IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 3, March 2020

Logistic Regression based Mass Classification using Feature Extraction

Nimmi Sudarsan[1], Nandakumar Paramparambath[2], Sidharth N[3]

Department of Electronics and Communication, NSS College of Engineering, Palakkad, India[1]

Professor, Department of Electronics and Communication, NSS College of Engineering, Palakkad, India[2]

Asst.Professor, Department of Electronics and Communication, NSS College of Engineering, Palakkad, India[3]

Abstract: Breast cancer holds the second position for cancer deaths in women [12]. There are several Computer Aided Detection and Diagnosis (CAD) systems used today in order to aid radiologists in detecting malignant cancers at the early stage. Such systems along with suitable classifiers yield better prediction of cancerous masses. This paper presents a logistic regression model based mass detection and classification based on selected geometrical features from breast DICOM images with an accuracy of 93%. Previous work of Alima et al, resulted in an accuracy of 91% using ANN[4]. The performance of the feature extraction and classification system is developed using the database collected as a part of the dream challenge[2]. Performance results are given in terms of confusion matrices.

Keywords: Microcalcifications, Compactness, Malignancy, Neoplasy, Craniocaudal, Mediolateral

1. INTRODUCTION

The morbidity and mortality rates of women affected with breast cancer are increasing day by day. The breast tissue comprises of glands used for milk production. These glands are otherwise called lobules. Breast cancer develops from these lobules and then spread to the ducts. The ducts are the connections between the lobules and the nipple. The rapid increase in mortality rate can be reduced as a result of an efficient detection of cancer. Namely, the breast regions classification is between mass and non-mass. There are several techniques for classifying benign and malignant neoplasy from digital images. Digital breast images are efficiently captured using mammography. X-Rays mammography is mainly used to detect asymmetry if any between the left and right breasts due to its Low amplitude [3]. This information is used by radiologist for detecting microcalcifications. In ultrasound technique, reflections from high frequency sound waves incident on breast tissues is used for imaging. Temperature variance in normal region and mass region is considered for breast thermography. It is an adjunct tool for accurate prediction of masses. Apart from these, for high risk women and to evaluate dense breasts, breast MRI of good resolution can be used [5]. Once the image acquisition phase is carried out, they have to be efficiently classified. For image classification, there are several conventional algorithms. Predictive analysis requires strong algorithms like SVM where a linear kernel is opted for a linear (separable) dataset and Gaussian kernel for a non-linear (non-separable) dataset classification. Binomial or multinomial logistic regression using sigmoid function is another method for classification. The two possible outcomes of binomial (binary) logistic regression give a straight forward interpretation of data. Input containing K closest training sets as neighbouring features is the K Nearest Neighbour (KNN) algorithm [3]. It is a non-parametric method for classification. Assumption of independence between each feature is the idea behind a Naïve Bayes classifier. It is named so because it works on the Bayes theorem for classification.

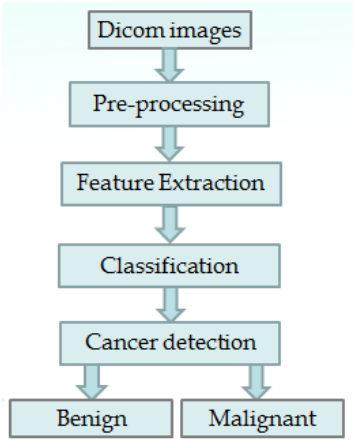

Fig 1.1: Cancer classification system model

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 3, March 2020

In this paper, feature extraction and classification of malignant and benign cancers based on a conventional as well as a recent algorithm is done. Both the algorithms are compared and the extracted features are validated using SVM, KNN and decision tree classifiers using R2015a. Figure 1.1 shows a block diagram describing the overall model for classification.

