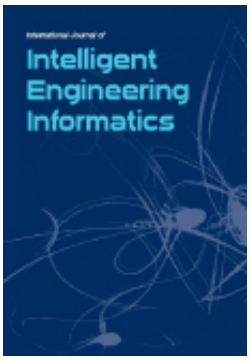


International Journal of Intelligent Engineering Informatics

ISSN online: 1758-8723 - ISSN print: 1758-8715
https://www.inderscience.com/ijiei

Efficient feature extraction on mammogram images using enhanced grey level co-occurrence matrix

S. Surya, A. Muthukumaravel

DOI: 10.1504/IJIEI.2023.10054613

Article History:

Received:	23 August 2022
Last revised:	30 January 2023
Accepted:	30 January 2023
Published online:	03 May 2023

Efficient feature extraction on mammogram images using enhanced grey level co-occurrence matrix

S. Surya* and A. Muthukumaravel

Department of Computer Applications,
Bharath Institute of Higher Education and Research,
Tamil Nadu-600126, India
Email: surya.mca@bharathuniv.ac.in
Email: dean.arts@bharathuniv.ac.in
*Corresponding author

Abstract: Breast cancer is among the most persistent malignant growths that can affect women, and it is now one of the leading causes of mortality. Mammography is a useful screening procedure for breast tumours, although it is difficult to identify and classify them. Textural (or shape-based) factors increased false positive and false negative rates in earlier research. This study proposes a multidimensional feature-based breast cancer identification technique from mammography pictures. Enhanced grey-level co-occurrence matrix (EGLCM) uses machine learning to classify mammography images and construct an accurate feature extraction approach. Contrast limited advanced histogram equalisation (CLAHE) pre-processing improves image contrast. Then, the suggested technique extracts feature from the region of interest (ROI). Since texture, intensity, and shape are needed to detect abnormalities, the EGLCM method captures them. K-nearest neighbour classifies the features (KNN). The given feature extraction methodology yielded accuracy rates of 92%, specificity rates of 90%, and sensitivity rates of 84% on the MIAS dataset, outperforming LBP and GLRLM methods. Python was utilised for evaluation.

Keywords: breast cancer; mammography; CAD system; feature extraction; GLCM; enhanced grey-level co-occurrence matrix; EGLCM; K-nearest neighbour; KNN.

Reference to this paper should be made as follows: Surya, S. and Muthukumaravel, A. (2023) 'Efficient feature extraction on mammogram images using enhanced grey level co-occurrence matrix', *Int. J. Intelligent Engineering Informatics*, Vol. 11, No. 1, pp.35–53.

Biographical notes: S. Surya is a PhD Research Scholar in Computer Applications from Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India. Her area of research specialises in digital image processing and the theory of artificial neural networks. She has attended more than 15+ webinars, workshops, FDP, and conferences at various levels. She has published more than five papers in various refereed international journals. She graduated with Gold Medalist in MCA and MPhil in Computer Applications from Vels University Chennai, India. She has worked as a Junior Programmer in the Asp.Net Platform for 1.5 years, as Guest Lectures for six months, and has 7+ years of teaching experience.

A. Muthukumaravel is a Professor and the Head at the Bharath Institute of Higher Education and Research. He has a Doctorate in Computer Science from Vels University, Chennai (2009–2012). His total teaching experience is 15 years, and he has five years of PhD research work at Vels University, Chennai. And his total number of international and national publications is 97, number of conferences, workshops, seminars, and FDP attended is 50. He has good administrative and interpersonal skills. He has more than 15 years of experience guiding Postgraduate students to carry out their projects, and his areas of specialisation are artificial neural networks and MATLAB.

1 Introduction

In a woman's lifetime, one in eight will develop breast cancer, one of the most serious and deadly diseases. Breast cancer diagnosis and survival are greatly improved by early detection. Breast cancer screening and early diagnosis are the two main methods for early detection (Bulletins, 2017). Early diagnosis involves three interconnected steps awareness, clinical assessment, and treatment, and must be given as soon as possible (Arabaci and Mohamed, 2020). Screening refers to the initial examination for diagnosis and therapy (Tao et al., 2020). It is used to detect irregularities of a particular form of cancer that have not yet manifested with any other symptoms (Virk and Maini, 2020). Iranmakani et al. (2020) proposed mammography as one of the most reliable and practical methods when considering all the diagnostic procedures that are accessible for finding breast cancer (El-Gamal et al., 2020). Digital and film mammography are the two forms of mammography that are now accessible (Puthige et al., 2021). When comparing the two types of mammography, digital is superior to film since the radiation output is reduced by up to 50% while maintaining the ability to detect cancer cells (Elkabbash et al., 2021).

Giess et al. (2014) proposed that a radiologist interpreting a mammography image incorrectly may make screening challenging. Radiographic interpretation can be challenging when the portrayal of normal tissue is unpredictable or complex, and the cancer cell manifestation is incredibly minute or vague (Adnan et al., 2021). To advance this process, the computer-aided detection (CAD) technique has been created to help radiologists efficiently identify masses on mammograms that could be signs of breast cancer (Ashraf et al., 2021).

Computer-aided diagnosis (CAD), which has advanced quickly, is now being deployed gradually in the clinical setting to help physicians identify diseases (Al Najdawi et al., 2019). Nevertheless, there are numerous false positive aspects that the computer has identified, which will consume more time for the clinicians to re-evaluate the data, lowering the quality and robustness (Sharma and Kumar, 2020). Therefore, machine learning (ML) has effectively been implemented in diagnosing breast cancer to enhance the overall efficiency of the detection approach (Singla and Kumari, 2021). It not only aids in cancer diagnosis but also aids in determining the stage of the disease, making it easier for doctors to prescribe the right medications to their patients (Kumari and Bhatia, 2021). The detection of cancer benefits greatly from ML approaches. Cancer can be classified as benign or malignant based on the abnormalities (Gigras et al., 2021). Identifying them is made simple by a standard approach through feature extraction,

which identifies the essential geometric data in an image (El-Hasnony et al., 2021). This paper proposes a system that combines computer-extracted information from mammographic pictures with a classification tool (Sharma et al., 2020).

Berbar (2018) proposed numerous feature extraction techniques based on shape, texture, and grey level features that have been put forth recently for classifying masses in mammography images. When determining whether a mammography image is benign or malignant, grey level characteristics, first-order statistics like mean, standard deviation, and variance, are utilised to quantify the intensity variation.

Rabidas et al. (2018) proposed that to discriminate between mass and normal breast tissue, feature extraction utilising shape characteristics is based on the collection of spatial patterns of pixels, such as area, perimeter, compactness, and circularity. Since it represents the spatial variability of pixel intensities in an image, the texture is crucial for categorising masses in mammography pictures. Additionally, it establishes the smoothness or roughness of visual characteristics. Structured, statistical, and transform methods extract texture features from the image.

