Breast Cancer Prediction and Tracking using Machine Learning and Image Processing

B M Harshitha1, Charitha K P2, Shwetha K 3, Keerthana V4

bmharshitha1997@gmail.com1, charithakp5@gmail.com2, shwethaharish1997@gmail.com3, [keerthanav.gubbi@gmail.com](mailto:keerthanav.gubbi@gmail.com)4

8th sem CSE students

Global Academy of Technology

Bengaluru

**ABSTRACT:**

***Today there are more than 1.15 million cases of breast cancer diagnosed worldwide annually. At present, only small numbers of accurate prognostic and predictive factors are used clinically for managing the patients with breast cancer. Early detection of this fatal disease is very important which helps in decreasing the morality rate and increasing the survival period of breast cancer patients. The project uses Mammography which is the main test used for screening and early diagnosis, and its analysis and processing are the keys to improving breast cancer prognosis. Texture features are extracted from surrounding tissue regions of interest (ST-ROIs) with Level Set and FCM technique. Morphological operators are used to distinguish masses and microcalcifications from the background tissue. KNN algorithm is used for classification. To discover the cancer growth in incremental unique person’s continuous MRI images is a big challenge with multiple cells wise growth. A new fuzzy level set algorithm is proposed to facilitate medical image segmentation. It is able to directly evolve from the initial segmentation by spatial fuzzy clustering. The controlling parameter of level set evolution are also estimated from the results of fuzzy clustering. Moreover, the fuzzy level set algorithm is enhanced with locally regularized evolution. Such improvements facilitate level set manipulation and lead to more robust segmentation. Performance evaluation of the proposed algorithm will be carried on medical images (breast cancer) from different modalities. The results confirm its effectiveness for medical image segmentation with incremental growth of decrease of the cells growth.***

***Keywords: Mammography, FCM, KNN, region of interest.***

**I.INTRODUCTION**

Machine learning is an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use [statistical analysis](https://whatis.techtarget.com/definition/statistical-analysis) to predict an output while updating outputs as new data becomes available.

**Image processing** is a method to convert an image into digital form and perform some operations on it, in order to get an enhanced image or to extract some useful information from it.

Cancer is a disease in which cells in the body grow out of control. Except for skin cancer, breast cancer is the most common cancer in women in the United States. Deaths from breast cancer have declined over time but remains the second leading cause of cancer death among women overall and the leading cause of cancer death among Hispanic women. The image representation process must tackle fundamental problems which is the ineffectiveness in capturing the textural information and poor discrimination capability of the features that result in low retrieval performance. Similarity measurement is one of the major tasks of a typical CBIR system, which has greater impact on the retrieval accuracy and retrieval time. The project aims to overcome the problems such as “Which similarity measure is suitable for particular feature type and how to minimize the computation for similarity measurement?” and “Which texture feature is more representative and discriminative for defining the given query mammogram?” by using feature extraction methods, level set and clustering techniques [6].

**II. REVIEW OF LITERATURE**

The paper [1] uses microarray breast cancer data for classification of the patients using machine learning methods. In the first case, eight different machine learning algorithms are applied to the dataset and the results of classification were noted. Then in the second case, two different feature selection methods such as Recursive Feature Elimination (RFE) and Randomized Logistic Regression (RLR) were applied on the microarray breast cancer dataset and 50 features were chosen as stop criterion. Again, the same eight machine learning algorithms were applied on the modified dataset. The results of the classifications are compared with each other and with the results of the first case. The methods applied are SVM, KNN, MLP, Decision Trees, Random Forest, Logistic Regression, Adaboost and Gradient Boosting Machines. After applying the two different feature selection methods, SVM gave the best results. MLP is applied using different number of layers and neurons to examine the effect of the number of layers and neurons on the classification accuracy.

