EAI Endorsed Transactions
on Pervasive Health and Technology

Research Article EAI.EU

A Robust Feature Extraction Technique for Breast Cancer Detection using Digital Mammograms based on Advanced GLCM Approach

L Kanya Kumari[1,*], B Naga Jagadesh[2]

[1]Research Scholar, Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur District, Andhra Pradesh, India. kanyabtech@yahoo.com
[2]Professor, Department of Computer Science & Engineering, Srinivasa Institute of Engineering and Technology, Amalapuram, Andhra Pradesh, India. nagajagadesh@gmail.com

Abstract

INTRODUCTION: Breast cancer is the most hazardous disease among women worldwide. A simple, cost-effective, and efficient screening called mammographic imaging is used to find the breast abnormalities to detect breast cancer in the early stages so that the patient's health can be improved.
OBJECTIVES: The main challenge is to extract the features by using a novel technique called Advanced Gray-Level Co-occurrence Matrix (AGLCM) from pre-processed images and to classify the images using machine learning algorithms.
METHODS: To achieve this, we proposed a four-step process: image acquisition, pre-processing, feature extraction, and classification. Initially, a pre-processing technique called Contrast Limited Advanced Histogram Equalization (CLAHE) is used to increase the contrast of images and the features are retrieved using AGLCM which extracts texture, intensity and shape-based features as these are important to identify the abnormality.
RESULTS: In our framework, a classifier called eXtreme Gradient Boosting (XGBoost) is applied on mammograms and the results are compared with other classifiers such as Random Forest (RF), K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN), and Support Vector Machine (SVM). The experiments are done on the Mammographic Image Analysis Society (MIAS) dataset.
CONCLUSION: The outcome achieved with CLAHE+ AGLCM+ XGBoost classifier is better than the existing methods. In future, we experiment on large datasets and also concentrate on optimal features selection to increase the classification.

Keywords: Contrast Limited Adaptive Histogram Equalization, Advanced Gray-Level Co-occurrence Matrix, K-Nearest Neighbor, Artificial Neural Network, Random Forest and eXtreme Gradient Boosting.

Received on 03 September 2021, accepted on 04 January 2022, published on 11 January 2022

doi: 10.4108/eai.11-1-2022.172813

1. Introduction

Breast Cancer (BC) is prominent cancer that occurs in women of 40 years age group [1]. The chances of survival are very remote if it reaches advanced stages. The survival rate in patients of BC is very low in India because of the delay in the detection of tumors. At present, several imaging modalities such as mammography, tomosynthesis, magnetic resonance, and ultra-sonography are used for BC detection. Among these modalities, mammography is the best cost-effective for detecting BC in the early stages [2]. But it is a challenge for radiologists to recognize the masses in the breast using mammogram images [3]. However, it is a tough task to classify mammograms [4]. These examine the breast to provide information like anatomy, morphology, contrast, etc. The survey reveals that radiologists fail in identifying the masses in the early stages [5]. Detection of masses is difficult because they are very pronounced in density, size, shape, similarity to the healthy tissue, and image contrast [6].

In recent years, the detection of disease is becoming a challenging task. Machine Learning (ML) techniques help to find the hidden patterns from a large amount of data

[*] Corresponding author. Email: kanyabtech@yahoo.com

which is otherwise difficult for radiologists or pathologists or physicians. It not only helps in diagnosing but also helps to find the stage of cancer thereby facilitating the doctors to give appropriate drugs to the patients [7]. ML techniques that are very much helpful in cancer detection are Artificial neural Networks (ANN) [8] and Decision Trees (DT) [9]. To predict and prognosis cancer the commonly used ML techniques are K-Nearest Networks (KNN), Bayesian Networks (BN), and Support Vector Machines (SVM) [10]. SVM is used to detect heart disease, breast cancer, ovarian cancer, multiple myeloma, and leukemia [11].

The importance of the proposed methodology are:

a. The mammogram images are pre-processed using CLAHE.

b. We proposed a new feature extraction technique called AGLCM which extracts texture (GLCM), intensity (entropy) and shape-based (Fourier descriptor) features. These features are fed to different classifiers for experimental comparison.

c. The proposed methodology (CLAHE+AGLCM) is experimented by using a classifier XGBoost and compared with other classifiers such as KNN, ANN, SVM, and RF.

d. The performance of these methods is evaluated by confusion matrix parameters and misclassification rate.

e. The results show that CLAHE+AGLCM with XGBoost is superior to previous works done by other authors [12] [13] [16].

The proposed methodology provides better mammogram classification using CLAHE+AGLCM.

The remaining paper is arranged as follows. Section 2 presents the related work and Section 3 discusses the proposed methodology in which the dataset, pre-processing technique CLAHE, feature extraction technique AGLCM, and about XGBoost classifier are described. Section 4 presents the results and at the end, the conclusion and future scope are provided.

2. Related Work

Extensive research was done in the domain of mammogram classification. The majority of the literature used different types of feature extraction techniques like texture, intensity, shape, or combination of these features.

The authors [12], Gray Level Co-occurrence Matrix (GLCM) was used for feature extraction. The feature selection techniques were applied and obtained 94.27% accuracy with a neural network classifier. To extract the features from Digital Database for Screening Mammography (DDSM) dataset, the authors used Gabor features [13]. These features were optimized using PSO and classified using SVM. The accuracy was 93.95% in classifying the images as benign and malignant. In [14], features were extracted by using an intensity histogram and feature radial distance. Enhanced Cuckoo Search (ECS) algorithm was experimented for feature selection and concluded that KNN with ECS achieved 99.13% and the Minimum Distance Classifier with ECS achieved 98.75% accuracy.

Global thresholding was used for pre-processing and classifying the mammogram images [15]. Features were extracted by using laws texture energy. To select the features Particle Swarm Optimized Wavelet neural networks were used. The sensitivity, specificity, and misclassification rates obtained were 94.167%, 92.105%, and 0.063 respectively.

The authors [16] proposed a model to classify mammogram images based on the CLAHE pre-processing technique and Histogram of Oriented Gradients (HOG) to extract features. They obtained a classification accuracy of 66% for the RF classifier.

