International Conference on Computational Modeling and Security (CMS 2016)

Image Mining using Association Rule for Medical Image dataset

Jyoti Deshmukh[a],* and Udhav Bhosle[b]

[a,b]*Department of Electronics and Telecommunication Engineering, Rajiv Gandhi Institute of Technology, Mumbai-400 053, India*

Abstract

The concept of data mining for discovering frequent image patterns in mammogram images using association rule is presented. Proposed method works in two phases. First phase is segmentation of digital mammogram to find region of interest (ROI). It consists of median filtering for noise removal, morphological processing for removing the background and suppressing artifacts, image enhancement techniques to improve image quality followed by region growing algorithm for complete removal of pectoral muscle. Second phase is image mining to find frequent image patterns present in mammogram images using Association rule. It consists of feature extraction, optimization by selecting most discriminating features among them, discretization of selected features and generation of transaction representation of input images. This is given as input to Apriori algorithm to generate association rules. The proposed method uses a new ESAR (Extraction of strong association rule) algorithm to obtain strong, effective and highly correlated association rules from the rules obtained using Apriori algorithm in previous step. Result shows that image mining is feasible and gives strong association rules. These association rules can be further used for effective diagnosis of mammogram images.

*Keywords :*Image mining*, Association rules, Support, Confidence, Region of interest, Correlation measures;*

1. Introduction

Every day, huge amount of satellite images, medical images and digital photographs are generated. Useful information can be revealed to the human users by analysing these images. As a result, there is a demand of image

———————

* Corresponding author. Tel.: +91-9892724542; Fax:022-26707026
E-mail address: jyoti.deshmukh@mctrgit.ac.in, udhav.bhosle@mctrgit.ac.in

mining systems, which can automatically review semantically meaningful information (knowledge) from huge amount of image data. A very large number of mammograms are generated every day in hospitals and medical centres. Process of analyzing and diagnosing mammogram has become critical. As a result, there is a need of computer aided diagnosis (CAD) system to help the physician's task. In this proposed method, we find association rules, which represent frequent pattern that occurs together in similar type of images i.e. benign or malignant mammogram images and these rules, can be used further for effective diagnosis of mammogram. In proposed method we use most discriminating features during mining process to get strong association rules. Our method makes the image mining algorithm faster, as it uses feature optimization by selecting most discriminating features and feature discretization. Proposed method works in two phases. First is segmentation of digital mammogram for finding Region of Interest (ROI) and second is finding frequent patterns in mammogram images using Apriori algorithm of Association Rule mining. We proposed a new ESAR algorithm for optimizing these generated association rules to get strong, effective and highly correlated association rules.

Agrawal et al. [1] discussed the problem of Association rule mining very first time. Beyer et al. [2] presented increasing the number of features to represent the image can create a problem. Thus to improve the discrimination accuracy, we have to keep the count of features as low as possible. Color distribution of image is expressed by histogram, but for breast lesions it provides poor identification. To differentiate mammogram lesions as malignant and benign, shape features can be used but it increases the computational complexity of the process. Textural variations in mammogram represent the differences in density of breast tissue. Properties of roughness, smoothness and regularity are described by texture feature.

One can query a database by using texture as a visual feature to retrieve similar patterns. The spatial arrangement of pixel intensities characterizes texture information [3][4]. Mammogram describes the tissue density differences and these differences are very important for analysis of mammogram. The differences in the density of breast tissue can be captured in a mammogram in the form of textural variations. [5] C. Ordonez et al. [6] introduced data mining for knowledge discovery in image database. They concentrate on the problem of finding association rules in 2-dimensional color images. Carson et.al. [7] presented image representation which provides a transformation from raw pixel data to a small set of localized coherent regions in color and texture space. Ji Zhang et al. [8] proposed various image mining research issues, frameworks used for image mining, current developments in image mining, state of the art techniques and systems. A framework for texture information of an image and achievement of higher retrieval efficiency than the shape features of an image is presented by Monika Sahu et al. [9]. Marcela Y. Ribeiro et al. [13] discussed a mammogram classification method based on association rule mining to improve the diagnosis of mammograms. This method produces both non-sensitive and sensitive association rules. For the diagnosis process, these non-sensitive association rules are not helpful. Also they are finding region of interest (ROI) of mammogram manually and then to these ROI, feature extraction techniques were applied.