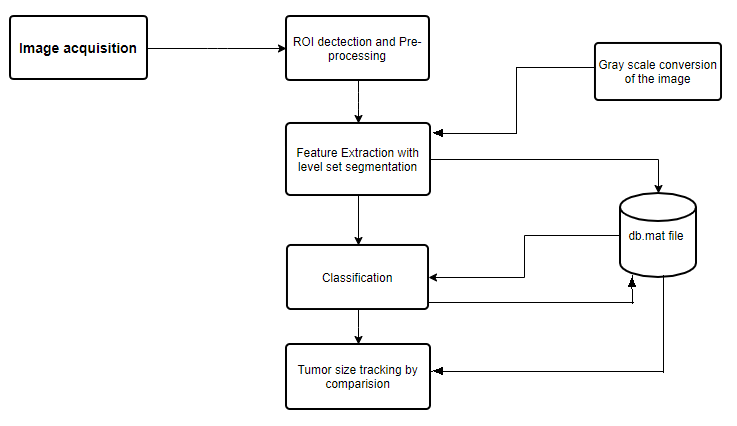
The paper [2] compares Four Novel approaches used for detection of Region of Interest in Mammographic images based on database and Real time images. In Approach I histogram equalization and dynamic thresholding techniques were used for preprocessing. Region of Interest (ROI) was partitioned from the preprocessed image by using particle swarm optimization and k-means clustering methods. In Approach II preprocessing was done using various morphological operations like erosion followed by dilation. For the identification of ROI, a modified approach of watershed segmentation was used. Approach III uses histogram equalization for preprocessing and an advanced level set approach for performing segmentation. Approach IV, which is considered to be the most efficient approach that uses different morphological operations and contrast limited adaptive histogram equalization for image preprocessing. A very novel algorithm was developed for detection of Region of Interest. Approaches I and II were applicable for Mammographic Image Analysis Society (MIAS) database images alone. Approaches III and IV were applicable for MIAS and Real time hospital images. The various graphs presented in the comparative study, clearly depicts that the novel approach that used a novel algorithm for detection of ROI is proved to be the most efficient, accurate and highly reliable approach that can be used by radiologists to detect tumors in MRM images [2].

**III. SYSTEM STUDY**

In the proposed scheme, the FCM (Fuzzy-C-means) clustering has been used for tumor region finding and segmentation. By the same thing segmented region is completely analyzed by using the Multi-level Wavelet-PCA along with GLCM features based on it [4]. After that features are extracted & process dataset is completely trained and classified by using the KNN algorithm of machine learning concept [5].

Raw features are the complex-valued responses of a set of multi-resolution Gabor filters were used on it. In that cross operator provides a simple approximation to the gradient magnitude and classical operator is detecting edges and their orientations. Have higher accuracy and less computational time. In the proposed system, classification after the extraction of texture features depends on the shapes of cancer cells in the region of interest and system also takes mean and variance measures into account for classification [7]. System will examine similarity measures and consider feature dataset reduction for similarity measurement improvement by using clustering and feature extraction techniques such as histogram of oriented gradients and speeded-up robust features [3].