2. DATASET DESCRIPTION

The dataset for classification consist of 500 mammographic images of size 16.2MB (malignant) and 26MB (benign) of 58 patients taken from over 640k images as a part of the digital mammography dream challenge [2]. The images are in DICOM format (.dcm) comprising of two images per breast, one in the craniocaudal (CC) view and one in the mediolateral oblique (MLO) view of both the left and right breast. The additional information include: Patient's age, BMI (Body Mass Index), Race, Cancerous or non-Cancerous, Invasive or non-Invasive, Image resolution etc.

3. SYSTEM MODEL DESCRIPTION

The cancer detection model consist of mainly three modules: a) Pre-processing module, b) Feature Extraction module and c) Classification module

3.1 Pre-processing

The input to the pre-processing stage is the DICOM (digital imaging and communications in medicine) breast image file. The DICOM standard is useful for integrating all modern imaging equipment's, accessories, networking servers, workstations, printers, and picture archiving and communication systems (PACS) that may have been installed by multiple manufacturers [25]. In this stage, a two dimensional convolution of the input DICOM image over a square matrix of ones is done. Figure 3.1 shows the input image. Then the image contrast is stretched by mapping its intensity values to new values such that there is 1% of data saturation at low and high intensities[26]. Noises in a DICOM image involve Gaussian noise and speckle noise. Noise reduction can be done in two steps. First, a Gaussian white noise of mean = 0 and variance = 0.25 is added to this intensity image because the noise of the original DICOM image does not appear. Figure 3.2 shows the noisy image. Secondly, the noisy image is passed through a wiener filter which is a 2 dimensional adaptive noise filtering technique that low pass filters the intensity image that has been degraded by constant power additive noise. Figure 3.3 shows the filtered image. The intensity image is then converted to binary. Figure 3.4 shows the final intensity image used for classification.

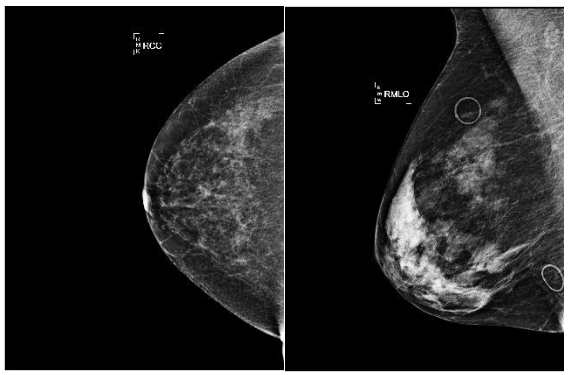

Fig 3.1: Input DICOM image for pre-precessing (a) Malignant (b) Benign.

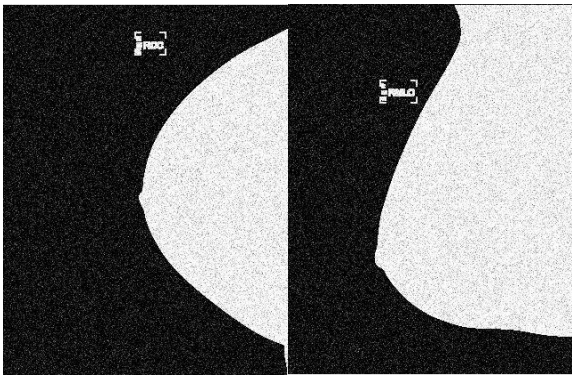

Fig 3.2: Intensity image after adding White Gaussian noise.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 3, March 2020

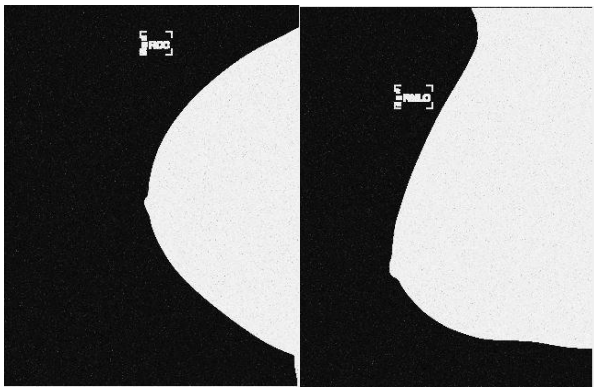

Fig 3.3: Intensity image after low pass filtering using wiener filter.

