Metrol. Meas. Syst., Vol. 30 (2023) No. 1, pp. 139–156
DOI: 10.24425/mms.2023.144400

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS
Index 330930, ISSN 0860-8229
www.metrology.wat.edu.pl

M&MS

ENSEMBLE OF CLASSIFIERS BASED ON DEEP LEARNING FOR MEDICAL IMAGE RECOGNITION

Fabian Gil[1)], Stanisław Osowski[1,2)], Bartosz Świderski[3)], Monika Słowińska[4)]

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 (*stanislaw.osowski@ee.pw.edu.pl*)
3) University of Life Sciences, ul. Nowoursynowska 166, 02-787 Warsaw (*jbswiderski@wp.pl*)
4) Central Clinical Hospital Ministry of Defense, Military Institute of Medicine – National Research Institute, ul. Szaserów 128, 04-141 Warsaw (*monika.slowinska@yahoo.com*)

Abstract

The paper presents special forms of an ensemble of classifiers for analysis of medical images based on application of deep learning. The study analyzes different structures of convolutional neural networks applied in the recognition of two types of medical images: dermoscopic images for melanoma and mammograms for breast cancer. Two approaches to ensemble creation are proposed. In the first approach, the images are processed by a convolutional neural network and the flattened vector of image descriptors is subjected to feature selection by applying different selection methods. As a result, different sets of a limited number of diagnostic features are generated. In the next stage, these sets of features represent input attributes for the classical classifiers: support vector machine, a random forest of decision trees, and softmax. By combining different selection methods with these classifiers an ensemble classification system is created and integrated by majority voting. In the second approach, different structures of convolutional neural networks are directly applied as the members of the ensemble. The efficiency of the proposed classification systems is investigated and compared to medical data representing dermoscopic images of melanoma and breast cancer mammogram images. Thanks to fusion of the results of many classifiers forming an ensemble, accuracy and all other quality measures have been significantly increased for both types of medical images.

Keywords: breast cancer, CNN, deep learning, ensemble of classifiers, feature selection, melanoma.

1. Introduction

Medical imaging is a widely applied approach in the diagnostic process, especially in cancer and other disorder identifications. Such methods are developing rapidly now due to very fast progress in computer technology, allowing for development of new very efficient image processing techniques. Nowadays, the most exploited technique in image analysis is based on deep learning

Article history: received September 9, 2022; revised November 12, 2022; accepted November 29, 2022; available online January 24, 2023.

using different types of *convolutional neural networks* (CNNs). The application of these networks in medical image analysis can significantly facilitate the analysis and improve the level of accuracy and timeliness of diagnostics.

Especially interesting is the application of an ensemble of classifiers whose votes are integrated into one final decision. Thanks to such organization, it is possible to analyze the images from different points of view and develop a more objective decision [1]. However, the improvement of the accuracy of the ensemble is possible only when its constituent members are independent. Therefore, the diversity among component classifiers is a very important factor in obtaining a well-performing ensemble. Up to now [2] there has been no theory explaining how to satisfy the independence conditions.

Instead, different methods and techniques are applied, based on heuristic experience of researchers. Such methods apply different types of classification units, for example, neural networks, decision trees, or support vector machines [3]. Very important is the application of different learning datasets, usually drawn randomly from the available database. This is typical, for example, in a random forest of decision trees [5]. In some methods, for instance, gradient boosting, the learning samples for the units are selected according to their performance on the learning data [4]. Another method is generating different input attributes to the classification units which form an ensemble [3, 6–9]. This paper will combine numerous different approaches to provide independent operation of constituent members of the ensemble in the task of the recognition of medical images related to melanoma and breast cancer.

Two different problems in the machine approach to medical image analysis can be pointed out. One are very high differences between images belonging to the same class and, at the same time, high similarity of representatives of different classes. The other problem is the limitation of data samples, which results in deteriorating the generalization ability of the developed systems. The remedy for this is application of the ensemble of classifiers.

The paper proposes two different methods to create an ensemble system which can deal effectively with the problems mentioned above. The solutions presented in the paper will apply the following approaches":

– The first investigated system will apply the ensemble based on the application of many feature selection methods combined with the classical (shallow) classifiers: *support vector machine* (SVM), *Random Forest* (RF), and softmax. In the first stage, the images are analyzed by the convolutional neural network to generate numerical descriptors. These descriptors are formed by the elements of the flattened vector of the last convolution layer of the CNN. In the next stage, the class discrimination ability of descriptors is assessed using different criteria. A limited number of descriptors of highest discrimination ability is selected and used as the input attributes to these three types of classifiers, forming an ensemble. Independence of ensemble members has been achieved thanks to the application of different structures and mechanisms of decision making by the classifiers as well as different sets of input attributes. Among three groups of classifiers, the SVM and RF are supplied by different sets of specially selected features. On the other hand, softmax classifiers are supplied by features randomly selected from the elements of the flattened vector. Thanks to the varied principles of the applied feature selection methods, the probability of independence of the classifiers is significantly increased.
– The second arrangement of the ensemble is formed by direct using different structures of deep *convolutional neural networks* (CNNs). These networks are responsible for simultaneous generation of features and final classification. Thanks to the diversity of CNN network architectures, the independence between the ensemble members is achieved.

Irrespective of the approach to ensemble creation, the final verdict of the ensemble is elaborated on the principle of majority voting.

The validation of the proposed classification systems will be performed using databases corresponding to dermoscopic images of melanoma and mammographic images of breast cancer. Different quality measures of the proposed system, like accuracy, sensitivity, precision, F1, and area under the ROC curve (AUC) are presented. They show a significant advantage of the ensemble over individual performance of classifiers.

The rest of the paper is organized as follows. Section 2 is devoted to the description of databases of melanoma and breast cancer images used in numerical experiments. Section 3 presents the feature selection methods used in the solution. Section 4 presents details of the ensemble of classifiers. Section 5 presents and discusses the results of numerical experiments concerning the recognition of dermoscopic images of melanoma and mammograms of breast cancer. The concluding Section 6 summarizes and discusses the obtained results and indicates the future directions of research.

2. Database description

The numerical experiments have been conducted using medical images representing two different classification problems: melanoma and breast cancer.

2.1. Melanoma database

The database of melanoma images was created at the Warsaw Maria Skłodowska-Curiebreak National Research Institute of Oncology, Department of Soft Tissue/Bone Sarcoma and Melanoma [3].

