Breast Cancer Computer-Aided Detection System based on Simple Statistical Features and SVM Classification

Yahia Osman[1], Umar Alqasemi[2]
Biomedical Engineering Program
Department of Electrical and Computer Engineering
King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia

***Abstract*—Computer-Aided Detection (CADe) systems are becoming very helpful and useful in supporting physicians for early detection of breast cancer. In this paper, a CADe system that is able to detect abnormal clusters in mammographic images will be implemented using different classifiers and features. The CADe system will utilize a Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) as classifiers. Adopting mammographic database from Mammographic Image Analysis Society (MIAS), for training and testing, the performance of the two types of classifiers are compared in terms of sensitivity, specificity, and accuracy. The obtained values for the previous parameters show the efficiency of the CADe system to be used as a secondary screening method in detecting abnormal clusters given the Region of Interest (ROI). The best classifier is found to be SVM showed 96% accuracy, 92% sensitivity and 100% specificity.**

Keywords—Breast cancer; MIAS; features extraction; SVM; mammogram; clusters; computer-aided detection systems; KNN; ROI

I. INTRODUCTION

Breast cancer is a disease occurred when the cells in the female breast grow randomly and out of control. The type of breast cancer depends on morphology and proliferation. A female human breast is made up of three main parts: lobules, ducts, and connective tissue. The lobules are the glands that produce milk. The ducts are the tubes that carry milk to the nipple. The connective tissue (which consists of fibrous and fatty tissue) surrounds and holds everything together [1]. Most breast cancers begin in the ducts or lobules.

Breast cancer can spread in later stages outside the breast through blood vessels and lymph vessels. When breast cancer spreads to other parts of the body, it is said to have metastasized [1]. Breast cancer is known to be the most lethal among abnormal masses leading to deaths of 2.09 million women globally in 2018 according to World Health Organization (WHO) [2]. Recently, the survival rates have been increased due to more awareness about the disease from social media and more availability and advancement of healthcare technology especially mammography and other diagnostic imaging techniques [3]. Mammography is commonly used as a diagnostic imaging technique for detecting breast cancer due to its availability, less imaging duration, and lower cost than other methods such as Magnetic Resonance Imaging (MRI). On the other hand, Ultrasound Imaging is lower in cost but worse in terms of reproducible mapping to physical location.

In mammography, there are some factors that can lead to wrong decisions among the physicians, such as the appearance of microcalcifications. Furthermore, biopsy is painful for patients to support surgery decision. Hence, the use of CADe systems may ensure the decision without the need of biopsy. Our CADe system assumes known ROI by radiologist and supposed to aid at least as a secondary diagnosis method to support surgery decision.

II. LITERATURE REVIEW

Several previous studies have been published involving CADe system for breast cancer using mammography, contributed in presenting preprocessing algorithms, new features of more statistical significance or more relevant to the morphology of the abnormal images, and classifiers of better performance combined with set of features.

Arai et al. [4] separated the database taken from Japanese Society of Computer Aided Medical Imaging Technology into two parts, training and testing with the data proportion were 74% and 26%, respectively. The author used the features that are mostly statistical including mean, variance, max, coefficient of variation, standard deviation, and two additional features, 7 Hu moments and centroid. These features are extracted from Wavelet decomposition results of each detail, horizontal, approximation, diagonal, and vertical details. Support Vector Machine (SVM) classifier is used, and obtained sensitivity and specificity of 90% and 91.43%, respectively. This study included features obtained after image transformation that may complicate the training process of the classifier and it is more computationally expensive.

Khaoula et al. [5] proposed a Computer Aided Diagnosis system using Mini-MIAS database to detect the abnormal areas in digital mammograms, using only the dense breast category and classifies them into abnormal (benign and malignant) and normal. Then, electromagnetism-like (EML) optimization algorithm, followed by the edge-based detection algorithm FIS (Fuzzy Inference System) were used to identify the suspicious structures. As a result, the performance of this method with SVM classifier in terms of accuracy is 86.36%. The features used in this study are computationally expensive while accuracy attained is low.

Pratiwi et al. [6] found that Radial Basis Function Neural Network (RBFNN) is more accurate in classifying digital mammogram image with sensitivity of 97.22% and specificity of 91.49% for normal and abnormal classification (CADe), while in classifying benign and malignant lesions (Computer Aided Diagnosis or CADx), RBFNN's sensitivity is 100% and specificity is 89.47%. The author used features from Gray-level Co-occurrence Matrix (GLCM) and suggested that using another texture-based feature extraction, such as wavelet or curvelet, may be used in breast cancer classification in the purpose of improving the accuracy.