In [17], the authors extracted GLCM features from the MIAS dataset. These features are passed to a hybrid classifier called KNN with SVM for classification. They achieved 94% accuracy for classifying the images. The authors [18] used the Gaussian filtering pre-processing technique and features were extracted using GLCM and Gray Level Run Length Matrix (GLRLM). They achieved 98% and 97.8% as sensitivity and specificity for feed-forward network classifier.

The authors presented [19] mammogram classification based on spiculation index, fractional concavity and compactness features and achieved 80% accuracy. The masses were detected by extracting GLCM features and obtained Area Under the Curve (AUC) as 0.79 [20]. Local descriptors were played an important role in mammogram classification. The authors [21] classified parenchymal tissue by extracting local descriptors and probabilistic Latent Semantic Analysis (pLSA) and obtained an accuracy of 95.42%.

The breast density classification was done based on the local descriptors such as Square Invariant Feature Transforms (SIFT), Local Binary Patterns (LBP) and texton histograms. The feature vector was classified using SVM and obtained an accuracy of 93% [22]. The authors [23] focused on extracting Speeded-Up Robust Feature (SURF) descriptors for mammogram classification and reported that they obtained 92.3% accuracy.

A CAD system was designed for tumor detection which extracted GLCM features. These features were fed to the SVM classifier to classify the tumors and reported that achieved 92.3% accuracy [24]. To detect breast cancer in early stages, the authors [25] extracted the Hough transform features and classified the features using SVM. They have analysed and concluded that they obtained 94% accuracy for early detection.

From the literature, it can be seen that an effective and efficient Computer Diagnosis System (CAD) is required for BC detection in the early stages. Furthermore, textual features are giving better results in the existing methodologies including pre-processing technique. In the literature, neural networks, SVM, RF classifiers are widely used for classification. Further, the existing methods are based on the MIAS dataset. Based on the above factors, it is required to propose an improved feature extraction technique to increase the overall performance of the CAD framework. So, we proposed a novel method called AGLCM to extract texture, intensity and shape-based features as all these features are very much helpful in the detection of breast cancer. These features are classified as normal or abnormal.

3. Proposed Methodology

The proposed CAD methodology is a 4-step process: image acquisition, pre-processing, feature extraction, and classification. In our methodology, CLAHE is applied as a pre-processing technique which increase the contrast of images so that better features can be extracted.

To extract the features, the AGLCM technique is used. AGLCM is a novel feature extraction technique that extracts texture, intensity and shape-based features of the tumor in mammogram images. These classifies the mammograms into normal or abnormal (0 indicates normal and 1 indicates abnormal). The efficiency of the proposed methodology is carried out by using a confusion matrix and misclassification rate. Figure 1 depicts the framework of proposed methodology.

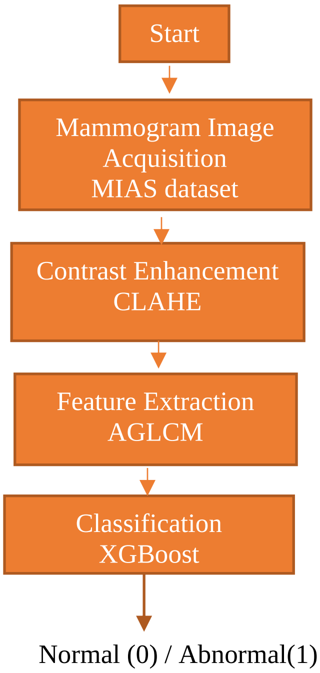


Figure 1. Framework of proposed methodology

3.1. Image Acquisition

Our proposed model is applied on a benchmark dataset called MIAS (Mammogram Image Analysis Society) which comprises of 322 grayscale images in which 112 images are normal and 210 images are abnormal. The images are 8-bit grayscale images with 1024×1024 size [26]. The images are represented in Portable Network Graphics (.PNG) format. The information is available in a separate file that contains character of background tissue, class of abnormalities, abnormality severity, central coordinate of abnormal and radius of circle. The MIAS dataset mammograms are represented in Figure 2.

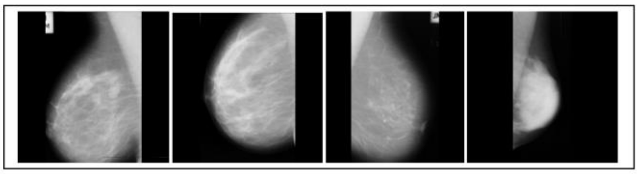


Figure 2. Images from MIAS dataset: mdb019, mdb049, mdb204, mdb262.

3.2. Contrast Limited Advanced Histogram Equalization

Pre-processing is an important phase that enhances some image features which are important for further processing. It plays a vital role in medical imaging which leads to the extraction of better features such as masses and tumors etc. In literature, the authors proposed a variety of contrast enhancement techniques such as Histogram Equalization (HE) [27], Median Filtering [28], filtering with morphological operators and un-sharp masking [29] to improve the visual contents of mammograms [30]. In [31] the authors used Local Contrast Enhancement (LCE) to enhance the contrast in images and achieved better results. In the same manner, we too used a pre-processing technique called CLAHE to enhance the contrast in images as it can overcome the problems in, HE and Adaptive Histogram Equalization (AHE) [27]. The over enhancement in AHE is reduced by using CLAHE[32]. CLAHE is an improvement of AHE where contrast is improved by user-defined clip-level. This method reduces the noise and edge-shadowing generated in consistent locations and is designed for medical imaging [33][34]. It is used to remove artifacts like wedges, labels, and markers in mammograms and it makes suspicious or hidden regions more visible.

In CLAHE, the image is split into small parts called tiles. By applying this technique, each tile contrast is enhanced. To combine the tiles, bipolar interpolation is used to eliminate the artifacts in the borders. The CLAHE steps are explained in algorithm 1. The clip limit is considered as 6.0 and window size considered as 8×8.

```
Algorithm 1  :   CLAHE ( Im, N, NB )
 INPUT        :Im : input image
                  Nr: Number of regions
              Nb : number of bins
OUTPUT    :    Contrast enhanced image

Begin
1. The image Im is divided into equal-sized regions of $8 \times 8$
2.  For each region, a histogram is calculated.
3.  A clip limit is used which is a hyperparameter for altering
the contrast in the image
```

Clip Limit is calculated as $N_{avg} = {(N_{Xaxis} \times N_{Yaxis})}/{N_{gn}}$ Where,

N_{avg} is the average number of pixels

N_{Xaxis}, N_{Yaxis} are the in X-axis and Y-axis.

