Journal of Biomedical Informatics 86 (2018) 25–32

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Skin lesion classification with ensembles of deep convolutional neural networks

Balazs Harangi

Faculty of Informatics, University of Debrecen, POB 400, 4002 Debrecen, Hungary

ARTICLE INFO

Keywords:
Deep convolutional neural network
Ensemble-based system
Melanoma detection
Information fusion

ABSTRACT

Skin cancer is a major public health problem with over 123,000 newly diagnosed cases worldwide in every year. Melanoma is the deadliest form of skin cancer, responsible for over 9000 deaths in the United States each year. Thus, reliable automatic melanoma screening systems would provide a great help for clinicians to detect the malignant skin lesions as early as possible. In the last five years, the efficiency of deep learning-based methods increased dramatically and their performances seem to outperform conventional image processing methods in classification tasks. However, this type of machine learning-based approaches have a main drawback, namely they require thousands of labeled images per classes for their training. In this paper, we investigate how we can create an ensemble of deep convolutional neural networks to improve further their individual accuracies in the task of classifying dermoscopy images into the three classes melanoma, nevus, and seborrheic keratosis when we have no opportunity to train them on adequate number of annotated images. To achieve high classification accuracy, we fuse the outputs of the classification layers of four different deep neural network architectures. More specifically, we propose the aggregation of robust convolutional neural networks (CNNs) into one framework, where the final classification is achieved based on the weighted output of the member CNNs. For aggregation, we consider different fusion-based methods and select the best performing one for this problem. Our experimental results also prove that the creation of an ensemble of different neural networks is a meaningful approach, since each of the applied fusion strategies outperforms the individual networks regarding classification accuracy. The average area under the receiver operating characteristic curve has been found to be 0.891 for the 3-class classification task. For an objective evaluation of our approach, we have tested its performance on the official test database of the IEEE International Symposium on Biomedical Imaging (ISBI) 2017 challenge on Skin Lesion Analysis Towards Melanoma Detection dedicated to skin cancer recognition.

1. Introduction

Skin cancer is a common and locally destructive cancerous growth of the skin. Every fifth person in the USA is in a risk of skin cancer, mainly with pale skin and living in a region under strong sunshine [1]. For 2017, an estimated 87,110 incipient cases of melanoma are predicted to be diagnosed in the United States [2].

The human skin is a structured tissue consisting of epidermis, dermis and hypodermis. The epidermis contains melanocytes which can produce melanin at a highly abnormal rate in case of e.g. strong ultraviolet radiation. The malignant tumor due to abnormal growth of melanocytes is known as melanoma [3]. As pigmented lesions appear on the surface of the skin, melanoma can be recognized early via appropriate visual inspection of a clinical expert. Dermoscopy is an imaging technique that eliminates the surface reflection of the skin and makes in vivo evaluation of the structures possible within the epidermis and dermis. With a larger magnification and by removing reflection artifacts, more visual information can be collected from the deeper levels of the skin to help the development of more accurate computer-aided diagnostic (CAD) systems.

The detection of melanoma can also be addressed by using efficient automated image processing methods. As affordable mobile dermatoscopes are becoming available to be attached to smart phones, the possibility for the automated assessment is expected to positively influence corresponding patient care for a wide population. Given the widespread availability of high-resolution cameras, automated algorithms assessing suspicious lesions can be of great value.

CAD systems and their components dedicated to skin lesion detection were introduced at the beginning of 1990 [4]. Since then numerous methods have been published to address this challenging task. Several algorithms (e.g. [5,6]) follow the commonly used manual evaluation method based on the ABCD rules proposed by Nachbar et al. in [7]. This

E-mail address: harangi.balazs@inf.unideb.hu.

https://doi.org/10.1016/j.jbi.2018.08.006
Received 16 January 2018; Received in revised form 14 June 2018; Accepted 7 August 2018
Available online 10 August 2018

protocol is based on the criteria asymmetry (A), border (B), color (C), and differential structure (D). For the proper integration of this rule-based method into a CAD system several problems should be handled. The first and perhaps the most important one is the precise segmentation of the skin lesion which step is the basis for the analysis of asymmetry and the border. Considering this task, we can find many recommendations such as for thresholding [8], region [9] and edge-based approaches [10], soft computing techniques [11] and deformable models [12]. For the final classification by the ABCD protocol, we also have to evaluate the color and structure information of the lesion as it is proposed in [13]. For this aim, we can borrow techniques like color histogram analysis [14] or the consideration of texture features to detect blotches, streaks [15] and pigment networks [16] to train a machine learning-based system. After GPU cards with high computational performance became available at a reasonable price some years ago, several methods based on deep convolutional neural networks (CNNs) have been released (e.g. [11,17]).

The methods proposed in [13,14,15,16] to classify skin lesion images use traditional hand-crafted feature sets; however, the dominance of the powerful, self-extracted, deep convolutional neural network-based features is hardly questionable in these days. For example, a strong evidence of the current superiority of CNN-based approaches in the field of dermoscopy image analysis is that 22 from 23 participants of the 2017 ISBI Challenge on Skin Lesion Analysis Towards Melanoma Detection [18] have considered deep neural network-based methods for an image classification problem.