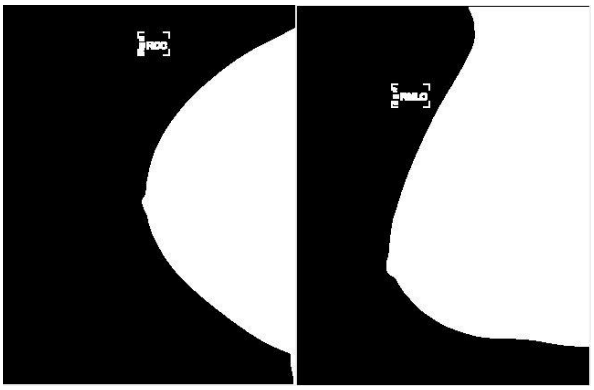

Fig 3.4: Output DICOM binary image after pre-precessing
(a) Malignant (b) Benign.

3.2 Feature Extraction

Once the pre-processing of the DICOM image is done, features are extracted and reduced using the logistic regression algorithm. Circles from the intensity image of radius between 15 and 30 are found out using the Circular Hough transform. The average radius value from all found values is then selected. Now different properties of the image region (area, perimeter, centroid, eccentricity, orientation) are calculated.

The texture of an image provides information about the spatial arrangement of intensities in the image. It helps in better image classification.

$$\text{Texture} = \frac{\text{Standard Deviation(Image)}}{10} \quad (1)$$

The smoothness of an image is a function of its color gradients. Smoothness can also be calculated from area of a particular region in an image.

$$\text{Smoothness} = \frac{\text{Perimeter}^2}{10(4\times\pi\times\text{Area})} \quad (2)$$

The compactness of an image is a dimensionless quantity that defines the degree to which a shape is compact.

$$\text{Compactness} = \frac{\text{Perimeter}^2}{100(\text{Area}-1)} \quad (3)$$

The absolute values of all these features are then compared with the features of images with malignant and benign cancer using a classifier and further predictions are made. The features are reduced from a larger set of features using logistic regression for better accuracy and easier prediction. Table 3.1 shows the typical values of the features of the first ten DICOM breast images taken for classification.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 3, March 2020

3.3 Classification

The final step of the cancer prediction model is the classification of benign and malignant cancer. For this a logistic regression algorithm is used. The logistic regression algorithm makes use of the sigmoid function for classification (Figure 3.5).

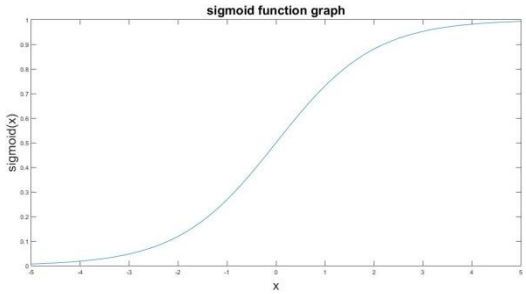

Fig 3.5: Sigmoid Function graph

It is used to estimate discrete values based on given set of independent variable(s). Logistic regression is also known as logit regression since it makes use of a logit function into which a given set of data is mapped or fitted. Thereby, much more accurate predictions on the probability of occurrence of an event is done. Since, it predicts the probability, its output values lies between 0 and 1. Figure 3.6 and 3.7 shows the result for classification using logistic regression algorithm.

Table 3.1: Typical values of features extracted for predicting malignancy.

