear kernel consumes less time than other variants. Next, we discuss with more details the obtained results in term of different evaluation metrics mentioned in equations (25), (26), (27) and (28).

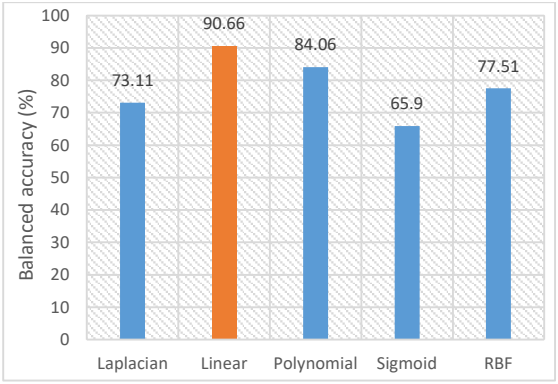


Fig. 3. Twin-SVM performance on balanced accuracy.

Fig. 3 shows the obtained results in term of balanced accuracy. The Linear Twin-SVM shows superior performance than all other variants. In addition, Twin-SVM with Laplacian kernel performs poorer than RBF and Polynomial kernels. Twin-SVM-Sigmoid shows a very bad performance.

Table III. Performance of Twin-SVM with different kernels in training and testing phases.

Kernel	Parameters	Training phase		Testing phase	
		Accuracy (%)	Time (s)	Accuracy (%)	Time (s)
Laplacian	γ =-3	98.75	1.3	69.53	0.03
Linear	-	99.43	1.1	**90.72**	0.009
Polynomial	γ =1, L=1, D=2	90.06	1.7	83.44	0.01
Sigmoid	γ=1, L= -1	72.84	2.2	66.22	0.04
RBF	γ =-1	90.06	2.3	76.15	0.04

Table IV. Twin-SVM against conventional SVM with different kernels.

Method	Parameters	Training phase		Testing phase	
		Accuracy (%)	Time (s)	Accuracy (%)	Time (s)
SVM-Laplacian	γ =-2	86.19	1.5	65.34	0.02
SVM-Linear	-	86.21	0.6	81.20	0.007
SVM-Polynomial	γ =1, L = 1, D = 2	98.67	0.9	77.57	0.008
SVM-Sigmoid	γ=1, L= -3	55.34	2.1	54.45	0.02
SVM-RBF	γ =-1	94.80	1.5	80.19	0.01
Twin-SVM-Linear	-	99.43	1.1	**90.72**	0.009

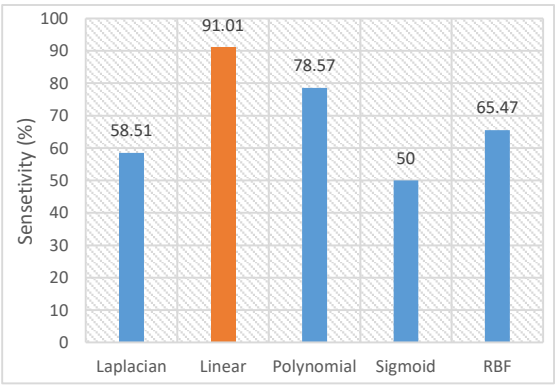


Fig. 4. Twin-SVM performance on Sensitivity rate.

Fig. 4 shows the performance of the proposed method on sensitivity rate with five different kernels. Twin-SVM-Linear shows a better sensitivity value than all considered kernels, with Twin-SVM-Polynomial is a close second. Twin-SVM-Sigmoid provides the worst performance.