The base contains 112 RGB images of verruca seborrhoica representing non-melanoma and 134 images of basal cell carcinoma (melanoma). The images were acquired using dermoscopy of the magnification of 20×. They have been registered from different parts of the body. The database was annotated by expert dermatologists applying the ABCDE criteria and confirmed by exact pathomorphological inspection. All images are stored in the JPEG format. The images were created at different times and recorded with different resolutions. Therefore, the size of the images varies from 767×576 to 4273×2848 pixels. Before supplying them to CNN networks, they have been rescaled to the common size needed by the applied architecture. Fig. 1 shows some exemplary images representing melanoma and non-melanoma cases.

The images on the left (a, b, c, d) of Fig. 1 represent melanoma and on the right (e, f, g, h) – are the non-melanoma samples. There are significant differences in the shape, size, and distribution of colors in the lesion regions within the same class. Some statistical parameters characterizing melanoma and non-melanoma samples are presented in Table 1. It can be seen that the ratio of the standard deviation of mean values of images to the mean value of means of all images is very large for both classes. It varied from 0.26 to 0.72 for melanoma and from 0.26 to 0.62 for non-melanoma. Another problem is similarity among the samples belonging to different classes. The mean values of statistical parameters characterizing images representing both classes are very similar, as shown for melanoma and non-melanoma in Table 1. The only significant difference is observed for the mean value of skewness. However, for this parameter the standard deviation is relatively very large. Therefore, the problem of image recognition for these types of medical images is difficult and needs special attention.

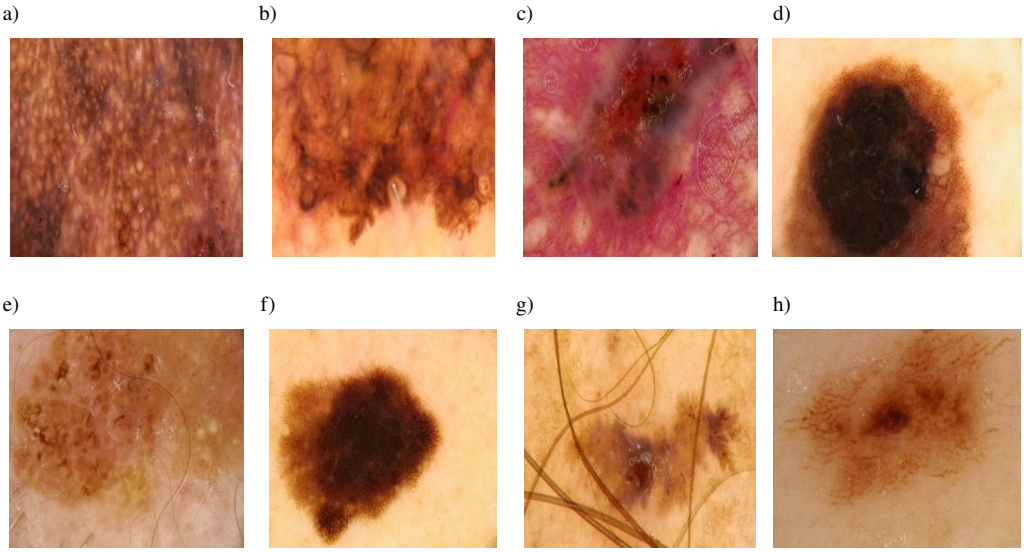

Fig. 1. Exemplary images from the melanoma database. The images on the left represent melanoma and those on the right – non-melanoma cases.

Table 1. The statistical parameters of images presenting melanoma and non-melanoma (mean and standard deviation over all images).

	mean	**std**	**energy**	**skewness**	**kurtosis**
Melanoma	129.42 ± 13.73	61.43 ± 12.11	20961 ± 3859	0.31 ± 0.24	1.95 ± 0.30
Non-melanoma	132.51 ± 13.62	61.09 ± 10.51	21580 ± 3802	−0.17 ± 0.28	2.03 ± 0.33

2.2. Mammogram database

Another problem is recognition of breast cancer images. This time we have used the publicly available database of mammographic images – *Digital Database for Screening Mammography* (DDSM) created by the University of South Florida [10].

The dataset contains 2802 cases, each composed of 4 mammograms (left and right breast from above, representing the cranial-caudal view, and oblique representing the mediolateral-oblique view). The images are prepared in the form of ROI, extracted by medical experts. The size of the ROI images was 128 × 128 pixels. Normal tissues are represented by 9215 images. Among abnormal, there are 888 benign and 1115 malignant cases. The database is seriously unbalanced since among all images 81% represent the same (normal) class.

Figure 2 presents several examples of ROI of mammograms typical for normal (top images) and abnormal cases representing either benign (middle images) or malignant (bottom images) cases. Similarly to melanoma images, we observe significant differences among images belonging to the same class of data and at the same time close similarities among samples of different classes. Therefore, the problem of class recognition is also very difficult to solve, similarly as is the case of melanoma image recognition.

Metrol. Meas. Syst., Vol. 30 (2023), No. 1, pp. 139–156
DOI: 10.24425/mms.2023.144400

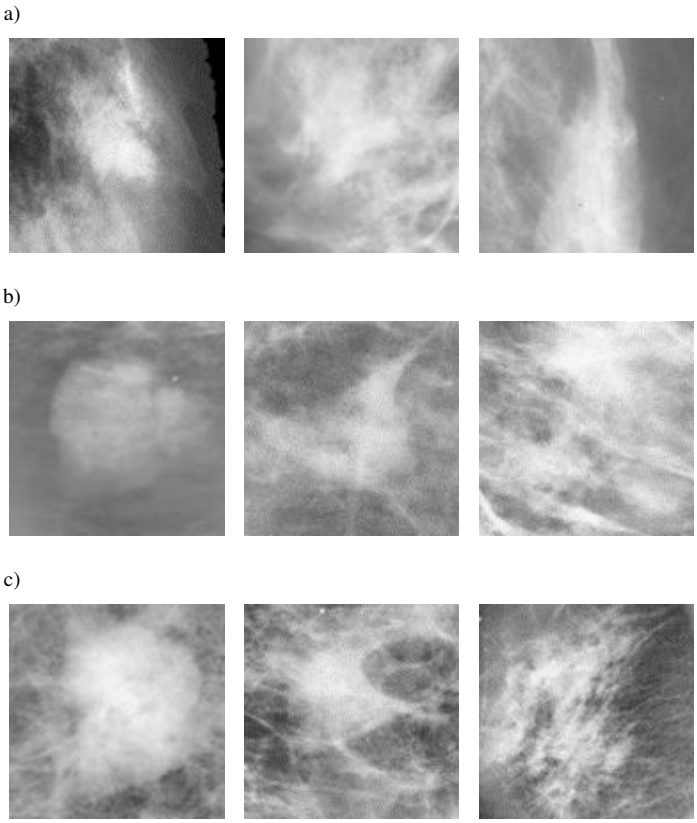

Fig. 2. Examples of ROI of mammograms belonging to normal (a), benign (b), and malignant (c) cases.

