TOWARDS A SUPPORT SYSTEM FOR DIGITAL MAMMOGRAM CLASSIFICATION

ADÉL BAJCSI

ABSTRACT. Cancer is the illness of the 21th century. With the development of technology some of these lesions became curable, if they are in an early stage. Researchers involved with image processing started to conduct experiments in the field of medical imaging, which contributed to the appearance of systems that can detect and/or diagnose illnesses in an early stage. This paper's aim is to create a similar system to help the detection of breast cancer. First, the region of interest is defined using filtering and two methods, Seeded Region Growing and Sliding Window Algorithm, to remove the pectoral muscle. The region of interest is segmented using k-means and further used together with the original image. Gray-Level Run-Length Matrix features (in four direction) are extracted from the image pairs. To filter the important features from resulting set Principal Component Analysis and a genetic algorithm based feature selection is used. For classification K-Nearest Neighbor, Support Vector Machine and Decision Tree classifiers are experimented. To train and test the system images of Mammographic Image Analysis Society are used. The best performance is achieved features for directions $\{45°, 90°, 135°\}$, applying GA feature selection and DT classification (with a maximum depth of 30). This paper presents a comprehensive analysis of the different combinations of the algorithms mentioned above, where the best performance repored is 100% and 59.2% to train and test accuracies respectively.

1. Introduction

Computer-aided diagnosis (CADx) and -detection (CADe) systems are frequently researched by the scientists to help doctors and to give patients higher chance to recover from their illness. Based on a report by the European Cancer Information System [1] in 2020, amongst females breast cancer causes the the most deaths. To diagnose the abnormality of the breast tissue often digital mammograms are used. Creating a system to process mammograms and to decide whether it is normal or abnormal (also deciding if the lesion is benign or malignant) can be divided into two blocks: (1) including preprocessing and segmentation and the (2) classification block (including feature extraction, -selection and classification).

The focus of the current paper is on constructing a CADx system by implementing each of the steps mentioned above. As input data MLO (medio-lateral oblique) view mammograms are used. For the first two steps we propose the use of unsupervised learning methods (filters, k-means), then use the result of this first block (the segmented image) as input, together with the original image to the second block. For feature extraction we propose the use of Gray-Level Run-Length Matrices (GLRLM), for feature selection Principal Component Analysis (PCA) and a genetic feature selection method (GA) and for classification K-Nearest Neighbors (KNN), Support Vector Machines (SVM) and Decision Tree (DT) classifiers are used. The novelty of the research consists in the classification approach (using one classifier to separate normal, benign and malignant mammogram images) and also in the combination of methods (k-means, GLRLM, PCA/GA, KNN/SVM/DT).

This paper is arranged as follows. In section 2 methods from the literature are presented to the steps mentioned previously. The details of the current experiment are presented in section 3 followed by the numerical results in section 4. Section 5 includes the conclusions of the experiment and the future work. Our acknowledgement is expressed in the last section.

2. Related work

CADe systems for cancerous cells are essential, because with an early detection there is a higher chance for the patients survival. Thus, methods for the steps mentioned in section 1 have been widely investigated in the literature. In the following paragraphs solutions from recent studies are presented.

2.1. **Preprocessing.** In case of digital mammogram analysis consists of the definition of the ROI, which is the region of the breast without the pectoral muscle. First, the foreground (labels, breast region, other artifacts) and the

---

[1]`https://ecis.jrc.ec.europa.eu/explorer.php`

background has to be separated. Second, border between the pectoral muscle and the breast has to be defined.

2.1.1. *Remove label and image artifacts.* The solution for the first step in the literature is implemented using thresholding, which is one of the simplest and easiest methods proposed to segment the image into foreground and background. In some studies [25, 12, 22] simple binary thresholding is used while others [15, 24, 19, 6, 18] use Otsu's thresholding with filters to prevent removing less dense tissues. In the previous researches Rahimeto et al. [15] used Wiener filter [20] while Salam et al. [19] used median filtering [20].

2.1.2. *Remove pectoral muscle.* The pectoral muscle is part of the breast but its intensity and density is similar to the abnormal tissues. Therefore, it can cause misclassifications. Hence, we want to remove the achieve better results.

A large number of existing studies in the broader literature have examined the use of segmentation methods to remove the pectoral muscle from mammograms. In some studies unsupervised methods are used like region growing [12, 6, 24, 18], thresholding [15, 25, 24] and k-means [24]. Other used supervised methods [24].

Maitra et al. [12] focuses on a method to remove the pectoral muscle based on a seeded region growing algorithm where the seeds are selected from a line from the pectoral muscle. Compared to Maitra et al. [12], Esener et al. [6] used only a single seed. To enhance the results of the proposed method they applied a straight-line approximation to define the boundary of the pectoral muscle.

