Metrol. Meas. Syst., Vol. 32 (2025), No. 3
DOI: 10.24425/mms.2025.154682

PAN
POLSKA AKADEMIA NAUK
METROLOGY AND MEASUREMENT SYSTEMS
Index 330930, ISSN 0860-8229
www.metrology.wat.edu.pl
M&MS

DEEP NETWORK ENSEMBLE IN BREAST CANCER RECOGNITION

Fabian Gil[1)], Stanisław Osowski[1,2)]

1) *Military University of Technology, Faculty of Electronics, Institute of Electronic Systems, ul. gen. Sylwestra Kaliskiego 2, 00-908 Warsaw, Poland* (✉ *fabian.gil@wat.edu.pl*)
2) *Warsaw University of Technology, Faculty of Electrical Engineering, pl. Politechniki 1, 00-661 Warsaw, Poland*

Abstract

This paper presents an ensemble-based approach utilizing convolutional neural network (CNN) classifiers for the recognition of breast cancer from mammogram images. The proposed method integrates several well-performing CNN models, each selected based on its individual efficiency, into a unified ensemble framework. The classifiers are aggregated using a majority voting strategy that considers the predicted class probabilities to make the final decision. This ensemble technique aims to enhance robustness, reduce the impact of misclassifications by individual models, and improve overall diagnostic reliability. The system was evaluated on a dataset containing mammograms categorized into three diagnostic classes: malignant, benign, and normal. Numerical experiments demonstrated that the proposed approach significantly improves classification performance compared to both classical machine learning methods and standalone CNN models. The ensemble achieved higher accuracy, sensitivity, and specificity, particularly in distinguishing between benign and malignant cases—a critical challenge in breast cancer diagnostics. These results highlight the potential of CNN ensembles in supporting more accurate computer-aided diagnosis in breast cancer screening.

Keywords: breast cancer, mammogram, CNN, ensemble of classifiers.

1. Introduction

Breast cancer is the most commonly diagnosed cancer in females. According to the statistics of the GLOBOCAN 2020, it leads to a 6.9% mortality rate worldwide [1]. The typical way of discovering this type of cancer is mammography, the image of an x-ray of the breast [2]. Analyzing this image, it is possible to find breast cancer early, before there are signs or symptoms of the disease. Early discovery makes the treatment easier and, in this way, lowers the risk of dying.

Screening mammography is nowadays the most used method to detect cancer in the early stage. Based on its results the radiologists classify the lesion into either benign or malignant classes. However, due to the massive scale of screening this process needs support, which can be provided by computer-aided systems. Nowadays, such systems use the newest achievements in machine learning, especially the application of the neural networks of the deep structure [3-6].

Many different solutions have been proposed in the past. The classical approaches are based on the manual extraction of the image descriptors, transforming them into diagnostic features and applying them as the input attributes to the neural classifiers, responsible for class recognition [3]. The most difficult problem in this approach is to find the proper method of image description. Different propositions based on texture, statistical measures of color distribution, or geometric characterization have shown their limitations.

The deep learning approach is the way to define the image features automatically [4, 5, 7-9]. The process of their creation is based on the multilayer neural network structure.

Article history: received April 16, 2025; revised June 8, 2025; accepted July 25, 2025; available online August 13, 2025.

The analyzed image is subjected to multiple processing in the cascading layers, using such operations as convolution, *Rectified Linear Unit* (ReLU) activation, pooling, etc. As a result, the original input image is converted in the last locally connected layer to many small-size images representing the features delivered to the output classifier stage (usually the softmax layer), which is responsible for the final classification [3, 4].

This paper shows the application of the deep learning approach based on *convolutional neural networks* (CNN). The CNN classifiers represent the typical multilayer structure responsible for the generation of the diagnostic features and at the same time for the final classification using these features [3, 4]. To increase the generalization ability of the system the team of many different CNN structures, called the ensemble, is proposed in this paper.

The important point in this approach is the creation of an efficient ensemble, that is the proper choice of its members. The paper proposes the original method, considering the results of the candidate units, related to accuracy, sensitivity, and precision. The ensemble created in this way has shown the increased quality of the class recognition in breast cancer pathology. The results of its application outweigh the results presented in the other papers for the same base of breast cancer.

The paper is organized as follows. Section 2 is devoted to the comprehensive literature review devoted to breast cancer recognition. The next section introduces the *Digital Database for Screening Mammography* (DDSM) [2] database of mammograms. In section 4 we introduce the details of creating the ensemble composed of CNN classifiers. In section 5 the results of breast cancer recognition at application of the developed ensemble are presented and discussed. Conclusions and further study directions are given in the last section.

2. State of research in breast cancer recognition

The computerized approach to the recognition of breast cancer based on mammograms has been investigated in many works in the past. The proposed methods are based either on conventional structures, like multilayer perceptron, support vector machine, or decision trees [10, 11], or on the deep approaches applying different solutions of CNN [12-17]. The quality of results depends on the proposed method, the database investigated as well as the number of samples used in experiments. To be most objective we will limit the comparison of our results to the same DDSM database investigated by different authors.

The paper [10] has applied two forms of classical classifiers *support vector machine* (SVM) and *multi-layer perceptron* (MLP) supplied by manually generated and selected features. The proposed system was used to recognize small subsets of malignant (337) from benign (314) mammogram images, selected from the DDSM database. The best-declared results obtained for these subsets were as follows: sensitivity 98.22%, specificity 97.45, accuracy 97.85%. However, the significant question is how the samples used in experiments have been selected from the entire dataset (337 chosen from 1115 malignant and 314 from 888 benign cases). Moreover, the presented results have been related to very small subsets of data chosen randomly from the whole database. Therefore, they are not representative of the problem.

