HANDLING MISSING DATA IN IPSILATERAL MAMMOGRAMS FOR COMPUTER AIDED BREAST CANCER DIAGNOSIS

N. Kalimatha[1], G.Vinutna[2], B. Namrata[3], A. Abinaya[4], Y. Anjana[5], R. Lavanya[6]

[1] UG student, Electronics and Communication Engg., Amrita Vishwa Vidyapeetham, Tamilnadu, India
[2] UG student, Electronics and Communication Engg., Amrita Vishwa Vidyapeetham, Tamilnadu, India
[3] UG student, Electronics and Communication Engg., Amrita Vishwa Vidyapeetham, Tamilnadu, India
[4] UG student, Electronics and Communication Engg., Amrita Vishwa Vidyapeetham, Tamilnadu, India
[5] UG student, Electronics and Communication Engg., Amrita Vishwa Vidyapeetham, Tamilnadu, India
[6]Assistant Professor, Electronics and Communication Engg., Amrita Vishwa Vidyapeetham, Tamilnadu, India

Abstract

Breast cancer is the most fatal among the cancers detected in women. Mammography is the most common and efficient tool for early detection of breast cancer. In mammography images of breast are captured in two standard views namely Mediolateral oblique (MLO) view and Craniocaudal (CC) view. Radiologists generally analyze both the views during diagnosis. In recent years as the number of cases to be diagnosed is increasing significantly, computer aided diagnosis (CAD) systems were developed. In the datasets used by these systems, often there is a chance for the presence of missing values (MVs) in any one of mammographic views due to various reasons. Some of these include obscuration of mass by a dense breast tissue, region of interest being out of frame during the image acquisition etc. This results in the use of data from single view alone for diagnosis. But the diagnostic performance of CAD systems is better when multi-view data is used. In this paper the use of Iterative Singular Value Decomposition (ISVD) imputation is proposed to handle missing values in order to preserve the advantage of using multi-view data during the diagnosis. Classification accuracy and Kappa statistics are the metrics used to assess the performance of ISVD scheme for different percentages of MVs ranging from 1-15%. Experimental results demonstrated that diagnosis using multi-view following ISVD performed at least as well as and most of the time better than the systems using single view.

Key Words: *Computer aided diagnosis, Imputation, Iterative singular value decomposition, Mammography, Missing values, and State vector machine.*

---***---

1. INTRODUCTION

Breast cancer is one of the most prevailing diseases causing women deaths in developed countries. Based on the statistics from the World Health Organization (WHO) [1], Breast cancer is also the most common cause of cancer death among women (522000 deaths in 2012) and the most frequently diagnosed cancer [1].

Generally cells in human body divide, grow, and die in an orderly fashion. But sometimes growth of the cells can go out of control, and may result in the formation of a mass or a lump called a tumor. These tumors are of two types, benign and malignant. Benign tumors are not harmful, but seldom can cause health risks. So they are usually removed. Malignant tumors have malformed cells which can infect nearby tissues. These malignant cells in their early stage can travel from breast through the blood and invade other parts of the body causing loss of life. So, early detection and diagnosis will be helpful in minimizing the evolution of cancer and thus reducing fatality.

Of all the existing techniques for the detection of breast cancer mammography is the best method for detection in early stages [2]. In mammography X-rays are used to take images of breast. These images are called mammograms. Mammograms are recorded in two standard views (ipsilateral): a head to toe view called craniocaudal (CC) view and a side to side view at an angle called mediolateral oblique (MLO) view. These mammograms are analyzed by the radiologists for the classification of breast cancer as benign or malignant. In early stages mammographic features of the malignant cells are very subtle and complex which makes detection a difficult task for the radiologists [3]. This difficulty may result in two problems, one being benign cases classified as malignant resulting in unnecessary biopsies and the other being malignant classified as benign which results in loss of lives. To overcome these difficulties and to assist the radiologists, computer aided diagnosis (CAD) systems have been developed. In general there are two types of CAD systems: CAD_e and CAD_x. CAD_e systems highlight suspicious regions assisting the radiologists for diagnosis. CAD_x systems diagnose the suspicious region into benign or malignant, giving second opinion to the radiologists.

