

Deep learning versus classical neural approach to mammogram recognition

J. KUREK¹, B. SWIDERSKI¹, S. OSOWSKI^{2*}, M. KRUK¹, and W. BARHOUMI³

¹Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences, 166 Nowoursynowska St., 02-787 Warsaw, Poland

²Faculty of Electrical Engineering, Warsaw University of Technology and Faculty of Electronic Engineering, Military University of Technology, Warsaw, 75 Koszykowa St., 00-662 Warsaw, Poland

³Research Team on Intelligent Systems in Imaging and Artificial Vision (SIIVA) – LimTic Laboratory, ISI, University of Tunis El Manar, Tunisia

Abstract. Automatic recognition of mammographic images in breast cancer is a complex issue due to the confusing appearance of some perfectly normal tissues which look like masses. The existing computer-aided systems suffer from non-satisfactory accuracy of cancer detection. This paper addresses this problem and proposes two alternative techniques of mammogram recognition: the application of a variety of methods for definition of numerical image descriptors in combination with an efficient SVM classifier (so-called classical approach) and application of deep learning in the form of convolutional neural networks, enhanced with additional transformations of input mammographic images.

The key point of the first approach is defining the proper numerical image descriptors and selecting the set which is the most class discriminative. To achieve better performance of the classifier, many image descriptors were defined by means of applying different characterization of the images: Hilbert curve representation, Kolmogorov-Smirnov statistics, the maximum subregion principle, percolation theory, fractal texture descriptors as well as application of wavelet and wavelet packets. Thanks to them, better description of the basic image properties has been obtained. In the case of deep learning, the features are automatically extracted as part of convolutional neural network learning. To get better quality of results, additional representations of mammograms, in the form of nonnegative matrix factorization and the self-similarity principle, have been proposed. The methods applied were evaluated based on a large database composed of 10,168 regions of interest in mammographic images taken from the DDSM database. Experimental results prove the advantage of deep learning over traditional approach to image recognition. Our best average accuracy in recognizing abnormal cases (malignant plus benign versus healthy) was 85.83%, with sensitivity of 82.82%, specificity of 86.59% and AUC = 0.919. These results are among the best for this massive database.

Key words: convolutional neural networks, breast cancer diagnosis, mammogram recognition, diagnostic features.

1. Introduction

Breast cancer belongs to the most dangerous cancers affecting women. More than 18% of all cancer deaths, both in males and females, are from breast cancer. Over 1.67 million new cases were in 2012 worldwide [1]. Early detection of cancer is crucial for treatment, since it translates into better perspectives for recovery.

Screening mammography programs are organized to cope with the problem and to reduce the mortality rates [2, 3]. However, mammography interpretation is a difficult task due to the subtle signs of breast abnormalities which can be observed at an early stage. According to statistics, 10–15% of cancer cases are still left undetected.

Due to the huge amount of screening mammograms, which should be analyzed by two independent experts, and due to the limited number of expert radiologists, a bottleneck forms in all screening programs. Therefore, computer aided detection (CAD) systems are urgently required. Such systems could replace the second reader and alert the expert radiologist as to the suspicious

regions. However, accuracy of the systems developed to date remains unsatisfactory. Different solutions have been applied to computer aided mammogram recognition. They differ by the image preprocessing stages, which lead to different diagnostic features, and also by the solutions of classification systems used in the recognition of patterns formed by these features.

Paper [4] reviews different methods of feature definition and application of classification tools. The diagnostic features are based on characterization of the texture, edge orientation, statistical analysis of a map of pixels in the mammographic image, etc. Different mathematical tools are used to define these features. They include wavelet decomposition, mathematical morphology, thresholding methods, template matching, neural networks and many others. Paper [4] presents a comparison of actual results of different approaches to distinguishing between normal and abnormal mammograms, obtained for limited numbers of mammograms (from 128 to 280). However, the quality factors defined in the form of true positive rate TPR = 75.7%, false positive rate FPR = 73.5% and AUC (area under ROC curve), oscillating for different solutions between 0.76 and 0.89, were not satisfactory. Meanwhile, paper [5] presents application of an extreme learning machine to tumor detection in double-views mammography.

Most research presented in the literature used only small databases of mammographic images. Paper [6] presents the ap-

*e-mail: sto@iem.pw.edu.pl

Manuscript submitted 2017-12-15, revised 2018-02-26 and 2018-04-19, initially accepted for publication 2018-04-19, published in December 2018.

plication of principal and independent component analyses to generating diagnostic features and using a radial basis function network as a classifier. The accuracy rate of 88.23% in detection of all types of abnormalities in the analyzed 119 regions of suspicion for mammogram images in the Mini Mammographic Database of MIAS has been reported. In [7], features based on estimation of the probability density function of the gray-level differences in the image were defined. After applying the genetic algorithm and forward sequential selection, these features have been used as the input signals for the multilayer perceptron operating in the classification mode. The classification accuracy of 89%, with 88.6% sensitivity and 83.3% specificity, have been reported for 410 mammograms from the Digital Database for Screening Mammography (DDSM). The 600 cases taken from the DDSM were analyzed in [8] using three different methods of feature problem solution: genetic algorithm, greedy selection and random mutation hill climbing. Different commercial CAD products for mammography analysis, including AccuDetect Parascript® [9], R2 ImageChecker and iCAD Second Look [10], have been tested in recognizing the abnormal cases. It was shown that all of them suffer from limited accuracy. The best results of AUC stood at 0.789.

In [11], the recognition results of abnormality cases in all mammograms from the DDSM base by using the curvelet moments was presented. Only the accuracy rate was reported. It changed from 81.26% to 86.46%, depending on the feature set applied. However, no sensitivity, specificity or AUC information have been presented. In [12], the application of deep learning to the recognition of mammograms was proposed.

