PressAcademia Procedia

2nd World Conference on Technology, Innovation and Entrepreneurship
May 12- 14, 2017, Istanbul, Turkey. Edited by Sefer Şener

ENHANCING BREAST CANCER DETECTION USING DATA MINING CLASSIFICATION TECHNIQUES

DOI: 10.17261/Pressacademia.2017.605
PAP-WCTIE-V.5-2017(43)-p.310-316

Florije Ismaili[1], Luzana Shabani[2], Bujar Raufi[3], Jaumin Ajdari[4], Xhemal Zenuni[5]
[1]South East European University. f.ismaili@seeu.edu.mk
[2]State University of Tetovo. Luzana.shabani@unite.edu.mk
[3]South East European University. b.raufi@seeu.edu.mk
[4]South East European University. j.ajdari@seeu.edu.mk
[5]South East European University. xh.zenuni@seeu.edu.mk

ABSTRACT

Cancer is one of the crucial causes of death for both men and women. All over the world, breast cancer is one of the leading cause of cancer deaths in women. The most effective way to reduce cancer death is to detect it earlier but the detection of cancer in early stages is not an easy process. As result, many researches are focused on developing different systems for breast cancer detection. In this paper we have discussed various data mining approaches that have been utilized for breast cancer diagnosis and prognosis. We have proposed a breast cancer prediction framework consisting of four main modules: Data Collection, Data Preprocessing, Feature Selection, and Classification. Evaluation results are provided as well. The goal is to find the best combination for feature extraction algorithm and classification algorithm, which will improve the accuracy of mammograms classification process.

Keywords: Breast cancer, mammograms classification, data mining

1. INTRODUCTION

Breast Cancer is among the leading causes of cancer death in women. Although mammography is currently the most effective tool for early detection of breast cancer, it has some restrictions. Radiologists visually search mammograms for specific abnormalities, but detection of suspicious abnormalities is a repetitive and wearing task, thus an abnormality may be unnoticed. Consequently, research for breast cancer detection is focused in finding computer aided methods, developed to aid radiologists in detecting mammographic lesions that may indicate the presence of breast cancer [1-5].

In this paper we propose a breast cancer prediction framework based on data mining classification techniques. The proposed framework consists of four major steps of determining the breast cancer: collection of mammogram images, image preprocessing, classification and result evaluation. To evaluate the accuracy of the proposed model, experimental results are provided as well.

The rest of the paper is organized as follows: section two provides an overview of the related work done for breast cancer identification using data mining techniques. In section three the proposed framework is presented followed by classification methodology in section four. Section five elaborates the experimental results while section six concludes the paper.

2. LITERATURE REVIEW

A literature survey showed that there have been several studies on the breast cancer detection using data mining techniques.

Salama, Gouda I., M. B. Abdelhalim, and Magdy Abd-elghany Zeid [6] applied various classification algorithms on three different breast cancer databases: Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC). Experiments are performed using 10-fold cross validation method combined

with tree, Multi-LayerPerception, Naive Bayes, Sequential Minimal Optimization, and Instance-Based for K-Nearest neighbor. According to their results, classification using fusion of MLP and J48 shows better results than other classification approaches.

Mittal, Dishant, Dev Gaurav, and Sanjiban Sekhar Roy [7] proposed a hybrid method of breast cancer diagnosis which gave a significant accuracy over training set and testing set. The proposed hybrid method combines unsupervised self-organizing maps (SOM) with a supervised classifier called stochastic gradient descent (SGD). The experimental results are conducted by comparing their results with three supervised machine learning techniques: decision tree (DTs), random forests (RF) and support vector machine (SVM).

M. Vasantha et al., are concentrated in classifying the mammogram images into three categories (normal image, benign image and malignant image) decision[8]. Halawani et al. have applied different clustering algorithms in order to detect breast cancer. Experiments were conducted using Digital mammograms in the University of Erlangen-Nuremberg between 2003 and 2006 [9].

