Breast Cancer Detection and Classification Using Machine Learning

R.R.Prianaka[1], Megala B[2], Nithyashree R[3]
[1]Asst. prof, Dept of computer science engineering
[2, 3]Dept of computer science engineering
[1, 2, 3]RMK College of Engineering and Technology

Abstract- *In the field of radiology, mammographic screened images (i.e. X-rays image sensing) square measure terribly difficult and difficult to interpret. The skilled radiotherapist visually hunts the mammograms for any specific abnormality. However, human factor causes an occasional degree of preciseness which frequently ends up in biopsy and anxiety for the patient concerned. This paper proposes a novel Computer-Aided Detection (CAD) system to scale back the human issue involvement and to assist the radiotherapist in automatic diagnosing of benign/malignant breast tissues by utilizing the basic morphological operations. The input Region of Interest (ROI) is extracted manually and subjected to additional variety of preprocessing stages. The geometrical and texture features are used for feature extraction of suspicious region. After that a KNN classifier is introduced to classify the required class of the breast cancer.*

I. INTRODUCTION

Breast cancer is a serious threat to women's life and health, and the morbidity and mortality of breast cancer are ranked first and second out of all female diseases. Early detection of lumps can effectively reduce the mortality rate of breast cancer. The mammogram is widely used in early screening of breast cancer due to its relatively low expense and high sensitivity to minor lesions. In the actual diagnosis process, however, the accuracy can be negatively affected by many factors, such as radiologist fatigue and distraction, the complexity of the breast structure, and the subtle characteristics of the early-stage disease. The computer-aided diagnosis (CAD) for breast cancer can help address this issue. The classical CAD for breast cancer contains three steps: (a) finding the Region of Interest (ROI) in the preprocessed mammogram, and hence locating the region of the tumor. (b) Then, extracting features of the tumor based on expert knowledge, such as shape, texture, and density, to manually generate feature vectors. (c) finally, diagnosing benign and malignant tumors by classifying these feature vectors.LGP is the other method for texture classification. Here the local pattern is computed based on the local gradient flow from one side to another side through centre pixel in a 3x3 centre pixel [9] The use of classifiers in medical diagnosis is increasing rapidly. Here we use support vector machine (SVM) classifiers which is considered as an effective learning method for classification .They rely on support vector for classification [10]. In this paper, we have computed features for the detection and classification of mammogram using LBP and LGP.

II. THE IMAGE PROCESSING SYSTEM

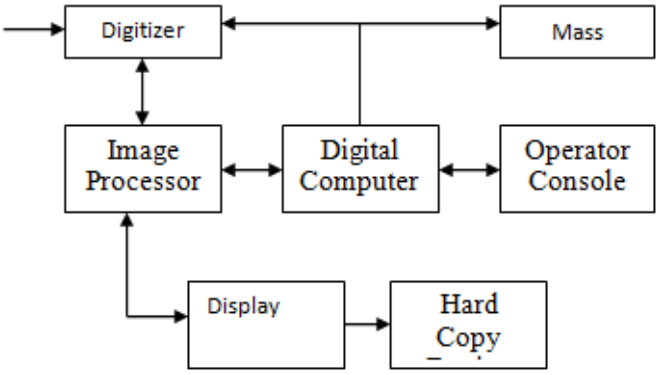


Fig 1.1 Block Diagram For Image Processing System

DIGITIZER:

A digitizer converts an image into a numerical representation suitable for input into a digital computer. Some common digitizers are

1. Microdensitometer
2. Flying spot scanner
3. Image dissector
4. Videocon camera
5. Photosensitive solid- state arrays.

IMAGE PROCESSOR:

An image processor does the functions of image acquisition, storage, preprocessing, segmentation, representation, recognition and interpretation and finally displays or records the resulting image. The following block

diagram gives the fundamental sequence involved in an image processing system.

As detailed in the diagram, the first step in the process is image acquisition by an imaging sensor in conjunction with a digitizer to digitize the image. The next step is the preprocessing step where the image is improved being fed as an input to the other processes. Preprocessing typically deals with enhancing, removing noise, isolating regions, etc. Segmentation partitions an image into its constituent parts or objects. The output of segmentation is usually raw pixel data, which consists of either the boundary of the region or the pixels in the region themselves. Representation is the process of transforming the raw pixel data into a form useful for subsequent processing by the computer. Description deals with extracting features that are basic in differentiating one class of objects from another. Recognition assigns a label to an object based on the information provided by its descriptors. Interpretation involves assigning meaning to an ensemble of recognized objects. The knowledge about a problem domain is incorporated into the knowledge base. The knowledge base guides the operation of each processing module and also controls the interaction between the modules. Not all modules need be necessarily present for a specific function. The composition of the image processing system depends on its application. The frame rate of the image processor is normally around 25 frames per second.