The paper [11] has proposed an ensemble of classical classifiers composed of SVM, *k-nearest neighbors* (KNN), decision trees, and autoencoder. All are supplied by the features defined in various ways. The method was applied to recognize the malignant from the rest (benign + normal) as well as to recognize lesions (malignant + benign) from normal cases. The reported sensitivity of malignant detection was 83.3%, specificity 79.8%, accuracy 80.2%, and the *area under the curve* (AUC) value 0.890. In recognition of cancer lesions from the normal cases, the results were as follows: sensitivity 82.9%, specificity 84.8%, accuracy 84.5%, and AUC 0.920.

Nowadays most papers apply deep learning techniques in image recognition. The paper [12] has proposed the application of a CNN network supported by a modified GAN to recognize abnormal versus normal cases. The declared results are as follows: sensitivity 93.54%, specificity 80.58%, accuracy 89.71%, and AUC 0.9410.

The paper [13] has presented the results of the recognition of malignant from benign cases on the DDSM repository by using Alexnet and Googlenet. The best of them is related to the Googlenet model which resulted in a sensitivity of 93.4% and precision of 92.4%.

Ansar et al. [14] have applied a pre-trained Mobilenet CNN using transfer learning to recognize malignant from benign on the DDSM database showing an accuracy of 86.8%.

The paper [15] has proposed solutions based on deep CNN classifiers to recognize breast cancer lesions from normal cases, declaring 95.53% accuracy.

The paper [16] has presented an approach based on data integration, feature extraction, and CNN model development and applied it to the recognition of malignant and benign lesions. The declared quality measures are as follows: accuracy of 96%, sensitivity and precision of 95%, AUC of 0.96. The results correspond only to the small subset of data from the repository.

The paper [17] has shown an ad-hock-built ensemble of deep CNN classifiers, showing its usefulness in medical image recognition. However, its results are of limited accuracy due to the lack of efficient way in the ensemble creation.

The interesting direction of research in breast cancer is breast mass segmentation in mammography. In this case very useful is the special structure of CNN network called U-Net. The paper [18] has shown that combining this form with transformers allows for achieving superior accuracy, dice similarity coefficient, and intersection over union in the DDSM database.

This paper develops the advanced procedure for forming the ensemble composed of a few CNN classifiers of different architectures. The numerical experiments performed for the recognition of three classes of breast cancer have shown increased accuracy compared to the results presented for the same database in other papers.

3. Database

The numerical experiments are performed using the publicly available database DDSM [2] created by the medical teams from a few institutions: the Massachusetts General Hospital, the University of South Florida, Sandia National Laboratories, Washington University School of Medicine, Wake Forest University School of Medicine (Departments of Medical Engineering and Radiology), and Sacred Heart Hospital and ISMD, Incorporated. The database is maintained by the University of South Florida to keep it accessible on the web [2].

It contains 2802 examples, composed of 4 mammograms: left and right breast from above representing the Cranial-Caudal view and oblique representing the Medio-Lateral-Oblique view. Each mammogram is associated with a description of its abnormality. The base covers three types of mammograms: normal cases represented by 9215 cases, lesions of benign type (888 cases), and malignant state (1115 cases).

The images taking part in numerical experiments are in the form of Region of Interest (ROI), created as a binary mask image. They have been prepared by the medical experts and offered for public use. The size of the images is 128x128 pixels. Fig. 1 presents some exemplary images representing 3 types of abnormal states, treated as the classes.

High similarity between images representing different classes can be observed. At the same time, the similarity within the same class is limited. This is well visible in the values of the statistical description of the images. Table 1 presents the statistical characterization of the images in the analyzed database. They are given in the form of mean value, standard deviation, kurtosis, and energy (the sum of the squared pixel values). The distribution of samples is far from normal (kurtosis smaller than 3). Within each parameter, we observe the significant value of standard deviation.

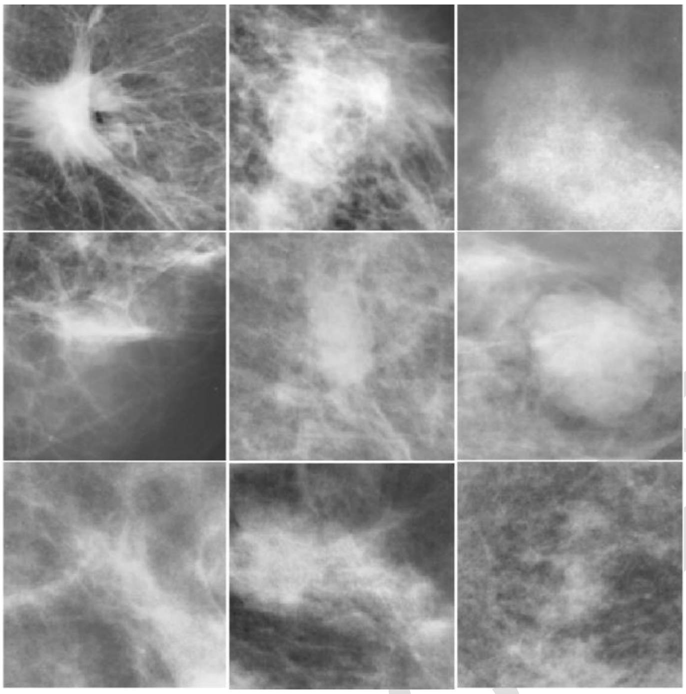

Fig. 1. The exemplary ROI images representing different states of abnormality: the upper row – malignant state, the middle row – benign state, and the lower row – normal cases.

