COMPARISON OF DATA MINING ALGORITHMS FOR MAMMOGRAM CLASSIFICATION

[1]MONIKA HEDAWOO, [2]ABHINANDAN JAISAWAL, [3]NISHITA MEHTA

[1,2,3]Department of Information Technology, SRM University, Chennai, Tamil Nadu
E-mail: [1]monikahedawoo@gmail.com, [2]abhinandan.jaisawal0508@gmail.com, [3]nishitamehta6261@gmail.com

Abstract-- This paper describes a breast cancer classification performance trade-off analysis using two computational intelligence system. The proposed system has been implemented in four stages: (a) Region of interest (ROI) which identifies suspicion regions, (b) feature extraction stage locally processed image (ROI) to compute important features of each breast cancer. (c) Feature selection stage by using forward stepwise linear regression method (FSLR). (d) Classification stage which classifies between cancer and non-cancer case. In the classification stage we are applying two computational intelligence paradigms. K- Nearest Neighbor and Naïve Bayes Algorithm are used for classification of data whether it is cancer or non- cancer.

Keywords— Naïve Bayes, k- Nearest Neighbor, Region of Interest, Feature Extraction.

I. INTRODUCTION

Cancer is the most wide spread disease around the world. Among this breast cancer is the most common and harmful disease among women. Breast cancer is the uncontrolled growth of breast tissues. It begins in the form of "cysts" in breast tissues. Cysts if huge in number or size can lead to breast cancer. Also micro calcification in breasts can show a possibility of cancer in breasts. Cancer tumour can be classified in two ways- benign (not cancerous) or malignant (is potentially cancer). The term "breast cancer" refers to malignant tumours that are developed from cells in breast. Mammography is the most effective, low cost, contemporary option of premature detection and highly sensitive technique for detecting small lesions resulting in at least a 30% reduction in breast cancer deaths. Mammograms are used as screening tool to detect early breast cancer in women experiencing no discharge.

II. PROPOSED METHOD

The experiment involves image or mammogram collection, preprocessing the mammograms, finding region of interest, extracting features and classification of the features using various data mining algorithms. Three features are selected which are categorized under physical features. Classification algorithms are used to segregate the normal cancer cases from the abnormal cancer cases. Here single classification algorithms which are K nearest neighbour and Naïve Bayes are used for classification.

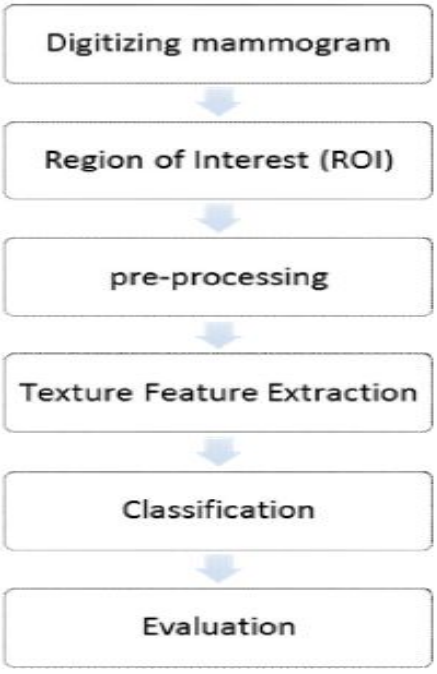

Figure 1: Proposed Method

III. DIGITIZING MAMMOGRAMS

Mammograms are an x-ray picture of breasts. It is used to check and identify cancer cells. It is one the efficient, affordable and quick methods to check for cancer. There are two kinds of mammography; screening mammography and digitized mammography. Screening mammography is the type of mammogram that checks you when you have no symptoms. It can help reduce the number of deaths from breast cancer among women ages 40 to 70. The dataset used in this paper is Digital Database for Screening Mammograms (DDMS). The left CC (cranial-caudal) view is chosen for each case. Two cases have been considered: 1. Normal is a collection of mammogram images that do not have breast cancer and 2. Cancer type is a collection of mammogram images have breast cancer.

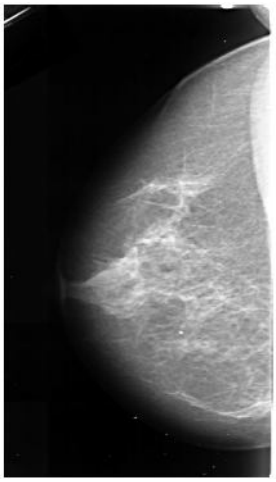

Figure 1: Digitized Mammogram from Left CC View

IV. PRE-PROCESSING

Pre-processing is an analysis and manipulation of digitized image to enhance the image and remove unwanted objects from the image. Pre-processing is required to overcome this problem and make efficient feature extraction of images as possible. A lot of pre-processing is done to bring out the required feature from the image. In this case background removal and contrast enhancement is done in each image.

