IOSR Journal of Engineering
e-ISSN: 2250-3021, **p**-ISSN: 2278-8719,
Vol. 2, Issue 12 (Dec. 2012), ||V1|| PP 41-48

Image Classification Approach for Breast Cancer Detection Based on a Complex Event Processing

Chetana Srinivas[1],Mangala.C.N[2],Sahana Devi.K.J[3]
[1]*(Asst Prof, Department of CSE, East West Institute of Technology,Bangalore-91)*
[2]*(Asst Prof, Department of CSE, East West Institute of Technology, Bangalore-91)*
[3]*(Asst Prof, Department of CS, East West Institute of Technology, Bangalore-91)*

Abstract:- Breast cancer is considered to be one of the leading causes of deaths among females in United States. All over the global level approximately 10000 women are diagnosed with this disease per year and approximately 3500 of these women are die from this types of cancer. In this paper, we propose a Complex Event Processing (CEP) Engine based on Support Vector Machine. Currently the Effective method for early detection and screening of Breast Cancer is Mammography Techniques. The detection of Tumor method follows the scheme of a) Mammogram image preprocessing b) The Segmentation Tumor area c) The Extraction of features and the use of support Vector Machine classification method.The monogram enhancement and segmentation techniques play an important role to improve the detection and diagnosis of breast cancer. The results reveal that the application of Event processing techniques improves the classification of images in medical domain and produces accurate results in order to help radiologist assessment.

Keywords: Mammography,CEP,EventProcessing,SVM,KernelFunction, separating hyper plane

I. INTRODUCTION

Breast cancer is the most common non skin malignancy in women and the second leading cause of female cancer mortality. Breast tumors and masses usually appear in the form of dense regions in mammograms. A typical benign mass has a round, smooth and well circumscribed boundary; on the other hand, a malignant tumor usually has a speculated, rough, and blurry boundary.Computer aided detection (CAD) systems in screening mammography serve as a second opinion for radiologists by identifying regions with high suspicious of malignancy [4]. The ultimate goal of CAD is to indicate such locations with great accuracy and reliability. Thus far, most studies support the fact that CAD technology has a positive impact on early breast cancer detection.Event processing technologies are being used successfully to extract information, over disparate event sources in different domains, such as supply chain management, decision making in financial services, health care domain, traffic management systems, short-lived opportunities in the stock market and intrusion detection networks, most of the possibilities still unexplored. For this purpose CEP provides attractive functionality such as data aggregation, timely situation detection and pattern matching. Basically the CEP based engine submits the events to the system by constructing relevant objects, based on the information received, then submits them to the query processor, next based on a predefined set of rules in the shape of queries, the engine filter, and correlates and enrich these elements. New events are created in response to situations on the information, these elements can be enhanced within the query context (search scope of a given pattern) providing a deeper understanding of the data. A CEP engine stores queries and let the information run through them, acting as a filter of information, thus this engine's purpose is not to store information, but to capture meaningful objects and transform them into more suitable representations of the system for a given time, the concept is presented in Fig.1.

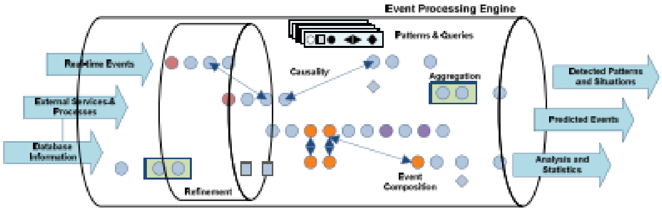

Fig. 1. CEP Conceptual Diagram

Mammography is the most common screening technique used for early detection of breast cancer. Typically a Radiologist evaluates the X-ray screening mammogram image of the breast to determine the presence or absence of tumors. After the evaluation, there are three possible outcomes for the mammogram: A normal one indicates no presence of high density bodies within the breast, a "benign" mammogram suggests the presence of a non-cancerous tumor and finally a malign one adverts the presence of a cancerous tumor on the patient's breast. In this paper, we propose a new classification approach for breast cancer using statistical features and an event processing engine. SVM is a learning machine used as a tool for data classification, function approximation, etc, due to its generalization ability and has found success in many applications [7-11]. Feature of SVM is that it minimizes and upper bound of generalization error through maximizing the margin between separating hyper plane and dataset. SVM has an extra advantage of automatic model selection in the sense that both the optimal number and locations of the basis functions are automatically obtained during training. The performance of SVM largely depends on the kernel [12], [13].Our results display considerable classification rate accuracy and an inexpensive and easy implementation, which can be integrated with other systems.