3. Feature selection methods

The classifiers applied in the ensemble should be supplied by the diversified sets of features selected from a large set of numerical descriptors of the image. The classical approach to the definition of descriptors uses expert knowledge of the images. Typically, it is based on texture analysis, statistics in the color description, or some geometric parameters [3, 7, 11]. Such an approach in the case of melanoma tries to follow the ABCD descriptive system used by medical experts in melanoma recognition [12].

However, nowadays such a manual approach to the definition of numerical description has been replaced by automatic methods and application of deep learning techniques. Such descriptors are generated by many preprocessing, locally connected layers, as is done in a CNN. They are finally available in the flattened layer of the CNN structure. For example, in Alexnet the flattened layer delivers 4096 descriptors, correlated in some way with the classes subjected to recognition.

Irrespective of the method applied in the definition of numerical descriptors of the image, the next step is to assess their class discrimination ability, which is strictly dependent on the selection procedure. The simplest selection method is to use a limited set of randomly selected descriptors. It has been found that such an approach is quite efficient with a very high population of descriptors. Such a selection method is typically associated with softmax in deep learning.

The choice of randomly selected features is changing at each learning cycle. Statistically, with many cycles, all descriptors are involved in making the classification decisions, albeit in different proportions depending on the drawing process.

The other, more advanced approach, is to select a limited number of descriptors as features based on a deterministic evaluation of their ability to discriminate between classes. There are many well-known selection methods based on different principles, like statistical hypotheses, THE correlation principle, the distance of neighboring samples, genetic algorithms, decision trees, etc. [9, 11, 13]. As a result of their application different sets of chosen descriptors can be selected.

Based on some introductory experiments the following selection methods have been applied in this work: the *Fisher discriminant method* (FD), *two-sample Student t-test* (T-test), *Kolmogorov-Smirnov test* (KS), *Kruskal-Wallis test* (KW), *correlation of data with the class* (CDC), *step-wise fit selection* (SWF), the *nearest neighbor analysis* (NNA), and *relieff* (REL) [9, 13].

They represent different approaches to feature selection, including filter, wrapper, and embedded methods. Each of these methods applies a different principle to evaluate the class discrimination ability of the feature. Some methods evaluate features based on their independent performance (FD, T-test, KS, KW, CDC), while others consider their cooperation in classification tasks (SWF, NNA, REL). The other approach is based on the results of hypothesis tests (T-test, KS, KW) and still another applies special formulas relating the values of the features to class membership (REL, NNA, FD, SWF). Irrespective of the approach, the chosen numerical descriptors are associated with the class membership according to the applied selection mechanism, which varies from method to method.

Figure 3 shows the exemplary results of class discrimination ability of the first 100 numerical descriptors (out of 4096 generated by CNN) according to the Kruskal-Wallis test, Fisher

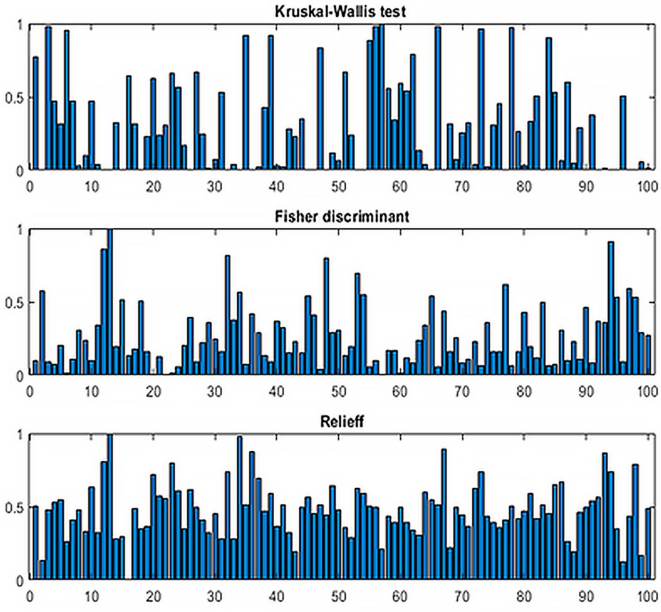

Fig. 3. Examples of estimated values of class discrimination ability of the first 100 numerical descriptors generated by the CNN in recognition of melanoma versus non-melanoma.

discriminant, and relieff. The results correspond to the database of melanoma images. It should be noticed here that the class discrimination ability of the first 100 features has been assessed in various ways with different methods. A similar situation is among the rest of the descriptors. Therefore, in the next step, after arranging them in decreasing order and limiting their number to some chosen value, we obtain 8 different sets of features, corresponding to 8 selection methods, which supply the SVM and RF. Their cooperation with these classifiers will result in 16 units of the ensemble.

To check how different the sets of selected descriptors are, we have estimated the statistical characterization of them. Table 2 shows the statistical values of the most important 300 descriptors selected by the applied selection algorithms. All of them have been normalized to the common range $\{0, 1\}$ by applying min-max normalization.

Table 2. Statistical parameters of the normalized quality measures corresponding to 8 selection methods.

Method	**mean**	**median**	**std**	**iqr**	**skewness**	**kurtosis**
SWF	0.3763	0.3196	0.2042	0.142	1.02066	4.216
T-test	0.2536	0.1879	0.2094	0.3322	0.96572	3.456
KS	0.2067	0.09598	0.25981	0.29824	1.56156	4.785
KW	0.2040	0.0802	0.2627	0.2821	1.368999	3.7128
CDC	0.210	0.0876	0.2674	0.3553	1.48718	4.3775
FD	0.261	0.196	0.2105	0.3429	0.94611	3.3415
NNA	0.260	0.2345	0.2462	0.4368	0.6830	2.703
REL	0.195	0.217	0.208	0.375	0.587	2.881

The estimated values of the mean, median, standard deviation, interquartile range, skewness, and kurtosis estimated values are presented for all 8 selection methods. The presented statistics confirm the existing differences in the selected sets of features. Especially large changes are observed for median and skewness values.