N_{gn} is the number of gray levels in contextual region

4. The Clip Limit is represented as $N_{ClLm} = N_{ClLm} \times N_{gn}$

Where N_{ClLm} is the actual clip limit

N_{ClLm} is normalized clip limit in between [0,1]

5. Redistribution of the histogram is done like height should not exceed the clip
limit.

6. Redistribution of pixels is given by

$$\text{Re}\,distribution = {N_{gn}}/{N_{remain}}$$

Where, N_{remain} is the remaining number of clipped pixels and the redistribution should
be at least 1.

7. Transformation function for all the histograms as given below.

$$h(r_i) = \sum_{j=0}^{i} P_r(r_j) \quad \text{where } P_r(r_j) = \frac{n_j}{n}$$

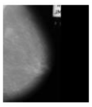
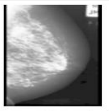
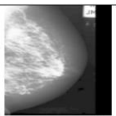
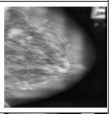
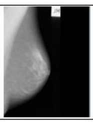
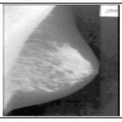
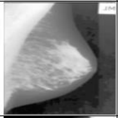
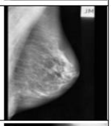
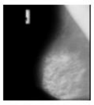
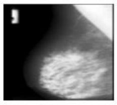
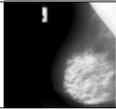
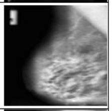
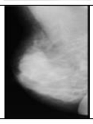
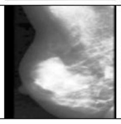
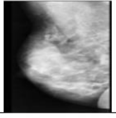
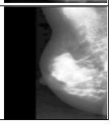
The probability density function of j^{th} n_j grayscale, n is the pixel count in the mammogram, n_j is the pixel number which is given as input of grayscale value j.

8. Bipolar linear interpolation is applied to combine neighboring tiles and the values are modified based on the new histograms in order to eliminate boundary artifacts.

End

The above Algorithm 1 is applied on random images from the MIAS dataset named as mdb304, mdb076, mdb099 and mdb241. The following Table 1 consists of MIAS images and contrast-enhanced (HE, AHE, and CLAHE) images.

Table 1. Contrast enhancement techniques (HE, AHE, CLAHE) applied on MIAS mammogram images

Image number	Original image	HE image	AHE Image	CLAHE image
mdb304				
mdb076				
mdb099				
mdb241				

These contrast-enhanced images are very much helpful to extract better features to detect breast cancer. Hence, these images are given as input to the feature extraction technique called AGLCM.

3.3 Advanced Gray Level Co-Occurrence Matrix

The features are extracted from the pre-processed images using AGLCM technique. The performance of the classifier is depending on how well the feature vector is calculated. The feature extraction technique examines the images to extract the features that signifies the several classes. These features are given as input to the classifier that assigns the class label for test data.

In our proposed methodology, the features are extracted based on the AGLCM technique in which texture, intensity and shape-based approaches are used as these features are crucial in detecting tumors or masses in mammograms [35]. Texture features in an image are the spatial distribution of gray levels whereas shape features describe the lesion boundaries (rounded, spiculate or stellate). Texture features are extracted using Gray Level Co-occurrence Matrix (GLCM), entropy is used to extract intensity-based features [36] and Fourier Descriptor is used for shape-based features. The combination of all these features is named AGLCM.

GLCM is a commonly used technique to extract texture features [37]. This technique results the distribution of gray-level pixel pairs in the image. The spatial distribution between two pixels is computed based on reference pixel and neighbour pixel. In GLCM, the matrix form of Gray values is called the Co-occurrence Matrix (CM). This matrix represents the relative frequencies of the neighboring pixels which are separated by the distance 'd'. The values in the matrix represent the frequency variations in the pixel intensities. The probability occurrences are calculated in 8 different directions 'ɵ' (0^0,45^0, 90^0, 135^0, 180^0, 225^0, 270^0, and 315^0) with distance 'd'. The working procedure of the GLCM technique is represented in Figure 3.

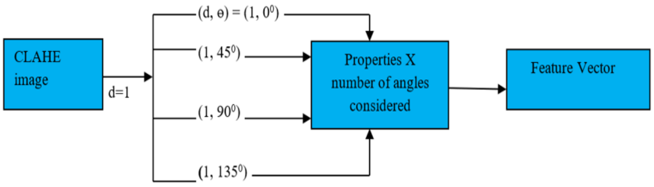


Figure 3. Model of GLCM

From CM, the calculated features are energy or uniformity, homogeneity, correlation, dissimilarity and contrast. The energy feature calculates the textual uniformity whereas the homogeneity feature calculates the variations in an image. The correlation of GLCM indicates the linear dependency of neighboring pixels. The dissimilarity feature measures the distance between pair of pixels and the contrast feature calculates the spatial frequency of an image. These are represented in the following equations 1 to 5. Total 20 GLCM features are considered (5 GLCM measures in 4 directions).

Energy = $\sum_{u,v}^{n} p^{2}_{uv}$ where p_{uv} = element x, y (2 samples of intensities) (1)

$$\text{Homogeneity} = \sum_{u,v=1}^{n} \frac{P_{uv}}{1+(u-v)^2} \qquad (2)$$

Correlation $= \sum_{u,v=1}^{n} \frac{(u-m)(v-m)}{s^2}$ where m= $\sum_{u,v=1}^{n} xP_{uv}$ and $s^2 = \sum_{u,v=1}^{n} P_{uv}(1-m)^2$ (3)

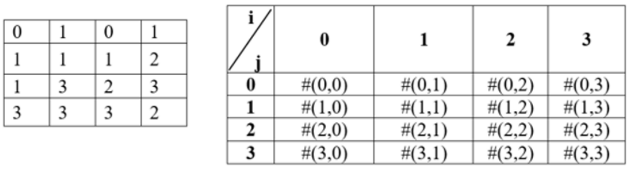
$$\text{Dissimilarity} = \sum_{u}\sum_{v}|u-v|p(u,v) \qquad (4)$$

$$\text{Contrast} = \sum_{u,v=1}^{n} p_{uv}(u-v)^2 \qquad (5)$$

Entropy gives information about the contents of the image. It gives the image uncertainty or randomness or intensity levels [38]. This is calculated as in equation 6 and it is added to the GLCM feature vector as another feature.

$$\text{Entropy} = \sum_{u,v=0}^{n-1} -\ln(p_{uv})p_{uv} \qquad (6)$$

The calculation of the Co-occurrence Matrix (CM) is described with the below example. Suppose the image consists of 4 possible values 0, 1, 2, and 3. So, CM is a 4×4 matrix and with the assumption of distance d=1. So, N=4 is considered and ɵ= 0⁰,45⁰, 90⁰ and 135⁰. The GLCM in 4 angles is mentioned in Figure 4.



