ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2017

Breast Cancer Detection Technique and a Way to Data Provenance

Ritu T[1*], Ankit Kumar M[2]

G.D. Goenka University, India[1]
International Atomic Energy Agency, Vienna, Austria[2]

Abstract: Breast cancer which is one of the most common diseases is mostly detected at 3rd or 4th stage when the mortality rate becomes high. Hence in order to detect this at an earlier stage the concern person needs to go through mammography (cost of this procedure is lower than biopsy) screening at regular intervals. In case a radiologist is unable to decide he may take assistance from a computer aided system. A CAD system with GLCM and RBFNN with accuracy of 97.3% is designed with improved confidence (efficiency and reliability) by the use of data provenance. In this design all the relevant medical data (demographic, medical history) is stored and processed for making a better and an informed decision.

Keywords: CAD System; Data provenance; GLCM; RBFNN; Cancer detection; MIAS; Artificial intelligence; Neural network; Breast cancer.

I. INTRODUCTION

In many countries, breast cancer is the most common form of cancer. This can be analyzed by using mammography, magnetic resonance imaging, thermography and ultrasound images [1]. It is a disease with a very high detection and treatment cost (e.g. biopsy, which is very risky and costly). But with mammography breast cancer detection becomes highly accurate and low cost. Because biopsy is interfering procedure that makes patient discomfort and sometimes causes death [2]. So, mammography is now considered as standard procedure for breast cancer detection [3]. Many artificial intelligence techniques are most widely in use for classification problems in medical diagnosis. Artificial neural networks and fuzzy logic are used for classification with image feature extraction in mammogram. Feature extraction is done using image processing. Various techniques are used for extracting features from a mammogram but for achieving more accuracy we used texture feature extraction that is GLCM with feed forward neural networks RBFNN classifier. GLCM has been proved as a best method for textural features extraction from various medical images [4]. Radial Basis Function Neural Network (RBFNN) is a type of feed forward neural networks in Artificial Intelligence (AI) system which is used for tissue classification [5]. This process is known as CAD System that is 'Computer Aided Detection System'. For a radiologist CAD system works as second pair of eyes whose firmness is very good. According to national laws, in medical domain demographic and clinical information of patients plays a key role. So, here comes the term Data Provenance, it is a term that is the actual description of the history of a data set. While taking mammography film it is important to store machine information, name of the technical person who is taking mammogram and also patient information. For making best and accurate decision machine configuration must be recorded. If a mammogram is pre-processed then at what level and which operations has been applied. So, provenance commonly means the history, ownership and management of data and its processing in the areas of interest. The possibility of storing provenance information is equally important as the results of the scientific analysis itself [6]. All of this information, normally generated through execution of scientific workflows enable the traceability of the origins of data (and processes); can identify event causality; enable broader forms of sharing, reuse, and long-term security of scientific data; can be used to quality control and resolve the quality of particular data set [7]. One might wish to recognize the path that mammograms took before reaching its current location. This information is known as the provenance information which will help to authenticate a piece of mammogram [8]. Here the mammogram indicates the term data because we are working with mammogram images. Knowledge about the origin and history of a mammogram/image is necessary for beginning data and results quality. The precise information about how a mammogram/image was processed, including the specific algorithms, software routines that were used, is significant for proper analysis, high integrity reproduction and re-use. So, while doing pre-processing and feature extraction it is important to store relevant information as data provenance. We have prepared a mechanism for specifying provenance in a simple and easy way for use in medical images for maintaining their source and relevant information about patient, disease, machine information in which mammogram is taken etc. This useful sequence of highly detailed and easy utility of metadata should largely simplify the collection of provenance and following distribution of data. In medical images, data provenance needs to be stored in a database to give relevant information about the validity, accuracy and utility of the data, so data provenance was stored in a database to give information about the authority, accuracy and timeliness of the data, which authorize data purity framework that is very important in the process of image analysis and disease detection.