Table 1. Statistical parameters describing the representatives of all three classes of mammograms.

Class	**Mean**	**Std**	**Energy**	**Kurtosis**
Malignant	160.79 ± 19	33.3 ± 9.33	27412 ± 6051	2.48 ± 0.69
Benign	158.69 ± 19.64	31.34 ± 9.34	26638 ± 6200	2.39 ± 0.51
Normal	160.79 ± 16.61	27.44 ± 9.33	26968 ± 5185	2.79 ± 0.99

Very interesting is also the local structural similarity (SSIM) of images within the class and between classes. This function implemented in Matlab [7] was applied in this study. It is calculated as the structural similarity value for each pixel based on its relationship to the pixels in its 11x11 neighborhood. These values calculated within class, and between classes for the DDSM base are presented in Table 2.

Table 2. The structural similarity values between the image representatives of classes.

SSIM measure	**Mean**	**Std**
Malignant	0.3896 ± 0.0609	0.0638 ± 0.0157
Benign	0.4193 ± 0.0554	0.0585 ± 0.0126
Normal	0.4386 ± 0.0501	0.0527 ± 0.0089
Malignant vs benign	0.4037 ± 0.0530	0.0664 ± 0.0140
Malignant vs normal	0.4085 ± 0.0641	0.0486 ± 0.0112
Benign vs normal	0.4250 ± 0.0556	0.0510 ± 0.0100

The mean value of the structural similarity is very similar for the images within the class and between classes. This is the confirmation of the difficulty in recognition of class membership of the images. Additional difficulty follows from a large class imbalance (9215 normal cases, against 888 lesions of benign type and 1115 of malignant state). We have decided not to interfere with the contents of the sets to check the tolerance of our system to this problem.

4. Deep neural network ensemble

To solve the problem of image recognition we propose the application of the CNN classifiers organized in the form of an ensemble [19]. The CNN is a multilayer deep structure defined especially for image recognition. It is composed of many layers organized in a feedforward manner. The CNN architecture is responsible at the same time for the automatic generation of image features and the final classification.

The first succeeding layers are locally connected. They apply such operations as linear convolution with small-size moving filters, the ReLU nonlinear activation operating on the convolution results, normalization of data, and the pooling operation responsible for the reduction of the size of images. Many output images are created simultaneously to compensate for the loss of information associated with the size reduction. The sets of images in each layer are represented in the form of a tensor. The size-reduced images in the succeeding layer try to capture the most significant features of the input images.

The images of the last locally connected layer are flattened and converted to the vectorial form by reshaping or global pooling [3-5]. This vector represents the features of the analyzed images. The elements of it are the input attributes to the fully connected final classifier structure, called softnet with the softmax function as shown in Fig. 2.

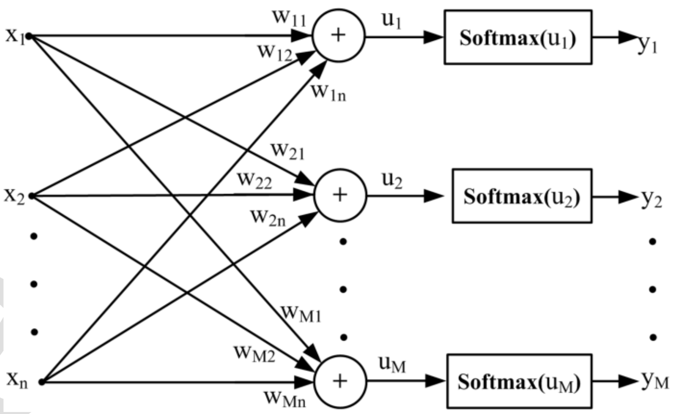

Fig. 2. The softnet classifier with softmax output function used in CNN architectures.

The signals $u_i(\mathbf{x})$ of the softnet represent the regression form described by

$$u_i(\mathbf{x}) = \sum_j w_{ij} x_j + w_{i0}. \tag{1}$$

The softmax activation function used on the output of softnet is described by

$$y_i(\mathbf{x}) = \mathrm{softmax}(u_i(\mathbf{x})) = \frac{\exp(u_i)}{\sum_{k=1}^{M} \exp(u_k)}. \tag{2}$$

It represents the probability of membership of the vector $\mathbf{x}$ to the ith class for $i = 1, 2, \ldots, M$. The position of the highest value of this probability determines the final class membership.

Many different CNN architectures starting from the first proposed Alexnet [5] have been developed nowadays [3, 7]. The important problem that occurs at increasing the depth of CNN is the observed process of vanishing/exploding gradients, which hampers the convergence and makes accuracy saturated. The improvement of this is the introduction of the residual connection (Fig. 3) proposed in the work [8]. Thanks to such additional connection it was possible to ease the training of networks that are substantially deeper than those used previously.