Image	Radius (mm)	Area (mm^2)	Eccentricity	Perimeter (mm)	Texture	Smoothness	Compactness	Symmetry
1	22.0552	2603452	0.8510	7402.844	0.4617	1.6750	20.0497	-0.8982
2	22.9264	2573423	0.8708	7391.5670	0.4601	1.6894	20.2305	-0.8872
3	24.6845	3807580	0.8964	8613.0790	0.4973	1.5504	18.4835	-0.8806
4	24.6845	4082192	0.8839	8797.1080	0.4995	1.50860	17.9577	0.8662
5	19.5633	3435329	0.8039	7945.4480	0.4908	1.462	17.3767	0.8332
6	20.5844	3361277	0.7808	7844.3990	0.4890	1.4568	17.3069	-0.8235
7	21.6393	4019484	0.8661	9503.1390	0.4991	1.7879	21.4679	-0.8394
8	21.6393	4118781	0.8572	9353.7380	0.4996	1.6904	20.2423	0.8633
9	23.0018	3876084	0.8902	9528.6220	0.4518	1.8640	22.4243	-0.8853
10	23.4185	3636398	0.8576	7958.1960	0.4430	1.3859	16.4163	-0.8821

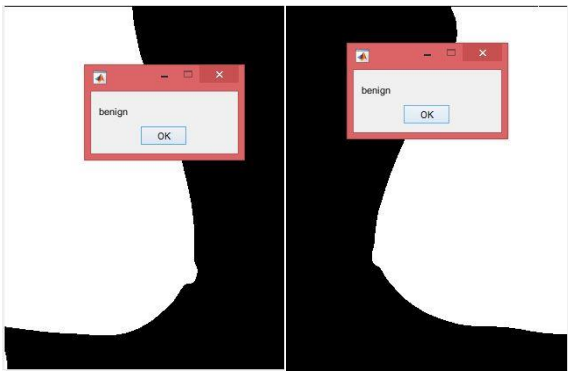

Fig 3.6: Classification of benign cancer in left and right breasts.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 3, March 2020

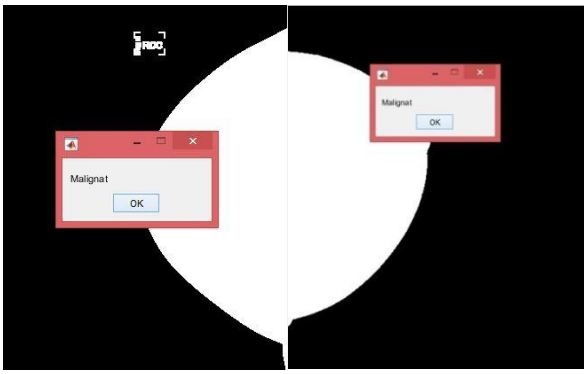

Fig 3.7: Classification of malignant cancer in left and right breasts.

4. SYSTEM PERFORMANCE

The performance of the system model was evaluated by performing a tenfold cross validation of the extracted features on different classifiers like SVM, KNN and Decision tree. The cross validation of the acquired features was done fifty times and the accuracy rates as well as the error rates were found. The performance evaluation of the features on the classifiers can be explained with the help of the following results. The confusion matrices shows the true results over the predicted results of the first set of cross validation where the features of the first ten images where taken as the test data and the remaining features here taken as the training data. Figure 4.1, 4.2 and 4.3 gives the confusion matrices for support vector machine, k-nearest neighbour and decision tree algorithm.

