Decision Support System for Diagnosis of Heart Disease using PCA and SVM Classifier

Gagan kumar[1], Rohit Kalra[2]

[1]Assistant Professor, Modern Institute of engineering & technology
[2]Research scholar, Modern Institute of engineering & technology

Abstract: The accurate diagnosis of life-threatening diseases such as heart disease is a very crucial task in medical science. The humans and computers can be integrated together to achieve best results for correct diagnosis of diseases by balancing the knowledge of human experts in related domains with the vast search potential of computers. Computer based decision support system can play an important role in accurate and timely diagnosis. Machine learning automatically learns through experience and performance of algorithm gets improved with each experience. In this paper, we have developed a decision support system for diagnosing heart disease using PCA and SVM. PCA can achieve high dimensionality reduction with usually lower noise than the original data pattern. The results obtained demonstrate that proposed decision support system predicts the disease of new patients with higher accuracy.

Keywords: SVM, Machine learning, Classifier, Heart Disease, Diagnosis

I. INTRODUCTION

The Heart is the center of circulatory system and is treated as most crucial organ in the human body as it pumps the blood to different parts of the human body through a network of blood vessels, supplying a constant supply of oxygen as well as other vital nutritional components. Many other organs may collapse without its proper functioning. If the heart ever stops functioning and ceases to pump blood, the body will shut down and within very less time a person will expire. According to the Heart Disease and Stroke Statistics Update [1], cardiovascular disease is the leading worldwide cause of death, accounting for 17.3 million deaths per year and by 2030 number of deaths will increase to 23.6 million. The count of people dying every year from cardiovascular disease is increasing drastically.

Machine learning presents various algorithms for analysis of medical data. It helps in diagnosis and prediction of healthcare problems untimely. Patient data is gathered with the help of data collection equipment and stored in a computer system in the form of medical records for treatment. Machine learning algorithms help in the diagnosis process of a new patient by analyzing the data pattern of the patient admitted in the past. It examines the disease, symptoms faced and the adequate treatment provided to the patient and uses that information for a newly admitted patient. Machine learning has attained notable results and can be successfully used in the healthcare industry.

In this paper, we have designed and developed a heart disease prediction system that is highly precise, efficient and useful in early diagnosis which lessens the patient mortality rate. The proposed system is based on Support vector machine (SVM) for the accurate classification of heart disease. This paper is structured as follows: Section II gives an overview of literature. Section III presents the heart disease diagnosis system. Section IV provides simulation results. Section V concludes the paper and discusses future scope.

II. LITERATURE REVIEW

Many machine learning and data mining algorithms have been discussed in literature for prediction and diagnosis of various diseases. Zhang et al. [2] proposed an efficient coronary heart disease prediction system using Support Vector Machine. In this, Principal Component Analysis (PCA) was used to extract the important features and different kernel functions were utilized as a classifier. The highest classification accuracy is achieved with Radial Basis Function (RBF). To find the optimal parameters values, Grid search method was employed and optimal values were found to be c=1 and g=0.0909. The highest classification accuracy reached is 88.6364%. It was used for prediction of two classes.

Naib et al. [3] suggested classification system of primary tumors using multiclass classifier with Random Forest. The dataset comprises of total 22 classes of tumor. The classification is performed with different machine learning algorithms. The result shows that random forest with 10 random trees outperforms with the accuracy of 85.7% and ROC area of 0.997.

Ismail et al. [4] presented a classification approach called GA-SVM for lymph disease diagnosis in which genetic algorithm (GA) is used to reduce the number of features of the dataset from 18 features to 6 features. The experiments were performed with 10-fold cross validation. Different kernel functions were employed and for each function,

performance was evaluated by measures like accuracy, sensitivity, area under curve (AUC), F-measure. The result indicates that GA-linear classifier achieved best results of 83.1% accuracy with 82.6% sensitivity, 82.7% F-measure and 84.9% AUC.