A survey done by [10] have analyzed a significant number of research done in the field of breast cancer detection. Their focus was on analyzing the different data mining techniques applied in breast cancer classification along with their advantages and disadvantages. Particularly, this survey discusses about use of the classification algorithms ID3 and C4.5 in breast cancer analysis.

A review on different machine learning applications in cancer prediction and prognosis is presented by Kourou et al. [11]. In the presented review of around 70 approaches, they came to conclusion that last years the research is focused on the development of predictive models using supervised ML methods and classification algorithms where the integration of multidimensional heterogeneous data is combined with the application of different techniques for feature selection.

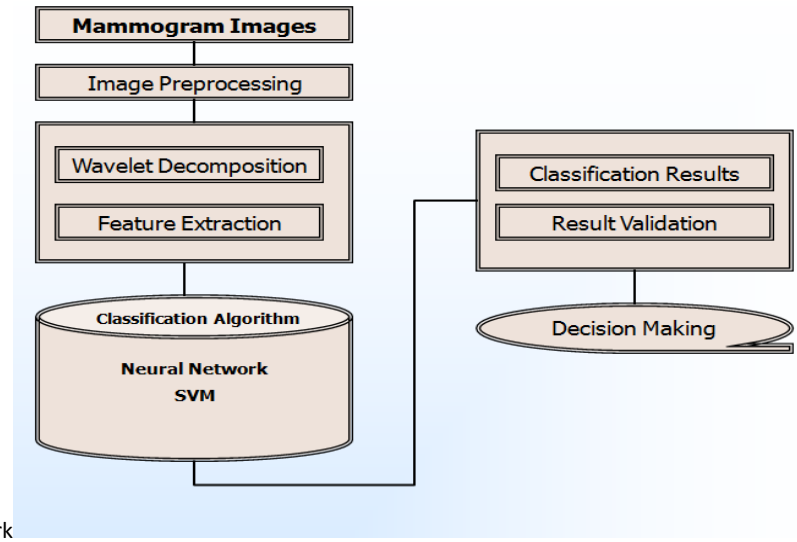
In the review done by Cruz et al. the performance of different machine learning that are being applied to cancer prediction and prognosis have been explained, compared and assessed [12]. They identified a number of trends with respect to the types of machine learning methods being used, the types of training data being integrated, the kinds of endpoint predictions being made, the types of cancers being studied and the overall performance of these methods in predicting cancer susceptibility or outcomes.

From related work it is obvious that even a huge work is done in the field of breast cancer detection, researches are still focused on enhancing the performance of cancer detection by using data mining techniques combined with the application of different techniques for feature selection, which proves the validity of the work presented in this paper.

3. PROPOSED FRAMEWORK

The aim of this research is to develop a tool for the prediction of breast cancer at its initial stage. Therefore we propose a “Breast Cancer Prediction Framework” which will contribute in decreasing mortality rate due to breast cancer. The overall architecture of the proposed framework is given in Figure 1.

Figure 1. Breast Cancer Prediction



Framework

The following are the key steps of the proposed framework:

- A corpus of data consisting of mammogram images is created.
- Mammograms are preprocessed in order to create feature vectors appropriate for classification. Feature vectors are created in two manners:
    - using global histogram equalization to obtain a uniform histogram for the output image where images were subjected to a decomposition process by wavelet transform and
    - feature extraction from image properties like radiologist's "truth"-markings on the locations of any abnormalities that may be present, character of background tissue, class of abnormality present, (x, y) image-coordinates of centre of abnormality and approximate radius (in pixels) of a circle enclosing the abnormality.
- Classifiers are trained using train data then classification algorithms are applied one by one to find out which one is producing better result in terms of accuracy for the given data set.
- Decision for breast cancer presence is made according to classification results evaluation.

4. DATA AND METHODOLOGY

This section, discusses the methodology which has been used for the proposed work. Breast Cancer Prediction Framework consists of four main modules namely: Data Collection, Data Pre-processing, Feature Selection, and Classification.

Database has been taken from Mammographic Image Analysis Society (MIAS), in order to perform experiments and evaluate the obtained results [13]. The details of all stages are discussed below.