Rahimeto et al. [15] used multilevel (Otsu's) thresholding to segment the breast's tissue into three groups: background, less dense tissue and highly dense tissue. The pectoral muscle as well as the abnormality are taking part of the highly dense segment. Then they define the pectoral muscle by measuring the perimeter and the portion of the perimeter on the edges. Finally, they propose the use of quadratic polynomial curve fitting to smooth the boundary of the muscle.

Shrivastava et al. [25] presented another unsupervised method to remove the pectoral muscle using a sliding window algorithm. While the given conditions are satisfied (minimum total intensity, maximum difference) the pixels inside the window are set to 0.

Shinde and Rao [24] proposed using Support Vector Machines (SVM) [1]. To define a possible location of the pectoral muscle they applied three segmentation methods: k-means clustering, Otsu's thresholding and region growing (the seed is selected based on the assumed location of the pectoral muscle and the region's intensity). They used Gray Level Co-occurrence Matrix (GLCM)

on each result to extract texture and statistical features. These feature were fed to an SVM which decided which one defines best the pectoral muscle.

The literature review conducted by Moghbel et al. [13] lists solutions (including the ones mentioned) presented for this step from the state-of-the-art.

2.2. **Segmentation.** Haty et al. [7] used Otsu's thresholding [3] to segment the breast tissue and in the last step kNN classification was used to decide if the breast is normal or abnormal. Sadeghi et. al [18] used a two step segmentation. In the first step they created a binary mask with a global threshold (relative maximums from the histogram) from the histogram normalized image. After applying the mask got from the first step they used morphological operations to enhance the texture of the remaining area. In the second step they go through the mammography image with two windows. Based on the average and the difference between the minimum and maximum intensity in the windows respectively they segment the tissue which is probably cancerous.

A recent study by Kim et al. [9] presented the use of convolutional neural networks (CNN) for unsupervised image segmentation. They alternatively predict the labels for each pixel and optimize the network's parameters until they become spatially continuous, or similar pixels are assigned to the same label and the number of labels is the highest. Li et. al [10] presented a dual CNN which segments the image and predicts the diagnosis simultaneously. The networks are running parallel and while the first one defines the semantic features the other one defines structural features. At the end they used the structural feature to segment the tumor and a fusion of features to decide if it is malignant or benign.

2.3. **Feature extraction.** Feature extraction has a key role in a CADe system. The extracted information consist of the input of the final classification. Mammograms are gray-scale images, therefore specific feature extraction methods need to be examined.

Vijayarajeswari et al. [28] applied Hough transform to the mammogram and then calculated statistical features like mean, variance, entropy and standard deviation. Similar features, are the wavelet features proposed by Rashed and Awad [16].

Chaieb and Kalti [5] conducted a literature survey on feature extraction from mammograms. They compared five statistical features (First Order Statistics, Gray-Level Co-occurrence Matrices, Gray-Level Difference Matrices, Tamura, GLRLM features) and concluded that the best result was achieved using GLRLM features.

Arora et al. [2] proposed the use of Convolutional Neural Networks (CNNs) for feature extraction. The ROI is embedded using five networks (AlexNet,

VGG16, ResNet, GoogLeNet, InceptionResNet) and the resulting features are concatenated and fed for the classification module. The advantage of using CNNs is that with the supervised manner of feature extraction mitigates the problem of class imbalance in the dataset. On the other hand, a disadvantage of the proposed method is that the ROI is defined as a minimum bounding box containing the abnormality.

2.4. **Feature selection.** The result of extraction can result in many feature. This can have an impact on the classifier's complexity. Thus, feature selection methods are used to reduce the number of features without too much information loss.

Different unsupervised feature selection methods are compared in a review by Solorio-Fernández et al. [26]. The most widely used unsupervised method for dimensionality reduction is PCA [11]. It also appears in recent studies to filter features extracted from mammograms [5].

In the literature supervised feature selection methods are also investigated. The survey by Chaieb and Kalti [5] also discuses the problem of feature selection. Tabu search, genetic algorithm (GA), ReliefF algorithm, sequential forward selection and sequential backward selection are compared as feature selection methods. They concluded that using GA selection had the best performance.

2.5. **Classification.** Classification is the final step of a CAD system. It aims to distinguish different types of breast tissues (normal/benign/malignant).

Deep Learning (DL) is an extensively researched field of computer science and in recent studies [29] it is used for mammogram classification. Wang et al. [29] compared the performance of six deep learning models: AlexNet, VGG16 and ResNet, two classifiers presented by Shen [23] (one using VGG16 and another with ResNet) and an instance-based learning method using r-CNN. They concluded that the CNN classifiers have a good performance on the training data, but the model can not be generalized to unseen data (regardless of the model's structure).