Maria-Luiza Antonie et al. [14, 15, 16] proposed a mammogram classification method using association rules. Authors used Neural Network as a classifier and association rule mining as the data mining algorithm. Combination of a rough set theory along with association rules is used for mammogram clarification by Jiang Yun et al. [17]. Sumeet Dua et al. [18] presented weighted association rule based classification. It uses inter-class and intra-class weight of each association rule for classification. Jawad Nagi et al. [19] proposed a method using morphological processing and seeded region growing algorithm for automated breast profile segmentation.

Although many of the researchers have developed different techniques for mining of mammogram images to find strong and efficient association rules, still it is a challenging task. Hence we proposed a texture based image mining method using association rule for mammogram images. The rest of this paper is organized as follows; section 2 represents proposed image mining method. Section 3 represents Experimental result. Section 4 summarizes the conclusion and the future work.

2. Proposed image mining method

Figure 1 shows the block diagram of proposed method. It works in two phases. First is a segmentation of digital mammograms for finding region of interest (ROI) and second is finding frequent patterns in mammogram images using association rule mining. Each input mammogram image is associated with a keyword i.e. benign or malignant. Figure 2 shows algorithm for proposed method.

```
┌──────────────┐   ┌──────────────┐   ┌──────────────┐   ┌──────────────────┐
│    Input     │──▶│     Pre-     │──▶│ Thresholding │──▶│  Region Growing  │
│  Mammogram   │   │  processing  │   │ and contrast │   │                  │
│              │   │              │   │  stretching  │   │                  │
└──────────────┘   └──────────────┘   └──────────────┘   └──────────────────┘
                                                                  │
                   ┌──────────────────┐   ┌──────────────────┐    ▼
                   │   Keyword of     │◀──│ Region of        │
                   │   Mammogram      │   │ Interest (ROI)   │
                   │ Benign/Malignant │   │                  │
                   └──────────────────┘   └──────────────────┘
                            │                      │
                            │             ┌──────────────────┐
                            │             │    Feature       │
                            │             │   Extraction     │
                            │             └──────────────────┘
                            │                      │
                            │             ┌──────────────────┐
                            │             │ Feature Vector   │
                            │             │   Generation     │
                            │             └──────────────────┘
                            │                      │
                            │             ┌──────────────────┐
                            ├────────────▶│ Feature          │
                            │             │ Optimization and │
                            │             │ Discretization   │
                            │             └──────────────────┘
                            │                      │
                            │             ┌──────────────────┐
                            │             │ Optimized        │
                            │             │ Feature Vector   │
                            │             └──────────────────┘
                            │                      │
                            │             ┌──────────────────┐
                            └────────────▶│ Formation of     │
                                          │ Transaction      │
                                          │ Database         │
                                          └──────────────────┘
                                                   │
                                          ┌──────────────────┐
                                          │ Association Rule │
                                          │     Mining       │
                                          └──────────────────┘
                                                   │
                                          ┌──────────────────┐
                                          │ Association Rules│
                                          └──────────────────┘
                                                   │
                                          ┌──────────────────┐
                                          │ Extraction of    │
                                          │ Strong and       │
                                          │ Effective        │
                                          │ Association Rules│
                                          │ (ESAR Algorithm) │
                                          └──────────────────┘
```
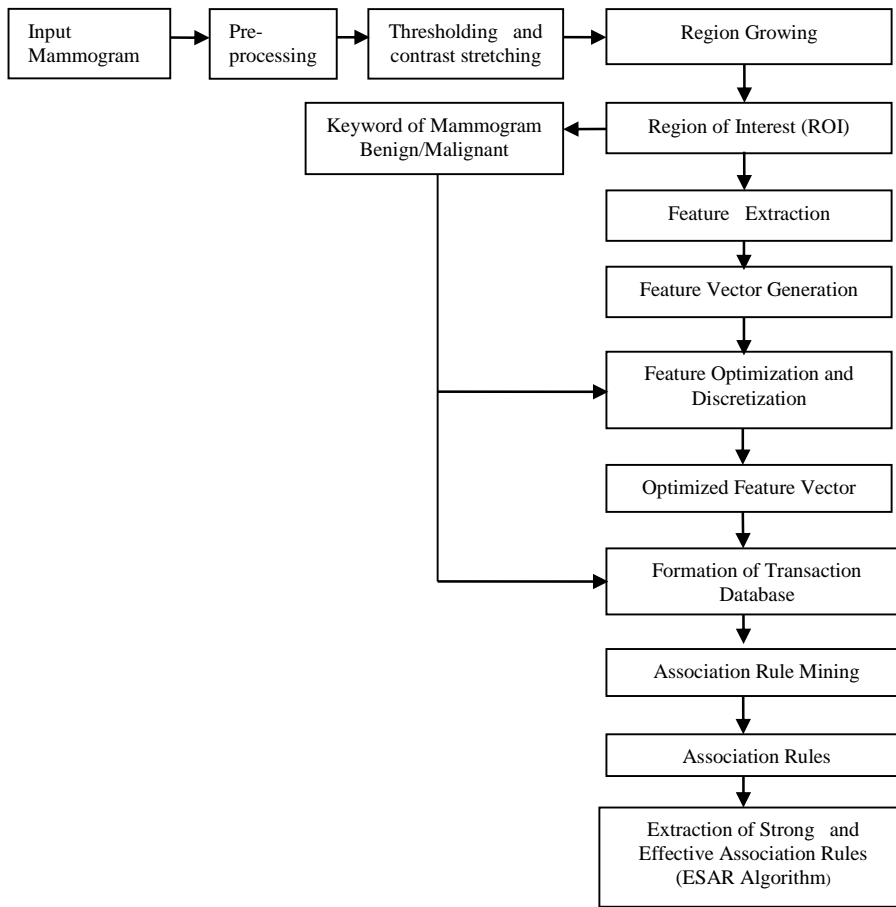
Fig. 1. Block diagram of proposed Image Mining method

Algorithm 1: Proposed Image Mining method

Input:　　Mammogram images along with keywords, minimum support and confidence threshold
Output: Strong and Effective Association Rules.
　　Step 1　: Segmentation of input mammogram.
　　1.1.　　Apply pre-processing on input images
　　1.2.　　Perform thresholding and contrast stretching
　　1.3.　　Execute region growing algorithm
　　1.4.　　Extract region of interest (ROI).
　　Step 2: Extract texture feature of segmented mammogram.
　　Step 3: Generate feature vector
　　Step 4: Apply pre-processing step for Association rule mining
　　Step 5: Generate transaction database.
　　Step 6: Apply Association rule mining using Apriori algorithm
　　Step 7: Extract strong and effective association rules using ESAR algorithm.

Fig. 2. Algorithm for proposed method

2.1. Segmentation of digital mammogram

In pre-processing step for segmentation of mammogram images, we use median filtering for the removal of digitization noise such as straight lines. The median value of the 3 by3 neighbourhood around the corresponding pixel in input images gives the corresponding output pixel value. However, we replace edges of the images by zeroes. Then we apply thresholding and morphological operations on the pre-processed mammogram for the removal of rediopaque artifacts such as labels and wedges. Through experimentation we set a global threshold with a value of T=100, which is selected for transforming gray scale image into binary [0, 1] format. For suppression of artifacts, labels and wedges, morphological operations such as dilation, erosion, opening and closing are carried out on thresholded binary image. We perform contrast enhancement on the processed mammogram image. Using the Region growing technique, pectoral muscles are segmented. For implementing region growing technique, we placed a seed inside the pectoral muscle of input mammogram image [19].

2.2. Feature extraction using GLCM and feature vector generation

Texture Features are extracted from the segmented mammogram (ROI) and organized into feature vectors. Features are extracted using Grey Level Co-occurrence Matrix (GLCM) method. For every input mammogram a GLCM matrix is generated. A co-occurrence Matrix M (d, θ) is given by the relative frequency of occurrences of two grey level pixels i and j separated by d pixels in the θ orientation. Co-occurrence matrices are calculated for the directions of 0^0, 45^0, 90^0 and 135^0 and for the 1,2,3,4 and 5 distances. Every input image is expressed by its unique feature vector. Twenty matrices of 16 x 16 integer elements per image are produced. For each matrix, seven features presented in Table 1 are calculated, producing a feature vector of 140 elements to represent each image [4], [12].

2.3. Pre-processing step for association rule mining

The feature vector and the keyword of the input mammogram images i.e. benign or malignant are submitted to pre-processing step used for association rule generation. In the pre-processing step, we perform feature selection and discretization of the continuous values of the features. Based on inconsistency we decide exact number of intervals and feature selection. As the inconsistency in feature value decreases, the number of feature value intervals becomes smaller. Pre-processing step for Association Rule Mining aims to have minimum number of feature value intervals with minimum inconsistency in feature value. Thus, from the feature vector of all input mammogram images, inconsistent and irrelevant features are removed and the most discriminative features are selected to form feature vector. These selected most discriminative feature presents the smallest class variation and then discretization of these selected features is done. Optimized feature vector for each image is produced and is given as input to transaction database.

2.4. Formation of transaction database and Association rule mining

Keywords of input mammogram images i.e. benign or malignant and optimized feature vectors are used to build the transaction database. Transaction database has a transaction record for every input image and it is submitted to Apriori algorithm. Association rule mining problem was first discussed in [1] as Let I = {i_1,…,i_n} be a set of literals called items. D be a set of data cases. An association rule is an expression of the form A$\rightarrow$ B. Where A and B are item sets. A is called body or antecedent of the rule and B is called head or consequent of the rule. An item set is a set of items, the antecedent plus the consequent. To determine the rules returned by the mining process, support and confidence measures ((1)-(2)) are used. Support value explains how frequently an association rule is applicable to a given transaction data set. Confidence explains how often items in B appear in transactions that contain A. These matrices are defined as,

$$\text{Support, S (A}\rightarrow\text{B)} = \frac{\text{No.of tuples (transactions)containing both A and B}}{\text{Total no.of tuples(transactions)}} \qquad (1)$$

$$\text{Confidence, C }(A \rightarrow B) = \frac{\text{No.of tuples (transactins)containing both A and B}}{\text{Total no.of tuples(transactions)containing A}} \tag{2}$$

If S% of the data case in D contains both A and B then the rule A→ B has support S in D. If C% of the data case in D that holds A also hold B, then the rule A→B has confidence C in D. The problem of mining association rules is to find all rules that have support and confidence value greater than some minimum support and minimum confidence threshold value specified by the user. Transaction record of all input images is given as an input to Apriori algorithm, for generating association rules. By applying minimum support and confidence threshold, Apriori algorithm generates association rules. Through experimentation minimum support and confidence threshold values are set.

2.5. Extraction of strong and effective association rules(ESAR algorithm)

Association rule mining by using Apriori algorithm gives all rules in data which satisfy the minimum support and confidence threshold value. Information interpreted varies according to rule. Many times, rule having high value of support and confidence gives conflicting or redundant information, which makes it uninteresting rule. The support and confidence measures are insufficient at filtering out uninteresting association rules. To overcome this limitation, interestingness correlation measures can be used to augment the support–confidence framework for association rules [10]. This leads to correlation rules of the form;

A→ B [support, confidence, correlation measures] (3)

There are many correlation measures listed in literature, we select interestingness correlation measures such as lift, certainty factor and completeness, as they show feasible results to find strong and effective rules generated in image mining.

$$\text{Lift }(A, B) = \frac{P(AUB)}{P(A)P(B)} \tag{4}$$

$$\text{Completeness }(A, B) = C1 = \frac{P(A \cup B)}{P(B)} \tag{5}$$

$$\text{Certainty factor }(A, B) = C2 = \max \left(\frac{P\left(\frac{B}{A}\right) - P(B)}{1 - P(B)}, \frac{P\left(\frac{A}{B}\right) - P(A)}{1 - P(A)} \right) \tag{6}$$

We apply ESAR (Extraction of strong association rules) algorithm explained in Figure 3, on all association rules generated by Apriori algorithm in earlier step to determine strong, effective and highly correlated association rules. For every rule we calculate support, confidence, and correlation measures as Lift, Completeness and Certainty Factor. Through experimentation, we set the minimum threshold value for resultant correlation measure (RCM). The rule which satisfy the minimum threshold value of RCM is called strong rule.

Algorithm 2: ESAR Algorithm (Extraction of strong association rule)

Input:　n: total number of rules, S = 0, Support, Confidence, Lift, Completeness and Certainty factor value for all rules and threshold value of RCM
Output: Strong association Rules
　　For every rule r_i, and $i \le n$
　　　　Compute supp (r_i), Conf (r_i), C1(r_i), C2(r_i)
　　　　　　If Lift $(r_i) > 1$
　　　　　　　　Then RCM $(r_i) = [\text{supp }(r_i) * \text{Conf }(r_i) * \text{C1}(r_i) * \text{C2}(r_i)]$
　　　　　　　　　　If RCM $(r_i) >$ Threshold value of RCM
　　　　　　　　　　　　Print 'Strong Association Rule'

```
                    S + +
                End
        Else RCM (rᵢ) = [supp (rᵢ) * Conf (rᵢ) * C1 (rᵢ) * C2(rᵢ)] / 4
            If RCM (rᵢ) > Threshold value of RCM
                Print 'Strong Association Rule'
                    S + +
            End
        End
    End For
```

Fig.3. ESAR algorithm

3. Experimental Result

We use Mammogram data set from Mammography Image Analysis Society (MIAS) to test proposed image mining method using association rule. We selected total 92 mammogram images, out of these 51 are benign and 41 are malignant images. In step 1 we perform automated segmentation of mammogram images to obtain region of interest (ROI). Figure 4(a-e) shows results of segmentation step. In step 2 we perform automatic textual feature extraction from the ROI by using Grey Level Co-Occurrence Matrix (GLCM) method and these features are organized into feature vectors. Feature vector of 140 elements is produced to represent each input image. Table 1 gives grey level texture features used and their positions in feature vector space. Figure 5 shows snapshot of feature vector generated where first value in each row is image number and second value onward is its feature values. In step 3 the feature vector and the keyword of the input mammogram images i.e. benign or malignant are submitted to pre-processing step used for association rule generation. In this step we perform feature selection and discretization of selected features. The most discriminative features are selected to form optimized feature vector. Total 140 features are obtained from feature extraction step for each input image. From these140 features, we get 17 most discriminative features by applying feature optimization process. Figure 6(a) shows the optimized feature vector where first column gives 17 most discriminative features i.e. feature number 22, 24,...,139. Also figure shows 32 intervals of these feature values. Second column above gives values of 32 intervals. We assign unique label for each interval which increases sequentially for next interval.

In step 4 we generate the transaction database and perform Association rule mining by using Apriori algorithm. We use 1001 as a keyword for benign image and 1002 as a keyword for malignant image. Thus each input mammogram image is having a keyword i.e. benign or malignant along with unique labels for intervals for 17 optimized feature values. Transaction database has a record for every input image. Figure 6(b) shows snapshot of a notepad where transaction database is stored. Transaction record of all input images is given as input to Apriori algorithm for generating association rules. The value of minimum support is set to 5% and value of minimum confidence is set to 90%. We get total 591 rules for this dataset.

Fig. 4. Mammogram segmentation process results: (a) original mammogram image; (b) filtered mammogram image after noise removal; (c) thresholded mammogram image; (d) mammogram image after contrast enhancement; (e) final segmented mammogram image

Table 1. Texture features and their positions in feature vector

Feature	Equation	Meaning	Position
Step	$\sum_i \sum_j P(i,j)$	Distribution	1-20
Variance	$\sum_i \sum_j P(i-j)^2 P(i,j)$	Contrast	21-40
Entropy	$\sum_i \sum_j P(i,j)\log(P(i,j))$	Suavity	41-60
Energy	$\sum_i \sum_j P(i-j)^2$	Uniformity	61-80
Homogeneity	$\sum_i \sum_j \frac{p(i-j)}{(1+|i-j|)}$	Homogeneity	81-100
3° Moment	$\sum_i \sum_j (i-j)^3 P(i,j)$	Distortion	101-120
Inv. Variance	$\sum_i \sum_j P(i,j)/(i-j)^2$	Inv. Contrast	121-140

Fig.5. Feature vector

a

b

Fig.6. (a) optimized feature vector; (b) transaction database

Examples of association rules mined are:

8,120 ->1001 i.e. Benign Image (Support=10% and Confidence=100%)

This rule explains that image having optimized feature value interval label as 8 and 120 tend to be benign image.

3, 35, 67 ->1002 i.e. Malignant image (Support=8% and Confidence=100%)

This rule explains that image having optimized feature value interval label as 3, 35 and 67 tend to be a malignant

image. In step 5 we apply ESAR (Extraction of strong association rules) algorithm on all 591 association rules generated by Apriori algorithm in earlier step to determine strong, effective and highly correlated association rules. For every rule we calculate support, confidence, and correlation measures as Lift, Completeness and Certainty Factor. Through experimentation, the minimum threshold value for resultant correlation measure (RCM) is set to 12%. By using ESAR algorithm, we get 252 strong and effective rules for this dataset. These strong, effective and highly correlated rules can be used further for effective diagnosis of mammogram images.

4. Conclusion

We presented an image mining method using texture features for finding associations, important hidden information from medical image data set. We applied the proposed algorithm to find association rules, which help to find frequent patterns, present in digital mammograms. All generated association rules using Apriori algorithm for mammogram image database are not strong rules and few of them gives redundant information. To overcome this problem, proposed method uses a new ESAR algorithm for optimization of rules generated in earlier step to get strong, effective and highly correlated association rules. Result shows that image mining is feasible and gives 252 strong and effective association rules from 591 association rules obtained from Apriori algorithm in previous step. Thus the proposed method enhances and brings more confidence to the diagnosis process of mammograms. Further this algorithm can be easily applied on other medical image data set such as MRI images. Future work includes optimization of Association rules using Genetic algorithm and effective diagnosis of mammogram images.

References

1. Agrawal R et al., "Mining association rules between sets of items in large databases", in *proceedings of the ACM SIGMOD ICMD*, Washington DC,1993,pp 207-216.
2. Beyer,K. et al. "When is nearest neighbor meaningful?' In *Proc. Int. Conf. Database Theo. (ICDT),* 1999 pp 217-235.
3. Felipe.J. C. et al., " Retieval by content of medical images using texture for tissue identification", in *Proc. 16th IEEE Symp. Computer -Based Med Systems CBMS 2003,* New York, 2003 pp 175-180.
4. Haralick. R. M. et al.," Textural features for image classification", *IEEE Trans Syst. Man. Cybern.,* Vol. SMC-3.ppc610-621, 1973.
5. Mudigonda . N. R. and Rangayyan R.M. , " Detection of breast masses in mammograms by density slicing and texture flow-field analysis", *IEEE Trans Med Imag.* vol 20 no 12 pp 1215-1227, 2001
6. Ordonez C. and Omiecinski E., "Discovering Association Rules Based on Image Content", in *Proceedings of the IEEE Advances in Digital Libraries Conference (ADL'99)*, 1999.
7. Carson, Chad, Serge Belongie, Hayit Greenspan, and Jitendra Malik. "Region-Based Image Querying", in *Content-Based Access of Image and Video Libraries, 1997. Proceedings. IEEE Workshop on*, pp. 42-49. IEEE, 1997.
8. Ji Zhang, Wynne Hsu, Mong Li Lee, "Image Mining: Trends and Developments", in *Proceedings of Journal of Intelligent Information Systems,* 19:1, pp. 7-23, 2002.
9. Sahu Monika, Shrivastava Madhup, "Image Mining: A New Approach for data mining based on texture", in *Proceedings of IEEE International Conference on Computer and Communication Technology,*2012
10. Jiawei Han, M. Kamber, J. Pei, "Data Mining Concepts and Techniques", Morgan Kaufmann publishers, 2012.
11. R. Gonzalez and R. Woods., "Digital Image Processing", Pearson Addison-Wesley Publications Co., Second Edition, March 1992.
12. Felipe J. C., Traina A. J. M., and Traina C., "Retrieval by content of medical images using texture for tissue identification", in *Proceedings of 16th IEEE Symp. Computer based Medical System, CBMS 2003,* New York, 2003, pp. 175-180.
13. Marcela X. Ribeiro, Agna J.M. Traina, Caetano Traina,Jr., and Paulo M. Azevedo-Marques, "An Association Rule-Based method to support medical images diagnosis with efficiency.", *IEEE Transactions on Multimedia*, Vol.10,No. 2, pp. 277-285, February 2008.
14. Maria-Luiza Antonie, Osmar R. Zaiane, Alexandru Coman, "Application of Data Mining Techniques for Medical Image Classification", in *Proceedings of Second International Workshop on Multimedia Data Mining (MDM/KDD'2001)* in conjunction with ACM SIGKDD conference, SanFrancisco, USA, Aug 26, 2001; pp. 94-101 .
15. Maria-Luiza Antonie, Osmar R. Zaiane, Alexandru Coman, "Associative Classifier for Medical Images", LNICS, Vol. 2797, MMCD, Berlin/Heidelberg: Springer; 2003, pp. 68-83.
16. Maria-Luiza Antonie, Osmar R. Zaiane, Alexandru Coman, "Mammography classification by an association rule-based classifier", in *Proceedings of Third International Workshop on Multimedia Data Mining*, 2002, pp. 62-69.
17. Jiang Yun et al., "Joining associative classifier for medical images," *Hybrid Intelligent Systems, 2005. HIS '05.*
18. Sumeet Dua, Harpreet Sigh and H. W. Thompson, "Associative Classification of Mammograms using weighted rules", Expert System Application: An International Journal, Volume 36, Issue 5, 2009 July, pp. 9250-9259
19. Jawad Nagi et al., "Automated Breast Profile Segmentation for ROI Detection Using Digital Mammograms", in *Proceedings of IEEE EMBS Conference on Biomedical Engineering and Science(IECBES 2010), KulaLumpur, Malaysia,* pp87-92, December 2010.