4. Ensemble systems of classifiers

An ensemble of classifiers is a known method for increasing the efficiency of classification systems [1]. It uses multiple learning models to obtain better performance compared to any of the constituent models alone. The research in this area has shown that significant diversity among the classification models is needed to obtain better results for the validation/testing data [1, 14]. One can notice that each unit looks at the classification problem from different points of view, considering various aspects of the mechanisms governing the analyzed process.

The paper proposes two approaches to ensemble creation. The first system uses classical (shallow) classifiers supplied by features chosen from the full set of descriptors by applying special deterministic selection methods as well as a randomly chosen set. The second model applies the features automatically generated by the deep structure of the convolutional neural network. It uses many CNN networks of different structures, providing in his way different sets of features. Such arrangement has shown high independence of the constituent units and, as a result, high efficiency in class recognition.

4.1. Ensemble based on shallow classification units

The first proposed system is based on three types of classification units: SVM, RF, and softmax. It applies the complex hierarchical structure of classification based on deep learning. The pre-trained CNN neural network used in the transfer learning mode delivers a wide set of numerical descriptors of the image These descriptors are subjected to different selection procedures generating a diversified set of features and creating the input attributes to SVM [15], the random forest of the decision tree [5], and softmax classifiers [16].

The SVM is a supervised classification model strictly associated with a very special learning algorithm developed by V. Vapnik. It applies nonlinear mapping of a set of original vectors $\mathbf{x}$ into a hyperplane using a kernel function $K(\mathbf{x}, \mathbf{x}_i)$, which allows much better discrimination between the data of two opposite classes. Thanks to this, the SVM classifiers perform very well (good generalization ability) in difficult high-dimensional classification problems with a relatively small population of learning data. The SVM of the Gaussian kernel was used in our solution. The regularization constant C and Gaussian kernel width have been adjusted in the introductory stage by repeating the learning experiments for the set of their predefined values and choosing the best one based on the validation data set. The SVM network has applied the kernel function $K(\mathbf{x}_i, \mathbf{x}) = \exp\left(-\gamma \|\mathbf{x} - \mathbf{x}_i\|^2\right)$ of the hyperparameter $\gamma = 0.5$ and regularization constant C = 1000. Both values have been adjusted by trials in the introductory experiments using the validation data.

The Breiman random forest [5] represents another high-quality classification system based on the ensemble of decision trees. It constructs many decision trees at training time. Their verdicts are integrated by the majority voting. The good generalization ability of RF is provided by the random selection of the subsets of the learning data, as well as using a limited set of randomly selected features, chosen in each decision node of the trees. The applied random forest was composed of 100 decision trees with the number of variables in the decision nodes equal to the square roots of the dimension of the input vector.

The softmax classifier, extensively applied in deep learning, elaborates its classification decision on the probability theory. Contrary to the previous selection approach, the input attributes to the softmax are usually formed from randomly selected descriptors which change with every learning cycle. The ith output unit calculates first the weighted sum u_i of the input signals (elements of descriptor vector $\mathbf{x}$):

$$u_i(\mathbf{x}) = \sum_j w_{ij} x_j + w_0 \tag{1}$$

and then the probability of the membership of this vector to the ith class ($i = 1, 2, \ldots, M$) based on the softmax operation:

$$\text{softmax}\,(\mathbf{u}_i) = \frac{\exp(u_i)}{\sum_{j=1}^{M} \exp(u_j)}. \tag{2}$$

The label of the output unit of the highest value of softmax represents the recognized class. Applying different dropout ratios in the random selection process allows to create many members of this type.

The diversity of the classification units is additionally increased thanks to a different definition of the objective function and learning algorithms. The SVM cost function is defined as the quadratic optimization problem with constraints. RF is based on building many weak decision trees and applying an ensemble approach to elaborate final classification decisions. Softmax

classifier is based on minimization of cross-entropy function and operates in a probabilistic space.

Based on the considerations presented above, we have developed an ensemble system of the structure presented in Fig. 4 [32]. The signals of the hidden convolution layers of the CNN architecture are created by the pre-trained structure. After some introductory experiments, we applied AlexNet architecture in the role of CNN [17].

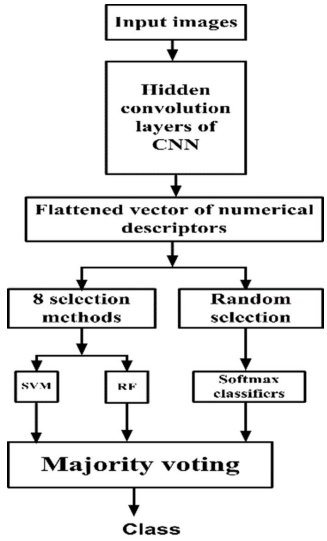

Fig. 4. First CNN-based ensemble structure used in the experiments. The pre-trained CNN delivers the set of 4096 numerical descriptors of the data. They are subjected to 8 selection procedures delivering a much smaller number of descriptors as the diagnostic features. These features represent the input signals to SVM and RF classifiers. The set of softmax classifiers applies different numbers of hidden units and is supplied by features randomly selected from the whole set of descriptors.

The first part of the applied AlexNet is composed of 6 locally connected hidden convolution layers responsible for the generation of the set of 4096 numerical descriptors of the image. The ReLU nonlinear activation function and MAX pooling were used in each layer. The ADAM learning algorithm with an initial learning rate of $1.3e-4$ was applied in learning. The mini-batch size was equal to 10 and the number of epochs limited to 50 was used.

The set of all descriptors represents the input to the ensemble system composed of 26 classifiers. The 16 classification units are based on SVM and RF supplied by the sets of features chosen by 8 deterministic selection methods.