II. RELATED WORK

In this literature, different techniques are explained to detect, classify and examine the presence of breast cancer in digital mammograms. Many researchers are working on textural feature analysis on mammogram images. Benign and Malignant classification using GLCM features derived by scientist Harlick and the achieved sensitivity and specificity are more than 90%. Indra Kanta Maitra et al. [9] used GLCM features to identify the abnormal masses and their study experimented on mammograms from the MIAS database. A mammogram breast cancer classification on MIAS (Mammogram Image Analysis Society) database uses intensity histogram features. They experimented on 350mammograms for classification. A proposed mammogram segmentation using texture features given by HB Kekre et al. [10]. Their study involved mammograms from the MIAS database. A comparison of texture and shape features for micro calcification classification is presented by Hamid Soltanian Zadeh et al. [11]. GLCM texture features extraction to identify masses in mammogram on 100 mammograms from the MIAS database done by A Mohd Khuzi et al. [2]. Some researchers used neural networks to classify like U. Rajendra Acharyaet al. [12] experimented on MIAS and then BN Prathibha et al. [13] used texture feature extraction with kernel discriminant analysis for mammogram also on MIAS database. “Position Paper: Provenance Data Visualization for Neuroimaging Analysis” Bilal Arshad, Kamran Munir, Richard McClathchey and Saad Liaquat. They presented framework on provenance data visualization for neuroimaging analysis on data coming from various different sources. They present a framework to visualize provenance information in order to benefit the flow of verification of analyses, scientific outputs, progression and evolution of results for neuroimaging analysis.

III. METHODOLOGY

CAD system follow a three step basic process, these are Pre-processing, Feature Extraction and then Classification of Benign and Malignant. We also followed the same procedure in the previous work with GLCM feature extraction technique and RBFNN classification technique [14]. In proposed work we would like to talk about the data provenance term to store the information of mammogram like machine name, mammogram specification, patient information and the hospital or diagnostic centre details. This information needs to be stored for the accurate analysis. But the authentication is much required of the medical data provenance because it is against the law and rules to use patient’s information without any permission.

3.1. Data Provenance

While developing CAD System, store data provenance of mammograms which will be used for testing purpose. It will help in making some decisions about the machine used while taking mammograms. Machines of different companies are with different configurations which produce output of different features. Mammograms can be different in contrast level, correlation etc., which causes issues with pre-processing and feature extraction. So, to resolve these issues we need to maintain record of mammograms generation sources. The information of patient’s demographic specially age, address, symptom which helps in making predictions like which age group women most commonly develop breast cancer. For storing data provenance any of the database can be used from excel, my SQL, hbase. It depends on the size of data and the security/authentication required.

3.2. Pre-Processing

In pre-processing step, quality of mammograms is improved by removing noise and removing some areas which are not required at the time of analysis. Breast mammogram can be done in one of two views either MLO or Cranio Caudal (CC) view. We worked on mediolateral oblique view. So, at the second step of pre-processing background area is being removed, removal of pectoral muscle and rib portion from the breast mammogram. At the final stage of pre-processing Region of Interest (ROI) is figured out using segmentation to use the remaining mammogram for feature extraction. This step proved to be demanding success factor in the CAD system and helps in suppressing distortions and neglecting those parts which are not part of breast and this ensures a better classification accuracy of CAD algorithm [15].

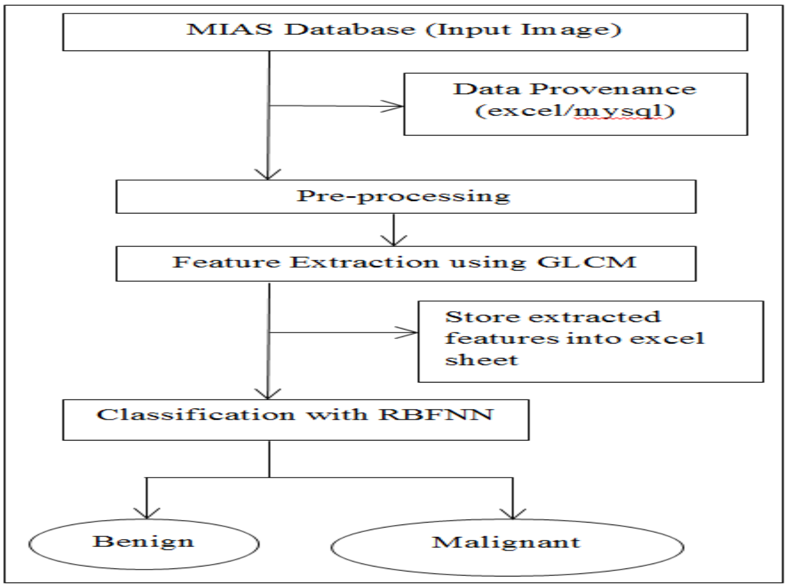
MIAS Database (Input Image)
Data Provenance (excel/mysql)
Pre-processing
Feature Extraction using GLCM
Store extracted features into excel sheet
Classification with RBFNN
Benign
Malignant

Figure 1: An overview of CAD system with data provenance.