Nagarajan et al. (2019) proposed the following key problems with the current feature extraction methods:

1 The grey level feature-based techniques had poor accuracy in classifying masses.

2 The mass cannot be classified as benign or malignant based solely on its shape. It serves only to separate the mass region from healthy breast tissue.

3 Due to the use of basic functions or filtering schemes, most texture feature extraction approaches are non-adaptive and rely on transform methods.

The proposed method incorporates a strategy that uses texture, intensity, and shape-based feature extraction methodologies to overcome the abovementioned issues (Rajpurohit et al., 2020).

This paper presents a robust and discriminatory feature extraction methodology that combines computer-extracted information from mammographic pictures with a classification tool for false positive reduction (normal and abnormal classification) and to optimise the performance of CAD systems in terms of accuracy, sensitivity, and specificity. The main contribution is that CLAHE is used to pre-process the mammography images, and the ROI is obtained. A novel feature extraction method known as EGLCM that extracts features combining GLCM (texture), entropy (intensity), and Fourier descriptor (shape-based) is proposed (Bibi et al., 2022). The KNN classifier receives these features as input. The suggested feature extraction methodology EGLCM is compared to previous approaches like LBP and GLRLM. Confusion matrix parameters are used to assess these approaches' performance (Iranmakani et al., 2020).

The entire paper is divided into the following sections. Section 1 deals with the introduction of the work, Section 2 deals with the detailed related works done so far by the earlier researchers, Section 3 deals with the proposed methodology, Section 4 deals with Results and Discussions of the proposed work, and last, and Section 5 deals with a conclusion.

2 Related works

Wisudawati et al. (2021) propose optimising the given method's performance metrics by using GLCM texture characteristics from 3 tiers of mammography image decomposition. Using DWT, the mammography images are turned into wavelet coefficients, which GLCM then employs to acquire the texture characteristics needed to generate the test dataset. The classification procedure examines the performance features using the test database. Tier after tier, the performance characteristics are maximised. It is concluded that the findings will improve the levels of deconstruction of the mammography images. The earlier work is then contrasted with the proposed approach, which is determined to be superior.

Xu and Li (2020) created to examine breast cancer features to increase breast tumour identification's effectiveness and lower the charge of breast tumour identification. Additionally, it attempted to develop a 2D and 3D detection model for breast malignancies using a classic ultra-wideband microwave technique.

Teramoto et al. (2019) additionally, the study applied a three-dimensional textural feature extraction approach to analyse the image and used dimensional translation to refine the two-dimensional textural feature extraction approach. Finally, this study's feature model is validated by comparison tests in this work, which also gathers distinct images of 2D and 3D detection performances.

Using several techniques for textural analysis to extract a feature from the mammography picture, the classification results were compared in this study by Novitasari et al. (2019). Certain textural analysis techniques, such as the first-order technique consisting of the GLCM, GLDM, and GLRLM, have successfully extracted features depending on their properties. The ECOC SVM classification uses the statistical characteristics of these techniques as input and builds the classification using three kernel comparisons: linear, RBF, and polynomial. The findings demonstrate that the best kernels containing statistical features created by GLRLM and greater accuracy values are polynomial kernels.

In this analysis, Militello et al. (2022) attempt to assess the potential effects that various segmentation techniques and quantisation grades would have on the stability of radiomic characteristics. Specifically, the intra-class correlation coefficient (ICC) and mean variances between segmentation evaluators were used to assess the stability of texture characteristics generated using PyRadiomics and related to five feature extraction methodologies. Each feature category's overall index was quantified to the durability of each feature. The investigation revealed that a crucial factor in defining sturdy features is the level of quantisation. Furthermore, while automated segmentation produced reduced ICC values, both manual segmentation approaches showed strong stability and consistency.

Htay et al. (2020) presented a strategy for categorising breast cancer in this work utilising image processing methods to aid physicians in improving the mammography screening procedure and extending the lives of people with cancer. There are five stages in this method. First, pre-processing is done to an image to eliminate noise and artefacts, crop the image to make it smaller, and enhance the image to make the portrayal of the image stand out. Median filters are employed to eliminate background aberrations, noise, high-frequency elements, and undesirable aspects of mammography images. Second, the breast region is separated from the background image using Otsu segmentation. A third stage involves applying augmentation to segmented output images to create effective

features that increase classification accuracy. The improved image retrieves first-order statistical and second-order textured GLCM features. A support vector machine classifier distinguishes between aberrant and normal images.

Tian and Zhang (2022) suggest a hybrid approach that integrates principal component analysis (PCA) with boosted C5.0 decision tree approach with a penalty factor to address the dual class mismatch issue in breast cancer detection. The dimension of the feature subset is decreased using PCA. The ensemble classifier used for classification is the boosted C5.0 decision tree method. The categorisation result is optimised using the penalty factor. Implementing the suggested strategy on biased representational breast cancer databases from the UCI repository allows us to see its effectiveness. According to the experimental findings and subsequent study, a concept is a potential approach for treating breast cancer and can be utilised as a substitute approach in class imbalance learning.

According to the suggested study by Singla and Kumari (2021), the fuzzy-NN algorithm outperforms other classification methods for the bior3.7 wavelet group when there is no feature selection. Following feature selection, the MLP algorithm's wavelet defines the breast cancer tumours' for the db8 wavelet group with the highest degree of accuracy. fuzzy-NN techniques have the maximum accuracy when comparing methods for the bior3.7 wavelet group with and without feature selection. It shows that the wavelet groups vary based on whether features are selected. A db8 wavelet group is a superior option when features are selected, and the bior3.7 wavelet family is when features are not selected.

3 Proposed methodology

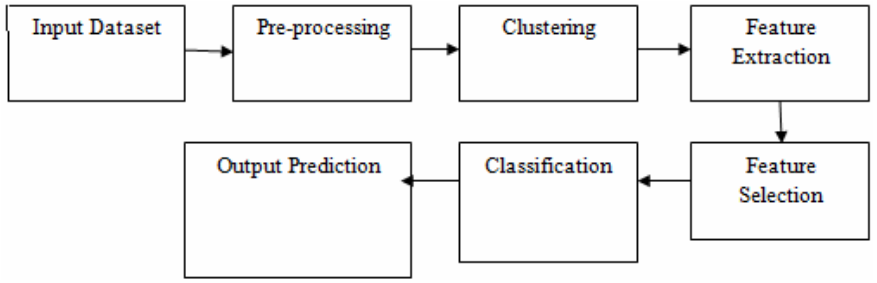
The proposed methodology is based on various stages that detect or categorise tumours as either normal (tumour-free) or abnormal (presence of tumour). Image acquisition, which involves acquiring a mammography image using MIAS, is the first step in diagnosing breast tumours'. To improve the quality of feature extraction, pre-processing is done using the CLAHE approach, which improves the contrast in images. The mammography image's remaining portions are removed during this stage, and the region of interest (ROI) is extracted. The mammography image's characteristics are then obtained through feature extraction. The proposed enhanced GLCM (EGLCM) approach extracts the features. The tumour's texture, intensity, and shape-based features are extracted from mammography images using this unique feature extraction technique known as EGLCM. Its effectiveness is determined by contrasting the suggested methodology's findings with other feature extraction methods using a confusion matrix. The last step is to classify the image to ascertain if normal or abnormal patterns exist. Figure 1 represents the proposed flow diagram representing each step.