The other 10 softmax classifiers are supplied by a limited set (from 30% to 50% randomly drawn) of the randomly selected features chosen from the set of descriptors with the assumed dropout ratio. They represent the input attributes for the final neural classification subsystem, which is based on the softmax operation. The CNN members of this set apply also different numbers (from 200 to 600) of hidden neurons of Rectified Linear Units (ReLU) activation. The number of output neurons is equal to the number of recognized classes.

4.2. Ensemble based on deep CNN classifiers

The second system developed in the paper is based on direct application of deep CNN classifiers of different structures. The CNN architecture of many locally connected layers is

responsible at the same time for generating the features and undertaking the final classification decision. The hidden convolutional layers apply the shared-weight architecture of the convolution kernels of filters that slide along input features providing the translation-equivariant responses (feature maps). The input image applied to the network is processed by many such layers. In general, the CNN multilayer structure takes advantage of the hierarchical pattern in data and patterns of increasing complexity using smaller and simpler patterns embossed in their filters.

Many different arrangements of CNN architectures can be built. They result in various preprocessing ways of input data. Thanks to this, the CNN members forming the ensemble are significantly independent in the decision regarding the class membership of input data. Nowadays, there are many pre-trained CNN structures that are ready to use in a transfer learning mode. After some introductory experiments the following CNN structures have been selected for this ensemble: GoogleNet, Inceptionv4, DenseNet201, ResNet50, InceptionResNetv2, NasNetlarge, EfficientNetb0, AlexNet and ShuffleNet [17–19]. All members of the ensemble accept the same input images, however, the way of preprocessing applied by them is significantly different. Fig. 5 presents the proposed ensemble of deep CNN units applied in the numerical experiments.

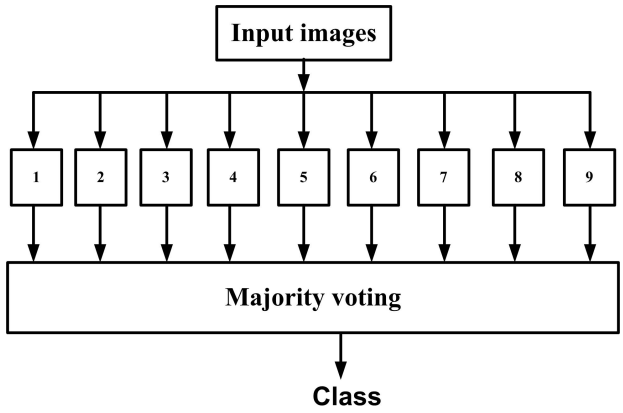

Fig. 5. Ensemble of deep CNN structures used for medical image recognition. The CNN structures are denoted by the numbers: 1 – GoogleNet, 2 – Inceptionv4, 3 – DenseNet201, 4 – ResNet50, 5 – InceptionResNetv2, 6 – NasNetlarge, 7 – EfficientNetb0, 8 – AlexNet and 9 – ShuffleNet.

5. Results of numerical experiments

Numerical experiments using the developed systems have been conducted for the recognition of the medical images corresponding to melanoma and breast cancer. All experiments have been performed in MATLAB [20]. Due to small populations of samples in databases, we resigned from scratch learning and used the pre-trained CNN architectures.

In the first approach to ensemble creation, only one CNN structure used only for the generation of image descriptors is chosen. The choice of it was based on the accuracy of the class recognition by using only a typical softmax learning algorithm, common to all architectures. The highest accuracy in both, melanoma and breast cancer problems on the randomly selected validation set was obtained with the application of AlexNet. The results related to other net structures were slightly worse. Fig. 6 presents a comparison of mean accuracy in melanoma recognition obtained individually by applying different CNN architectures. The names of CNN structures were coded numerically in the same way as in Fig. 5.

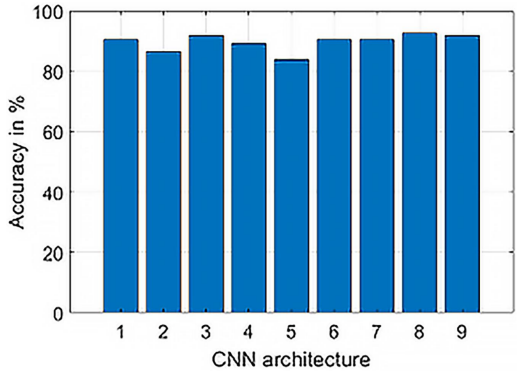

Fig. 6. Comparison of accuracy of class recognition in the melanoma problem obtained in application of single CNN architectures. The pre-trained AlexNet (8) was found the best.

In the database of mammogram images representing breast cancer, the advantage of AlexNet was also confirmed. Therefore, in further experiments, the pre-trained AlexNet was used as the basic structure in the generation of the numerical descriptors of the images in both tasks.

5.1. Results of melanoma recognition

Two different arrangements of the ensemble were used in the experiments. One, based on the application of shallow classifiers was presented in Fig. 4 and the other formed from only deep CNN networks was given in Fig. 5. The first one (called a shallow ensemble) contained 26 classifiers following from a combination of results of feature selection methods and three classifiers: a set of SVM, RF, and softmax classifiers (the later cooperating directly with randomly selected numerical descriptors). Note that the softmax classifier accepting the random choice of descriptors introduces the element of randomness in the classification process (some sort of implicit regularization of the classification task). The second solution (deep ensemble) was formed from only 9 selected CNN classification units as shown in Fig. 5.

The numerical results will be presented using the following quality measures: *accuracy of the system* (ACC), sensitivity (*true positive rate* – TPR), specificity (*true negative rate* – TNR), *true positive precision* (TPP), *true negative precision* (TNP), *F1 measure corresponding to true positive* (F1_TP) and *true negative* (F1_TN) as well as the *area under the ROC curve* (AUC) [21]. These results will correspond to 10 repetitions of experiments with randomly changing contents of the learning subset (70% of database samples) and testing samples (30% of the database). Only the results of testing data will be discussed here. Table 3 represents the mean values of the quality measures corresponding to the shallow and deep ensembles in the recognition of melanoma versus non-melanoma cases. The middle column of the table, related to application of shallow ensemble depicts the results already presented by the authors at the CPEE 2022 conference [32].

All results of the integrated ensembles are much better than the average results of individual members of the ensemble. The average accuracy of non-integrated classifiers of the shallow ensemble was equal to 92.06% at a 2.57% of standard deviation, compared to 94.59% and the standard deviation of 1.08% of the integrated ensemble. In the case of the deep ensemble, the mean accuracy obtained in 10 repetitions of experiments was 98.6% with a small standard deviation equal to 1.05%. The advantage of fusing the results of members of the ensemble into the final verdict of class recognition is evident.