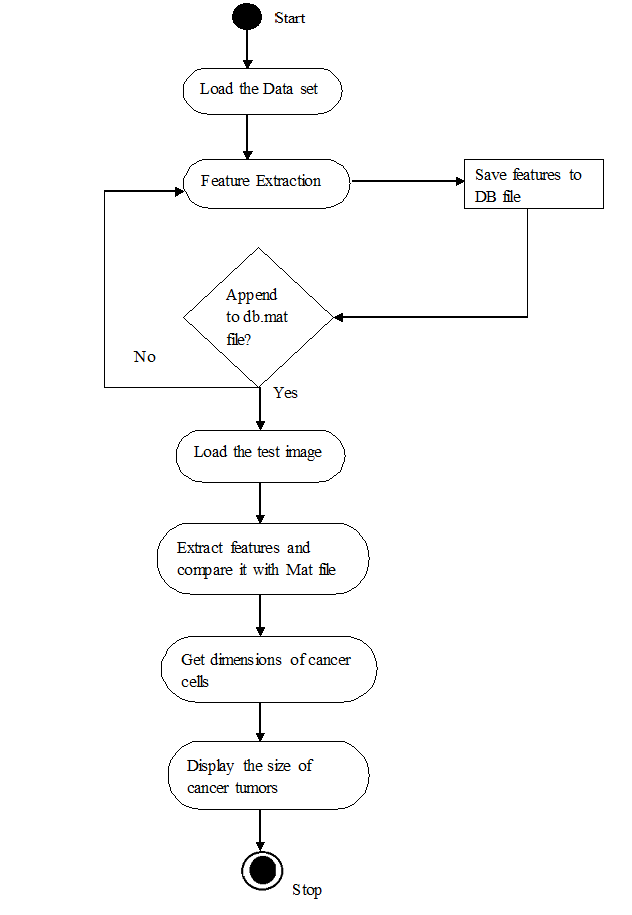
**IV. SYSTEM ARCHITECTURE**

The system architecture is the [conceptual model](https://en.wikipedia.org/wiki/Conceptual_model) that defines the structure, behaviour, technology and other views of any system.

**Figure 1: System Architecture**

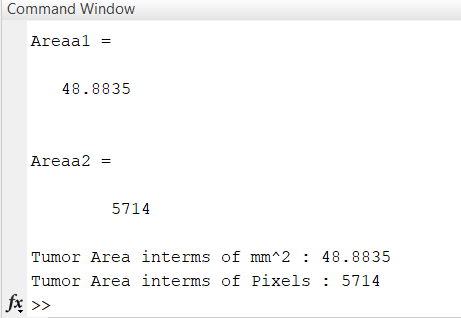
The system mainly consists of four processes. Once the image is acquitted, the system detects ROI and the image is enhanced using pre-processing techniques. The pre-processed image is then converted to its equivalent gray level ad the features are extracted by using feature extraction methods. The features are then updated to db.mat file and level set segmentation is applied on the image. Then image undergoes classification process with respect to dataset in db.mat file. If the image has cancer cells, then the size of tumor is computed and updated in db.mat file. Based on the db.mat file, size of concurrent tumors of the person can be compared and hence, the cancer cells can be tracked.

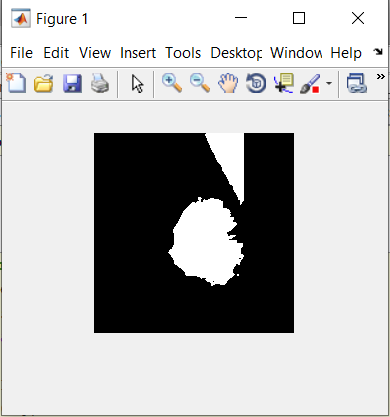
**V. ACTIVITY DIAGRAM**

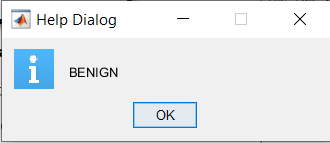
****Activity diagram is an important behavioral diagram in UML to describe dynamic aspects of the system.

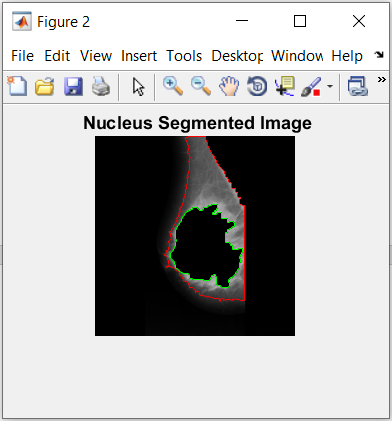
**Figure 2: Activity Diagram**

Once the data set is loaded into the system, feature extraction techniques are applied to the training data present in data set and the required features are extracted. The extracted features are appended into Mat file. If features are successfully appended to Mat file, then the doctor can upload the patient’s MRI image and features of the test image are extracted. Extracted features are then compared with the data in Mat file to classify the patient’s MRI image as either having cancer or non-cancerous. If the patient has cancer, then size of the cancerous tumors is discovered. If the extracted features cannot be successfully appended to Mat file, then the process of feature extraction repeats.

