Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 7, July 2021: 6659- 6668

Research Article

Comparative Analysis of ANN and SVM in Digital Mammogram based on Feature Selection

K.K. Kavitha,
Research Scholar (Part Time),
PG and Research Department of Computer Science
Government Arts College(A) Salem-7, India.
kavithakkcs@gmail.com

A. Kangaiammal

Assistant Professor

Department of Computer Applications
Government Arts College(A), Salem-7, India.
indurath2007@gmail.com

Abstract

This research paper focuses on feature extraction and selection strategies for detecting and classifying malignant tumors in mammograms. Features, the distinctiveness of the objects of attention, if chosen carefully, are envoy of the greatest relevant information so that the image has to suggest for absolute characterization of an abrasion. Feature extraction techniques analyze the images of an object to extract the most important features that are representative characteristics of the various classes of objects. Features are used as inputs to distinguish that allocate them to the class that they signify. GLCM (Gray Level Co-occurrence Matrix) is the universally employed technique for texture examination and it compares the gray-level divergence of any two neighboring pixels in a specified displacement and direction on an image. GLCM of an image encompasses a function of the angular connection and an interval between pixels in the neighborhood. In this paper GLCM method has been used to extract features in the proposed CAD system. Feature selection is a technique generally employed for data mining and knowledge exploration to minimize dimensionality and allows the removal of redundant features, keeping the essential hidden information simultaneously, and selection of features requires a reduced amount of data show and efficient data mining. In respect of packet collisions, data rate, and storage, it also brings potential communication benefits. Feature selection is a significant key point in machine learning and other related medical fields. In this paper, the proposed feature selection method is employed with existing classification techniques for the results, before and after feature selection. Support Vector Machine (SVM) and Artificial Neural Network (ANN) classification algorithms were used for validation. The findings show that the suggested strategy is capable of outperforming existing feature selection techniques in terms of classification performance.

Keywords: Classification, Feature Extraction, Feature Selection, GLCM, CAD, Data Mining, Machine Learning, SVM, ANN.

I INTRODUCTION

In general, a CAD system comprises of four stages: (a) Pre-processing, which removes labels, pectoral muscle, and noise sources; (b) Region of Interest (ROI) segmentation, (c) Feature extraction and selection, (d) classification. CAD systems for breast cancer diagnosis have been developed in several researches. This paper focuses on third stage feature extraction and selection.

For the classification of masses in mammography images, numerous feature extraction algorithms based on gray level, shape and texture features have been presented in recent years [1]. The first order statistical and GLCM based textural feature extraction methodologies are emphasized more [2]. In image processing, an image texture is a group of metrics determined to measure the superficial texture of an image. Image Texture provides the details on the spatial color arrangement of an image. One direct application of image texture is the recognition of image region using texture features. In the identification of these kinds of homogeneous areas, the texture is the significant visual signal [3].

While selecting images, the method of feature selection is addressed in three steps: Screening, Ranking and Selecting. 1) Screening eradicates irrelevant and challenging predictors and data. 2) Ranking, sorting the left over predictors, and allocating position based on significance. 3) Selecting: By retaining only the most relevant predictors and filtering or deleting all others; further, it recognize the subset of features. The most dominant predictors in feature selection are the Ranks, Screens and Selects. Proper use of the features will provide optimal classification performance, in addition to reducing the processor's burden of unimportant data processing in computation [4].

Preprocessing, Segmentation, and Feature Extraction and Selection approaches are the three main subjects covered in this work. CAD systems rely largely on a feature selection stage for classification in addition to these three primary categories [5]. A flowchart for a common CAD system schema is shown in Figure 1.

