Development of An Automatic Breast Cancer Detection System Using Randomized Feature Selection Technique

*Adepoju, T.M., Oladele, M.O. , Sobowale, A.A., Jimoh, K.O. & Ayofe, O.A.
Department of Computer Engineering Technology
Federal Polytechnic Ede
Ede, Osun State, Nigeria
E-mails: atemilola@gmail.com
Phone: 07037659707

ABSTRACT

Breast cancer is becoming a leading cause of death among women in the whole world, meanwhile, it is confirmed that the early detection and accurate diagnosis of this disease can ensure a long survival of the patients. Feature selection is an important part of most learning algorithms especially for high dimensional data sets. Due to the existence of irrelevant and redundant attributes, selecting only the relevant attributes of the data gives higher predictive accuracy expected from a machine learning method. This project was therefore designed to develop an automatic feature construction technique that will select relevant features from extracted features for classification of breast cancer. Breast images used in the developed system were acquired from segmented images from an available database. Extraction of discriminate features from the segmented breast images was carried out using Gray Level Co-occurrence Matrix. Selection of features from the extracted features was carried out using Randomized feature selection method. Based on the selected features, segmented breast images were classified into normal, benign and malignant using K- Nearest Neighbour. The developed system was evaluated using sensitivity, specificity and overall accuracy. The randomized feature selection algorithm with KNN classifier achieved a sensitivity of 87.04%, specificity of 87.65% and overall accuracy of 87.41%. The developed system has improved the technique of detection and classification of breast cancers in terms of overall accuracy. Therefore, this system can be adopted in medical field for better classification of breast cancer.

Keywords: Mammogram, masses, KNN, segmentation, Benign, Gray Level Co-occurrence Matrix (GLCM), Classification, , Malignant, Sensitivity, Specificity, Accuracy.

SMART-iSTEAMS Conference Proceedings Paper Citation Format
*Adepoju, T.M., Oladele, M.O., Sobowale, A.A., Jimoh, K.O. & Ayofe, O.A. (2018): Development of An Automatic Breast Cancer Detection System Using Randomized Feature Selection Technique.
Proceedings of the 10th SMART- iSTEAMS Multidisciplinary Conference, February, 2018, Ogwuashi-uku, Delta State, Nigeria. Pp 17-26

1. BACKGROUND TO THE STUDY

Breast cancer is the most prevalent cancer that affects women all over the world. The initial symptom of a breast cancer is the formation of a lump. This is due to tiny deposits of calcium called micro-calcifications and tumors called masses (Anu and Sindhu, 2015). Breast mass is used to indicate a localized swelling or tumor in the breast which is usually described by its location, size, shape and margin characteristics.

A mass can either be benign (breast-abscess, fat or necrosis) or malignant (cancer) (McGraw, 2015). Early detection and treatment of breast cancer could decline the mortality rate; it often leads to more effective treatment with fewer side effects. However, its early detection is difficult since there are no symptoms during the first stages of breast cancer development (Adeyemo, 2017). Different imaging techniques such as magnetic resonance, thermography, mammography and ultrasound images are possible for early detection of breast cancer (Adeyemo, 2017). Mammography is at present, the best available examination for the detection of early signs of breast cancer (Adepoju, 2015).

Randomization is an algorithmic technique that has been used to produce provably efficient algorithms for a wide variety of problems. Randomized variable elimination avoids the cost of evaluating many variable sets by taking large steps through the space of possible input sets. The number of variables eliminated in a single step depends on the number of currently selected variables (Ladha et al., 2011).

Randomized variable elimination is a wrapper method motivated by the idea that, in the presence of many irrelevant variables, the probability of successfully selecting several irrelevant variables simultaneously at random from the set of all variables is high. The algorithm computes the cost of attempting to remove k input variables out of n remaining variables given that r are relevant. A sequence of values for k is then found by minimizing the aggregate cost of removing all N −r irrelevant variables which provide a broader and widely applicable introduction to randomized algorithms (Ladha et al., 2011). Randomized quicksort selects a pivot point at random, but always produces a correctly sorted output. The goal of randomization is to avoid degenerate inputs, such as a pre-sorted sequence, which produce the worst-case $O(n^2)$ runtime of the deterministic (pivot point always the same) quicksort algorithm. The effect is that randomized quicksort achieves the expected runtime of O (nlogn) with high probability, regardless of input.