The aim of this work is to develop and compare two new approaches to mammographic image recognition, able to recognize the abnormal cases (benign + malignant) from normal ones with an increased accuracy. Both will be used to analyze the regions of interest (ROI) in the mammograms. The first approach consists in typical steps used in classical pattern recognition: generation of numerous numerical image descriptors, selection of the most discriminative ones, which will serve as diagnostic features for the classifier, and the final classification step involving the support vector machine (SVM). To get the most objective and independent description of the image, we have proposed different feature extraction methods. They include representation of the image by means of the Hilbert curve and definition of special descriptors based on the self-similarity of vectors, Kolmogorov-Smirnov statistics, maximum subregion principle, percolation theory, the gray-level co-occurrence matrix (GLCM) analysis, fractal texture description as well as application of wavelet and wavelet packets in creating numerical descriptors. To the best of our knowledge, most of them are applied for the first time in mammographic image analysis. In the next step, a sequential feature selection method is used to choose the most class-discriminative subset of features. The SVM has been applied in the classification step.

In the second approach, we will use the deep learning strategy based on the convolutional neural network (CNN) as the workhorse. CNN plays the role of the unsupervised feature selection and final classification mechanism at the same time. However, direct application of the set of mammograms avail-

able in the DDSM base to the CNN is not fully successful due to the limited number of sample images. Therefore, we propose to expand the input data by providing additional images created by applying non-negative matrix factorization (NMF) and statistical self-similarity. They fulfill the significant role in the classification system and allow to increase the accuracy of image recognition.

The numerical experiments have been performed on a large DDSM database containing more than 10,000 mammograms. The results of these investigations have confirmed good accuracy of class recognition. A comparison of the classical and deep learning approaches has shown the advantage of the deep learning strategy. The main contribution of this work is thus as follows:

- Proposition and application of novel methods for extracting the numerical descriptors of mammographic images in the classical neural approach to image recognition. Diversity of descriptions allows characterizing details of the images from many different points of view.
- Successful application of deep learning strategy, in the form of the convolutional neural network, to the analysis of mammographic image. The important element in this representation is the application of non-negative matrix factorization and statistical self-similarity, which are capable of enhancing the differences between classes of mammograms and increasing the accuracy of class recognition this way.
- Experimental application of the proposed solution to the DDSM set of mammograms and proving its better performance in comparison to other results presented in different papers to date. Our best average accuracy in recognizing abnormal cases from normal ones was 85.83%, with sensitivity of 82.82%, specificity of 86.59% and $AUC = 0.919$.

These results are one of the best obtained so far for this set. The rest of the paper is organized as follows. Section 2 provides a brief description of the database of the mammograms used. Section 3 presents the classical approach to image recognition and the results of numerical experiments. Section 4 is devoted to the deep learning approach to mammogram recognition. Section 5 compares the obtained results using both methods along with the others, reported in previous publications. The concluding section summarizes the considerations presented herein.

2. Database of mammograms applied in the investigations

The numerical investigations have been carried out using the largest publically available database of mammographic images, i.e. the “Digital Database for Screening Mammography” [13]. It is composed of 2,604 cases, each containing 4 mammograms (left and right breast from above, representing the cranial-caudal view, and an oblique image, representing the medio-lateral-oblique view). The dataset contains important information on each mammogram, including its diagnostic results (normal, benign or malignant) and the location of existing lesions, forming the ROI. For the abnormal cases (benign and malignant ones), manual cropping was done based on the in-

formation provided in the ground truth. The ROI corresponding to the masses observed represents a rectangular area with the lesion in the center. In the normal cases the ROI was extracted manually by the medical expert from normal tissues. The size of ROI images was the same and equal to 128×128 pixels, irrespective of their type. The number of ROI images that has been used in the experiments was 10,168. The DDSM database contained the following number of class representations:

- a. Normal tissue: 8,254,
- b. Benign lesions: 862,
- c. Malignant lesions: 1,052.

The above means that the abnormal tissue set, representing the benign and malignant cases, contains only 1,914 ROI images, distinctly less than the normal ones (8,254 samples). This presents some additional problems related to the unbalanced set of data.

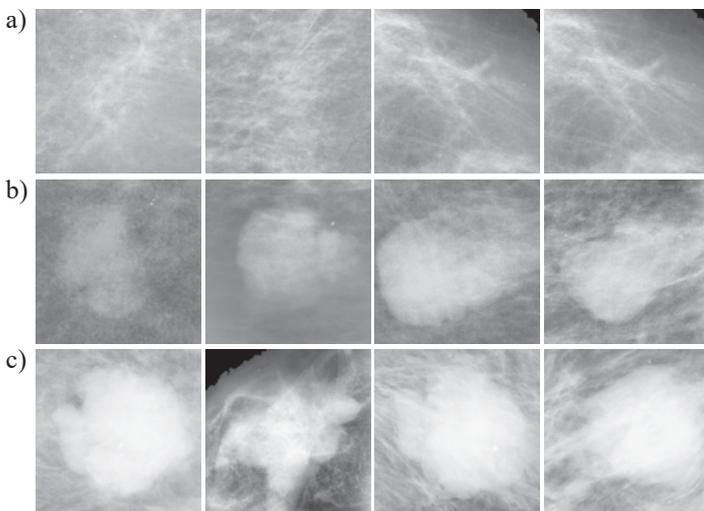


Fig. 1 ROI examples of mammograms representing normal (a), benign (b) and malignant (c) cases

This problem was solved by splitting the set of normal cases into 4 subsets, each confronted in classification with the same set of abnormal cases with applying the majority voting rule.

Figure 1 presents the examples of mammograms of normal (Fig. 1a) and abnormal cases: benign (Fig. 1b) and malignant (Fig. 1c) ones. We can observe the significant differences among the images representing the same class of data and close similarity of images representing different classes. This results in significant problems at the stage of class recognition.

3. Classical neural approach to mammogram recognition

Three independent steps are usually applied in the classical neural approach to image recognition: extraction of the numerical image descriptors, selection of the best set of the class discriminative diagnostic features and the classification step responsible for the final recognition of classes.