0	1	0	1
1	1	1	2
1	3	2	3
3	3	3	2

i / j	0	1	2	3
0	#(0,0)	#(0,1)	#(0,2)	#(0,3)
1	#(1,0)	#(1,1)	#(1,2)	#(1,3)
2	#(2,0)	#(2,1)	#(2,2)	#(2,3)
3	#(3,0)	#(3,1)	#(3,2)	#(3,3)

Figure 4.a. The sample image matrix **b.** Number of occurences of pixel i to the neighboring pixel j

The images from the MIAS dataset are considered where some images are normal and some abnormal. GLCM features' performance is efficient in finding breast cancer [39]. In our framework, texture, intensity-based features are combined with shape-based features as it is significant to know the intensity and shape of the tumor is important including the texture in finding the abnormality in mammograms. Depending on the tumor shape, it is easy to identify whether the tumor is normal or abnormal. Several shape descriptors are available in the literature to identify the shape of the tumor. Among them, Fourier Descriptors (FD's) are extremely useful for pattern recognition [40] and are used to recognize the shape of the tumor in mammograms. An FD is based on Fourier transformed boundary as the shape feature. These features are effective in representing the shape and it is invariant to rotation, translation, and scaling [14][41].

FD's are derived from Fourier transformations in which larger frequency values represent the fine details and smaller frequency values represent the global shape. Few FD's are used to capture the essence of a boundary. This carries shape information. So, these are used for differentiating distinct boundary shapes. A digital boundary is represented as a complex number. Starting at any point of the boundary denoted by (a_0, b_0) as coordinate pairs and moving clockwise direction. The x-axis is the real number axis and y is the imaginary number axis of a complex number. This complex number has a major advantage as it reduces a 2-dimensional problem to a 1-dimensional problem. This complex co-efficient is nothing but FD [42]. A discrete Fourier is defined as

$$F_n = \frac{1}{N}\sum_{i=0}^{N-1} s(i)\times e^{\left(\frac{-j2\pi ni}{N}\right)} \qquad (7)$$

where,

F_n =nth Fourier descriptor

$S(i)$ = 1- Dimensional contour signal

N= total points of the contour

$i = 0,1,2...N-1$.

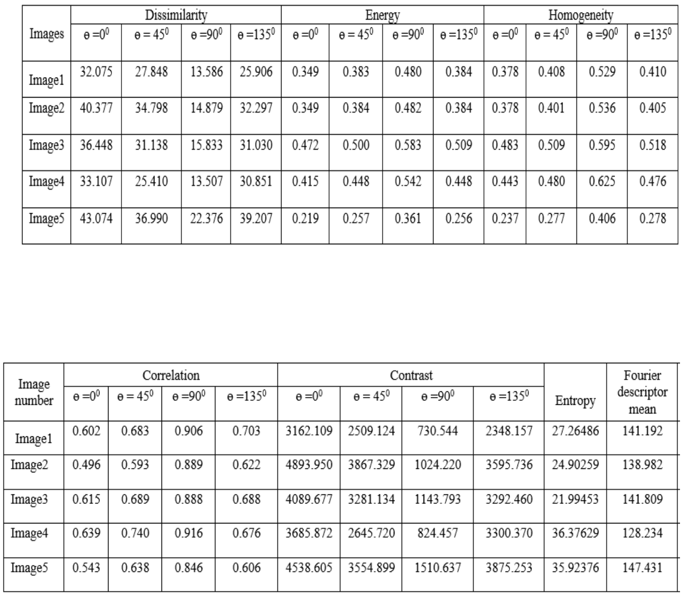
By using the above equation (7), FD's of size 'N' are calculated which are invariant to translation. In our proposed methodology the GLCM features, entropy and the mean value of the Fourier co-efficient are combined to obtain the feature vector. The step-by-step process of the AGLCM algorithm is described in Algorithm 2.

```
Algorithm 2 : AdvancedGLCM (Im, Ds, angles )
INPUT      :Im : Contrast Enhanced Image
                 Ds: Distance
                  angles=[$0^0$, $45^0$,$90^0$,$135^0$]
OUTPUT  :      Feature Vector

Begin
1. The contrast-enhanced images is given as input
2. Find the co-occurrence matrix with Ds=1 and ɵ=$0^0$
   $45^0$,$90^0$, $135^0$.
3. Make GLCM symmetric
4. Normalize the GLCM
5. Calculate GLCM features: energy, homogeneity,
   correlation, dissimilarity, contrast in 4 different angles
   with Ds=1
6. Find entropy from GLCM
7. Calculate the Fourier co efficient
8. Calculate the mean of Fourier coefficients
9. Combine GLCM features, entropy and statistical Fourier
   coefficients to get the final feature vector
10. Return feature vector.
End
```

The above Algorithm 2 is applied on contrast-enhanced images and features are extracted. The AGLCM feature vector for MIAS sample images is depicted in Table 2.