4.1. Data Collection

First step is the collection of data. For experimental purposes we have used the MIAS digital mammography database, which consists of total 330 images. The mammograms are collected by United Kingdom National Breast Screening Programme and all mammograms follow the same criteria: only the medio-lateral oblique view is available, all films taken have been digitalized to 50 micron pixel edge with a Joyce-Loebl scanning microdensitometer.

The mammograms database includes 123 images with both benign and malignant forms of abnormalities and 207 normal mammograms. Additional information like radiologist's "truth"-markings on the locations of any abnormalities that may be present, character of background tissue (Fatty, Fatty-glandular, Denseglandular), class of abnormality present (Calcification, Welldefined/circumscribed masses, Spiculated masses, ill-defined mass, Architectural distortion, Asymmetry, Normal), (x, y) image-coordinates of centre of abnormality and approximate radius (in pixels) of a circle enclosing the abnormality are provided as well.

4.2. Data Preprocessing

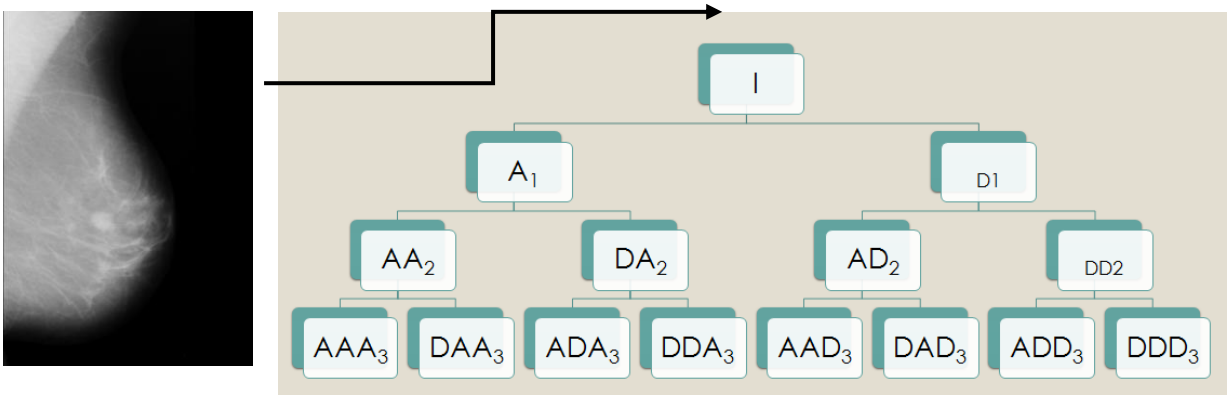
To improve the quality of the images, we use the global histogram equalization (GHE) in the pre-processing stage. The main objective of GHE is to obtain a uniform histogram for the output image. To perform histogram equalization, the running sum of the histogram values is discovered, normalized and then multiplied with maximum gray level value. These values are then mapped on the previous original values using one-to-one correspondence [14]. Basically, this method multiplies the scale factor from the normalized cumulative distribution of the brightness distribution of the original image with the original image to redistribute the intensity.

At this stage, images were subjected to a decomposition process by wavelet transform which involves decomposition of the signal. The decomposition of the multilevel wavelet transform can be expressed as follows [15]:

$$I = A_j + D_j + D_{j-1} + ... + D_2 + D_1 \quad (1)$$

where I represents the $i-th$ image, j represents the level of decomposition, A is approximation and D is detail coefficients. Figure2 demonstrates the decomposition process of the original image.