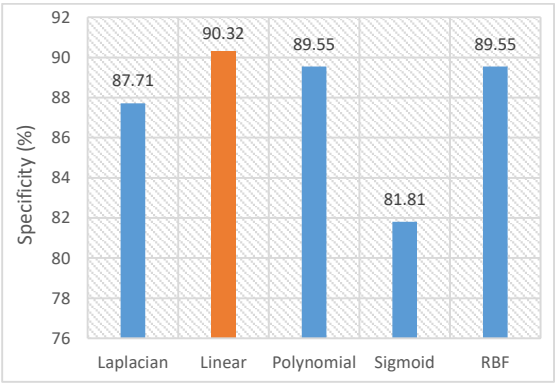


Fig. 5. Twin-SVM performance on Specificity rate.

Fig. 5 shows the performance of the proposed method on specificity rate. Linear based-Twin-SVM shows high specificity Rate with 90.32% followed by Twin-SVM-Polynomial with 89.55%. Twin-SVM-Sigmoid presents constantly the worst performance.

Fig. 6 shows the performance of the proposed method on Matthews Correlation Coefficient rate. We recall here that Matthews Correlation Coefficient is a correlation value between the actual and predicted classes that varies from -1 to +1. A value of +1 means complete identical prediction, 0 is random, -1 means complete disagreement. It is clear that Twin-SVM-Linear far outperforms all other Twin-SVM variants with range of 11.43% compared to the second. Besides, Sigmoid and Laplacian variants demonstrate poor performance.

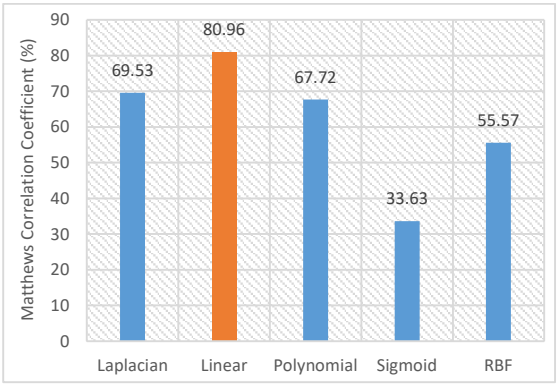


Fig. 6. Twin-SVM performance on Matthews Correlation Coefficient rate.

Now, in order to verify the credibility of these results, we have performed another experiments using conventional SVM with the same kernel functions and the obtained results are depicted in Tab. IV. We can clearly see that Twin-SVM with linear kernel outperforms conventional SVM with different kernels in term of accuracy. Moreover, the conventional SVM with Linear kernel gives more accuracy than other non-linear kernels in less time, which confirm the superiority of linear kernel-based Twin-SVM. Hence, in our case (i.e. heat disease data) Twin-SVM-linear minimizes the empirical risks of training samples so that providing more precise and faster results than conventional SVM with several kernels.

4. COMPARISON WITH OTHER WELL-KNOWN CLASSIFIERS

More importantly, the proposed method is benchmarked with some well-known classifiers such as Multilayer Perceptron (MLP), Logistic Regression, Decision Trees, Random Forest, and k-Nearest Neighbors (kNN). These algorithms have been widely used for automatic medical diagnosis [4], [5], [34]. In order to define the hyperparameters

of each algorithm, we performed the process of trial and errors. For MLP, we used Levenberg-Marquardt function to fit the network with one hidden layer of size 12 neurons. As activation function, the symmetric sigmoid and linear transfer functions was used for hidden and output layers, respectively. In the case of Decision Tree, the maximum depth equal the training sample size minus 1, the minimum sample leaves was 1 and the minimum parent size is 10. For Logistic regression, we used maximum likelihood in fitting the model. However, Random Forest was implemented with Bayesian optimization for tuning its hyperparameters, where the minimum number of observations per leaf was 6 and the number of predictors to sample at each node was 7. Finally, kNN used Hamming distance with three neighbors. As evaluation metrics, we only consider the accuracy and the balanced accuracy since this latter is combination between sensitivity and specificity. The obtained results are shown in Tab. V.

Table V. Comparison results.