3.1 Image acquisition

The mammography method adopted for breast cancer diagnostics produces mammogram images. The mini-MIAS database, where MIAS refers to the Mammographic Image Analysis Society, gathers mammogram pictures. This was given its name in honour of a UK research team whose primary goal was to identify breast cancer through digital

mammography images in their studies. It has 322 photos, including breast images from 161 benign, normal, and malignant people. The mammography picture is a PNG file with a resolution of 1,024 × 1,024 pixels (Suckling et al., 2015). While 210 of the mammography photos are either benign or cancerous, 112 of the images show normal breasts. In proportions of 60%, 20%, and 20%, respectively, the datasets were divided randomly as training, testing, and validation sets.

Figure 1 The proposed flow diagram representing each step



3.2 Pre-processing

Pre-processing is a crucial aspect of medical image analysis since it improves the quality of the image for feature extraction techniques. The noise exposure and artefacts in mammograms make them poor contrast images. Pre-processing is required to enhance images, protect picture edges, and optimise image quality. Different kinds of procedures are used in pre-processing. Here, a contrast enhancement is employed during pre-processing to improve the contrast in the photos. Following this, the ROI is derived from the mammography images by cropping the digital mammogram images to remove the unwanted portions of the image.

3.3 Image enhancement

Image enhancement procedures generally draw attention to specific elements in a mammography image. Image enhancement, in this sense, refers to the processing of the pictures to boost contrast and reduce noise to help radiologists find anomalies. The adaptive contrast enhancement method (AHE), one of many image enhancement approaches, can enhance local contrast and bring out additional details within the image. In this study, the contrast of images is enhanced by user-defined clip levels using the contrast-limited adaptive histogram equalisation method (CLAHE), a form of AHE (Jayandhi et al., 2022). Utilising CLAHE reduces the over-amplification in AHE. This technique, developed for medical imaging, lessens the noise and edge-shadowing produced in predictable regions. It is applied to eliminate artefacts from mammograms, such as wedges, labels, and markers, and it increases the visibility of concerning or hidden regions. The image in CLAHE is divided into discrete pieces called tiles. This method improves the contrast between each tile. Algorithm 1 describes the CLAHE processes.

Algorithm 1 CLAHE method

Input: Input image B

Output: Contrast-enhanced image

Start

1 Divide the original image B into contextual regions of size 8 x 8.

2 Obtain a local histogram for each pixel.

3 Limit this histogram based on the clip level to alter the image's contrast.

Clip Limit is calculated as

$$N_{avg} = \frac{Nr_X \times Nr_Y}{N_{grey}} \tag{1}$$

where N_{avg} is average pixel count

N_{gray} is the grey level numbers in the contextual region

Nr_X and Nr_Y are numbers of pixels in the X-axis and Y-axis.

$$N_{CL} = N_{clip} \times N_{grey} \tag{2}$$

Here, N_{CL} is the original clip limit and N_{clip} is the normalised clip limit.

4 Redistribute the histogram using binary search.

$$\text{Redistribution} = \frac{N_{grey}}{N_{remain}} \tag{3}$$

5 Obtain the enhanced pixel value by histogram integration.

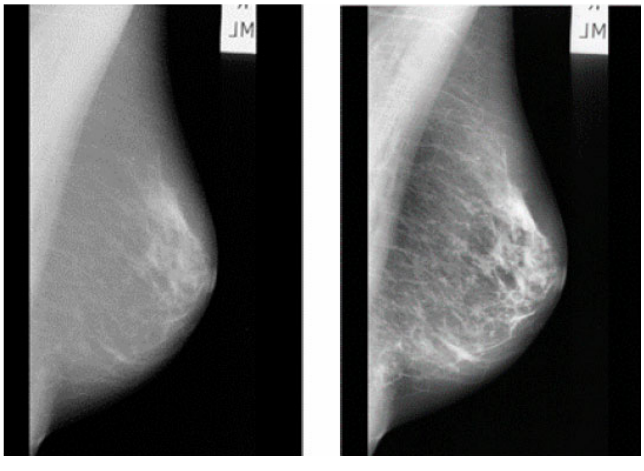
$$h(r_i) = \sum_{j=0}^{i} \frac{n_j}{n} \tag{4}$$

n is the mammogram pixel count

n_j is the pixel number given as input of greyscale value j.

End

Figure 2 First is the original abnormal mass, and second is the Enhanced image using the CLAHE method



The enhanced image is presented in Figure 2 after the algorithm above was applied to images in the MIAS dataset.

3.4 Region of interest

After enhancing the image, the ROI is automatically and adaptively extracted. The suggested method carries out the automatic cropping procedure and dynamically adapts to the mammography properties. Algorithm 2 illustrates the proposed technique to extract ROI for the left/right orientation image.

Algorithm 2 Automatic Image Cropping

Input: The result after contrast enhancement ($B_{p,q}$)

Output: Region of interest

Start

1 Check the input image ($B_{p,q}$)

2 Specify the start point (x_0, y_0) and endpoint (x_1, y_1) where $x_0 = 0$, $y_0 = 0$ and $x_1 = 0$, $y_1 = 0$.

3 Obtain the new locations x_0 and y_0 at every place in the input image in which the intensity was found by reading every point from the right direction (horizontal):

$x_0 = p$ and $y_0 = q$ if $B_{(p,q)} > 0$ and $x_0 = 0$.

4 Get the new location x_1 and y_1 at the place in which the intensity is obtained by reading every point in the source images from the bottom location (vertical).

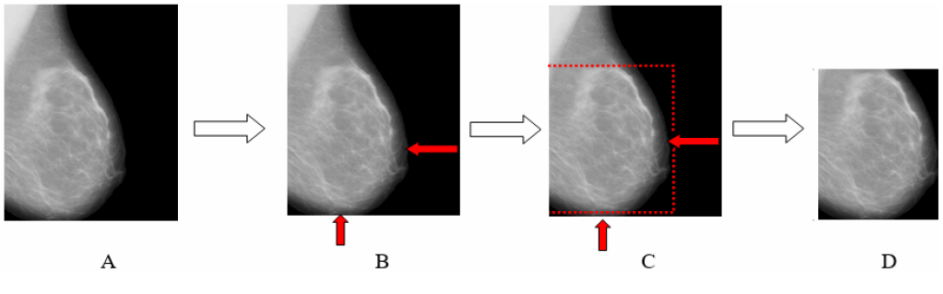
$x_1 = p$ and $y_1 = q$ if $B_{(p,q)} > 0$ and $x_1 = 0$.