In the second part of this paper we will discuss related work about mammography, SVM classification and event processing. In section III the proposed method is described, the overall stages are illustrated; in section IV we present the results of the classification using our system. In the last part we provide our conclusions and a brief idea of our future work is expressed.

II. RELATED WORK

Breast cancer is one of the leading causes of mortality on women; despite it can be treated efficiently if detected in an early stage. Several works have been done to develop computer aided breast cancer detection and diagnosis tools. The detection process is mostly done based on the Radiologist judgment. Therefore, it is subject of many factors as experience, exhaustion, and instruction. In many cases the prediction may reveal an undetected tumor or the detection of a nonexistent threat, the latter ends up in a biopsy, which is expensive and highly unpleasant for the patient.

Computer aided detection (CAD) systems in screening mammography serve as a second opinion for radiologists by identifying regions with high suspicious of malignancy. The ultimate goal of CAD is to indicate such locations with great accuracy and reliability. A CAD system is proposed to detect and classify masses on ultrasound breast images, using fuzzy support vector machines. A comprehensive comparison for mammogram analysis is done in judging three different methods i.e. pixel-based, region-based and physics based analysis of densities; density quantifies well the risk of a cancerous tumor. Their results point out the region based method as more consistent and accurate. Ferreira and Borges address two problems on mammogram analysis: sample classification based on the tumor geometry, and classification between normal, benign and malign areas. Their work proposed a supervised classifier and a wavelet transformation to join the set of features. This work provides a novel approach and conceives pattern recognition to achieve the classification. Presents a two-layered model to detect a micro calcification (µCA++) on X-mammograms which are one of the primary indicators of breast cancer. Firstly the location and shape of the micro calcification is calculated. Finally, a set of features is extracted and classified using two types of classifiers: general regression neural network and support vector machines. Any kind of classification needs a set of features and classification system.SVM learning is based on the principle of structural risk minimization [13]. Instead of directly minimizing learning error, it aims to minimize the bound on the generalization error. As a result, an SVM is able to perform well when applied to data outside the training set. In recent years SVM learning has been applied to a wide range of realworld applications where it has been found to offer superior performance to that of competing methods [14].Pattern recognition via imprecisely formulated knowledge representation is presented in a rule-based recognition algorithm is used to weight the rule set and improve the accuracy of the measurement. The effectiveness of the concept was proven using computer generated data. Zaïane, Antonie and Coman [22] propose a mammography classification method based on association rule mining, three phases compose their system: pre processing, mining and organization, their result were tested on a set of 322 images, and reported accuracy of 80.33% with false negatives and false positives tending to zero in the majority of the splits.

Event Processing techniques are being explored intensively on diverse scenarios with very different operational requirements. Diverse constraints are present: source of events, types, patterns to detect, complexity, throughput, latency, reliability. In a health care system outlined using CEP, the system proposes a chain of communication between patients and specialists, efficient enough to filter and correlate meaningful information and responsive to deliver timely updates to the relevant parties. A traffic management network is depicted, defining a structural event model and illustrating its use in a real scenario for the retention problem pattern in a highway, the simplicity of the rules in the shape of queries and event composition of the event processing engine, provided to be effective in this work. RFID technologies and EPC Global standard is used to track and manage supplies safely, a complex event engine is proposed to manage the volume of information inherent to this case; filtering, grouping and aggregation of information were extremely valuable in this scenario and provided significant information about the status of the products. In traditional SVM, every data contributes equally to the classification task. Though the isolated data contributes nothing to classification, it influences the results of the classifier. To avoid this problem, fuzzy C-means clustering algorithm is coupled with SVM to evolve the Fuzzy SVM [4].Fuzzy SVM is capable of increasing the contribution of effective data and suppressing the role of isolated data in classification. Thus fuzzy C- means clustering improves the generalization ability of SVM. Then the adaptive optimization algorithm is employed to compute the SVM parameters. This algorithm, not only increases the flexibility of parameter selection for SVM, but also speeds up the convergence.