3.3. Feature Extraction

Feature Extraction is the ultimate relevant process used for extracting value of different features and shows a crucial role in pattern classification. Visual contents from images can be taken for indexing and retrieval. Image features can be anything either simple features like color, size, shape, contrast etc. But for medical diagnosis mammogram consist of dissimilar information that shows various types of tissues, glandular ducts, blood vessels, breast edges and mammogram machine characteristics. So to have a better diagnostic approach for detecting normal and abnormal tissues, we have to choose such a system which generates accurate result with much better accuracy. So we used GLCM technique for feature extraction. GLCM is Grey-Level Co-occurrence Matrix which is basically a texture analysis application. It is a tabular representation of occurrence of grey levels occurs in an image. It works on the number of pixels or dots in combination. On mammograms we have used feature extraction based on GLCM is the second order statistics which is used for analysing mammograms as a texture. So, based on the number of intensity points (pixels) in every combination, statistics are classified into first-order, second order and higher-order statistics. Harlick has defined 14 features [16], in many CAD systems only 5 or 7 features have been used to extract features. Many researchers worked on GLCM, so while using 5 features system gives accuracy of 96% with 50 test cases and with 7 features 95.50% accuracy. So, for this work we have used 8 features namely Contrast, Correlation, Autocorrelation, Sum Entropy, Variance, Information Measure of Correlation, Inverse Difference Momentum, Difference Entropy to measure the accuracy and other performance factors and after extracting, these features were stored into an excel sheet as data provenance for classification purpose (Figure 1).

3.4. Classification

We used artificial neural networks for the classification of breast cancer. Because of machine vision approach, a problem of detecting masses in digital mammograms usually occurs. This problem is slightly solved with artificial neural networks [16]. Classification process of neural network consists of two processes: Training and Testing. The accuracy of classifier is depends on training process, so training of CAD system must be done with more accurate mammogram data. Because neural network generates output based on its training experience. So, data provenance also needs before giving training to a CAD System. A proper classification method must be used to identify the potential micro calcification pixel based on features extracted using GLCM [17]. We choose RBFNN that is Radial Basis Function Neural Networks [18]. Neural network is based on the three layer architecture: input layer, hidden layer and output layer [19, 20]. Input layer is based on the number of features extracted in feature extraction step, hidden layers depend on input and hardware configuration of the system and then finally out layer which diagnose whether it is benign or malignant.

IV. PERFORMANCE EVALUATION

There are three terms based on which performance of the CAD system is measured and they are Accuracy, Sensitivity and specificity (Tables 1 and 2). Accuracy measures the binary classification. Binary classification stands for classifying true or false, positive and negative. Specificity deals with negative cases and sensitivity deals with positive cases.

Measures	**Formula**
Accuracy	(TP+TN)/(TP+FP+TN+FN)
Sensitivity	TP/(TP+FN)
Specificity	TN/(TN+FP)

Table 1: Measure formula.

Actual	**Predicted**	
	Normal(Negative)	**Cancer(Positive)**
Normal	TN	FN
Cancer	FP	TP

Table 2: Confusion matrix.

Cases which are accurately identified termed as true positive (TP), negative cases which are identified as negative termed as true negative/normal (TN). False positive (FP) is the term use for cases which are actually negative but identified as positive/cancer and false negative (FN) is vice versa (Table 3).

4.1. Confusion Matrix for Tested Database

A second pair of eyes is developed for radiologist to make better decisions about whether the mammogram is normal or cancerous. For doing analysis using GLCM and RBFNN, system is tested on MIAS (Mammogram Image Analysis Society) database of 112 mammograms. In this 56 were normal and 56 were abnormal/ cancerous. The process started with mammogram taken as input and stores its information into database as data provenance, the mammogram which was used for giving training to a system and mammograms which were used for testing the system. Start with pre-processing and then pass ROI to extract features from them. Store these extracted features and the pass into neural network as an input, 8 input layers, 111 m hidden layers based on input and 1 unit in output, which gives information either benign or malignant.

Actual	**Predicted**	
	Normal	***Cancer***
Normal	56(TN)	3 (FN)
Cancer	0(FP)	53 (TP)

Table 3: Confusion matrix for tested database.

Test Cases	Accuracy	Specificity	Sensitivity
112	97.3%	100%	94%

Table 4: Performance analysis.

4.2. Discussion

Table 3 shows the confusion matrix for tested database and Table 4 shows the performance measure. During training system has identified all the benign images correctly but cannot identify all malignant images correctly, because of which accuracy goes down. The classification accuracy obtained by using 8 features was 97.3% whereas sensitivity and specificity were 94% and 100%. The CAD system is developed for the classification of mammogram into normal and cancer pattern with the aim of supporting radiologists in visual diagnosis .This work has investigated a classification of mammogram images using GLCM features. The maximum accuracy rate for normal and cancer classification is 97.3% by using 8 extracted features. For future work, we can use more than one feature extraction technique like, extract feature using GLCM and then pass it to another feature extraction technique. So, we can get new feature. Using proper feature selection method accuracy may be improved efficiently.

