Expert Systems With Applications 164 (2021) 113968

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Deep neural system for supporting tumor recognition of mammograms using modified GAN

B. Swiderski [a], L. Gielata [a], P. Olszewski [a], S. Osowski [b,c,*], M. Kołodziej [b]

[a] Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences, 166 Nowoursynowska Street, 02-787 Warsaw, Poland
[b] Faculty of Electrical Engineering, Warsaw University of Technology Warsaw, 75 Koszykowa, 00-662 Warsaw, Poland
[c] Faculty of Electronic Engineering, Military University of Technology, Warsaw, Kaliskiego 2, 00-908 Warsaw, Poland

ARTICLE INFO

Keywords:
Deep learning
GAN
CNN
Mammogram recognition

ABSTRACT

This paper presents the autoencoder-generative adversarial network (AGAN) in the analysis of mammograms. The AGAN architecture is used to augment the data by generating additional representations of the mammogram images, enhancing this way the information of the analyzed problem. The images generated by this this deep network are appended to the original set of mammograms and fed to the input of convolutional neural network, which plays the role of the final classifier. The proposed system was used to recognize the mammograms belonging to two classes: normal and abnormal. The investigations were performed using a large database consisting of 11,218 regions of interest of mammographic images from the DDSM base. The results demonstrate the advantage of this proposed deep learning system over other known approaches to mammogram recognition. Our average accuracy in detecting abnormal cases (malignant plus benign versus healthy) was 89.71%, sensitivity 93.54%, specificity 80.58% and AUC = 0.9410. These results are among the best for this large database.

1. Introduction

Mammograms are used to check for breast cancer in women, who have no symptoms or signs of the disease. Early detection of breast cancer with screening mammography is very useful in the treatment of cancer, because it can be started earlier in the course of the disease. In this way it helps to reduce the number of deaths from breast cancer (Ferlay et al., 2012).

Due to the limited number of expert radiologists and the very large number of screening mammograms, the mammogram detection procedure is a bottle neck in all screening programs. Therefore, computer supported systems are of interest (Christoyiani et al., 2002). Such systems can alert expert radiologists to the suspicious regions. Despite the fact, that various solutions to this problem have been reported, the efficiency of the systems actually proposed is still not satisfactory.

The known solutions to this problem can be divided into two streams. The first stream focused on the search for specially defined mammography descriptors based on expert knowledge. Selected descriptors serve as input attributes for the final classifier or ensemble of classifiers. The diagnostic features defined in this way are based on the statistical characterization of the image, including texture, edge orientation, analysis of a map of pixels in the mammographic image, etc. Many different tools are used to describe the image. They include thresholding techniques, mathematical morphology tools, wavelet decomposition, template matching, neural networks and many others (Christoyiani et al., 2002; Jiang et al., 2015; Kooi & Litjens, 2017; Samal et al., 2017; Swiderski et al., 2017). The main problem of such an approach to image analysis is the narrow characterization of the data, which is strictly related to the applied method. They are not able to cover the scale of the variety of mammograms that make up the database.

The second stream of research is based on the application of deep learning (Teare et al., 2017; Yi et al., 2017; Guan & Loew, 2019). In this approach the deep structure is simultaneously responsible for generation of diagnostic features and classification. The most known example of such a solution is application of the convolutional neural network (CNN) (Krizhevsky et al., 2012). The user delivers the original raw mammograms to the input of CNN and the network is responsible for all stages of image processing and delivers the class membership of the analyzed image to its output. This approach to mammogram analysis is much more universal. The local convolutional layers of CNN analyze different details of the image and form the overall view of various aspects of the image, which are then used as diagnostic features by the classifier, which is a fully connected part of the CNN.

Many research results presented in the past are dedicated to small

* Corresponding author at: Warsaw University of Technology, Faculty of Electrical Engineering, Koszykowa 75, Warsaw, Poland.
E-mail addresses: bartosz_swiderski@sggw.pl (B. Swiderski), stanislaw.osowski@ee.pw.edu.pl (S. Osowski).

https://doi.org/10.1016/j.eswa.2020.113968
Received 15 November 2019; Received in revised form 10 August 2020; Accepted 2 September 2020
Available online 17 September 2020

databases of mammographic images, for example Mini Mammographic Database of MIAS or limited set of mammograms chosen from Digital Database for Screening Mammography (DDSM). They are not reliable, due to the small size of the data sets. The results are presented in various forms: accuracy, sensitivity, area under ROC curve (AUC), etc. Paper (Christoyiani et al., 2002) has used a radial basis function network cooperating with PCA and declared the accuracy rate of 88.23% in detection of all kinds of abnormalities in the analyzed 119 regions of suspicion for MIAS database. The classification accuracy of 87%, with 88.6% sensitivity and 78.6% specificity was reported for 410 mammograms randomly selected from DDSM (Elfarra & Abuhaiba, 2012). They have used run difference and spatial grey level dependence methods for feature extraction and multilayer perceptron as the classification tool. The recognition results of abnormal cases for all mammograms of DDSM base (over 10,000 ROI of mammograms) by using the curvelet moments in feature generation and the K-nearest neighbor classifier were presented in Dhahbi et al. (2015). There only the accuracy rate was given. This measure changed from 81.26% to 86.46% depending on the applied feature set. However, no information on sensitivity, specificity and AUC information was presented. The authors of the paper (Jiang et al., 2015) developed the vocabulary tree framework to retrieve mammographic masses using scale-invariant feature transformation and proposed the vocabulary tree refinement to select the specific mammographic mass. The best declared accuracy in recognition abnormal from normal mammograms was 86.9% for the DDSM base.

Deep learning is the most commonly used method in mammographic image analysis today. Many research directions have been proposed in this field (Kurek et al., 2019). Some of them use the networks that have been trained from scratch and other use transfer learning from CNN (Krizhevsky et al., 2012). Individual classifier or the number of them organized in the form of an ensemble are used. The work Yi et al. (2017) has used Google Le Net system and an ensemble of 100 parallel networks, declaring 85% of accuracy and AUC = 0.91 in recognition of normal from abnormal mammograms in DDSM base. The results presented in Samala et al. (2017) for DDSM base focused only on ROC and declared the best value of AUC = 0.82. The results of Teare et al. (2017) were obtained for 6000 mammographic images from DDSM and 1739 from Zebra Mammography Dataset (ZMDS). The best results for these images were AUC = 0.922, sensitivity 90.1% and specificity 78.3%. The paper Kooi and Litjens (2017) presented a deep CNN network in combination with a candidate ROI detector for recognition of mammograms using very large Dutch database (over 44,000 mammographic views). The best reported AUC with augmentation (context, location, patient information) and manual feature support was 0.941 and without augmentation 0.929. No sensitivity and specificity values were reported.