For classification of breast cancer, the main cause of difficulty is arisen by diversification of features (Adepoju, 2015). In clinical diagnosis, the signs of abnormality observed by radiologists are very diverse. They include the size, contrast, intensity, density and shape of edges (Adepoju et al., 2017). Feature construction is one of the key steps in the data analysis process, largely conditioning the success of any subsequent statistics or machine learning endeavor (Guyon and Elisseeff, 2003). Feature Selection, extraction and construction can be used in combination. In many cases, feature construction expands the member of features with newly constructed ones that are more expressive but they may include useless features. The feature extraction and selection is a key step in mass detection since the performance of CAD depends more on the optimization of the feature selection than the classification method. Feature selection can be used to automatically reduce those excessive features (Adepoju et al., 2017).

Classification techniques play the main role in medical field for diagnosing the disease and identifying the early treatment. Salama, Abdelhalim, and Zeid (2012) compared different classification techniques to find accuracy among three different breast cancer datasets for which confusion matrix based on 10-fold cross validation method is used. The classification process is subsequently divided into two phases: training set (used to build the model) and testing set (used to determine the accuracy of the model). Subashini, Ramalingam and Palanivel (2010) proposed a method to assess breast density in digital mammogram.

Using Support Vector Machine (SVM) classifier, the work classifies mammogram tissues into fatty, glandular and dense. Sampaio et al., (2011) present a computational methodology that assists specialists to detect breast masses in mammogram images.

The k-NN algorithm is one of the most famous classification algorithms used for predicting the class of a record or (sample) with unspecified class based on the class of its neighbor records.

The algorithm is made of three steps as follows (Wahbeh, Al-Radaideh, Al-Kabiand, and Al-Shawakfa, 2011):

1. Calculating the distance of input record from all training records.
2. Arranging training records based on the distance and selection of K-nearest neighbor.
3. Using the class that owns the majority among the k-nearest neighbors (this method considers the class as the class of input record that is observed more than all the other classes among the K-nearest neighbors).

KNN computes the similarity (distance) between the nearest k neighbors. The neighbors are taken from a set of features extracted for which the correct classification. The k-nearest neighbor algorithm is sensitive to the local structure of the data (Kukhan, 2016). Wang, Hua and Bai (2012) proposed a hybrid approach of rough set and genetic algorithm for web page classification. This approach employs the rough set technique for feature reduction and applies genetic algorithm based on an analogy for biological evolution, and then an initial population is created by randomly generated rules. The results of this paper show that support vector machine with rough set and genetic algorithm seems to be a useful tool for inductive learning. In this paper, they used rough set method for generating classification rules for set of 360 samples of the breast cancer data. The results of this study show that rough set is a promising tool for machine learning and for building the expert system.

Sridevi and Murugan (2012) proposed a rough set-based attribute reduction for breast cancer diagnosis. The effectiveness of rough set reductive algorithm is analyzed on breast cancer dataset and the result shows that this method yields a better attribute reduction. As compared to the traditional approach, this method requires less processing time because it uses dimensionality reduction technique. This paper indicates that the rough set feature selection approach can improve the classification accuracy. Moayedi, Azimifar, Boostani and Katebi (2010) presented a study of contourlet based mammography mass classification using support vector machine (SVM). In their study, a set of statistical properties of contourlet coefficients from four decomposition levels, co-occurrence matrix features and geometrical features are used as feature vector for the region of interest (ROI).

Genetic algorithm was used for feature selection based on neural network pattern classification. They concluded that the contourlet features offer an improvement of the classification process. Eltoukhy and Faye (2014) introduced a method for feature extraction from multi resolution representations (wavelet, curvelet) for classification of digital mammograms. The proposed method selects the features according to its capability to distinguish between different classes. The method starts with both performing wavelet and curvelet transform over mammogram images.