A. *Background removal:* Sometimes the image contains certain portions which are not requires for processing. These components in the image may give false or incorrect values. Thus background removal is done for removing the unnecessary components in the image.

B. *Contrast Enhancement*: Image enhancement techniques have been widely used in many applications of image processing where the subjective quality of images is important for human interpretation.

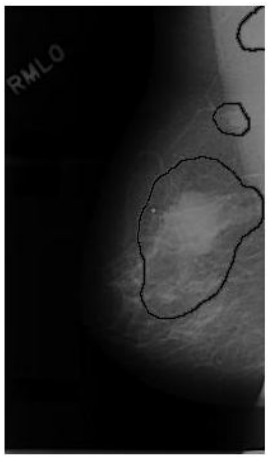

Figure 2: Original Mammogram

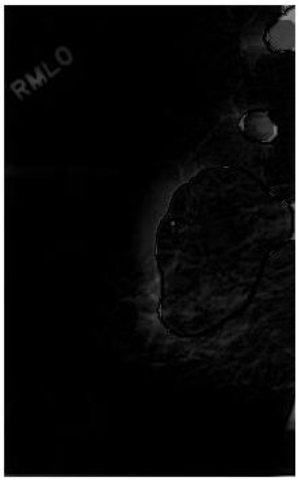

Figure 3: Background removed image

Figure 4: Contrast Enhanced Image

V. REGION OF INTEREST

A region of interest (ROI), is a selected subset of samples within a dataset identified for a particular purpose. The concept of a ROI is commonly used in many application areas. In medical imaging, the

boundaries of a tumor may be defined on an image or in a volume, for the purpose of measuring its size. ROI extracted by entering coordinates X, Y and radius in pixels, according to data provided by the DDSM database for each abnormal mammogram image.

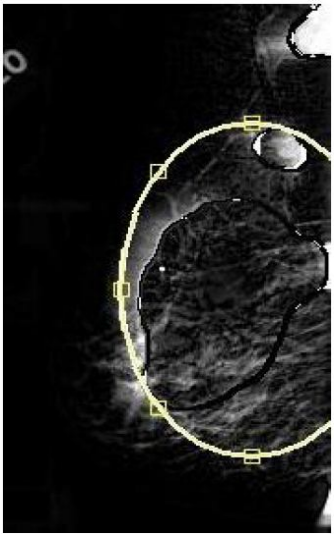

Figure 5: Manually Selected ROI

VI. FEATURE EXTRACTION

The relevant features are extracted from the digitized mammograms. Majorly two kind of features are highlighted here; Physical feature and GLCM feature. In this paper we are analyzing only physical features

Physical features are features common to almost all the images in image processing. They basically give an idea about the basic nature of the image. The physical features also tell about how the image is categorized and easily classified. The physical features are: mean, Entropy, standard deviation and Variance. The formula for each of the features is given below:

Table 1: Physical Features

Features	Formula
Mean	$M = 1/MN \sum^{M}_{i=1} \sum^{N}_{i=1} (p(i, j)-\mu)^2$
Standard Deviation	$\sigma = \sqrt{(1/MN \sum^{M}_{i=1} \sum^{N}_{i=1} (p(i, j)-\mu)^2)}$
Entropy	$F_8 = -\sum_I \sum_j p(I,j) \log(p(I,j))$

Here the values for mean, entropy and standard deviation are mentioned for cancer and non-cancer cells.

Table 2: Features values for cancer and non-cancer data

	Non- Cancer	Cancer
Mean	0.421387	0.291627
Entropy	0.982094271	0.870813049
Standard Deviation	0.493785515	0.454514911

VII. CLASSIFICATION ALGORITHMS

There are various algorithms for automated classification. In this paper used several classification algorithms to compare their performance: k-Nearest Neighbour (kNN) and Naive Bayes (NB)

A. *K-Nearest Neighbour (kNN)* is used to classify the unknown value into the known classifiers by finding the nearest available character in the graph. If k=3, it will check the three nearest values present around the unknown value. If number of class1 elements is more than the number of class2 elements, then the unknown entity belongs to class1. kNN is best used in multi-case models as its decision is based on the number of surrounding classifiers.

B. *The Naive Bayesian (NB)* is based on the Bayesian theorem .The Naïve Bayesian Classifier assumes that features are independent. This method is important for several reasons. It is very easy to construct, does not need any complicated iterative parameter estimation schemes. This means it may be readily applied to huge data sets. This classification technique analyses the relationship between each attribute and the class for each instance to derive a conditional probability for the relationships between the attribute values and the class.

VIII. RESULT

Confusion matrix is a matrix which represents a relation between actual values and predicted values.

