Prediction Of Coronary Artery Disease Using Core Principal Component Analysis Based Support Vector Machine

Omprakash Subramaniam, Dr. Ravichandran Mylswamy

Abstract—Data mining plays vital role in many fields. In medical field, the usage of data mining is getting increased day by day to predict the disease and classify its severity. Coronary Artery Disease (CAD) is becoming major reason for sudden death, which is getting increased in the South Asian countries. Hence, there exist a need to predict CAD by utilizing the patients history by using data mining algorithms. In order to solve this issue, this paper proposes a core framework for finding the indicators and fixing the thresholds to classify the patterns in the dataset; it utilizes the feature based mechanism which integrate principal-component-analysis (PCA) and support-vector-machine (SVM) for productive detection of patterns in the dataset. In dataset multiple features may be available, where few or more features might not be used in classification, even if used it may reduce the classification accuracy. The proposed classification algorithm wisely eliminates the features that are not required for performing the classification by introducing the new features. The results shows that the proposed algorithm outperforms the existing algorithms with 97.34%.

Keywords—CAD, classification, feature selection, PCA, SVM

1. INTRODUCTION

Data mining is the process of seeking a important information from the available data which is stored in the repositories. Pattern recognition, statistical and mathematical methods are utilized for processing the data. It can also be said as; Data mining is the process of observing the dataset in order to find the relationship that is not identified before, and performing the summarization which will be helpful for the data owner. The intention of data mining is not to use for a specific field like computer science, it can be applied to different fields like marketing, engineering, education and even in medical field. In medical field, data mining is utilized to classify the patients affected by disease and it is used to predict the future diseases for the persons. Analyzing the multidimensional data for reducing the features by using the principal component analysis (PCA) is one of the superior cum widely used method. The main concept of PCA is to minimize the dimension of actual data by seeking a novel variables which will be lesser than the actual variables, but still remains with majority of the sample information. The variation that are present in the sample are provided by the relationship that exist between the variables. To describe PCA in a mathematical manner, principal components are considered as the eigen-vectors having the symmetric correlation with the dataset. This clearly indicates that the matrix is necessary to have standard and numeric data. Eigen-vectors of the matrix is orthogonal by default. The eigen-vectors communicate in the direction where there exist a maximum variance in the dataset. When there exist a maximum eigen-value, it represent the better performance towards the eigen-vector, with minimum eigen-value it represents the worst performance. Considering future risk complication towards health, CAD results as a major problem among prediction of diseases. CAD is one of the major types of disease, where 25% of affected people gets dead suddenly without having any symptoms. Heart attacks become severe in patients who are all affected by CAD. It is estimated that CAD may affect the teenage population. World Health Organization has estimated 11.1 millions of death all throughout the world due to CAD in the year 2020. By having the awareness of CAD and its symptoms can aid in timely treatment, and it will reduce the severity and its side effects. Cardiovascular diseases are becoming the most common reasons of death across South Asia countries. Correct and in-time diagnosis of CAD is very important. Many high-level treatments (i.e., Angiography) are available for CAD, but it has many side effects and it is costly. Performance degradation exists due to more number of records. The healthcare industry is having huge volumes data and it need mining to discover hidden information for making effective decision. Data mining algorithms are available to analyze the dataset and predict the results towards disease, but the performance and results gets varied for datasets. Most data mining algorithms are available for constant or specific dataset. When applying the algorithms in different datasets, the performance of the algorithm gets weak which leads to a major problem in medical field. Time taken for analyzing and predicting the results are higher. The main objective of this research work is to propose a better classifier based on principal component analysis in order to make prediction by classifying the patients affected by CAD. In the proposed work, feature reduction will be done in order to increase the accuracy, where null values too will be checked for existence.