The important point in deep network approach to medical image analysis is the augmentation of the data. Various methods have been proposed in the past. Such methods include: Boltzmann machine, deep believe network (Goodfellow et al., 2016), diffusion inversion and various types of generative models, including moment matching method, denoising autoencoder, variational autoencoder and generative adversarial networks (Alyafi et al., 2019; Guan & Loew, 2019; Kingma & Welling, 2019; Yoo et al., 2020; Yia et al., 2019). Generative models represent a very powerful methods of learning any kind of data distribution. They aim to learn the true data distribution of the training set in order to generate new data points with some variations (Wu et al., 2018).

The paper Guan and Loew (2019) examined the application of the Generative Adversarial Network (GAN) in the augmentation process of images in CNN learning and declared very good results of its application. The stable average validation accuracy for the classification of abnormal vs. normal cases in the DDSM database converged at about 91.48%. The paper presented the application of Deep Convolutional Generative Adversarial Networks in Breast Mass Augmentation in X-ray Mammography and declared the F1 score from 86% to over 98% (dependent on the size of learning data). These results were obtained for OPTIMAM Mammography Image Database.

The other aspect of mammography analysis is also the improvement of image quality and mass segmentation. The paper (Zhu et al., 2017) proposed application of adversarial deep structured net for this problem. The dice index metric used to evaluate the results reached the value 91.30%. The paper (Korkinof et al., 2019) presented the methodology for the generation of highly realistic, high-resolution synthetic mammograms using a progressively trained generative adversarial network. They could be included in learning process of deep classifier.

This paper will also focus on the application of deep learning in mammogram recognition (normal versus abnormal). However, our approach differs significantly from all others presented so far. We propose special augmentation of the analyzed images by training the deep learning structure, so called autoencoder-generative adversarial network (AGAN). The deep generator of this structure is responsible for the reproduction of the additional images of mammograms. This type of augmentation was proposed recently in the paper Guan and Loew (2019), where synthetic images produced by GAN were used to enrich the database.

Our proposal is significantly different. Firstly, the modified structure of GAN, called AGAN, has been proposed. Secondly, only normal images were used when learning AGAN. Thanks to such an organization the reproduction of normal images is almost perfect. However, in the abnormal cases the reproduced images differ significantly from the originals. In this way we increase the differences between the two classes.

The most important difference of our solution to other approaches is at the concept level. We present a new idea that does not interfere with the simultaneous use of the standard GAN. The key here is the assumption that normal data is more consistent than abnormal. By teaching only this normality (on a separate set of normal data) and then requesting its reproduction (image reconstruction), we can rightly expect that the reconstruction error will be greater for abnormal images.

In our solution the CNN classifier is fed simultaneously by the stream of two images: the original ones and the images reproduced by the generator. They form the tensor of the duplicated depth. The set of normal images representing a class (original and images reconstructed by AGAN) used to learn the final CNN classifier is much more uniform than the set of images representing the abnormal class, for which the reconstructed images have a greater difference from their prototypes. Thanks to such organization the quality of the recognition results is much better. Our best average accuracy in detecting abnormal cases (malignant plus benign) versus healthy for all 11,218 images in DDSM database was 89.71%, sensitivity 93.54%, specificity 80.58%, AUC = 0.941 and positive predictive value PPV = 0.920. These results prove, that our method is superior to the compared approaches with respect to all evaluation metrics, including accuracy, sensitivity and AUC.

The main contribution of this work is as follows:

- Proposition and application of novel deep structures called AGAN, based on GAN and autoencoder system for the augmentation of mammographic images. Only images of normal cases are involved in learning process of this system. The reproduced images for these cases are almost perfect, in contrast to abnormal images, where such differences are much greater. Thanks to this the differences between analyzed images of diverse classes are enriched. This helps in the final classification and increases the quality measures in class recognition.
- Successful application of convolutional neural network in the analysis of mammographic images. The important element in this solution is application of an additional set of images reproduced by the generator of the autoencoder-GAN architecture. They duplicate the depth of input tensor applied to CNN.
- Experimental evaluations of the proposed solution performed on the large DDSM set of mammograms, have shown a significantly better performance compared to other results actually presented in various

studies. The quality measures of recognition obtained on very large DDSM database belong to the best for this set.

The rest of the paper is structured as follows. Section 2 briefly describes the proposed AGAN structure, which plays an important role in the solution. Section 3 describes the applied deep neural system for mammogram recognition, built on the basis of AGAN and CNN. Section 4 discusses various aspects of databases used in learning the system. Section 5 presents the details of all applied deep learning structures of the system and the results of numerical experiments. The concluding section summarizes the presented considerations.

2. Modified GAN system

The generative adversarial network is a very clever deep learning structure, training simultaneously two models: a generative model G to capture the unknown data distribution represented by input noise vector **z** and discriminative model D. The generator tries to map this noise vector to data space G(**z**) in a way to imitate the image X. The discriminator D is supplied by the original image **X** and image G(**z**), and is trained to maximize the probability that the analyzed image **X** came from the training set, rather than from G (Goodfellow et al., 2014). The general structure of GAN system is presented in Fig. 1.

Learning process of G is directed to maximize the probability of D making a mistake. The adversarial discriminator D tries to distinguish between real and generated samples as exactly as possible. The generator G tries to "fool" discriminator by producing samples which are very similar to real data samples. The learning procedure of generator attempts to create its results in such a way, that discriminative model D cannot distinguish samples from real data distribution and from generative distribution. The learning task of GAN system is defined as the minimax optimization problem of the function $V(G,D)$ (Goodfellow et al., 2014)

$$\min_G \max_D V(G,D) = \mathbb{E}_{\mathbf{x}\sim p_x}[log\mathrm{D}(\mathbf{X})] + \mathbb{E}_{\mathbf{z}\sim p_z}[log(1-\mathrm{D}(\mathrm{G}(\mathbf{z})))] \quad (1)$$

where symbol represents the expectation operator, while $\boldsymbol{p_x}$ **and** $\boldsymbol{p_z}$ the real data and generator data distribution, respectively. $D(\mathbf{X})$ and $G(\mathbf{X})$ are the values created by discriminator D and generator G for particular image **X**, respectively. The optimization process searches for minimum value with respect to G and maximum with respect to D.

The generator G is fed by the random vector **z** of chosen dimensionality and tries to produce the image resembling the real image of the data. The discriminator fed by either real image or image generated by G, produces the single scalar, representing the probability that the data generated by G are not distinguishable from the real data set. The probability level of 0.5 means that both data set are not distinguishable. Both models: G and D may take any form of neural network, for example multilayer perceptron. However, in our solution they take the form of more efficient deep convolutional neural network.