The classification accuracies of the different solutions are almost equal considering some common metrics. The small variance in performance should come from the different sizes of image sets used for training the CNNs. In [11], CNN-based features were also considered; however, a pre-trained neural network model was applied trained with only 900 images which hardly seems sufficiently large to perform efficient training of a deep learning-based method. The main bottleneck of CNN approaches is that they require large training sets. Esteva et al. published a study [17], where a GoogLeNet Inception V3 CNN architecture was trained on a dataset containing 129,450 clinical images labeled by dermatologists. The authors showed that if the training dataset was sufficiently large, a deep neural network-based method was able to outperform the clinical experts regarding the classification accuracy of the dermoscopy images. Unfortunately, in most medical domains such a huge number of manually annotated training images is not yet available to let an individual CNN to extract and learn all the discriminative texture-based descriptors for high classification accuracy.

In this paper, we propose a deep learning-based approach for skin lesion classification via the fusion of different individual CNN architectures that have already proven their efficiencies in pattern recognition scenarios. We elaborate an ensemble-based framework which can be successfully applied to improve the accuracy of individual CNNs, if the size of the training image set cannot be extended. A possible solution to overcome the challenges related to the applicability of a single CNN for the given task is the fusion of CNNs, so we let more classifiers (based on different CNNs) to vote to compensate the weaknesses of each other. Namely, we show how we can create an ensemble of CNNs in order to outperform the accuracies of the individual neural networks which are trained on the given and insufficient number of annotated images. We have trained four different CNN architectures (GoogLeNet [19], AlexNet [20], ResNet [21] and VGGNet [22]) which were used the most often also in the skin lesion challenge and fused their outputs to reach a higher classification accuracy.

The rest of the paper is organized as follows. In Section 2, we give a short description about the dataset and a brief overview on the four deep neural networks. In Section 3, we introduce our methodology for the creation of ensembles of CNNs. Section 4 is dedicated to our experimental results. In Section 5, we describe further technical details according to the training and fine-tuning of CNNs. Finally, some conclusions are drawn in Section 6.

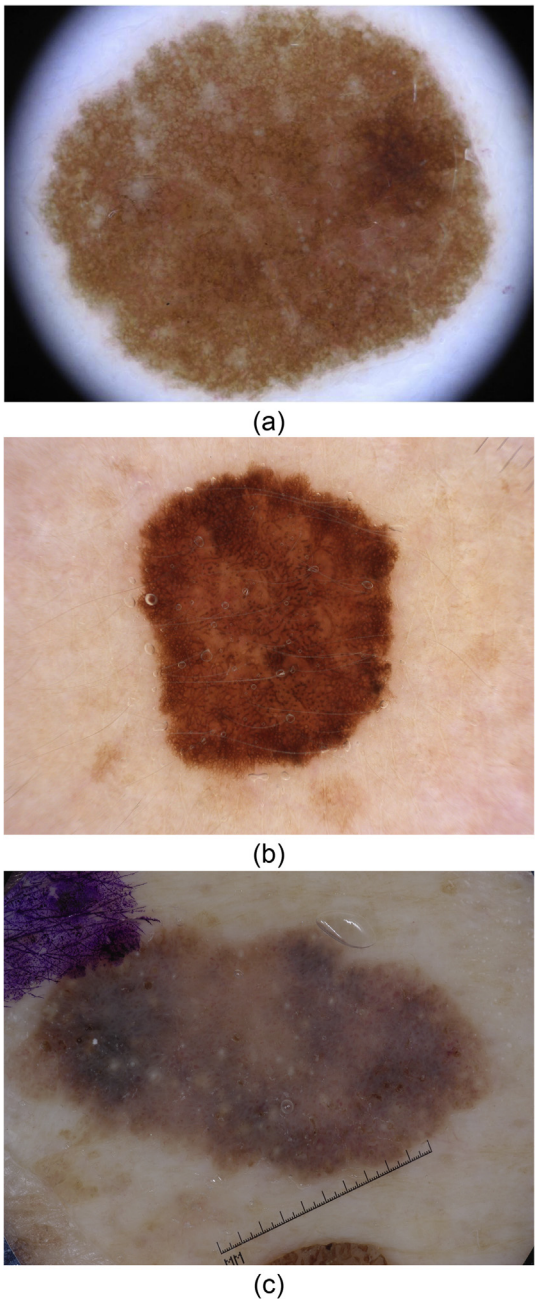

Fig. 1. Special types of skin lesions: (a) nevus; (b) melanoma; (c) seborrheic keratosis.

2. The materials

In this study, our aim is to develop an efficient method which classifies skin lesion images into three classes. Namely, the first class contains the nevus (see Fig. 1a), which is a benign skin tumor. The second one labels cancerous melanoma cases (see Fig. 1b) corresponding to malignant skin tumors derived from melanocytes. The lesions in the third class are labeled as seborrheic keratosis (see Fig. 1c), which is also a benign skin tumor derived from keratinocytes.