- *S.Omprakash M.Sc, M.Phil Assistant Professor, Department of Computer Science Sankara College of Science and Commerce, Coimbatore, Tamil Nadu, India-641020. Subramaniamomprakash@gmail.com*
- *Dr. M.Ravichandran M.Sc.,M.Phil.,M.Phil.,P.hD AssociateProfessor, Department of Computer Science SRMV College of Arts and Science, Coimbatore, Tamil Nadu, India-641020.*

2. LITERATURE REVIEW

Modified Differential Evolution Algorithm [1] was proposed to select the significant features from enormous set of features available for predicting the heart disease. From the available 10 strategies of differential evolution algorithm, 7th strategy was used with fuzzy logic to increase the accuracy, but the results came with inversely proportional. Clustering based Analytical Method [2] was proposed to predict disease using different data mining techniques. Classification with Regression Tree algorithm was utilized to generate the rules for fuzzy system. When applying the method won different medical dataset it decreased the classification accuracy. CART [3] method was used to analyze the acute rheumatic fever which was related to cardiac disease. Significant attributes which are identified as the root cause of cardiac disease were identified and classification was processed. The result with increased false positive shows CART is not efficient in predicting the heart disease.

Ensemble Method [4] was proposed to find the important attributes to increase the classification accuracy of CAD. The low accuracy result shows that the ensemble of classification techniques won't give any improvement towards classification accuracy. Modified Probabilistic Neural Networks [5] was proposed to find the hidden information in the attributes towards the prediction of heart disease. Each features where analyzed individually in order to eliminate the low ranked attribute. The result with increased false positive and false negative shows that the algorithm is not fit for the prediction of heart disease like CAD. Frequent Itemsets based Prediction Model [6] was proposed to identify the level of risk with the patients who have heart disease. It generates the itemsets by analyzing the symptoms and low level support value. Later, the itemsets are extracted and risk levels are determined, where it does not provide the increased classification accuracy.

Sequential Minimal Optimization (SMO)[7] was evaluated to detect the CAD by reducing the number of features available in the dataset, where the dataset named Z-Alizadeh-Sani was introduced which holds 303 patient records and 54 features. The results shows that the algorithm is not effective in predicting the CAD due to low classification accuracy. Ant Colony Optimization based Support Vector Machine (ACO-SVM) [8] was proposed to detect the CAD among the patients in an optimized manner. Random classification with threshold value fixing enhanced the classification accuracy, but it is not sufficient for predicting and classifying the patients who have CAD.A Cox proportional hazards regression model was used to develop risk prediction model [9] for cardiovascular diseases. The risk assessment ability of the developed model was evaluated, and a bootstrapping method was used for internal validation. The predicted risk was translated into a simplified scoring system. A decision curve analysis was used to evaluate clinical usefulness. Feature Identification Method [10] aimed to identify significant features and data mining techniques that can improve the accuracy of predicting cardiovascular disease, where prediction models were developed using different combination of features and classification techniques.

Table 1. Previous Classification Methodologies and Drawbacks

Modified Differential Evolution Algorithm [1]	Low accuracy
Clustering based Analytical Method [2]	Low accuracy
Classification and Regression Tree [3]	High false positive
Ensemble Method [4]	Low accuracy
Modified Probabilistic Neural Networks [5]	High false positive and false negative
Frequent Itemsets based Prediction Model [6]	Low accuracy and F-Measure
Sequential Minimal Optimization (SMO)[7]	Low accuracy
Ant Colony Optimization based Support Vector Machine (ACO-SVM) [8]	Low accuracy
Validation Based Prediction Model [9]	Low Sensitivity
Feature Identification Method [10]	Low Positive Rate

3. PROPOSED METHODOLOGY

3.1 Support Vector Machine (SVM)

Supervised Machine Learning (SML) is assumed as a specific technique which expects the required input and expected output from the user. The data provided by the user clearly labeled for the making better classification with the aim of providing ensured potential data processing. Specifically SML algorithms provide the measuring ability in order to support future dimension. Shortly, all SML have the input variables as X and output variables as Y, where the user utilize the algorithms for studying the classification with the function $Y = f(X)$.

SVM is a specific category in SML algorithms which can be fully utilized for performing the classification and regression. In the current world, Multiple domains like education, medicine, business, etc., started utilizing SVM algorithms for the classifying and predicting the data. But, in real time SVM algorithm is fully utilized for classifying the text, spam emails, analysis of sentiment, making opinion, taking decisions, classifying the disease, recognizing the image, and etc.

SVM algorithm perform based on discovering the hyperplane which can divide the input or dataset (i.e., X) into two classes. Data points that close to the hyperplane are treated as the support vectors. If data points that are close to the hyperplane are removed then there exists a modification in the position of hyperplane. In short, the hyperplane is treated as a line which classifies the dataset in a linearly.