The paper Donahue et al. (2017) has proposed the modified GAN structure called BIGAN by including autoencoder into the basic GAN model. In this solution the encoder E maps the real data **X** to latent representations. The discriminator D discriminates now not only in image data space (**X** versus G(**z**)) but jointly in data and latent space (**X**, E(**X**)) versus (G(**z**), **z**). In this way the information used in comparison is enhanced. To get good results of system performance (to " fool" discriminator), the encoder E should be learned to invert the generator G. The learning procedure of BIGAN is transformed to the minimax problem of the objective function $V(D,E,G)$ depending on discriminator D, autoencoder E and generator G (Donahue et al., 2017)

$$\min_{G,E}\max_D V(D,E,G) \quad (2)$$

where

$$V(D,E,G) = \mathbb{E}_{\mathbf{x}\sim p_x}[log\mathrm{D}(\mathbf{X},\mathrm{E}(\mathbf{X}))] + \mathbb{E}_{\mathbf{z}\sim p_z}[log(1-\mathrm{D}(\mathbf{X},\mathbf{z}))] \quad (3)$$

In our solution we have applied different modification of GAN, which also includes an autoencoder. This time the generator G is supplied subsequently either by **z** (as in GAN) or by autoencoder signal of the real image. In this way we increase the latent information of the object delivered to generator and then to discriminator. As a result, the probability of a better reconstruction of the image by the generator is higher than with GAN, which had no direct access to the distribution of original images. The image generating process in learning stage is sequential, i. e., input signals delivered to generator G are interlaced every phase: first come from autoencoder and then from **z**. The learning task is defined now as the following minimax problem

$$\min\ \max\ \{\mathbb{E}_{\mathbf{x}\sim p_x}[log\mathrm{D}(\mathbf{X})] + \mathbb{E}_{\mathbf{x}\sim p_x,\mathbf{z}\sim p_z}[log(1-\mathrm{D}(\mathrm{G}(\mathbf{z},\mathrm{E}(\mathbf{X}))))]\} \quad (4)$$

where the expression $G(z,E(\boldsymbol{X}))$ means that generator is learned also on the basis of autoencoder signals. The general structure of such system is presented in Fig. 2. **X** represents the input image and **z** is the random vector of defined size. The model is composed of two basic parts: autoencoder cooperating with generator G and GAN subnetwork, which uses the same generator G and discriminator D.

The generator G is trained alternatively:

- either by the latent representation generated by the encoder E(**X**) (part of the autoencoder), the aim being to restore original (normal) image **X**,

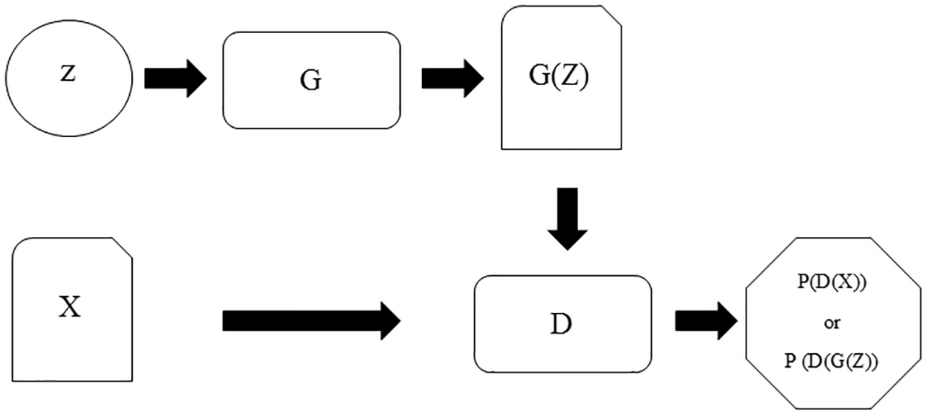

Fig. 1. The general structure of GAN system. Vector **z** represents the noise, G is the generator trained to produce the image G(**z**) very similar to real image **X**. D is the discriminator trying to distinguish samples from real data distribution and from generative distribution. It is trained to maximize the probability of assigning the correct label to training samples **X** and samples from generator G.

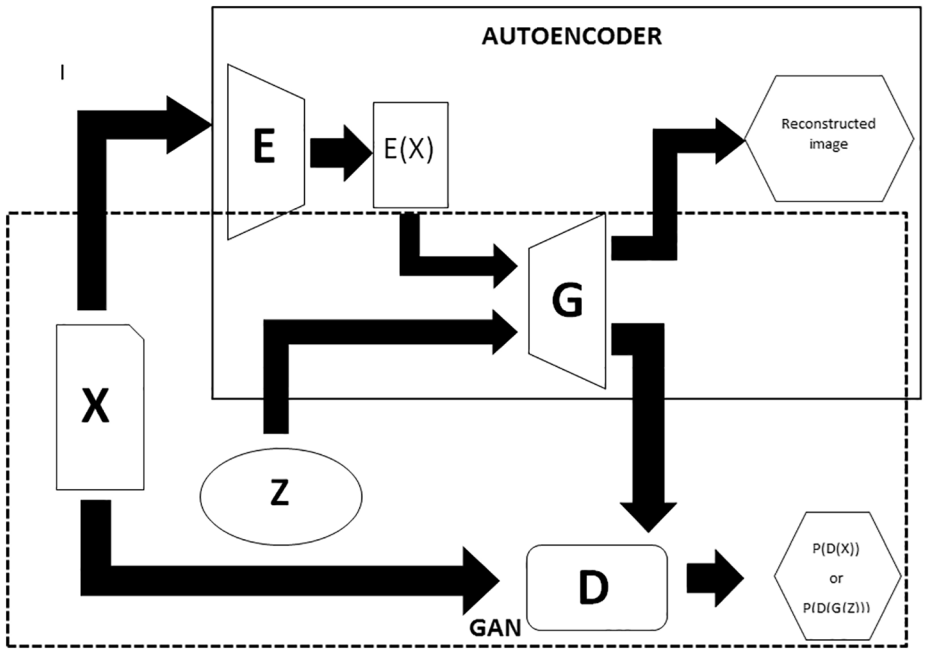

Fig. 2. The proposed structure of modified GAN used in the solution. Block E represents autoencoder producing signal E(**X**). Generator G is supplied subsequently either by signal of autoencoder E(**X**) or directly by z. The result is the reconstructed image and probability P(D(**X**)) and P(D(G(**z**))).

- or by random latent variable **z,** the aim of which is to produce synthetic real-like normal image (cheat the discriminator D).

This means that the parameters of the generator are adapted twice per cycle. The discriminator D tries to distinguish between synthetic and real images. The result of D represents now the probability that both images: the real **X** and the reconstructed G(**z**) are the same. Its result is used to train generator G.

3. AGAN based deep system for mammogram recognition

Mammograms are recognized by the Neural Convolution Network (CNN), which is fed by two sets of images: the original images of the mammograms and the images reproduced by AGAN's generator G. They form the tensor of the depth equal to 2. The aim of the AGAN application is to increase the information of images contents and enrich the differences between the normal images representing healthy persons and abnormal images representing cancer (both benign and malignant). To achieve these goals the AGAN system is learned only with normal images selected from the original database. In this way the generator G of AGAN is able to represent the normal cases in an efficient way. Its reproductions of the abnormal cases will be significantly different from the originals.

The input for CNN in its learning and testing processes is made up of two channels. One consists of the original images and the second of the images reconstructed by generator G of AGAN, as shown in Fig. 3. In case of normal mammograms both images will present similar inputs. In abnormal cases, however, the original and reconstructed images differ significantly. In this way, we reinforce the differences between the two classes of data.