	Predicted: NO	Predicted: YES
Actual: Yes	TN	FP
Actual: No	FN	TP

Figure 6: Confusion Matrix

TN= True Negative; values which are correct but are not predicted
TP= True positive; Values which are correct and predicted
FP= False Positive; values which are incorrect but are predicted
FN= False Negative; values which are incorrect and not predicted

A. *Accuracy:* Classification accuracy is the percentage of instances that are correctly classified by the model. It is calculated as the sum of correct classification divided by the total number of samples. It is given by the formula:

$$Accuracy = \left(\frac{(TP+TN)}{(TP+TN+FP+FN)} \right)$$

B. *Recall:* It is the measure of the ability of a classification model to select instances of certain class from the dataset. It is the proportion of actual positive which are predicted positive.

$$\text{Re}call = \frac{TP}{TP+FN} = \frac{C_{ii}}{\sum_{j=1}^{n+1} C_{ij}} = C_{ii} / ATotal_i$$

C. *Specificity:* This is a measure that is commonly used in two class problems where the focus is on a particular class. It is the proportion of the negative class that was predicted negative and it is also known as the true negative rate. Specificity formula:

$$Specificity : \frac{TN}{TN+FP} = \frac{T - ATotal_i - PTotal_i + C_{ii}}{T - ATotal_i}$$

Table 3: Comparison of kNN and NB algorithms

Classifiers	KNN	NB
accuracy in %	78.5714	64.2857
recall/sensitivity/TPR	0.8571	1
Specificity (TNR)	0.4545	0.2222

CONCLUSION

It is observed that the performance of an algorithm depends majorly on the datasets we are working on. For the DDSM dataset which is for breast cancer detection, it is k-nearest Neighbor in this case. It has more accuracy and specificity. But Naïve Bayes is more sensitive.

It is noticed that the values may differ if number of datasets for factors for data. And also it should be kept in mind that all the algorithms work appropriately, given the fact that they are in the optimum conditions and datasets.

The ultimate goal of this paper is to form an algorithm to detect the presence of the tumor cells in breasts and to give an early prediction of breast cancer so that the lives of many woman could be saved.

ACKNOWLEDGEMENT

We would like express their deepest gratitude to their guide, **Dr. A. Shanthini** for her valuable guidance, consistent encouragement, personal caring, timely help and providing an excellent environment for doing research. All through the work, in spite of her busy schedule, she has extended cheerful and cordial support to us for completion of the research work.

REFERENCES

[1] Adegoke, B., Ola, B., & Omotayo, M. (2014). Review of Feature Selection Methods in Medical Image Processing. IOSR Journal of Engineering, 5.

[2] Chadha, A., Mallik, S., & Johar, R. (2012). Comparative Study and Optimization of Feature-Extraction Techniques for Content based Image Retrieval. International Journal of Computer Applications, 8.

[3] Chapelle, O. (1998). Support Vector Machines et. Juin-Aoˆut.

[4] GADKARI, D. (2004). IMAGE QUALITY ANALYSIS USING GLCM. Florida: Thesis report.

[5] Gebejes, A., & Huertas, R. (2013). Texture Characterization based on Gray Level Co-occurance Matrix. Conference of Informatics and Management Sciences, 4.

[6] Jehlol, H. B., Abdalrdha, Z. k., & Oleiwi, A. S. (2015). Classification of Mammography Image Using Machine Learnig Classifiers and Texture Features. International Journal of Innovative Research in Advanced Engineering, 8.

[7] KIM, J., KIM, B.-S., & SAVARESE, S. (2008). Comparing Image Classification Methods: K-Nearest-Neighbor and Support-Vector-Machines

[8] Kumar, G., & Bhatia, P. K. (2014). A Detailed Review of Feature Extraction in Image Processing Systems. Fourth International Conference on Advanced Computing & Communication Technologies, (p. 8). Haryana.

[9] M.A, A., & Kiran, P. (2014). Feature Extraction Values for Digital Mammograms. International Journal of Soft Computing and Engineering, 5.

[10] Mohanaiah, P., Sathyanarayana, P., & GuruKuma, L. (2013). Image Texture Feature Extraction Using G Image Texture Feature Journal of Scientific and Research Publications, 5.

[11] Nandgaonkar, M., Jagtap, M., & Anarase[, M. (2010). Image Mining of Textual Images Using Low-Level Image Features. IEEE, 5.

[12] Srivastava, D. k., & Bhambhu, L. (2009). DATA CLASSIFICATION USING SUPPORT VECTOR MACHINE. Journal of Theoretical and Applied Information Technology, 7.

[13] Zulpe, N., & Pawar, V. (2012). GLCM Textural Features for Brain Tumor Classification. IJCSI International Journal of Computer Science, 6.

★★★