Table 2. AGLCM features for MIAS images



Images	Dissimilarity				Energy				Homogeneity			
	ɵ =0^0	ɵ = 45^0	ɵ =90^0	ɵ =135^0	ɵ =0^0	ɵ = 45^0	ɵ =90^0	ɵ =135^0	ɵ =0^0	ɵ = 45^0	ɵ =90^0	ɵ =135^0
Image1	32.075	27.848	13.586	25.906	0.349	0.383	0.480	0.384	0.378	0.408	0.529	0.410
Image2	40.377	34.798	14.879	32.297	0.349	0.384	0.482	0.384	0.378	0.401	0.536	0.405
Image3	36.448	31.138	15.833	31.030	0.472	0.500	0.583	0.509	0.483	0.509	0.595	0.518
Image4	33.107	25.410	13.507	30.851	0.415	0.448	0.542	0.448	0.443	0.480	0.625	0.476
Image5	43.074	36.990	22.376	39.207	0.219	0.257	0.361	0.256	0.237	0.277	0.406	0.278

Image number	Correlation				Contrast				Entropy	Fourier descriptor mean
	ɵ =0^0	ɵ = 45^0	ɵ =90^0	ɵ =135^0	ɵ =0^0	ɵ = 45^0	ɵ =90^0	ɵ =135^0		
Image1	0.602	0.683	0.906	0.703	3162.109	2509.124	730.544	2348.157	27.26486	141.192
Image2	0.496	0.593	0.889	0.622	4893.950	3867.329	1024.220	3595.736	24.90259	138.982
Image3	0.615	0.689	0.888	0.688	4089.677	3281.134	1143.793	3292.460	21.99453	141.809
Image4	0.639	0.740	0.916	0.676	3685.872	2645.720	824.457	3300.370	36.37629	128.234
Image5	0.543	0.638	0.846	0.606	4538.605	3554.899	1510.637	3875.253	35.92376	147.431

3.4. Classification - Extreme Gradient Boosting Classifier

Classification is the process of finding a class label based on existing data. A classifier is constructed on training data in which the class label is already known. Based on this information, the classifier learns the properties of each subset and finds the labels for test data. in our framework, we have two classes as normal and abnormal. The extracted AGLCM features are classified as normal or abnormal by using several classification algorithms namely KNN, ANN, SVM, RF and XGBoost. Among these, XGBoost is efficient in automatic parallel computation and gives good results for most of the datasets. The XGBoost classifier is implanted by using gradient boosting which has special characteristics like the more regularized model to control overfitting, which results in a better way [43]. XGBoost is an optimized combination of hardware and software by using fewer computing resources in less amount of time [44]. This model is a good combination in terms of prediction, performance, and processing speed compared to other algorithms. It is more effective than deep learning techniques if a limited number of training samples are available [6].

4. Results and Discussion

In our methodology, pre-processing is an important phase in extracting better features. The pre-processing techniques (HE, AHE, CLAHE) are applied on the MIAS dataset images discussed in section 3.2. The performance of these pre-processing techniques is measured by Mean Square Error (MSE) and Peak Signal to Noise Ratio (PSNR) [45][46][47]. MSE is the most used form of measuring image quality which is the Mean Squared Error between the actual and pre-processed image. PSNR is the image quality measurement between the actual image and the processed image. Mathematically, they are represented as in equations 8 and 9.

$$MSE\left(O,P\right)=\frac{1}{MN}\sum_{u=1}^{M}\sum_{v=1}^{N}\left(O\left(u,v\right)-P\left(u,v\right)\right)^{2} \qquad (8)$$

Where,

$O\left(u,v\right)$ is the original mammogram image

$P\left(u,v\right)$ is the pre-processed image

M, N are image dimensions

$$PSNR=10\log_{10}\frac{P^{2}}{MSE(O,P)} \qquad (9)$$

where,

P is maximum possible intensity value

MSE (O, C) Mean Square Error

The MSE and PSNR values are calculated in between the original image and pre-processed images using equations 8 and 9. The small value of PSNR represents poor quality whereas a greater value indicates a good-quality image. The PSNR value ranges for HE, AHE, and CLAHE pre-processing techniques are 20db to 30db, 31db to 35db and 36db to 45db respectively and are represented in Table 3.

By analysing Table 3, it can be inferred that PSNR values of CLAHE are greater than other pre-processing techniques. Hence, the image quality is improved by using the CLAHE technique and also, it is understood that pre-processing is an important phase in mammogram classification.

Table 3. MSE and PSNR values of HE, AHE and CLAHE techniques.

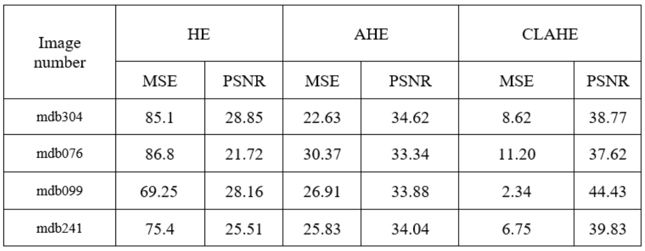
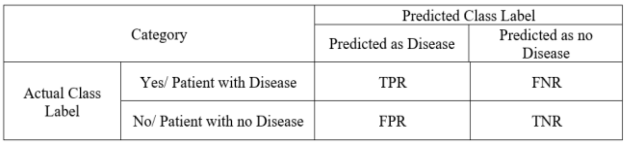


Image number	HE		AHE		CLAHE	
	MSE	PSNR	MSE	PSNR	MSE	PSNR
mdb304	85.1	28.85	22.63	34.62	8.62	38.77
mdb076	86.8	21.72	30.37	33.34	11.20	37.62
mdb099	69.25	28.16	26.91	33.88	2.34	44.43
mdb241	75.4	25.51	25.83	34.04	6.75	39.83

The contrast-enhanced images are fed to the AGLCM feature extraction algorithm which is discussed in section 3.3. The MIAS mammograms consists of abnormality in the middle, Region of Interest (ROI) is considered as 256×256 for feature extraction [48]. The feature extraction algorithm extracts texture, intensity and shape-based features from pre-processed images. A total of 22 features are extracted from the MIAS dataset and are represented in Table 2 for sample images. This table consists of the image as a row and features are in columns with a class label. Among these 22 features, the first 20 features are GLCM texture-based features namely dissimilarity, energy, homogeneity, correlation, contrast (five properties in four different angles 0^0,45^0, 90^0 and 135^0 and distance d=1), GLCM intensity-based entropy, and finally shape-based feature called mean of Fourier descriptor are represented. The extracted features are classified using the XGBoost algorithm and the results are compared with other classifiers like KNN, ANN, SVM and RF. The dataset is divided into 70%, 30% training and testing respectively.

The effectiveness of the methodology is validated using several performance measures such as sensitivity, specificity, precision, f1-score, accuracy, and misclassification rate and are calculated using the equations 10 to 14. They are calculated based on the confusion matrix. Diagrammatically the confusion matrix is represented in Table 4 [49].