2.1. The dataset and augmentation

For the training, validation and experimental evaluation of our approaches, we have used a skin lesion image set consisting of a training, a validation, and a test part. The available training/(test) set

contained 2000/(600) images with manual annotations regarding the three different classes in the following compounds: 1372/(393) images with nevus, 254/(90) images with seborrheic keratosis, and 374/(117) with malignant skin tumors. These volumes of images in the certain classes are not sufficiently large to train CNNs [23]. In one way, the extension of the training set by any publicly available data set [24] it may be a possible solution; however, in this study we do not consider neither public nor private additional image sets to highlight the efficiency of the ensemble-based approach on a smaller dataset to raise classification accuracy.

Naturally, to avoid the overfitting of the CNNs, we have followed the commonly proposed recommendation for the augmentation of the training dataset. There are several possibilities for data augmentation, such as cropping random samples from the images or horizontal flipping or rotating them with different angles (e.g. 90°, 180° and 270°). Using these transformations, we have generated 14,300 images from the originally provided 2000 training ones. That is, we have increased the number of images from 1372 to 8200 for nevus, from 374 to 4600 for melanoma, and from 254 to 1500 for seborrheic keratosis.

2.2. Convolutional neural networks

Recently, in the field of natural image classification, several CNN architectures have been published, like GoogLeNet [19], AlexNet [20], ResNet [21], VGGNet [22] beside others. Some of these architectures (e.g. GoogLeNet, ResNet) are available as pre-trained models also which were initially trained on approximately 1.28 million natural images from the dataset ImageNet [25]. Thus, we can use the weights and biases from these pre-trained models. That is, if we fine-tune all the layers of these models by going on with the backpropagation using our data, they can be applied to our specific classification task, as well. Other architectures as AlexNet and VGGNet are initialized so that their weights and biases are not influenced by visual information which may differ from skin images. Now, we give a brief overview of these commonly used CNNs.

2.2.1. The GoogLeNet architecture

Usually, the CNN architectures consider a sequence of convolutional and sub-sampling layers followed by a fully connected one. In 2015, Szegedi et al. in [19] proposed the GoogLeNet consisting of 22 convolutional layers including 9 Inception modules. An Inception module has 3 different sizes of kernel filters 5×5, 3×3, and 1×1 for convolution and a 3×3 filter for pooling. The size of the receptive field in this network is $224 \times 224 \times 3$ using the RGB color space with the given parameters. Similarly to other CNNs, GoogLeNet learns the convolutional filter entries by stochastic gradient descent (SGD) algorithms during the training stage and extracts higher-level features, when an image goes through the hierarchical structure of the network. In this way, GoogLeNet can automatically extract offline the characteristic features of the different classes during the time-demanding training phase, which also requires a huge number of training images. For simplicity, we will refer to the GoogLeNet as CNN_1 in the later sections.

2.2.2. The AlexNet architecture

In 2012, the AlexNet (CNN_2) architecture proposed by Alex Krizhevsky et al. [20] won the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). AlexNet consists of 5 convolutional layers, some of which are followed by max-pooling layers and 3 fully-connected layers with a final softmax one. The first convolutional layer filters the $224 \times 224 \times 3$ input image with 96 kernels of size $11 \times 11 \times 3$ with a stride of 4 pixels. The second convolutional layer takes the output of the first convolutional layer and filters it with 256 kernels of size $5 \times 5 \times 48$. The third, fourth, and fifth convolutional layers are connected to each other without any intervening pooling or normalization layers. The third convolutional layer has 384 kernels of size $3 \times 3 \times 256$ connected to the normalized/pooled outputs of the second convolutional layer. The fourth convolutional layer has 384 kernels of size $3 \times 3 \times 192$, and the fifth convolutional layer has 256 kernels of size $3 \times 3 \times 192$. The fully-connected layers have 4096 neurons.

2.2.3. The Residual Network (ResNet) architecture

It is well-known that the depth of a CNNs is crucial in visual recognition tasks; however, deeper neural networks are more difficult to train. He et al. reformulated the layers as learning residual functions with reference to the layer inputs instead of learning unreferenced functions and proposed the ResNet (CNN_3) in [21] which has a maximal depth of 152 layers. It means that ResNet is eight times deeper than VGGNet [22], but still has a lower complexity. In 2015, ResNet won the ILSVRC classification challenge.

2.2.4. The VGGNet architecture

In 2015, Simonyan et al. proposed a deep convolutional network in [22], which had depth between 16 and 19 layers and consisted of very small convolution filters. In our setup, we used the configuration consisting of 13 convolutional layers, with filters of size 3×3. Spatial pooling is carried out by five max-pooling layers, followed by some of the convolutional layers and max-pooling is performed on a 2×2 pixels mask with stride 2. A stack of convolutional layers is followed by three fully-connected layers. In the following sections, VGGNet will be denoted by CNN_4.

3. Ensemble of CNNs

Esteva et al. showed in [17] that CNNs can outperform a human expert in a classification task after an exhausted learning phase on a huge annotated training set. However, in many cases, a sufficient amount of annotated images (ground-truth) is not available, so we should improve the accuracy by other approaches.

The fields of decision making and risk analysis, where information derived from several experts and aggregated by a decision maker, have a well-established literature (see e.g. [26,27]). In general, the aggregation of the opinions of the experts increases the precision of the forecast. In order to achieve the highest possible accuracy considering our image classification scenario, we have investigated and elaborated an automated method considering the ensemble of deep convolutional neural networks.