Nurtanto Diaz et al. [14] used KNN to determine if a lesion is benign or malignant. They used first order features, extracted from the ROI, as input to the classification and achieved an accuracy of 91.8%. Vijayarajeswari et al. [28] proposed the use of SVM classifier and achieved an accuracy of 94%.

Decision trees (DTs) [1] for mammogram classification (benign/malignant) were proposed by Kamalakannan and Rajasekhara Babu [8]. The used input was extracted from the bounding box containing only the abnormality.

3. Proposed approach

The current paper focuses on defining an approach for each of the above identified steps in order to facilitate the creation of a support system for digital mammogram classification. In the following sections the implemented algorithms are presented.

3.1. **Preprocessing.** Mammograms are X-ray images taken from the breast tissue. Most of the mammograms also contain informative labels: the view of the image (MLO or CC – cranio-caudal) and the side (L – left or R – right) from which the image was taken. Besides the labels, small numbers can appear on them as well. The preprocessing's aim is to separate a subset of pixels needed to define whether the breast tissue is ill or not from the rest of the image.

First, the region of the breast has to be defined. The labels and artifacts outside this region are not relevant for a cancer detecting system. Morphological opening and histogram equalization are used in the first step to emphasize image features and to remove noise. After applying simple binarization to the resulting image we select the biggest region (the region of the breast) [25]. The threshold for the binarization is set to 50 (pixels with intensity less than this value does not contain information related to the breast): pixels with value greater than the given threshold will result in 1, otherwise in 0.

Next, we want to remove noninformative regions, which are placed inside the breast's area. This means the removal of the pectoral muscle, a triangular shaped area on the mammogram. It has similar intensity as the lesions. Therefore, with a very high probability the pectoral muscle will be detected as abnormal tissue. Hence, we want to separate it.

First, we implemented a sliding window algorithm proposed by Shrivastava et al. [25]. The method consists in the traversal of the image with a 5×5 window. While (1) the total intensity of the window is greater than a given value ($total\_intensity$) and (2) the difference between the top left and lower right corner is less than another value ($max\_difference$) the content of the window in the resulting image will be set to 0. One drawback of this method is that the intensity of the pectoral muscle and the soft tissue near can vary. Thus, the definition of $total\_intensity$ and $max\_difference$ is not straightforward. To overcome this problem we defined $total\_intensity$ separately for each image based on the intensities in the first window.

The other implemented method is based on seeded region growing (SRG) and it is proposed by Maitra et al. [12]. First, the range, where the pectoral muscle can be located, is decreased. For this four lines are defined: (1) AB on the left-, (2) CD on the right side of the breast, (3) CO connects the top of
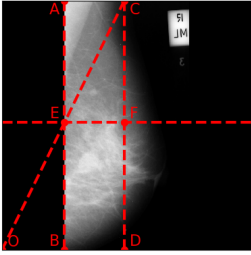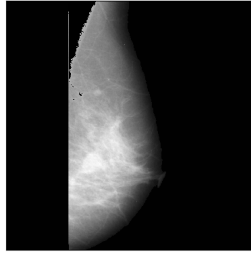
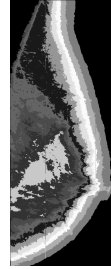FIGURE 1. Guide lines used for SRG.

FIGURE 2. Result of SRG.

FIGURE 3. Segmentation of the ROI with k-means in 11 clusters.

the right side to the lower left corner and defines E (intersection of AB and CO) and (4) EF perpendicular to CO. The lines are marked in Figure 1. In most of the cases the pectoral muscle is inside triangle ACE. As recommended in the article seeds are selected from the the upper half of the first diagonal. After the seeds (S) are defined we calculate the average (S_{avg}) and maximum (S_{max}) intensity of the set. Next, we have to define which pixels to add to the region. For this we recalculate each pixel in triangle ACE by subtracting the average intensity and dividing it with the difference between the maximum and the average ($I' = \frac{I_{x\,y} - S_{avg}}{S_{max} - S_{avg}}$). The pixels with new value from interval $(0, 1]$ will be added to the region. The result will be the mask of the pectoral muscle. Figure 2 shows a mammogram after applying the got mask.

3.2. **Segmentation.** After defining the ROI, the next step towards a CAD system is the segmentation of the breast tissue. For this scope there are solutions in the literature both from the field of supervised and unsupervised learning. In this study we focus on the unsupervised segmentation methods.