The resulting coefficients of each image are used to construct a matrix. Each row in the matrix corresponds to an image. The most significant features, in terms of capabilities of differentiating classes are selected. The method uses threshold values to select the columns that will maximize the difference between the different classes' representatives. The proposed method is applied to the mammographic image analysis society (MIAS) dataset. The results calculated using 2x5-folds cross validation show that the proposed method is able to find an appropriate feature set that lead to significant improvement in classification accuracy. The obtained results were satisfactory and the performances of both wavelet and curvelet are presented and compared. From these literatures, it was observed that identifying a subset of the features that are the most important in determining a property leads to computational efficiency as well as better accuracy. Therefore, this project adopts the randomized method for selecting the significant features from the dataset.

This project is motivated by the fact stated in (Guyon and Elisseeff, 2003) that feature selection is primarily performed to; select relevant and informative features, limit storage requirements and increase algorithm speed, save resources in the next round of data collection or during utilization, improve performance and gain knowledge about process that generates the data or simply visualize the data. It involves feature extraction and feature selection in which Gray Level Co-occurrence Matrix (GLCM) is employed to extract features while Randomized feature selection method is used to select subset of relevant features and an optimal subset for the classification of cancer regions from normal regions when fed into K-nearest neighbor (KNN) classifier.

2. STATEMENT OF THE PROBLEM

Feature selection is an important part of most learning algorithms especially for high dimensional data sets. Due to the existence of irrelevant and redundant attributes, selecting only the relevant attributes of the data gives higher predictive accuracy expected from a machine learning method. For classification of breast cancer, the main cause of difficulty is arisen by diversification of features (Adepoju, 2015). In clinical diagnosis, the signs of abnormality observed by radiologists are very diverse. They include the size, contrast, intensity, density and shape of edges (Adepoju et al., 2017). Feature construction is one of the key steps in the data analysis process, largely conditioning the success of any subsequent statistics or machine learning endeavor (Guyon and Elisseeff, 2003).

3. OBJEJCTIVE

The main objective of this study is to design a feature extraction and selection techniques using GLCM and Randomized feature selector and classify mammogram into normal and abnormal masses, benign and malignant,

4. METHODOLOGY

The complete framework for the proposed technique, which involves feature extraction and selection, is presented in Fig.1. The first step in this research is acquisition of the segmented image from an available database. In the second step, gray level co-occurrence matrix (GLCM) feature extraction algorithm is performed on the loaded dataset and some features were extracted.

The second step involves the use of randomized algorithm for selection of significant features from the extracted features. Then the dataset with the obtained features were fed into K-Nearest Neighbor (KNN) for classification and the cancerous region were classified into normal, abnormal, benign and malignant.

Finally, the evaluation of the system was carried out such that selected features were classified in order to evaluate the performance with respect to accuracy, sensitivity and specificity using True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) as performance indices. The research was developed in a MATLAB environment. This is because MATLAB is a very powerful computing system for handling scientific and engineering calculations.