3.1. Theoretical model of the ensemble

For an ensemble of CNNs, we have considered the outputs of their classification layers which take the output of the fully-connected layers and determine confidence values for each class (in our case there are $n = 3$ classes). For the adequate formalization, we consider a CNN to be a function $h: x \to \mathbb{R}^n$ that assigns n confidence values $p_i \in \mathbb{R}$ to a new, formerly unseen image x, where $p_i \in [0, 1]$for $i = 1, \cdots, n$, and $\sum_{i=1}^{n} p_i = 1$. In our concrete task, the values p_1, p_2, p_3 indicate the confidence of the given CNN that x should be classified as nevus (c_1), melanoma (c_2) or seborrheic keratosis (c_3), respectively. As a simple decision, the CNN classifies x to belong to the class having the maximal likelihood:

$$x \to c_i, \quad if \quad p_i = \max(h(x)) \tag{1}$$

In this scenario, the CNNs should assign 3 confidence values to each test image to give the probabilities that the image was labeled by nevus, melanoma or seborrheic keratosis. Thus, we had to derive p'_i probabilities ($p'_i \in [0, 1]$, $i = 1, \cdots, 3$ and $\sum_{i=1}^{3} p'_i = 1$) for each image from the confidence values of the individual CNNs in our ensemble. The possible aggregation approaches are discussed next.

3.1.1. Sum of the probabilities (SP)

According to the literature (see e.g. [26,27]) the most commonly used aggregation model considers the sum of the individual

confidences as:

$$p'_i = \frac{\sum_{j=1}^{m} p_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}}, \quad i = 1, \cdots, n, \tag{2}$$

where p_{ij} stands for the confidence of CNN_j that x belongs to class c_i. Notice that in the present application we consider $j = 1, \cdots, m = 4$ different CNN classifiers. The normalization term $\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}$ is applied to have $p'_i \in [0, 1]$ for $i = 1, \cdots, n$ with $\sum_{i=1}^{n} p'_i = 1$. In this fusion model we calculate the sum of the confidence values of the members CNN_j ($j = 1, \cdots, m$) regarding each class separately and the final label of the image is determined by selecting the maximum normalized sum. Unfortunately, a misclassification can easily occur with using this model, if a weak classifier with low overall accuracy misses an image with assigning a large probability, while the other classifiers also supply low but not zero probabilities to the same wrong class.

3.1.2. Product of the probabilities (PP)

The combined confidence levels p'_i can also be derived as a product of the individual outputs of the CNNs as:

$$p'_i = \frac{\prod_{j=1}^{m} p_{ij}}{\sum_{i=1}^{n}\prod_{j=1}^{m} p_{ij}}, \quad i = 1, \cdots, n. \tag{3}$$

In this case, the p'_i ($i = 1, \cdots, n$) terms are also normalized. In this model an inaccurate, very low or zero probability can strongly influence the final predicted label even though the majority of the classifiers determines the true label of the test image.

3.1.3. Simple majority voting (SMV)

Another basic fusion model of the networks corresponds to classical majority voting, when the final class label of an input image is derived as the majority of the ones provided by the individual CNNs. For example, if the CNN classifiers are not completely correlated, then when (say) $CNN_1(x)$ is wrong, $CNN_2(x)$, $CNN_3(x)$ and $CNN_4(x)$ may be correct, so majority voting will correctly classifies x (see also [28]). Based on this voting rule we can derive the p'_i probabilities as follows:

$$p'_i = \frac{\sum_{j=1}^{m} G(p_{ij})}{m}, \quad i = 1, \cdots, n, \tag{4}$$

$$\text{where} \quad G(p_{ij}) = \begin{cases} 1, & \text{if } p_{ij} = \max(CNN_j(x)), \\ 0, & \text{otherwise}. \end{cases} \tag{5}$$

In this way, each CNN is forced to assign a single class label to x and formula (4) aggregates these votes. A division by the number of the voters is considered for normalization again. In this model the members can assign the input image to a single class and their votes are considered with the same weight regardless their individual accuracies. Moreover, in this fusion method ties are also common issues to be resolved. These are the two main reasons why we have modified the *SMV* model and allowed the members to vote to their strongest candidate classes by their confidence levels.

3.1.4. Sum of the maximal probabilities (SMP)

Similarly to the *SMV* model, the *SMP* allows the j^{th} CNN to vote only for one class. However, in this case we preserve its original confidence value p_{ij} for the derivation of p'_i. Namely, the final probability values are calculated as:

$$p'_i = \frac{\sum_{j=1}^{m} G'(p_{ij})}{\sum_{i=1}^{n}\sum_{j=1}^{m} G'(p_{ij})}, \quad i = 1, \cdots, n, \tag{6}$$

$$\text{where} \quad G'(p_{ij}) = \begin{cases} p_{ij}, & \text{if } p_{ij} = \max(CNN_j(x)), \\ 0, & \text{otherwise}. \end{cases} \tag{7}$$

Notice that normalization has been applied again in (6). In this model we consider only the highest probabilities assigned by the individual classifiers and calculate the normalized average of them for each possible class. The final class label of the input image is derived based on selecting the maximum average probabilities.