Table 4. Confusion Matrix



Category		Predicted Class Label	
		Predicted as Disease	Predicted as no Disease
Actual Class Label	Yes/ Patient with Disease	TPR	FNR
	No/ Patient with no Disease	FPR	TNR

where: TPR is True Positive Rate which represents unhealthy persons are diagnosed correctly

FPR is False Positive Rate which represents healthy persons are incorrectly diagnosed

TNR is True Negative Rate which represents healthy persons are diagnosed correctly

FNR is False Negative Rate which represents sick persons are diagnosed incorrectly.

$$Sensitivity = \frac{TPR}{TPR + FNR} \tag{10}$$

$$Specificity = \frac{TNR}{FPR + TNR} \tag{11}$$

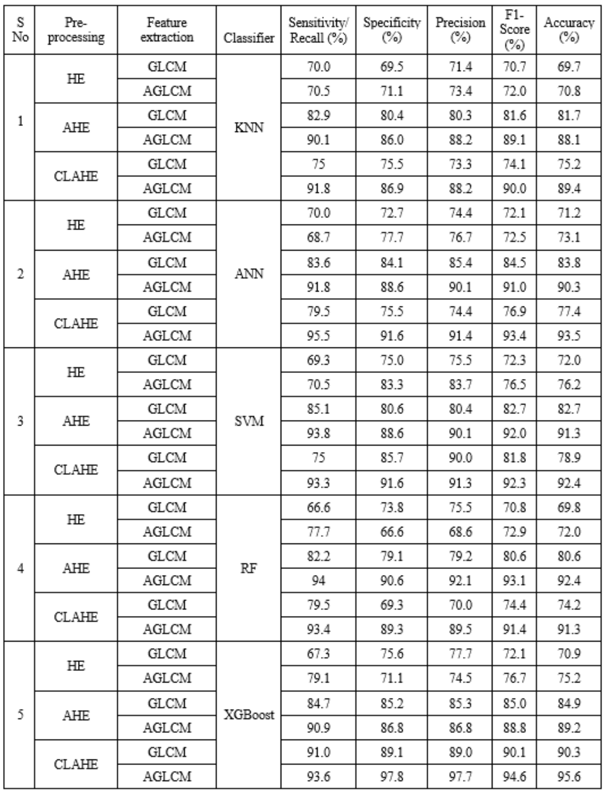
$$\Pr ecision = \frac{TPR}{TPR + FPR} \tag{12}$$

$$f1 - score = \frac{2 \times Precision \times Sensitivity}{\Pr ecision + Sensitivity} \tag{13}$$

$$Accuracy = \frac{TPR + TNR}{TPR + FPR + TNR + FNR} \tag{14}$$

The measure of true positives rate is Sensitivity whereas the measure of true negatives rate is Specificity. Precision is a measure of correct positive rates, and f1-score is a harmonic mean of precision and sensitivity. A publicly available dataset called MIAS is used for experimentation.

Table 5. Comparison results of various pre-processing and feature extraction techniques using different classifiers.



S No	Pre-processing	Feature extraction	Classifier	Sensitivity/ Recall (%)	Specificity (%)	Precision (%)	F1-Score (%)	Accuracy (%)
1	HE	GLCM	KNN	70.0	69.5	71.4	70.7	69.7
		AGLCM		70.5	71.1	73.4	72.0	70.8
	AHE	GLCM		82.9	80.4	80.3	81.6	81.7
		AGLCM		90.1	86.0	88.2	89.1	88.1
	CLAHE	GLCM		75	75.5	73.3	74.1	75.2
		AGLCM		91.8	86.9	88.2	90.0	89.4
2	HE	GLCM	ANN	70.0	72.7	74.4	72.1	71.2
		AGLCM		68.7	77.7	76.7	72.5	73.1
	AHE	GLCM		83.6	84.1	85.4	84.5	83.8
		AGLCM		91.8	88.6	90.1	91.0	90.3
	CLAHE	GLCM		79.5	75.5	74.4	76.9	77.4
		AGLCM		95.5	91.6	91.4	93.4	93.5
3	HE	GLCM	SVM	69.3	75.0	75.5	72.3	72.0
		AGLCM		70.5	83.3	83.7	76.5	76.2
	AHE	GLCM		85.1	80.6	80.4	82.7	82.7
		AGLCM		93.8	88.6	90.1	92.0	91.3
	CLAHE	GLCM		75	85.7	90.0	81.8	78.9
		AGLCM		93.3	91.6	91.3	92.3	92.4
4	HE	GLCM	RF	66.6	73.8	75.5	70.8	69.8
		AGLCM		77.7	66.6	68.6	72.9	72.0
	AHE	GLCM		82.2	79.1	79.2	80.6	80.6
		AGLCM		94	90.6	92.1	93.1	92.4
	CLAHE	GLCM		79.5	69.3	70.0	74.4	74.2
		AGLCM		93.4	89.3	89.5	91.4	91.3
5	HE	GLCM	XGBoost	67.3	75.6	77.7	72.1	70.9
		AGLCM		79.1	71.1	74.5	76.7	75.2
	AHE	GLCM		84.7	85.2	85.3	85.0	84.9
		AGLCM		90.9	86.8	86.8	88.8	89.2
	CLAHE	GLCM		91.0	89.1	89.0	90.1	90.3
		AGLCM		93.6	97.8	97.7	94.6	95.6

We have done different experiments on MIAS dataset images. They are 1. Applying pre-processing techniques namely HE with GLCM and AGLCM 2. Applying AHE as pre-processing with GLCM and AGLCM as feature extraction techniques and finally 3. CLAHE with GLCM and AGLCM. These experiments are done by using the classifiers such as KNN, ANN, SVM, RF and XGBoost. The results of these experiments are shown in Table 5.

It is observed that mammogram image classification accuracies are improved by using CLAHE as a pre-processing technique, AGLCM as a feature extraction technique with XGBoost classifier. This work has proved that pre-processing with a good feature extraction technique positively impacts the accuracy of classifiers. We have used another good performance measure called misclassification rate to evaluate the proposed feature extraction technique called AGLCM. It is the ratio of FPR+FNR to a total which is given in equation 15.

$$\text{Misclassification Rate} = \frac{FPR+FNR}{TPR+FPR+TNR+FNR} \quad (15)$$

The misclassification rates are represented for HE, AHE, CLAHE as pre-processing techniques and GLCM, AGLCM as feature extraction techniques. These are calculated for the XGBoost classifier and compared with other classifiers such as KNN, ANN, SVM, and RF. The graphical representation is given in Figure 5.