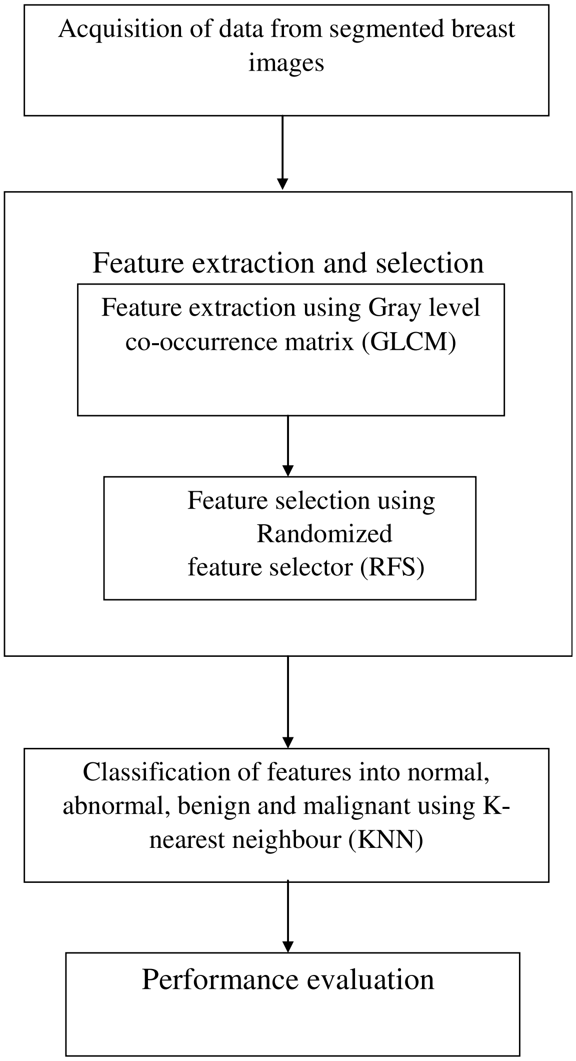

Acquisition of data from segmented breast images
Feature extraction and selection
Feature extraction using Gray level co-occurrence matrix (GLCM)
Feature selection using Randomized feature selector (RFS)
Classification of features into normal, abnormal, benign and malignant using K-nearest neighbour (KNN)
Performance evaluation

Fig 1: Framewok for the Developed System

The algorithm of the randomized feature selection is shown in Table 1

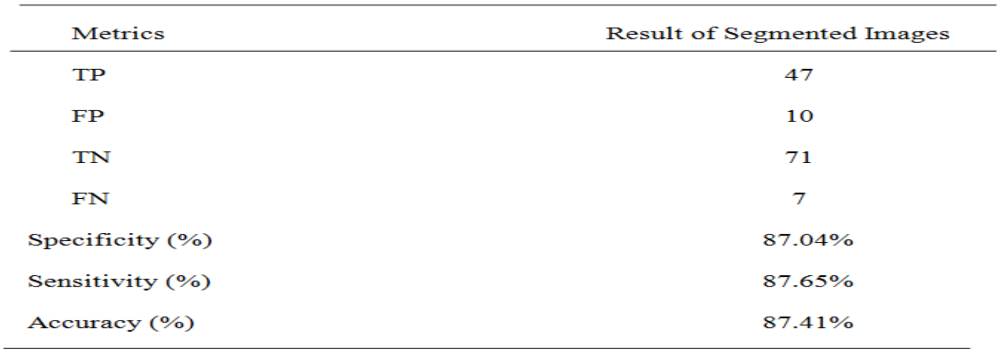
Table 1: Algorithm of the Randomized feature Selection technique

S/N	Steps
[1]	Input: The set F of all possible features and an Inductive Learning Algorithm
[2]	Randomly sample a subset F' of features from F.
[3]	Run the inductive learning algorithm L using the features in F'.
[4]	Compute the accuracy A of the concept C learnt by L.
[5]	Flip an unbiased three-sided coin with sides 1, 2, and 3.
[6]	If the outcomes of the coin flip is 1
[7]	Choose a random feature f from F - F' and add it to F'.
[8]	Remove a random feature f' from F' to get F".
[9]	Else if the outcomes of the coin flip is 2,
[10]	Choose a random feature f from F - F' and add it to F' to get F".
[11]	Else if the outcomes of the coin flip is 3,
[12]	Remove a random feature f from F' to get F".
[13]	Run the inductive learning algorithm L using the features in F".
[15]	Compute the accuracy A' of the concept C' learnt by L.
[16]	If (A' > A)
[17]	F’:= F" and A: = A'; Perform the search from F'.
[18]	Else
[19]	With probability p, perform the search from F' and
[20]	With probability 1- p, perform the search from F" with A: = A'.
[21]	Until no significant improvement in the accuracy can be obtained;
[22]	Output: A near optimal subset F' of features.

5. DATA PRESENTATION

Algorithm based on feature extraction technique was applied on 135 images to extract some parameters and features needed in the classification stage. The extracted parameters include Energy, contrast, correlation, entropy and homogeneity while 32 features were successfully extracted from the images using gray level co-occurrence matrix. Out of the 32 features, 17 features were randomly selected as relevant features using randomized feature selection technique. The efficacy of the selected features subset was measured by applying K-nearest neighbor classifier, which only used the selected features.