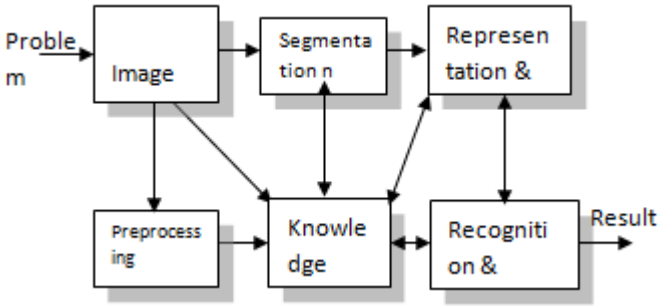


Fig 1.2 Block Diagram Of Fundamental Sequence Involved In An Image Processing System

DIGITAL COMPUTER:

Mathematical processing of the digitized image such as convolution, averaging, addition, subtraction, etc. are done by the computer.

MASS STORAGE:

The secondary storage devices normally used are floppy disks, CD ROMs etc.

HARD COPY DEVICE:

The hard copy device is used to produce a permanent copy of the image and for the storage of the software involved.

OPERATOR CONSOLE:

The operator console consists of equipment and arrangements for verification of intermediate results and for alterations in the software as and when require. The operator is also capable of checking for any resulting errors and for the entry of requisite data.

III. EXISTING SYSTEM

This paper proposes a mass detection method based on CNN deep features and Unsupervised Extreme Learning Machine (US-ELM) clustering. Second, they build a feature set fusing deep features. Third, an ELM classifier is developed using the fused feature set to classify benign and malignant breast masses.

Disadvantages:

1. Here the segmentation is not done. Because the area occupied by the disease is important factor for better diagnosis.
2. The deep learning processes need some advanced hardware requirements like high amount of RAM, graphics cards and etc.

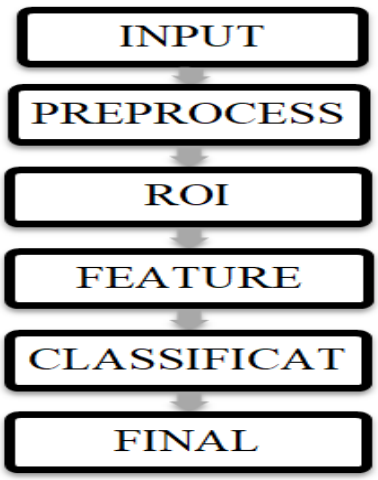
IV. PROPOSED SYSTEM

This paper proposes a novel Computer-Aided Detection (CAD) system to reduce the human factor involvement and to help the radiologist in automatic diagnosis of benign/malignant breast tissues by utilizing the input Region of Interest (ROI) is extracted manually and subjected to further number of preprocessing stages. The geometrical and texture features are extracted for feature extraction of suspicious region. After that a KNN classifier is introduced to classify the required class of the breast cancer.

Advantages:

1. Using of different Morphological operations we have to segment the affected part accurately.
2. Using of various feature extraction methods we have to identify the appropriate differences between benign and malignant breast cancer diseases.

OVERALL DIAGRAM



V. MODULES DESCRIPTION

1. INPUT:

Read and Display an input Image. Read an image into the workspace, using the imread command. In image processing, it is defined as the action of retrieving an image from some source, usually ahardware-based source for processing. It is the first step in the workflow sequence because, without an image, no processing is possible. The image that is acquired is completely unprocessed.

2. PREPROCESSING:

Data sets can require preprocessing techniques to ensure accurate, efficient, or meaningful analysis. This technique consists of **resize** the input image and converting the input image into **gray scale** image and **black and white** image using filters. Data cleaning refers to methods for finding, removing, and replacing bad or missing data. Detecting local extreme a and abrupt changes can help to identify significant data trends. Smoothing and detrending are processes for removing noise and linear trends from data, while scaling changes the bounds of the data. Grouping and binning methods are techniques that identifyrelationships among the data variables.