Table 3. The results of recognition of melanoma versus non-melanoma cases.

Quality measure	Shallow ensemble	Deep ensemble
ACC	94.59% ± 1.08%	98.6% ± 1.05%
TPR	0.982	1.00
TNR	0.911	0.971
TPP	0.930	0.975
TNP	0.973	1.000
F1_TP	0.954	0.985
F1_TN	0.942	0.986
AUC	0.9794	0.9996
Mean ACC of individual classifiers	92.06% ± 2.57%	90.95 ± 4.75%

It is generally not easy to objectively compare our results with those presented in different papers, because the databases used in the experiments are usually different. For example, [22] has shown a sensitivity TPR = 0.83 and a specificity TNR = 0.90% of the cascade classifiers in detecting melanoma in clinical images. In [8], a sensitivity TPR = 0.96 and a specificity TNR = 0.80 are reported for the proposed system. A recent paper [14] has presented a deep classification system in melanoma recognition by application of 9 CNN networks and declared an accuracy of 86.71%, obtained on the HAM10000 dataset. [23] has shown the classification system based on several CNN architectures. The experimental results based on 7146 images have shown an accuracy of 76.08% in melanoma recognition.

[24] has shown the results of the application of the current market version of a CNN (Moleanalyzer-Pro®, FotoFinder Systems GmbH) used for classifications (malignant/benign) in six dermoscopic image sets. Each set included 30 melanomas and 100 benign lesions. The best results obtained in these sets were as follows: TPR = 0.93%, TNR = 0.65, AUC = 0.926.

However, all the results presented above refer to different datasets. Therefore, it is difficult to draw objective conclusions. The only fair comparison can be made in the case of the application of the same database. Such a case was presented in [3]. The solution shown there was based on the ensemble but created differently. It combined an extended set of descriptors based also on texture (Haralick GLCM descriptors), chaotic theory (fractal texture analysis), analysis of pixel intensity (maximum subregion statistics, percolation theory, the Kolmogorov-Smirnov statistics). Two types of classical classifiers (SVM and RF) have been used in ensemble creation. The best accuracy reported in [3] for the same database was ACC = 93.76% and AUC = 0.923, while our best results presented in this work are ACC = 98.6% and AUC = 0.9996.

5.2. Results in breast cancer recognition

The numerical experiments for breast cancer have been directed at solving two tasks. The first one is recognizing the abnormal cases (the benign and malignant instances forming one class) from the normal group. The second task represents the recognition of malignant cases from the rest (benign and normal cases treated as a common class). Once again, the experiments were repeated 10 times with randomly selected learning and testing data (70% – learning and 30% – testing).

The additional problem with this database is the very large unbalance of the classes (9215 images representing normal cases, and 888 benign and 1115 malignant cases). To counteract such

a situation the normal data set used in learning was split into 4 equal parts and four classification systems were trained. Each system was supplied with one part of normal cases against the same (whole) set of malignant ones. In the testing mode, the results of all these 4 systems were fused into a final decision using majority voting.

Table 4 presents the numerical results of the ensemble concerning the recognition of normal mammograms against malignant and benign cases treated together. The results obtained by the ensemble are better than the average of individual members not integrated into a common decision (accuracy of 88.93% of deep ensemble compared to 83.72% as the mean of individual classifiers). A significant improvement is observed also for the stability of results in the runs. In the case of a deep ensemble, the standard deviation of results in 10 runs was only 0.55%, compared to 2.90% in the case of non-integrated results.

Table 4. Results of recognition of normal versus abnormal (malignant and benign) cases of breast cancer.

Quality measure	**Shallow ensemble**	**Deep ensemble**
ACC	**84**, **94**% ± 0.94%	**88.93**% ± 0.55%
TPR	0.836	0.863
TNR	0.850	0.910
TPP	0.812	0.906
TNP	0.881	0.869
F1_TP	0.823	0.884
F1_TN	0.865	0.889
AUC	0.9201	0.9644
Mean ACC of individual classifiers	79.12% ± 2.11%	83.72% ± 2.90%

Figure 7 shows the ROC curve obtained in a single run of the experiments. After integration of the ensemble, the average area under this curve obtained in 10 repetitions of the experiment reached the value AUC = 0.9644. This value is significantly larger in comparison with AUC = 0.9201,

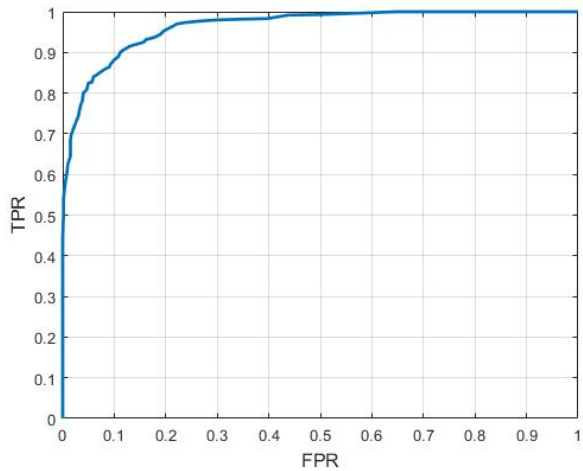

Fig. 7. ROC curve obtained in one run of the experiments in mammogram recognition.

which was calculated as the average of the individual values of the non-integrated members of the ensemble.

The second experiments for mammograms were conducted in recognition of malignant cases from the normal and benign together. This task is slightly more difficult, because of some similarities between malignant (one class) and benign (opposite class) cases.

Table 5 presents the detailed results corresponding to the considered quality measures in the second task. Once again, both ensembles have provided improved results compared to the average value of all individual members taking part in the classification task. For example, the accuracy of the deep ensemble was 86.54% compared to an average of the single classification deep units equal to 79.63%.

Table 5. The results of recognition of malignant versus normal joined with benign cases of breast cancer.