Bascil et al. [5] presented a comparative analysis of methods used in the hepatitis disease diagnosis. The dataset comprises of 155 instances and 19 features. The system is applicable for classification of two classes that are die and live. The dataset is taken from UCI data repository. In this study, probabilistic neural network (PNN) was proposed using 10 fold cross validation technique. The LDA-ANFIS structure [6] obtained the best results followed by FS-FUZZY-AIRS [7]. The PNN approach can be used effectively in the prediction of hepatitis disease. Decision trees are prone to overfitting of data and may not be able to generalize well due to the presence of noise in the training data. This problem can be solved by SVM. SVM are less prone to overfitting because of the presence of regularization parameter.

III. PROPOSED DIAGNOSIS SYSTEM

This section described the proposed diagnosis model for heart disease prediction. The proposed system is based on Support vector machine (SVM) for the accurate classification of heart disease. The proposed system consists of following 6 steps: Selecting and pre-processing data set, normalizing, applying PCA for dimensionality reduction, K-fold for selecting training and testing set and SVM as binary classifier.

a. Selecting and Pre-processing the data set: The Cleveland Heart Dataset is taken from UCI Machine Learning Dataset Repository which was contributed by Detrano [8]. The dataset comprises of 297 instances and 14 attributes of disease as shown in Table 1.

Table 1. Attributes in Cleveland Heart Dataset

Sr No.	Attribute Name	Attribute Description
1.	age	age in years
2.	sex	sex (1 = male; 0 = female)
3.	cp	chest pain type --Value 1: typical angina --Value 2: atypical angina --Value 3: non-anginal pain -- Value 4: asymptomatic
4.	trestbps	resting blood pressure (in mm Hg on admission to the hospital)
5.	chol	serum cholestoral in mg/dl
6.	fbs	fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7.	restecg	resting electrocardiographic results -- Value 0: normal -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8.	thalach	maximum heart rate achieved
9.	exang	exercise induced angina (1 = yes; 0 = no)
10.	oldpeak	ST depression induced by exercise relative to rest
11.	slope	slope: the slope of the peak exercise ST segment -- Value 1: upsloping -- Value 2: flat -- Value 3: downsloping
12.	ca	number of major vessels (0-3) colored by flourosopy
13.	thal	3 = normal; 6 = fixed defect; 7 = reversable defect
14.	num	Predicted attribute healthy or diseased

Data cleansing (or pre-processing) includes dealing with missing values, purging of redundant information, removing inconsistencies and errors which make the quality of data better and efficient to find useful patterns from the data. It is a time-consuming step and very important step because the solution is highly affected by the quality of data. It also

converts continuous valued variables to discrete values using discretization. This method was applied to reduce distinct values of continuous valued variables by allowing to have limited numbers of labels to represent the original variables.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
2	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
3	67	1	4	160	286	0	2	108	1	1.5	2	3	3	1
4	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
5	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
6	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
7	56	1	2	120	236	0	0	178	0	0.8	1	0	3	0

Figure 2. Pre-processed Cleveland Heart Dataset

b. Normalizing the data set: The Cleveland Heart Dataset consists of various attributes having different units and scales. For example, thalach ranges from 71 to 202 while the fbs being 0 or 1, age ranges from 29 to 77, resting blood pressure is in mm Hg and the cholesterol is in mg/dl ranges from 126 to 564. Normalization makes the data scalable into a small specific numeric range to have fair comparison. The dataset after normalization of values is shown in Figure 3.

If $data = (d_1, d_2, \ldots . d_k)$ are the data points, bsxfun in MATALB will normalize the dataset using the following method:

$$Normalized\ data\ (n_i) = \frac{d_i - mean(d)}{var(x)}$$

where

d_i= Data point i where $1 \leq i \leq k$

$mean(d)$= The average of all the data values

$var(d)$ = The sample deviation of all the data values

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
0.934603	0.68993	-2.23685	0.749116	-0.27598	2.426332	1.008496	0.017465	-0.69525	1.067164	2.26033	-0.72076	0.654772
1.376605	0.68993	0.872408	1.593577	0.743301	-0.41076	1.008496	-1.81327	1.433497	0.38113	0.642696	2.474249	-0.89271
1.376605	0.68993	0.872408	-0.65832	-0.3529	-0.41076	1.008496	-0.8979	1.433497	1.324427	0.642696	1.409246	1.170601
-1.93841	0.68993	-0.16401	-0.09535	0.050961	-0.41076	-1.00173	1.630258	-0.69525	2.096215	2.26033	-0.72076	-0.89271
-1.49641	-1.44454	-1.20043	-0.09535	-0.8337	-0.41076	1.008496	0.976423	-0.69525	0.295376	-0.97494	-0.72076	-0.89271
0.1611	0.68993	-1.20043	-0.65832	-0.21828	-0.41076	-1.00173	1.237957	-0.69525	-0.21915	-0.97494	-0.72076	-0.89271

