International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 07 | July 2020 www.irjet.net p-ISSN: 2395-0072

DMFS-PCA APPROACH FOR BREAST CANCER PREDICTION

Bhagyabhanu M.P[1], Anu Maria Joykutty[2]

[1]*Student, Rajagiri School of Engineering and Technology, Kerala, India*
[2]*Professor, Dept. of Computer Science, Rajagiri School of Engineering and Technology, Kerala, India*

Abstract – *Nowadays breast cancer is one of the most widespread causes of death in women. The difference in the DNA methylation values can be considered as an important biomarker for identifying cancer. According to an estimation, approximately 40,920 women died in 2018 only due to breast cancer, which is a highly startling number. This situation could be reduced if the cancer can be diagnosed at an early stage. With the emerging technologies, such predictions has become an easier task. Machine learning is one of the latest technology, which helps to make predictions related to diseases based on physical or behavioural characteristics. The cascaded approach is used for Cancer prediction. First, Standard Deviation Threshold based Differential Mean Feature Selection (DMFS) is applied on the input data to select most discriminative features based on threshold. Threshold value is set as Standard Deviation of weight vectors. Value of the threshold is set to get the best prediction accuracy. Post which Principal Component Analysis (PCA) is applied on the set of the selected features after which Cancer prediction is performed.*

Key Words: Breast Cancer Prediction, DMFS, PCA, Random Forest Classifier, DNA Methylation, Accuracy, F measure, RMSE

1. INTRODUCTION

Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics. So it makes a significant public health problem in the new era. The early diagnosis of Breast Cancer can significantly improve the prognosis and chance of survival. It can also promote timely clinical treatment to patients. Further accurate classification of benign tumours can prevent patients unnecessarily undergoing for treatments.

2. LITERATURE SURVEY

Many methods have been developed to predict breast cancer. The existing methods can be classified into image processing based approach and DNA methylation based approach.

Mandeep Rana et.al. [1] have compared the accuracies of different classification techniques. They found SVM most suited for predictive analysis and KNN performed best for the overall methodology. Morteza H. et. al. [2] used a locally preserving projection (LPP) based approach for cancer prediction. They utilised computer-aided image processing scheme to segment dense fibro-glandular breast tissue regions automatically in each mammogram. Vahid et.al. [3] used Wireless Capsule Endoscopy Images for prediction. The image is divided into several patches and Discrete Wavelet Transformation is applied. Finally SVM classifier is used for predicting cancer.

Abeer et.al. [4] method uses F score based feature selection method and Fast Fourier Transform algorithm for feature extraction. Abdulmajid et.al. [5] utilised cascaded DMFS-DWFE approach for early cancer prediction. In this approach, threshold value for feature selection needs to be tuned based on the dataset.

3. PROPOSED METHODOLOGY

This section deals with the various stages in the proposed cascaded DMFS-PCA approach for breast cancer prediction.

3.1 Pre-processing

Initial stage of the proposed approach is pre-processing. Figure 4.1 represents pre-processing stage. The input dataset used is TCGA HumanMethylation450 dataset which is not in comma separated format. For easy processing of data we have converted the input dataset to CSV format. As part of pre-processing we have considered 32000 features and 888 samples. Samples comprises of 790 Cancer and 98 Normal samples. Cleansing of data is done by removing features having null values for all samples as those features are irrelevant ones for Classification. Transposing of data is performed so that row represents sample and column represents feature.

3.2 Standard Deviation Threshold Based Differential Mean Feature Selection (DMFS)

First each feature is analysed vertically across all samples and the mean value of the DNA methylation values of both normal and cancer samples are found separately [5]. After that absolute difference between the mean values of normal and cancer samples are found out. The resulted vector is named weighting vector. Finally, the weighting vector is used to select features. The threshold value is set as standard deviation of the weight vector. We select those features that are having weight vector value greater than this threshold.

This Standard Deviation Threshold indicates that the difference between the two mean values of normal and cancer samples is adequate to say that feature is discriminative for cancer classification. Hence this approach is known as Standard Deviation Threshold based Differential Mean Feature Selection (DMFS).