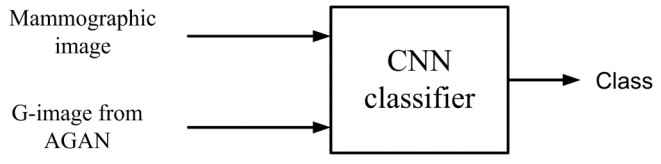

Fig. 3. The general structure of final CNN classifier system used in recognition of mammograms. CNN is fed from original mammographic images and from images reconstructed by AGAN.

Fig. 4 shows an example of a normal mammographic image (a) and its reconstructed representation (b). As can be seen, both images are very similar. They form the duplicated input for the CNN classifier, which enhances the information provided in training and testing phases.

Fig. 5 shows two abnormal cases (benign and malignant) and their reconstructions performed by the generator G of AGAN. This time we can see significant differences between the corresponding images. These two different images (original and reconstructed) delivered to the input of CNN classifier reinforce the internal differences between mammograms of the second class.

As a result two similar images representing the normal case (one class) are contrasted with two quite different images representing the second (abnormal) class. Such set of images delivered to the input of CNN increases the probability of being correctly recognized between these two classes.

4. Experimental setup

4.1. Database of mammograms

The numerical experiments have been performed using the largest publically available database of mammographic images "Digital Database for Screening Mammography" (Heath et al., 1998). It contains 2604 cases, each characterized by at least 4 mammograms (left and right breast from above representing Cranial-Caudal view and oblique representing Medio-Lateral-Oblique view). The mammograms are accompanied by the diagnostic results (normal, benign or malignant) and also the location of lesions, which form the regions of interest (ROI). The manual cropping of images for abnormal cases (benign and malignant) was done on the basis of information provided by the regions indicated by experts. The segmented ROI in each abnormal mammogram represents rectangular area with the lesion in the center. The ROIs in normal cases were extracted manually by the medical expert from normal tissues. The size of each ROI image was unified and reduced to 32×32 pixels. The total number of ROI images, used in experiments was 11218. The DDSM database used in the experiments consisted of the following

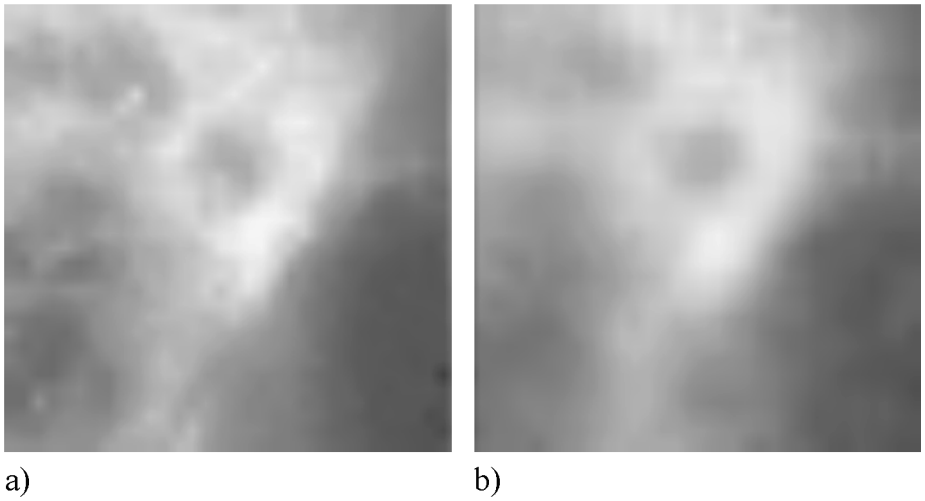

Fig. 4. The effect of reproduction of normal image by AGAN: a) mammogram corresponding to healthy person, b) reproduction of this image by generator G.

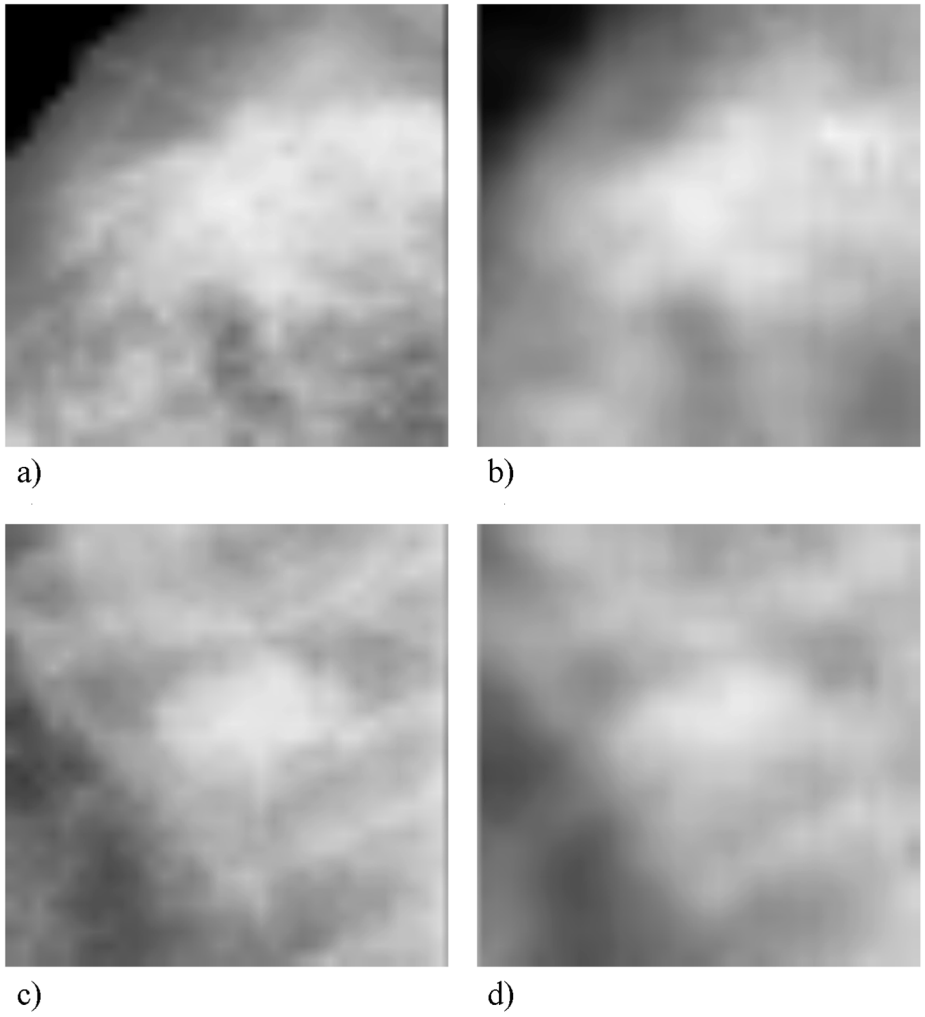

Fig. 5. The effect of reproduction of abnormal images by AGAN: a) mammogram corresponding to malignant case, b) the reproduction of malignant image by generator G, c) mammogram corresponding to benign case, b) the reproduction of benign image by generator G.