In this paper a CAD_x system for classifying masses, the most common indicator of breast cancer, in the presence of missing data is considered. During the extraction of features from the mammograms there is a chance for occurrence of missing data in any one of the views. This is common in dense breast when a mass is obscured by a glandular tissue or when the region of interest is out of frame during the screening. In such cases diagnosis has to be restricted to using information from single view. However as performance of the CAD systems can be better when both views are used [4], the use of imputation is proposed in this work to estimate the missing data. Before diagnosis an iterative singular value decomposition (ISVD) algorithm is employed in the proposed work for imputing the missing data followed by two-view analysis for mass classification.

2. PROPOSED WORK

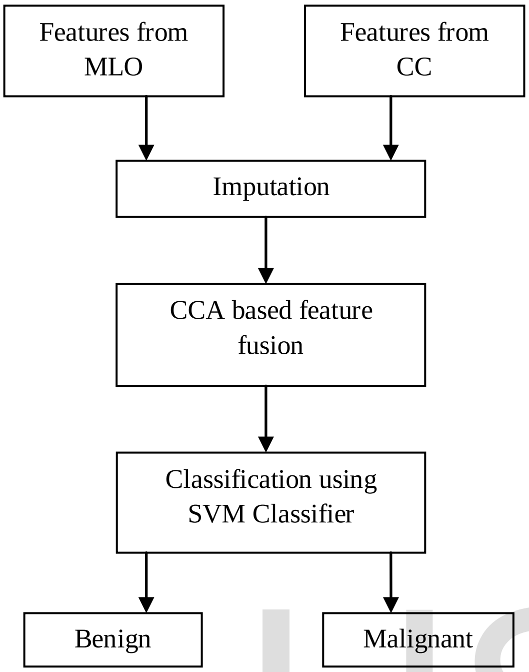

Fig –1: Flow chart of proposed work

2.1 Image Acquisition

Mammograms used to validate the proposed scheme were collected from a private hospital. There are totally 53 cases out of which 37 are benign and 16 are malignant. Region of Interest (ROIs) containing the masses from the mammograms are cropped manually.

2.2 Feature Extraction

Image acquisition is followed by extraction of shape and texture descriptors. Shape features include circularity, area and perimeter of the mass, statistical measures of the mass radius, long axis to short axis ratio, Hu's movements, zero crossing, roughness, average length of subsequences of freeman chain code, depth to width ratio and texture features Haralick's features, entropy of segmented mass[5].

2.3 Iterative Singular Value Decomposition

Different percentages of MVs (1%, 3%, 5%, 10%, and 15%) are introduced in the dataset and these MVs are estimated using the iterative SVD imputation algorithm as discussed below.

Assume the dataset with MVs as X.

Step1: MVs in each feature of the dataset are replaced with the average of available values in that feature forming a complete dataset $X_{complete}(0)$. These form the initial values.

Step2: In every i^{th} imputation cycle (i=1, 2…) the following steps are done.

a) Singular value decomposition is employed on $X_{complete}(i-1)$

$$[U,S,V] = SVD(X_{complete}(i-1)) \quad (1)$$

where,

U and V are unitary matrices and S is a diagonal matrix with eigen values in decreasing order such that

$$X_{complete}(i-1) = U * S * V^T \quad (2)$$

b) MVs in feature j are estimated using regression.

$$B = regress(P,Q) \quad (3)$$

where,

B is a row matrix with regression coefficients

P is a matrix obtained from U by taking K significant eigen observations corresponding to K significant eigen values containing 95% of the total variance.

Q is a column vector obtained from $X_{complete}(i-1)$ by considering available values corresponding to the feature j.

c) After imputing MVs in all the features, $X_{complete}(i)$ is obtained.

d) Compute the sum of squared differences between the estimated MVs of $X_{complete}(i)$ and $X_{complete}(i-1)$.