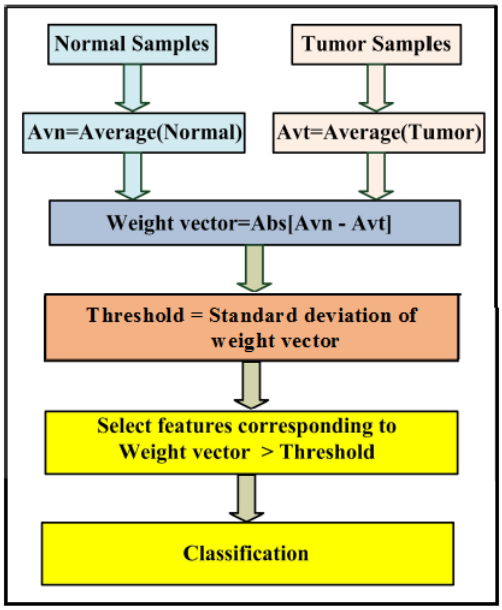

Fig -1: Standard Deviation threshold based Differential Mean Feature Selection (DMFS)

3.3 Principal Component Analysis (PCA)

Dimensionality reduction is done using PCA which is one of the most popular techniques to remove irrelevant and redundant features [6]. PCA is a linear transformation technique. It transforms data in such a way that the first coordinate represents data with highest variance, second coordinate represents data with second highest variance and so on. Therefore, PCA successfully reduces the large dimension of datasets by considering the coordinates having high variance values and ignore the data that has low variance.

PCA Algorithm:

Step 1: Calculate the covariance matrix

Step 2: Calculate the Eigen vectors and Eigen values of the covariance matrix

Step 3: Sort Eigen vectors in decreasing order of Eigen values

Step 4: Identify Principal Component as the Eigen vector with the largest Eigen value

Step 5: Perform Dimensionality reduction by eliminating the principal components with minor significances

3.4 Cascaded DMFS-PCA Approach

Pre-processed data is used as input for Feature Selection using DMFS approach. Selected features are then used for feature extraction. Feature Extraction is performed by Principal Component Analysis (PCA). PCA is a data reduction technique by removing irrelevant and redundant features without losing its properties. After feature extraction, Classification is done using Random Forest method.

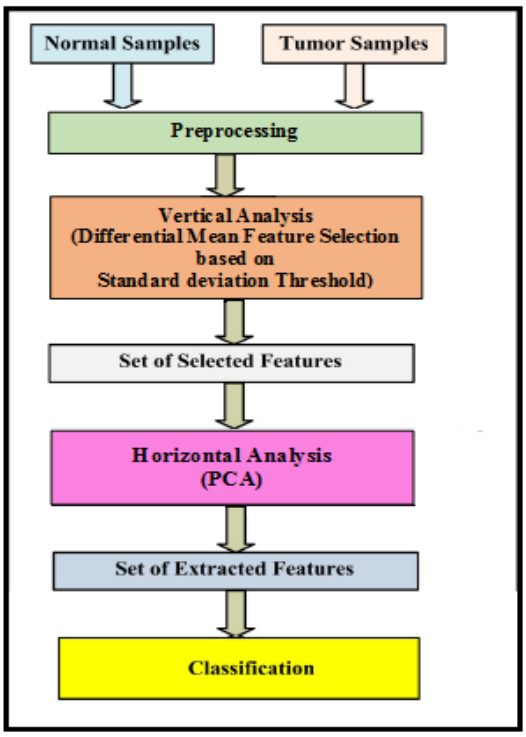

Fig -2: Cascaded DMFS-PCA Architecture

4. EXPERIMENTAL RESULTS AND ANALYSIS

The Cancer Genome Atlas (TCGA) dataset [7][8][9] from Max Planck Institute for Informatics (MPI) is used in this study. Breast cancer is studied with DNA Methylation values.

We have considered 32000 features and 888 samples. Samples comprises of 790 Cancer and 98 Normal samples. For training phase of classification, we have used 75% of the dataset. Remaining 25% of the dataset is used for classification testing.

Evaluating the model accuracy is an essential part of the process in creating machine learning models to describe how well the model is performing in its predictions.