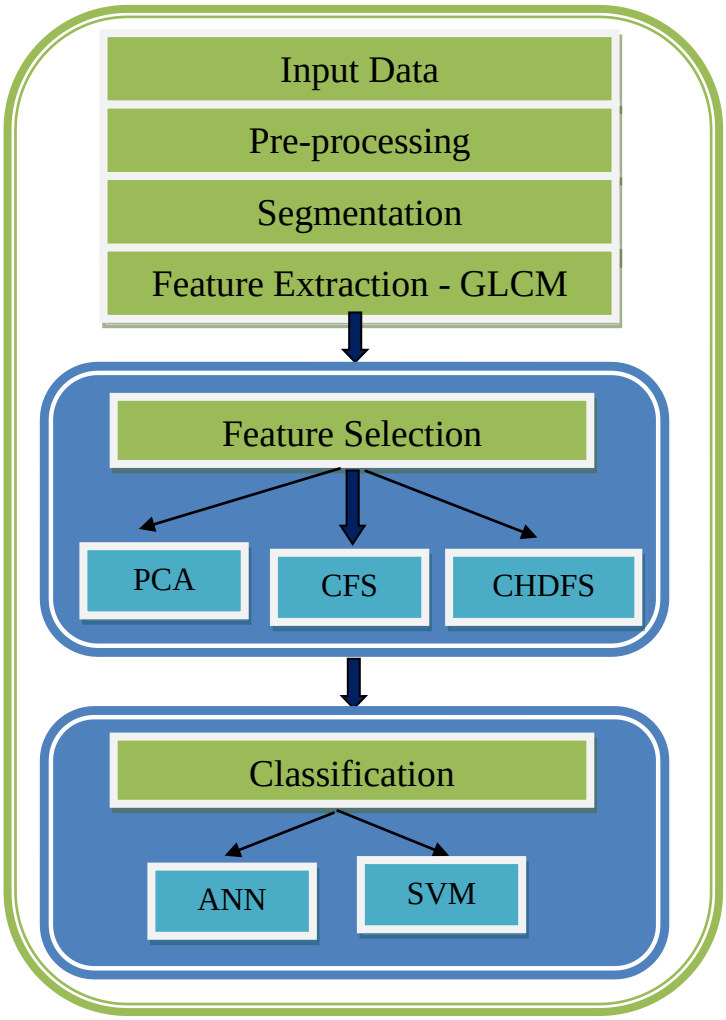

Figure 1 Framework of CAD

For a sample of unusual features, the correlation-based High Distinction Feature Selection (CHDFS) technique has been introduced and compared to existing strategies to overcome the disadvantages of execution time, different feature inter-relations, and dependency on the evaluation classifier [6]. The achievement of the proposed approach is assessed using the mini-MIAS data set.

II LITERATURE REVIEW

The GLCM is characterized in the conventional approach by Causi *et al.*, (2002) [7] as holding in its components and the number of couples of pixels with two precise values of the gray levels such as g1 and g2 and they placed at a gap which is elucidate by a dislocation vector with the n^{th} order GLCM and calculated following features like Contrast, Correlation, Entropy, Homogeneity and Variance. Higher-order GLCM features are able to assist to get better categorization rate of micro-calcification in breast tissue and it is very complicated to recognize due to its tiny size. GLCM texture measurements encompass the workhorse of image consistency because they were projected by Haralick *et al.*, (1973) and introduced 14 statistical features [8].

Haralick describes 14 textural variables measured from the probability matrix to separate the characteristics of texture information of inaccessible sensor images using the co-occurrence matrix. The

GLCM technique is employed for takeout 4 Statistical Texture constraints such as Entropy, Difference and Moment, Angular Second Moment, Inverse, and Correlation [9]. Features can be segregated based on their nature of Color, Texture and Shape.

Features that measure coarseness, smoothness and texture-based data that have high discriminatory power can be determined from the Co-occurrence matrix. The preprocessing phase helps to reduce any unnecessary noise in the image to improve the image quality. Improvement after segmenting the breast ROI enhances the result by effectively instructing the extraction of the feature. Compared to existing, the second-order GLCM features offer a better outcome than each other [10].