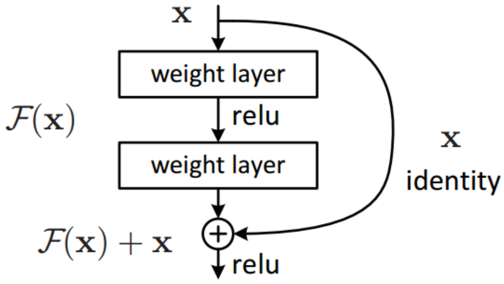

Fig. 3. The idea of the residual connection introduced in Resnet-type architectures [8]. *F*(x) represents the stacked two nonlinear layers fit of the input data.

The other modifications are aimed at minimizing the number of adapted parameters or increasing the sensitivity to the statistical properties of the particular regions of the pixel distribution in the image. The last idea was implemented by introducing a few different-sized filters working simultaneously to create the final output image. The typical example proposed in [8] is in the form of a cell called inception (Fig. 4),

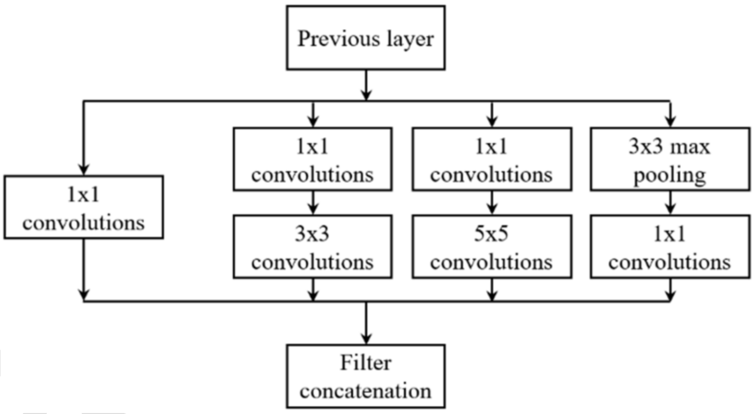

Fig. 4. The structure of the inception cell employing filters of different sizes: 1x1, 3x3, and 5x5 [9].

Such structure allows the network to capture information at various scales and complexities. The smallest filters are responsible for small size details in the analyzed image and the largest for image regions of larger size. Thanks to such an organization more information is passed from the preceding layer to the next one. The use of 1x1 convolutions serves as a method for reducing computational complexity and the number of parameters without losing depth in the network.

The differences in signal processing included in the existing CNN architecture allow us to pay attention to various aspects of the analyzed images, and lead to the diversified conclusion concerning the class membership. This creates the space to arrange them in the form of an ensemble since the independence of the unit operation is the most significant condition for the team to operate properly [19].

Different CNN architectures are used in the creation of the ensemble [3, 10, 11]. To accelerate the training phase, the transfer learning of the pre-trained networks is applied. The fine-tuning of pre-trained architectures by using the actual DDSM dataset involves the adaptation of the softnet parameters and the weights of kernels in 2/3 of the locally connected layers closest to the softnet (the first 1/3 of the locally connected layers are frozen). The ADAM algorithm [20] was used with an adaptive learning rate and a mini-batch size of 30. No augmentation procedure was applied.

In our paper, we have tried 19 pre-trained CNN classifiers available actually in Matlab [7]. Their names with numerical notations are presented in Table 3. These notations will be used to represent them in the team. All networks are of different architectures, differing by the size of input images, number of layers (from a few up to several hundred), type of filter arrangements in the layers, presence or lack of residual connection, etc. These differences provide a good premise for their independent functioning, which is important for the ensemble.

Table 3. The CNN classifiers used in further experiments.

Classifier notation	CNN architecture
1	Squeezenet
2	Googlenet
3	Inceptionv3
4	Densenet201
5	Mobilenetv2
6	Resnet18
7	Resnet50
8	Resnet101
9	Xception
10	Inceptionresnetv2
11	Shufflenet
12	Nasnetmobile
13	Nasnetlarge
14	Darknet19
15	Darknet53
16	Efficientnetb0
17	Alexnet
18	Vgg16
19	Vgg19

Fig. 5 depicts the proposed general structure of an ensemble composed of such CNN units. The number *N* of the members and their composition is subjected to the choice, based on the introductory experiments.

The analyzed image is simultaneously delivered to all CNN classifiers. Each of them works out their own decision of class membership. Their verdicts in the form of the class probability values are merged by majority voting. At *N* classifiers the summed *i*th class probability is calculated.

$$\alpha(i) = \sum_{j=1}^{N} prob(i,j) \tag{3}$$

The index *i* corresponding to the maximum value of $\alpha(i)$ represents the recognized class.

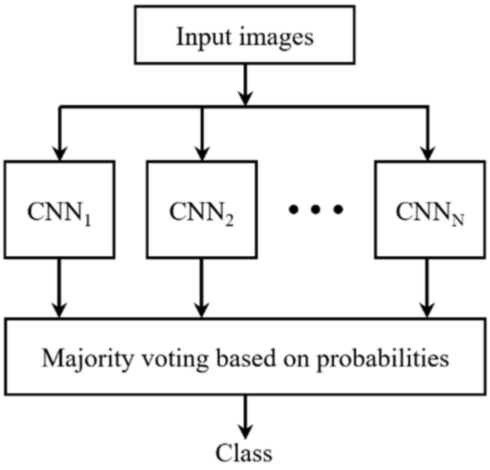

Fig. 5. The general arrangement of CNN classifiers in an ensemble.