4.1 Feature Reduction Analysis

The dataset contains 32000 features. In pre-processing stage it is reduced to 23925 features after removing unwanted features. In DMFS stage number of feature gets further reduced from 23925 to 7172 features. Finally, number of features in the proposed cascaded approach reduced to 100. So, the proposed method has reduced the

number of features from 32000 to 100. Figure 3 depicts that using the proposed method the number of features will be reduced by 0.31%.

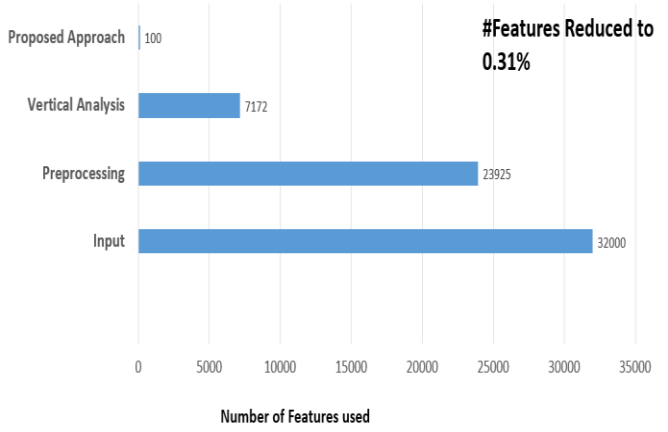

Fig -3: Feature Reductions in various stages

4.2 Performance Analysis

The performance of classification done in various stages of the proposed method is compared. Table 1 depicts this performance improvement in various stages. From this table, we could infer that the accuracy and F-measure of the proposed cascaded DMFS-PCA approach is higher than both DMFS and PCA approaches. Also, MAE and RMSE measures are significantly reduced in the proposed method compared to DMFS and PCA approaches.

Table -1: Performance improvement in various stages of the proposed method

	DMFS	PCA	Cascaded DMFS-PCA
Accuracy	0.97747	0.98198	0.99099
F-Measure	0.94968	0.95913	0.97957
MAE	0.02252	0.01802	0.00901
RMSE	0.02252	0.01802	0.00901

We have then analysed the performance of classification done based on the features extracted from the proposed method with the simple classification done based on the original pre-processed dataset. The table 2 depicts the fact that classification done based on features extracted from the proposed method was able to improve accuracy and F-measure compared to simple classification which is nearer to the ideal accuracy and F-measure. The proposed method was also able to reduce MAE and RMSE close to zero compared to the simple classification.

Table -2: Performance comparison of the proposed method with simple classification

	Simple Classification	Cascaded DMFS-PCA
Accuracy	0.98198	0.99099
F-Measure	0.95913	0.97957
MAE	0.01802	0.00901
RMSE	0.01802	0.00901

The performance of the proposed method is compared with existing DWFE approach. Figure 4 signifies that the accuracy and F-measure has been improved after classifying the features extracted using the proposed method compared to DWFE approach. From the figure 5, it is evident that the proposed system was able to reduce the MAE and RMSE compared to the DWFE approach.

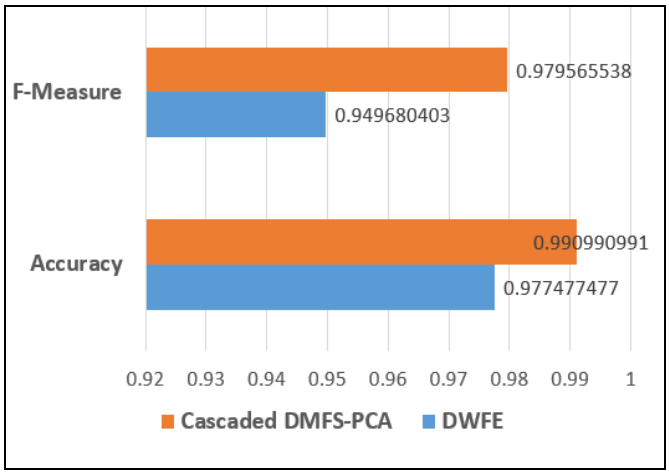

Fig -4: Comparison of Proposed Cascaded DMFS-PCA method and DWFE approach in terms of F-measure and Accuracy