III. PROPOSED SYSTEM

Detection of tumors in mammogram is divided into three main stages. The first step involves pre-processing and an image enhancement procedure where pre-processing stage considers the removal of noise and image enhancement techniques are used to improve an image. After the mammogram enhancement, segment the tumor area, based on threshold value. Then finally it involves three arguments. First, we introduce the features to be extracted from the mammographic image. Secondly, a rule based classification method will be discussed and these rules will be expressed as a query to a SVM for classification. Finally, an event processing engine will enhance the features used in the classification process.

3.1 Preprocessing

Noise Removal We have used Gaussian filter. In this approach, there are two major steps:

1. Detect noisy pixel using new impulse detector
2. Utilize weighted directional information to calculate the Gaussian value for removing impulse noise and preserve details

Enhancement of mammogram In this step, contrast limited adaptive histogram equalization (CLAHE) technique has been applied. In CLAHE, the pixel's intensity is transformed to a value within the display range proportional to the pixel intensity's rank in the local intensity histogram. The enhancement is condensed in flat areas of the image, which prevent over enhancement of noise. It also reduces the edge shadowing effect. The CLAHE operates on small regions in the image called tiles rather than the entire image. Each tiles 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 neighbouring tiles are then combined using bilinear interpolation to eliminate artificially induced boundaries. First of all, input image is divided into equal size of number of non-overlapping regions. Then, the histogram of each region has been calculated. Clip limit has to be set f or clipping histograms. In our case, we have set t = 0.002. Each histogram has been redistributed in such a way that its height does not exceed the clip limit. All histograms were modified by the transformation function of normal histogram. Then, using bilinear interpolation, neighbouring tiles has been combined. At the end, image gray scale values have been altered according to the modified histograms.

3.2 Segmentation

The enhanced mammogram images are converted to binary images through thresholding at different values. The segmented images are filtered again with Gaussian smoothing filter to eliminate noise. The thresholding is an important step to improve the detection of breast cancer segmentation subdivides an image into its constituent regions.

operates on small regions in the image, called tiles, rather than the entire image. For each region computes a threshold (level) that can be used to convert an intensity image to a binary image. Level is a normalized intensity value that lies in the range [0, 1]. For each region uses Otsu's method, this chooses the threshold to minimize the intra class variance of the black and white pixels. Multidimensional arrays are converted automatically to 2-D arrays using reshape. The Adaptive Thresholding function ignores any nonzero imaginary part of Image. Adaptive Thresholding returns the effectiveness metric, EM, as the second output argument. The effectiveness metric is a value in the range [0 1] that indicates the effectiveness of the thresholding of the input image. The lower bound is attainable only by images having a single gray level, and the upper bound is attainable only by two-valued images. After finding threshold value converts the greyscale image to a binary image. The output image replaces all pixels in the input image with luminance greater than level with the value 1 (white) and replaces all other pixels with the value 0 (black). Level specify in the range [0, 1], regardless of the class of the input image. The Adaptive Thresholding function can be used to compute the level argument automatically. Use these level segments the Biopsy Image. The original image is divided into an array of overlapping sub-images. A gray-level distribution histogram is produced for each sub-image, and the optimal threshold for that sub-image is calculated based on this histogram. Since the sub-images overlap, it is then possible to produce a threshold for each individual pixel by interpolating the thresholds of the sub-images. An alternative approach is to statistically examine the intensity values of the local neighborhood of each pixel. The first problem facing us when choosing this method is the choice of statistic by which the measurement is made. The appropriate statistic may vary from one image to another, and is largely dependent on the nature of the image.