Step3: Step 2 is repeated iteratively until the computed error converges to a value less than 10^{-3}.

2.4 Feature Fusion

Following ISVD, Canonical correlation analysis (CCA) followed by feature fusion technique is employed. CCA is used to combine information from the ipsilateral views to estimate optimal directions X_A and X_B such that the two data sets S_A (features from MLO view) and S_B (features from CC view) are maximally correlated [6].

$$max_{X_A,X_B}\{corr[Y_A, Y_B]\} = \max_{X_A,X_B} \frac{X_B^T R_{BA} X_A}{\sqrt{X_A^T R_{AA} X_A}\sqrt{X_B^T R_{BB} X_B}} \quad (4)$$

where Y_A and Y_B are the transformed canonical variates, R_{AA} and R_{BB} are the autocorrelation matrices of X_A and X_B, respectively and R_{BA} is the cross correlation matrix of (X_A, X_B) . Feature fusion is employed by concatenating the transformed canonical variates Y_A and Y_B obtained after CCA.

2.5 Classification

Data set obtained after CCA based feature fusion is given as an input for the classification system which uses SVM classifier and is validated using 10 fold cross validation.

3. RESULTS

Generally classification accuracy is used as a metric to assess the diagnostic performance of a system. It is the ratio of sum of true positives (TPs) where malignant cases classified as malignant and true negatives (TNs) where benign classified as benign to that of total number of cases i.e., the percentage of cases that are diagnosed correctly.

$$classification\ accuracy = \frac{TNs+TPs}{no.of\ cases} \qquad (5)$$

Table -1: Classification performance of two-view system and single view systems.

System	Accuracy (%)
MLO alone	77.66
CC alone	82.66
Two-View analysis	85

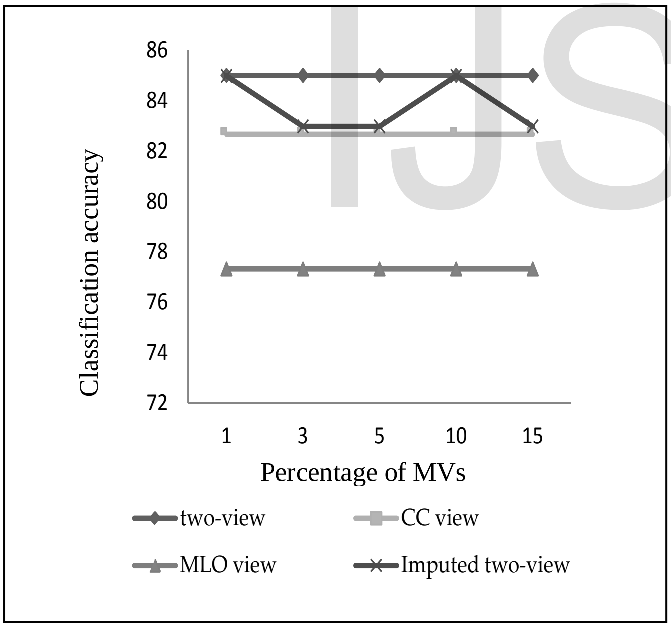

Fig -2: Classification performance of two-view system, single view systems and imputed two-view system for different percentages of MVs.

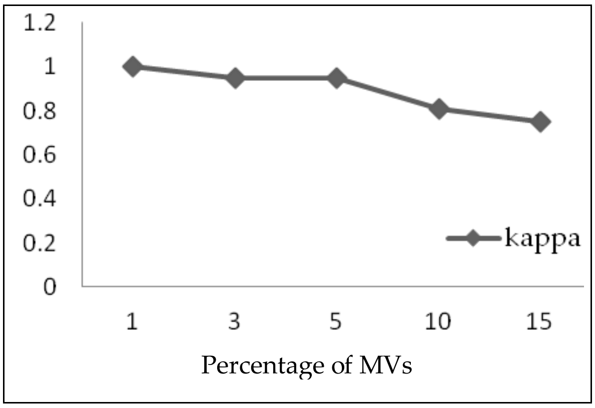

Fig -3: Kappa statistics for different percentages of MVs.