As a further refinement described next, we have introduced weight factors in the fusion models to take the individual accuracies of the classifiers into account, as well.

3.2. Weighted ensemble of CNNs

As mentioned in Section 3.1.3, we must handle the problem when two classes receive the same number of votes (i.e. we have a tie). On the other hand our further motivation is that we would like to integrate the individual accuracies of the classifiers in the fusion models to determine the final label. To resolve this issue, we can assign weights to the voters which way we consider models based on weighted majority voting and the sum of the weighted maximal probabilities. To adjust the proper weights, we recommend two options. First, we can rely on the individual accuracies of the CNNs. Second, we can consider the weights as parameters of the ensemble, and find their optimal adjustment via stochastic search methods like a genetic algorithm [29] or simulated annealing [30].

After having found the proper weights ω_j ($j = 1, \cdots, 4$) for all the individual CNNs, we multiply the confidence values p_{ij} of CNN_j ($i = 1, \cdots, 3$) by ω_j ($j = 1, \cdots, 4$) and calculate the class probabilities p'_i according to (6) using the weighted confidence values $\omega_j p_{ij}$ instead of the original p_{ij} ones. In other words, we can interpret these weights as information about the reliability of the corresponding CNNs.

4. Experimental results

After training the individual CNNs on the augmented dataset, we have composed ensembles from them based on all the models described in Section 3.1 and also the weighted ones in Section 3.2 to increase the overall accuracy of classification. We have investigated numerous fusion models in this way and after their comprehensive evaluation we have selected the most accurate one for this task; for the sake of completeness individual CNN performances have also been checked.

Since the involved CNNs are also applied in [11,17,31] as appropriate architectures for skin lesion classification, we have considered them as state-of-the-art solutions and included them in the quantitative comparison. However, the authors of [11,17,31] trained ResNet, GoogLeNet and AlexNet on image sets of sizes 900, 129,450 and 2000, respectively. Thus, for a fair comprehensive evaluation, we have considered their final individual accuracies after training them on the same dataset described in Section 2.1 with the uniform training parameter settings.

Each individual CNN and their ensembles have been evaluated according to the overall score on the test set calculated as the average of the area under the receiver operating characteristic curve (AUC) corresponding to the melanoma and the seborrheic keratosis classification results. As common performance measures, accuracy, sensitivity and specificity also have been calculated at the confidence threshold 0.5. The quantitative results are summarized in Table 1, where M/SK denotes melanoma/seborrheic keratosis, while ACC/SE/SP stand for accuracy/sensitivity/specificity, respectively. For the visual comprehensive evaluation, the receiver operating characteristic (ROC) curves of the four CNNs and their ensembles are plotted in Fig. 2 regarding the melanoma (Fig. 2(a)) and seborrheic keratosis (Fig. 2(b)) classification results.

As a possible further improvement for this classification task, we have also investigated the influence of applying weights to the individual CNNs as it was discussed in Section 3.2. To determine the weights ω_j ($j = 1, \cdots, 4$), on the one hand, we have considered the individual accuracies of GoogLeNet (CNN_1), AlexNet (CNN_2), ResNet

Table 1
Experimental classification results on the test set.

	AVG_ACC	M_ACC	SK_ACC	AVG_AUC	M_AUC	SK_AUC	AVG_SE	M_SE	SK_SE	AVG_SP	M_SP	SK_SP
GoogLeNet	0.842	0.818	0.865	0.848	0.794	0.902	0.592	0.496	0.689	0.722	0.613	0.831
AlexNet	0.848	0.823	0.872	0.830	0.802	0.859	0.518	0.470	0.567	0.645	0.561	0.728
ResNet	0.828	0.822	0.833	0.809	0.757	0.861	0.520	0.385	0.656	0.598	0.437	0.759
VGGNet	0.813	0.802	0.825	0.808	0.766	0.849	0.434	0.256	0.611	0.636	0.585	0.686
SP	0.867	0.845	0.888	0.875	0.832	0.918	0.516	0.376	0.656	0.746	0.654	0.838
PP	0.861	0.845	0.877	0.875	0.836	0.913	0.524	0.393	0.656	0.738	0.650	0.826
SMV	0.903	0.875	0.930	0.839	0.789	0.888	0.632	0.487	0.778	0.822	0.766	0.878
SMP	0.868	0.848	0.887	0.884	0.843	0.925	0.551	0.402	0.700	0.763	0.689	0.836
$\omega^{(1)}$*SMP*	0.869	0.850	0.888	0.886	0.844	0.928	0.580	0.427	0.733	0.772	0.689	0.855
$\omega^{(2)}$SMP	0.866	0.852	0.880	**0.891**	0.851	0.930	0.556	0.402	0.711	0.785	0.719	0.851