Quality measure	Shallow ensemble	Deep ensemble
ACC	80.15% ± 1.03%	86.54% ± 0.61%
TPR	0.83	0.867
TNR	0.78	0.863
TPP	0.73	0.863
TNP	0.86	0.866
F1_TP	0.78	0.865
F1_TN	0.82	0.865
AUC	0.8663	0.9319
Mean ACC of individual classifiers	73.22% ± 3.95%	79.63% ± 3.71%

The comparison of our results with other works will be limited to the same DDSM database applied in several recent works. All these papers have proposed different CNN structures as a solution to the problem. However, the presented results strongly depend on the number of samples taken from DDSM database since this base is very strongly unbalanced (9215 of normal class out of all 11218 images existing in the database) A problem with the comparison is the presented set of quality factors, usually different in the papers. For example, the lack of sensitivity results does not allow assessing the real quality of the methods since for the unbalanced data set it is very easy to achieve good accuracy at the expense of sensitivity. In general, the only real comparison can be done based on AUC (often omitted by authors).

For example, [25] declared an accuracy of 92.5% for a set of only 600 images selected from DDSM and representing benign and malignant cases. [26] presented results for 2085 images. They include an accuracy of 85% and AUC = 0.91 in recognition of normal, benign, and malignant classes. [27] applied 1318 images and declared an accuracy of 66% in recognition of normal, benign, and malignant cases. [28] used 10480 images and declared an accuracy of 92% in the recognition of normal, benign, and malignant cases. [29] declared the sensitivity of the value TPR = 0.95 in recognition of 3568 images representing benign and malignant cases of breast cancer.

In [30], the numerical results of the system based on the curvelet application were presented for the whole database containing 11218 images. The authors stated that the accuracy of detection of malignant and benign cases (treated together) versus normal cases ranges from 81.3% to 86.4%, depending on the feature set used. The accuracy for detecting malignancy compared with normal and benign (treated together) ranged from 50.7% to 60.4% for the best solution.

Another paper, [31], presented an ensemble of 10 classification units based on SVM, random forest, k-nearest neighbors, nonlinear autoencoder, and logistic classifiers, all powered by a set of diagnostic features organized in a different order. The experiments also used the whole database containing 11218 images. For the detection of normal versus abnormal (malignant and benign) cases, the reported accuracy was 84.5% with a sensitivity of 0.829 and an AUC of 0.920. For the detection of malignant cases from the rest (benign and healthy cases combined), the reported accuracy was 80.2%, a sensitivity of 0.833, and an AUC of 0.890.

Our best results for the recognition of normal from abnormal (malignant and benign) cases by the deep ensemble are as follows: accuracy of 88.93%, the sensitivity of 0.863, and AUC = 0.9644. For malignant case detection, our best results are an accuracy of 86.54%, a sensitivity of 0.867, and AUC = 0.9319.

6. Conclusions

The paper has proposed two multi-stage ensemble systems to solve the classification problem in medical image recognition representing melanoma and breast cancer. The first system is organized into a few stages of data processing. The first stage applies a preprocessing CNN for the generation of a large set of numerical descriptors of the images. Based on the introductory experiments, AlexNet was chosen as the source of 4096 descriptors. In the second phase, different selection methods were applied to generate a limited set of the best features (input attributes to the classifiers) from the full set of image descriptors. The applied methods represent different approaches to the problem of selection, including filters, wrappers, embedded methods, and even a special form of random selection. Thanks to their independent operation, they analyze the data from different points of view and in this way enrich our information about the problem.

In the last step, different sets of features, which are generated by individual selection methods, are used as input attributes to the few classifiers forming the ensemble. Three types of classification systems have been used: a support vector machine, a random forest of decision trees, and softmax. Combining them with the varying sets of features has resulted in 26 members of the ensemble. Their results are fused into a final classification decision using majority voting.

The second ensemble system was formed from many different deep CNN structures whose results are fused into a common decision by majority voting. Due to significant differences in signal processing in the hidden layers, the classifiers forming the ensemble achieved very high independence in their operation. Thanks to this, the ensemble created in this way is very efficient and allows for improving the quality of its operation.

The results of numerical experiments have shown the high efficiency of the proposed approach to building the ensemble system in application to medical image recognition. Two types of image recognition tasks have been investigated. One is related to the dermoscopic melanoma images and the other to the images of screening mammography in breast cancer. In both cases, the ensemble system based on deep neural networks has shown a high advantage over the individual member results, significantly increasing the accuracy, sensitivity, and precision of image recognition.

The main contribution of the paper is proposing new ways of creating deep learning ensemble systems which are relatively insensitive to the number of data samples used in learning, and at the same time can cope effectively with the problem of large similarity between representatives of opposite classes. The proposed approaches are based on image descriptors automatically generated by the CNN architecture. The numerical experiments have shown their advantage over the existing solutions.

Future research will be directed at the following tasks:

– One is to find the best combination of different CNN structures associated with the feature selection methods. In this way, we expect to increase further the independence of operation of individual units and, as a result, improve the performance of the ensemble.
– More study is needed to determine the optimal number of component classifiers of an ensemble. Prior determining the ensemble size and volume of the needed learning data is crucial in creating an optimal ensemble system. Some theoretical papers like [2] have suggested that there should be an ideal number of component classifiers for an ensemble, however, the surplus number of classifiers might also lead to the deterioration of the accuracy.

References

[1] Kuncheva, L. (2014). *Combining Pattern Classifiers*. Wiley.

[2] Bonab, H., & Can, F. (2019). Less is more: a comprehensive framework for the number of components of ensemble classifiers. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(9), 2735–2745. https://doi.org/10.48550/arXiv.1709.02925

[3] Kruk, M., Świderski, B., Osowski, S., Kurek, J., Słowińska, M., & Walecka, I. (2015). Melanoma recognition using extended set of descriptors and classifiers. *EURASIP journal on Image and Video Processing*, *2015*(1), 1–10. https://doi.org/10.1186/s13640-015-0099-9

[4] Brownlee J. (2020). *Master Machine Learning Algorithms*. Machine Learning Mastery.