It is noticed that less misclassification rate and highest accuracy are obtained for the proposed methodology and it is observed that CLAHE+ AGLCM+ XGBoost classifier is better than the state-of-the-art methods as signified in the following Table 6.

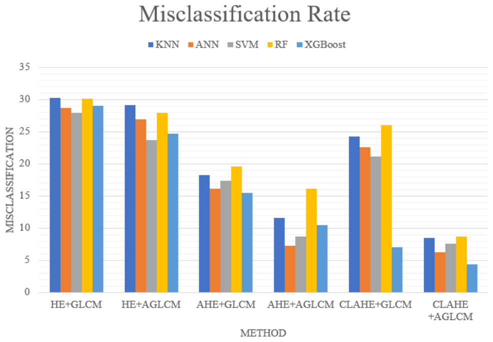


Figure 5. Misclassification rate of proposed methods

Table 6. Results comparison with state-of-art methods

Reference number	Methodology	Accuracy (%)
[12]	GLCM	94.2
[13]	Gabor features	93.9
[16]	CLAHE+HOG	66
[19]	Spiculation index, fractional concavity and compactness	80%
[20]	GLCM	AUC=0.79
[21]	Local descriptors, pLSA	95.42
[22]	SIFT, LBP and texton histograms	93
[23]	SURF	92.3
[24]	GLCM	81
[25]	Hough transform	94
Proposed method	CLAHE+AGLCM	95.6

4. Conclusion

In our research, an improved diagnostic system is designed to encounter the challenges in breast cancer detection. In the proposed framework, firstly, a pre-processing technique called CLAHE is applied to increase the contrast in mammograms. It is followed by a feature extraction technique called GLCM, the combines texture, intensity and shape-based features. These features are classified using the XGBoost technique and the results are compared with KNN, ANN, SVM, and RF classifiers. The experiments are done by using a dataset called MIAS. The performance of the proposed methodology reflects that our framework achieves better performance concerning confusion matrix parameters including misclassification rate. The better accuracy and less misclassification rate are obtained for CLAHE+ AGLCM with XGBoost classifier. Designing a CAD system for BC detection remains a research problem. There are some directions that might improve our research in the future. They are 1. The method is to be applied on large databases 2. The optimal features are to be selected from the extracted features for better classification.

References

[1] Meenakshi M. Pawara, Sanjay N. Talbar. Genetic Fuzzy System (GFS) based wavelet co-occurrence feature selection in mammogram classification for breast cancer diagnosis. Perspectives in Science. 2016; 8:247-250.

[2] Daniel O. Tambasco Bruno, Marcelo Z. do Nascimento, Rodrigo P. Ramos, Valerio R. Batista, Leandro A. Neves, Alessandro S. Martins. LBP operators on curvelet coefficients as an algorithm to describe texture in breast cancer tissues. Expert Systems with Applications. 2016;55: 329-340.

[3] Yu-Dong Zhang, Shui-HuaWang, Ge Liu , Jiquan Yang. Computer-aided diagnosis of abnormal breasts in mammogram images by weighted-type fractional Fourier transform. Advances in medical Engineering. 2016; 8(2).

[4] Shankar Thawkar and Ranjana Ingolikar. Classification of Masses in Digital Mammograms Using the Genetic Ensemble Method. Journal of Intelligent Systems. 2018;29(1).

[5] Figlu Mohanty, Suvendu Rup, Bodhisattva Dash, Banshidhar Majhi, MNS Swamy. Mammogram classification using contourlet features with forest optimization-based feature selection approach. Multimed Tools Appl. 2019; 78: 2805–12834.

[6] R. Song, T. Li and Y. Wang. Mammographic Classification Based on XGBoost and DCNN With Multi Features. IEEE Access. 2020;8:75011-75021.

[7] Machine learning in cancer diagnostics. EBioMedicine. 2019; 45:1–2.

[8] M. Pérez, M. E. Benalcázar, E. Tusa, W. Rivas and A. Conci. Mammogram classification using back-propagation neural networks and texture feature descriptors. In: IEEE Second Ecuador Technical Chapters Meeting; 16-20 October; IEEE; 2017. p. 1-6.

[9] Aswini Mohanty et al. Texture-based features for classification of mammograms using decision tree. Neural Computing and Applications. 2013; 23: 3-4.

[10] T. Hossain, F. S. Shishir, M. Ashraf, M. A. AI Nasim, F. Muhammad Shah. Brain Tumor Detection Using Convolutional Neural Network. In:1st International Conference on Advances in Science Engineering and Robotics Technology; 19 December; IEEE; 2019. p.1-6.

[11] L Kanya kumari, B N Jagadesh. A Review on Big Data Analytics in Multiple Levels of Health Informatics. International Journal of Sciences and Research. 2017;73(6):172-184.

[12] Hajar Alharbi, Gregory Falzon. A Novel feature reduction framework for digital mammogram image classification. In: IAPR Asian Conference on Pattern Recognition; 3-6 November; Kuala Lumpur, Malaysia. IEEE; 2015. p. 221-225.

[13] SalabatKhana, Muhammad Hussainb, Hatim Aboalsamhb, Hassan Mathkourb. Optimized Gabor features for mass classification in mammography. Applied Soft Computing. 2016;44: 267-280.

[14] M. N. Sudha, S. Selvarajan. Feature Selection Based on Enhanced Cuckoo Search for Breast Cancer classification in mammogram image. Circuits and Systems. 2016; 7:327-338.

[15] J. DheebaN. Albert Singh S. Tamil Selvi. Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach. Journal of Biomedical Informatics. 2014; 49: 45-52.

[16] B. Bektaş, İ. E. Emre, E. Kartal and S. Gulsecen. Classification of Mammography Images by Machine Learning Techniques. In: 3rd International Conference on Computer Science and Engineering; 20-23 September; Sarajevo, Bosina and Herzegovina. IEEE;2018 p. 580-585.

[17] Sonar, U, Bhosle and C Choudhury. Mammography classification using modified hybrid SVM-KNN. In: International Conference on Signal Processing and Communication; 28-29 July; Coimbatore, India. IEEE; 2017. p. 305-311.

[18] S. PunithaA. AmuthanK. Suresh Joseph. Benign and malignant breast cancer segmentation using optimized region growing technique. Future computing and informatics journal. 2018;3: 48-358.