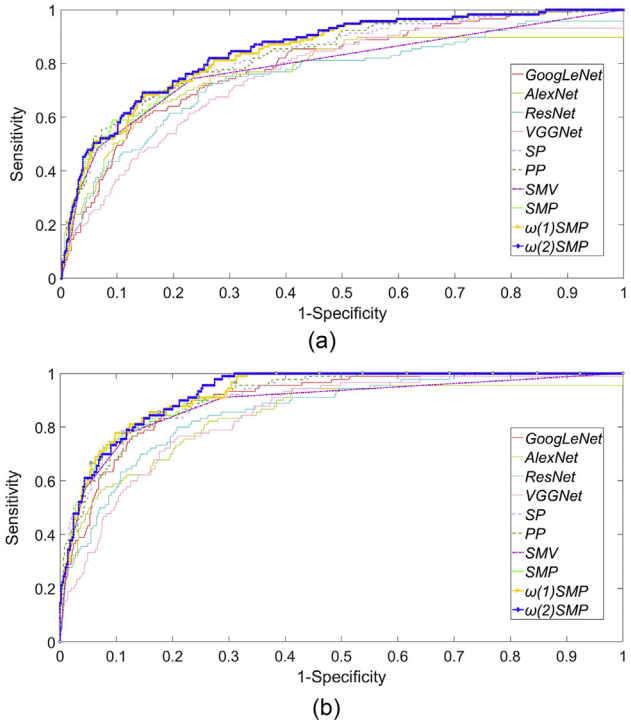

Fig. 2. ROC curves of the four CNNs and their ensembles regarding the (a) melanoma and (b) seborrheic keratosis classification results.

(CNN_3), and VGGNet (CNN_4) based on the *AVG_AUC* values considering the average of the melanoma and the seborrheic keratosis detection results on the validation set. In this way, we have set the weights as follows: $\omega_1^{(1)} = 0.848$, $\omega_2^{(1)} = 0.830$, $\omega_3^{(1)} = 0.809$ and $\omega_4^{(1)} = 0.808$. As we can see in the $\omega^{(1)}$*SMP* row of Table 1, the overall accuracy could be slightly raised using the *SMP* fusion model with the accuracy-based weighting.

On the other hand, we have tried to optimize the ensemble of the CNNs by finding the appropriate weights with simulated annealing (SA) [30]. The SA algorithm considered the accuracy of the ensemble on the validation set and searched for the weights to minimize the cost function −AVG_AUC. Based on the result of this stochastic search, the following weights are obtained: $\omega_1^{(2)} = 0.924$, $\omega_2^{(2)} = 0.659$, $\omega_3^{(2)} = 0.603$ and $\omega_4^{(2)} = 0.594$. The performance of *SMP* with these weights is enclosed in the $\omega^{(2)}$*SMP* row of Table 1. As it can be seen from the results, the most efficient way for fusion has been found to weight the CNNs by the weights determined by SA search and to consider the sum of the maximal confidence levels of their outputs with allowing them to vote to only one class.

For the sake of completeness, Figs. 3 and 4 show sample test images which are misclassified by one of the individual CNNs, but labelled properly by the proposed ensemble-based classification frameworks. Moreover, we provide some statistics about the number of properly classified test images which also justify the fusion of the individual CNNs in a single ensemble framework. The test set used for evaluation contained 600 images with manual annotations as described in Section 2.1 broken down to 393 images with nevus, 90 ones with seborrheic keratosis, and 117 cases with malignant skin tumors. The GoogLeNet, AlexNet, ResNet and VGGNet have found 63/68, 59/52, 54/65 and 36/56 cases from the 117/90 malignant/seborrheic keratosis ones. To determine the label of an image, we select the class with the highest confidence value among the predicted probabilities of an individual CNN. When a test image is classified based on the output of the ensemble of CNNs, 68 and 71 images are labeled properly from the same two classes which indicates that the appropriate combination can outperform the members in terms of accuracy.

As an additional comparative analysis between our method and other approaches in this specific field, we have involved all the state-of-the-art approaches into the quantitative comparison which were developed and presented during the 2017 ISBI Challenge on Skin Lesion Analysis Towards Melanoma Detection [18]. The participants were required to develop automatic methods in this competition to classify skin lesion images as nevus, melanoma or seborrheic keratosis. Notice that, in this challenge the same dataset was considered for the evaluation and the AVG_AUC error measurement was used for the final scoreboard [32]. According to this metric, our approach outperforms 19 proposed solutions from the 23 submitted to the challenge without using additional data for training. We note that, the learning on extended data set was allowed in this challenge and the 1st and 4th participants used external annotated dataset for the training procedure. For the final official result of the 2017 ISBI Challenge on Skin Lesion Analysis Towards Melanoma Detection, see Table 2, which also includes our previous method [33] submitted to the challenge.

5. Discussion

Before creating an ensemble of CNNs, we have to train or fine-tune them. Both are very time consuming procedures, so efficient implementations are also highly recommended because of the limited computational resources and the vast amount of training data. Several open source machine learning libraries are available, such as CudaConvNet [34], Torch [35], Theano [36], Caffe [37], and MatConvNet [38]. MatConvNet is a MATLAB toolbox implementing CNNs, which are optimized for both CPU and GPU platforms. It provides a friendly environment for research purposes together with high computational performance thanks to the C++ and CUDA-based implementations. For these advantages, we have also considered MatConvNet.