3.3 Feature Extraction

The images must be subject of a feature extraction process in order to classify them. The chosen parameters are five statistical measurements, commonly used in the study of mammograms; these make a good estimation of the intensity histogram of the image and the Texture Features Area, Centriod, Perimeter etc used to identify the objects and regions of interest in an image. The structure of the Histogram features can be represented by:

$(\mu, \sigma, \sigma^2, \gamma, \beta)$

$$Mean = \frac{\sum(x \quad x)^n}{N} \tag{4}$$

$$SD = \sqrt{\mu_2^1 - \mu^2} \tag{5}$$

$$Variance = (X - \mu^2) \tag{6}$$

$$Skewness = \frac{1}{N} * \left(\frac{(x - x)}{\sigma}\right)^3 \tag{7}$$

$$Kurtosis - \frac{1}{N} * \left(\frac{(x - x)}{\sigma}\right)^4 - 3 \tag{8}$$

3.4 Rule Based Classification using SVM

The main requirement of a rule-based classifier is a set of rules that represent the connection between the features and the classification groups. Each classification is done via implications where the precedent ends up in an estimated effect. The precedent then must be evaluated as a condition to test the validity of the statement. The effect (rule consequent) is a label on the classification problem. The rules can be simple or composite. The rules are the representation of the expert knowledge. Consequently, are subject of precision variations, the rules must be as precise as possible.

```
Algorithm 1: Rule Based Feature Classification
RuleSet: {r1, r2, r3, .......rn}
FeatureSet: {μ, σ, σ², γ, β}
for each rule r in RuleSet do
   if Pre(r) ⊆ FeatureSet then
      Label ← Pos(r)
   end if
end for
```

Given a set of rules and a set of feature values, for each rule in the set of rules each precondition must be matched against the feature set values, if there is a match, i.e. the values of the feature set are present in the rule, the occurrence must be marked as dictated by the rule set post condition. The output then must be a label, which classifies the feature values based on the precedent of the rule.

3.4.1 SVM Classifier

The basic idea of an SVM classifier is illustrated in Fig. 2. This figure shows the simplest case in which the data vectors (marked by 'X's and 'O's) can be separated by a hyper plane. In such a case there may exist many separating hyper planes. Among them, the SVM classifier seeks the separating hyper plane that produces the

largest separation margin [13,14]. The different types of SVM Classifiers are linear, Polynomial, Radial Basis Function, Multi-Layer Perception etc. Linear classification. When used for classification, the SVM algorithm creates a hyper plane that separates the data into two classes with the maximum-margin. Given training examples labeled either "yes" or "no", a maximum-margin hyper plane is identified which splits the "yes" from the "no and the closest examples (the margin) is maximized. Non-linear classification with the "kernel trick" The original optimal hyperplane algorithm proposed by Vladimir Vapnik in 1963 was a linear classifier. However, in 1992, Bernhard Boser, Isabelle Guyon and Vapnik [13,14] suggested a way to create nonlinear classifiers by applying the kernel trick (originally proposed by Aizerman) to maximum margin hyper planes. The resulting algorithm is formally similar, except that every dot product is replaced by a non-linear kernel function. This allows the algorithm to fit the maximum-margin hyper plane in the transformed feature space. In the more general case in which the data points are not linearly separable in the input space, a nonlinear transformation is used to map the data vector **x** into a high-dimensional space
(called *feature space*) prior to applying the linear maximum-margin classifier.

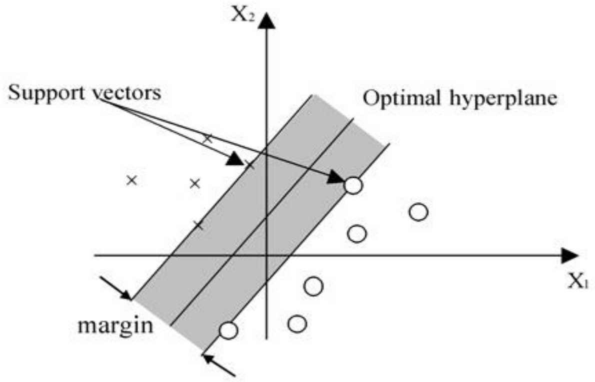

Fig. 2. Support vector machine classification with a linear hyper plane that maximizes the separating margin between the two classes.