The most important problem is to determine the proper selection of the particular units and their population. This will be done by analyzing the quality of each CNN classifier and selecting the best for the team composition. The typical quality measures used in the assessment of the classifiers are the average accuracy in *all class recognition* (ACC), *true positive rate* (TPR) representing class sensitivity, *true negative range* (TPR) called specificity, *positive class precision value* (PPV), *negative class precision value* (NPV), and F1 measure [21]. All of them are calculated based on the confusion matrix [3, 21]. Additionally, the area under the ROC curve (AUC) is treated as a good tool to compare different solutions of the classifiers [3, 21].

5. Results of numerical experiments

The numerical experiments have been directed to solve three tasks, two of them represent the 2-class problem and one case 3-class problem:

- Recognition of the malignant state of the cancer from the benign, except normal state (2-class problem).
- Recognition of lesions (malignant and benign represent one class) from the normal state (2-class problem).
- Recognition of 3 classes: class 1 – malignant, class 2 – benign, and class 3 – normal state.

The significant imbalance of classes is visible in the second and third cases since the normal class is the most numerous (9215 images) compared to 1115 images representing malignant and 888 benign cases. This makes the recognition problem more difficult.

All experiments have been performed using the K-fold cross-validation [3], the most objective method of assessment of the solution. In the experiments, the value of $K = 5$ was applied. Thanks to this approach all data are used also in the testing phase. The numerical results will be presented only for testing data not taking part in the learning phase of the networks.

5.1. Recognition of malignant from benign lesions

Table 4 presents the results of individual CNN architectures in recognition of malignant (class 1) from benign (class 2). They are given in the form of statistical results concerning AUC, ACC, TPR, TNR, PPV, and NPV corresponding to the testing images in the 5-fold cross-validation mode. A high variety of results is observed. The best units concerning all considered quality measures belong to Nasnetlarge, Densenet, and Resnet101 (in bold). The worst results have been obtained by using Alexnet and Squeezenet.

Table 4. The results of malignant versus benign cases recognition using individual CNN classifiers.

CNN	AUC	ACC	TPR	TNR	PPV	NPV
Squeezenet	0.8254	0.7474	0.7848	0.7005	0.7669	0.7216
Googlenet	0.8911	0.7958	0.8413	0.7387	0.8017	0.7875
Inceptionv3	0.9696	0.9006	0.9058	0.8941	0.9149	0.8832
Densenet201	**0.9735**	**0.9146**	**0.9220**	**0.9054**	**0.9245**	**0.9024**
Mobilenetv2	0.9381	0.8557	0.8933	0.8086	0.8542	0.8578
Resnet18	0.9473	0.8752	0.8951	0.8502	0.8824	0.8658
Resnet50	0.9682	0.9061	0.9220	0.8863	0.9105	0.9005
Resnet101	**0.9695**	**0.9151**	**0.9291**	**0.8975**	**0.9193**	**0.9098**
Xception	0.9380	0.8542	0.8673	0.8378	0.8704	0.8341
Inceptionresnetv2	0.9688	0.8917	0.8969	0.8851	0.9074	0.8724
Shufflenet	0.9356	0.8552	0.8924	0.8086	0.8541	0.8568
Nasnetmobile	0.8874	0.7873	0.8179	0.7489	0.8035	0.7661
Nasnetlarge	**0.9782**	**0.9201**	**0.9130**	**0.9291**	**0.9417**	**0.8948**
Darknet19	0.9504	0.8707	0.9148	0.8153	0.8615	0.8840
Darknet53	0.9591	0.8817	0.9013	0.8570	0.8878	0.8737
Efficientnetb0	0.8677	0.7703	0.8691	0.6464	0.7553	0.7972
Alexnet	0.7714	0.7024	0.7892	0.5935	0.7091	0.6916
Vgg16	0.9361	0.8462	0.8861	0.7962	0.8452	0.8477
Vgg19	0.9470	0.8637	0.9148	0.7995	0.8514	0.8820

Table 5 presents the statistical characterization of results for the set of all 19 CNN classifiers. These values confirm the significant differences existing among analyzed architectures of CNN classifiers. Irrespective of the quality measure the distance between the worst (minimum value) and the best units (maximum value) is very large.

Table 5. Statistical characterization of the results of class recognition of the whole set of CNN classifiers.

Quality measure	Mean	Std	Minimum	Maximum
AUC	0.9275	0.0555	0.7714	0.9782
ACC	0.8502	0.0618	0.7024	0.9201
TPR	0.8819	0.0437	0.7848	0.9291
TNR	0.8105	0.0909	0.5935	0.9291
PPV	0.8559	0.0635	0.7091	0.9417
NPV	0.8436	0.0625	0.6916	0.9098

Different arrangements of the CNN classifiers have been tried to find the most efficient ensemble composition. The selection process of the ensemble members is based on the performance quality of the individual units. According to medical practice the most important quality measures belong to the average accuracy ACC and sensitivity TPR. Additionally, the AUC is also included in the selection procedure as a unique tool for comparing the classifiers. The results corresponding to these three measures have been considered in the selection process. Table 6 provides the details of the possible choices that were found to be potentially best in the introductory experiments. The CNN models participating in the subsequent teams are coded by numbers, as shown in Table 3. The last set contains all 19 units that form an ensemble.

Table 6. The compositions of the analyzed ensembles.