Figure 3. Normalized Cleveland Heart Dataset

c. Applying PCA: Principal Component Analysis (PCA) is used for dimensionality reduction i.e. to select the subset of features which best reflects the original heart dataset. Each feature has its own contribution. Some features are more significant to others while some features are irrelevant and add no useful information to the data which degrades the efficiency of the system. Moreover, high dimension of data results in more computation cost. So, there is a need to reduce the dimensions without affecting the quality of data. The goal of PCA is to transform a number of correlated variables of a dataset to a new set of a small number of variables which are linear combinations of original variables called Principal Components [9]. The original dataset is replaced by its principal components after the application of PCA. The pseudocode for PCA is given in Figure 4 and principal component score in figure 5.

Algorithm for PCA

Input: The input data matrix X of size $N \times D$ where N is the number of instances and D is the number of dimensions or components.

Output: Principal Components coefficients, Score, Latent

Principal component coefficients, returned as a D X D matrix. Each column of coeff contains coefficients for one principal component. Principal component scores, is a N X D matrix where rows of score correspond to instances, and columns to number of components. The column vector, latent, stores the variances of the D principal components i.e. the eigenvalues of the covariance matrix of X. The columns in coeff matrix are in the order of descending component variance.

Procedure:

1. Calculate and subtract the mean in every dimension d of the dataset to centralize the data.
2. Construct the covariance matrix Cov of d*d as:

$$Cov = \frac{1}{N}\sum_{p=1}^{N}(x_p - \mu)\,(x_p - \mu)^T$$

where $\{x_p,\ p = 1,2, \ldots . N\}$ is given N input data records with mean μ.

3. Calculate the eigen values $(\lambda_1, \lambda_2, \ldots \lambda_D)$ and $(e_1, e_2, \ldots e_D)$ eigen vectors from the covariance matrix Cov such that

$$\lambda \times e = Cov \times e$$

4. Choose the M eigen vectors corresponding to m largest eigen values where M ≤ D.
5. Compute the $D \times M$ dimensional matrix W from the above selected m eigen vectors where eigen vectors are represented by columns.
6. The original dataset X is transformed via W onto M-dimensional new subspace Y.

$$y = W^T \times x$$

where x is a $D \times 1$dimensional vector representing one data record and y is transformed $M \times 1$ dimensional vector representing data record in the new subspace Y.

Figure 4. Pseudocode for PCA

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
2	1.128759	-1.08582	3.158932	2.289188	0.023136	0.577839	0.663733	-0.53588	-1.49287	-0.49864	0.337016	0.478775	0.1404
3	3.18555	-1.4155	-0.53282	-0.85652	-0.00628	0.744092	-0.25863	1.067975	0.341947	1.429097	-1.14333	-0.88992	1.020542
4	3.119075	0.655901	-0.28465	-0.62558	0.152536	1.128275	-0.32445	0.208947	0.043132	0.461525	0.430938	0.861774	-0.202
5	-0.48352	1.408595	0.397135	2.827968	0.720094	-0.38771	-0.52134	-2.1499	0.758798	0.227994	-1.54246	0.342382	0.587389
6	-2.28069	-0.32948	-0.07214	1.207281	0.769536	0.624533	0.378927	0.014711	1.048611	0.627052	0.795867	-0.3224	-1.14841
7	-2.201	0.344498	0.4914	0.013841	-0.3796	0.030723	-0.92618	0.013119	-0.54561	0.47407	-0.60452	0.87946	-0.24082
8	1.919177	-1.68264	-0.92105	1.935805	0.071256	1.219604	-0.4568	-1.37547	1.673323	-0.02712	-0.45354	0.669155	0.522965