3.4.2 SVM Kernel Functions

The kernel function plays the central role of implicitly mapping the input vector into a high-dimensional feature space, in which better separability is achieved. In this study the following five types of kernel functions are considered:

1. Linear Kernel
   $Ø = x_i * x_j$
2. Polynomial Kernel
   $K(\mathbf{x},\mathbf{y}) = (\mathbf{x}^T \mathbf{y} \cdot 1)\, p$, where $p > 0$ is a constant .
3. Radial Basis Kernel (RBK)
   Exp (-γ|xi-xj|2)
4. Quadratic Kernel
   Tanh (γxixj+coeff)
5. Multi-Layer Perceptional Kernel
   $K(\mathbf{x},\mathbf{y}) = (\mathbf{x}^T \tilde{\mathbf{y}}\ 1)\, p$, where $p > 0$ is a constant .

3.4.3 Training examples

Training examples are gathered from the mammograms as follows: for the “Cancer present” class (designated Class 1), image windows of size *M .M* are collected at the centers of tumor of the Cancers identified in the database; for the “Non-Cancer absent” class (designated Class 2), image windows are collected from those regions of the image containing no tumor. Because there are typically far more background regions than regions containing Ms, a random sampling scheme is adopted for Class-2 examples so that the training examples are representative of all the mammograms.

3.4.4 SVM Training

During the training phase, the following variables need to be determined for the SVM classifier: the kernel function to use, and the regularization parameter *C*. For this purpose, we adopt a widely used statistical method called *m*-fold cross-validation, which consists of the following steps: 1) divide randomly all the available training examples into *m* equal-sized subsets; 2) use all but one subset to train the SVM; 3) use the held out subset to measure classification error; 4) repeat Steps 2 and 3 for each subset, 5) average the results to get an estimate of the generalization error of the SVM classifier. The SVM was tested using this procedure for various parameter settings. In the end, the model with the smallest generalization error was adopted. The kernel is then modified in data dependent way by using the obtained support vectors. The modified kernel is used to get the final SVM classifiers.

3.5 event-based architecture

In the search of new possibilities for event processing technologies, we explore the medical diagnosis domain. Mammogram image classification can be achieved using event processing technologies. The proposed scenario includes the Radiologist as the end user of the system, on which based on features extracted from the mammogram images, the system will classify whether cancer is present or not. This classification will be done based on the defined set of rules and enhancing the information available during the execution of the engine, synthesizing an accurate evaluation of the patient's condition. Fig.3. shows the overall architecture of the proposed system. In order to classify information a set of features is needed as evaluation criteria, in our system the classification groups are two: Cancer or Non-Cancer. The image features are being statistically enriched via aggregation and correlation. Our system doesn't require training data. Whatsoever, the classification is done from the start based in the rules defined; the novel proposal is the constantly growing pool of knowledge provided when all the information that passed through the engine can be used to further improve the classification accuracy. After the classification, the information will be presented to the radiologist to assist the evaluation of the patient, as well as a set of enhanced values which will refine the model; simultaneously this enhanced information might be stored in the feature repository for further analysis.