The preliminary phase in the process of determining the best subsets of qualities is called feature selection [11]. Attributes will be highly compatible with target variables in the CFS algorithm, but not with each other. The degree of correlation between the attributes will be between 0 and 1, with 1 indicating strongly connected attributes and 0 indicating no correlation.

PCA turns the features into principal components, which are a set of linearly not correlated variables [12]. It aids in the reduction of the original feature set's dimensionality. It reduces the amount of redundant features by mapping data from a superior dimensionality space to a minor dimensionality space. The resulting reduced set with maximum variability [13].

III METHODOLOGY

Figure 1 depicts the designed CAD system. Using the GLCM Feature Extraction method, the feature extraction phase has been performed. This final stage of the system is categorization, which allowed abnormalities to be identified.

Feature Extraction

Algorithms were used to extract aspects of possible importance. The features must be rotation and translation invariant, such that a mammography will show the same features regardless of the breast's location and orientation. Texture refers to the interaction between pixels in a neighborhood in image processing. Texture gives information on the types of patterns existing in an image by providing second-order information [14].

GLCM

A prominent texture-based feature extraction approach is Gray level co-occurrence matrix. The GLCM calculates the textural link between pixels by using second-order statistics in the images to conduct an operation. For this process, two pixels are usually employed [15]. An image's GLCM attributes are expressed as a matrix with the same number of columns and rows as the image's gray values. In this paper Haralick's definitions [16] were used to compute fourteen texture features. All of the features are numbered and presented in Table 1 for ease of reference.

Table 1 List of texture features

Features
F1 : Energy

F2 : Contrast
F3 : Correlation
F4 : Sum of squares
F5 : Inverse difference moment
F6 : Sum_average
F7 : Sum_variance
F8 : Sum_entropy
F9 : Entropy
F10 : Difference variance
F11 : Difference entropy
F12 : Information measure of correlation
F13 : Information measure of correlation
F14 : Maximal correlation coefficient

Feature Selection

Feature selection is a data mining and knowledge exploration strategy that reduces dimensionality and permits the elimination of unnecessary features while maintaining the crucial hidden information. Feature selection necessitates fewer data transmission and systematic data mining. In machine learning and other relevant medical domains, feature selection is critical. The irrelevant noisy characteristics can be removed, resulting in improved dataset quality and learning system efficiency [17]. In addition to minimizing the processor's burden of irrelevant data processing in computing, proper utilization of the features will provide best classification performance [18]. The technique finds undesired image characteristics that lower image processing efficiency, which leads to the suggested superior classification of the algorithm, and it eliminates inappropriate and recurring features from the mammographic image. Existing methods Correlation-based Feature Selection (CFS), Principal Component Analysis (PCA) has been applied. A unique feature selection Correlation-based High Distinction Feature Selection (CHDFS) technique was presented to reduce the drawbacks of diverse feature inter-relations, and dependence on the evaluation classifier.

Classification

Artificial Neural Networks

The categorization process was carried out using ANN, a supervised learning method. ANN is a computational technique based on the connections of neuronal in the human brain and other creatures' neural systems. Layers are commonly used to arrange NN. Layers are made up of a series of interconnected nodes, each with a different activation function. The network receives patterns from the input layer, which communicates with number hidden layers via weighted connections, which execute the actual processing. The hidden layers are then connected to an output layer, which outputs the answer. This study intends to build up an understanding of a specific kind of ANN said to be multi-layer perception [19].

Support Vector Machine

The SVM classifier is well-known for its ability to efficiently categories a dataset [20]. SVM is used here for mass classification. Large classification data are analyzed by SVM. By creating a hyper-plane in high-dimensional space, data is separated into two classes such as normal and abnormal. Then the abnormal cases are classified as benign and malignant. The margin between the classes must be large so that there will be fewer generalization errors. SVM performance is largely dependent on the kernel and discriminating features [21].