Figure 5. Principal Component Scores

d. Selecting training and testing set: The 10-fold cross validation method is used for selecting the training and testing set. In 10-fold cross validation, the complete dataset is randomly divided into 10 mutually exclusive subsets of approximately equal size. The classification model is trained and tested 10 times but tested on different fold each time to reduce the bias associated with hold-out method. It is normally trained on nine folds and tested on the remaining single fold.

e. Classifying using SVM: Support Vector Machine is a supervised method of classification invented by Vladimir Vapnik and Chervonenkis in 1963 and proposed as a kernel based learning method for classification of non linear data in 1993 [10]. Support vector machines (SVM) are binary classifiers which can be applied to linearly separable datasets [11]. A classifier is implemented to classify the data into their respective classes. Classification mainly includes two phases. The first phase is the training step and building classifier in which a classifier is trained to analyze the given data records and the class with which they are associated. It analyzes the pattern in the training set. The second phase is the testing step in which model classifies the test dataset on the basis of pattern analyzed in the first step. SVM divide the dataset into two classes using a hyperplane. Hyperplane is the decision surface that separates the data from two classes in such a manner that data of one class are on one side of the hyperplane and of other class are on other side.

Let the dataset be given as {X, Y} where

$X = \{x_1, x_2 \ldots \ldots x_n\}$ represents a set of n training tuples
$Y = \{y_1, y_2 \ldots \ldots y_n\}$ represents associated class label of training tuple
Each y_i belongs to either +1 or -1, that corresponds to two classes of dataset.

$$y_i \in \{+1, -1\}$$

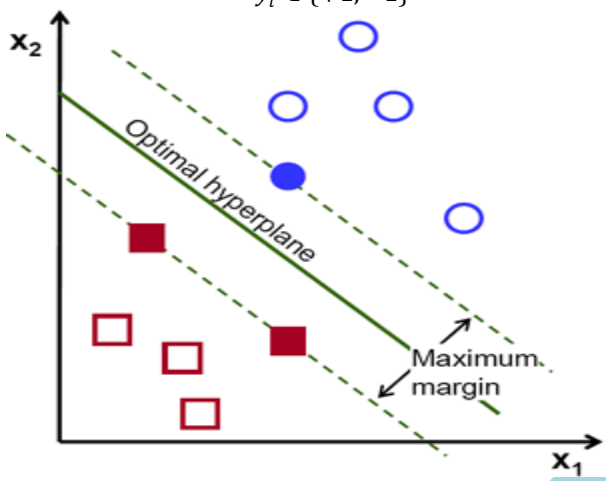

Figure 2.6: Decision boundary and margins of SVM
The hyperpane can be formally described as:

$$f(x) = \beta_0 + \beta^T \mathrm{x}$$

where β is known as the *weight vector* and β_0 as the *bias*.
The optimal hyperplane can be represented in an infinite number of different ways by scaling of β and β_0. Among all the possible representations of the hyperplane, the one chosen is

$$|\beta_0 + \beta^T \mathrm{x}| = 1$$

where x symbolizes the training examples closest to the hyperplane. This hyperlane has the largest margin between two decision boundaries. In general, the training examples that are closest to the hyperplane are called support vectors. The distance between a point x and a hyperplane (β, β_0):

$$Distance = \frac{|\beta_0 + \beta^T x|}{||\beta||}$$

In particular, for the hyperplane, the numerator is equal to one and the distance to the support vectors is

$$Distance_{Support\ \ Vectors} = \frac{|\beta_0 + \beta^T x|}{||\beta||} = \frac{1}{||\beta||}$$

The margin between the hyperplanes denoted as M, is twice the distance to the closest examples:

$$M = \frac{2}{||\beta||}$$

The problem of maximizing M is equivalent to the problem of minimizing a function $L(\beta)$ subject to some constraints. The constraints model the requirement for the hyperplane to classify correctly all the training examples x_i. Formally,

$$\min L(\beta) = \frac{||\beta||^2}{2} \ subject\ to\ y_i(\beta_0 + \beta^T x) \geq 1 \quad \forall i$$

This is a problem of Lagrangian optimization that can be solved using Lagrange multipliers to obtain the weight vector and the bias of the optimal hyperplane. We have used Sequential Minimal Optimization (SMO) [12] for solving a large quadratic programming problem encountered while training SVM. SMO breaks large quadratic programming problem into multiple small quadratic programming problems that are solved analytically. It consumes less memory and suits well for large training sets.
SVMs can also be used non-linearly by mapping the data to a higher dimensional space, thus making the data separable. This mapping is done by a kernel function. SVMs perform well with large feature spaces, as long as the data is separable with a wide margin. They also do well with sparse datasets.