[19] Rangayyan, R.M, Mudigonda, N.R, Desautels, J.L. Boundary modelling and shape analysis methods for classification of mammographic masses. Med. Biol. Eng. Comput 2000;38: 487–496.

[20] Mudigonda, N.R, Rangayyan, R.M, Desautels, J.L. Detection of breast masses in mammograms by density slicing and texture flow-field analysis. IEEE Trans. Med. Imaging.2001; 20:1215–1227.

[21] Bosch, A, Munoz, X, Oliver, A, Marti J. Modeling and Classifying Breast Tissue Density in Mammograms. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 17-22 June; New York, USA.IEEE; 2006. p.1552–1558.

[22] Liasis, G, Pattichis C, Petroudi S. Combination of different texture features for mammographic breast density classification. In: International Conference on Bioinformatics Bioengineering; 11-13 November; Larnaca, Cyprus. IEEE; 2012.p.732–737.

[23] Deshmukh, J, Bhosle, U. SURF features-based classifiers for mammogram classification. In: International Conference on Wireless Communications Signal Processing and Networking; 22-24 March; Chennai, India. IEEE; 2017.p.134–139.

[24] Ancy, C,Nair, L.S. An efficient CAD for detection of tumour in mammograms using SVM. In: International Conference on Communication and Signal Processing; 22-24 March; Chennai India. IEEE; 2017.p.1431–1435.

[25] Vijayarajeswari, R, Parthasarathy P, Vivekanandan S, Basha A.A. Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform. Measurement 2019;146:800–805.

[26] Suckling J, Parker J, Dance D, Astley S, Hutt I, Boggis C, Ricketts I, et al. Mammographic Image Analysis Society (MIAS) database. 2015.

[27] K Akila L S, Jayashree A, Vasuki. Mammographic Image Enhancement Using Indirect Contrast Enhancement Techniques – A Comparative Study. Procedia Computer Science. 2015; 47:255-261.

[28] Mohanty A K, Senapati M R, Lenka S. K. A novel image mining technique for classification of mammograms using hybrid feature selection. Neural Computing and Applications. 2013; 22(6):1151–1161.

[29] Nijad A, Najdawi, Mariam Biltawi, Sara Tedmori. Mammogram image visual enhancement, mass segmentation and classification. Applied Soft Computing. 2015;35: 175-185.

[30] Meenakshi Pawar, Sanjay Talbar. Local entropy maximization-based image fusion for contrast enhancement of mammogram. Journal of King Saud University - Computer and Information Sciences. 2021;33(2):50-160.

[31] Narayanan BN, Hardie RC, Krishnaraja V, Karam C, Davuluru VSP. Transfer-to-Transfer Learning Approach for Computer Aided Detection of COVID-19 in Chest Radiographs. AI. 2020;1(4):539-557.

[32] Jayandhi G, Jasmine J L, Joans S M. Mammogram Learning System for Breast Cancer Diagnosis Using Deep Learning SVM. Computer Systems Science and Engineering. 2022; 40(2):491–503.

[33] Maitra IK, Nag S, Bandyopadhyay SK. Technique for preprocessing of digital mammogram. Comput Methods Programs Biomed. 2012;107(2).

[34] Kalyani, G., Janakiramaiah, B., Karuna, A. et al. Diabetic retinopathy detection and classification using capsule networks. Complex Intell. Syst. 2021.

[35] R. Parekh. Using Texture Analysis for Medical Diagnosis. IEEE MultiMedia. 2012;19: 28-37.

[36] Sidney M. L de Lima, Abel G. da Silva-Filho, Wellington Pinheiro dos Santos. Detection and classification of masses in mammographic images in a multi kernel approach. Computer Methods and Programs in Biomedicine. 2015;134(C).

[37] B N Jagadesh, L Kanya kumari. A GLCM based Feature Extraction in Mammogram Images using Machine Learning Algorithms. International Journal of Current Research and Review. 2021;13(5):145-149.

[38] R. Parekh. Using Texture Analysis for Medical Diagnosis. IEEE Multi Media. 2012;19(2):28-37.

[39] Julio César Mello Román 1, José Luis Vázquez Noguera, Horacio Legal-Ayala, Diego P. Pinto-Roa, Santiago Gomez-Guerrero, and Miguel García Torres. Entropy and Contrast Enhancement of Infrared Thermal Images Using the Multiscale Top-Hat Transform. Entropy. 2019;21.

[40] Sathish D et al. Medical imaging techniques and computer aided diagnostic approaches for the detection of breast cancer with an emphasis on thermography – a review. Int. J. Medical Engineering and Informatics. 2016;8(3):275–299.

[41] Frejlichowski D., Gościewska K. Application of 2D Fourier Descriptors and Similarity Measures to the General Shape Analysis Problem. In: International Conference on Computer Vision and Graphics; 24-26 September; Poland: Springer Link; 2012.p.371-378.

[42] https://www.sciencedirect.com/topics/engineering/fourier-descriptor. (accessed on 15 August 2020).

[43] El-ghazal, Akrem, Basir, Otman, Belkasim, S. A New Shape Signature for Fourier Descriptors. In: International Conference on Image Processing; 16-19 September; San Antonio, USA: IEEE; 2007.p.161-164.

[44] https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm/(accessed on 5 September 2020).

[45] https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/ (accessed on 9 September 2020).

[46] Ramani, R. & Vanitha, Suthanthira&Valarmathy, S. The Pre-Processing Techniques for Breast Cancer Detection in Mammography Images. International Journal of Image. Graphics and Signal Processing. 2013;5: 47-54.

[47] Bhupendra, Gupta & Tiwari, Mayank. A tool supported approach for brightness preserving contrast enhancement and mass segmentation of mammogram images using histogram modified grey relational analysis. Multidimensional Systems and Signal Processing. 2017; 28.

[48] M Velmurugan, K. Thangavel. A Novel Wavelet Fractal Feature Extraction Method For Mammogram Image Classification. International Journal of Grid and Distributed Computing 2020;13(2).

[49] G. Kalyani, B. Janakiramaiah. Trends in Deep Learning Methodologies Algorithms, Applications and Systems. 1st Edition. Academic Press; 2021.11, Deep learning-based detection and classification of adenocarcinoma cell nuclei; p. [241-264].