ROI class representations:

- Normal tissues: 9215,
- Benign lesion: 888 and malignant lesion: 1115.

In our experiments, benign and malignant mammograms were combined and form the class of abnormal cases. This means that this class was composed of 2003 ROI images. This is minority compared to 9215 samples representing the class of normal cases. Such an imbalance

creates additional problems associated with such a data set.

Fig. 6 presents some exemplary images of mammograms representing normal (upper row) and abnormal cases: benign (middle row) and malignant (bottom row). The significant differences among the images belonging to the same class and close similarity of images representing different classes can be observed

4.2. Details of applied system

The numerical experiments have been performed by using deep structure of all subsystems, including generator, autoencoder, discriminator and final CNN classifier. All of them have been designed separately from scratch. They differ by the architecture and parameters. These particular structures and parameter values have been found after series of introductory experiments.

The generator is composed of 4 sublayers. The input layer represents 100 randomly generated signals of uniform distribution U(0,1). The first fully connected layer is composed of 256 neurons with ReLU activation. It is learned by Adam optimization algorithm with momentum and initial learning rate 0.0002. The next layer is also fully connected and contained 625 neurons with ReLU activation. The output signals of this layer are re-interpolated by bicubic method to 35 × 35 size (Odena et al., 2016; Yu et al., 2016). The next is a convolution layer applying one filter 4 × 4, stride 1 × 1, of VALID convolution form. As a result, the output layer reproduces 32 × 32 image.

The structure of autoencoder is also of deep form. The input is formed by the original 32 × 32 image. The first convolution layer applies one filter 4 × 4, stride 1 × 1, VALID type convolution, ReLU activation and MAX pooling. Next two fully connected layers are composed of 100 neurons of ReLU activation. The output is taken from the second layer and contains 100 output signals.

The discriminator architecture is as following. The single 32 × 32 images are delivered to its input. The convolution layer is locally connected and applies one filter 4 × 4, stride 1 × 1, no zero padding, ReLU activation and 2 × 2 MAX pooling with stride 2 × 2. The first fully connected layer contains 100 neurons with ReLU activation. The second fully connected layer is composed of 2 neurons with ReLU activation. The output of discriminator contains one neuron with sigmoidal activation function. It predicts probability of real (normal) image. It should return value close to 1 for real images, and value close to 0 for images generated by generator.

The set of 4444 randomly chosen mammographic images, representing only normal cases, has been used in AGAN architecture training. In this way only the normal cases are well reconstructed by the system, increasing their population. However, the abnormal cases which did not take part in learning, are subject to significant changes. Thanks to this the differences between normal and abnormal cases are enriched. After finishing learning procedure its parameters are fixed and system is ready to generate the reproduced images, which are used in learning the CNN classifier, responsible for final recognition of mammograms.

The real recognition of mammograms is performed by the classical CNN network designed from scratch. Its input is formed by the original mammographic images and the images reconstructed by the generator of the AGAN network, as shown in Fig. 3. The size of these images is 32 × 32. The first and second convolution layers apply 32 filters of the size 3 × 3, stride 1 × 1 and no zero padding. The activation function is softplus. The 3 × 3 MAX pooling with stride 3 × 3 is applied. The third convolution layer applies 64 filters of the size 3 × 3, stride 1 × 1 and 1 × 1 zero padding. The activation function is also softplus. The 3 × 3 MAX pooling with stride 3 × 3 is used in this layer. The first fully connected layer contains 64 neurons of softplus activation. The second fully connected layer is formed of 2 neurons and applies softmax classifier, predicting the probability of normal or abnormal case of the actually analyzed mammographic image. The highest value of probability indicates the recognized class.

The details of experimental setup are presented in Fig. 7. The data set is divided into subset representing 4444 ROI images of normal cases used only in AGAN training and the rest containing 4771 normal, 1115 malignant and 888 benign images. The rest of data is restored by the trained AGAN and forms the additional channel of input data to CNN. 6774 samples of original mammograms and the same number of restored mammograms form the input data to CNN classifier. The blocks

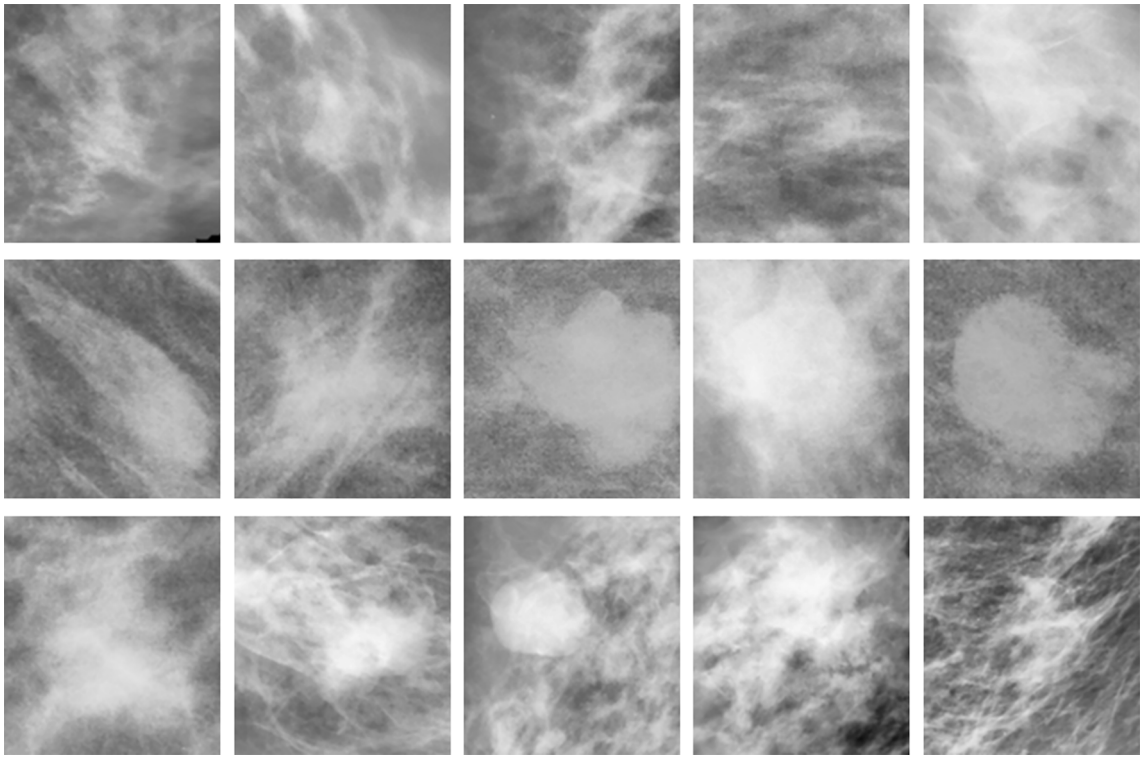

Fig. 6. The ROI examples of mammograms representing normal (upper row), benign (middle row) and malignant (bottom row) cases.