Method	Accuracy (%)	Balanced Accuracy (%)
MLP	86.15	86.34
Decision Tree	71.80	69.75
Logistic Regression	81.94	81.73
Random Forest	82.71	80.25
kNN	84.48	85.32
Our method	**90.72**	**90.66**

From Tab. V, we clearly see that the highest accuracy and balanced accuracy achieved are when applying Twin-SVM for diagnosing the heart disease data with 90.72% and 90.66%, respectively. The MLP classifier comes in second class with 86.15% and 86.34% in term of accuracy and balanced accuracy, respectively. The kNN is the third, while the Decision Tree, Logistic Regression and Random Forest classifiers are not competitive. Therefore, Twin-SVM can deal butter with binary classification problem by finding two non-parallel hyper-planes, such that each one is closer to the first class and is as far as possible from the second class. With this improvement, Twin-SVM has lower computational complexity and better generalization ability with linear kernel compared to conventional SVM and its variants.

5. COMPARISON OF STATE-OF-THE-ART METHODS

In order to give an idea on where our proposed method ranks performance-wise, we made a comparison with several state-of-the-art methods that used the same heart disease dataset, the same experimental protocol, and the same performance metrics. The disease diagnosis results obtained for the proposed method and other approaches have been presented in Tab. VI. It is worth noting that this comparison was based only on the accuracy metric because the other evaluation metrics (balanced accuracy, sensitivity, specificity, and Matthews's correlation coefficient) are not available. As observed from Tab. VI, the proposed Twin-SVM based diagnosing model outperforms methods reported in literature. The other techniques also exhibit sensibly good results but are slightly low in terms of prediction accuracy compared to Twin-SVM method.

6. CONCLUSION

In this study, we proposed an effective heart disease diagnosing method based on Twin support vector machines. The performance evaluation of the proposed system was conducted on real cardiovascular disease dataset, which contains clinical data from trial subjects and whether or not they have heart disease. In fact, our system can predict the presence or absence of heart disease with given a new subject's data providing a good accuracy. The proposed diagnostic system demonstrated its superiority on different performance evaluation metrics. This superiority is justified by the ability of Twin-SVM in dealing with complex data (i.e., contains imbalanced continuous and discrete attributes) that is nonlinearly inseparable where single hyper-plane cannot efficiently maximize the margin between the classes. Furthermore, a comparison between the proposed method and several well-known classifiers as well as the state-of-the-art methods has been performed. This comparison proved that our proposed method based on Twin-SVM classifier can significantly give promising performances better than the state-of-the-art in heart disease diagnosing.

Table VI. Comparison of various methods to estimate the diagnosing accuracy on the heart UCI dataset.

Author	Method	Reported accuracy (%)
Wang et al. [22]	Ensemble of SVMs	83.37
Srinivas et al. [6]	Naïve Bayes	83.70
Shouman et al. [7]	Decision tree	84.10
Peter et al. [30]	Multilayer perceptron	82.22
Nahar et al. [31]	Naïve Bayes	69.11
Tomar and Agarwal [23]	Least square SVM	85.59
Ismaeel et al. [32]	Extreme learning machine	80.00
Amin et al. [12]	Logistic regression	78.03
Padmanabhan et al. [13]	Auto machine learning	85.00
Djerioui et al. [33]	Feature selection with SVM	85.43
Our method	T win-SVM	**90.72**

In the future work, we plan to perform some powerful algorithms for selecting the most pertinent features to find which one is more suitable for our purpose. Likewise, it is very interesting to integrate this proposed method in medical diagnostic systems, which can positively provide an economic and life-saving impact in healthcare services.