Setiawan et al. [7] studied the usage of Law's Texture Energy Measure (LAWS) features as descriptors for classifying mammogram images. Based on result of the experiment, LAWS features give better accuracy when classifying mammogram images compared to GLCM features. The true accuracy value of benign-malignant classification (CADx) is 78.21%, but using GLCM feature, the accuracy less than 55% for each degree. In this study, the author used ANN as classifier, suggested improvement can be done by changing the architecture of neural network model or by changing the number of nodes in the hidden layer.

Saad et al. [8] introduced an algorithm using Otsu's method for detection of Microcalcifications (MCs) and automatic diagnoses of breast cancer has been developed. The enhancement evaluation parameters such as contrast improvement index (CII), peak signal-to-noise ratio (PSNR), and Edge Preservation Index (EPI) conclude that enhancement algorithm significantly improved the contrast of MCs against the background and hence improved detection of MCs. The algorithm implemented also shows that adaptive boosting (Adaboost) classification is more sensitive and accurate for the detection of both single and clustered MCs as compared to the ANN [14]. The algorithm was tested for The Digital Database for Screening Mammography (DDSM), MIAS and local database and showed high level of overall accuracy (98.68%) and sensitivity (80.15%).

Pavel et al. [9] proposed a breast cancer detection method which uses Local Binary Patterns (LBP) features for breast representation. The proposed method was evaluated on a set created from MIAS and DDSM databases. The method showed accuracy close to 84% using SVM classifier only. This study used only LBP features which showed attractive accuracy [13]. The overall performance of the classifier can be improved if the ROI has been specified in this study.

Table. I summarize the previous studies involving breast cancer images using mammogram.

III. Database

MIAS is organized by U.K research groups that are interested in the understanding of mammograms and for image processing and recognition [10]. MIAS database consists of 322 images, which belong to three classes normal, benign and malignant. There are 208 normal, 63 benign and 51 malignant mammograms.

The detailed information about MIAS database included in an introduction file in seven information columns for each mammogram, for more information, refer to [11].

The dataset used in this study is part of MIAS database, includes 72 normal images non-cancerous and 72 abnormal ones cancerous (total 96 of which are used for training including 48 normal and 48 abnormal, and total 48 of which are used for testing including 24 normal and 24 abnormal) (diagnostic details of abnormal cases are shown in Table II). The software used in this study is MATLAB V2019b network licensed through the university system.

TABLE. I. Previous Studies Summarization

Author and date	**Features Used**	**Features Elimination Technique**	**Classifiers**
Arai et al. [4]	Max, Mean, Variance, STD, CV, Centroid, 7 Hu, and Wavelet	N/A	SVM
Khaoula et al. [5]	FIS and Zernike Moments	N/A	SVM
Pratiwi et al. [6]	GLCM (ASM, Correlation, Sum Entropy, and Sum Variance)	T-test	Back-PNN and RBFNN
Setiawan et al. [7]	Laws' texture, energy measures, and GLCM	T-test for GLCM only	ANN
Saad et al. [8]	LAWS, GLCM, Kurtosis, and Skewness	N/A	ANN and Adaptive boosting
Pavel et al. [9]	LBP	N/A	SVM
Author and date	**Accuracy**	**Sensitivity**	**Specificity**
Arai et al. [4]	84.44%	90.00%	91.43%
Khaoula et al. [5]	86.36%	81.81%	90.9%
Pratiwi et al. [6]	92.1%	97.22%	91.49%
Setiawan et al. [7]	93.90%	91%	100%
Saad et al. [8]	97.92%	64.33%	74.16%
Pavel et al. [9]	84%	------	------

TABLE. II. CLASS OF ABNORMALITY PRESENTS

Class of Abnormality	**Abbreviation**	**Number of Images**
Well-defined/circumscribed masses	CIRC	22
Speculated masses	SPIC	14
Architectural distortion	ARCH	11
Asymmetry	ASYM	13
Other, ill-defined masses	MISC	12

IV. METHODOLOGY

1) Preprocessing: Region of Interest (ROI) of 32x32 pixels is cropped around the marked center of the suspicious area marked by radiologist for all the dataset images. This is to reduce the computational load and to make feature computation more concentrated in the ROI not distracted by other details in the whole breast image [12]. During our study, we tried using the full size of the mammogram images (1024x1024), but the results were not significant.

2) Features Extraction: Initially we computed 94 features starting from the first order statistics (14 features) and texture features (64 Histogram features and 16 GLCM features). Then, using the T-test (significance p-value < 5%) and classifiers performance, the added features are eliminated manually after checking both the P-value of t-test and classifiers' performance parameters including accuracy, sensitivity and specificity. At the end, the final most contributing features used in this study after rounds of trial and error are first order statistical ones including mean, median, mode, and quantile (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9), those showed best t-test significance along with best classification performance.

3) Classifiers: The classifiers used in this study are shown in Table. III:

Fig. 1 illustrates the iterative steps used while designing the CADe system. Each block/step will be explained more within the following text.

TABLE. III. THE CLASSIFIERS USED

Classifier	**Abbreviation**	**Parameters**
Support Vector Machine	SVM	Linear, Polynomial, and Radial Basis Function
K-Nearest Neighbor	KNN	1, 2, 3, 4, and 5

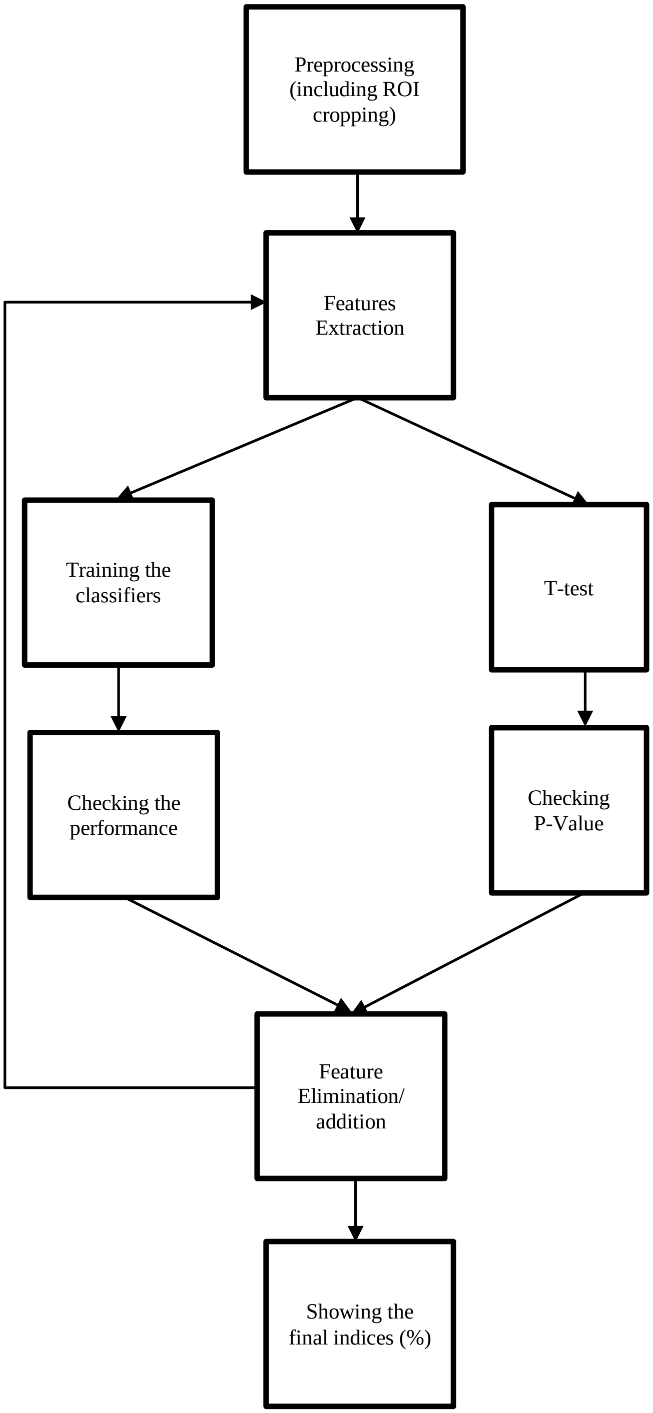
Preprocessing (including ROI cropping)
Features Extraction
Training the classifiers
T-test
Checking the performance
Checking P-Value
Feature Elimination/ addition
Showing the final indices (%)

Fig. 1. Methodology Chart.

V. Results

The final results showed that the T-Test has a number of useful features (P-Value < 0.05) = 12 out of 12. Which means that all the used first order statistical features are significantly useful. The final results are shown in Table IV.

TABLE. IV. Final Results

Indices (%)	SVM rbf	SVM Poly	SVM Linear	KNN 1
Sensitivity	NAN	88%	92%	88.5%
Specificity	50%	91%	100%	95%
PPV	0%	92%	100%	96%
NPV	100%	87.5%	92%	87.5%
Accuracy	50%	89.5%	96%	92%
Error	50%	10%	4%	8%
Indices (%)	**KNN 2**	**KNN 3**	**KNN 4**	**KNN 5**
Sensitivity	86%	89%	86%	88.5%
Specificity	100%	100%	100%	95%
PPV	100%	100%	100%	96%
NPV	83%	87.5%	83%	87.5%
Accuracy	92%	94%	92%	92%
Error	8%	6%	8%	8%

Table IV shows that the best classifier was SVM-Linear with accuracy = 96% and Sensitivity = 92%. Followed by KNN-3 with an error that is equal to 6% only. The results in comparing to the previous studies were satisfying as the features used were only the simple first order statistics.

VI. Conclusion and Discussion

In this study, the final results were impressive in comparing to previous studies those used SVM classification. Given that, we used here simple first order statistics features. On the other hand, with Neural Network based classifiers, previous studies showed that computationally more expensive features gave comparable results to what we got here. Future studies can contribute by adding the microcalcifications (MCs) to the dataset (we excluded MCs in this study) and using the sophisticated classifiers, such as, ANN and RBFNN.

Table V shows the results of our study in comparing to the previous studies that used SVM classifier:

TABLE. V. Comparing Results with SVM Classifiers' Studies

Author and date	Classifiers	Accuracy	Sensitivity	Specificity
Arai et al. [4]	SVM	84.44%	90.00%	91.43%
Khaoula et al. [5]	SVM	86.36%	81.81%	90.9%
Pavel et al. [9]	SVM	84%	------	------
Our Study	SVM	96%	92%	100%

References

[1] D. o. C. P. a. Control, "Centers for Disease Control and Prevention," 11 September 2018. [Online]. Available: https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm. [Accessed 26 December 2019].

[2] W. H. Organization, "World Health Organization," 12 September 2018. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer. [Accessed 26 December 2019].

[3] P. U. S. Alqasemi, Interviewee, Assistant Professor of Biomedical Engineering. [Interview]. 12 December 2019.

[4] I. N. A. H. O. R. K. Kohei Arai, "Improvement of Automated Detection Method for Clustered Microcalcification Based on Wavelet Transformation and Support Vector Machine," International Journal of Advanced Research in Artificial Intelligence, vol. 2, no. 4, pp. 23-28, 2013.

[5] M. N. S. A. T. Khaoula Belhaj Soulami, "A CAD system for the detection and classification of abnormalities in dense mammograms using Electromagnetism-like optimization algorithm," in 3rd International Conference on Advanced Technologies for Signal and Image Processing, Morocco , 2017.

[6] A. ,. J. H. ,. S. N. Mellisa Pratiwi, "Mammograms Classification using Gray-level Co-occurrence Matrix and Radial Basis Function Neural Network," in International Conference on Computer Science and Computational Intelligence, Indonesia, 2015.

[7] E.,.J. W. a. Y. P. Arden Sagiterry Setiawan, "Mammogram Classification using Law's Texture Energy Measure and Neural Networks," in International Conference on Computer Science and Computational Intelligence, Indonesia, 2015.

[8] A. K. Q. K. Ghada Saad, "ANN and Adaboost application for automatic detection of microcalcifications in breast cancer," The Egyptian Journal of Radiology and Nuclear Medicine, vol. 47, pp. 1803-1814, 2016.

[9] L. L. Pavel Kra´l, "LBP Features for Breast Cancer Detection," in International Conference on Image Processing, Arizona, 2016.

[10] M. A. S. Al-antari, "Computer-Aided Breast Cancer Detection and Diagnosis from Digital Mammograms," Faculty of Engineering, Cairo University Giza, Egypt, Egypt, 2015.

[11] A. F. Clark, "The mini-MIAS database of mammograms," 11 December 2012. [Online]. Available: http://peipa.essex.ac.uk/info/mias.html. [Accessed 25 July 2019].

[12] N. Petrick, H. Chan, B. Sahiner, and Datong Wei, "An adaptive density-weighted contrast enhance- ment filter for mammographic breast mass detection," Medi- cal Imaging, IEEE Transactions on, vol. 15, no. 1, pp. 59–67, 1996.

[13] A. Oliver, X. Llado´, J. Freixenet, and J. Mart, "False positive reduction in mammographic mass detection us- ing local binary patterns," MICCAI, pp. 286–293, 2007.

[14] A. AbuBaker Mass lesion detection using wavelet decomposition transform and support vector machine. International Journal of Computer Science & Information Technology. 2012 Apr 1;4(2):33.