3.2 Principal Component Analysis (PCA)

The core intention of PCA is to eliminate the noisy cum highly correlated variables, where it preserves the significant information in the dataset. PCA method depends on the data gathered in general working condition.

Let W belongs to Q with the order $M \times N$ matrix, where M indicates the measures and n indicate its variables. In PCA method, W (the matrix) is balanced to the mean value zero with a component variance. Data matrix W's covariance matrix ζ.The covariance matrix $\zeta$of the data matrix W and its depreciation eigen value are mathematically expressed as:

$$\zeta = (i + M)W^S W$$

$$\zeta = (\bar{O}\ddot{O})\begin{pmatrix} \ddot{\Pi} & 0 \\ 0 & \ddot{\Pi} \end{pmatrix}(\bar{O}\ddot{O})^S \quad (1)$$

where $\bar{O}$ belongs to $Q^{n/k}$ and $\ddot{O}$ belongs to $Q^{n/n+k}$, and it represents the initial and final eigen-vectors of ζ. Correspondingly, k indicates the count of principal components. $\bar{\Pi}$ and $\ddot{\Pi}$ indicates the diagonal matrices which contains maximum $(i.e.,k)$ and minimum $(i.e.,(n+k))$ values of ζ. The W matrix can be rewritten as

$$W = \bar{W} - \ddot{W} \quad (2)$$

Where $\bar{W}$ and $\ddot{W}$ projects the subspaces of the principal components, traversed by $\bar{O}$ columns, subspaces occupied by the remaining $\bar{O}$ columns that are traversed.

3.3 Core Principal Component Analysis based Support Vector Machine

Traditional PCA gives its performance only when there exist a linear dataset. In order to perform well in non-linear dataset CPCA is proposed. The main intention of CPCA is to locate the space-of-input into the feature space (FS)G through non-linear mapping methodology $\propto$, and finally performing PCA in G.

Consider the general sets available for training be $w^{i+1}, w^{i+2}, \dots, w^M$ where all belongs to Q^n. The deliberated inputs are made to project in space-of-feature by utilizing the mapping methodology $\propto$:

$$\propto\ w^j \text{belongs to} Q^n > \propto [w^j] \text{belongs to} Q^g \quad (3)$$

where g is equivalent to M, and it indicates the FS dimensionality. It is considered as the significant property of FS, where it involves the dot operation between the two vectors $\propto [w^j]$ and $\propto [w^i]$, where $j, i = 1,2,3 \dots, M$. It is mathematically expressed as:

$$\propto [w^j]^S \propto [w^j] = |\propto [w^j], \propto [w^i]| = l[w^j, w^i] \quad (4)$$

where l is treated the core function for handling the non-linear dataset. There exist different kinds of functions for handling dataset. Most utilized function in handling the dataset is RBF (Radial-Basis-Function), it can be mathematically defined as

$$l[w^j, w^i] = exp\begin{pmatrix} \ll w^j + w^i \gg^2 \\ \emptyset + 2 \end{pmatrix} \quad (5)$$

where $\emptyset$ represent the width of a function related to Gaussian method. In FS, the vectors are extended to reach the mean with the value zero. The data mapping in G can be expressed as

$$w = (\propto [w^{i+1}] \propto [w^{i+2}] \dots \propto [w^M])^S \quad (6)$$

The calculation of covariance matrix R that exist in FS is given as

$$R = (i+1) * \frac{(M+1)}{w^S w} = i + 1 * M + 1 \int_{i+1}^{M} \propto [w^i] \propto [w^i]^S \quad (7)$$

where it has been made an assumption that $\int_{i+1}^{M} \propto [w^i]$ will be equivalent to 0. As like PCA handling linear dataset, CPCA in FS will be equal in elucidating the eigen-value issues, which is

$$\rho^l U^l = RU^l \quad (8)$$

$$= i + 1 * M + 1 \int_{i=1}^{M} |\propto [w^i], U^l| \propto [w^i] \quad (9)$$

where ρ^l and U^l indicates the corresponding l^{th} eigen-value and R's eigen-vector, and $|.,.|$ is the dot operation performed. When ρ^l equals the value 0, then individual eigen-vectors are assumed as linear sequence of $\propto [w^{i+1}], \dots, \propto [w^M]$.Therefore. There exist co-efficient value β^{l+i},where $i = 12,3, \dots, M]$.