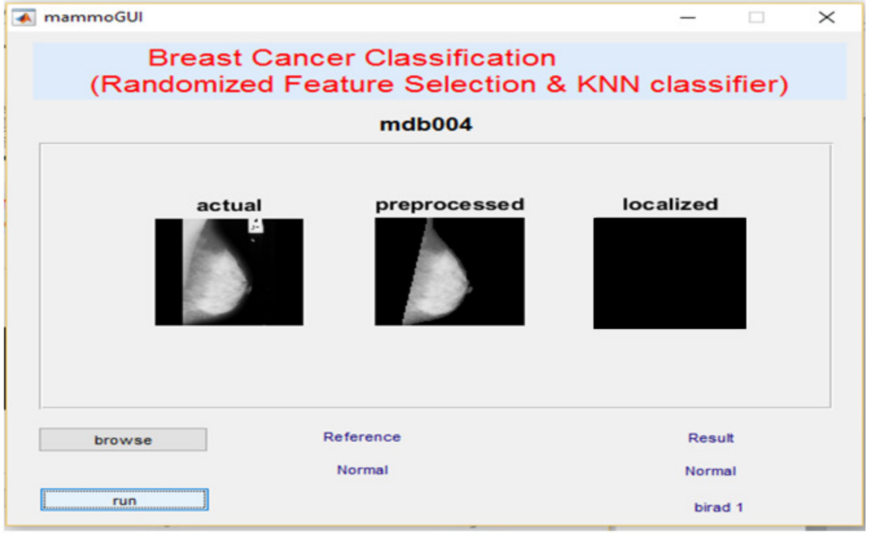
Table 2: Performance summary of the proposed system



Metrics	Result of Segmented Images
TP	47
FP	10
TN	71
FN	7
Specificity (%)	87.04%
Sensitivity (%)	87.65%
Accuracy (%)	87.41%

6. DISCUSSION OF RESULTS

An interactive graphic user interface database (GUI) was developed for the mammogram database to enable an easier use. The outlooks of the GUIs are in Fig. 2, Fig. 3 and Fig. 4. The sensitivity (87. 04%) shows the level of severity of the detected breast cancer i. e. Malignancy while the specificity (87.65%) shows that the patients are free of caner (normal). This specificity in a cancer diagnosis is the good news every patients may want. The accuracy (87.41%) of the developed system shows the level of classification of the cancer. This accuracy is the system is a reliable result for cancer diagnosis.


mammoGUI

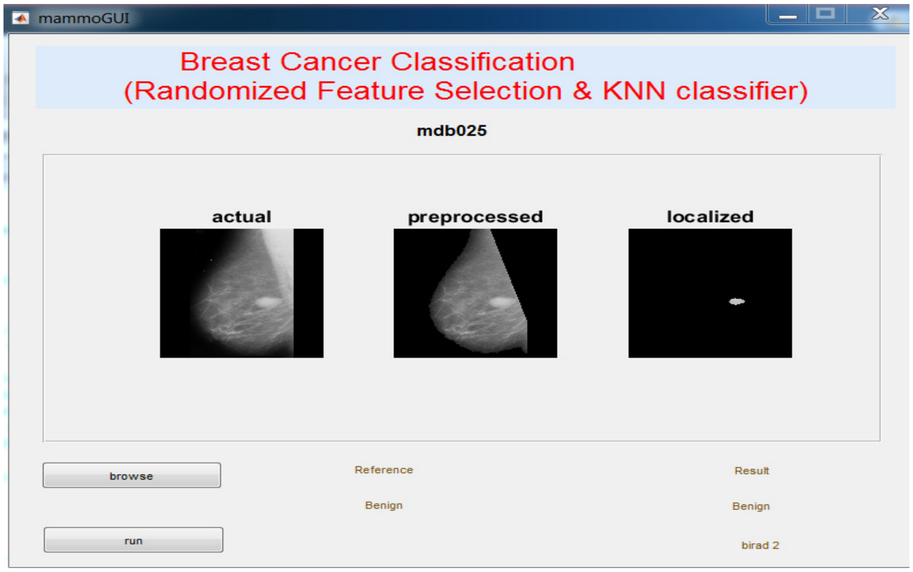
Breast Cancer Classification
(Randomized Feature Selection & KNN classifier)

mdb004

Reference
Normal

Result
Normal
birad 1

Fig. 2: Matlab GUI for feature selection and classification of a normal mammogram.