We segmented each image with k-means algorithm [11] (see Figure 3). K-means is a clustering method that aims to split n observations into k clusters. The basic steps of the algorithm are (1) calculating the centroid of each cluster and (2) assigning each point from the input to the cluster with the closest centroid. The method stops when the difference between the new and the old cluster centroids falls below a given threshold. There are different researches in the literature on how to define the centroids in the first iteration ([21]). In this experiment we used random initialization. Also, the input for the algorithm in our case is a preprocessed mammogram image.

The application of the algorithm results in a segmentation of the mammogram (example in image Figure 3). We will use this result image together with the original one as input to the second block. Due to the use of clustered

image in feature extraction the system will have more information about the shape of the cancerous tissue.

3.3. **Feature extraction.** Mammograms are gray-scale images. Hence, specific feature extraction is selected. Textural features are calculated from Gray-Level Run-Length Matrices (GLRLM) [5]. In case of a 2-dimensional images four GLRLMs can be calculated for directions: horizontal, vertical, first- and second diagonals. The result matrix will be a 2-dimensional array. As its name suggests the first axis corresponds to the intensity values and the second axis corresponds to the run length values. From each GLRLM 11 features can be derived, characterizing the distribution of short- and long runs in the input image in the specified direction. Chaieb and Kalti [5] included in their review the equations used to calculate these features. Consequently, for each image the descriptor contains $44 = 4 \times 11$ (4 directions and 11 features) elements.

3.4. **Feature selection.** The size of the result of the feature extraction can increase the computational cost, thus the classification progress can show a slow down. Therefore, filtering the best descriptive and independent features is the next step of our system. We propose PCA and a genetic algorithm (GA) based feature selection algorithm.

PCA [11] is the most used feature selection method. It consists in the calculation of covariance matrix from the input data and then applying eigen-decomposition on it. The results of the decomposition will be eigenvalue and -vector pairs, where the higher the eigenvalue the better the descriptiveness of the feature. The principal components are the first $n_{components}$ number eigenvectors with the highest eigenvalues. Using the calculated principal components the input data can be projected into a lower dimensional space.

GAs [11] are nature inspired, stochastic search algorithm. Its basic concept is to start with an initial population (usually randomly defines) and in each iteration create new individuals using genetic operations (selection, crossover and mutation). Each individual is evaluated and the best performing ones are added to the population. The algorithm stops when an individual exceeds a given threshold of fitting the result.

In case of feature selection individuals are binary vectors, where 1 means that the respective feature is selected and 0 otherwise. For the fitness function we selected a classifier (DT) and a performance measure (accuracy score). For each individual a DT is trained and its accuracy will correspond to the individual's fitness value. The method shown in the pseudocode in   algorithm 1 represents the fitness function.

3.5. **Classification.** Classification is the progress of labeling observations based on examples. The built model extracts the characteristics of each class,

---

Algorithm 1 GA - genetic algorithm fitness function for feature selection

---

Require: $estimator, scorer, cv, individual$
Ensure: performance of the individual
 1: **BEGIN**
 2: $total\_metrics \leftarrow 0$
 3: **for** $train\_indices, test\_indices \in cross\_validation\_folds(cv)$ **do**
 4:     X_train, y_train @ define train of features and ground truth based on $train\_indices$
 5:     X_test, y_test @ define train of features and ground truth based on $test\_indices$
 6:     $current\_model \leftarrow estimator.fit(X\_train, y\_train)$
 7:     $prediction \leftarrow current\_model.predict(X\_test)$
 8:     $total\_metrics \leftarrow total\_metrics + scorer(prediction, Y_test)$
 9: **end for**
10: **return** $\frac{total\_metrics}{cv}$
11: **END**

---

thus it will be able to differentiate classes for new inputs. In our system there are three classes: normal, benign and malignant.

In our experiments we considered Decision Trees (DT) [1] which are nature inspired classifiers. DTs are usually represented as binary trees where each leaf represents a predicted class, otherwise each node contains a condition.

The depth of a DT defines the performance of the classifier. However, there is a compromise between the performance and the chance of overfitting. The deeper the tree is the classification on the training data is more accurate, but this can cause poor performance on the test data (caused by overfitting).

K-nearest neighbors [1] is another classification method considered in out experiments. The class of a new observation is defined based on the majority class between its K nearest neighbors. The disadvantage of this method is that in case of high dimensional input the distance between any two points will be one (curse of dimensionality). Hence, feature selection is crucial for this type of classification.

Support Vector Machines (SVM) [1] are widely used classifiers based on statistical learning. The scope of the method is to define a hyperplane, which has the largest distance from the samples of each class (functional margin). For this the support vectors are used (perpendicular sections from the sample point to the plane). The advantage of the this approach is that it works well in high dimension.