IV. SIMULATION RESULTS AND DISCUSSIONS

The proposed system is implemented by using MATLAB simulator. We have used radial basis as kernel function for SVM and 10-fold cross validation for dividing data set into training and testing set. A confusion matrix obtained illustrates the accuracy of the solution to a classification problem. Given 2 classes, a confusion matrix is a 2 X 2 matrix, where C[i, j] indicates the number of tuples from dataset of class i that were assigned to class C[i, j]. The ideal solution will have only zero in non-diagonal entries.
Table 2. Confusion Matrix entries

		Predicted value	
		Negative	**Positive**
Actual Value	**Negative**	TN	FP
	Positive	FN	TP

Where,

- True positive (TP_i) for a particular class is the number of positive cases that were correctly identified.
- False positive (FP_i) for a particular class is the number of negatives cases that were incorrectly classified as positive.
- True negative (TN_i) for a particular class is the number of negatives cases that were classified correctly.
- False negative (FN_i) for a particular class is the number of positives cases that were incorrectly classified as negative.

The performance of proposed system is evaluated in terms of accuracy, precision and recall using the above parameters.

The overall accuracy is the proportion of the total number of predictions that were correct.

$$\text{Accuracy} = \frac{\sum_{i=1}^{2} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}}{2}$$

The overall precision is the proportion of the predicted positive cases that were correct.

$$\text{Precision} = \frac{\sum_{i=1}^{2} \frac{TP_i}{TP_i + FP_i}}{2}$$

The overall specificity is the proportion of the predicted negative cases that were correctly identified.

$$\text{Specificity} = \frac{\sum_{i=1}^{2} \frac{TN_i}{TN_i + FP_i}}{2}$$

The overall recall or sensitivity is the proportion of positive predicted samples that were correctly identified.

$$\text{Recall} = \frac{\sum_{i=1}^{N} \frac{TP_i}{TP_i + FN_i}}{N}$$

Table 3: A sample confusion matrix for SVM classifer

		Predicted value	
		No disease (0)	**Disease (1)**
Actual Value	**No disease (0)**	12	4
	Disease (1)	0	13

TP for class 0 means the person is not suffering from disease and test also says no disease = 12
TN for class 0 means person is having disease and test also detects disease = 13
FP for class 0 means test predicts no but person has disease = 0
FN for class 0 means test predicts yes and person is not having disease = 4

$$Accuracy\ for\ class\ 0 = \frac{12+13}{12+13+4+0} = 0.8621$$

$$Precision\ for\ class\ 0 = \frac{12}{12+0} = 1$$

$$Recall\ for\ class\ 0 = \frac{12}{12+4} = 0.75$$

$$Specificity\ for\ class\ 0 = \frac{13}{13+0} = 1$$

TP for class 1 means the person is having disease and test also predicts yes = 13
TN for class 1 means person is not having disease and test also predicted no = 12
FP for class 1 means test predicts yes but person doesn't has disease = 4
FN for class 1 means test predicts no and person is having disease = 0

$$Accuracy\ for\ class\ 1 = \frac{12+13}{12+13+4+0} = 0.8621$$

$$Precision\ for\ class\ 1 = \frac{13}{13+4} = 0.7647$$

$$Recall\ for\ class\ 1 = \frac{13}{13+0} = 1$$

$$Specificity\ for\ class\ 1 = \frac{12}{12+4} = 0.75$$

Overall accuracy of system = 0.8621
Overall Precision of system = 0.8824
Overall Recall of system = 0.875
Overall Specificity of system = 0.875

Table 4 gives final values of accuracy, precision, recall and specificity for 10-fold cross validation after taking average of results from 10 different folds. The overall accuracy varies with principal components considered. With 8 principal components, an accuracy of 97% is achieved which decreases to 94.28% with 10 components. The variability of the data can be captured by a relatively small number of PCs, and, as a result, 99% accuracy is achieved with 6 PC's using SVM classifier.