$$U^k = \int_{j=1}^{M} \beta^{l,j} \propto [w^j] = w^S \beta^l \quad (10)$$

where β^l equals $(\beta^{l+1}, \beta^{l+2}, \dots, \beta^{l+M})^S$. By performing the product operation in Eq. (7) with $\propto [w^l]$, and by introducing Eq. (10) in Eq. (7), the result will be:

$$\rho^l \int_{j=1}^{m} \beta^{l+j} |\propto [w^l], \propto [w^j]| = (i+1) * (M+1) \int_{j=1}^{M} |\beta^{l+j} \propto [w^l], \int_{i=1}^{M} \propto [w^i]|| \propto [w^i], \propto [w^j]| \quad (11)$$

By utilizing the core function $L^{j+i} = l[w^j, w^i] = |\propto [w^j], \propto [w^i]|$, Eq. (11) can be updated as:

$$\rho^l \int_{j=1}^{M} \beta^{l+j} L^{l+j} = i+1 * M+1 \int_{j=1}^{M} \beta^{l+j} \int_{i=1}^{M} L^{l+i} L^{i+j} \quad (12)$$

Hence, the issue of eigen-value in can be mathematically written as:

$$\pi^l \beta^l = L\beta^l \quad (13)$$

where$\pi^l = [M+1]\rho^l$,where L belongs to Q^{M+N}which will be the core matrix that can be defined by utilizing L^{i+j} and β^l. In G, the eigen-vector U^l forms the matrix U^e that can be defined as $U^e = (U^{i+1}, U^{i+2}, \dots, U^k U^{k+1} \dots U^M)$, where k indicates the preserved number of principal components (PCs) in the non-linear dataset.

The first part $\bar{O}^e = (U^{i+1}, U^{i+2}, \dots, U^k)$ belongs to $\mathbb{Z}^{M/k}$ and the last part $\ddot{O}^e = (U^{k+1}, \dots, U^M)$ belongs to $\mathbb{Z}^{\frac{M}{M+k}}$ indicates the defined eigen-vectors corresponding to initial and final. For providing the assurance for G towards the normalization constraint, Eq. (10) and Eq. (13) are used in the derivation

$$|U^l, U^l| = |\int_{j=1}^{M} \beta^{l+j} \propto [w^j], \int_{i=1}^{M} \beta^{l+i} \propto [w^i]| = \int_{j=1}^{M} \int_{i=1}^{M} \beta^{l+j} \beta^{l+i} \beta^{i+j} = \int_{j=1}^{M} \int_{i=1}^{M} \beta^{l+j} \beta^{l+i} \beta^{i+j} = 1 \quad (14)$$

Hence, β^l must be normalized to $(1+\pi^l)$. Assume $\hat{\beta}^l$ be the single normalized eigen-vector with respect to π^l, where:

$$\beta^l = i+1 * \pi^l = \hat{\beta}^l \quad (15)$$

where l ranges from 1,2,3..M

By utilizing the Eq. (10) and Eq.(15), $\bar{O}^e$ matrix can be written as:

$$\bar{O}^e = \left(w^S \hat{\beta}^{i+1} \pi^{i+1}, \dots, w^S \hat{\beta}^k \pi^k\right) == w^S O \prod^{i+2} \quad (16)$$

where O is equivalent to $(\hat{\beta}^{i+1}, \hat{\beta}^{i+2}, \dots, \hat{\beta}^k)$ and $\prod$ is equivalent to $diag\,[\pi^{i+1}, \dots, \pi^k]$. The corresponding initial and final components, s belongs to Q^k and $\ddot{s}$ belongs to Q^{M+k} of a analyzing vector x, were extracted by utilizing $\propto [w]$ in the initial and final spaces, as Eq. (17):

$$\begin{cases} s = \bar{O}^{S/e} \propto [w] \\ \prod^{i+2} O^S l[w] \\ \ddot{s} = \ddot{O}^{S/e} \propto [w] \end{cases} \quad (17)$$

where $l[w]$ is equivalent to $(l[w^{i+1}, w] l[w^{i+2}, w] \dots, l[w^M, w])^S$. In FS vector $\propto [w]$ has the mean value 0. Else $\hat{\propto}\,[w]$ calculated and expressed as:

$$\hat{\propto}\,[w] = \propto [w] + (i+1*M) * \int_{j=1}^{M} \propto [w^j] = \propto [w] + (\propto [w^{i+1}] \propto [w^{i+2}], \dots, \propto [w^{i+M}])] J^M \quad (18)$$

Where $J^M = (i+1*M)(i+1, \dots, i+n)^S$ and it belongs to Q^M.