4. Experiments

At the beginning of this section we describe the used dataset. In the second part the ran experiments and the achieved results are presented.

4.1. **MIAS.** Mammographic Image Analysis Society contains 161 pair (322) MLO mammograms. From the samples 207 are from healthy breast tissues, 64 are from benign lesions and 54 are from malignant cancerous tissues. This highlights the imbalance of the dataset. It is included a ground truth file

		True class	
		Positive	Negative
Predicted class	Positive	TP	FP
	Negative	FN	TN

TABLE 1. Contingency table, where TP and TN denote true positive/negative samples, while FP/FN denote false positive/negative samples.

in the dataset defined by radiologists. This file specifies the class of each mammogram (normal/benign/malignant).

We used simple train and test split to train our classifiers. Both in the training- and test sets the ration of the classes will be preserved. Due to this property of the split overfitting caused by missing classes from the set is prevented. The split is randomly defined by assigning 75% (241) of the data to the training set and the remaining 25% (81) to the test set.

4.2. **Metrics.** To evaluate the performance of the first block (preprocessing – pectoral muscle removal and segmentation) precision, recall and quality are used. For the second block instead of quality accuracy and f1-score are used. To calculate these values the elements of the contingency table (see table 1) are used, which contains the relation between the ground truth and the prediction.

4.3. **Results.** In this section we discuss the performance of each part in the constructed CAD system. In the experiment we implemented multiple methods, but in the previous sections we discussed the best performing one. In the following paragraphs we present the result of all the used methods.

4.3.1. *Pectoral muscle removal.* For the evaluation of the removal of the pectoral muscle first we need a ground truth. Maitra et al. [12] segmented the pixels in the ACE triangle (Figure 1) and selected the cluster corresponding to the pectoral muscle. These clusters were validated by radiologists and were taken as ground truth. The achieved results are presented in Table 2 (columns 2-4). In our experiment we re-implemented the used methods, but without the validation from radiologists we could achieve the results presented in the same table's last three columns (Table 2 – columns 5-7). Besides SRG we implemented a sliding window algorithm proposed by Shrivastava et al. [25] and the results achieved by our implementation are presented in Table 2 – column 8. This algorithm has a bit better performance than our SRG, but it could not outperform the original method.

4.3.2. *Segmentation.* To evaluate the segmentation, first we have to select the cluster of the abnormality. For this we use the ground truth given in the dataset. The cluster of the lesion will be defined by the maximum number

	Original SRG[12]			SRG			SWA
	Fatty	Gladural	Dense	Fatty	Gladural	Dense	
Precision	0.963	0.978	0.991	0.8125	0.7378	0.7637	0.8125
Recall	0.971	0.975	0.994	0.8930	0.8978	0.8627	0.8930
Quality	0.936	0.954	0.985	0.7415	0.6762	0.6731	0.7415

TABLE 2. Results of the pectoral removal on MIAS

	k						
	4	8	10	12	14	16	18
Precision	0.8370	0.6522	0.5765	0.5230	0.4872	0.4531	0.4212
Recall	0.0771	0.1339	0.1464	0.1556	0.1646	0.1648	0.1763
Quality	0.0692	0.1043	0.1057	0.1068	0.1059	0.1021	0.0990

TABLE 3. Mean measures calculated from mammograms' segmentation containing tumors from MIAS for different number of clusters

of pixels overlapping with the ground truth. After defining this cluster we compare it with the ground truth and we calculate the measures mentioned in Section 4.2. The results are presented in table 3. These numbers are calculated from the mammograms that contain abnormality. As we can see the results are not satisfactory, and with the growth of the cluster number the precision drops faster compared to how recall increases. The cause of this phenomenon is the small number of clusters. With a low k value it is likely that the pixels in the ground truth are part of the same cluster (high precision value). On the other hand, other pixels outside of the ground truth are likely to be selected (low recall value).

4.3.3. *Classification.* In our research the input of the classification consists of the GLRLM features calculated from each image and its segmented version. GLRLM matrices are constructed in all four directions ($0°$, $45°$, $90°$, $135°$). To reduce the dimensionality of the classification's input feature selection is applied. In the experiments the explained variance of the PCA is set to 99%. Consequently, the first two principal components are selected in general. For the evaluation of the feature selection we looked at the system's over all performance. In our approach, we built a single classifier with labels: normal, benign and malignant.