3.1. Definition of numerical descriptors of mammographic images. The main problem in efficient numerical characterization of mammographic images is their diversity inside the same class of images and close similarity between normal and abnormal tissues. To cope with this problem, we propose application of different mechanisms of feature definition, characterizing the image from different points of view. The methods applied will refer to characteristics of chaotic systems using fractal measures, texture description using the co-occurrence Haralick's matrix and Kolmogorov-Smirnov statistics, self-similarity of images, percolation theory, different types of statistical description as well as description based on wavelet representation. The following methods are used to generate numerical image descriptors:

- description based on Hilbert's curve representation of the image,
- statistical description based on the coaxial rings image representation and their characterization by applying the Kolmogorov-Smirnov distance,
- maximum subregion principle,
- description based on percolation theory,
- texture description based on the gray-level co-occurrence matrix,
- application of self-similarity principle of the image in connection with the box-counting dimension,
- segmentation-based fractal texture analysis,
- application of wavelet and wavelet packet decomposition.

3.1.1. Kolmogorov-Smirnov descriptors. Kolmogorov-Smirnov (KS) descriptors belong to the statistical parameters. They are defined on the basis of pixel intensity in the coaxial rings of the increasing diameters [14]. The successive regions of the image are split into a number of concentric rings around the central point. Individual regions contain approximately equal numbers of pixels in each ring. The central point travels around the whole image. In each of its positions, the KS statistics, describing the difference between the pixel populations in the rings placed at equal distances from each other, are estimated. The KS statistics check if the pixels belonging to two rings belong to the same population. KS distance is defined on the basis of their cumulative distributions, $F(x_i)$ and $F(x_j)$:

$$d_{KS} = \max |F(x_i) - F(x_j)| \quad (1)$$

over all x . This distance represents the measure of difference between the pixel statistics in both rings.

Four coaxial rings have been constructed for each mammographic image. Every coaxial ring contains approximately the same number of pixels. The sets of KS distances corresponding to the combinations of these four levels have been estimated. Level 1 represents KS distance of two successive rings, i.e. rings 1 and 2, 2 and 3, 3 and 4, etc. Level 2 describes the statistics of rings distant by 2, for example 1 and 3, 2 and 4. The cumulative mean and median values of KS distance between the intensity of pixels belonging to two different rings, generated over the whole image, have both been estimated. The functions contrasting the mean and median values of KS

distance d_{KS} with the level l are linearly approximated in the following forms:

$$mean_d_{KS} = \alpha_{0mean} + \alpha_{1mean}l + \varepsilon \quad (2)$$

$$med_d_{KS} = \alpha_{0med} + \alpha_{1med}l + \varepsilon \quad (3)$$

where α_0 and α_1 are the regression coefficients corresponding to equations (2) and (3). The following KS parameters were used for description of the image:

d_{KS12} (mean and median values of KS distances between rings 1 and 2),

- d_{KS13} (mean and median values of KS distances between rings 1 and 3)
- d_{KS14} (mean and median values of KS distances between rings 1 and 4)
- the d_{KS13}/d_{KS12} ratio in mean and median representation
- the d_{KS14}/d_{KS12} ratio in mean and median representation
- coefficient α_{0mean} and α_{0med} of linear approximations (2) and (3)
- slope coefficient α_{1mean} and α_{1med} of linear approximations (2) and (3)

This way, we have obtained 14 descriptors resulting from KS statistics.

3.1.2. Maximum subregion descriptors. The main idea behind this method is to observe the process of disaggregating the image into smaller consistent subgroups by using thresholding at different values of bias [14]. The process of splitting aims to find the level of thresholding which provides the largest number of consistent subgroups. Many thresholding processes are performed on the image to achieve the goal.

In the searching procedure we apply the idea of quantile representation of pixel's intensity, i.e. 0.01, 0.02, ..., 0.99. We search for quantile q and its corresponding intensity threshold value th_q , which splits the image into the largest number of compact groups of pixels (the group is understood as the compact area isolated completely from the other pixels). The value of quantile q and its normalized threshold nth_q will form the diagnostic features.

The normalized threshold is defined as $nth_q = (th_q - f_1) \frac{255}{f_{99} - f_1}$, where f_1 is the lowest intensity level of the pixels corresponding to the first quantile and f_{99} is the intensity level corresponding to 99th quantile. The third descriptor takes the form of the relative area of the largest compact subgroup of pixels in the image after thresholding. For two types of sub-images after thresholding (the sub-image of pixel intensity higher or lower than the assumed threshold value), the number of these features is duplicated (six descriptors in total).

3.1.3. Percolation descriptors. Percolation descriptors focus on differences in the complexity of the borders (smoothness, raggedness, etc.) of the structure formed by the pixels in the image being analyzed. The image is first binarized into multiple sub-images using different threshold values and then "fire" is set to each segment [14, 15]. In each iteration the pixels adja-

cent to the region under fire enlarge the fired area. The number of iterations needed to illustrate the whole image at different binarization thresholds are determined. This process is performed on an image resized to the dimension of 1024×1024 . In the first phase, the image is covered by horizontal and vertical lines, located at every 100 pixels. The fire, initiated in each node created by the crossing points of horizontal and vertical lines, spreads simultaneously in all directions (horizontal, vertical and diagonal one). The process is repeated simultaneously on all sub-images, which are obtained by means of binarization generated by different threshold values. The more jagged the image, the longer the fire duration. The threshold values are changed step by step in the intensity range [0–255] of the pixels, according to the decile steps from $q = 1$ up to $q = 9$. The fire duration (measured by the number of iterations) is registered for each threshold value. The percolation descriptor of the image is assumed to take the form of the weighted average measure q_w of quantiles, defined as follows:

$$q_w = \frac{\sum_{i=1}^9 q_i d_i}{\sum_{i=1}^9 d_i} \quad (4)$$

where q_i is the quantile changing from 0.1 to 0.9 by means of 0.1 steps, and d_i is the number of iterations of the fire at the threshold value corresponding to the i th decile. The segmentation is repeated many times on the sub-images formed in the thresholding process, assuming pixel intensity higher or lower than the assumed threshold value. This results in two numerical descriptors q_w , corresponding to these two percolation processes.