Based on Eq. (18), core function of calculated vectors $\hat{\propto}[w^j]$ and $\hat{\propto}[w^i]$ are mathematically expressed as:

$$\hat{l}[w^j, w^i] = \hat{\propto}[w^j]^S \hat{\propto}[w^i] = l\left[w^j, w^i + l^S[w^j]\right]J^M + l^S[w^i]J^M - J^{S/M}LJ^M \quad (19)$$

As like the calculation of core function vector (i.e., Eq. (19)), calculation of core vector can be proceeded as:

$$\begin{cases} \hat{l}[w] = (\hat{\propto}[w^{i+1}], \dots, \hat{\propto}[w^M])^S \hat{\propto}[w] \\ = E[l[w] + LJ^M] \\ E = J + F \end{cases} \quad (20)$$

Identity matrix is indicated as J in Eq. (20). $F \subset Q^{N*M}$ is considered as a matrix which have1 $* M$ elements. The authoritative core matrix $\hat{L}$ deliberated as:

$$\hat{L} = (\hat{\propto}[w^{i+1}], \dots, \hat{\propto}[w^M])^S * (\hat{\propto}[w^{i+1}, \dots, \hat{\propto}[w^M]]) \quad (21)$$

4. Z-ALIZADEHSANI DATASET

The dataset [7] consists of 303 patients records. Each record holds 54 features. Every feature in the dataset are considered as the indicators of CAD, which is according to medical history, but some of the specific features are never used in the approaches of data mining in diagnosing the CAD. The features regarding the CAD are arranged in 4 groups: demographic, symptom and examination, ECG, and laboratory and echo features. Each patient could be in two possible categories CAD or Normal. Patients are categorized as CAD, if his/her diameter range arrows greater than or equal to 50%, else the patient is treated as normal. Some features are used to identify the history of (i) hypertension,(ii) Diabetes Mellitus, (iii) consumption of cigarettes, (iv) previous consumption of cigarettes, and (v) heart disease in first-degree relatives.

5. ABOUT MATLAB:

To apply the proposed work, Matlab R2013a was used. Matlab is a tool for mining of machine learning, DM, text, and business analytics. It is mainly utilized for applications in research, edification, guidance, and engineering due to user-friendly and ease of use. Matlab comprises of inbuilt mathematical functions aim to solve scientific problems. It is well suited for designing, exploring and solving the iterative problems. The functions and applications available in Matlab are easy-to-use and help the researchers to design the predictive models in an accurate and rapid manner.

6. PERFORMANCE MEASURES

In DM, algorithms performance are measured using specificity, sensitivity, and accuracy. It is considered as most important due to its applicability in the field of medicine. The confusion matrix is a type of table which provides visualization to the algorithms performance. Considering 2 class problem (Class1 and Class2), the matrixes will have rows and columns in the number 2, it identify the count of true positives (TP), true negatives (TN). false positives (FP), and false negatives (FN). These measures [7] are defined as follows:

- ✓ TP – It's the statistic count of Class1 samples which has been accurately classified.
- ✓ TN - It's the statistic count of Class2 samples which has been accurately classified.
- ✓ FN – It's the statistic count of Class1 samples which has been unjustifiably classified as Class2.
- ✓ FP – It's the statistic count of Class2 samples which has been unjustifiably classified as Class1.
  - $Sensitivity = \frac{TP}{TP+FN}$
  - $Specificity = \frac{TN}{TN+FP}$
  - $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
  - $False\ Positive\ Rate = \frac{FP}{FP+TN}$
  - $True\ Positive\ Rate = \frac{TP}{TP+FN}$
  - $Precision = \frac{TP}{TP+FP}$
  - $Recall = \frac{TP}{TP+FN}$
  - $F - Measure = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$

7. RESULTS AND DISCUSSIONS

Fig.1 to Fig. 6 corresponding values are shown in table 2.