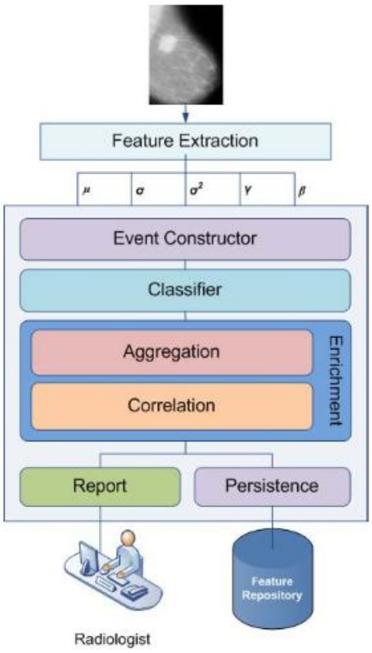

Fig. 2. Engine Conceptual Architecture

3.5.1 Event Constructor

The features extracted previously are sent to the processing engine, that shape them as domain knowledge, some part of the information is cleaned in this stage using weight factors to determine if the values are suitable for event construction. Raw data is sent to the engine, and then normalized; each object is composed of an identifier of the image source for each feature, as well as a time stamp to correlate information in subsequent steps. We implement an abstract object Mammogram which extends these attributes to other objects in the event model. Each feature extracted from the mammogram is an event in the engine context; in addition the diagnosis event needs to be correlated to the mammogram identifier.

3.5.2 Classifier

The domain features are shaped in the construction phase and then are passed into the classifying block of the engine for rule-based analysis. The set of features is linked via identifier, and then using the rules defined and the thresholds for each feature a particular image status is predicted. The classification is done. The rules are evaluated as queries, passed to different SVM Classifiers and the Testing and training datasets are compared to give better accuracy.

3.5.3Aggregation and correlation

The set of features can be enriched via aggregation, it means a defined group of statistic information can be obtained using the actual, previous and incoming information, the context for this aggregation can be done using windows, like a batch of images, or time based windows, e.g. last month features. After a set of features is aggregated into the high level statistics, it is correlated to detect occurrences on the information. The process of aggregate and correlate provides enrichment to the system, making it valuable and providing benefit to the Radiologist.

3.5.4 Report and Persistance

The information in the engine corresponds to three domain categories, feature events, diagnosis events and report events; the first two represent the actual prediction of the system about the presence of cancer in the image. The last one represents the enriched information based on the statistic values extracted from the events passed to the engine. Once the report is presented to the radiologist, the features extracted can be stored in a conventional repository, nonetheless, part of the information as defined by the query windows and the statistical values remain in the engine to provide enhanced accuracy in the further readings.

IV. RESULTS AND DISCUSSION

We present experimental results to show the effectiveness of the proposed method, 300 mammogram images were used for the test; the images were taken from the Digital Database for Screening Mammography (DDSM). The cases were picked randomly between all the criteria available, samples in this database are classified as normal, cancer, benign. Our classification groups are only two: Cancer or Non-cancer.

Since our system is based on statistical information, it is necessary to express the respective statistical error, for the classification purposes we discriminate errors of Type I and Type II:

Type I error: or false positive, when an image is marked as Cancer, given that is not. In any case the image might have a benign tumor, but this is left out of our scenario. This error is associated with the specificity of the classifier as the percentage of images correctly identified as Non-Cancer.

Type II error: also known as false negative. A set of features corresponding to an image is marked as Non-Cancer when in reality there is presence of Cancer. Error II is associated with the sensitivity of the classifier, which is the percentage of images classified as Cancer appropriately.
In the classification problem, we need to compute the trade-off between true positive and false positive. ROC (Receiver Operating Characteristics) is being used successfully for binary classification evaluation, specifically in medical diagnostic; it can be observed that our work performs well for the best possible predication using SVM. In figure (5) we present the relation between the predicted values and the actual labels of each set of features, true-positives, which is the number of correctly predicted cancers and false-positive rate as the number of detected normal mammograms. The plotted lines in the upper segment indicate the Cancer and Non-cancer values, these are very close to the maximum value of the y-axis (True-positive rate), this indicates that the classified features were placed in the proper group and is likely that the rules are suitable for the upcoming features.