The result of the KNN classifier is shown in figure 4a and figure 4b. Figure 4a shows the performance of KNN classifier, where the five nearest neighbor is considered in the decision making. On x axis the different combinations of feature sets are visible. It shows that the KNN reaches its best performance with features $\{0°, 45°\}$ using PCA feature selection. The accuracy with GA feature selection is a bit below on the training set, but it's better on the test set. On the next figure (figure 4b) the feature set is fixed to $\{0°, 45°\}$ and the result of experiments with the number of neighbors is presented. We can see
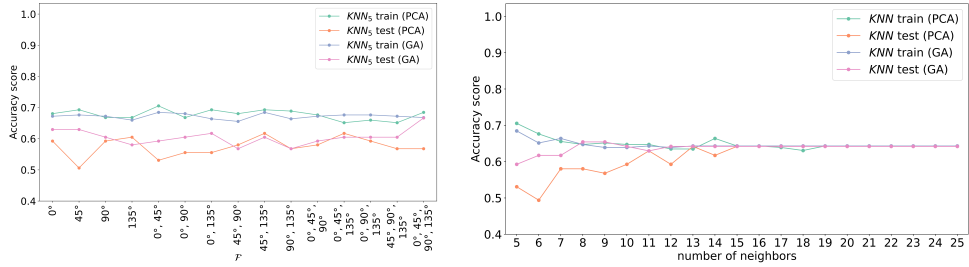
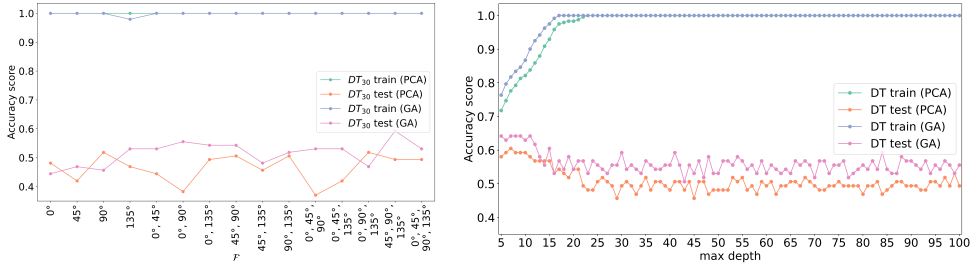(A) Number of neighbors fixed to 5 for all the possible combination of features.

(B) Features fixed to $\{0°, 45°\}$ with different number of neighbors considered.

FIGURE 4. Performance of the KNN classifier.

that from higher than 15 for the number of neighbors in the classification the train an the test result is the same. This can be explained with the fact that the model predicts "normal" for 99% of the images and the ratio of the labels is the same in the train and the test set. Figure 4b shows that KNN on input filtered by GA has lower variance than on input filtered by PCA.

The same experiments were performed with the other classifiers. On figure 5a the performance of the DT model is shown using different feature combinations (using a maximum depth of 30). The result on the training set (with both selection function) is 100%, while on the test set using GA outperforms the results with PCA. The best test result is on set $\{45°, 90°, 135°\}$ using GA (59.2%). Figure 5b shows the result of changing the maximum depth of the DT classifier. It can be seen that with the increase of the maximum depth the accuracy calculated on the training set it is also increasing. On the other hand the metrics calculated on the test set are deceasing. This can be explained with the overfitting generated by the depth of the model.

Table 4 shows the results of the classifications. It can be seen that filtering features with GA and using DT clearly outperforms the rest of the classifiers on the training data. On the test set the difference in the metrics is smaller, and the combination PCA-KNN, GA-SVM have the highest accuracy. As mentioned above, this can be caused by overfitting in the DT because the best train accuracy is achieved when the depth of the tree is 30. The best performance, considering both the train and test results, was achieved by using GLRLM features for directions $\{45°, 90°, 135°\}$, GA feature selection with DT and accuracy used in its fitness function and 30 as the depth of the tree. With this setting the train accuracy and test accuracy was 100% and

(A) Maximum depth fixed to 30 for all the possible combination of features.

(B) Features fixed to $\{45°, 90°, 135°\}$ with different maximum depth considered.

FIGURE 5. Performance of the DT classifier.

		PCA				GA			
train	KNN	0.6556	0.6357	0.6556	0.5568	0.6722	0.7061	0.6722	0.5800
	SVM	0.6515	0.7316	0.7734	0.6197	0.6473	0.7296	0.7685	0.6089
	DT	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
test	KNN	0.6420	0.5211	0.6420	0.5410	0.6049	0.4866	0.7206	0.5809
	SVM	0.6173	0.4840	0.7353	0.5837	0.6420	0.6420	1.0000	0.7820
	DT	0.4938	0.5048	0.4938	0.4976	0.5926	0.5853	0.5926	0.5884

TABLE 4. Results of the feature selection and classification combinations for MIAS. From each feature selection group the columns correspond to accuracy, precision, recall and f1-score respectively.

59.2% respectively for classifying the mammograms into normal, benign and malignant classes.