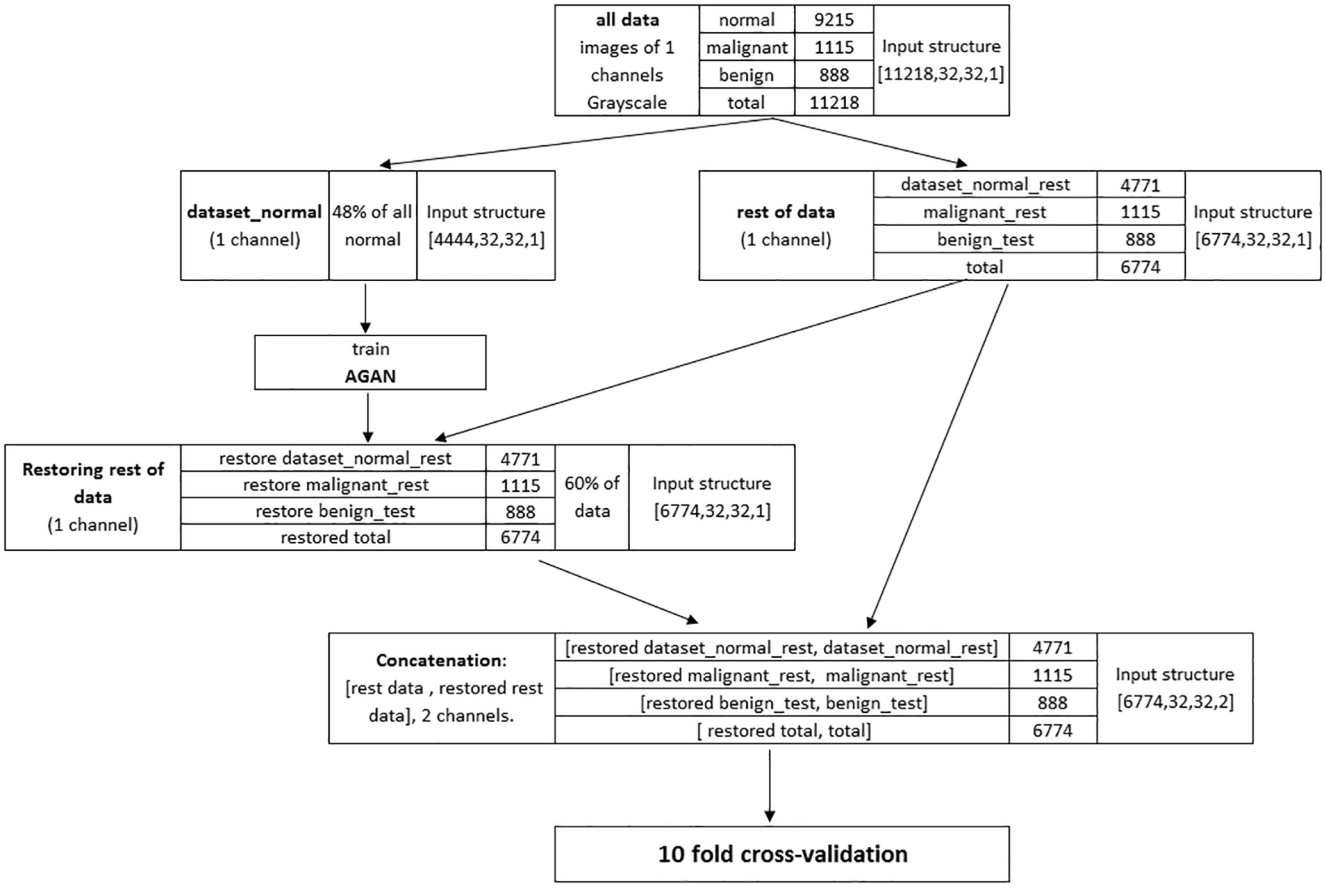

Fig. 7. The organization structure of the proposed classification system. The depth of input data tensor has been duplicated. The training data of AGAN are separated from the rest, which (after augmentation) is used in learning the CNN classifier. Process is organized in 10-fold cross validation mode.

representing input structures (for example 6774, 32,32,2 delivered to CNN) represent in sequence: the number of images in each channel, the size of image and the number of channels. Classifier is trained and tested in 10-fold cross validation mode.

All training procedures have been performed via Python 3.7 implementation of TensorFlow backend, with software executing on Windows. The process of single learning the whole system (AGAN and CNN) performed on single PC (RAM 64 GB, processor: 12 × Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60 GHz, graphic card GeForce GTX 960-4 GB) took about 2 h. This process could be significantly shortened by applying cluster system architecture and parallel computing.

4.3. Statistical results of numerical experiments

The numerical experiments were performed in 10-fold cross validation mode. The dataset which contained 4771 normal images (not taking part in AGAN training), 1115 malignant and 888 benign cases was divided into 10 subsets, each containing the same proportion of both classes, according to their populations in the database. Nine parts of samples have been used in learning procedure of CNN classifier and the last one only for testing the learned system. The experiments were repeated ten times, each time changing the test subset. The splitting into 10 parts of both normal and abnormal sets was random. The AGAN was trained separately in each fold. The test vectors enriched by the images generated by AGAN were delivered to the CNN classifier. Only testing results taken from 10-fold cross validation will be presented in the paper.

Results are evaluated based on average accuracy, sensitivity, specificity, class precision and AUC in testing mode of operation (Tan et al., 2006). Ten experimental sessions were performed (each session organized in 10-fold mode). The results of them are depicted in table 1. The first 10 rows show the actual average quality measures obtained in each session. Two last rows represent the mean and standard deviation of the results corresponding to these 10 sessions. In this table sensitivity represents the accuracy of recognition of class 1 (abnormal), i.e., the ratio of well detected cases to total number of cases representing this class. Specificity refers to detection accuracy of the normal (class 2) samples. The accuracy column refers to both classes taken together. The columns: Precision 1 and Precision 2 refer to precision of class 1 and class 2, respectively.

The results presented above are among the best for this large DDSM database. As we can see the results of particular sessions are very stable (change of accuracy from 89.37% to 90.73%, change of sensitivity from 93.10% to 93.80%). The standard deviations of accuracy, sensitivity, specificity and precision are very low (change from 0.25% to 0.82%).

The additional experiments have been also done at application of standard CNN structure without AGAN using only the original mammographic images. The results of such experiments are presented in Table 2. The structure and parameters of final CNN classifier in each case were the same. The superiority of the proposed AGAN-based solution is evident in all quality measures.

Fig. 8 shows the comparison of ROC curves corresponding to 100 implementations of deep systems in mammogram recognition (results of 10 sessions in 10-fold cross validation mode) (Matlab user manual, 2019). Red color is associated with the results of AGAN application in the system and blue the results without AGAN. Two solid curves (red and blues) represent the results of single chosen run of both systems. The advantage of the proposed solution applying AGAN is evident.

The additional experiments were carried out to compare our solution with the application of the original GAN used to supplement the data.

Table 1
The results of tumor recognition obtained in 10 different measurement sessions by the system employing AGAN.