Ensemble	Quality measure	Composition of CNN ensemble
1	AUC	[3,4,13]
2	ACC	[4,8,13]
3	TPR	[4,7,8]
4	AUC	[3,4,8,10,13]
5	ACC	[3,4,7,8,13]
6	TPR	[4,7,8,14,19]
7	AUC, ACC	[3,4,7,8,10,13,15]
8	TPR	[3,4,7,8,13,14,19]
9	AUC, ACC, TPR	[3,4,5,6,7,8,10,13,14,15,19]
10	AUC, ACC	[2,3,4,5,6,7,8,9,10,11,13,14,15,18,19]
11	TPR	[3,4,5,6,7,8,9,10,11,13,14,15,16,18,19]
12	-	All 19 units

Table 7 depicts the results of these ensembles after aggregation of the results of its members based on the class probability principle.

Table 7. The results of different compositions of an aggregated ensemble in recognition of malignant versus benign cases.

Ensemble	AUC	ACC	TPR	TNR	PPV	NPV
1	0.9836	0.9301	0.9300	0.9302	0.9436	0.9137
2	0.9838	0.9356	0.9390	0.9313	0.9449	0.9240
3	0.9801	0.9331	0.9417	0.9223	0.9383	0.9265
4	0.9846	0.9336	0.9363	0.9302	0.9439	0.9208
5	0.9846	0.9356	0.9390	0.9313	0.9449	0.9240
6	0.9798	0.9311	0.9462	0.9122	0.9312	0.9310
7	**0.9843**	**0.9381**	**0.9445**	**0.9313**	**0.9452**	**0.9392**
8	0.9833	0.9351	0.9423	0.9223	0.9386	0.9307
9	0.9824	0.9356	0.9444	0.9245	0.9402	0.9298
10	0.9803	0.9316	0.9390	0.9223	0.9382	0.9233
11	0.9807	0.9316	0.9417	0.9189	0.9358	0.9262
12	0.9771	0.9281	0.9399	0.9133	0.9316	0.9237

Irrespective of the composition, all of them (except the ensemble composed of all 19 units) represent improved results compared to the best individual. Their quality measures are very similar to each other. However, ensemble number 7 (denoted in bold) composed of Inceptionv3, Densenet201, Resnet50, Resnet101, Inceptionresnetv2, Nasnetlarge, and Darknet19 might be treated as the best.

5.2. Recognition of cancer lesions from normal cases

In these experiments, the malignant and benign cases form class one, and normal cases the opposite class. Introductory experiments have been performed to find the best units as possible members of the ensemble. Based on these results the following six CNN classifiers presented in Table 8 have been selected for the ensemble. All of them represent similar quality values. The results of their aggregation by the majority voting based on the probability of class membership are presented in Table 9.

Table 8. The results of 6 individual CNN classifiers selected for the ensemble in recognition of cancer lesions from normal cases.

CNN	AUC	ACC	TPR	TNR	PPV	NPV
Inceptionv3	0.9969	0.9853	0.9641	0.9899	0.9541	0.9922
Densenet201	0.9868	0.9558	0.9396	0.9593	0.8339	0.9865
Resnet50	0.9921	0.9532	0.9596	0.9518	0.8123	0.9908
Resnet101	0.9940	0.9641	0.9661	0.9636	0.8524	0.9924
Inceptionresnetv2	0.9972	0.9736	0.9785	0.9725	0.8857	0.9952
Darknet53	0.9920	0.9615	0.9451	0.9651	0.8546	0.9878

Table 9. The aggregated results of the ensemble composed of six selected CNN classifiers depicted in Table 8.

AUC	ACC	TPR	TNR	PPV	NPV
0.9975	0.9877	0.9786	0.9907	0.9578	0.9943

The ensemble has improved the best individual results of the quality measures, however, this statistical improvement is of a limited level. The most significant advantage is the radical reduction of the misclassification cases. The total misclassification number of the best individual classifier (Inceptionv3) equal to 165 was reduced to only 138 by an ensemble. The most dangerous cases for patients (recognizing cancer as normal) have been reduced from 72 (the best CNN classifier) to only 52 by the ensemble.

5.3. Recognition of three classes in breast cancer

The last series of experiments have been conducted in recognition of three classes: class 1 - malignant, class 2 – benign, and class 3 – normal case. Similarly to the previous tasks, the first step was to find the best composition of the ensemble. As a result, the set of 6 best CNN classifiers has been selected. The results of their application in class recognition are presented in Table 10.

They depict the considered quality measures (ACC, AUC, TPR, TNR, PPV, and NPV) as well as the confusion matrix and the number of misclassifications in recognition of these three classes.

The best classifier is Inceptionv3 characterized by the largest value of ACC and the smallest total number of misclassifications. Although the values of quality measures seem to be similar for all CNN classifiers, the most important difference is observed in the total number of misclassifications. The worst classifier (Resnet50) committed 561 errors, while the best one (Inceptionv3) was only 318.

Aggregation of these results using majority voting of all classifiers has improved the final results of recognition. They are presented in Table 11.

The results of aggregation have shown an improvement in all quality measures. This is well seen by comparing the confusion matrices of the best CNN classifier and the aggregated ensemble. Such a comparison is depicted in Table 12. The misclassification errors in all three classes have been significantly reduced by the ensemble. The highest reduction rate is observed in the recognition of malignant cases (reduction from 109 to 75) and normal cases (reduction from 93 to 32).

The highest misclassification rate is observed in recognition of benign cases (this is the intermediate state between malignant and normal class).

Table 10. The results of the 3-class recognition problem using individual CNN classifiers selected for the ensemble.