4.4. **Discussions.** Based on the results shown above, from the investigated methods the combination of SRG, GA and DT achieved the best result for a breast cancer detecting CAD system. The reported test accuracy for classification in three classes (normal, benign and malignant) is 59.2%. In this section we compare the results of methods from the state-of-the-art using the same database, feature extraction, -selection or classification.

Srivastava et al. [27] proposed the use of GA feature selection on features extracted from mammograms in MIAS. They used histogram, texture, geometric, wavelet and Gabor features, and after feature selection an SVM-MLP is used to classify the data into normal and abnormal classes. They reported an accuracy of 87%. A combination of GA and RF was proposed by Rouhi et al. [17] on GLCM, shape and intensity histogram features. They achieved a result of 74.44% on classifying images in benign and malignant classes.

Recent researches using GLRLM features used different label in the classification or multiple classifiers compared to the proposed approach. Candra et al. [4] built three models to classify mammograms in normal, benign and malignant classes. They used SVM models with different kernels, but the best result (93.97%) was reported using polynomial kernel.

The mentioned result above can not be directly compared to the proposed approach because the aim of the classification is different. The deviation of the result metrics can be caused because the difference in the train test split or in the used dataset (MIAS or DDSM).

5. Conclusions and future work

The scope of the current paper was to construct a system that helps detecting breast cancer by classifying mammograms into normal, benign and malignant categories. This process involved the usage of preprocessing, segmentation, feature extraction and selection. Based on the exhaustive experiments conducted on methods for the steps of image classification it is concluded the best performance achieved is by using a combination of SRG, GLRLM $\{45°, 90°\}$, GA and DT with a 100% training accuracy and 59.2% test accuracy.

In the current approach MLO mammograms are used as input to the presented system. Therefore, it will not work with CC images. In future work other preprocessing methods and ROI definition are planned to be investigated. In future work, investigating pruning might improve the performance of DT classifiers by reducing overfitting. We will further investigate the use of other feature selection methods and classifiers (for instance neural networks). In the current paper, we built a single classifier. Hence, in a future work we can build two binary classifiers (one to decide if the breast tissue is normal or abnormal and a second classifier to decide if the lesion is benign or malignant) equivalent to the proposed one. Furthermore, cross validation is planned to be investigated instead of simple train test split. MIAS is an unbalanced dataset, therefore investigating different balancing techniques might increase the performance of the system. In addition, in furute work a detailed comparative analysis will be included between the methods presented in the literature and the proposed one.

References

[1] AGGARWAL, C. C. *Data Classification: Algorithms and Applications*, 1st ed. Chapman & Hall/CRC, 2014.

[2] ARORA, R., RAI, P. K., AND RAMAN, B. Deep feature–based automatic classification of mammograms. *Medical & Biological Engineering & Computing 58*, 6 (June 2020), 1199–1211.

[3] BALI, A., AND SINGH, S. N. A review on the strategies and techniques of image segmentation. In *Proceedings of the 2015 Fifth International Conference on Advanced Computing & Communication Technologies* (USA, 2015), IEEE Computer Society, p. 113–120.

[4] CANDRA, D., NOVITASARI, R., LUBAB, A., ET AL. Application of feature extraction for breast cancer using one order statistic, GLCM, GLRLM, and GLDM. *Advances in Science, Technology and Engineering Systems Journal 4*, 4 (2019), 115–120.

[5] CHAIEB, R., AND KALTI, K. Feature subset selection for classification of malignant and benign breast masses in digital mammography. *Pattern Analysis and Applications 22*, 3 (Aug. 2019), 803–829.

[6] ESENER, I. I., ERGIN, S., AND YUKSEL, T. A novel multistage system for the detection and removal of pectoral muscles in mammograms. *Turkish Journal of Electrical Engineering and Computer Sciences 26* (2018), 35–49.

[7] HTAY, T. T., AND MAUNG, S. S. Early stage breast cancer detection system using glcm feature extraction and k-nearest neighbor (k-nn) on mammography image. In *18th International Symposium on Communications and Information Technologies* (2018), pp. 171–175.

[8] KAMALAKANNAN, J., AND BABU, M. R. Classification of breast abnormality using decision tree based on GLCM features in mammograms. *International Journal of Computer Aided Engineering and Technology 10*, 5 (2018), 504–512.

[9] KIM, W., KANEZAKI, A., AND TANAKA, M. Unsupervised learning of image segmentation based on differentiable feature clustering. *IEEE Transactions on Image Processing 29* (2020), 8055–8068.

[10] LI, H., CHEN, D., NAILON, W. H., ET AL. Dual convolutional neural networks for breast mass segmentation and diagnosis in mammography, 2020.