mammoGUI

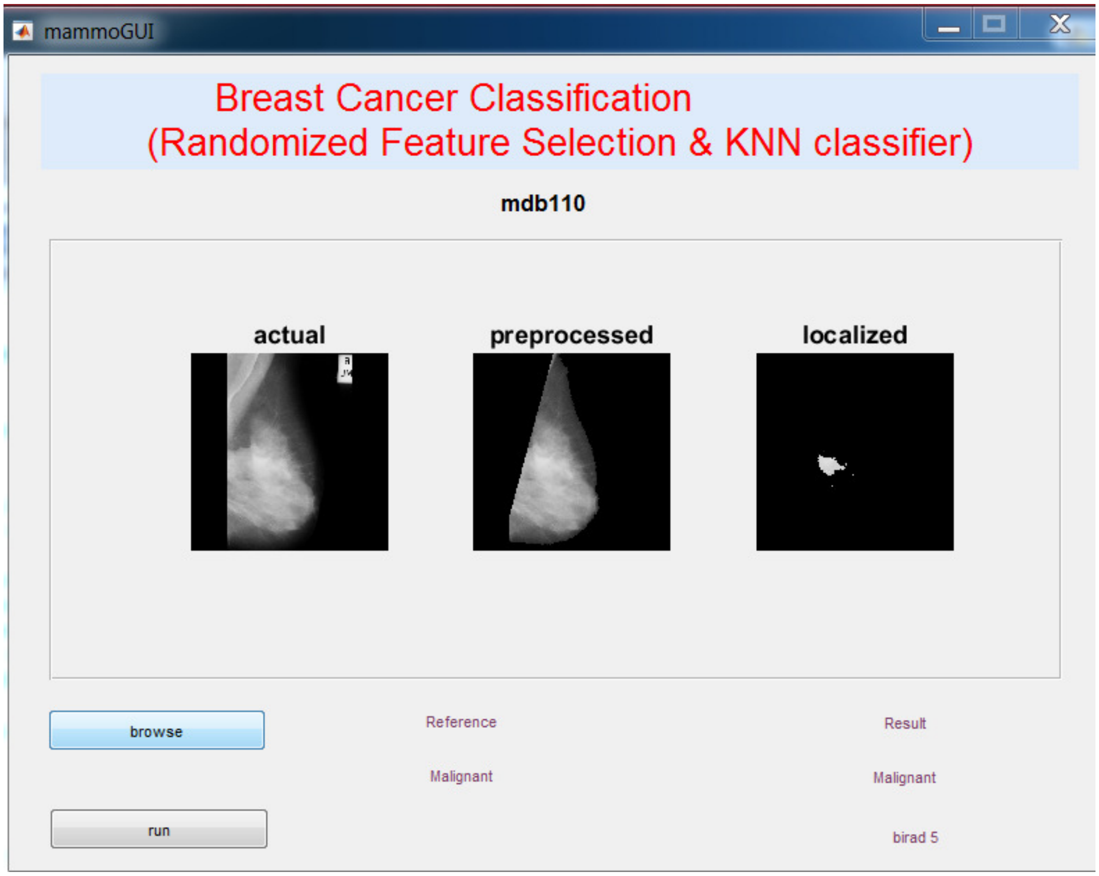
Breast Cancer Classification
(Randomized Feature Selection & KNN classifier)

mdb025

Reference
Benign

Result
Benign
birad 2

Fig. 3: Matlab GUI for feature selection and classification of a benign mammogram


mammoGUI
Breast Cancer Classification
(Randomized Feature Selection & KNN classifier)
mdb110
actual
preprocessed
localized
browse
run

Reference

Malignant

Result

Malignant

birad 5

Fig. 4: Matlab GUI for feature selection and classification of a malignant mammogram

7. CONCLUDING REMARKS

In conclusion, the developed system removed all the irrelevant features to select useful features for computation. The classification of the segmented mammogram into normal, malignant and benign achieved was an encouraged results.

8. CONTRIBUTION TO KNOWLEDGE

The developed system has improved the technique of detection and classification of breast cancers in terms of overall accuracy. Therefore, this system can be adopted in medical field for better classification of breast cancer.