[5] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[6] Grochowski, M., Wąsowicz, M., Mikołajczyk, A., Ficek, M., Kulka, M., Wróbel, M. S., & Jędrzejewska-Szczerska, M. (2019). Machine learning system for automated blood smear analysis. *Metrology and Measurement Systems, 26*(1)., 81–93. https://doi.org/10.24425/mms.2019.126323

[7] 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*, 26(1), 65–80. https://doi.org/10.24425/mms.2019.126327

[8] Barata, C., Ruela, M., Francisco, M., Mendonça, T., & Marques, J. S. (2013). Two systems for the detection of melanomas in dermoscopy images using texture and color features. *IEEE Systems Journal*, *8*(3), 965–979. https://doi.org/10.1109/JSYST.2013.2271540

[9] Stańczyk, U., Zielosko, B., & Jain, L. C. (Eds.). (2018). *Advances in feature selection for data and pattern recognition*. Springer International Publishing. https://doi.org/10.1007/978-3-319-67588-6

[10] Heath, M., Bowyer, K., Kopans, D., Kegelmeyer, P., Moore, R., Chang, K., & Munishkumaran, S. (1998). Current status of the digital database for screening mammography. In *Digital mammography* (pp. 457–460). Springer, Dordrecht. https://doi.org/10.1007/978-94-011-5318-8_75

[11] Li, Y., Li, T., & Liu, H. (2017). Recent advances in feature selection and its applications. *Knowledge and Information Systems*, *53*(3), 551–577. https://doi.org/10.1007/s10115-017-1059-8

[12] Abbas, Q., Emre Celebi, M., Garcia, I. F., & Ahmad, W. (2013). Melanoma recognition framework based on expert definition of ABCD for dermoscopic images. *Skin Research and Technology*, *19*(1), e93–e102. https://doi.org/10.1111/j.1600-0846.2012.00614.x

[13] Aziz, R., Verma, C. K., & Srivastava, N. (2017). Dimension reduction methods for microarray data: a review. *AIMS Bioengineering*, *4*(2), 179–197. https://doi.org/10.3934/bioeng.2017.1.179

[14] Popescu, D., El-Khatib, M., & Ichim, L. (2022). Skin Lesion Classification Using Collective Intelligence of Multiple Neural Networks. *Sensors*, *22*(12), 4399. https://doi.org/10.3390/s22124399

[15] Scholkopf, B., & Smola, A. J. (2018). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

[16] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

[17] 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

[18] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*. https://doi.org/10.48550/arXiv.1602.07261

[19] Tan, M., & Le, Q. (2019, May). EfficientNet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114). PMLR. https://doi.org/10.48550/arXiv.1905.11946

[20] MathWorks. (2021). *Matlab user manual*.

[21] Tan, P. N., Steinbach, M. & Kumar, V. (2013). *Introduction to Data Mining*, Pearson Education Inc.

[22] Sabouri, P., GholamHosseini, H., Larsson, T., & Collins, J. (2014, August). A cascade classifier for diagnosis of melanoma in clinical images. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 6748–6751). IEEE. https://doi.org/10.1109/embc.2014.6945177

[23] Aljohani, K., & Turki, T. (2022). Automatic Classification of Melanoma Skin Cancer with Deep Convolutional Neural Networks. *AI*, *3*(2), 512–525. https://doi.org/10.3390/ai3020029

[24] Winkler, J. K., Sies, K., Fink, C., Toberer, F., Enk, A., Deinlein, T., ... & Haenssle, H. A. (2020). Melanoma recognition by a deep learning convolutional neural network—performance in different melanoma subtypes and localisations. *European Journal of Cancer*, *127*, 21–29. https://doi.org/10.1016/j.ejca.2019.11.020

[25] Jiao, Z., Gao, X., Wang, Y., & Li, J. (2018). A parasitic metric learning net for breast mass classification based on mammography. *Pattern Recognition*, *75*, 292–301. https://doi.org/10.1016/j.patcog.2017.07.008

[26] Yi, D., Sawyer, R. L., Cohn III, D., Dunnmon, J., Lam, C., Xiao, X., & Rubin, D. (2017). Optimizing and visualizing deep learning for benign/malignant classification in breast tumors. In *29th Conference on Neural Information Processing Systems (NIPS 2016)*. https://doi.org/10.48550/arXiv.1705.06362

[27] Hang, W., Liu Z. & Hannun, A. (2017). GlimpseNet: attentional methods for full-image mammogram diagnosis. *Proceedings Hang 2017 Glimpse Net A*.

[28] Lotter, W., Sorensen, G., & Cox, D. (2017). A multi-scale CNN and curriculum learning strategy for mammogram classification. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 169–177). Springer, Cham. https://doi.org/10.48550/arXiv.1707.06978

[29] ur Rehman, K., Li, J., Pei, Y., Yasin, A., & Ali, S. (2022, October). A Deep Learning-Based Approach for Mammographic Architectural Distortion Classification. In *Innovative Computing: Proceedings of the 5th International Conference on Innovative Computing (IC 2022)* (pp. 3–14). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-4132-0_1

[30] 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. https://doi.org/10.1016/j.compbiomed.2015.06.012

[31] 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

[32] Gil, F., Osowski, S., & Slowinska, M. (2022, September). Melanoma recognition using deep learning and ensemble of classifiers. In *2022 23rd International Conference on Computational Problems of Electrical Engineering (CPEE)* (pp. 1–4). IEEE. https://doi.org/10.1109/CPEE56060.2022.9919681

Fabian Gil received his M.Sc. degree in electronic engineering from the Military University of Technology, Warsaw, Poland in 2019. Currently, he is a Ph.D. student at the Faculty of Electronics of the Military University of Technology, Warsaw. His research interest is computational intelligence, especially machine learning methods using the deep approach.

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 Theory of Electrical Engineering and Electrical Measurements of the same university and at the Faculty of Electronics of the Military University of Technology, Warsaw, Poland. His research and teaching interests are 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.

Bartosz Świderski received his M.Sc. degree from Lodz University, Poland in 2002. He also received his Ph.D., and D.Sc., both in electrical engineering, the Warsaw University of Technology in 2007 and 2018, respectively. He is a professor at the Warsaw University of Life Sciences, Faculty of Applied Informatics and Mathematics. His research and teaching interests are in the areas of computational intelligence, machine learning, neural networks, and deep learning, including applications in various areas of medical engineering. He is an author or co-author of numerous scientific papers published in international journals and conferences.

Monika Słowińska graduated in medicine from Collegium Medicum of the Jagiellonian University in 1998. She received her Doctor of Medicine (MD) degree from the Warsaw Medical University in 2010. Her specialization is dermatology and venerology. She is with the Central Clinical Hospital of the Ministry of Defence, Military Institute of Medicine – National Research Institute, Department of Dermatology, Warsaw, Poland.