3.1.4. GLCM texture descriptors. GLCM texture description is a well-known approach to characterization of images. It is based on the co-occurrence matrix [16], which reflects statistical relationships between the intensity of the neighboring pixels in the image. In this particular application, our texture characterization is limited to four statistical descriptors of the co-occurrence matrix of the image. They include: local contrast of the image, which characterizes the intensity difference between a pixel and its neighbors over the whole image, correlation existing between different pixel pairs, energy representing the occurrence of repeated pairs in the image, and the homogeneity coefficient, the latter characterizing the distribution of elements in GLCM matrix.

3.1.5. Statistical image descriptors. Statistical image descriptors have been created directly on the basis of the pixel intensity level. They include the mean, median, standard deviation (std), kurtosis, minimum, maximum, cumulants of the second, third and fourth orders, the ratio of the difference of 0.75 and 0.25 quantiles related to the maximum of the median (or the value 0.001 if the maximum is less than 0.001) and the std to maximum ratio (or the value 0.001 if the maximum is less than 0.001). This way, the total number of these descriptors stands at 11.

3.1.6. Self-similarity descriptors. This family of descriptors is the generalization of the box counting dimension applied to the gray scale image. The original ROI image, resized to the dimension of 1024×1024 , is first covered by the grid of horizontal and vertical lines, separating it into $s \times s$ small regions. In the next step, the similarity of each region to the whole image is estimated. This is done by using statistics of Kolmogorov-Smirnov distance d_{KS} [17]. The higher the value of this distance, the lower the similarity index of the analyzed subregion to the whole image. After performing such calculations for all regions of the original image, a new image, $n \times n$ in size, is created. The ij th element of this image represents the similarity of this particular region to the whole image and is described by $y_{ij} = 1 - d_{KS}$. All similarity values are in the range of $[0, 1]$.

Three different grids have been applied: 64×64 , 128×128 and 256×256 . Each of them generates the corresponding self-similarity images described by matrices of appropriate size. The next step is similar to the classical box-counting dimension of fractals [18]. The sum of elements corresponding to the appropriate matrices is calculated. At the three sizes of the grid applied, we get three pairs of points, representing the scale s (here $s = 64, 128, 256$) and the sum $N(s)$ of the values of elements in the corresponding matrix. Linear regression in logarithmic scale is estimated for these results:

$$\log_2(N(s)) = a \log_2(s) + b + \varepsilon. \quad (5)$$

The slope a and intercept point b represent two image descriptors. The next 6 descriptors represent the mean value and standard deviation of the self-similarity matrices corresponding to the following sizes: 64×64 , 128×128 and 256×256 . The total number of these descriptors is 8.

3.1.7. Segmentation-based fractal texture descriptors. This method generates descriptors on the basis of the multi-thresholding level Otsu algorithm and is referred to shortly as SFTD [20]. The image is binarized using different pairs of upper and lower threshold values, which are selected by using the so-called two threshold binary decomposition technique. Then the recursive algorithm is applied to each image region until the desired number of threshold values n is obtained, where n is the user-defined parameter. As a result, the image is decomposed into a set of binary images. The more jagged the edges of the segmented regions, the higher their fractal dimension. Therefore, the box-counting dimension of boundaries is a good candidate for being the numerical descriptor characterizing the image. Two additional descriptors are defined in the form of the size of and mean gray-level of the sub-images. For n threshold values, the number of descriptors equals $3n$. In this application, we have used 12 threshold values selected in this manner. As a result, 36 numerical image descriptors have been defined.

3.1.8. Hilbert's descriptors. The Hilbert space-filling method, known otherwise as the Hilbert curve, is a continuous fractal space-filling curve providing mapping between the 1D and 2D space that preserves local regions of the image [19–21] fairly effectively. The 1-D Hilbert curve of the image represents pixel

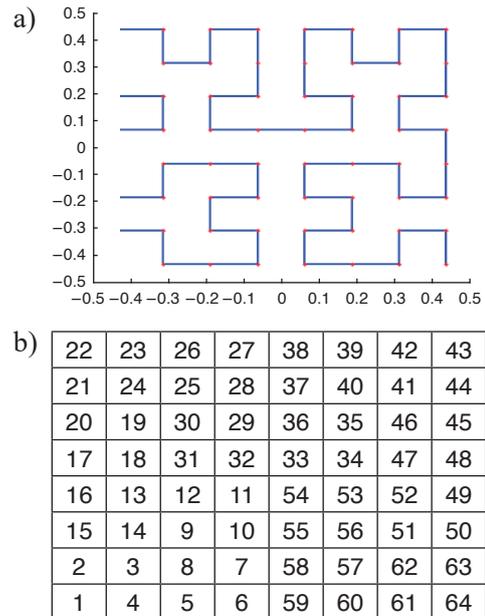


Fig. 2 Example of a Hilbert curve (a) and the order of 64 pixels in a 1-D vector representation of the image

intensity in the points specified by the nodes as shown in Fig. 2, where we have limited representation for the grid that is 8×8 in size.

In this work the Hilbert representation of the mammogram, containing 1,024 elements, has been used. As a result of such representation, each analyzed image has been substituted by its vector form, 1,024 in length.

The family of descriptors is defined using the KS statistics estimated for two Hilbert sub-vectors traveling along the Hilbert curve with a step that equals one. Both vectors have the length of 256 elements and occupy neighboring positions in space. KS distance d_{KS} for each position of these two vectors is estimated. On the basis of these distances and the corresponding significance levels p , additional statistical descriptors are defined. They include the mean, median, std values of d_{KS} and p and also the corresponding ratio of std-to-mean calculated for both parameters. The set is supplemented by the values of 0.25 and 0.75 quantiles along with their differences and the ratio of their difference related to the maximum of the median at the assumed significance level of 0.001 estimated for both parameters. 16 Hilbert descriptors have been defined this way.

3.1.9. Multiscale wavelet transform descriptors. The Hilbert curve of the image is transformed into wavelet decomposition [22] and represented by means of detailed coefficients on different levels and the residual signal on the last level of decomposition. Each detailed and residual signal has been characterized by 4 parameters: energy, variance, standard deviation and waveform length. In this particular application, db4 wavelet and 10 levels of decomposition have been used. The wavelet function type and number of decomposition levels have been selected after a series of introductory experiments, in which Fisher discriminant measure was used to assess the quality of resulting descriptors.