CNN	**Class**	**AUC**	**ACC**	**TPR**	**TNR**	**PPV**	**NPV**	**Confusion matrix**			**Number of errors in classification**
Inceptionv3	1	0.9938	**0.9717**	0.9022	**0.9897**	0.9063	0.9892	1006	61	48	109
	2	0.9899		0.8694	0.9892	0.8733	0.9888	62	772	54	116
	3	0.9958		0.9899	0.9491	0.9889	0.9534	42	51	9122	93
Densenet201	1	0.9909	0.9671	0.9184	0.9838	0.8620	**0.9909**	1024	51	40	91
	2	0.9852		0.8525	0.9892	0.8711	0.9873	78	757	53	131
	3	0.9935		0.9840	0.9536	0.9898	0.9285	86	61	9068	147
Resnet50	1	0.9880	0.9500	0.9130	0.9751	0.8016	0.9902	1018	77	20	97
	2	0.9810		0.8367	0.9766	0.7543	0.9858	98	743	47	145
	3	0.9922		0.9654	0.9666	**0.9925**	0.8585	154	165	8896	319
Resnet101	1	0.9893	0.9531	0.9157	0.9758	0.8071	0.9906	1021	64	30	94
	2	0.9829		0.8615	0.9801	0.7878	0.9880	77	765	46	123
	3	0.9925		0.9665	0.9621	0.9915	0.8618	167	142	8906	309
Inceptionresnetv2	1	0.9933	0.9685	0.8619	0.9892	0.8981	0.9848	961	86	68	154
	2	0.9905		0.8390	**0.9897**	0.8754	0.9862	73	745	70	143
	3	**0.9967**		**0.9939**	0.9311	0.9852	0.9708	36	20	9159	56
Darknet53	1	0.9889	0.9561	0.9103	0.9756	0.8043	0.9900	1015	64	36	100
	2	0.9842		0.8266	0.9851	0.8266	0.9851	99	734	55	154
	3	0.9926		0.9742	0.9546	0.9900	0.8893	148	90	8977	238

Table 11. The results of recognition of 3 classes using the aggregated ensemble of CNN classifiers.

Class	**AUC**	**ACC**	**TPR**	**TNR**	**PPV**	**NPV**
1	0.9959	0.9815	0.9327	0.9928	0.9344	0.9926
2	0.9943		0.8863	0.9947	0.9347	0.9903
3	0.9978		0.9965	0.9601	0.9914	0.9836

Table 12. The comparison of the confusion matrices and committed errors in three class recognition problems of the best CNN classifier (Inceptionv3) and the aggregated ensemble.

	Inceptionv3				**Aggregated ensemble**			
Class	1	2	3	**Errors**	1	2	3	**Errors**
1	1006	61	48	**109**	1040	47	28	**75**
2	62	772	54	**116**	49	787	52	**101**
3	42	51	9122	**93**	24	8	9183	**32**

5.1. Summary of the class recognition results

The results presented in this paper deal with the most general problems of class recognition in breast cancer (malignant versus benign, cancer cases against normal, and simultaneous recognition of 3 classes). The results are presented for the most often used DDSM dataset. Moreover, they consider all samples in the DDSM database and apply the 5-fold cross-validation approach, which performs the testing phase on the whole database and not on a very narrow (usually around 20%) set of data left for testing, as was shown in many papers mentioned in the second section. Therefore, our results belong to the most objective.

In the case of cancer lesions versus normal, the presented quality measures are as follows: accuracy of 98.73%, sensitivity of 97.86%, specificity of 99.07%, and AUC of 0.9975. Recognition of malignant versus benign cases has delivered the results: accuracy of 93.81%, sensitivity of 94.45%, specificity of 93.13%, and AUC of 0.9843. In the case of recognition of three classes at the same time, the average results are as follows: accuracy of 98.15%, sensitivity of 93.85%, specificity of 98.25%, and AUC of 0.9960. All of them belong to the best for this database.

6. Conclusions and future directions of study

The paper has presented a novel approach to the recognition of breast cancer based on mammogram images. The significant difference between the actual approaches to this problem is proposing the efficient method of creation of an ensemble composed of the set of potential CNN classifiers, working simultaneously on the same database.

Thanks to the transfer learning technique applying the pre-trained architectures, it is possible to adapt different architectures of CNNs forming an ensemble in a reasonable time and aggregating their results into a final verdict. However, still, the total time of fine-tuning this very large set of classifiers in 5-fold cross-validation is rather long (the training time of individual CNN was changing from a few minutes using Alexnet up to 40 minutes for Nasnetlarge. Nonetheless, the testing stage in each run is very short and is counted in seconds. All these results have been obtained using the 64-bit PC operating under Windows 10 Pro, Procesor Intel Core i7-2700K, CPU 3.50 GHz, 16Gb RAM, and graphic card NVIDIA GeForce GTX 1080, VRAM 8 GB.

The best composition of the ensemble applied in the paper is based on the analysis of the quality measures such as accuracy, sensitivity, and AUC shown by the particular classifier selected from the available set. The proposed method was checked on the DDSM database and shows its great potential.

The paper has shown a very good performance of the proposed ensemble considering the recognition of all possible combinations of classes in the DDSM repository of mammograms. The presented results are the best for this DDSM repository of mammograms.

In the future study, we will check and compare its performance using other available breast cancer repositories, like MIAS, BCDR, INbreast, or CMMD.