Session	Accuracy [%]	Sensitivity [%]	Specificity [%]	Precision 1 [%]	Precision 2 [%]	AUC
1	89.67	93.52	80.48	91.94	83,91	0.9392
2	89.46	93.23	80.48	91.92	83,31	0.9424
3	89.55	93.31	80.58	91.96	83,50	0.9424
4	89.37	93.10	80.48	91.91	83,05	0.9386
5	89.73	93.80	80.03	91.79	84,41	0.9402
6	90.05	93.75	81.23	92.25	84,52	0.9457
7	89.42	93.21	80.38	91.88	83,25	0.9368
8	90.73	94.36	82.08	92.61	85,94	0.9468
9	89.71	93.73	80.13	91.83	84,30	0.9383
10	89.40	93.36	79.98	91.74	83,48	0.9393
mean	**89.71**	**93.54**	**80.58**	**91.98**	**83,97**	**0.9410**
std	0.39	0.36	0.60	0.25	0,82	0.0031

Table 2
The results of mammogram recognition obtained in 10 different measurement sessions without GAN.

session	Accuracy [%]	Sensitivity [%]	Specificity [%]	Precision 1 [%]	Precision 2 [%]	AUC
1	88.66	93.21	77.83	90.92	82,79	0.9292
2	88.65	92.96	78.38	91.11	82,37	0.9239
3	88.60	92.92	78.33	91.08	82,28	0.9275
4	87.61	92.24	76.59	90.37	80,57	0.9206
5	88.34	93.02	77.18	90.66	82,28	0.9294
6	88.29	92.52	78.23	91.01	81,44	0.9228
7	88.43	92.58	78.53	91.13	81,63	0.9272
8	87.87	92.68	76.39	90.34	81,43	0.9252
9	87.85	92.56	76.64	90.42	81,22	0.9209
10	88.03	92.12	78.28	90.99	80,66	0.9236
mean	**88.23**	**92.68**	**77.64**	**90.80**	**81,66**	**0.9250**
std	0.35	0.33	0.81	0.31	0,71	0.0030

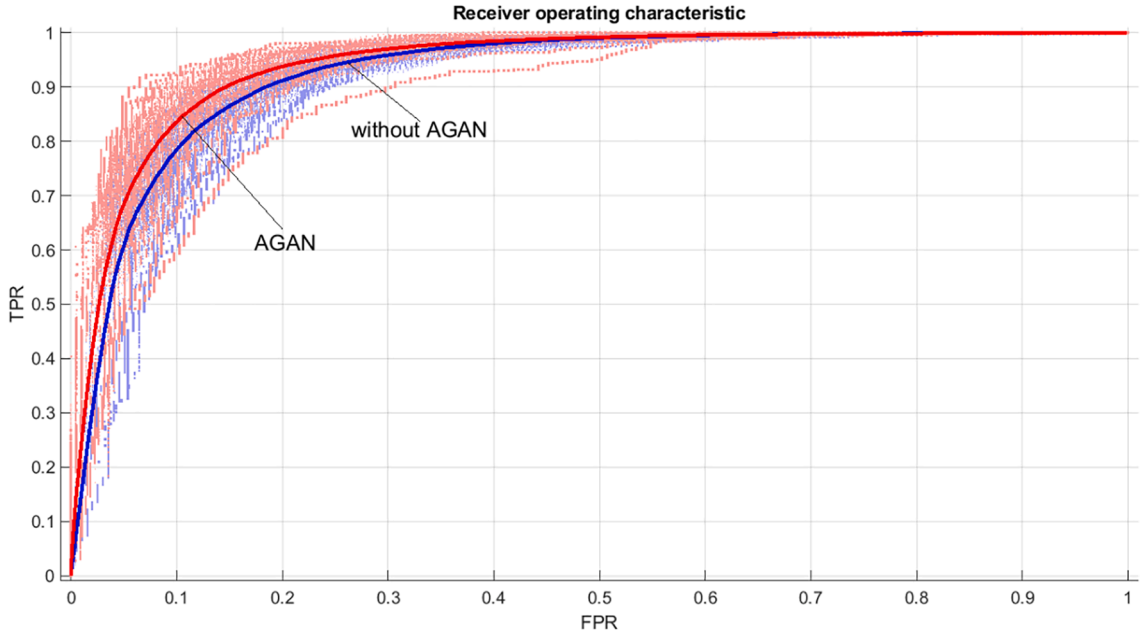

Fig. 8. ROC curves corresponding to 100 experiments. Red color is associated with the results of AGAN and blue without AGAN. Two solid curves (red and blue) represent the results of single chosen run of both systems.

The same strategy was applied to the database. The results obtained are as follows: average accuracy 88.36 ± 0.37%, sensitivity 92.57 ± 0.29%, specificity 78.34 ± 0.82%, precision of the first class 91.06 ± 0.32%, precision of the second class 81.67 ± 0.65% and AUC 0.9206 ± 0.0034. All these values are lower than those obtained by using AGAN.

4.4. Discussion

This section compares recent results in recognizing the mammographic lesions by applying deep learning methods to the DDSM base. This deep approach seems to be the most efficient nowadays. Our results presented in the previous section area among the best for this large database. Also note, that all experiments were performed in the most objective way, by applying 10-fold cross validation mode and many

repetitions of experiments. The most important quality measures: sensitivity, specificity, accuracy, precision and AUC in class recognition were determined, which is not always the case in other works. In the paper Guan and Loew (2019) only the average validation accuracy of 91.48% was reported, which is slightly higher than our test score (average 89.71% of 10 sessions organized in 10-fold cross validation) obtained for the test data that did not participate in any training. However, it is impossible to compare other quality measures of both solutions.

The quality measures of class recognition presented in this paper seem to be better than the results of other works, obtained at application of deep learning approach for the DDSM base. The accuracy of 85% and AUC = 0.91 were reported in Yi et al. (2017) by using Google Le Net system and an ensemble of 100 parallel networks for recognition of normal versus abnormal mammograms. The corresponding results presented in Samala et al. (2017) showed the best value of AUC = 0.82 (only this quality measure was given).

It should be noted, that the quality of the classification system depends strongly on the type and size of the database analyzed. In the work (Alyafi et al., 2019), for example, Deep Convolutional Generative Adversarial Networks were applied in mammography detection for the OPTIMAM mammography image database. The declared F1 score depended strongly on the population of images used in the experiments. Its value changed from 86% to over 98% with different sizes of the database (from 100 k to 1000 k).

The presented results as well as the results of other authors have shown that adding different forms of GAN augmentation can help the classifier prevent from over-fitting and improve the generalization ability of the classification system.

5. Conclusions

The paper has proposed a novel neural system for mammogram recognition by applying the deep learning approach. The significant novelty was the application of AGAN in augmenting the input set of images delivered to the CNN classification system. AGAN trained on the normal set of images is able to reconstruct the normal test images in a correct way. However, the abnormal images are reconstructed in a non-similar way. In this way we have increased the differences between the set of normal images (forming uniform set of images) and the abnormal image set, represented now by the originals and their non-similar reconstructions. In addition, we have proposed some modification to the GAN structure (called AGAN), that have led to a better distinction between normal and abnormal mammographic images. These modifications have enabled a better differentiation between normal and abnormal ROI images. The additional experiments with the system using the classical GAN have shown worse results. The average accuracy was decreased by about (0.6–1)%.