Table 4. Final Parameters of proposed system after 10-fold cross-validation

Principal Components	Accuracy	Precision	Recall	Specificity
6	0.9966	0.9965	0.9969	0.9969
7	0.9864	0.9867	0.9875	0.9875
8	0.97	0.9709	0.9719	0.9719
10	0.9428	0.9489	0.9469	0.9469
12	0.9054	0.9157	0.9118	0.9118
13	0.8955	0.9053	0.9015	0.9015

V. CONCLUSION AND FUTURE SCOPE

Heart disease is a fatal disease and misdiagnosis of this disease can cause life threatening complications such as heart attack and death. This study showed that PCA and SVM can be used efficiently to model and predict heart disease cases. SMO is used for solving quadratic programming for determining parameter for SVM. It consumes less memory and performs well with large data sets. The outcome of this study can be used as an assistant tool by cardiologists to help them to make more consistent diagnosis of heart disease. SVM are less prone to overfitting because of the presence of regularization parameter. The parameters of SVM are. Furthermore, the resulting model has a high specificity rate which makes it a handy tool for junior cardiologists to screen out patients who have a high probability of having the disease and transfer those patients to senior cardiologists for further analysis. The variability of the data can be captured by a relatively small number of principal components, and, as a result, 99% accuracy is achieved with 6 components.
Missing values, noisy data, inconsistent data, and outliers pose a great challenge in the data mining process. Therefore, statistical and machine learning techniques should be applied to control the overall quality of the data. Future work also involves optimization of SVM parameters with other methods such as scatter search method, ant colony optimization etc and comparing results with our proposed algorithm.

REFERENCES

[1]. S. Goenka et al.,"Preventing cardiovascular disease in India-translating evidence to action,'' *Current science*, vol. 97, no. 3, pp. 367– 377, 2009.

[2]. Y. Zhang et al. , "Studies on application of Support Vector Machine in diagnose of coronary heart disease," *Electromagnetic Field Problems and Applications 2012 Sixth International Conference (ICEF)*, Dalian, IEEE 2012.

[3]. M. Naib, "Predicting Primary Tumors using Multiclass Classifier Approach of Data Mining," *International Journal of Computer Applications* ,vol. 96, no. 8, pp. 9–13, 2014.

[4]. H. I. Elshazly, A. M. Elkorany, and A. E. Hassanien, "Lymph diseases diagnosis approach based on support vector machines with different kernel functions," *Computer Engineering & Systems 9th International Conference (ICCES)*, Cairo, pp. 198–203, 2014.

[5]. M. S. Bascil and H. Oztekin, "A study on hepatitis disease diagnosis using probabilistic neural network," *Journal of medical systems*, vol. 36, no. 3, pp. 1603–1606, 2012.

[6]. E. Dogantekin, A. Dogantekin, and D. Avci, "Automatic hepatitis diagnosis system based on Linear Discriminant Analysis and Adaptive Network based on Fuzzy Inference System," *Expert Systems with Appications*, vol. 36, no. 8, pp. 11282–11286, Elsevier 2009.

[7]. K. Polat and S. Güneş, "Medical decision support system based on artificial immune recognition immune system (AIRS), fuzzy weighted pre-processing and feature selection," *Expert Systems with Applications*, vol. 33, no. 2, pp. 484–490, Elsevier 2007.

[8]. "Cleveland dataset." [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data. [Accessed: Jun. 7, 2016].

[9]. A. Ilin and T. Raiko, "Practical approaches to principal component analysis in the presence of missing values," *Journal of Machine Learning Research*, vol. 11, pp. 1957–2000, 2010.

[10]. C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[11]. S. Ding, Z. Shi, D. Tao, and B. An, Recent advances in Support Vector Machines. *Neurocomputing*, 2016 [In Press].

[12]. R.E. Fan, P.H. Chen, and C.J. Lin, C.J., "Working set selection using second order information for training support vector machines", *Journal of machine learning research*, vol. *6*, pp.1889-1918.