V. CONCLUSION

This research work is under progress with data provenance and authorizes a support for the requirement to analyze provenance for mammography analysis. We have referred actual provenance visualization systems, so over the literature survey we have found that they are not satisfactory for provenance analysis in mammograms. So, based on the experience we have drawn a flow of information for the need of data provenance. This will help the radiologist in making better decision. In the process, as relevant information is being stored into database which can further help to radiologist in breast cancer diagnosis. Sometimes mammogram or medical information gets changes in between from source to destination but radiologist is not aware about change. So keeping record of where this mammogram is taken, by whom (technician) it is taken, machine configuration, age of patient, address etc. This work can be done within single flow. If all the information is correct, it will help the radiologist in making better decisions and diagnosis of present mass whether it is cancer/malignant or non-cancerous/benign.

VI. REFERENCES

1. K Ali, K Ayturk, et al. Expert system based on neuro-fuzzy rules for diagnosis breast cancer. Experts Systems with Applications 2011; 38: 5719-5726.
2. KM Azlindawaty, R Besar, et al. Identification of masses in digital mammogram using gray level co-occurrences matrices. Biomedical Imaging and Intervention Journal 2009; 5.
3. DP Atam, C Yateen, et al. Radial-Basis-Function Based Classification of Mammographic Microcalcifications Using Texture Features. Engineering in Medicine and Biology Society IEEE 17th Annual Conference 1995; 1: 535-536.
4. BN Prathibha, V Sadasivam A kernel discriminant analysis in mammogram classification using with texture features in wavelet domain. International Journal on Computational Intelligence 2010; 1.
5. ZS Hamid, RR Farshid, et al. Comparison of multiwavelet, wavelet, Haralick and shape features for microcalcification in mammograms. Pattern Recognition 2004; 37; 1973-1986.
6. S Hadi, P Brajendra, Data Authentication and the Corresponding Provenance Information Management Journal of Digital Information Management 2007; 7: 74-82.
7. SS Holalu, K Arumugam Experimental investigation on breast tissue classification based on statistical feature extraction of mammograms. Computerized Medical Imaging and Graphics 2007; 31: 46-48.
8. HB Kekre, G Saylee, Texture based segmentation using statistical properties for mammographic images. International Journal of Advanced Computer Science and it Applications 2010; 1: 102-107.

9. MK Indra, N Sanjay, et al. Identification of abnormal masses in digital mammography images International Journal of Computer Graphics 2011; 2: 17-30.
10. HB Kekre, G Saylee, Texture based segmentation using statistical properties for mammographic images. International Journal of Advanced Computer Science and it Applications 2010; 1: 102-107.
11. ZS Hamid, RR Farshid, et al. Comparison of multiwavelet, wavelet, Haralick and shape features for microcalcification in mammograms. Pattern Recognition 2004; 37: 1973-1986.
12. BN Prathibha, V Sadasivam A kernel discriminant analysis in mammogram classification using with texture features in wavelet domain. International Journal on Computational Intelligence 2010; 1.
13. YM Norlia, IMA Nor, et al. Computer-Aided Detection and Diagnosis for Microcalcifications in Mammogram: A Review. IJCSNS International Journal of Computer Science and Network Security 2007; 7: 202-208.
14. A Rajendra, EYK Ng, et al. Computer-based identification of breast cancer using digitized mammograms. Journal of Medical Systems 2008; 32: 499-507.
15. T Ritu, K Indu, Efficient Technique For Texture Features Based CAD System for Mammogram Images using GLCM and RBFNN (Breast Cancer Detection). Advances in Computer Science and Information Technology ACSIT 2015; 2: 68-71.
16. HM Harlick, K Shannugam, et al. Textural features for image classification. IEEE, Transaction on Systems Man & Cybernetics 1973; 3: 610-621.
17. M Simon, G Paul, et al. Provenance: The Bridge between experiments and data. Computer in Science & Engineering 2008; 10: 38-46.
18. SS Basha, KS Prasad, Automatic detection of breast cancer mass in mammograms using morphological operators and fuzzy c-means clustering. Journal of Theoretical and Information Technology 2009; 5: 704-709.
19. S Shekar, PR Gupta, Breast cancer detection and classification using neural network. International Journal of Advanced Engineering Sciences and Technologies 6: 4-9.
20. SL Yogesh, P Beth et al. Towards a quality model for effective data selection in collaboratories. Workshop on Workflow and Data Flow for Scientific Applications (SciFlow06) held in conjunction with ICDEW. IEEE 2006: 72.