Algorithms	SMO	ACO-SVM	CPCA-SVM
TP	159	164	177
TN	126	129	116
FP	8	4	3
FN	10	6	5
Acc	94.06	96.70	97.34
Sen	94.08	96.47	97.25
Spec	94.03	96.99	97.48
FPR	5.97	3.01	2.52
TPR	94.08	96.47	97.25
PRECISION	95.21	97.62	98.33
RECALL	94.08	96.47	97.25
F-MEASURE	94.64	97.04	97.79

Table 2: *performance metrics result values*

7.1 True Positive and True Negative Analysis

In Fig. 1, the metrics TP and TN are plotted in x-axis, and the corresponding percentages are plotted in y-axis. Fig. 1 clearly indicates that the proposed algorithm CPCA-SVM has outperformed the previous algorithms namely SMO [7] and ACO-SVM [8], due to performing the feature reduction. The other algorithms performs classification by taking all the available features.

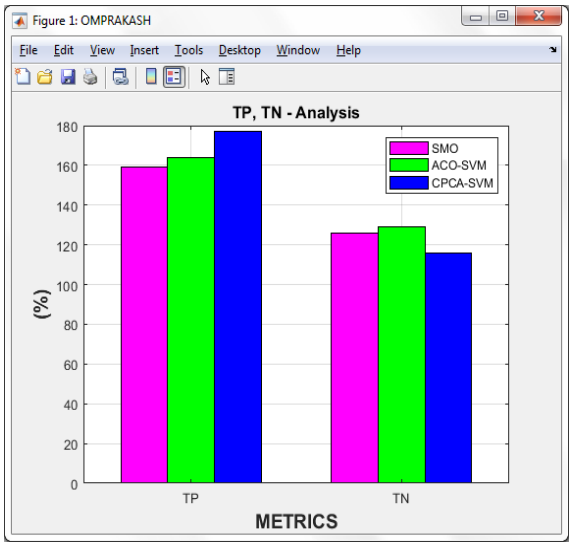

Fig. 1 TP, TN – Analysis

7.2 False Positive and False Negative Analysis

In Fig.2, the metrics FP and FN are plotted in x-axis, and the corresponding percentages are plotted in y-axis. Fig. 2 clearly indicates that the proposed algorithm CPCA-SVM has outperformed the previous algorithms namely SMO [7] and ACO-SVM [8], by giving the result in a low value where other algorithms have increased value which is not accepted in the medical field. Effective utilization of Gaussian method in the proposed work has remarkable reduced the false rates.

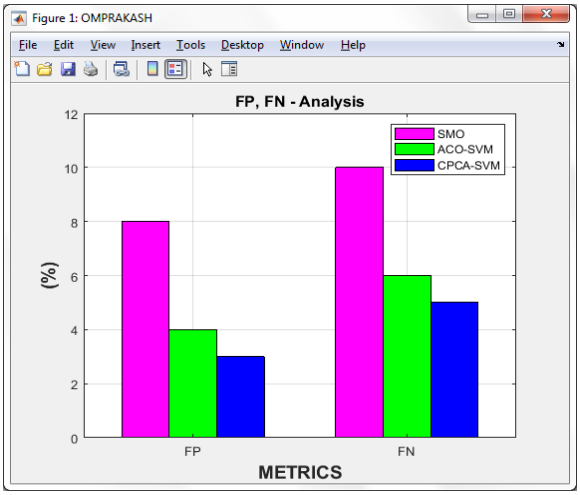

Fig. 2 FP, FN – Analysis

7.3 Sensitivity and Specificity Analysis

In Fig.3, the metrics Sensitivity and Specificity are plotted in x-axis, and the corresponding percentages are plotted in y-axis. Fig. 3 clearly indicates that the proposed algorithm CPCA-SVM has given its performance than the previous algorithms namely SMO [7] and ACO-SVM [8]. The covariance matrix utilization plays a major role in the proposed work in classification tending to give the best value in sensitivity and specificity.