The developed system was used to recognize two classes of mammograms: normal from abnormal for the large DDSM database consisting of 11,218 ROI mammographic images. The results have shown the advantage of the proposed approach over other known methods of mammogram recognition (including also other deep learning approaches). The average quality measures (accuracy, sensitivity, specificity, AUC) in recognizing abnormal cases (malignant plus benign versus healthy) belong to the best for this large DDSM database. The presented system outperforms the known state-of-the-art systems and therefore has great potential to advance the field of research.

There are also some limitations of our method. AGAN is learned only on normal data, which obviously reduces the amount of data that can be effectively used to teach the final classifier. However, this is not a serious problem, as there is usually much more “normal” data than the abnormal data.

Future research in this area will be focused on the following tasks:

- Increasing the number of classifiers by incorporating different CNN architectures (Szegedy et al., 2016) and coupling them in an ensemble. The ensemble thus built will represent the expert system that will be able to view this recognition problem from different points of view, which hopefully will increase the final accuracy.
- In future works we could also apply our system to recognize the benign and malignant ROIs instead of abnormal versus normal ROIs.
- Additional experiments should be conducted to check the performance of the system on other databases of mammograms and to implement this procedure on a large scale in medical practice for supporting the breast cancer diagnosis.

CRediT authorship contribution statement

B. Swiderski: Conceptualization, Methodology, Formal analysis, Supervision. **L. Gielata:** Software, Validation. **P. Olszewski:** Software, Investigation. **S. Osowski:** Supervision, Writing - original draft, Writing - review & editing. **M. Kołodziej:** Visualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Alyafi, B., Diaz, O., & Marti, R. (2019). DCGANs for realistic breast mass augmentation in X-ray mammography. arXiv:1909.02062v1 [eess.IV].

Christoyiani, I., Koutra, A., Dermata, E., & Kokkinakis, G. (2002). Computer aided diagnosis of breast cancer in digital mammograms. *Computerized Medical Imaging and Graphics, 26*, 309–319.

Dhahbi, S., Barhoumi, W., & Zagrouba, E. (2015). Breast cancer diagnosis in digitized mammograms using curvelet moments. *Computers in Biology and Medicine, 64*(1), 79–90.

Donahue, J., Krähenbühl, P., & Darrell, T. (2017). Adversarial feature learning. *Proc. international conference on learning representations, arXiv:1605.09782v6 [cs.LG]*.

Elfarra, B. K., & Abuhaiba, I. S. (2012). New feature extraction method for mammogram computer aided diagnosis. *International Journal of Signal Processing, Image Processing and Pattern Recognition, 6*(1), 1–81.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., & Bray, F. (2012). Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN. *International Journal of Cancer, 136*(5), E359–E386.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Massachusetts, USA: MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair S., & et al. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems. arXiv:1406.2661.

Guan, S., & Loew, M. (2019). Using generative adversarial networks and transfer learning for breast cancer detection by convolutional neural networks. In Proc. SPIE 10954. Medical imaging 2019: Imaging informatics for healthcare, research, and applications 10954C (25 March 2019). http://doi.org/10.1117/12.2512671.

Heath, M., Bowyer, K., Kopans, D., Moore R., & Kegelmeyer, P. (1998). The digital database for screening mammography. In Digital mammography (pp. 457–460). Netherlands: Springer.

Jiang, M., Zhang, S., Li, H., & Metaxas, N. (2015). Computer-aided diagnosis of mammographic masses using scalable image retrieval. *IEEE Transactions on Biomedical Engineering, 62*(2), 783–792.

Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends in Machine Learning, 12*, 307–392. https://doi.org/10.1561/2200000056.

Kooi, T., & Litjens, G. (2017). Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis, 35*, 303–312.

Korkinof_D., Rijken, T., O'Neill, M., Yearsley, J., Harvey, H., Glocker,B. (2019). High-resolution mammogram synthesis using progressive generative adversarial networks. arXiv:1807.03401v2 [cs.CV].

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Image net classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1–9.

Kurek, J., Swiderski, B., Osowski, S., Kruk, M., & Barhoumi, W. (2019). Deep learning versus classical neural approach to mammogram recognition. *Bulletin of the Polish Academy of Sciences, Technical Sciences, 67*(3), 517–525.

Matlab user manual. (2019). *MathWorks*. USA: Natick.

Odena, A., Dumoulin, V., & Olah, C. (2016). Deconvolution and checkerboard artifacts, http:// distill.pub/2016/deconv-checkerboard.

Samala, R. K., Chan, H. P., & Hadjinski, L. M. (2017). 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.

Swiderski, B., Osowski, S., Kurek, J., Kruk, M., Lugowska, I., Rutkowski, P., & Barhoumi, W. (2017). Novel methods of image description and ensemble of classifiers in application to mammogram analysis. *Expert Systems with Applications, 81*, 67–78.

Szegedy, C., Ioffe, S., & Vanhoucke V. (2016). Inveption-v4, Inception-ResNet and the impact of residual connections on learning. arXiv:1602.07261v2.

Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Boston: Pearson Education Inc.

Teare, P., Fishman, M., Benzaquen, O., & Toledano, E. (2017). Malignancy detection on mammography using dual deep convolutional neural networks and genetically discovered false color input enhancement. *Journal of Digital Imaging, 30*, 499–505.

Yi, D., Sawyer, R.L., Cohn, III D., Dunmon, J., & Lam, C. (2017). Optimizing and visualizing deep learning for benign/malignant classification in breast tumors. NIPS 2016, arXiv preprint, arxiv.org.

Yu, H., Huang, T. Z., Deng, L. J., & Zhao, X. L. (2016). Superresolution via fast deconvolution with kernel estimation. *Eurasip Journal on Image and Video Processing, 2017*, 3. https://doi.org/10.1186/s13640-016-0125-6.

Wu, E., Wu, K., Cox, D., & Lotter W. (2018). Conditional infilling GANs for data augmentation in mammogram classification. In Image analysis for moving organ, breast, and thoracic images (part of lecture notes in computer science book series (LNCS, Vol. 11040, pp. 98–106).

Yia, X., Walia, E., & Babyn, P. (2019). Generative adversarial network in medical imaging: a review. arXiv:1809.07294v4 [cs.CV].

Yoo, K. M., Lee, H., Dernoncourt, F., Bui, T., Chang, W., Lee, S. G. (2020). Variational hierarchical dialog autoencoder for dialog state tracking data augmentation. arXiv: 2001.08604v2 [cs.CL].

Zhu, W., Xiangy, X., Trany, T. D., Hagery, G. D., & Xie, X. (2017). Adversarial deep structured nets for mass segmentation from mammograms. arXiv:1710.09288v2 [cs. CV].