The MatConvNet toolbox provides some pre-trained CNN models and some functions to create and initialize new neural networks. As it was described in Section 2.2, we have considered the models GoogLeNet and ResNet, which were initially trained on the dataset ImageNet. Thus, we could use the weights and biases from these models

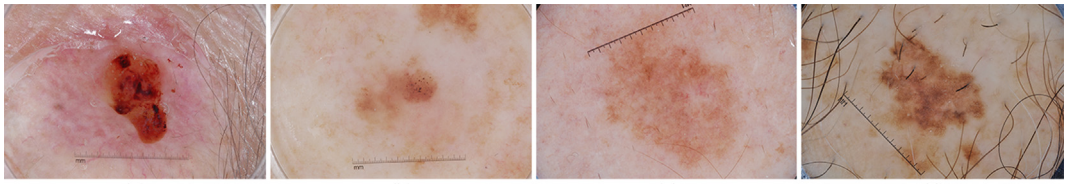

Fig. 3. Melanoma images from the test set misclassified by (a) GoogLeNet, (b) AlexNet, (c) ResNet and (d) VGGNet, but classified properly by the ensemble of CNNs.

trained on ImageNet, and just fine-tuned their layers. Moreover, we have included AlexNet and VGGNet, whose weights and biases were initialized by random values.

Fine-tuning and training have been performed on a computer equipped with an NVIDIA TITAN X GPU card with 7 TFlops of single precision, 336.5 GB/s of memory bandwidth, 3072 CUDA cores, and 12 GB memory. The convolutional filters of the CNNs were found by stochastic gradient descent algorithm iterated through 500 training epochs per neural network.

During these fine-tuning and training processes, the results of backpropagation and the classification of the validation images after each epoch was determined (see Fig. 5) using the top-1 error rate [20]. It can be observed that the overfitting problem is successfully resolved by the usually applied dataset extension methods, since the validation curves move together with the tests curves in a descending way.

6. Conclusion

The application of CNNs for skin lesion detection is promising; however, the lack of large annotated datasets to train their models is still a barrier to develop methods unquestionably suitable for clinical application. In this paper, we have investigated the possibilities of creating ensembles of deep neural networks to raise classification accuracy with combining their architectures to benefit from their strengths while overcoming their weaknesses when the number of annotated images available for training is insufficient.

The main practical motivation was to develop an automatic method to classify skin lesion images as nevus, melanoma or seborrheic keratosis. In this study, we used an image set that contained 2000 images for training, 150 images for validation, and 600 images for testing. The extension of the training set by additional images is not considered, because we have addressed the enhancement of the performance of the CNNs in another way. Namely, we have proposed the consideration of creating ensembles of them. We have found that if we weight the CNNs appropriately, let them to vote to only one class, and determine the final class label of the input image based on the sum of the maximum confidence levels then this ensemble outperforms the accuracy of the individual CNNs in this classification task.

We note that the proposed ensemble based CNN framework is modular, that is it can be extended using additional CNN networks and it can be customized for the actual usage. In other words, the framework can be set as it should not cause unnecessary worry for the patient or miss early malignant lesions that are easier to cure successfully. For example, the output of the framework should favor sensitivity to avoid missing any potential melanoma, at the expense of false positives that may be detected by a regular visit to the dermatologist. On the other hand, patients at high risk for melanoma could monitor themselves e.g. using their smartphones; in this case specificity should be favored. In both scenarios, the proposed computerized framework should be adapted to the goal of its application.

Table 2
The official result of the 2017 ISBI skin lesion challenge extended by the accuracy of the proposed $\omega^{(2)}$SMP combination model.

Rank	User name	Organization	AVG_AUC
1	Kazuhisa Matsunaga	Casio and Shinshu University joint team	0.911
2	monty python	Multimedia Processing Group - Universidad Carlos III de Madrid	0.910
3	RECOD Titans	RECOD Titans / UNICAMP	0.908
4	Lei Bi	USYD-BMIT	0.896
	B. Harangi [$\omega^{(2)}$*SMP*]	**University of Debrecen**	**0.891**
5	Xulei Yang	Institute of HPC + National Skin Center, Singapore	0.886
6	T D	University of Guelph - MLRG	0.886
7	Cristina Vasconcelos	icuff	0.851
8	Cristina Vasconcelos	icuff	0.850
9	Euijoon Ahn	USYD-BMIT	0.836
10	x j	CVI	0.829
11	B. Harangi [33]	University of Debrecen	0.825
12	INESC TECNALIA	INESC TEC Porto / TECNALIA	0.823
13	Rafael Sousa	Universidade Federal de Mato Grosso	0.823
14	Dylan Shen	Computer Vision Institute, Shenzhen University	0.823
15	Vic Lee	Computer Vision Institute, Shenzhen University	0.816
16	Masih Mahbod	IPA	0.811
17	Matt Berseth	NLPLOGIX/WISEEYE.AI	0.804
18	Dennis Murphree	Dennis Murphree	0.750
19	Hao Chang	Yale	0.705
20	Wenhao Zhang	CSMedical	0.658
21	Jaisakthi S.M.	SSNMLRG	0.655
22	Wiselin Jiji	Dr. Sivanthi Aditanar College of Engineering	0.497
23	Yanzhi Song	song	0.465