Figure 2: DWT Decomposition Process of the Original Image



The detail coefficients consist of noise, so for feature extraction, only approximation coefficients are used. Information loss can occur after level four since informative coefficients cannot be detected properly. Thus, to avoid misclassification, each image was decomposed up to four levels, i.e. As a result, by summing all the approximation coefficients at each level, a one-dimensional matrix is obtained as follows:

$$Mat = A_4 + A_3 + A_2 + A_1 \quad (2)$$

where A indicates the approximation coefficients at each level of decomposition and Mat represents the one-dimensional matrix obtained from summation of all the approximation coefficients. The resultant matrix is subjected to wavelet transform to generate the feature vectors. The general form of the wavelet transform can be written as:

$$W_y(a,b) = C(a,b) = \int_{-\infty}^{\infty} y(t)\frac{1}{\sqrt{a}}\Psi\left(\frac{t-b}{a}\right)dt \quad (3)$$

where $W_y(a,b) = C(a,b)$ is the wavelet coefficient (approximately directly proportional to the amplitude of a specific mode) with the scale a (inversely proportional to the wavelet center frequency) and position b and Ψ is the complex conjugated wavelet function [16]. To generate the feature vectors from one-dimensional matrix (M), the wavelet coefficient is converted into the mode of frequency (i.e., f_m) as follows:

$$f_v = \frac{f_{avg}(\Psi_{f.e})}{a(\Psi_{f.e})\Delta} \quad (4)$$

where $f_{avg}(\Psi_{f.e})$ is the average frequency of the wavelet function, $a(\Psi_{f.e})$ represents the approximation coefficients at all levels of decomposition, Δ indicates the image decomposition period, and f_v indicates the resultant feature vector for an image. To reduce the amount of data produced by the wavelet transform, the discrete wavelet transform (DWT) that uses a certain subset of scales- a , and positions - b is used[16].

4.3. Classification

For classifying mammogram images which are associated with benign, malign and normal classes, we used SVMs and Artificial Neural Networks (ANNs).

Neural Networks are statistical learning models used in machine learning. They are capable of a wide range of classification or pattern recognition problems since they are able to perform a range of statistical (linear, logistic and nonlinear regression) and logical operations or inferences (AND, OR, XOR, NOT, IF-THEN) as part of the classification process. They are trained to generate an output as a combination between the input variables. Multilayer Artificial Neural Network contains several intermediary called hidden layers between its input and output layers [17,18]. The basic structure of the Neural Network is given in Figure 3 below.

Figure 3: Basic structure of the Neural Network

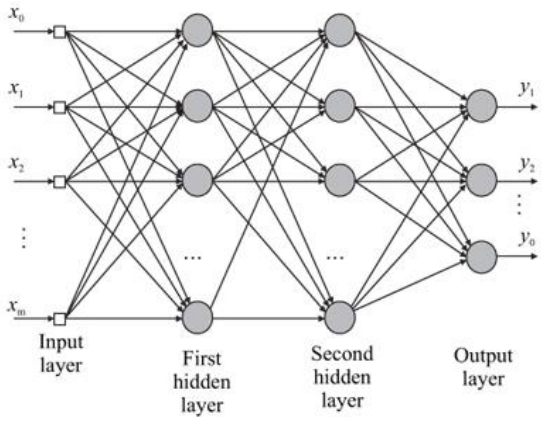
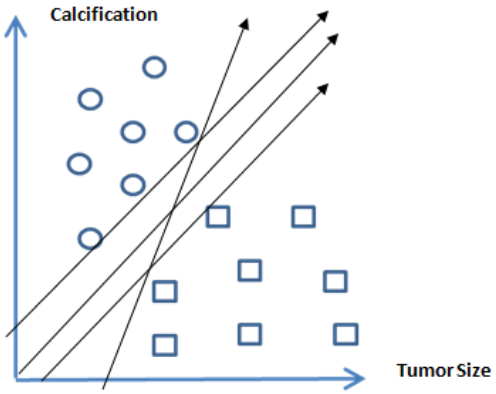


Figure 4. SVM data transformation into a higher dimensional space



Support Vector Machine (SVM) is another Data Mining technique used for data classification. Initially, SVM tries to identify the hyperplane that separates the data points into two classes by mapping the input vector into a feature space of higher dimensionality. Using non-linear kernel which is a mathematical function that transforms the data from a linear feature space to a non-linear feature space, SVM is able to perform nonlinear classification. Like ANNs, SVMs can be used in a wide range of pattern recognition and data classification [17,18]. Figure 4 illustrates how an SVM might classify benign and malignant tumors based on their size and calcification.