In table 1, the classification accuracies for single views and two view data for original data (no MVs) are first presented. From the table it is evident that the system using two views for diagnosis performs better when compared to the systems which use only single view.

For measuring the performance of the proposed system which utilizes imputation for estimating MVs, in addition to classification accuracy, Kappa statistics is also used as a metrics in this paper. Kappa statistics gives the agreement between the opinions of any two systems. Its value ranges from -1 to 1. The value of kappa being 1 shows total agreement and -1 shows total disagreement.

Here Kappa statistics is done to compare performances of proposed system and original two-view system (no MVs). From fig-2 it can be observed that, for different percentages of MVs considered, the classification accuracy for the imputed data set (which has both views) is higher than that of any of the single view systems. Specifically for 1% and 10% the classification accuracy becomes equal to original two-view system. For 3%, 5% and 15% the classification accuracy of proposed system is better than MLO view system by 5.34% and CC view system by 0.34%.

From fig-3 it can be observed that Kappa for lower percentages (up to 5%) is very close to '1'. And even beyond 5%, Kappa does not reduce below 0.8, which shows that most of the imputed cases are categorized into the same class as that of the corresponding original cases.

4. CONCLUSIONS

In this work ISVD imputation is proposed for estimating MVs in one of the views by using the other view for breast cancer diagnosis. The improvement in the classification accuracy for two-view analysis with imputation is utmost 3% when compared to single view system. From the results of our proposed work, it is concluded that in the case of occurrence of MVs in any one of the views, the advantage of using two views for diagnosis can be preserved by imputing the data in missing view using the other view.

REFERENCES

[1]. WHO Cancer Fact Sheets, 2013. [Online]. Available: http://www.iarc.fr/en/media-centre/pr/2013/pdfs/pr223_E.pdf

[2]. Rangaraj Mandayan Rangayyan, Fabio Ayres, and Leo Desautels J.E., "A Review of Computer-aided Diagnosis of Breast Cancer: Toward the Detection of Subtle Signs," IEEE Trans. Med. Imag., vol. 23, no. 3, pp. 313–320, Mar. 2004.

[3]. Byung-Woo Hong and Bong-Soo Sohn, "Segmentation of Regions of Interest in Mammograms in a Topographic Approach," IEEE Trans. Info. Tech., Biomed., vol. 14, no. 1, pp. 129-139, Jan. 2010.

[4]. Maurice Samulski and Nico Karssemeijer, "Optimizing Case-Based Detection Performance in a Multiview CAD System for Mammography," IEEE Trans. Med. Imag., vol. 30, no. 4, Apr. 2011.

[5]. Mencattini Arianna, Marcello Salmeri: "Metological Characterization of a CADx System for the Classification of Breast Masses in Mammograms," IEEE Trans. Instrum. Meas., vol. 59, pp. 2792-2799, 2010.

[6]. Harold Hotelling: "Relations Between Two Sets of Variates," Biometrika, vol. 28, pp. 321-377, 1936

BIOGRAPHIES

N. Kalimatha, Bachelor of Electronics and Communication in Amrita vishwa vidyapeetham, Coimbatore.

G. Vinutna, Bachelor of Electronics and Communication in Amrita vishwa vidyapeetham, Coimbatore.

B. Namrata, Bachelor of Electronics and Communication in Amrita vishwa vidyapeetham, Coimbatore.

A. Abinaya, Bachelor of Electronics and Communication in Amrita vishwa vidyapeetham, Coimbatore.

Y. Anjana, Bachelor of Electronics and Communication in Amrita vishwa vidyapeetham, Coimbatore.

R. Lavanya, Assiatant professor, Dept. of Electronics and Communication Engg., Amrita vishwa vidyapeetham, Coimbatore.