[11] LIU, H., AND MOTODA, H. *Computational Methods of Feature Selection*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press, 2007.

[12] MAITRA, I. K., NAG, S., AND BANDYOPADHYAY, S. K. Technique for preprocessing of digital mammogram. *Computer Methods and Programs in Biomedicine 107*, 2 (2012), 175–188.

[13] MOGHBEL, M., OOI, C. Y., ISMAIL, N., ET AL. A review of breast boundary and pectoral muscle segmentation methods in computer-aided detection/diagnosis of breast mammography. *Artificial Intelligence Review 53*, 3 (Mar. 2020), 1873–1918.

[14] NURTANTO DIAZ, R. A., NYOMAN TRIA SWANDEWI, N., AND PRADNYANI NOVIANTI, K. D. Malignancy determination breast cancer based on mammogram image with k-nearest neighbor. In *2019 1st International Conference on Cybernetics and Intelligent System* (2019), vol. 1, pp. 233–237.

[15] RAHIMETO, S., DEBELEE, T. G., YOHANNES, D., ET AL. Automatic pectoral muscle removal in mammograms. *Evolving Systems* (Nov. 2019).

[16] RASHED, E. A., AND AWAD, M. G. Neural networks approach for mammography diagnosis using wavelets features, 2020.

[17] ROUHI, R., JAFARI, M., KASAEI, S., ET AL. Benign and malignant breast tumors classification based on region growing and CNN segmentation. *Expert Systems with Applications 42*, 3 (2015), 990–1002.

[18] SADEGHI, B., KARIMI, M., AND MAZAHERI, S. Automatic suspicions lesions segmentation based on variable-size windows in mammography images. *Health and Technology 11*, 1 (Jan. 2021), 99–110.

[19] SALAMA, M. S., ELTRASS, A. S., AND ELKAMCHOUCHI, H. M. An improved approach for computer-aided diagnosis of breast cancer in digital mammography. In *IEEE International Symposium on Medical Measurements and Applications* (2018), pp. 1–5.

[20] SARKER, O., AKTER, S., AND MISHU, A. A. Review on the performance of different types of filter in the presence of various noises. *Engineering International 4*, 2 (dec 2016), 49–56.

[21] SAXENA, A., WANG, J., AND SINTUNAVARAT, W. An empirical study on initializing centroid in k-means clustering for feature selection. *International Journal of Software Science and Computational Intelligence 13*, 1 (jan 2021), 1–16.

[22] SELVATHI, D., AND AARTHY POORNILA, A. *Deep Learning Techniques for Breast Cancer Detection Using Medical Image Analysis.* Springer International Publishing, Cham, 2018, pp. 159–186.

[23] SHEN, L., MARGOLIES, L. R., ROTHSTEIN, J. H., ET AL. Deep learning to improve breast cancer detection on screening mammography. *Scientific Reports 9*, 1 (Aug. 2019).

[24] SHINDE, V., AND THIRUMALA RAO, B. Novel approach to segment the pectoral muscle in the mammograms. In *Cognitive Informatics and Soft Computing* (Singapore, 2019), pp. 227–237.

[25] SHRIVASTAVA, A., CHAUDHARY, A., KULSHRESHTHA, D., ET AL. Automated digital mammogram segmentation using dispersed region growing and sliding window algorithm. In *2nd International Conference on Image, Vision and Computing* (June 2017), pp. 366–370.

[26] SOLORIO-FERNÁNDEZ, S., CARRASCO-OCHOA, J. A., AND MARTÍNEZ-TRINIDAD, J. F. A review of unsupervised feature selection methods. *Artificial Intelligence Review 53*, 2 (Feb. 2020), 907–948.

[27] SRIVASTAVA, S., SHARMA, N., SINGH, S., ET AL. Quantitative analysis of a general framework of a CAD tool for breast cancer detection from mammograms. *Journal of Medical Imaging and Health Informatics 4*, 5 (Oct. 2014), 654–674.

[28] VIJAYARAJESWARI, R., PARTHASARATHY, P., VIVEKANANDAN, S., ET AL. Classification of mammogram for early detection of breast cancer using SVM classifier and hough transform. *Measurement 146* (2019), 800–805.

[29] WANG, X., LIANG, G., ZHANG, Y., ET AL. Inconsistent performance of deep learning models on mammogram classification. *Journal of the American College of Radiology 17*, 6 (2020), 796–803.

BABEŞ-BOLYAI UNIVERSITY, FACULTY OF MATHEMATICS AND COMPUTER SCIENCE, 1 MIHAIL KOGĂLNICEANU, CLUJ-NAPOCA 400084, ROMANIA
Email address: adel.bajcsi@ubbcluj.ro