Fig 4.1: SVM per predicted class

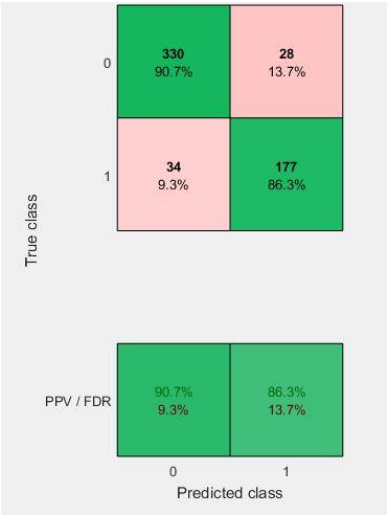

Fig 4.2: KNN per predicted class

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 3, March 2020

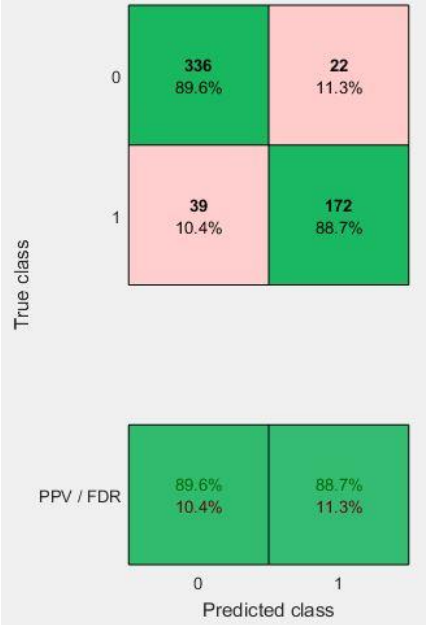

Fig 4.3: Decision tree per predicted class

Table 4.1 gives the minimum accuracy, maximum accuracy, overall accuracy and the overall error rates of the three classifiers used for cross validation.

Table 4.1: Accuracy rates and error rates of classifiers

Classifier	Min. accuracy	Max. accuracy	Overall accuracy	Overall error rate
SVM	91.2%	93.2%	92.4%	7.6%
KNN	86.3%	90.7%	89.1%	10.9%
Decision Tree	88.7%	89.6%	89.3%	10.7%

5. CONCLUSION

The rapid changes in the food habits and daily routine of human have evolved several invasive diseases among which breast cancer has a prominent position. There are several computational techniques to detect breast cancer in its early stage. In this paper, a logistic regression algorithm-based classification of benign and malignant cancers using selected features of breast DICOM images was done. An accuracy level 94.8% was observed. The extracted features where then passed over a tenfold cross validation using three different classifiers (SVM, KNN, Decision Tree) and an accuracy level of 92.4%, 89.1%, 89.3% were observed.

REFERENCES

[1]. Logistic Regression Model for Breast Cancer Automatic Diagnosis, Ahmed F. Seddik, Doaa M. Shawky, SAI Intelligent Systems Conference 2015
[2]. The digital mammography dream challenge, data from Breast Cancer Surveillance Consortium, www.dreamchallenges.org
[3]. Breast cancer diagnosis using machine Learning algorithms –a survey, B.M.Gayathri., C.P.Sumathi and T.Santhanam, International Journal of Distributed and Parallel Systems (IJDPS) Vol.4, No.3, May 2013
[4]. Robust mass classification-based local binary pattern variance and shape descriptors, Alima Damak Masmoudi, Norhen Gargouri Ben Ayed and Dorra Sellami Masmoudi,CIELS , Int. J. Signal and Imaging Systems Engineering, Vol. 8, Nos. 1/2, 2015
[5]. Breast imaging: A survey, Subbhuraam Vinitha Sree, Eddie Yin-Kwee Ng, Rajendra U Acharya, and Oliver Faust , World journal of clinical oncology, Published online 2011 Apr.
[6]. Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation, Turgay Ayer, Jagpreet Chhatwal, Oguzhan Alagoz, Charles E. Kahn, Ryan W. Woods and Elizabeth S. Burnside, The journal of continuing medical education in radiology, Jan 2010
[7]. Breast cancer diagnosis and recurrence prediction using machine learning techniques, Mandeep Rana , Pooja Chandorkar , Alishiba Dsouza , Nikahat Kazi , IJRET: International Journal of Research in Engineering and Technology.
[8]. Breast cancer analysis using logistic regression, H. Yusuff, N. Mohamad U.K. Ngah & A.S. Yahaya, IJRRAS, Jan 2012