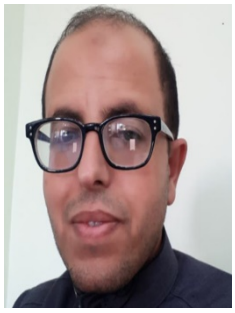
REFERENCES

1. World Health Organization (WHO), 2019. Cardiovascular diseases (CVDs)–Key Facts. http://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).
2. Raza, K. (2019). Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule. In U-Healthcare Monitoring Systems, pp. 179-196. Academic Press.
3. Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., ... & Turner, M. B. (2015). Heart disease and stroke statistics—2015 update: a report from the American Heart Association. Circulation, 131(4), e29-e322.
4. Desai, R. J., Wang, S. V., Vaduganathan, M., Evers, T., & Schneeweiss, S. (2020). Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. JAMA network open, 3(1), e1918962-e1918962.
5. Awaysheh, A., Wilcke, J., Elvinger, F., Rees, L., Fan, W., & Zimmerman, K. L. (2019). Review of medical decision support and machine-learning methods. Veterinary pathology, 56(4), 512-525.
6. Srinivas, K.; Rani, B.K.; Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. Int. J. Comput. Sci. Eng. (IJCSE), 2, pp. 250–255.
7. Shouman, M., Turner, T., & Stocker, R. (2011). Using decision tree for diagnosing heart disease patients. In Proceedings of the Ninth Australasian Data Mining Conference-Volume 121, pp. 23-30.
8. Chaurasia, V., & Pal, S. (2013). Early prediction of heart diseases using data mining techniques. Caribbean Journal of Science and Technology, 1, pp. 208-217.
9. Abushariah, M. A., Alqudah, A. A., Adwan, O. Y., & Yousef, R. M. (2014). Automatic heart disease diagnosis system based on artificial neural network (ANN) and adaptive neuro-fuzzy inference systems (ANFIS) approaches. Journal of software engineering and applications, 7(12), 1055.
10. Xiong, Z., Nash, M. P., Cheng, E., Fedorov, V. V., Stiles, M. K., & Zhao, J. (2018). ECG signal classification for the detection of cardiac arrhythmias using a convolutional recurrent neural network. Physiological measurement, 39(9), 094006.
11. Xiao, B., Xu, Y., Bi, X., Zhang, J., & Ma, X. (2020). Heart sounds classification using a novel 1-D convolutional neural network with extremely low parameter consumption. Neurocomputing, 392, pp. 153-159.
12. Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics, 36, pp. 82-93.
13. Padmanabhan, M., Yuan, P., Chada, G., & Nguyen, H. V. (2019). Physician-friendly machine learning: A case study with cardiovascular disease risk prediction. Journal of clinical medicine, 8(7), 1050.
14. Hasan, N. I., & Bhattacharjee, A. (2019). Deep learning approach to cardiovascular disease classification employing modified ECG signal from empirical mode decomposition. Biomedical Signal Processing and Control, 52, pp. 128-140.
15. Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., & Khan, J. A. (2019). An automated diagnostic system for heart disease prediction based on χ^2 statistical model and optimally configured deep neural network. IEEE Access, 7, pp. 34938-34945.
16. Sellami, A., & Hwang, H. (2019). A robust deep convolutional neural network with batch-weighted loss for heartbeat classification. Expert Systems with Applications, 122, pp. 75-84.
17. Vapnik, V. (2013). The nature of statistical learning theory. Springer science & business media.
18. Scholkopf, B., & Smola, A. J. (2018). Learning with kernels: support vector machines, regularization, optimization, and beyond. Adaptive Computation and Machine Learning series.
19. Tan, K. C., Teoh, E. J., Yu, Q., & Goh, K. C. (2009). A hybrid evolutionary algorithm for attribute selection in data mining. Expert Systems with Applications, 36(4), pp. 8616-8630.
20. Bouali, H. and Akaichi, J. (2014). Comparative study of different classification techniques: heart disease use case. In: 2014 13th International Conference on Machine Learning and Applications. pp. 482–486.
21. Otoom, A.F., Abdallah, E.E., Kilani, Y., Kefaye, A., Ashour, M. (2015). Effective diagnosis and monitoring of heart disease. International Journal of Software Engineering and Its Applications, 9(1), pp. 143–156.
22. Wang, S.J., Mathew, A., Chen, Y., Xi, L.F. (2009). Ma, L.; Lee, J. Empirical analysis of support vector machine ensemble classifiers. Expert Syst. Appl., Vol. 36, pp. 6466–6476.
23. Tomar, D. and Agarwal, S. (2014). Feature selection based least square twin support vector machine for diagnosis of heart disease. Int. J. Bio-Sci. Bio-Technol, Vol. 6, pp. 69–82.
24. Tang, L., Tian, Y., & Pardalos, P. M. (2019). A novel perspective on multiclass classification: Regular simplex support vector machine. Information Sciences, 480, 324-338.
25. Khemchandani, R., & Chandra, S. (2007). Twin support vector machines for pattern classification. IEEE Transactions on pattern analysis and machine intelligence, 29(5), 905-910.
26. Tanveer, M., Sharma, A., & Suganthan, P. N. (2019). General twin support vector machine with pinball loss function. Information Sciences, 494, 311-327.
27. Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and computing, 14(3), 199-222.
28. Steinwart, I., & Scovel, C. (2012). Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. Constructive Approximation, 35(3), 363-417.
29. Uci data homepage. [Online]. Available: Https: //archive.ics.uci.edu/ml/datasets/Heart+Disease
30. Peter, T. J., & Somasundaram, K. (2012). An empirical study on prediction of heart disease using classification data mining techniques. In IEEE-International conference on advances in engineering, science and management (ICAESM-2012), pp. 514-518.
31. Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P. (2013). Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. Expert Systems with Applications, 40(1), pp. 96-104.