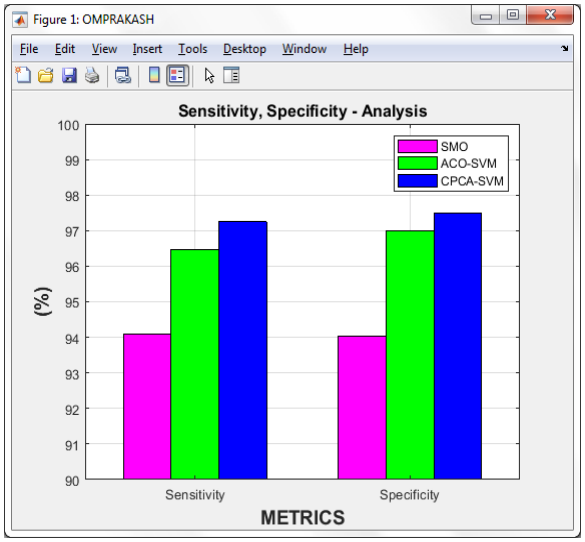

Fig. 3 Sensitivity, Specificity - Analysis

7.4 Positive Rate Analysis

In Fig.4, the metrics True Positive Rate and False Positive Rate are plotted in x-axis, and the corresponding percentages are plotted in y-axis. From Fig. 4, it is evident that the values given by proposed algorithm CPCA-SVM has maximum positive rates than the previous algorithms namely SMO [7] and ACO-SVM [8]. Reducing the features towards classifying CAD has given increased value in true positive and decreased value in false positive than other algorithms.

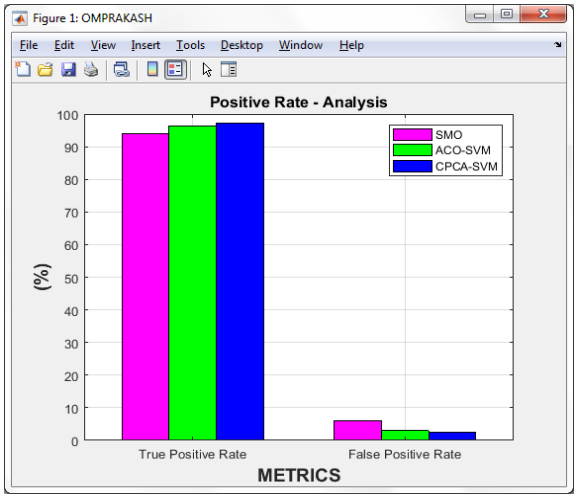

Fig. 4 Positive Rate - Analysis

7.5 Precision and Recall Analysis

In Fig.5, the metrics Precision and Recall plotted in x-axis, and the corresponding percentages are plotted in y-axis. From Fig. 5, it is evident that the previous algorithms SMO [7] and ACO-SVM [8] have low values than the proposed algorithm CPCA-SVM. Utilization of eigen vectors and values results in better result in precision and recall.

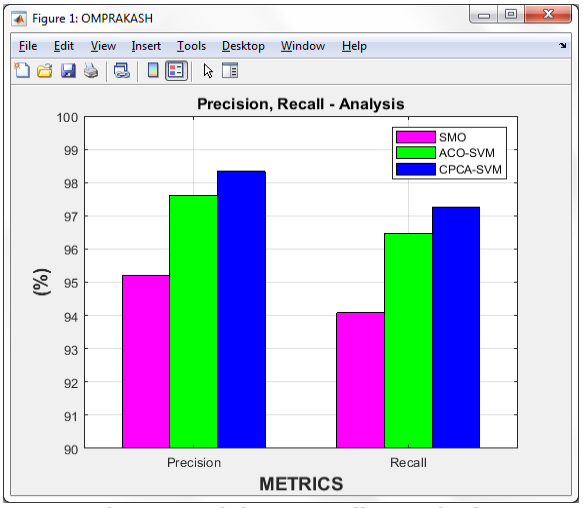

Fig. 5 Precision, Recall - Analysis

7.6 Accuracy and F-Measure Analysis

In Fig.6, the metrics Accuracy and F-Measure is plotted x-axis, and the corresponding percentages are plotted in y-axis. From Fig. 6, it is evident that the previous algorithms SMO [7] and ACO-SVM [8] have low performance in predicting CAD than the proposed algorithm CPCA-SVM. Existing algorithms simple perform classification with all the features available, also they don't even check whether data are available in the features. But the proposed work performs feature reduction with null value check. This tends to provide better accuracy and f-measure value.