5. FINDINGS AND DISCUSSIONS

WEKA version 3.8.1 was utilized as a data mining tool to evaluate the performance and effectiveness of the breast cancer prediction models. Mammogram data are divided in test data and train data, then performance of various classifiers in combination with data pre-processing methods is evaluated using the breast cancer data set. The performance of a chosen classifier is validated based on precision, recall, F – measure and ROC area. Accuracy of the model is measured by the area under the ROC curve. An area of 1 represents a perfect test. Precision or sensitivity is defined by TP / (TP + FN), Recall or specificity is defined by TN / (TN + FP) while F – Measure is the harmonic mean of precision and recall.

True positive (TP) = number of examples predicted positive that are actually positive. False negative (FN) = number of examples predicted negative that are actually positive. False positive (FP) = number of examples predicted positive that are actually negative. True negative (TN) = number of examples predicted negative that are actually negative [17,18].

The analyses have been carried on using two algorithms namely, SMO and Multilayer Perceptron - MP for two types of generated feature vectors: a) feature extraction from properties associated with mammogram images and b) feature vectors created from image wavelet decomposition.

The following table demonstrates the detailed analysis of both classifications algorithms where the correctly and incorrectly classified instances show the percentage of test instances. Kappa statistics value should be high for a good model since Kappa is a chance-corrected measure of agreement between the classifications and the true classes. The Mean absolute error, Root means squared error, Relative absolute error, Root relative squared error are used to assess performance as they are frequently used measure of the differences between classes predicted by a model and the classes actually observed.

Table1: Summary of Mammogram Classification Models

	SVM		Neural Networks	
	Feature Extraction	**Wavelet Transformation**	**Feature Extraction**	**Wavelet Transformation**
Correctly Classified Instances	80.303%	86.9697 %	83.9394 %	95.1515 %
Incorrectly Classified Instances	16.697%	13.0303 %	16.0606 %	4.8485 %
Kappa statistic	0.6317	0.7562	0.7012	0.9093
Mean absolute error	0.266	0.2512	0.1162	0.0529

Root mean squared error	0.3433	0.321	0.3021	0.149
Relative absolute error	74.2453 %	70.1342 %	32.4274 %	14.7782 %
Root Relative squared error	81.2024 %	75.9352 %	71.4685 %	35.2557 %
Total Number of Instances	330	330	330	330

As can be seen in Table1, Neural Network and Support Vector Machine have comparable performances. From experimental results using two classification algorithms combined with two preprocessing techniques is obvious that both of classification algorithms shows better performance when used with image wavelet decomposition preprocessing technique in terms of higher classification accuracy and lower error rate .

Figure 4. Accuracy of SMO Predicting Breast Cancer using Feature Extraction

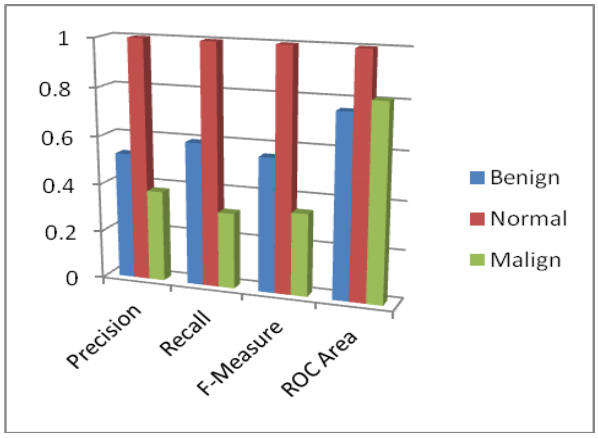


Figure 5. Accuracy of SMO Classifiers for Predicting Breast Cancer using Wavelet Decomposition

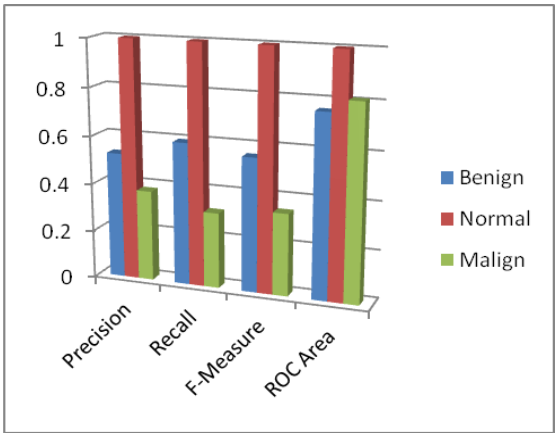


Figure 6. Accuracy of MP Predicting Breast Cancer using Feature Extraction

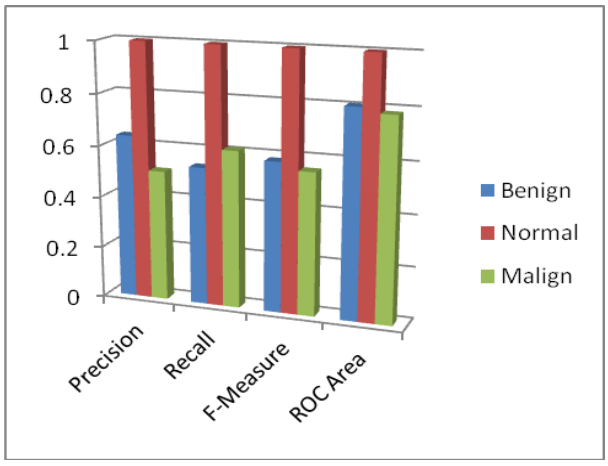


Figure 7. Accuracy of MP Classifiers for Predicting Breast Cancer using Wavelet Decomposition

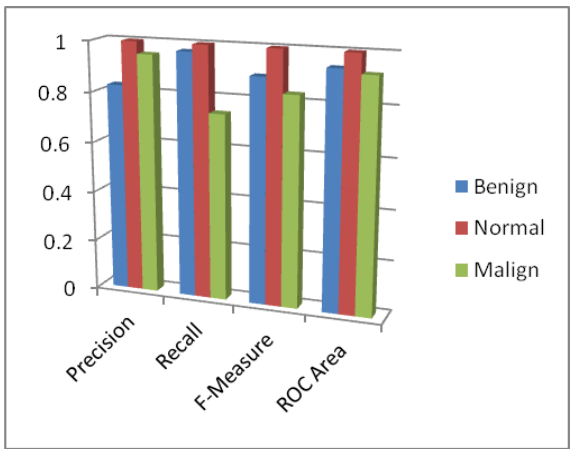


Figure 4, 5, 6, 7 shows the Precision, Recall, F – measure and ROC Area values for SMO and MP classification algorithms. It can be seen that normal mammograms are correctly classified in all cases while mammograms classified as benign and malign have higher precision (0.827, 0952), recall (0.971, 0741), F-measure (0.893, 0.833) and ROC Area (0.94, 0.923) values when Neural Network is used for classification and feature vectors are prepared using image wavelet decomposition technique.

Since cancer detection at very early stage is very sensitive, from experimental results we can conclude that the proposed framework is contributing in detecting predispositions of a patient for having breast cancer. A single user input data (mammogram) is fed into the system and gets preprocessed and classified according to the explained technique. For all

used technique combinations, non cancer data are correctly classified in all cases. If a mammogram is classified as benign or malign than the user should for surely proceed with further analyses. With each new entry getting appended to the model the process becomes intelligent and ensures accurate results.

6. CONCLUSION

This paper has outlined and discussed various data mining approaches and techniques for the problem of breast cancer detection. Although different classification techniques have been developed for cancer classification, there are still many drawbacks in their classification capability.

In order to enhance breast cancer classification, in this paper we proposed a new framework for breast cancer classification by combining mammogram wavelet transformation and neural network. According to results, classification based on locations of any abnormalities that may be present, character of background tissue, class of abnormality present, does not always shows the desired result. Finally, the evaluation and performance analysis of the proposed approach clearly shows that the preliminary results are promising in breast cancer discovery at early stage.

REFERENCES

[1] V. Gaike, R. Mhaske, S. Sonawane, N. Akhter, P. D. Deshmukh, "Clustering of breast cancer tumor using third order GLCM feature", , vol. 00, no. , pp. 318-322, 2015

[2] A. Sahar "Predicting the Serverity of Breast Masses with Data Mining Methods" International Journal of Computer Science Issues, Vol. 10, Issues 2, No 2, March 2013 ISSN (Print):1694-0814| ISSN (Online):1694-0784 www.IJCSI.org.

[3] S. Sondele and I. Saini, "Classification of Mammograms Using Bidimensional Empirical Mode Decomposition Based Features and Artificial Neural Network", International Journal of Bio-Science and Bio-Technology, Vol.5, No.6 (2013), pp.171-180.

[4] Z. K. Senturk and R. "Breast Cancer Diagnosis Via Data Mining: Performance Analysis of Seven Different Algorithms", Computer Science & Engineering: An International Journal (CSEIJ), Vol. 4, No. 1, February 2014.

[5] M. P. Sampat, M. K. Markey, and A. C. Bovik, "Computer-Aided Detection and Diagnosis in Mammography," in Handbook of Image and Video Processing, A. C. Bovik, Ed., 2nd ed, 2005.

[6] Salama Gouda I., M. B. Abdelhalim, and Magdy Abd-elghany Zeid. "Experimental comparison of classifiers for breast cancer diagnosis." Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on. IEEE, 2012.

[7] Mittal Dishant, Dev Gaurav, and Sanjiban Sekhar Roy. "An effective hybridized classifier for breast cancer diagnosis." 2015 IEEE International Conference on Advanced Intelligent Mechatronics (AIM). IEEE, 2015.

[8] M. Vasantha, S. Bharathi V and R. Dhamodharan, "Medical image feature, extraction, selection and classification". Intl. J. Engg.Sci. & Technol. 2(6), 2071-2076, 2010.

[9] S M Halawani "A study of digital mammograms by using clustering algorithms" Journal of Scientific & Industrial Research Vol. 71, September 2012, pp. 594-600.

[10] B. Padmapriya and T. Velmurugan, "A survey on breast cancer analysis using data mining techniques," 2014 IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, 2014, pp. 1-4.

[11] K. Kourou, T. P. Exarchos, K. P. Exarchos , M. V. Karamouzis, D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction", Computational and Structural Biotechnology Journal 13 (2015) 8–17.

[12] Cruz JA, Wishart DS. "Applications of Machine Learning in Cancer Prediction and Prognosis". Cancer Informatics. 2006;2:59-77.

[13] Mammographic Image Analysis, http://www.mammoimage.org/databases/. Last accessed: March, 2017.

[14] R. Sharmila and R. Uma, "A new approach to image contrast enhancement using weighted threshold histogram equalization with improved switching median filter," International Journal of Advanced Engineering Sciences and Technologies, Vol. 7, 2011, pp. 206-211.

[15] P. Kour, "Immage Processing using Discrete Wavelet Transformation", International Journal of Electronics & Communication (IIJEC), Volume 3, Issue 1, January 2015, ISSN 2321-5984.

[16] J. Turunen, "A wavelet-based method for estimating damping in power systems,"Ph.D. Thesis, Department of Electrical Engineering Power Transmission Systems, Aalto University, 2011.

[17] J. Han, M. Kamber, J. Pei, "Data Mining: Concepts and Techniques", Third Edition (The Morgan Kaufmann Series in Data Management), ISBN-13: 978-9380931913, 2011.

[18] P. N. Tan, M. Steinbach, V. Kumar," Introduction to Data Mining 1st Edition", ISBN-13: 978-0321321367, Pearson Education, 2014.