5 Identify the ROI zone and the mammae area: the x and y positions are twice as length and width (Bcrop)

End

The first step in Algorithm 2 is to define the start and finish points before analysing the input image, which is a left-oriented mammography image. The right horizontal position of the image serves as the starting point for the search for the emergence of the intensity value, giving rise to new locations x_0 and y_0. Then, find a new location x_1 and y_1 by looking for the intensity value at the image's bottom location vertically. Cropping a mammae area involves multiplying the x and y positions' length and width by two. Algorithm 2 follows the same procedure for the right orientation but reads the image point commencing from the left location of the image.

Figure 3 ROI extraction algorithm representation (see online version for colours)



When cropping, the mammography image's unnecessary areas are eliminated and the suspected areas are extracted (ROI) using Algorithm 2. Figure 3 represents Algorithm 2, where A is a mammography image (enhancement outcomes) with a left-facing orientation, and B illustrates how to discover the presence of intensity spots from the right and bottom. Area mammae can be obtained by multiplying the distance between the x and y points by two, as shown in C, and its extracted area is described in D.

3.5 Feature extraction

As numerous features are extracted, such as spectral, textural, and contextual attributes, feature extraction is extracting attributes from images for subjective interpretation. The accuracy of the feature vector calculation determines how effectively a classifier performs. There are numerous methods for extracting features. Here, the proposed enhanced GLCM approach extracts the features from the pre-processed images. The various feature extraction techniques are described in more detail below.

3.5.1 Enhanced grey-level co-occurrence matrix

The EGLCM technique is applied in the proposed methodology to extract the characteristics, which use texture, intensity, and shape-based elements since these aspects are essential for identifying tumours' or lesions in mammograms. While shape features represent the lesion borders, texture features in a picture are the spatial dispersion of grey levels. Grey level co-occurrence matrix (GLCM) is used to collect texture data, entropy is utilised to collect intensity-based characteristics, and Fourier descriptor is employed to collect shape-based features. EGLCM is the term given to this feature combination (Kumari and Jagadesh, 2022).

Algorithm 3 Enhanced GLCM (B_m, Ds, θ)

Input: B_m : ROI image, D: distance, angles = [0°, 45°, 90°, 135°]

Output: Feature vector

Start

1 The input is the contrast-enhanced images.

2 Using the parameters D = 1 and θ = 0°, 45°, 90°, 135°, construct the co-occurrence matrix.

3 Create a symmetric GLCM.

4 Make the GLCM normalised.

5 Use D = 1 to compute the GLCM features from four various angles.

6 Calculate entropy using GLCM.

7 Determine the Fourier coefficacy.

8 Determine the average Fourier coefficients.

9 To obtain the finalised feature vector, integrate the GLCM features, entropy, and statistical Fourier coefficients.

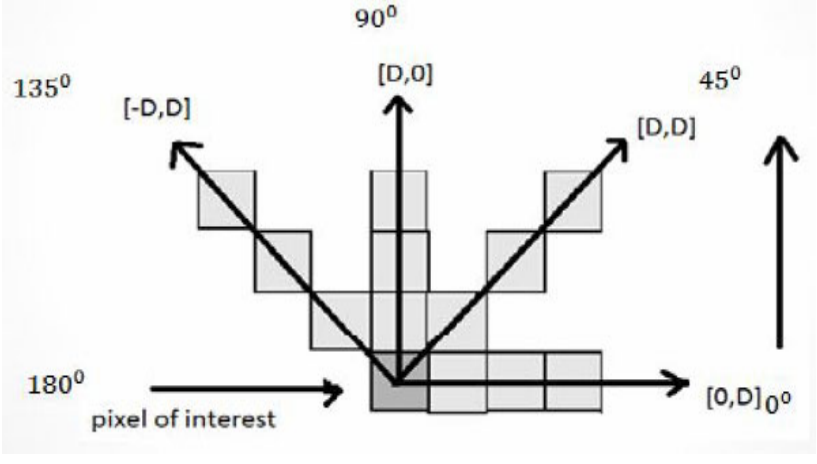
10 Provide the feature vector.

End

3.5.1.1 GLCM

Statistical second-order texture features can be extracted using this technique. This method yields the distribution of pixel pairs with different grey levels in the image. It creates a co-occurrence matrix to model the associations between pixels in the area (CM). The comparative frequencies of the adjacent pixels that are spaced apart by the value 'D' are represented by this matrix. The elements in the matrix represent the pixel intensities' frequency variability. The calculation of the conditional probability functions for different directions θ and varied distances, D, is the basis of the GLCM. When the distance between two pixels is minimal and the texture is rough, the grey values of the two pixels at that distance D should be identical. On the other hand, for a smooth texture, the set of pixels at distance D should frequently be very different, ensuring that the values in the GLCM are distributed fairly equally. The degree of spreading of the values around the main diagonal in the GLCM must also change with the direction of the texture rougher in one direction than the other. At a distance of 1 and in the direction of 0°, it depicts the creation of the GLCM of a grey-level (four levels) image (Figure 4).

Figure 4 GLCM directions



Below are the main GLCM technique steps.

Step 1 Quantise the image information.

Step 2 Construct the GLCM.

Step 3 The GLCM should be symmetric.

Step 4 Transpose the GLCM and make a copy of it.

Step 5 Include this copy in the GLCM.

Step 6 Make the GLCM normalised.

The computed features from CM are as follows:

3.5.1.2 Energy

The square of each matrix element depicts the roughness of the texture and the uniformity of the greyscale distribution of the images. Small energy profiles were produced when all the co-occurrence matrix values were the same; however, substantial energy could be anticipated when the co-occurrence matrix values differed.

The textural regularity is computed through the energy feature.

$$\text{Energy} = \sum_{k,l}^{n} p_{k,l}^2 \tag{5}$$

where p_{kl} – elements of x, y.

3.5.1.3 Homogeneity

The homogeneity feature computes an image's variability. The homogeneity of the image is measured, with bigger values being assumed for smaller variances in the grey tone between pair elements. The GLCM's homogeneity is more susceptible to the existence of near diagonal pieces. When all of the components of the image match, homogeneity is at its highest level. Because of the significant but negative correlation between contrast and homogeneity in the GLCM, homogeneity falls as contrast rises while energy remains constant.