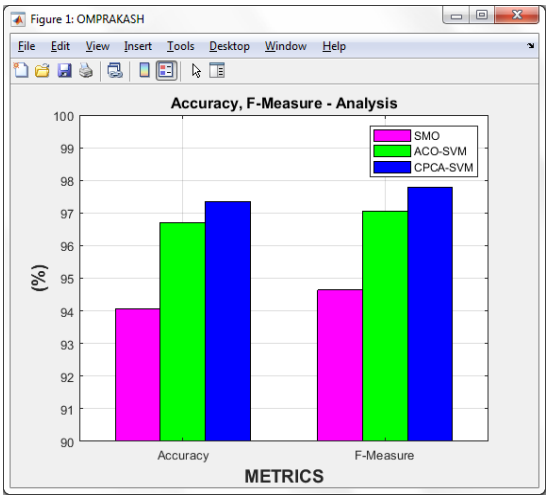

Fig. 6 Accuracy, F-Measure - Analysis

8. CONCLUSIONS

CAD is identified as a deadly disease causing sudden death. Multiple algorithms were proposed to classify CAD, but those algorithms were limited to specific dataset or small dataset. In this paper, ensemble of PCA and SVM has been performed to classify the CAD in a large dataset. The proposed method in this paper performs the updation by using monitoring process

that deal with dynamic change in behavior. In order to maximum accuracy, value of gamma used in RBF kernel has been dynamically updated based on probability of feature size. The proposed classifier is evaluated with Z-AlizadehSani dataset for classification accuracy for the prediction of coronary artery disease. Utilization of RBF kernel, the ensemble of PCA and SVM has resulted with 97.34% accuracy in detecting CAD which outperforms the baseline schemes. Future enhancement of this research work can be focused with statistical based model combined with data mining which can result in achieving improved classification accuracy.

References:

[1] Vivekanandan. T., Ch. S. N. I. “Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease,” *Computers in Biology and Medicine*, Volume 90, Pages 125-136, 2017.

[2] Mehrbakhsh. N., Othman. B. I., Hossein. A., Leila. S. “An analytical method for diseases prediction using machine learning techniques,” *Computers & Chemical Engineering*, Volume 106, Pages 212-223, 2017.

[3] İlkim. E. E., Nurdan. E., Yusuf. İ. A., Yalçın. Ö.,Çiğdem. E. “The analysis of the effects of acute rheumatic fever in childhood on cardiac disease with data mining,” *International Journal of Medical Informatics*, Volume 123, Pages 68-75, 2019.

[4] Mohammad. S. A., Yin. K. C., Kasturi. D. V. “Identification of significant features and data mining techniques in predicting heart disease,” *Telematics and Informatics*, Volume 36, Pages 82-93, 2019.

[5] El-Houssainy. A. R., Ayman. S. A. “Prediction of kidney disease stages using data mining algorithms,” *Informatics in Medicine Unlocked*, Volume 15, 2019.

[6] Ilayaraja. M., Meyyappan. T. “Efficient Data Mining Method to Predict the Risk of Heart Diseases Through Frequent Itemsets,” *Procedia Computer Science*, Volume 70, Pages 586-592, 2015.

[7] Alizadehsani. R, Habibi. J, Hosseini. M. J., Mashayekhi. H., Boghrati. R., Ghandeharioun. A., Bahadorian. B., Sani. Z. A. "A Data Mining Approach for Diagnosis of Coronary Artery Disease", *Computer Methods and Programs in Biomedicine*, Volume 111, Pages 52-61, 2013.

[8] Omprakash. S., Ravichandran. M. "Ant Colony Optimization Based Support Vector Machine Towards Predicting Coronary Artery Disease", International Journal of Recent Technology and Engineering, Volume 7, Pages 210-215, 2019.

[9] Amin. M. S., Chiam. Y. K., Varathan. K. D. “Identification of Significant Features and Data Mining Techniques in Predicting Heart Disease”, *Telematics and Informatics*, Volume 36, Pages 82-93, 2019.

[10] Honda. T., Yoshida. D., Hata. J., Hirakawa. Y., Ishida. Y., Shibata. M., Sakata. S., Kitazono. T., Ninomiya.T. “Development and Validation of Modified Risk Prediction Models for Cardiovascular Disease and its Subtypes: The Hisayama Study”, *Atherosclerosis*, Volume 279, Pages 38-44, 2018.