Data Set

The proposed methodology is developed using an imaging processing tool MATLAB. The Mammographic Image Analysis Society (MIAS) is an institute of study groups in the UK involved in mammogram understanding and has developed a digital mammogram database. The entire 322 optical mammogram images (161 breast pairs) in the mediolateral oblique vision are part of the MIAS database. In which 61 mammograms were determined as benign, 54 as malignant, and 207 as normal. The images have been reviewed by the radiologist to identify abnormalities. This research work has used mammogram images from the MIAS database to conduct the experiments.

Performance Evaluation

In this performance evaluation, the classification techniques ANN and SVM are applied before and after feature selection. The suggested technique assesses performance using well-known criteria such as accuracy, specificity and sensitivity. The sensitivity rate, also called as the true positive rate, is the amount of correctly detected positive instances to all positive cases [22], as in:

$$\text{Sensitivity (\%)} = TP / (TP + FN)$$

The specificity, also called as the true negative rate, is used to find the ratio of properly detected negative cases to total negative cases, as in:

$$\text{Specificity (\%)} = TN / (TN + FP)$$

The accuracy is examined by dividing the total number of true defined cases by the total number of cases, as in:

$$\text{Accuracy (\%)} = (TP + TN) / (TP + FN + TN + FP)$$

True positives, true negatives, false positives, and false negatives are represented as TP, TN, FP, and FN, respectively.

ANN and SVM are utilized in the suggested system to classify images as normal, benign, or malignant.

V EXPERIMENTAL RESULTS

The digital mammograms used in the proposed methodology are acquired from the MIAS database [23], which contains a categorized collection of pictures for cancer abnormalities. These images are fed into the GLCM texture-based feature extraction algorithm.

In Table 1, the 14 Haralick features are represented. These 14 features are given as input to ANN and SVM classifier. Table 2(a), 3(a) and Figure 2(b), 3(b) displays the outcomes of execution measures such as sensitivity, accuracy and specificity for feature selection based on ANN before and after feature selection.

Table 2(a) ANN before Feature Selection

lassifier/ Metrics	ANN		
	PCA	CFS	CHDFS
Accuracy	88.32	90.45	**95.68**
Specificity	89.67	91.23	**95.21**
Sensitivity	85.54	87.38	**90.62**

Fig. 2(b) ANN before Feature Selection

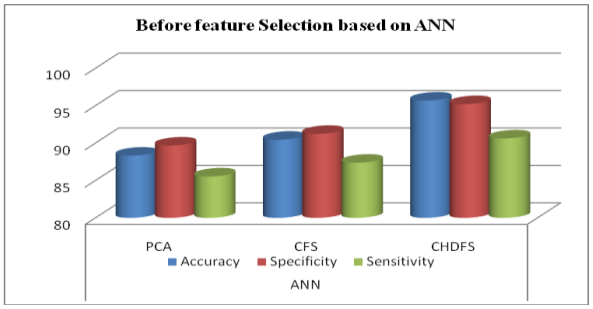

Table 3(a) ANN after Feature Selection

Classifier/ Metrics	ANN		
	PCA	CFS	CHDFS
Accuracy	93.68	94.00	**96.06**
Specificity	94.21	94.86	**96.40**
Sensitivity	87.62	89.00	**94.64**

Fig. 3(b) ANN after Feature Selection

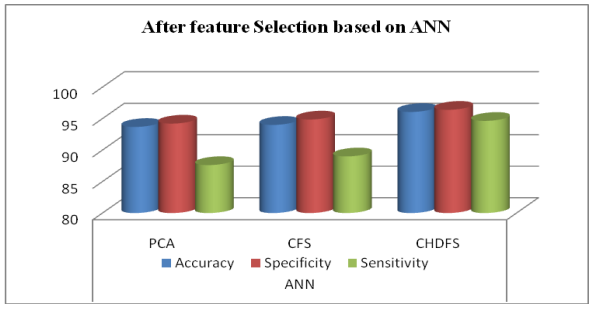

Table 2

Performance Measure of ANN before and after Feature Selection

Classifier/ Metrics	ANN					
	PCA(B)	PCA(A)	CFS(B)	CFS(A)	CHDFS(B)	CHDFS(A)
Accuracy	88.32	93.68	90.45	94.00	**95.68**	**96.06**
Specificity	89.67	94.21	91.23	94.86	**95.21**	**96.40**
Sensitivity	85.54	87.62	87.38	89.00	**90.62**	**94.64**