32. Ismaeel, S., Miri, A., & Chourishi, D. (2015). Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis. In 2015 IEEE Canada International Humanitarian Technology Conference (IHTC2015), pp. 1-3.
33. Djerioui, M., Brik, Y., Ladjal, M., Attallah, B. (2019). Neighborhood component analysis and support vector machines for heart disease prediction. Journal of Ingénierie des Systèmes d'Information, Vol. 24, No. 6, pp. 591-595. https://doi.org/10.18280/isi.240605
34. Richens, J. G., Lee, C. M., & Johri, S. (2020). Improving the accuracy of medical diagnosis with causal machine learning. Nature communications, 11(1), 1-9.

Youcef Brik was born in Algeria in 1984. He received his BEng degree in electronics from the University of M'sila, Algeria, in 2007. Then, he received his Magister and Ph.D degrees in signal and image processing from the Faculty of Electronic and Computer Science, University of Sciences and Technology Houari Boumediene, Algiers, Algeria, in 2010 and 2019, respectively. From 2012 to 2013, he was a Research Assistant with the systems architectures and multimedia division in CDTA (Algeria). Since Dec. 2013, he has been an Associate Professor with the Electronics Department, M'sila University, Algeria. His research interests include information retrieval, computer vision, machine learning, and healthcare informatics. Dr. Youcef Brik receipt the exceptional national program scholarship between 2015 and 2017 to finalize his P.hD research in MOIVRE laboratory, Université de Sherbrook, Canada.

Mohamed Djerioui was born in Algeria in 1979. He received his BEng degree in electronics from the University of M'sila, Algeria, in 2002. Then, he received his Magister and Ph.D degrees in industrial control from the same university in 2007 and 2019, respectively. Since 2009, he is an Associate Professor with the Electronics Department, M'sila University, Algeria. His research interests include systems control, computer vision, machine learning, and healthcare informatics.



Bilal Attallah was born in Algeria in 1985. He received his BEng degree in electronics from the University of M'sila, Algeria, in 2008. Then, he received his Magister and Ph. D degrees in signal and systems from the University of Sciences and Technology Houari Boumediene, Algiers, Algeria, in 2012 and 2018, respectively. Since Jan. 2014, he has been an Associate Professor with the Electronics Department, M'sila University, Algeria. His research interests include Biometrics, computer vision and machine learning.