[9]. Latest Advances in Computer-Aided Detection of Breast Cancer by Mammography, R.Bhanumathi, G.R.Suresh, IJITE Vol.01 Issue-06, Nov., 2013.
[10]. Breast Cancer Classification using Support Vector Machine and Genetic Programming, K.Menaka , S.Karpagavalli, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 7, September 2013
[11]. Segmentation of Breast Regions in Mammogram Based on Density: A Review, Nafiza Saidin, Harsa Amylia Mat Sakim, Umi Kalthum Ngah and Ibrahim Lutfi Shuaib, Imaging & Computational Intelligence Group (ICI) School of Electrical and Electronic Engineering Universiti Sains Malaysia
[12]. www.cancer.org/americancancersociety
[13]. Assessment of algorithms for mitosis detection in breast cancer histopathology images, Mitko Veta , Paul J. van Diest , Stefan M. Willems , Haibo Wang , Anant Madabhushi , Angel Cruz-Roa , Fabio Gonzalez, Image Sciences Institute, University Medical Center Utrecht, The Netherlands
[14]. Inferior Breast-Chest Contour Detection in 3-D Images of the Female Torso, lijuan zhao, audrey cheong, gregory p. Reece, michelle c. Fingere,shishir k. Shah and fatima a. Merchant
[15]. Classification and Immunohistochemical Scoring of Breast Tissue Microarray Spots Telmo Amaral, Stephen J. McKenna, Katherine Robertson, and Alastair Thompson, IEEE transactions on biomedical engineering, vol. 60, no. 10, october 2013
[16]. High Accuracy Gene Signature for Chemosensitivity Prediction in
[17]. Breast Cancer, Wei Hu, tsinghua science and technology, Volume 20, Number 5, October 2015
[18]. Topological Modeling and Classification of Mammographic Microcalcification Clusters, Zhili Chen, Harry Strange, Arnau Oliver, Erika R. E. Denton, Caroline Boggis, and Reyer Zwiggelaar, IEEE transactions on biomedical engineering, vol. 62, no. 4, april 2015
[19]. AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images, Shadi Albarqouni, Christoph Baur, Felix Achilles, Vasileios Belagiannis, Stefanie Demirci, and Nassir Navab, IEEE transactions on medical imaging, vol. 35, no. 5, may 2016
[20]. A low-cost screening method for the detection of the carotid artery diseases, Ahmed F. Seddik , Doaa M. Shawky, IEEE Journal Knowledge-Based Systems, Volume 52, November, 2013
[21]. Discover the Expert: Context-Adaptive Expert Selection for Medical Diagnosis, Cem Tekin, Onur Atan, and Mihaela Van Der Schaar, Department of Electrical Engineering, University of California at Los Angeles, Los Angeles, USA
[22]. A Radar-Based Breast Cancer Detection System Using CMOS Integrated Circuits, Hang Song, Hayato Kono, special section on bio-compatible devices and bio-electromagnetics for bio-medical applications
[23]. Computational Pathology to Discriminate Benign from Malignant Intraductal Proliferations of the Breast, Fei Dong, Humayun Irshad, research article December 9, 2014
[24]. www.trauma.org
[25]. American College of Radiology, ACR BI-RADS—Mammography, Ultrasound & Magnetic Resonance Imaging, 4th ed. Reston, VA: Amer. Coll. Radiol., 2003.
[26]. Managing DICOM images: Tips and tricks for the radiologist, Dandu Ravi Varma, Department of Radiology, Krishna Institute of Medical Sciences, Hyderabad, India, IJRI, v.22(1), Jan-Mar 2012.
[27]. DICOM Image Enhancement of Mammogram Breast Cancer, Dina.R.Elshahat, Dr .M .Morsy, Prof. MohyELdin A.Abo_ELsoud, AL Mansoura University Faculty of Engineering Electronics & Comm. Dept, IJRASET, Volume 4 Issue III, March 2016.

BIOGRAPHIES

Nimmi Sudarsan received her B.Tech degree in Electronics and Communication Engineering from Calicut University, Kerala in 2015 and Mtech degree in Communication Engineering from A P J Abdul Kalam Technological University in 2017.

Nandakumar Paramparambath is working as Professor, Department of Electronics and Communication Engineering, NSS College of Engineering. His current research interests include machine learning and biomedical signal processing.

Sidharth N is working as Assistant Professor, Department of Electronics and Communication Engineering, NSS College of Engineering. His research interests include bio medical signals and image processing.