The homogeneity feature computes an image's variability.

$$\text{Homogeneity} = \sum_{k,l=1}^{n} \frac{P_{kl}}{1+(k-l)^2} \tag{6}$$

The dissimilarity feature determines the distance between each pair of pixels.

3.5.1.4 Dissimilarity

An image's local variances can be quantified linearly by dissimilarity. The dissimilarity feature determines the distance between each pair of pixels.

$$\text{Dissimilarity} = \sum_{k} \sum_{l} |k-l|\, p(k,l) \tag{7}$$

The contrast feature determines the spatial frequency of an image.

3.5.1.5 Contrast

The contrast feature determines the spatial frequency of an image. The disparity between the highest and lowest values of the pixels next to each other is what this term refers to. The contrasting texture measures the local variations that are present in the image. Low spatial frequencies are present in an image with low contrast, and the GLCM concentration term is centred on the main diagonal.

$$\text{Contrast} = \sum_{k,l=1}^{n} p_{kl}\,(k-l)^2 \tag{8}$$

3.5.1.6 Entropy

It gives information about the contents of the image. It gives the image uncertainty, randomness, or intensity levels to characterise the input image. This is calculated and added to the GLCM feature vector as another feature.

$$\text{Entropy} = \sum_{k,l=0}^{n-1} -\ln\left(p_{kl}\right) p_{kl} \tag{9}$$

3.5.1.7 Fourier descriptors

The tumour's shape makes it simple to determine if it is normal or malignant. The lesion boundaries are described by shape attributes (rounded, spiculate or stellate). Fourier descriptors (FDs) are utilised to identify the morphology of the tumor in mammograms and are very helpful for pattern identification. Thus, shape-based characteristics are described using the Fourier descriptor (Sathish et al., 2016).FDs are obtained from Fourier transforms, in which the smaller frequency scores and the fine details by the larger frequency scores represent the overall shape. The contour of a boundary is carried by a small number of FDs encapsulating its essence. Thus, these are employed to distinguish between various boundary shapes. A complex number can be used to represent a digital boundary. A complex number is represented by the boundary at any location starting at the coordinate pair denotation. Because it transforms a 2d issue into a 1-d issue, this complex integer offers a significant benefit. There is only FD in this complex co-efficient. The definition of a discrete Fourier is,

$$F_n = \frac{1}{N} \sum_{i=0}^{N-1} C(i) \times e^{\frac{-j2\pi ni}{N}} \tag{10}$$

where F_n represents nth FD, C(i) denotes 1D contour signal, and N gives the total count of contour.

3.5.2 Local binary pattern

An operator called local binary pattern (LBP) is used to explain the local texture components of an image (Wan et al., 2017). Rotation and grey level invariance are just two of its many benefits. The central pixel in a 3×3 window serves as the threshold value for the primary LBP operator, which compares the grey levels of eight nearby pixels with it. The location of the pixel is labelled as 1 if the value of the nearby pixel is larger than or equal to the valuation of the central pixel; otherwise, it is labelled as 0. For example, consider pixel (x_c, y_c) on the image.

$$LBP_{U,V}\left(x_c, y_c\right) = \sum_{u=0}^{U-1} s\left(g_u - g_c\right) 2^u, \quad s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{11}$$

where g_c is the central pixel's grey value; g_u is the grey value of the pixel next to the central pixel; U is the number of sample points in the central pixel's neighbourhood; V is indeed the radius of the neighbourhood. The LBP value of the central pixel of the 3×3 window is generated in this manner, and this value is utilised to represent the texture data of the area. Doing so, eight points in the surrounding can be compared to produce a maximum of 256 8-bit binary integers.

3.5.3 Grey-level run-length matrix

In mammography images, high-order statistical features can be extracted using GLRLM (Humeau-Heurtier, 2019). It is a collection of continuous greyscale pixels. The amount of adjacent grey levels in a certain direction makes up the run length. It is determined by counting the instances of the same run in the image, such as two successive pixels with the same value. The identical procedure follows the next set of integers for the following three pixels. It should be noted that multiple run-length matrices, one for each selected direction, may be generated for a single image.

$$R(x, y) = (g(x, y)|x), \qquad 0 \le x \le N_g, 0 \le y \le R_{max} \tag{12}$$

where R_{max} is the longest length possible, and N_g is the highest possible grey-level. The GLRLM can be used to extract six textural features. These features employ the pixel's grey level one at a time and are designed to distinguish between textures that have equal value but differ in their grey level distribution.

3.5.4 Classification

K-nearest neighbour (KNN) is a widely used classifier because it is easy to use, intuitive, and very effective, especially with noisy data. It is nonetheless capable of reaching significant accuracy rates in medical applications despite its simplicity. According to the class of every data point among its k-nearest neighbours in the training set, K-NN classifies all the data points in the test set. The distance between every data point within the test set that must be categorised and subsequent data points in the training set is measured to achieve this. The distance shows how similar the test set and training set cases are. The Euclidean distance is employed in this method, and the value of k is two.

4 Results and discussion

The proposed feature extraction approach is assessed on mammography images from the MIAS database. A computer equipped with an Intel Core i7-7700k (4.2 GHz) processor, 32 GB of DDR4 RAM, and an NVIDIA GTX 1050Ti graphics card is used to execute feature extraction algorithms and classifiers. With OpenCV and other required libraries, the breast tumour classification and detection model is trained and tested using Python 3.6.

Several performance metrics confirm the methodology's efficacy, including sensitivity, specificity, and accuracy dependent on the confusion matrix. The confusion matrix's components are as follows:

- true positive rate (TPR) indicates properly diagnosed unhealthy individuals
- FPR stands for false positive rate indicates that healthy individuals are misdiagnosed
- true negative rate (TNR) indicates that healthy individuals have been properly diagnosed
- false negative rate (FNR) is the percentage of unhealthy individuals who have received an inaccurate diagnosis.

$$\text{Accuracy} = \frac{\text{TPR} + \text{TNR}}{\text{TPR} + \text{FPR} + \text{TNR} + \text{FNR}} \tag{13}$$

$$\text{Specificity} = \frac{\text{TNR}}{\text{FPR} = \text{TNR}} \tag{14}$$

$$\text{Sensitivity} = \frac{\text{TPR}}{\text{TPR} = \text{FNR}} \tag{15}$$

4.1 Accuracy analysis

Table 1 represents the accuracy obtained by EGLCM, LBP and GLRLM after classifying by the KNN algorithm.