The procedure results in 11 waveforms, representing 10 detailed coefficients and one residual signal. Since each waveform is characterized by 4 parameters, we have obtained 44 descriptors.

3.1.10. Wavelet packet descriptors. The wavelet packet decomposition was used to form the next set of descriptors [23]. It was applied also to the Hilbert vector form of the image. The db4 wavelet family and two levels of decomposition were used in the numerical experiments. As a result, 16 vectors, representing the details on four levels, and residual vectors for the last, fourth level, have been obtained. Each vector is characterized by the energy of its elements, i.e. $E_k = \sum_i (x_i^{(k)})^2$ for $k = 1, 2, \dots, 16$, where $x_i^{(k)}$ represents the value of the element in i -th position in k th detail or the residual vector. These values create the set of 16 wavelet descriptors.

3.2. Feature selection. The total number of descriptors defined in the previous sections equals 157. However, not all of them represent equally good features in class discrimination. Therefore, selection is required in order to come up with a set of the best diagnostic features representing the highest class recognition ability. The sequential forward and backward selection method [17, 24] has been applied. This approach was due to high effectiveness and relatively quick performance. As a result of its application, a specific set of optimum features is generated. This is in contrast with other methods, based on different informative or correlation measures.

Individual descriptors are added and removed from the current feature set in the selection process. After including or removing a feature, the newly created set of features is checked for class prediction accuracy. If the added or removed descriptor has increased the accuracy of the resulting set, the operation is accepted, otherwise it will be discarded [17]. In the process of checking the class discrimination ability of the current feature set, the support vector machine of the radial kernel was used as the classifier. For every candidate feature subset, the sequential feature selection was applied using 10-fold cross-validation, by repeatedly calling the function with different training subsets of learning data and the changing the validation subset of data. As a result of such a selection process, we get the logical vector indicating which features are finally chosen by means of the selection procedure.

Only 39 diagnostic features out of the 122 descriptors generated in the initial image description have been left after this selection process. The composition of the selected feature set included representatives of all types of descriptions. Among the selected features there were 9 representatives of multiscale wavelet transformation, 8 Hilbert descriptors, 5 wavelet packet descriptors, 4 SFTA descriptors, 3 percolation descriptors, 3 statistical descriptors, 3 fractal descriptors, 2 Haralick texture descriptors, one maximum subregion and one KS descriptor.

3.3. Results of numerical experiments. The dataset was split into 10 subsets, each containing the same proportion of both classes, related to their populations in the database. Nine parts of samples are used in the feature selection and learning the SVM classifier and the last one is used for testing the learned

system. The same experiments have been repeated ten times, exchanging the testing and learning subsets. To balance the number of classes in each experiment, the normal class was split into four parts, each associated with the same abnormal cases and the results were averaged. The training and testing sets have been chosen randomly from the database.

Different classifiers, including SVM, multilayer perceptron, decision tree and random forest, have been tried in the introductory experiments. However, best results have been obtained for SVM of the radial Gaussian kernel: $K(\mathbf{x}, \mathbf{x}_i) = \exp(-\gamma \|\mathbf{x} - \mathbf{x}_i\|^2)$ of $\gamma = 0.1$ at application of the regularization constant $C = 1000$, and only these results will be presented here. These parameters have been selected after introductory experiments performed on a small set of data using a set of predefined values for C and γ . Parameters leading to the best results of recognition have been selected. Classification experiments were done for the whole set of features and for the reduced set of features created by the stepwise fit.

Table 1 summarizes the statistical classification results for the testing data achieved by the SVM for all descriptors and after their selection. The first number represents the mean value and the term after the \pm sign stands for standard deviation. Both were obtained in repeated 10-fold cross validation experiments. We have applied this procedure since it is the approach regarded as the most objective one in estimation of the quality of the model applied. This is due to the fact that all data take part in learning and testing stages simultaneously.

Table 1
Results of mammogram recognition using the classical neural approach

	All features	Selected features
Accuracy	78.73% \pm 1.96	81.01% \pm 2.36
Sensitivity	79.73% \pm 1.53	82.48% \pm 1.97
Specificity	77.73% \pm 1.79	79.63% \pm 2.02

4. Deep learning application to mammogram recognition

Deep learning is a novel research technique, integrating the process of self-organizing feature selection and final classification of the images [25, 26]. In this research, we have applied the convolutional neural network (CNN) as a workhorse. An important drawback of this approach is the limited number of images representing the abnormal cases. To increase the information on class differences among the mammograms being analyzed, additional preprocessing of the images has been proposed. This was done by applying nonnegative matrix factorization (NMF) and statistical self-similarity of the images. Thanks to this additional view of the mammograms, diagnostic information contained in the original database has been enhanced.

4.1. Image representation using NMF. Non-negative matrix factorization is a decomposition technique representing the

given matrix \mathbf{P} by two other matrices \mathbf{W} and \mathbf{H} , both composed of non-negative elements [27, 28], i.e. $\mathbf{P} = \mathbf{WH}$. The columns of \mathbf{W} represent the basis vectors and the columns of \mathbf{H} stand for the encodings associated with them. Let us assume that the matrix \mathbf{P} is composed of the column vectors of the length N while M is the number of such vectors. Matrices \mathbf{W} and \mathbf{H} are of the $N \times r$ and $r \times M$ size, respectively, where the value of r is the factorization rank adjusted by the user (usually $(N + M)r < NM$).

The NMF will be performed here to enrich the representation of the mammographic images and to enhance the difference between mammograms representing normal and abnormal cases. All analyzed mammographic images are represented in the vector Hilbert form. They are grouped in the matrix \mathbf{P} .