**VI. RESULTS**

In the proposed system, the process begins with detection of Region Of Interest (ROI) in MRI image of a person, by using FCM technique. Upon successful detection of ROI, pre-processing techniques are applied and surrounding tissue ROI are subjected to texture feature extraction by using Histogram of Oriented Gradients. The gray level and wavelet coefficient texture features are extracted, analysed. The ST-ROI’s are classified based on shapes of breast tumors. Classification algorithms such as KNN along with mean and variance measures are used as decision makers for prediction and tracking of breast cancer tumors.

** Figure 6.1: Gray scale equivalent of input image**

** Figure 6.2: Classification result of input image**

**Figure 6.3: Tumor Area location in the input image**

**Figure 6.4: Tumor area measurement in console**

**VII. CONCLUSION**

Breast Cancer represents one of the diseases that makes highest number of deaths every year. At present, only few accurate prognostic and predictive factors are used clinically for managing the patients with breast cancer. Here, by making use of Clustering with Level Set approach, high accuracy can be achieved in detection of effected cell shapes with exact marking on detected contours. The proposed system helps to enhance the performance of mammogram retrieval by selecting optimal features. The system makes use of Histogram of Oriented Gradients (HOG) descriptor for the purpose of object detection. These techniques improve accuracy in tracking the breast cancer cells. The project aims to assess the correctness in classifying data with respect to

efficiency and effectiveness of each algorithm in terms of accuracy, precision, sensitivity and

specificity. Hence the proposed system is designed to provide high accuracy and maximum

efficiency in prediction and tracking of breast cancer.

**REFERENCES:**

[1] Siyabend Turgut, Mustafa Dagtekin, Tolga Ensari “***Microarray Breast Cancer Data Classification Using Machine Learning Methods***” 978-1-5386-5135-3/18/$31.00 ©2018 IEEE.

[2] Varalatchoumy.M, Ravishankar. M “***Four Novel Approaches for Detection of Region of Interest in Mammograms - A Comparative Study***” Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS 2017) IEEE Xplore Compliant - Part Number: CFP17M19-ART, ISBN:978-1-5386-1959-9.

[3] Ameer Nisha.S, Shajun Nisha.S, Dr.M.Mohamed Sathik “***A Study on Surf & Hog Descriptors for Alzheimer’s Disease Detection***” (IRJET 2017) e-ISSN: 2395 -0056, p-ISSN: 2395-0072.

[4] Ammu P K, Preeja V “***Review on Feature Selection Techniques of DNA Microarray Data***” International Journal of Computer Applications, Volume 61– No.12, January 2013.

[5] LiBn, Chui CK, Chang S, Ong SH “***Integrating spatial fuzzy clustering with levelset methods for automated medical image segmentation***” Elsevier Ltd. 0010-4825/© 2010.

[6] Anna N. Karahaliou, Ioannis S. Boniatis, Spyros G. Skiadopoulos, Filippos N. Sakellaropoulos, Nikolaos S. Arikidis, Eleni A. Likaki, George S. Panayiotakis, Lena I.

Costaridou “***Breast Cancer Diagnosis: Analyzing Texture of Tissue Surrounding Microcalcifications***” 1089-7771/© 2008 IEEE.

[7] P. M. Sampat, M. K. Markey, A. C. Bovik, “***Computer-aided detection and diagnosis in mammography***,” in Handbook of Image and Video Processing, A. C. Bovik, Ed., 2nd ed. New York: Academic, 2005, pp. 1195–1217