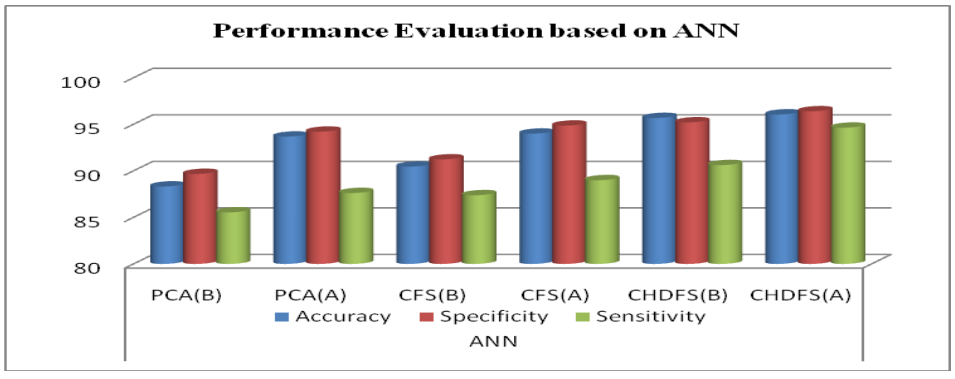

Figure 2: Performance Evaluation based on ANN

Table 2 and Figure 2 reveal that ANN has the highest accuracy, sensitivity, and specificity in the proposed Correlation-based High Distinction Feature Selection (CHDFS).Table 4(a), 5(a) and Figure 4(b), 5(b) displays the outcomes of performance measures such as accuracy, sensitivity and specificity for feature selection based on SVM before and after feature selection.

Table 4(a) SVM before Feature Selection

Classifier/ Metrics	SVM		
	PCA	CFS	CHDFS
Accuracy	85.55	90.45	**91.71**
Specificity	87.43	91.36	**92.59**
Sensitivity	79.87	82.28	**85.26**

Fig. 4(a) SVM before Feature Selection

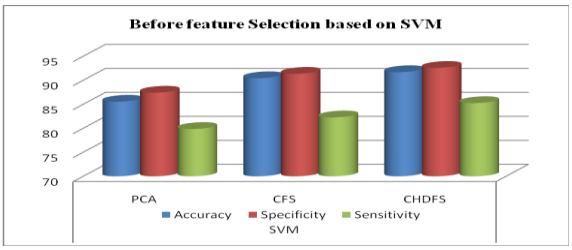

Table 5(a) SVM after Feature Selection

Classifier/ Metrics	SVM		
	PCA	CFS	CHDFS
Accuracy	91.71	93.57	**95.24**
Specificity	92.59	94.02	**95.7**
Sensitivity	85.26	85.80	**92.72**

Fig. 5(b) ANN after Feature selection

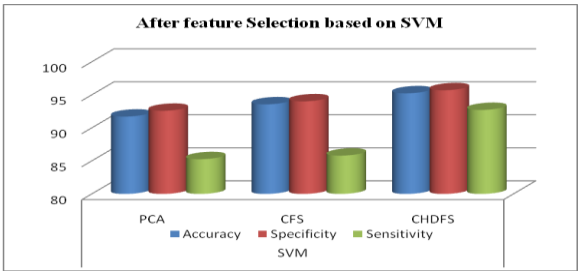

Table 3 Performance measure of SVM before and after feature selection

Classifier/ Metrics	SVM					
	PCA(B)	PCA(A)	CFS(B)	CFS(A)	CHDFS(B)	CHDFS(A)
Accuracy	85.55	91.71	90.45	93.57	**91.71**	**95.24**
Specificity	87.43	92.59	91.36	94.02	**92.59**	**95.7**
Sensitivity	79.87	85.26	82.28	85.80	**85.26**	**92.72**