The NMF operation will be performed on the set of mammograms representing only the normal cases, since these images are more similar to each other. Only half of the vectors belonging to the normal class has been used in this step of processing. According to the NMF procedure, matrix \mathbf{P} is decomposed into \mathbf{W} and \mathbf{H} of non-negative elements. Factorization means that i th vector \mathbf{p}_i (the i th column of \mathbf{P}) can be expressed as the weighted sum of basis vectors, and it might be presented in a Matlab notation [17] in the following manner:

$$\mathbf{p}_i = \sum_{j=1}^r \mathbf{W}(:, j) \mathbf{H}(j, i). \quad (6)$$

The whole set of original mammographic images representing normal and abnormal cases is converted to NMF factors and then reconstructed using only the limited number r of the basis vectors. In these investigations, we have applied only 10 basis vectors in reconstruction ($r = 10$). Since NMF decomposition was performed only on the normal cases, such reconstruction will represent images of this class more effectively. The abnormal cases reconstructed by means of basis vectors obtained as part of NMF decomposition of only the normal cases will show larger discrepancy with the original ones. This way, the differences between normal and abnormal cases have been increased. Thanks to it, the recognition of classes will become easier.

4.2. Statistical self-similarity for image representation. The next type of transformation applied to the original images is effected using the so-called statistical self-similarity. These images are defined on the basis of statistics of pixel intensity distribution in regions which are small in comparison with the whole image. In the first stage of processing, the image is resized to the dimension of 1024×1024 pixels and then split into small 5×5 compact overlapping regions. This way, the original mammographic image is represented by 256×256 small sub-images. In the next step, the similarity of these sub-images to the whole image is measured using the Kolmogorov-Smirnov d_{KS} distance [17]. As a result, the small subregions are represented by single values equal to $1 - d_{KS}$, with the range between 0 and 1. The lower the value of KS, the more similar the sub-image to the whole image. In the final stage, the set of 256×256 KS images is scaled back to the original dimension of 128×128 . Such transformation of images increases the differences between representatives of various classes.

4.3. Convolutional neural network in mammogram recognition. The CNN model is a very complex nonlinear structure, exploiting high-level abstraction by using multiple hidden layers [25, 26]. These layers are able to extract and identify different levels of detail in images. In the higher layers, more abstract concepts are learned on the basis of previous patterns extracted by the lower layers. The layer is composed of a group of neurons, performing the role of locally connected filters. Each neuron receives input signals from a set of compact units located in a small neighborhood of the previous layer. The neurons extract the elementary features, such as blobs, edges, crossings of edges, end points, corners, etc. The local reception field of each neuron is moved along all pixels of the image with the step (stride) defined by the user.

Eventually, the features combined by the subsequent layers create a fully connected layer, representing the input signals sent to the output classification layer. The output signals of this layer are generated by the softmax units and form the final class recognition. The softmax layer calculates the output value based on multinomial logistic regression [26], representing the probability of membership of the actual input vector in the relevant class. The number of units in the softmax layer is equal to the number of classes. The class of the highest probability is taken as the final winner. A detailed description of CNN can be found in [25].

In this paper, the CNN containing three convolution layers and two fully connected layers has been found most successful [12]. The details of the layers that followed are presented below.

- First convolution layer structure: 32 filters of dimension 5×5 with zero padding 2×2 and stride 1×1 ; Max pooling of the size = 3×3 , zero padding = 0×0 , stride = 2×2 ; rectified linear unit layer.
- Second convolution layer structure: 32 filters of dimension 5×5 , zero padding 2×2 , stride 1×1 ; Average pooling of the pooling size = 3×3 with zero padding = 0×0 , stride = 2×2 ; rectified linear unit layer.
- Third convolution layer structure: 64 filters of dimension 5×5 with zero padding 2×2 and stride 1×1 ; Average pooling of size = 3×3 with zero padding = 0×0 , stride = 2×2 ; rectified linear unit layer.
- First fully connected layer: 64 neurons with rectified linear units.
- The second fully connected layer contains two neurons (dependent on the number of recognized classes) with softmax. It performs the role of final classification.

The general organization of the CNN system for recognition of classes of mammograms is presented in Fig. 3 [12].

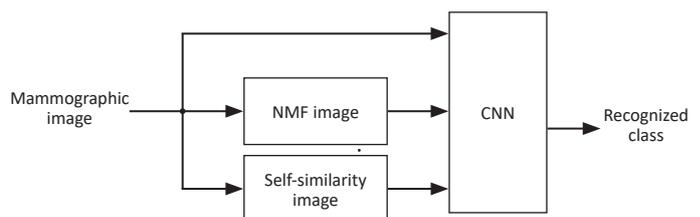


Fig. 3. Deep learning system used in mammogram recognition

The CNN is supplied by three images representing the mammograms being analyzed: the original mammographic image, the image reconstructed on the basis of NMF and the transformed image reconstructed on the basis of the self-similarity principle. Thanks to these additional transformations of the mammograms, more information regarding the structure of the analyzed images is delivered to CNN. At the same time, the number of input data is also triplicated. The information enhanced this way significantly increases the probability of correct classification.

4.4. Results of numerical experiments. The numerical experiments with mammogram recognition using CNN have been performed on the same samples of mammograms in DDSM database as those used in the previous section. The aim was to differentiate between normal cases (class 1) and the abnormal ones (benign and malignant together), representing class 2. The experiments have been performed using 10-fold cross validation organized in the same fashion as in the classical approach. Only half of the normal cases in the learning sets have been used in NMF decomposition to get the basis vectors of \mathbf{W} used in the reconstruction of all mammographic images.

The procedure of the final class recognition was performed in the manner described below. The continuous output of the classifier in the learning mode (before binarization) was subject to dynamic thresholding at different values of threshold. The threshold generating the highest quality measure for the learning data was fixed and applied in testing the remaining validation data (10% of data in each run of the cross validation procedure). The quality measure which was taken into account in this step included the value of AUC of the receiver operating characteristic. The maximized AUC measure is a good compromise between the sensitivity (ability to discover the minority class) and specificity (ability to discover the majority class) of the recognition system.