Table 1 Accuracy by EGLCM, LBP and GLRLM with KNN

No of images	*LBP+KNN accuracy %*	*GLRLM+KNN accuracy %*	*EGLCM+KNN accuracy %*
1	86.45	88.72	89.21
2	86.90	88.92	90.15
3	87.96	89.11	90.65
4	88.05	89.55	91.50
5	88.57	89.93	91.88

Figure 5 Accuracy by EGLCM, LBP and GLRLM with KNN graph (see online version for colours)

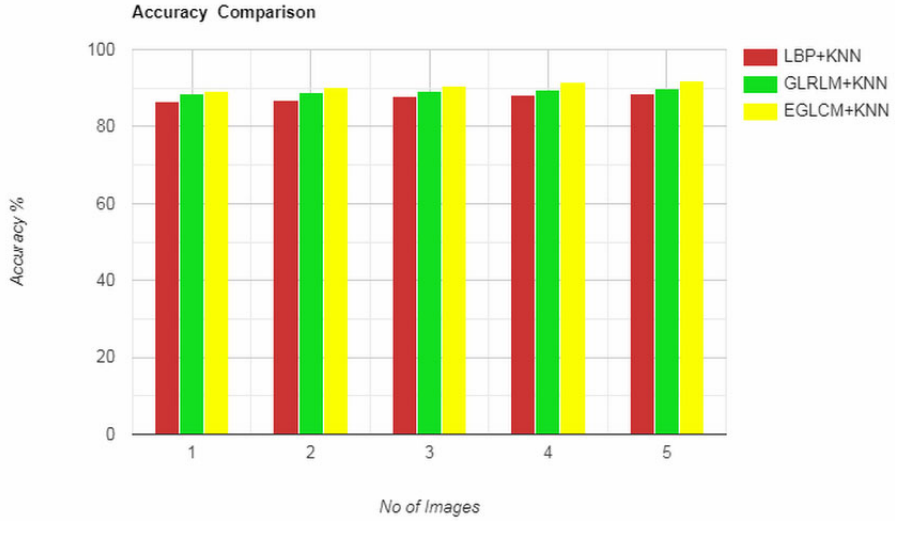


Table 1 and Figure 5 represent the accuracy obtained by EGLCM, LBP and GLRLM with KNN classification. The results proved that EGLCM+KNN produces accuracy ranging from 89% to 92%, which is higher than LBP+KNN ranging from 86% to 88% and GLRLM+KNN ranging from 88 % to 90 %, respectively (Sharma et al., 2022).

4.2 Specificity analysis

Table 2 represents the specificity obtained by EGLCM, LBP and GLRLM after classifying by the KNN algorithm (Sharma et al., 2022).

Table 2 Specificity by EGLCM, LBP and GLRLM with KNN

No. of images	*LBP+KNN specificity %*	*GLRLM+KNN specificity %*	*EGLCM+KNN specificity %*
10	84.91	86.91	88.65
20	85.32	87.17	89.56
30	85.64	87.42	89.99
40	86.12	88.10	90.25
50	86.54	88.54	90.88

Figure 6 Specificity by EGLCM, LBP and GLRLM with KNN graph (see online version for colours)

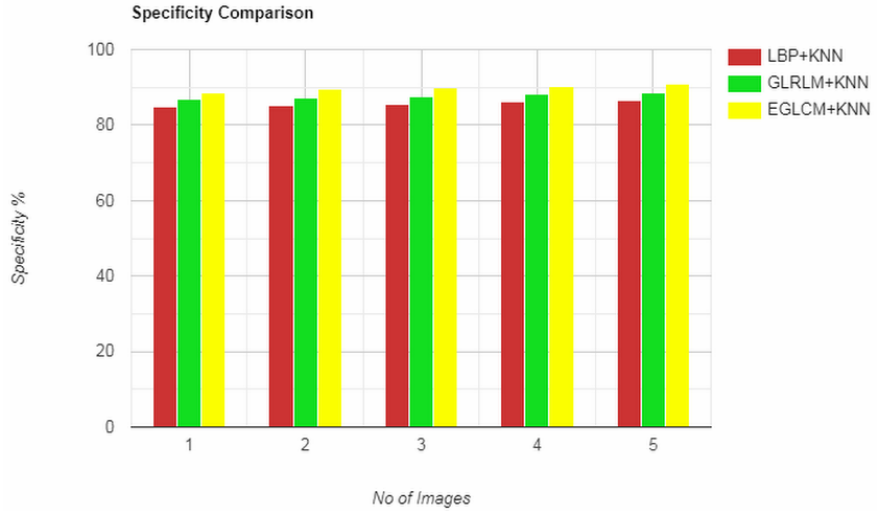


Table 2 and Figure 6 represent specificity obtained by EGLCM, LBP and GLRLM with KNN classification. The results proved that EGLCM+KNN produces specificity ranging from 88% to 91%, which is higher than LBP+KNN, ranging from 86% to 89% and GLRLM+KNN ranging from 88 % to 90 %, respectively (Sharma et al., 2022).

4.3 Sensitivity analysis

Table 3 represents the sensitivity obtained by EGLCM, LBP and GLRLM after classifying by the KNN algorithm (Sharma et al., 2022).

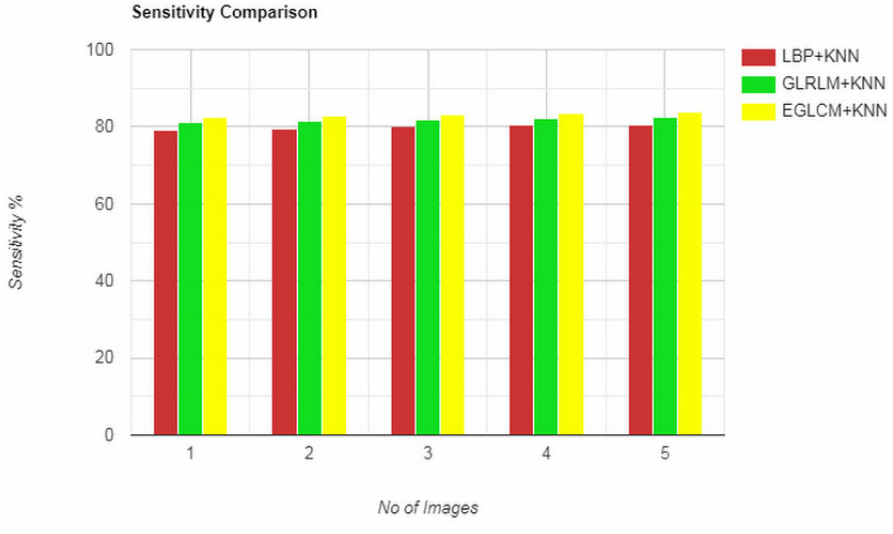
Table 3 and Figure 7 represent sensitivity obtained by EGLCM, LBP and GLRLM with KNN classification. The results proved that EGLCM+KNN produces sensitivity ranging from 82% to 84%, which is higher than LBP+KNN ranging from 79% to 81% and GLRLM+KNN ranging from 81% to 83%, respectively. The texture, intensity-based

characteristics, and shape-based features are combined in the framework since it's crucial to know the intensity and shape of the tumour as well as the texture while looking for abnormalities in mammograms. It has been found that employing CLAHE as a pre-processing method, EGLCM as a feature extraction method, and KNN classifier increases the accuracy of mammography image classification. This research has demonstrated a positive correlation between pre-processing and improved classifier accuracy using a robust feature extraction technique.