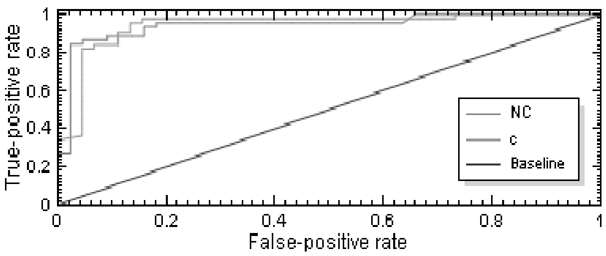

Figure.4. ROC Curve for Mammogram Data

V. CONCLUSION

A novel approach for mammogram classification, using an event processing engine using SVM was presented. We proposed a set of rules, presented as queries to evaluate different statistical features in a mammography; these features were classified, aggregated and correlated, to provide a better tool for the radiologist work.We demonstrated the potential if CEP technologies in the medical field, as a tool for increasing the value of the incoming information. The experimental results showed a good accuracy in the detection of cancer for a given feature set; also the rate of wrong predictions was satisfactory. In our future work, more features and variables will be considered; a more SVM classifiers and different filters will be included as well for better diagnosis of other types of disease.

ACKNOWLEGEMENTS

We express our sincere thanks to our Principal Dr.K.channakeshavlu for his constant guidance in paper publishing.Also we thank our guide and coordinator Dr.chandreshekar & Mrs.Bharathi.M for their motivation and support throughout the session.

REFERENCES

1). Early detection of breast cancer Using svm classifier technique Y.Ireaneus Anna Rejani , Dr.S.Thamarai Selvi
2). Breast Cancer Image Classification Based on a Complex Event ProcessingEngine, Cesar D. Velandia-Briflez, S. Don, Na-Yun-Cho, Eunmi Choi, Dugki Min.
3). Preprocessing and pectoral muscle separation from breast mammogram Muhammad Talha, and Ghazali Bin Sulong.
4). Biao Qin, Yuni Xia, Sunil Prabhakar, Yicheng Tu, "A Rule-Based Classification Algorithm for Uncertain Data," Data Engineering, International Conference on, pp. 1633-1640, 2009 IEEE International Conference on Data Engineering,2009.
5). Chang RF, Chang-Chien KC, Takada E, Suri JS, Moon WK, Wu JH, Cho N, Wang YF, Chen DR, "Three comparative approaches for breast density estimation in digital and screen film mammograms," Annual International Conference of the IEEE Engineering in Medicine an Biology Society, 2006.
6). Breast Cancer detection and Classification using Neural Network Shekhar Singh, Dr P. R. Gupta
7). J.C Fu, S.K Lee, STC Wong, JY Yeh, A.H Wang, HK Wu, Image Segmentation feature selection and pattern classification for mammographic microclassifications", Computerized Medical Imaging and graphics, 2005, pp419-429.
8). Yonit Magid, Asaf Adi, Maya Barnea, David Botzer, Ella Rabinovich, "Application Generation Framework for Real-Time Complex Event Processing," Computer Software and Applications Conference, Annual International, pp. 1162-1167, 2008 32nd Annual IEEE International Computer Software and Applications Conference, 2008.
9). Weixin Wang, Jongwoo Sung, Daeyoung Kim, "Complex Event Processing in EPC Sensor Network Middleware for Both RFID and WSN," Object-Oriented Real-Time Distributed Computing, IEEE International Symposium on, pp. 165-169, 2008 11th IEEE Symposium on Object Oriented Real-Time Distributed Computing (ISORC), 2008.
10). Support Vector Machine Learning for Detection of Microcalcifications in Mammograms, Issam El-Naqa, Yongyi Yang, Miles N. Wernick, Nikolas P. Galatsanos, and Robert Nishikawa* Dept. of Electrical and Computer Engineering, Illinois Institute of Technology 3301 S. Dearborn Street, Chicago, IL 60616 *Department of Radiology, University of Chicago 5841 South Maryland Avenue, Chicago, IL 60637.
11). Detection of Masses in Digital Mammograms using K-means and support Vector Machine Leonardo de Oliveira Martins, Geraldo Braz Junior+, and Marcelo Gattass, Federal University of Maranh˜ao, Av. dos Portugueses, SN, Campus do Bacanga, 65085-580, Received 23th May 2008; accepted 16th July 2009.