Table 2 presents the detailed results of recognition of mammograms obtained in the testing mode of the 10-fold cross validation [12]. The results of recognition of normal cases from abnormal ones are presented in the form of sensitivity, specificity and average accuracy. The sensitivity of recognizing the abnormal cases from normal ones equaled 82.82% and specificity stood at 86.59%. The obtained accuracy is found somewhere in the middle between the of them (85.83%). The area obtained under the ROC curve was $AUC = 0.919$. These results belong to the best ones reported for such a large DDSM database to date.

Table 2
Results of numerical experiments of mammogram recognition using CNN

Sensitivity	Specificity	Accuracy
82.82% ±0.95	86.59% ±1.12	85.83% ±1.08

Comparison with the classical results presented in Table 1 shows evident advantage of the deep learning approach. All quality measures have increased significantly. To assess the importance of including NMF and self-similarity images in

the recognition process, additional experiments have been performed using only the original mammogram images. The obtained AUC value for recognition of abnormal from normal cases was reduced to $AUC = 0.88$. According to the *ranscum* test, at a 5% significance level this difference is statistically significant.

5. Comparative study

The problem of mammogram recognition has been studied in many papers. However, most of them used either different databases or very limited images selected from DDSM. Different quality measures have been also applied in presentation of the results. Therefore it is difficult to present a comparison of all these works in an objective manner. We will thus limit our comparison here to the papers which have used the same DDSM database that we did.

Paper [7] has considered a very small set of 410 mammograms from the DDSM database and the overall accuracy achieved by authors was 87% with 88.6% sensitivity and 78.6% specificity. Due to the small number of samples, these results are not fully credible. The quality measure of the solution in the form of AUC value was presented in papers [8] and [28]. The AUC value of 0.789 for 600 cases was reported in [8] and of 0.871 for 1,000 screening mammograms in [29].

In [30], the deep CNN approach to recognition of normal from abnormal mammograms from a very large database from the Netherlands, containing over 44,000 mammographic views, has been presented. The results are represented by the ROC curve. The best AUC with the augmentation (context, location, patient information) and manual feature support was $AUC = 0.941$. The best result without augmentation was $AUC = 0.929$.

In [31], the results for DDSM declaring 85% of accuracy and $AUC = 0.91$ have been presented. They were obtained using Google Le Net system and an ensemble of 100 parallel networks.

The results for the DDSM base presented in [32] have covered 1,057 malignant and 1,397 benign cases. They all concentrated on ROC and declared the best value of $AUC = 0.82$.

In [33], the DDSM for more than 6,000 mammographic images and ZMDS (1,739 mammograms) have been considered. The best results for images declared $AUC = 0.922$, sensitivity of 0.901 and specificity of 0.783.

Our best average accuracy in recognizing abnormal cases (malignant plus benign versus healthy) for whole images from the DDSM database was 85.83%, with sensitivity of 82.82%, specificity of 86.59% and $AUC = 0.919$. The only recent results presented also for the whole DDSM database (2,003 abnormal and 9,215 normal mammograms) are given in [11]. The accuracy in abnormality detection (malignant plus benign versus healthy) reported in this paper by using the curvelets for the same DDSM database was in the range from 81.3% to 86.4%, depending on the applied feature set. However, sensitivity, specificity and AUC were not given. It is difficult to assess the quality of their solution on the basis of only accuracy

value, since it is very easy for this unbalanced data set (2,003 abnormal and 9,215 normal mammograms) to obtain high accuracy at the cost of sensitivity. In our additional experiments, by applying accuracy as the quality measure of the ensemble, we have obtained average accuracy equal to 89.4%. However, this was done at the cost of sensitivity, which then dropped to only 69.5%.

6. Conclusions

The paper has presented a comparative analysis of the classical and deep learning approach to differentiation between abnormal and normal cases on the basis of mammogram images. In the classical approach, an extended set of numerical descriptors has been proposed. They were defined on the basis of different principles of image characterization and included representations of image by using the Hilbert form and corresponding descriptors, i.e. Kolmogorov-Smirnov statistics, the maximum subregion principle, percolation theory, fractal texture descriptors as well as application of wavelet and wavelet packets. Thanks to the application of so many methods, different points of view were considered for the image in pattern recognition. However, in spite of such rich descriptive representation of the images and application of the efficient SVM classifier, the results were inferior in comparison to the application of the deep learning approach, enhanced by non-negative matrix factorization and self-similarity of the images.

The most important advantage of deep learning for mammogram recognition is the relatively simple way of preparation of input data for the convolutional neural network. The diagnostic features are self-defined in an unsupervised approach to the process of CNN learning. However, to get good results of recognition, a large number of learning samples should be used. The NMF and self-similarity transformations have not only enhanced the information of the image details, but also increased the population of samples involved in learning. Better results of recognition might be expected after further increasing the population of the original mammogram images.

Additional investigations are needed to increase accuracy to a level acceptable for everyday use in medical practice. Future investigations will explore both approaches. The classical one will be directed towards applying more classifiers arranged in an ensemble to increase the diversity of principles on the basis of which the final decision is made. In the case of deep learning, new perspectives open now with transfer learning [26]. More specialized ways of learning hidden neurons in CNN will be studied. And in both cases the accuracy of image recognition might certainly be increased by applying a larger database of abnormal cases.