Table 3 Sensitivity by EGLCM, LBP and GLRLM with KNN

No of images	*LBP+KNN sensitivity %*	*GLRLM+KNN sensitivity %*	*EGLCM+KNN sensitivity %*
10	79.15	81.07	82.57
20	79.58	81.36	82.95
30	80.08	81.97	83.10
40	80.37	82.12	83.65
50	80.55	82.50	83.95

Figure 7 Sensitivity by EGLCM, LBP and GLRLM with KNN graph (see online version for colours)



5 Conclusions

In this work, we introduced a developed feature extraction approach called enhanced GLCM. GLCM, entropy, and Fourier descriptors are combined to extract various features. This method uses an automatic image cropping methodology to obtain ROI after applying the pre-processing technique CLAHE to enhance contrast in mammograms. It then combines texture, intensity, and shape-based features in the proposed feature extraction technique EGLCM. The KNN classifiers are used to classify these features. The MIAS dataset is used for the tests, and the outcomes are contrasted with those of

other feature extraction techniques like LBP and GLRLM. In terms of accuracy, specificity, and sensitivity, the effectiveness of the proposed method shows that it improved performance concerning confusion matrix characteristics. A minimal feature collection is sufficient to enable class discrimination; hence, feature selection isn't required for this method. Even when using recent graphics processing units, training a deep learning network can take longer. In contrast, this work takes less time and can be implemented in any central processing unit that is accessible. Future research will suggest faster and more effective ways to identify breast cancer. The goal is to achieve more accurate multi-class breast cancer recognition.

References

Adnan, R.M., Mostafa, R.R., Islam, A.R.M.T., Gorgij, A.D., Kuriqi, A. and Kisi, O. (2021). Improving drought modeling using hybrid random vector functional link methods', *Water*, Vol. 13, No. 23, p.3379, DOI: 10.3390/w13233379.

Al Najdawi, N., Rashaideh, H. and Shaheen, A. (2019) 'Real-time image encryption and decryption methods based on the Karhunen-Loeve transform', *International Journal of Intelligent Engineering Informatics*, Vol. 7, No. 5, p.399, https://doi.org/10.1504/ijiei.2019.10025231.

Arabaci, H. and Mohamed, M.A. (2020) 'A knowledge-based diagnosis algorithm for broken rotor bar fault classification using FFT, principal component analysis and support vector machines', *International Journal of Intelligent Engineering Informatics*, Vol. 8, No. 1, p.19, https://doi.org/10.1504/ijiei.2020.10027093.

Ashraf, N.M., Mostafa, R.R., Sakr, R.H. and Rashad, M.Z. (2021) 'Optimizing hyperparameters of deep reinforcement learning for autonomous driving based on whale optimization algorithm', *PloS One*, Vol. 16, No. 6, p.e0252754, DOI: 10.1371/journal.pone.0252754.

Berbar, M.A. (2018) 'Hybrid methods for feature extraction for breast masses classification', *Egyptian Informatics Journal*, Vol. 19, No. 1, pp.63–73, https://doi.org/10.1016/j.eij.2017.08.001.

Bibi, A., Attique Khan, M., Younus Javed, M., Tariq, U., Kang, B.-G., Nam, Y., H. Sakr, R. (2022) 'Skin lesion segmentation and classification using conventional and deep learning based framework', *Computers, Materials & Continua*, Vol. 71, No. 2, pp.2477–2495, DOI: 10.32604/cmc.2022.018917.

Bulletins, G. (2017) 'Practice bulletin number 179: breast cancer risk assessment and screening in average-risk women: Breast cancer risk assessment and screening in average-risk women', *Obstetrics and Gynecology*, Vol. 130, No. 1, pp.e1–e16, https://doi.org/10.1097/aog.0000000000002158.

El-Gamal, A.H., Mostafa, R.R. and Hikal, N.A. (2020) 'Load balancing enhanced technique for static task scheduling in cloud computing environments', *Internet of Things – Applications and Future*, pp.411–430, Springer Singapore, Singapore, https://doi.org/10.1007/978-981-15-3075-3_28.

El-Hasnony, I.M., Mostafa, R.R., Elhoseny, M. and Barakat, S.I. (2021) 'Leveraging mist and fog for big data analytics in IoT environment', *Transactions on Emerging Telecommunications Technologies*, Vol. 32, No. 7, DOI:10.1002/ett.4057.

Elkabbash, E.T., Mostafa, R.R. and Barakat, S.I. (2021) 'Android malware classification based on random vector functional link and artificial Jellyfish Search optimizer', *PloS One*, Vol. 16, No. 11, p.e0260232, https://doi.org/10.1371/journal.pone.0260232.

Giess, C.S., Frost, E.P. and Birdwell, R.L. (2014) 'Interpreting one-view mammographic findings: minimizing callbacks while maximizing cancer detection', *Radiographics: a review publication of the Radiological Society of North America, Inc*, Vol. 34, No. 4, pp.928–940, https://doi.org/10.1148/rg.344130066.

Gigras, Y., Lamba, M. and Munjal, G. (2021) 'A MCDM-based performance of classification algorithms in breast cancer prediction for imbalanced datasets', *International Journal of Intelligent Engineering Informatics*, Vol. 9, No. 5, p.425, https://doi.org/10.1504/ijiei.2021.10044779.

Htay, T.T., Maung, S.S. and Zar, K.T. (2020) 'Analysis on results comparison of feature extraction methods for breast cancer classification', *International Journal of Advances in Scientific Research and Engineering*, Vol. 6, No. 3, pp.95–102, https://doi.org/10.31695/ijasre.2020.33753.

Humeau-Heurtier, A. (2019) 'Texture feature extraction methods: a survey', *IEEE Access: Practical Innovations, Open Solutions*, Vol. 7, pp.8975–9000, https://doi.org/10.1109/access.2018.2890743.

Iranmakani, S., Mortezazadeh, T., Sajadian, F., Ghaziani, M.F., Ghafari, A., Khezerloo, D. and Musa, A.E. (2020) 'A review of various modalities in breast imaging: technical aspects and clinical outcomes', *The Egyptian Journal of Radiology and Nuclear Medicine*, Vol. 51, No. 1, DOI: 10.1186/s43055-020-00175-5.

Jayandhi, G., Jasmine, J.S.L. and Joans, S.M. (2022) 'Mammogram learning system for breast cancer diagnosis using deep learning SVM', *Computer Systems Science and Engineering*, Vol. 40, No. 2, pp.491–503, https://doi.org/10.32604/csse.2022.016376.