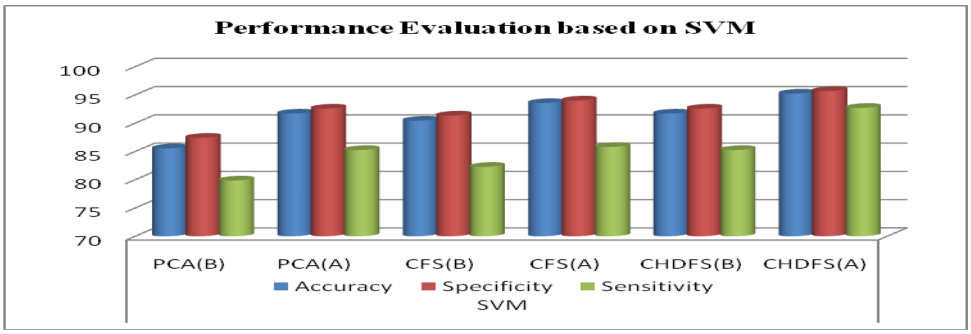

Figure 3: Performance evaluation based on SVM

In comparison to other current strategies, the accuracy of SVM utilizing the suggested feature selection technique (CHDFS) was determined to be the highest at 95.24% in Table 3.

After the execution, the table shows that ANN has the utmost accuracy of 96.06% followed by SVM with accuracy of 95.24%. In the case of before feature selection ANN has the accuracy of 95.68% followed by SVM with accuracy of 91.75%. Likewise, the outcome attained for Mini-mias database is revealed in Table 2 and 3.

V CONCLUSION

Breast cancer detection and diagnosis at an early stage aids to reduce the casualty rate to a higher level. As a result, developing a structured and consistent CAD system capable of reliably classifying mammograms becomes critical. A model CAD system is proposed in this paper. GLCM, a textural feature extraction approach, is used in the described system. With the help of CHDFS, a total of 14 features are extracted, which are subsequently reduced to a smaller feature set. To evaluate the performance measures, the condensed feature set is feed to different classifiers such as SVM and ANN. It has been discovered that ANN outperforms than SVM in the following aspects like sensitivity, accuracy and specificity. Furthermore, it has been discovered that the proposed strategy produces superior outcomes than the existing methods.

To improve classification accuracy, the presented work is able to be extending to the design of alternate feature extraction, feature diminution, and classification methods.

IV REFERENCES

[1] R. Nithya and B. Shanthi, “Comparative study on feature extraction method for breast cancer classification,” Journal of Theoretical and Applied Information Technology, vol. 12, pp. 220-226, 2011.

[2] T. T. Htay and S. S. Maung, "Early Stage Breast Cancer Detection System using GLCM feature extraction and K-Nearest Neighbor (k-NN) on Mammography image," 2018 18th International Symposium on Communications and Information Technologies (ISCIT), 2018, pp. 171-175, doi: 10.1109/ISCIT.2018.8587920.

[3] Chitalia, R. D., & Kontos, D. (2019). Role of texture analysis in breast MRI as a cancer biomarker: A review. Journal of magnetic resonance imaging, 49(4), 927-938.

[4] Shofwatul ‘Uyun , & Lina Choridah. (2018). Feature Selection Mammogram based on Breast Cancer Mining. International Journal of Electrical and Computer Engineering, 8(1), 60-69.

[5] Mani-Varnosfaderani A, Ghaemmaghami M. Assessment of the orthogonality in two-dimensional separation systems using criteria defined by the maximal information coefficient, Journal of Chromatography, 2015, 1415: 108-114.

[6] K.K. Kavitha, A. Kangaiammal,Correlation-based high distinction feature selection in digital mammogram, Materials Today: Proceedings, 2020.