REFERENCES

- [1] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, and F. Bray, "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN". *International Journal of Cancer* 136(5), E359-E386 (2012).
- [2] H. Nelson, K. Tyne, A. Naik, C. Bougatsos, B. Chan, P. Nygren, and L. Humphrey L, "Screening for breast cancer: Systematic evidence review update for the U. S. preventive services task force", *Ann. Intern. Med.* 151(10), 727-W242 (2009).
- [3] S. Hofvind, G. Ursin, S. Tretli, S. Sebuodegard, and B. Moller, "Breast cancer mortality in participants of the Norwegian breast cancer screening program", *Cancer* 119(17), 3106–12 (2013).
- [4] A. Jalalian, S.B. Mashohor, H.R. Mahmud, M.I. Sariipan, A.R. Ramli, and B. Karasfi, "Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound" *Clinical Imaging* 37, 420–426 (2013).
- [5] Z. Wang, Q. Qu, and G. Yu, "Breast tumor detection in double views mammography based on extreme learning machine", *Neural Computing and Applications* 27(1), 227–240 (2016)
- [6] I. Christoyiani, A. Koutra, E. Dermata, and G. Kokkinakis, "Computer aided diagnosis of breast cancer in digital mammograms", *Computerized Medical Imaging and Graphics* 26, 309–319 (2002).
- [7] B.K. Elfarrar and I.S. Abuhaiba, "New Feature Extraction Method for Mammogram Computer Aided Diagnosis", *Intern. Journal of Signal Processing, Image Processing and Pattern Recognition* 6(1), 1–81 2013.
- [8] M. Mazurowski, J. Zurada, and G. Tourassi, "Selection of examples in case-based computer-aided decision systems", *Physics in Medicine and Biology* 53, 6079–6096 (2008).
- [9] M. Lobbes, M. Smidt, K. Keymeulen, R. Girometti, C. Zuiani, R. Beets-Tan, J. Wildberger, and C. Boetes, "Malignant lesions on mammography: accuracy of two different computer-aided detection systems", *Clinical Imaging* 37, 283–288 (2013).
- [10] S. Leon, L. Libby Brateman, J Honeyman-Buck, and J. Marshall, "Comparison of two commercial CAD systems for digital mammography", *Journal of Digital Imaging* 22(4), 421–423 (2009) doi: 10.1007/s10278-008-9144-x.
- [11] S. Dhahbi, W. Barhoumi, and E. Zagrouba, "Breast cancer diagnosis in digitized mammograms using curvelet moments", *Computers in Biology and Medicine* 64(1), 79–90 (2015).
- [12] B. Swiderski, J. Kurek, S. Osowski, M. Kruk, and W. Barhoumi "Deep learning and non-negative matrix factorization in recognition of mammograms", in *Proc. SPIE 10225B, Eighth Int. Conf. Graphic and Image Processing*, (2016) doi:10.1117/12.2266335.
- [13] M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer, "The digital database for screening mammography", in: *Digital Mammography*, Springer, Netherlands, 457–460 (1998).
- [14] B. Swiderski, S. Osowski, J. Kurek, M. Kruk, I. Lugowska, P. Rutkowski, and W. Barhoumi, "Novel methods of image description and ensemble of classifiers in application to mammogram analysis", *Expert Systems with Applications* 81, 67–78 (2017).
- [15] D. Stauffer, *Introduction to percolation theory*, Taylor & Francis, London, 1985.
- [16] R. Haralick and L. Shapiro, Image segmentation techniques. *Computer Vision. Graphics and Image Processing* 29, 100–132 (1985).
- [17] *Matlab user manual*, MathWorks, Inc. Natick, USA, 2017.
- [18] M. Schroeder, *Fractals, Chaos, Power Laws*. W.H. Freeman and Company, New York, 2006.
- [19] B. Moon, H.V. Jagadish, C. Faloutsos, and J.H. Saltz, "Analysis of the clustering properties of the Hilbert space-filling curve", *IEEE Transactions on Knowledge and Data Engineering* 13(1):124–141 (2001) doi:10.1109/69.908985
- [20] A. Costa, G. Humpire-Mamani, and A. Traina, "An efficient algorithm for fractal analysis of textures", in *Proc. 25 SIBGRAPI Conference on Graphics, Patterns and Images*, 39–46 (2012).

- [21] M. Jiang, S. Zhang, H. Li, and N. Metaxas, "Computer-aided diagnosis of mammographic masses using scalable image retrieval", *IEEE Transactions on Biomedical Engineering* 62(2), 783–792 (2015).
- [22] I. Daubechies, *Ten lectures on wavelets*, SIAM, Philadelphia, 1992.
- [23] R.N. Khushaba, S. Kodagoa, S. Lal, and G. Dissanayake, "Driver drowsiness classification using fuzzy wavelet packet based feature extraction algorithm", *IEEE Transaction on Biomedical Engineering* 58(1), 121–131 (2011).
- [24] J. Kurek, M. Kruk, S. Osowski, P. Hoser, G. Wiczorek, A. Jegorowa, J. Górski, J. Wilkowski, K. Śmietańska, and J. Kossakowska, "Developing automatic recognition system of drill wear in standard laminated chipboard drilling process", *Bulletin of the Polish Academy of Sciences Technical Sciences*, 64(3), 633–640 (2016).
- [25] I. Goodfellow, Y. Bengio and A. Courville, *Deep learning*, MIT Press, Massachusetts, USA, 2016.
- [26] A. Krizhevsky, I. Sutskever, and G. Hinton, "Image net classification with deep convolutional neural networks", *Advances in Neural Information Processing Systems* 25, 1–9 (2012).
- [27] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization", *Nature* 401, 788–791 (1999).
- [28] A. Cichocki, R. Zdunek, A.H. Phan, and S.I. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*, Wiley, New York, 2009.
- [29] J.J. Fenton, S.H. Taplin, P.A. Carney, et al. "Influence of computer-aided detection on performance of screening mammography", *New England Journal of Medicine* 356, 1399–1409 (2007).
- [30] T. Kooi and G. Litjens, "Large scale deep learning for computer aided detection of mammographic lesions", *Medical Image Analysis*, 35, 303–312 (2017).
- [31] D. Yi, R.L. Sawyer, D. Cohn III, J. Dunnmon, and C. Lam, „Optimizing and visualizing deep learning for benign/malignant classification in breast tumors”, *29th Conf. NIPS 2016*, arXiv preprint, arxiv.org (2017).
- [32] R.K. Samala, H.P. Chan, and L.M. Hadjiiski, „Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms”, *Physics in Medicine & Biology* 62, 8894–8908 (2017).
- [33] P. Teare, M. Fishman, O. Benzaquen, and E. Toledano, "Malignancy Detection on Mammography Using Dual Deep Convolutional Neural Networks and Genetically Discovered False Color Input Enhancement", *Journal of Digital Imaging* 30, 499–505 (2017).