3. SEGMENTATION

The technique of partitioning the image into segment can be defined as image segmentation. Considering the similar property, segmentation is implemented. This similar property is our propounded approach implements the **Morphology** based segmentation techniques. This technique aids in the extraction of important image characteristics, based on which information can be easily perceived. Then we use different morphological operations like **DILATION, EROSION**, **AREA OPENING**, **CLOSING**, **BORDER CLEARING** and etc. From the all above different operations we have to segment the affected part.

4. FEATURE EXTRACTION

In pattern recognition and in image processing, feature extraction isa special form of dimensionality reduction. When the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant, then the input data will be transformed into a reduced representation set of features. Transforming the input data into the set of features is called feature extraction. If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. When performing analysis of complex data one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which overfits the training sample and generalizes poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy. Here we use **geometrical based** feature extraction method like Area, Diameter, Perimeter and **Texture based** feature extraction method like **GLCM** (Grey level co-occurrence matrix) for feature extraction. The glcm gives the texture features of the test image like contrast, correlation, energy and etc. Then the region based features gives the various different features of the input image like area, diameter etc. From the all above extracted features we have to identify the best features that are related to differentiate the benign and malignant cancers.

5. KNN CLASSIFICATION

The **KNN (K-Nearest Neighbor**) binary (as two class) is given more accurate data classification which beneficial to select k as an odd number which avoids the irregular data. The KNN procedure is the technique in ML procedures: It is an object which classified through a mainstream selection of its neighbors, with the determination assigned occurrence for most mutual class amongst its k nearest neighbors (k is a positive integer, classically small). Classically Euclidean distance is used as the distance metric; however, this is only suitable for endless variables. KNN is a

new process that deliveries all available cases and categorizes novel cases built on an evaluation quantity (e.g., distance functions). KNN procedure is identical simple. It works built on a minimum distance from the interrogation instance to the training samples to regulate the K-nearest neighbors. The information for KNN procedure contains numerous attribute which will be used to categorize. The information of KNN can be any dimension scale from insignificant, to measurable scale.

VI. CONCLUSION

Malicious URL detection plays a critical role for many cyber security applications, and clearly deep learning approaches are a promising direction. In this article, the support vector machine algorithm based on Term frequency – inverse document frequency is compared with the logistic regression algorithm and the CNN algorithm based on the word2vac feature. By comparing the three aspects (precision, recall, and f1 –score) of SVM, logical regression and CNN, we can get a conclusion. The use of Term frequency–inverse document frequency of SVM with logical regression method, SVM of these three aspects (precision, recall, and f1 – score) are slightly higher than the logical regression algorithm. The convolution neural network based on Word2vac is consistent with the SVM algorithm based on Term frequency–inverse document frequency.

REFERENCES

[1] R. L. Siegel, K. D. Miller, S. A. Fedewa, D. J. Ahnen, R. G. S. Meester, A. B. M. PhD, and A. J. D. PhD, "Colorectal cancer statistics, 2017," Ca A Cancer Journal for Clinicians, vol. 67, no. 3, pp. 177–193, 2017.

[2] J. B. Harford, "Breast-cancer early detection in low-income and middle-income countries: do what you can versus one size fits all," Lancet Oncology, vol. 12, no. 3, pp. 306–312, 2011.

[3] C. Lerman, M. Daly, C. Sands, A. Balshem, E. Lustbader, T. Heggan, L. Goldstein, J. James, and P. Engstrom, "Mammography adherence and psychological distress among women at risk for breast cancer," Journal of the National Cancer Institute, vol. 85, no. 13, pp. 1074–1080, 1993.

[4] P. T. Huynh, A. M. Jarolimek, and S. Daye, "The false-negative mammogram," Radiographics, vol. 18, no. 5, pp. 1137–54, 1998.

[5] M. G. Ertosun and D. L. Rubin, "Probabilistic visual search for masses within mammography images using deeplearning," in IEEE International Conference on Bioinformatics and Biomedicine, 2015, pp. 1310–1315.

[6] S. D. Tzikopoulos, M. E. Mavroforakis, and H. V. Georgiou, "A fully automated scheme for mammographic segmentation and classification based on breast density and asymmetry," Computer Methods and Programs in Biomedicine, vol. 102, no. 1, pp. 47–63, 2011.