[7] Causi, & David, A. (2002). An Analysis of Co-occurrence Texture Statistics as A Function of Gray Level Quantization. Canadian Journal of remote sensing, 28(1), 45-62.

[8] Haralick, Robert, M., & Karthikeyan Shanmugam. (1973). Textural Features for Image Classification. IEEE Transactions on systems, man, and cybernetics, 3(6), 610-621.

[9] Mohanaiah, P., Sathyanarayana, P., & GuruKumar, L. (2013). Image Texture feature Extraction using GLCM Approach. International Journal of Scientific and Research Publications, 3(5).

[10] Than Than Htay, Su Su Maung, & Khine Thin Zar. (2020). Analysis on Results Comparison of Feature Extraction Methods for Breast Cancer Classification. International Journal of Advances in Scientific Research and Engineering, 6(3).

[11] Deepa, B. G., Senthil, S., Gupta Rahil, M., & Shah Vishakha, R. (2019). Augmentation of Classifier Accuracy through Implication of Feature Selection for Breast Cancer Prediction. International Journal of Recent Technology and Engineering, 8 (2).

[12] Buciu, I., Gacsadi, A.: Directional features for automatic tumor classification of mammogram images. Biomed. Signal Process. Control. 6(4), 370–378 (2011)

[13] Bagchi, M. J., Mohanty, F., Rup, S., Dash, B., & Majhi, B. (2018). Digital Mammogram Classification Using Compound Local Binary Pattern Features with Principal Component Analysis Based Feature Reduction Approach. Advances in Computing and Data Sciences, 270–278.

[14] Daniel F. Schmidt, Enes Makalic, Benjamin Goudey, Gillian S. Dite, et al. Cirrus: An Automated Mammography-Based Measure of Breast Cancer Risk Based on Textural Features, JNCI Cancer Spectrum, 2018, Vol. 2, No. 4.

[15] Albregtsen, F., Nielsen, B., &Danielsen, H. E. (2000). Adaptive gray level run length features from class distance matrices. In Pattern Recognition, 2000. Proceedings. 15th International Conference on (Vol. 3, pp. 738-741). IEEE.

[16] Haralick RM (1979) Statistical and structural approaches to texture. Proc IEEE 67(5):786–804

[17] Kokula Krishna Hari Kunasekaran & Rajkumar Sugumaran. (2016). Exploratory Analysis of Feature Selection Techniques in Medical Image Processing. International Conference on Information Engineering, Management and Security, 33-37.

[18] Shofwatul 'Uyun , & Lina Choridah. (2018). Feature Selection Mammogram based on Breast Cancer Mining. International Journal of Electrical and Computer Engineering, 8(1), 60-69.

[19] Shankar Thawkar & Ranjana Ingolikar. (2017). Automatic Detection and Classification of Masses in Digital Mammograms. International Journal of Intelligent Engineering and Systems, 10(1).

[20] Bhateja V. et al. (2018) Haralick Features-Based Classification of Mammograms Using SVM. In: Bhateja V., Nguyen B., Nguyen N., Satapathy S., Le DN. (eds) Information Systems Design and Intelligent Applications. Advances in Intelligent Systems and Computing, vol 672. Springer, Singapore. https://doi.org/10.1007/978-981-10-7512-4_77

[21] Solanke, A. M., Manjunath. R., & Jadhav, D., V. (2019). Classification of Masses as Malignant or Benign Using Support Vector Machine. Advances and Applications in Mathematical Sciences, 18(9), 901-908.

[22] R. Rasti, M. Teshnehlab, and S. L. Phung,"Breast cancer diagnosis in DCE-MRI using mixture ensemble of convolutional neural networks," Pattern Recognition, vol. 72, pp. 381-390, 2017/12/01/ 2017.

[23] J Suckling et al. (1994) "The Mammographic Image Analysis Society Digital Mammogram Database" Exerpta Medica. International Congress Series 1069, pp. 375–378.