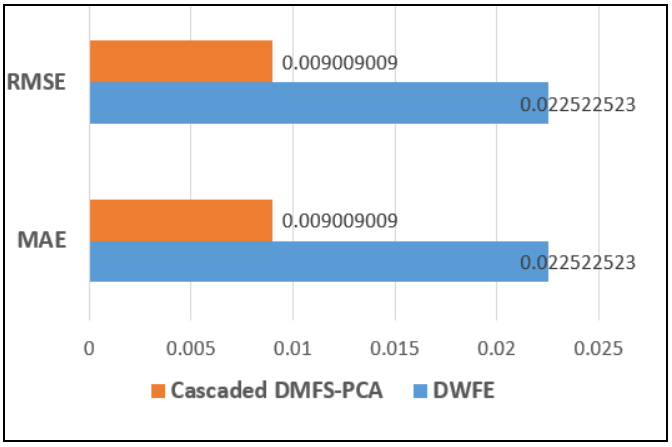

Fig -5: Comparison of Proposed Cascaded DMFS-PCA method and DWFE approach in terms of RMSE and MAE

5. CONCLUSION

The proposed method uses a cascaded approach for Cancer prediction with the help of DNA methylation values. First, Differential Mean Feature Selection (DMFS) is applied on the input data to select some discriminative features that will be helpful for classification. The features are then selected based on standard deviation threshold. After that to extract features Principal Component Analysis (PCA) is applied on the set of the selected features resulted from the vertical analysis. The features extracted from the cascaded approach is then used to classify cancer and normal samples. The proposed method is able to reduce the number of features used for classification with an improvement in accuracy and F-Measure & with a reduction in MAE and RMSE.

ACKNOWLEDGEMENT

I am thankful to Ms. Anu Maria Joykutty, Professor of Computer Science department, RSET Kerala for her valuable guidance in preparing my proposed method. I am also thankful to God almighty for showering his blessing on me for successful completion of this work.

REFERENCES

[1] Mandeep Rana, Pooja Chandorkar, Alishiba Dsouza, Nikahat Kazi, "Breast cancer diagnosis and recurrence prediction using machine learning techniques", IJRET: International Journal of Research in Engineering and Technology, 2015.

[2] Morteza Heidari, Abolfazl Zargari Khuzani, Alan B. Hollingsworth, "Prediction of Breast Cancer Risk Using a Machine Learning Approach Embedded with a Locality Preserving Projection Algorithm", Phys Med Biol. Author manuscript 2019.

[3] Vahid Faghih Dinevari, Ghader Karimian Khosroshahi, and Mina Zolfy Lighvan, "Singular Value Decomposition Based Features for Automatic Tumor Detection in Wireless Capsule Endoscopy Images", Applied Bionics and Biomechanics, Volume 2016.

[4] Abeer A. Raweh, Mohammed Nassef, And Amr Badr, "A Hybridized Feature Selection and Extraction Approach for Enhancing Cancer Prediction Based on DNA Methylation", IEEE, 2018.

[5] Abdulmajid F. Al-Juniad1, Talal S. Qaid1, Mohammad Yahya H. Al-Shamri, Mahdi H. A. Ahmed, And Abeer A. Raweh, "Vertical and Horizontal DNA Differential Methylation Analysis for Predicting Breast Cancer", IEEE, Vol. 6, 2018.

[6] Chris Albon. "Feature Extraction With PCA". chrisalbon.com. https://chrisalbon.com/machine learning/feature engineering/feature extraction with pca.

[7] HumanMethylation450 Dataset, TCGA breast invasive carcinoma (BRCA). [Online]. Available: https://xenabrowser.net/datapages/?cohort=TCGA\%20Breast%20Cancer%20(BRCA)&removeHub=http%3A%2F%2F127.0.0.1%3A7222.

[8] Biostars bioinformatics. "Cancer data download with normal sample". https://www.biostars.org/p/358889/.

[9] Sample Type Codes. National Cancer Institute Genomic Data Commons. https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes.