Data mining for Breast Cancer Classification and Diagnosis Using Texture Features

[1] Amalendu Bag, [2] Manmohan Sahoo, [3] Manas ranjan ojha, [4] Aswini Kumar Mohanty

[1] Regional College of Management, BBSR
[2] Kmbbcet,Khurda
[3] Silicon Institute of Technology,BBSR
[4] The Techno School

E-mail: [1] amalendu.bag@gmail.com, [2] for_manumohan@yahoo.co.in, [3] cet.manas@gmail.com, [4]asw_moh@yahoo.com

Abstract: The data mining technique deals with the extraction of implicit knowledge with data relationship or other patterns not explicitly stored in the dataset. The main objective of this paper is to apply data mining on features extracted from mammograms to classify and detect the cancerous tissue. The data mining techniques are generally more suitable to larger databases than the one used for small dataset tests. Generally association rule algorithms adopt an iterative method to discovery frequent item set, which requires very large calculations and a complicated transaction process. Because of this, a modified association rule algorithm is proposed in this paper. Experimental results show that this method can quickly discover frequent item sets and effectively mine potential association rules. Texture features are most vulnerable features that include histogram intensity features and GLCM features which are extracted from mammogram images. A new approach of feature subset selection FSFCN is proposed which approximately reduces 50 to 60% of the features and the proposed association rule is used for classification. The most interesting one is that oscillating search algorithm which is used for feature selection provides the best optimal features and no where it is applied or used for GLCM feature selection from mammogram, Experiments have been taken from a data set of 322 images of MIAS of different types with the aim to improving the accuracy by generating minimum no. of rules to cover more patterns. The accuracy obtained by this method is approximately 97.82% which is highly encouraging.

Keywords: Mammogram, Gray Level Co-occurrence Matrix feature, Histogram Intensity, Feature Selection based on Feature Correlation Networks FSFCN, Association rule mining.

1. Introduction

Breast Cancer is one of the most common cancers, leading to cause of death among women, especially in developed countries. There is no primary prevention since cause is still not understood. So, early detection of the stage of cancer allows treatment which could lead to high survival rate. Mammography is currently the most effective imaging modality for breast cancer screening. However, 10-30% of breast cancers are missed at mammography [1]. Mining information and knowledge from large database has been recognized by many researchers as a key research topic in database system and machine learning Researches that use data mining approach in image learning can be found in [2,3].

Data mining of medical images is used to collect effective models, relations, rules, abnormalities and patterns from large volume of data. This procedure can accelerate the diagnosis process and decision-making. Different methods of data mining have been used to detect and classify anomalies in mammogram images such as wavelets [4,5], statistical methods and most of them used feature extracted using image processing techniques [6].Some other methods are based on fuzzy theory [7,8] and neural networks [9]. In this paper we have used classification method called Decision tree classifier for image classification [10-12].

Classification process typically involves two phases: training phase and testing phase. In training phase the properties of typical image features are isolated and based on this training class is created .In the subsequent testing phase , these feature space partitions are used to classify the image. A block diagram of the method is shown in figure1.

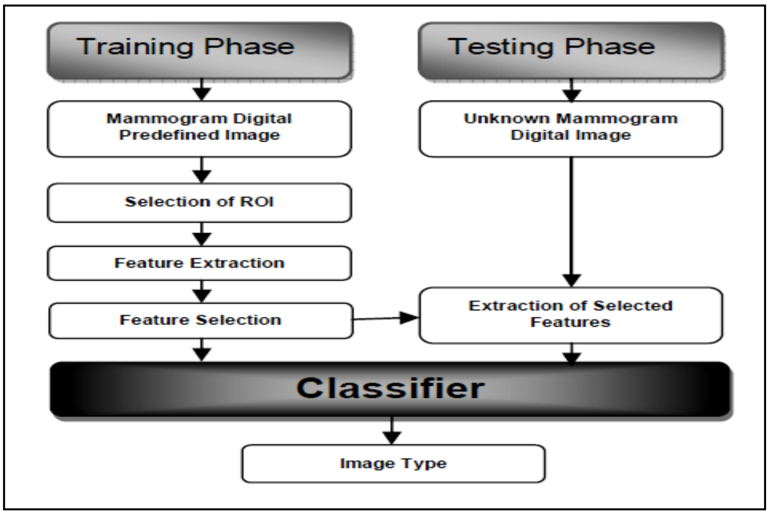


Fig.1: Block diagram for mammogram classification system

We have used association rule mining using image content method by extracting low level image features for classification. The merits of this method are effective feature extraction, selection and efficient classification. The rest of the paper is organized as follows. Section 2 presents the preprocessing and section 3 presents the feature extraction phase. Section 4 discusses the proposed method of Feature selection and classification. In section5 the results are discussed and conclusion is presented in section 6.

2. Methodologies

2.1 Digital mammogram database

The mammogram images used in this experiment are taken from the mini mammography database of MIAS (http://peipa.essex.ac.uk/ipa/pix/mias/). In this database, the original MIAS database are digitized at 50 micron pixel edge and has been reduced to 200 micron pixel edge and clipped or padded so that every image is 1024 X 1024 pixels. All images are held as 8-bit gray level scale images with 256 different gray levels (0-255) and physically in portable gray map (pgm) format. This study solely concerns the detection of masses in mammograms and, therefore, a total of 100 mammograms comprising normal, malignant and benign case were considered. Ground truth of location and size of masses is available inside the database.

.

2.2. Pre-processing

The mammogram image for this study is taken from Mammography Image Analysis Society (MIAS), which is an UK research group organization related to the Breast cancer investigation [13]. As mammograms are difficult to interpret, preprocessing is necessary to improve the quality of image and make the feature extraction phase as an easier and reliable one. The calcification cluster/tumor is surrounded by breast tissue that masks the calcifications preventing accurate detection and shown in Figure.3. .A pre-processing; usually noise-reducing step [14] is applied to improve image and calcification contrast figure 3. In this work the efficient filter (CLAHE) was applied to the image that maintained calcifications while suppressing unimportant image features. Figures 3 shows representative output image of the filter for a image cluster in figure 2. By comparing the two images, we observe background mammography structures are removed while calcifications are preserved. This simplifies the further tumor detection step.

Contrast limited adaptive histogram equalization (CLAHE) method seeks to reduce the noise produced in homogeneous areas and was originally developed for medical imaging [15]. This method has been used for enhancement to remove the noise in the pre-processing of digital mammogram [16]. CLAHE operates on small regions in the image called tiles rather than the entire image. Each tile’s contrast is enhanced, so that the histogram of the output region approximately matches the uniform distribution or Rayleigh distribution or exponential distribution. Distribution is the desired histogram shape for the image tiles. The neighboring tiles are then combined using bilinear interpolation to eliminate artificially induced boundaries. The contrast, especially in homogeneous areas, can be limited to avoid amplifying any noise that might be present in the image. The block diagram of pre-processing is shown in Figure 4.

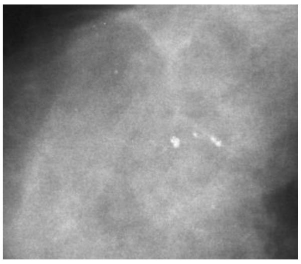


Fig 2: ROI of a Benign

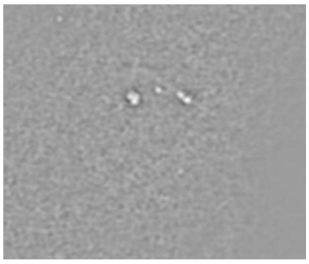


Fig 3: ROI after Pre-processing Operation

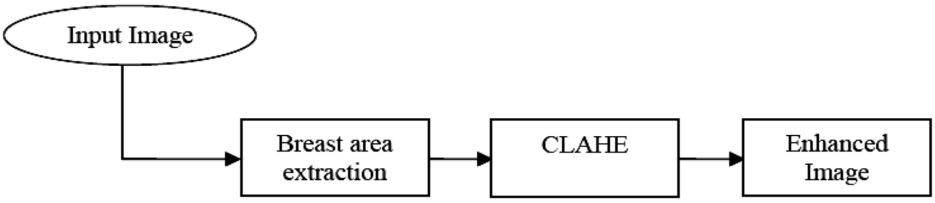


Fig 4: Image pre-processing block diagram.

2.3. *Histogram Equalization*

Histogram equalization is a method in image processing of contrast adjustment using the image's histogram [17]. Through this adjustment, the intensities can be better distributed on the histogram. This allows for areas of lower local contrast to get better contrast. Histogram equalization accomplishes this by efficiently spreading out the most frequent intensity values. The method is useful in images with backgrounds and foregrounds that are both bright or both dark. In particular, the method can lead to better views of bone structure in x-ray images, and to better detail in photographs that are over or under-exposed. In mammogram images Histogram equalization is used to make contrast adjustment so that the image abnormalities will be better visible.

[†] peipa.essex.ac.uk/info/**mias**.html

3. Feature extraction

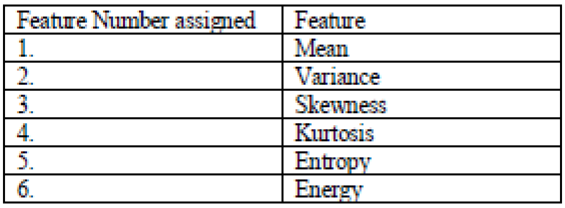
Features, characteristics of the objects of interest, if selected carefully are representative of the maximum relevant information that the image has to offer for a complete characterization a lesion [18, 19]. Feature extraction methodologies analyze objects and images to extract the most prominent features that are representative of the various classes of objects. Features are used as inputs to classifiers that assign them to the class that they represent.

In this Work intensity histogram features and Gray Level Co-Occurrence Matrix (GLCM) features are extracted.

3.1 Intensity Histogram Features

Intensity Histogram analysis has been extensively researched in the initial stages of development of this algorithm [18,20]. Prior studies have yielded the intensity histogram features like mean, variance, entropy etc. These are summarized in Table 1 Mean values characterize individual calcifications; Standard Deviations (SD) characterize the cluster. Table 2 summarizes the values for those features.

Table 1: Intensity histogram features



Feature Number assigned	Feature
1.	Mean
2.	Variance
3.	Skewness
4.	Kurtosis
5.	Entropy
6.	Energy

In this paper, the value obtained from our work for different type of image is given as follows:

Table 2: Intensity histogram features and their values

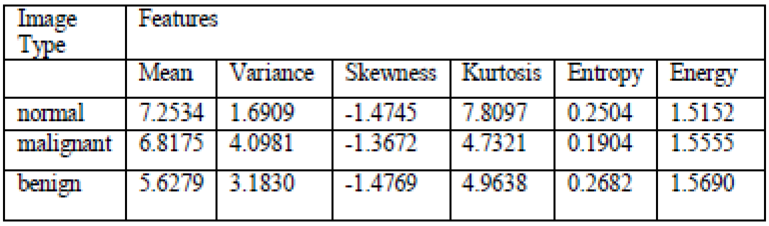


Image Type	Features					
	Mean	Variance	Skewness	Kurtosis	Entropy	Energy
normal	7.2534	1.6909	-1.4745	7.8097	0.2504	1.5152
malignant	6.8175	4.0981	-1.3672	4.7321	0.1904	1.5555
benign	5.6279	3.1830	-1.4769	4.9638	0.2682	1.5690

3.2 ***GLCM Features***

It is a statistical method that considers the spatial relationship of pixels is the gray-level co-occurrence matrix (GLCM), also known as the gray-level spatial dependence matrix [21,22]. By default, the spatial relationship is defined as the pixel of interest and the pixel to its immediate right (horizontally adjacent), but you can specify other spatial relationships between the two pixels. Each element (*I*, *J*) in the resultant GLCM is simply the sum of the number of times that the pixel with value *I* occurred in the specified spatial relationship to a pixel with value *J* in the input image.

3.2.1 GLCM Construction

GLCM is a matrix **S** that contains the relative frequencies with two pixels: one with gray level value i and the other with gray level j-separated by distance d at a certain angle θ occurring in the image. Given an image window W(x, y, c), for each discrete values of d and θ, the GLCM matrix **S**(i, j, d, θ) is defined as follows.

An entry in the matrix **S** gives the number of times that gray level i is oriented with respect to gray level j such that W(x_1, y_1)=i and W(x_2, y_2)=j, then

$(x_2, y_2) = (x_1, y_1) + (d\cos\theta,\ d\sin\theta)$

We use two different distances d={1, 2} and three different angles θ={0°, 45°, 90°}. Here, angle representation is taken in clock wise direction.

Example

Intensity matrix

$$\begin{vmatrix} 1 & 3 & 1 & 1 & 1 \\ 2 & 2 & 4 & 2 & 1 \\ 1 & 4 & 1 & 4 & 1 \\ 2 & 2 & 2 & 1 & 1 \\ 1 & 1 & 2 & 2 & 1 \end{vmatrix}$$

$\boxed{\text{for } \theta = 45^{\circ} \text{ and } d = 1}$

and

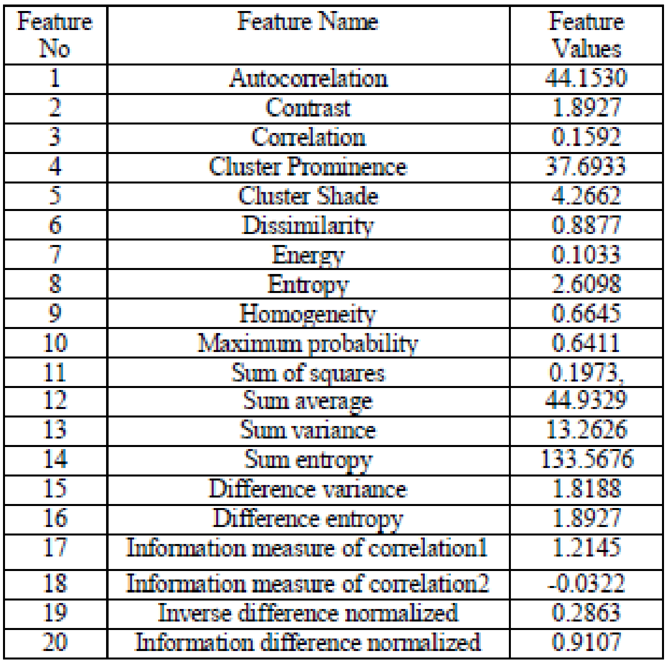
$$\begin{bmatrix} 3 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

$\boxed{\text{for } \theta = 45^{\circ} \text{ and } d = 2\,.}$

The Following GLCM features were extracted in our research work:

Autocorrelation, Contrast, Correlation, Cluster Prominence, Cluster Shade, Dissimilarity Energy, Entropy, Homogeneity, Maximum probability, Sum of squares, Sum average, Sum variance, Sum entropy, Difference variance, Difference entropy, information measure of correlation1, information measure of correlation2, Inverse difference normalized. Information difference normalized. The value obtained for the above features from our work for a typical image is given in the following table 3.

Table 3: GLCM Features and values Extracted from Mammogram Image(Malignant)



Feature No	Feature Name	Feature Values
1	Autocorrelation	44.1530
2	Contrast	1.8927
3	Correlation	0.1592
4	Cluster Prominence	37.6933
5	Cluster Shade	4.2662
6	Dissimilarity	0.8877
7	Energy	0.1033
8	Entropy	2.6098
9	Homogeneity	0.6645
10	Maximum probability	0.6411
11	Sum of squares	0.1973,
12	Sum average	44.9329
13	Sum variance	13.2626
14	Sum entropy	133.5676
15	Difference variance	1.8188
16	Difference entropy	1.8927
17	Information measure of correlation1	1.2145
18	Information measure of correlation2	-0.0322
19	Inverse difference normalized	0.2863
20	Information difference normalized	0.9107

4. Feature subset selection

Feature subset selection helps to reduce the feature space which improves the prediction accuracy and minimizes the computation time [23]. This is achieved by removing irrelevant, redundant and noisy features .i.e., it selects the subset of features that can achieve the best performance in terms of accuracy and computation time. It performs the Dimensionality reduction.

Features are generally selected by search procedures. A number of search procedures have been proposed. Popularly used feature selection algorithms are Sequential forward Selection, Sequential Backward Selection, Genetic Algorithm and Particle Swarm Optimization, Branch and Bound feature optimization. In this work a new approach of Feature Selection based on Feature Correlation Networks is proposed to select the optimal features. The selected optimal features are considered for classification. The FSFCN search has been fully exploited to select the feature from mammogram which is one of the best techniques to optimize the features among many features. We have attempted to optimize the feature of GLCM and statistical features.

4.1 FSFCN: Feature Selection based on Feature Correlation Networks

The method for feature selection proposed in this paper, denoted by FSFCN, is based on the notion of feature correlation networks. A feature correlation network describes correlations between features in a dataset that are equal or higher than a specified threshold. To formally define feature correlation networks, we will assume that a dataset is composed of data instances having numeric features and a categorical class variable. The below stated definition of feature correlation networks can be adapted in a straightforward manner for other types of datasets (categorical features, a mix of categorical and numeric features, continuous target variable) by taking appropriate correlation measures.

Definition 1 (Feature Correlation Network).

Let D be a dataset composed of data instances described by k real-valued features Definition 1 (Feature Correlation Network). Let D be a dataset composed of data instances described by k real-valued features f1, f2, . . . , fk ∈ R and a categorical class variable c. Let Cf : R×R → [−1, 1] denote a correlation measure applicable to features (e.g the Pearson or Spearman correlation coefficient) and let Cc be a correlation measure applicable to a feature and the class variable (e.g. the mutual information, the Goodman-Kruskal index, etc.). The feature correlation network corresponding to D is an undirected, weighted, attributed graph G = (V, E) with the following properties:

– The set of nodes V corresponds to the set of features (fi ∈ V for each i in[1 .. k]).

– Two features fi and fj , i = j, are connected by an edge ei,j in G, ei,j ∈ E, if |Cf (fi, fj)| ≥ T , where T is previously given threshold indicating a significant correlation between features. The weight of ei,j is equal to |Cf (fi, fj)|.

– Each node in the network has a real-valued attribute reflecting its association with the class variable which is measured by Cc.

The features in D can be ranked according to the Cc measure and highly ranked features can be considered as the most relevant for training a classifier.

Definition 2 (Subset of Relevant Features). A subset Fr of the set of features F is called relevant if (∀f ∈ Fr) Cc(f) ≥ R where R denotes a threshold indicating a significant association between a feature and the class variable.

Definition 3 (Pruned Feature Correlation Network). A pruned feature correlation network is a feature correlation network constructed from a subset of relevant features.

Our implementation of the FSFCN method for datasets with real-valued features and categorical class variables uses pruned feature correlation networks which are constructed without explicitly stating the threshold T. This means that the algorithm for constructing pruned correlation networks has only one parameter R separating relevant from irrelevant features. Additionally, the algorithm uses the Spearman correlation coefficient to determine correlations among relevant features (the Cf measure), while correlations between relevant features and the class variable are quantified by their mutual information (the Cc measure). The mutual information between a real-valued feature f and the categorical class variable c, denoted by I(f, c), can be approximated by

$$I(f,c) \approx \sum_{y \in c} \sum_{x \in f'} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right),$$

where f' is the set of discrete values obtained by a discretization of f, p(x, y) is the joint probability distribution function of f' and c, and p(x) and p(y) are the marginal probability distribution functions of f' and c, respectively. I(f, c) equal to 0 means that f and c are totally unrelated. A larger value of I(f, c) implies a stronger association between f and c.

The algorithm for constructing pruned correlation network consists of the following steps (see Algorithm 1):

1. The subset of relevant features Fr is determined using the mutual information measure. Then, the nodes of the network are created such that each node corresponds to one feature from Fr.
2. For each pair of relevant features fi and fj , the algorithm forms a list L,

where elements are tuples in the form (fi, fj , Sij), where Sij denotes the value of the Spearman correlation coefficient between features fi and fj .

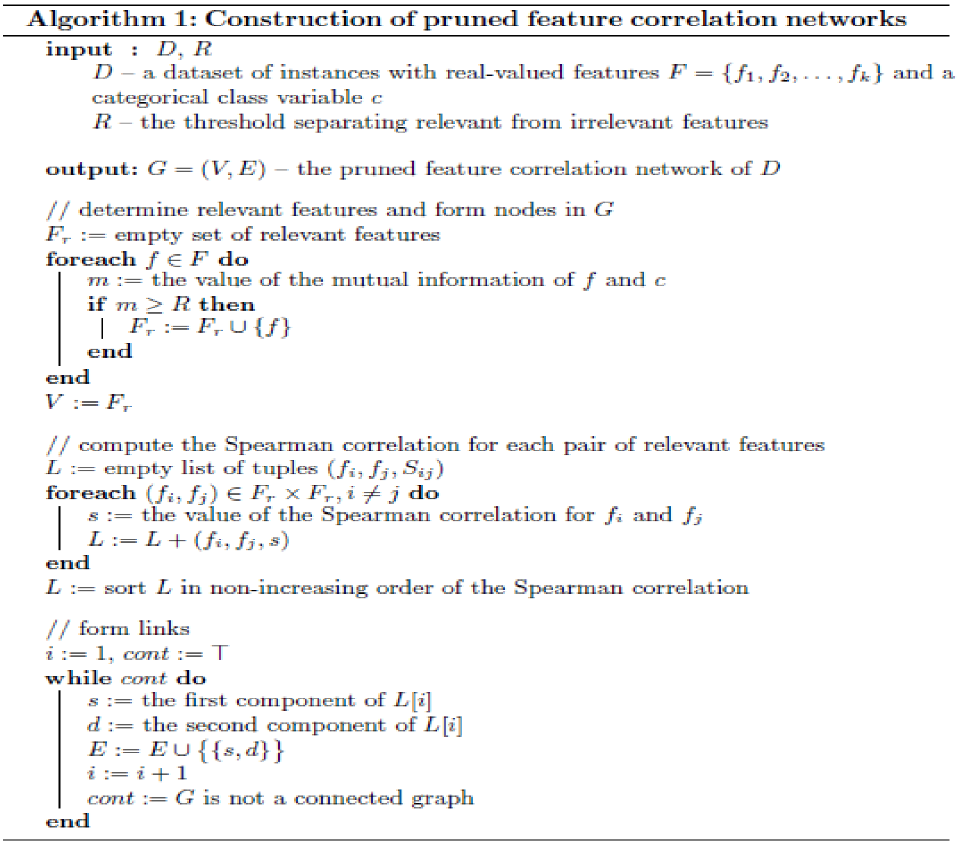
3. L is sorted by the third component (Sij) in decreasing order, i.e. the first element of the sorted list is the pair of features exhibiting the highest correlation, while the last element is the pair of features with the lowest correlation.

4. In the last step, the algorithm forms the links of the network by iterating through the sorted list L beginning from the first element. Let ek = (fi, fj , Sij) denotes the element processed in the k-th iteration. The algorithm forms a link lij connecting fi and fj with weight Sij . If the addition of lij results in a connected graph (i.e., a graph that has exactly one connected component or, equivalently, a graph in which there is path between each pair of nodes) then the algorithm stops, otherwise it goes to the next element in the sorted list and repeats the same procedure. In other words, the algorithm iteratively builds the network by connecting features having the highest correlation until the network becomes a connected graph. Consequently, the weight of the last added link determines the value of the threshold T

The basic idea of the FSFCN method is to cluster a pruned feature correlation network in order to obtain cohesive groups of relevant features such that correlations between features within a group are stronger than correlations between features belonging to different groups. The FSFCN method leans on com- munity detection techniques to identify clusters in feature correlation networks. The development of community detection techniques started with Newman and Girvan [18] who introduced a measure called modularity to estimate the quality of a partition of a network into communities. The main idea behind the modularity measure is that a subgraph can be considered a community if the actual number of links connecting nodes within the subgraph is significantly higher than the expected number of links with respect to some null random graph model. In the case of weighted networks, modularity accumulates differences between the total weight of links within a community and the mathematical expectation of the previous quantity with respect to a random network having the same degree and link weight distribution [17].

Definition 4 (Modularity). For weighted networks modularity Q is defined as

$$Q = \sum_{c=1}^{n_c} \left[\frac{W_c}{W} - \left(\frac{S_c}{2W} \right)^2 \right],$$



Algorithm 1: Construction of pruned feature correlation networks

```
input : $D$, $R$
        $D$ – a dataset of instances with real-valued features $F = \{f_1, f_2, \dots, f_k\}$ and a
        categorical class variable $c$
        $R$ – the threshold separating relevant from irrelevant features

output: $G = (V, E)$ – the pruned feature correlation network of $D$

// determine relevant features and form nodes in $G$
$F_r$ := empty set of relevant features
foreach $f \in F$ do
    $m$ := the value of the mutual information of $f$ and $c$
    if $m \geq R$ then
        $F_r := F_r \cup \{f\}$
    end
end
$V := F_r$

// compute the Spearman correlation for each pair of relevant features
$L$ := empty list of tuples $(f_i, f_j, S_{ij})$
foreach $(f_i, f_j) \in F_r \times F_r, i \neq j$ do
    $s$ := the value of the Spearman correlation for $f_i$ and $f_j$
    $L := L + (f_i, f_j, s)$
end
$L$ := sort $L$ in non-increasing order of the Spearman correlation

// form links
$i := 1$, $cont := \top$
while $cont$ do
    $s$ := the first component of $L[i]$
    $d$ := the second component of $L[i]$
    $E := E \cup \{\{s, d\}\}$
    $i := i + 1$
    $cont$ := $G$ is not a connected graph
end
```

where nc is the number of communities in the network, Wc is the sum of weights of intra community links in c, Sc is the total weight of links incident to nodes in c, and W is the total weight of links in the network. Four different community detection algorithms provided by the iGraph library [24] are used to detect non-overlapping communities in feature correlation networks:

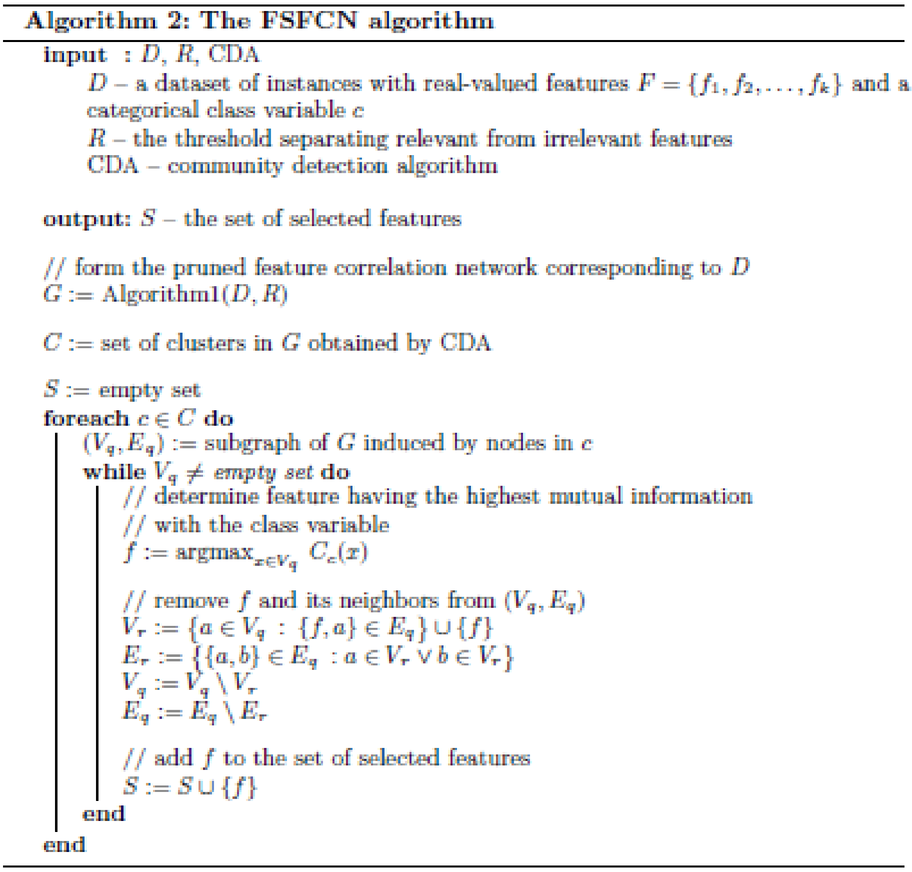
1. The Greedy Modularity Optimization (GMO) algorithm [25]. This algorithm relies on a greedy hierarchical agglomeration strategy to maximize modularity. The algorithm starts with the partitioning in which each node is assigned to a singleton cluster. In each iteration of the algorithm, the variation of modularity obtained by merging any two communities is computed. The merge operation that maximally increases (or minimally decreases) modularity is chosen and the merge of corresponding clusters is performed.
2. The Louvain algorithm [26]. This method is an improvement of the previous method. The algorithm uses a greedy multi-resolution strategy to maximize modularity starting from the partition in which all nodes are put in different communities. When modularity is optimized locally by moving nodes to neighboring clusters, the algorithm creates a network of communities and then repeats the same procedure on that network until a maximum of modularity is obtained.
3. The Walktrap algorithm [27]. This algorithm relies on a node distance mea- sure reflecting probability that a random walker moves from one node to another node in exactly k steps (k is the only parameter of the algorithm having default value k = 4). The clustering dendrogram is constructed by Ward’s agglomerative clustering technique and the partition which maxi- mizes modularity is taken as the output of the algorithm.

4. The Infomap algorithm [28]. This method reveals communities by optimally compressing descriptions of information flows on the network. The algorithm uses a greedy strategy to minimize the map equation which reflects the expected description length of a random walk on a partitioned network.

Each of used community detection algorithms defines one concrete implementation instance (i.e. one variant) of the FSFCN method.

The final step in the FSFCN method is the selection of features according to obtained community partitions in pruned feature correlation networks. The main idea is to select one or more features within each community such that:

1. selected features have a strong association with the class variable, and
2. any two selected features belonging to the same community are not directly connected.



```
Algorithm 2: The FSFCN algorithm
input  : $D$, $R$, CDA
      $D$ – a dataset of instances with real-valued features $F = \{f_1, f_2, \dots, f_k\}$ and a
      categorical class variable $c$
      $R$ – the threshold separating relevant from irrelevant features
      CDA – community detection algorithm

output: $S$ – the set of selected features

// form the pruned feature correlation network corresponding to $D$
$G :=$ Algorithm1($D, R$)

$C :=$ set of clusters in $G$ obtained by CDA

$S :=$ empty set
foreach $c \in C$ do
    $(V_q, E_q) :=$ subgraph of $G$ induced by nodes in $c$
    while $V_q \neq$ empty set do
        // determine feature having the highest mutual information
        // with the class variable
        $f := \text{argmax}_{x \in V_q} \; C_c(x)$

        // remove $f$ and its neighbors from $(V_q, E_q)$
        $V_r := \{a \in V_q \; : \; \{f, a\} \in E_q\} \cup \{f\}$
        $E_r := \{\{a, b\} \in E_q \; : a \in V_r \vee b \in V_r\}$
        $V_q := V_q \setminus V_r$
        $E_q := E_q \setminus E_r$

        // add $f$ to the set of selected features
        $S := S \cup \{f\}$
    end
end
```

The procedure for forming the set of selected features is described in Algorithm 2.

After the pruned correlation network is constructed and clustered, the FS- FCN method forms subgraphs of the network corresponding to identified com- munities where one subgraph is induced by nodes belonging to one community. For each of community subgraphs the following operations are performed:

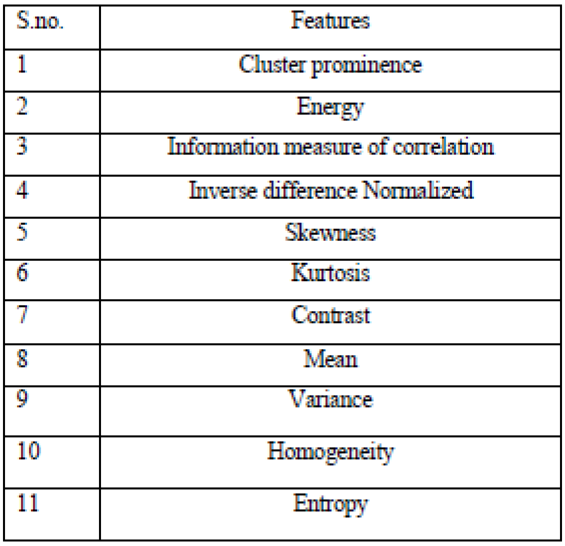
1. A feature having the highest association with the class variable is identi- fied and put in the set of selected features. Then, it is removed from the community subgraph together with its neighbors.

2. The previous step is repeated while the community subgraph is not empty.

In other words, for each of identified communities the FSFCN method selects one or more features which represent the whole community. The method also takes into account the size of communities – for larger communities a higher number of features is selected. Also, when a feature is added to the set of selected features its neighbors are removed from the community subgraph which implies that the set of selected features will not contain features having a high mutual correlation (otherwise, such two features would be directly connected in the community subgraph).

This search algorithm must give productive results in conjunction with better accuracy positively. The features selected by the method are given in table 4.

Table 4: Feature selected by proposed method



S.no.	Features
1	Cluster prominence
2	Energy
3	Information measure of correlation
4	Inverse difference Normalized
5	Skewness
6	Kurtosis
7	Contrast
8	Mean
9	Variance
10	Homogeneity
11	Entropy

5. Classification

5.1 Preparation of Transactional Database:

The selected features are organized in a database in the form of transactions [31], which in turn constitute the input for deriving association rules. The transactions are of the form

[Image ID, F1*; F*2*; :::; F11*] where *F*1*:::F11* are *11*features extracted for a given image.

5.2 Association Rule Mining: Discovering frequent item sets is the key process in association rule mining.

In order to perform data mining association rule algorithm, numerical attributes should be discretized first, i.e. continuous attribute values should be divided into multiple segments. Traditional association rule algorithms adopt an iterative method to discovery, which requires very large calculations and a complicated transaction process. Because of this, a new association rule algorithm is proposed in this paper. This new algorithm adopts a Boolean vector method to discovering frequent item sets. In general, the new association rule algorithm consists of four phases as follows:

1. Transforming the transaction database into the Boolean matrix.
2. Generating the set of frequent 1-itemsets L1.
3. Pruning the Boolean matrix.
4. Generating the set of frequent k-item sets Lk(k>1).

The detailed algorithm, phase by phase, is presented below:

1. *Transforming the transaction database into the Boolean matrix:* The mined transaction database is *D*, with *D* having m transactions and *n* items. Let T={T1,T2,…,Tm} be the set of transactions and I={I1,I2,…,In}be the set of items. We set up a Boolean matrix Am*n, which has m rows and n columns. Scanning the transaction database *D*, we use a binning procedure to convert each real valued feature into a set of binary features. The 0 to 1 range for each feature is uniformly divided into k bins, and each of *k* binary features record whether the feature lies within corresponding range.

2. *Generating the set of frequent 1-itemset L1:* The Boolean matrix Am*n is scanned and support numbers of all items are computed. The support number Ij.supth of item Ij is the number of '1s' in the jth column of the Boolean matrix Am*n. If Ij.supth is smaller than the minimum support number, itemset {Ij} is not a frequent 1-itemset and the jth column of the Boolean matrix Am*n will be deleted from Am*n. Otherwise itemset {Ij} is the frequent 1-itemset and is added to the set of frequent 1-itemset L1. The sum of the element values of each row is recomputed, and the rows whose sum of element values is smaller than 2 are deleted from this matrix.

3. *Pruning the Boolean matrix:* Pruning the Boolean matrix means deleting some rows and columns from it. First, the column of the Boolean matrix is pruned according to Proposition 2. This is described in detail as: Let I• be the set of all items in the frequent set LK-1, where k>2. Compute all |LK-1(j)| where j belongs to I2, and delete the column of correspondence item j if |*LK* – 1(*j*)| is smaller than *k* – 1. Second, re-compute the sum of the element values in each row in the Boolean matrix. The rows of the Boolean matrix whose sum of element values is smaller than k are deleted from this matrix.

4. *Generating the set of frequent k-itemsets Lk:* Frequent k-item sets are discovered only by "and" relational calculus, which is carried out for the k-vectors combination. If the Boolean matrix *Ap*q* has q columns where 2 < *q* £ *n* and *minsupt*h £ *p* £ *m, k q c,* combinations of k-vectors will be produced. The 'and' relational calculus is for each combination of k-vectors. If the sum of element values in the "and" calculation result is not smaller than the minimum support number *minsupth*, the k-itemsets corresponding to this combination of kvectors are the frequent k-itemsets and are added to the set of frequent k-itemsets Lk.

6. Experimental results

In this paper we used association rule mining using image contents for the classification of mammograms. The average accuracy is 97.67 %. We have used the precision and recall measures as the evaluation metric for mammogram classification. Precision is the fraction of the number of true positive predictions divided by the total number of true positives in the set. Recall is the total number of predictions divided by the total number of true positives in the set. The testing result using the selected features is given in table 5. The selected features are used for classification. For classification of samples, we have employed the freely available Machine Learning package, WEKA [32]. Out of 322 images in the dataset, 208 were used for training and the remaining 114 for testing purposes.

Table 5: Results obtained by proposed method

Normal	98.08%
Malignant	96.07%
Benign	100%

The confusion matrix has been obtained from the testing part .In this case for example out of 97 actual malignant images 07 images was classified as normal. In case of benign and normal all images are correctly classified. The confusion matrix is given in Table 8.

Table 6: Confusion matrix

Actual	Predicted class		
	Benign	Malignant	Normal
Benign	62	0	0
Malignant	51	48	03
Normal	209	04	205

The following graph shows the comparative analysis of our method and various other methods.

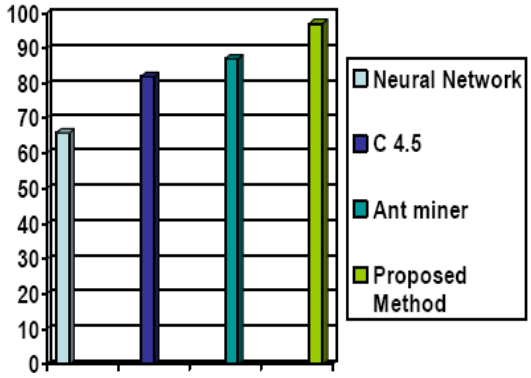


Fig 5: Performance of the Classifier

7. Conclusion

Mammography is one of the primary methods to detect breast cancer, but in some cases radiologists find difficulty in diagnosing the tumors. We have described a comparatively suitable method for the specific application in detection of micro calcifications in mammogram images. In this paper, a new method for association rule mining is proposed. The main positive impact of this method are that it only scans the transaction database once, it does not produce candidate jtemsets, and it adopts the Boolean vector “relational calculus” to discover frequent itemsets. In addition, it stores all transaction data in binary form, so it needs less memory space and can be applied to mining large databases.

Although till date a few better progress has been achieved, there are still challenges and directions for future research, such as, developing better preprocessing, enhancement and segmentation techniques; designing better feature extraction, selection and classification algorithms; integration of classifiers to reduce both false positives and false negatives; employing high resolution mammograms and investigating 3D mammograms. Mammogram image analysis society database is a standard test set but defining different standard test set (database) and better evaluation criteria are still very important. With some rigorous evaluations, and objective and fair comparison could determine the relative merit of competing algorithms and facilitate the development of better and robust systems. The methods like one presented in this paper could assist the radiologists and improve the accuracy of detection. This proposed method can reduce the computation cost of mammogram image analysis and time complexity. The algorithm uses simple statistical techniques in collaboration with a standard and efficient feature selection technique for mammogram image analysis. The value of this technique is that it not only tackles the measurement problem but also provides a visualization of the relation among features. In addition to ease of use, this approach effectively addresses the feature redundancy problem. The method proposed has been proven that it is easier and it requires less computing time than many existing methods.

References

[1]. Majid AS, de Paredes ES, Doherty RD, Sharma N Salvador X. “Missed breast carcinoma: pitfalls and Pearls”. Radiographics, pp.881-895, 2003.

[2]. Osmar R. Zaïane,M-L. Antonie, A. Coman “Mammography Classification by Association Rule based Classifier,” MDM/KDD2002 International Workshop on Multimedia Data Mining ACM SIGKDD, pp.62-69,2002,

[3]. Xie Xuanyang, Gong Yuchang, Wan Shouhong, Li Xi ,”Computer Aided Detection of SARS Based on Radiographs Data Mining “, Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference Shanghai, China, pp7459 – 7462, 2005.

[4] C.Chen and G.Lee, “Image segmentation using multitiresolution wavelet analysis and Expectation Maximum(EM) algorithm for mammography” , International Journal of Imaging System and Technology, 8(5): pp491-504,1997.

[5] T.Wang and N.Karayaiannis, "Detection of microcalcification in digital mammograms using wavelets", IEEE Trans. Medical Imaging, 17(4):498-509, 1998.

[6] Jelena Bozek, Mario Mustra, Kresimir Delac, and Mislav Grgic "A Survey of Image Processing Algorithms in Digital mammography"Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 631–657,2009

[7]. Shuyan Wang, Mingquan Zhou and Guohua Geng, "Application of Fuzzy Cluster analysis for Medical Image Data Mining" Proceedings of the IEEE International Conference on Mechatronics & Automation Niagara Falls, Canada,pp. 36 – 41,July 2005.

[8]. R.Jensen, Qiang Shen, "Semantics Preserving Dimensionality Reduction: Rough and Fuzzy-Rough Based Approaches", IEEE Transactions on Knowledge and Data Engineering, pp. 1457-1471, 2004.

[9]. I.Christiyanni et al ., "Fast detection of masses in computer aided mammography", IEEE Signal processing Magazine, pp:54- 64,2000.

[10]. Walid Erray, and Hakim Hacid, "A New Cost Sensitive Decision Tree Method Application for Mammograms Classification" IJCSNS International Journal of Computer Science and Network Security, pp. 130-138, 2006.

[11]. Ying Liu, Dengsheng Zhang, Guojun Lu, Regionbased "image retrieval with high-level semantics using decision tree learning", Pattern Recognition, 41, pp. 2554 – 2570, 2008.

[12]. Kemal Polat , Salih Gu¨nes, "A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems", Expert Systems with Applications, Volume 36 Issue 2, pp.1587-1592, March, 2009, doi:10.1016/j.eswa.2007.11.051

[13]. Etta D. Pisano, Elodia B. Cole Bradley, M. Hemminger, Martin J. Yaffe, Stephen R. Aylward, Andrew D. A. Maidment, R. Eugene Johnston, Mark B. Williams,Loren T. Niklason, Emily F. Conant, Laurie L. Fajardo,Daniel B. Kopans, Marylee E. Brown • Stephen M. Pizer "Image Processing Algorithms for Digital Mammography: A Pictorial Essay" journal of Radio Graphics Volume 20,Number 5,sept.2000

[14] Pisano ED, Gatsonis C, Hendrick E et al. "Diagnostic performance of digital versus film mammography for breast-cancer screening". NEngl J Med 2005; 353(17):1773-83.

[15] Wanga X, Wong BS, Guan TC. 'Image enhancement for radiography inspection". International Conference on Experimental Mechanics. 2004: 462-8.

[16]. D.Brazokovic and M.Nescovic, "Mammogram screening using multisolution based image segmentation", International journal of pattern recognition and Artificial Intelligence, 7(6): pp.1437-1460, 1993

[17]. Dougherty J, Kohavi R, Sahami M. "Supervised and unsupervised discretization of continuous features". In: Proceedings of the 12th international conference on machine learning.San Francisco:Morgan Kaufmann; pp 194–202, 1995.

[18]. Yvan Saeys, Thomas Abeel, Yves Van de Peer "Towards robust feature selection techniques", www.bioinformatics.psb.ugent

[19] Gianluca Bontempi, Benjamin Haibe-Kains "Feature selection methods for mining bioinformatics data", http://www.ulb.ac.be/di/mlg

[20] Li Liu, Jian Wang and Kai He "Breast density classification using histogram moments of multiple resolution mammograms" Biomedical Engineering and Informatics (BMEI), 3rd International Conference, IEEE explore pp.146–149, DOI: November 2010, 10.1109/ BMEI.2010 .5639662,

[21] Li Ke,Nannan Mu,Yan Kang Mass computer-aided diagnosis method in mammogram based on texture features, Biomedical Engineering and Informatics (BMEI), 3rd International Conference, IEEE Explore, pp.146 – 149, November 2010, DOI: 10.1109/ BMEI.2010.5639662,

[22] Azlindawaty Mohd Khuzi, R. Besar and W. M. D. Wan Zaki "Texture Features Selection for Masses Detection In Digital Mammogram" 4th Kuala Lumpur International Conference on Biomedical Engineering 2008 IFMBE Proceedings, 2008, Volume 21, Part 3, Part 8, 629-632, DOI: 10.1007/978-3-540-69139-6_157

[23] S.Lai,X.Li and W.Bischof "On techniques for detecting circumscribed masses in mammograms", IEEE Trans on Medical Imaging , 8(4): pp. 377-386,1989.

[24] Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E 69, 026113 (Feb 2004)

[25]. Newman, M.E.J.: Analysis of weighted networks. Physical Review E 70056131 (Nov 2004).

[26] Csardi, G., Nepusz, T.: The igraph software package for complex network research. Inter Journal Complex Systems p. 1695 (2006)

[27] Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in ver large networks. Physical Review E 70, 066111 (Dec 2004)

[28] Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of Communities in large networks. Journal of Statistical Mechanics: Theory and Ex-periment 2008(10), P10008 (2008)

[29] Pons, P., Latapy, M.: Computing communities in large networks using random walks. Journal of Graph Algorithms and Applications 10(2), 191–218 (2006)

[30] Rosvall, M., Bergstrom, C.T.: Maps of information flow reveal community structure in complex networks. Proceedings of the National Academy of Sciences of the United States of America 105(4), 1118–1123(2007)

[31] Deepa S. Deshpande “ASSOCIATION RULE MINING BASED ON IMAGE CONTENT” International Journal of Information Technology and Knowledge Management January-June 2011, Volume 4, No. 1, pp. 143-146

[32] Holmes, G., Donkin, A., Witten, I.H.: WEKA: a machine learning workbench. In: Proceedings Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia, pp. 357-361, 1994.