REFERENCES

1. Adepoju T. M., Ojo J. A., Omidiora E. O., and Olabiyisi S. O. (2015). "Pixel-Based Morphological Technique for Breast Tumour Detection". International Journal of Scientific & Engineering Research, Volume 6, Issue 6, pp. 1416-1420.
2. Adepoju T. M., Ojo J. A., Bello O. T., Omdiora E. O., and Olabiyisi S. O. (2015). " Detection of Tumour Based on Breast Tissue Categorization", British Journal of Science and applied Technology, Volume 11, Issue 5, pp. 1-12.
3. Adepoju, Temilola Morufat, Adeyemo, Temitope Tosin, Oladele, Mattiahs Omotayo, Sobowale, Adedayo Aladejobi, Omidiora, Elijah Olusayo and Olabiyisi, Stephen Olatuide 'Segmentation of mass in Digital mammograms using Active Contour techniques (2017), Proceedings of the Seventh International conference on Mobile e-services, Vol.7, pp. 159-170.
4. Adeyemo T. T., Adepoju T. M., Sobowale A. A., Oyediran M. O., Omidiora E. O. and Olabiyisi S. O., (2017) "Feature Extraction Techniques for Mass Detection in Digital Mammogram" Journal of Scientific Research and Reports Vol. 17 No. 1 Pp 1-11.
5. Anu A, Poonjar S. (2015). Breast cancer early detection and classification techniques (Survey). International Journal of Computer Application. Vol. 132 Pp1201- 1029.
6. Eltoukhy and Faye (2014). "An Optimized Feature Selection Method For Breast Cancer Diagnosis in Digital Mammogram using Multiresolution Representation", An International Journal of Applied Mathematics and Information Science, Vol. 8, No. 6, p. 2922.
7. Guyon I., Elisseeff A., (2003). "An introduction to variable and feature selection", Journal of Machine Learning Research., Vol. 3, Pp.1157–1182.
8. Kukhan M., (2016). "A Method to Improve the Accuracy of K-Nearest Neighbor Algorithm" International Journal of Computer Engineering and Information Technology (IJCEIT), Vol. 8, No. 6, Pp.90-95.
9. Ladha et al., (2011). "Feature Selection Methods and Algorithms", International Journal on Computer Science and Engineering (IJCSE), Vol. 3, No. 5, Pp.1787-1797.
10. Moayedi F., Azimifar Z., Boostani R., and S. Katebi, (2010). "Contourlet-based Mammography Mass Classification Using the SVM Family," Computers in Biology and Medicine, Vol. 40, Pp. 373-383.
11. Salama, G. I., Abdelhalim, M., and Zeid, M. A.-e. (2012). Breast cancer diagnosis on three different datasets using multi-classifiers. Breast Cancer (WDBC), Vol. 32 No. 2 p 569.
12. Setiono R., and Liu H., (2001) "Neural-Network Feature Selector" Term Paper for Department of Information System and Computer Science National University of Singapore. Kent Publl.
13. Sridevi T. and Murugan A. (2012), Indian Journal of Innovations and Developments, Vol. 1, No. 5, P 15.
14. Subashini T.S., Ramalingam V. and Palanivel S., (2010). "Automated assessment of breast tissue density in digital mammograms", Computer Vision and Image Understanding", No.114, Pp.33-44.

15. Sampaio, W. Borges, Diniz, E. Moraes, Silva, A. Corrêa , Cardoso de Paiva, A. and Gattass, M. (2011). "Detection of masses in mammogram images using CNN, geostatistic functions and SVM," Computers in biology and medicine, Vol. 41, Pp. 653-664.

16. Wahbeh, A. H., Al-Radaideh Q. A., Al-Kabiand, M. N. and Al-Shawakfa¸ E. M. (2011) "A Comparison Study Between Data Mining Tools Over Some Classification Methods", International Journal Of Advanced Computer Science And Applications. Vol. 35¸, pp. 18-26.

17. Wang X., Hua Z. and Bai R., (2012). “A Hybrid Text Classification model based on Rough Sets and Genetic Algorithms”, Proceedings of IEEE Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, Pp. 971-977.