The presented method of creating an ensemble of classifiers is universal and can find application in the recognition problems of other types of images, not necessarily medical. The interesting direction of the research is to check its efficiency on images created by other methods of image representation, like multispectral, infrared, etc.

References

[1] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA a Cancer Journal for Clinicians*, *71*(3), 209–249. https://doi.org/10.3322/caac.21660

[2] http://www.eng.usf.edu/cvprg/mammography/database.html

[3] Osowski S., Szmurlo R., Matematyczne modele uczenia maszynowego w językach Matlab i Python. 2023, OWPW, Warszawa

[4] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. In *MIT Press eBooks*. https://dl.acm.org/citation.cfm?id=3086952

[5] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

[6] Grochowski, M., Mikołajczyk, A., & Kwasigroch, A. (2019). Diagnosis of malignant melanoma by neural network ensemble-based system utilising hand-crafted skin lesion features. *Metrology and Measurement Systems*, 65–80. https://doi.org/10.24425/mms.2019.126327

[7] Matlab user interface. MathWorks, 2024.

[8] He, K., Zhang, X., Ren, S., & Sun, J. (2015a). Deep residual learning for image recognition. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1512.03385

[9] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-V4, Inception-ResNet and the impact of residual connections on learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1). https://doi.org/10.1609/aaai.v31i1.11231

[10] Thawkar, S. (2022). Feature selection and classification in mammography using hybrid crow search algorithm with Harris hawks optimization. *Journal of Applied Biomedicine*, *42*(4), 1094–1111. https://doi.org/10.1016/j.bbe.2022.09.001

[11] 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. https://doi.org/10.1016/j.eswa.2017.03.031

[12] Swiderski, B., Gielata, L., Olszewski, P., Osowski, S., & Kołodziej, M. (2020). Deep neural system for supporting tumor recognition of mammograms using modified GAN. *Expert Systems With Applications*, *164*, 113968. https://doi.org/10.1016/j.eswa.2020.113968

[13] Logan, J., Kennedy, P. J., & Catchpoole, D. (2023). A review of the machine learning datasets in mammography, their adherence to the FAIR principles and the outlook for the future. *Scientific Data*, *10*(1). https://doi.org/10.1038/s41597-023-02430-6

[14] Ansar, W., Shahid, A. R., Raza, B., & Dar, A. H. (2020). Breast cancer detection and localization using MobileNet based transfer learning for mammograms. In *Communications in computer and information science* (pp. 11–21). https://doi.org/10.1007/978-3-030-43364-2_2

[15] Alhsnony, F. H., & Sellami, L. (2024). Advancing breast cancer detection with convolutional neural networks: A comparative analysis of MIAS and DDSM datasets. In *IEEE 7th International Conference on Advanced Technologies, Signal and Image Processing (ATSIP)* (pp. 194–199). https://doi.org/10.1109/atsip62566.2024.10638886

[16] Murty, P. S. R. C., Anuradha, C., Naidu, P. A., Mandru, D., Ashok, M., Atheeswaran, A., Rajeswaran, N., & Saravanan, V. (2024). Integrative hybrid deep learning for enhanced breast cancer diagnosis: leveraging the Wisconsin Breast Cancer Database and the CBIS-DDSM dataset. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-74305-8

[17] Gil, F., Osowski, S., Świderski, B., & Słowińska, M. (2022). Ensemble of classifiers based on deep learning for medical image recognition. *Metrology and Measurement Systems*, 139–156. https://doi.org/10.24425/mms.2023.144400

[18] Mohammadi, S., & Livani, M. A. (2024). Enhanced breast mass segmentation in mammograms using a hybrid transformer UNet model. *Computers in Biology and Medicine*, *184*, 109432. https://doi.org/10.1016/j.compbiomed.2024.109432

[19] Kuncheva, L. I. (2014). *Combining pattern classifiers: Methods and Algorithms*. John Wiley & Sons. https://doi.org/10.1002/9781118914564

[20] Kingma, D. P., & Ba, J. L. (2014). Adam: A method for stochastic optimization. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1412.6980

[21] Tan P.N., Steinbach M., Kumar V., (2014) *Introduction to data mining*, Pearson Education Inc., Boston.

Metrol. Meas. Syst., Vol. 32 (2025), No. 3
DOI: 10.24425/mms.2025.154682

Fabian Gil received his MSc and PhD in Electronic Engineering from the Military University of Technology in Warsaw in 2019 and 2024, respectively. He is currently an academic teacher as an assistant at the Faculty of Electronics, Military University of Technology in Warsaw. His research interests include artificial intelligence methods, in particular machine learning, deep learning, data mining and their application in a wide range of engineering especially in medical applications.

Stanisław Osowski received the M.Sc., Ph.D., and D.Sc. degrees from the Warsaw University of Technology, Warsaw, Poland, in 1972, 1975, and 1981, respectively, all in electrical engineering. Currently, he is a professor of electrical engineering at the Institute of the Theory of Electrical Engineering and Electrical Measurements of the same university and at the Electronic Faculty of the Military University of Technology, Warsaw, Poland. His research and teaching interests are in the areas of computational intelligence, especially machine learning, neural networks and deep learning, data mining, and their applications in various areas of engineering. He is an author or co-author of more than 200 scientific papers and many books

---

Article history: received April 16, 2025; revised June 8, 2025; accepted July 25, 2025; available online August 13, 2025.