Kumari, L. and Jagadesh, B. (2022) 'A robust feature extraction technique for breast cancer detection using digital mammograms based on advanced GLCM approach', *EAI Endorsed Transactions on Pervasive Health and Technology*, Vol. 8, No. 30, p.172813, DOI: 10.4108/eai.11-1-2022.172813.

Kumari, N. and Bhatia, R. (2021) 'Systematic review of various feature extraction techniques for facial emotion recognition system', *International Journal of Intelligent Engineering Informatics*, Vol. 9, No. 1, p.59, https://doi.org/ 10.1504/ijiei.2021.10039255.

Militello, C., Rundo, L., Dimarco, M., Orlando, A., D'Angelo, I., Conti, V. and Bartolotta, T.V. (2022) 'Robustness analysis of DCE-MRI-derived radiomic features in breast masses: assessing quantization levels and segmentation agreement', *Applied Sciences (Basel, Switzerland)*, Vol. 12, No. 11, p.5512, DOI: 10.3390/app12115512.

Nagarajan, V., Britto, E.C. and Veeraputhiran, S.M. (2019) 'Feature extraction based on empirical mode decomposition for automatic mass classification of mammogram images', *Medicine in Novel Technology and Devices*, Vol. 1, No. 100004, p.100004, https://doi.org/10.1016/j.medntd.2019.100004.

Novitasari, D.C.R., Lubab, A., Sawiji, A. and Asyhar, A.H. (2019) 'Application of feature extraction for breast cancer using one order statistic, GLCM, GLRLM, and GLDM', *Advances in Science Technology and Engineering Systems Journal*, Vol. 4, No. 4, pp.115–120, DOI: 10.25046/aj040413.

Puthige, I., Bansal, K., Bindra, C., Kapur, M., Singh, D., Mishra, V.K. and Mostafa, R. (2021) 'Safest route detection via danger index calculation and K-means clustering', *Computers, Materials & Continua*, Vol. 69, No. 2, pp.2761–2777, DOI:10.32604/cmc.2021.018128.

Rabidas, R., Midya, A. and Chakraborty, J. (2018) 'Neighborhood structural similarity mapping for the classification of masses in mammograms', *IEEE Journal of Biomedical and Health Informatics*, Vol. 22, No. 3, pp.826–834, https://doi.org/ 10.1109/jbhi.2017.2715021.

Rajpurohit, V.S., Huddar, M.G. and Sannakki, S.S. (2020) 'Attention-based word-level contextual feature extraction and cross-modality fusion for sentiment analysis and emotion classification', *International Journal of Intelligent Engineering Informatics*, Vol. 8, No. 1, p.1, https://doi.org/10.1504/ijiei.2020.10027092.

Sathish, D., Kamath, S., Rajagopal, K.V. and Prasad, K. (2016) 'Medical imaging techniques and computer aided diagnostic approaches for the detection of breast cancer with an emphasis on thermography – a review', *International Journal of Medical Engineering and Informatics*, Vol. 8, No. 3, p.275, DOI: 10.1504/ijmei.2016.077446.

Sharma, H., Sharma, S. and Sharma, J.B. (2020) 'Artificial intelligence-based watermarking in hybrid DDS domain for security of colour images', *International Journal of Intelligent Engineering Informatics*, Vol. 8, No. 4, p.331, https://doi.org/10.1504/ijiei.2020.10034279.

Sharma, M., Shaji, C. and Unni, S.N. (2022) 'Optical polarization response of collagen: role in clinical cancer diagnostics – part I', *ISSS Journal of Micro and Smart Systems*, Vol. 11, pp.3–30, DOI:10.1007/s41683-022-00090-z.

Sharma, S. and Kumar, V. (2020) 'Low-level features based 2D face recognition using machine learning', *International Journal of Intelligent Engineering Informatics*, Vol. 8, No. 4, p.305, https://doi.org/10.1504/ijiei.2020.10034278.

Singla, S. and Kumari, U. (2021) 'Analysis of performance of two wavelet families using GLCM feature extraction for mammogram classification of breast cancer', *Recent Advances in Computer Science and Communications*, Vol. 14, No. 6, pp.1919–1925, https://doi.org/10.2174/2666255813666191218111850.

Suckling, J., Parker, J., Dance, D., Astley, S., Hutt, I., Boggis, C. and Savage, J. (2015) *Mammographic Image Analysis Society (MIAS) Database v1.21* [online] https://www.repository.cam.ac.uk/handle/1810/250394 (accessed 15 May 2022).

Tao, H., Al-Sulttani, A.O., Ameen, A.M.S., Ali, Z.H., Al-Ansari, N., Salih, S.Q. and Mostafa, R.R. (2020) 'Training and testing data division influence on hybrid machine learning model process: application of river flow forecasting', *Complexity*, Vol. 2020, pp.1–22, DOI: 10.1155/2020/8844367.

Teramoto, A., Yamada, A., Kiriyama, Y., Tsukamoto, T., Yan, K., Zhang, L. and Fujita, H. (2019) 'Automated classification of benign and malignant cells from lung cytological images using deep convolutional neural network', *Informatics in Medicine Unlocked*, Vol. 16, No. 100205, p.100205, DOI: 10.1016/j.imu.2019.100205.

Tian, J-X. and Zhang, J. (2022) 'Breast cancer diagnosis using feature extraction and boosted C5.0 decision tree algorithm with penalty factor', *Mathematical Biosciences and Engineering: MBE*, Vol. 19, No. 3, pp.2193–2205, https://doi.org/10.3934/mbe.2022102.

Virk, I.S. and Maini, R. (2020) 'Medical image segmentation based on fuzzy 2-partition Kapur entropy using fast recursive algorithm', *International Journal of Intelligent Engineering Informatics*, Vol. 8, No. 4, p.346, https://doi.org/10.1504/ijiei.2020.10034282.

Wan, S., Lee, H-C., Huang, X., Xu, T., Xu, T., Zeng, X. and Zhou, C. (2017) 'Integrated local binary pattern texture features for classification of breast tissue imaged by optical coherence microscopy', *Medical Image Analysis*, Vol. 38, pp.104–116, DOI: 10.1016/j.media.2017.03.002.

Wisudawati, L.M., Madenda, S., Wibowo, E.P., Abdullah, A.A., (2021) 'Benign and malignant breast tumors classification based on texture analysis and backpropagation neural network', *Computer Optics*, Vol. 45, No. 2, pp.227–234, https://doi.org/10.18287/2412-6179-co-769.

Xu, N. and Li, C. (2020) 'Image feature extraction in detection technology of breast tumor', *Journal of King Saud University. Science*, Vol. 32, No. 3, pp.2170–2175, https://doi.org/10.1016/j.jksus.2020.02.018.