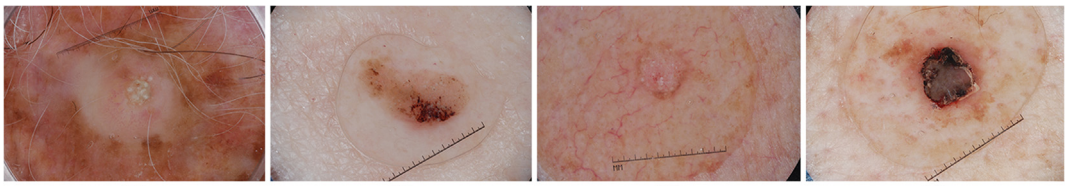

Fig. 4. Seborrheic keratosis images from the test set misclassified by (a) GoogLeNet, (b) AlexNet, (c) ResNet and (d) VGGNet, but classified properly by the ensemble of CNNs.

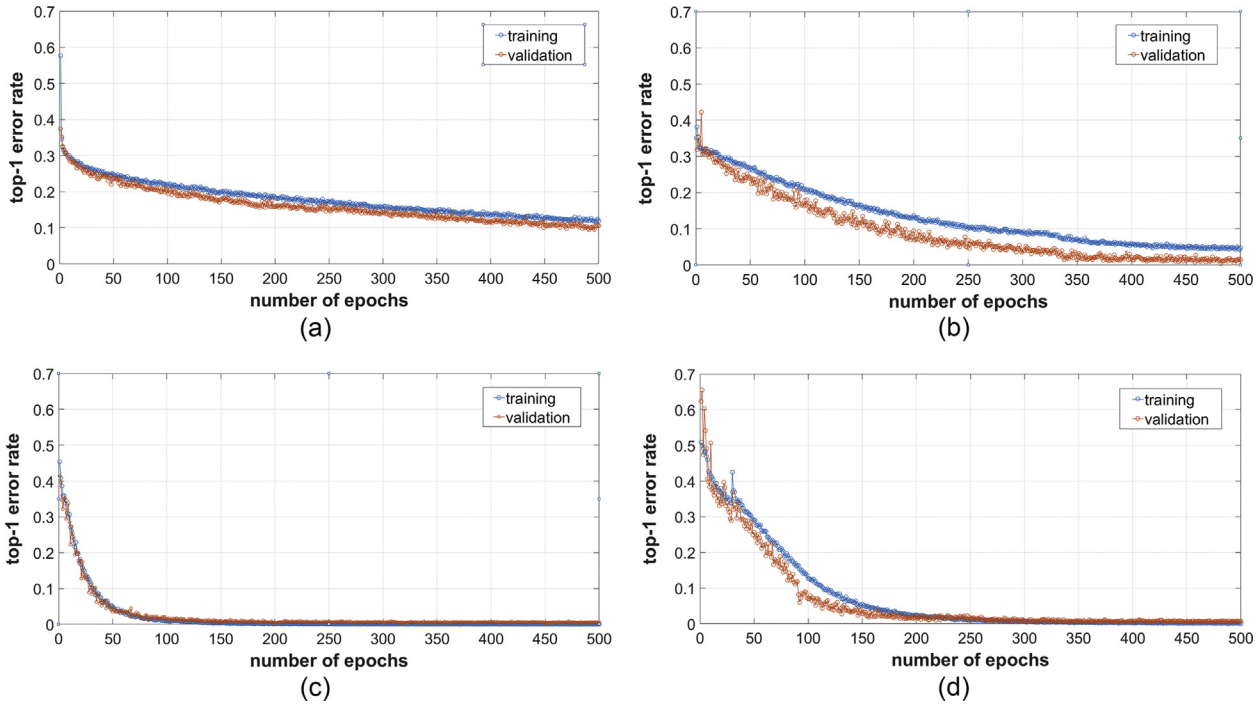

Fig. 5. Training and validation results during fine-tuning of (a) GoogLeNet, (b) ResNet, (c) AlexNet and (d) VGGNet.

Conflict of interest

None declared.

Acknowledgement

This work was supported in part by the projects GINOP-2.1.1-15-2015-00376 and EFOP-3.6.2-16-2017-00015 supported by the European Union and the State of Hungary, co-financed by the European Social Fund.

References

[1] L.A.G. Ries, D. Harkins, M. Krapcho, A. Mariotto, B.A. Miller, E.J. Feuer, L. Clegg, M.P. Eisner, M.J. Horner, N. Howlader, M. Hayat, B.F. Hankey, B.K. Edwards, SEER Cancer Statistics Review, 1975–2003, National Cancer Institute, Bethesda, 2006.
[2] Cancer Statistics Center 2017. < https://cancerstatisticscenter.cancer.org/#/ > (accessed: 07.09.2017).
[3] J. Feng, N.G. Isern, S.D. Burton, J.Z. Hu, Studies of secondary melanoma on C57BL/6J mouse liver using 1H NMR metabolomics, Metabolites 3 (4) (2013) 1011–1035.
[4] R. White, D.S. Rigel, R. Friedman, Computer applications in the diagnosis and prognosis of malignant melanoma, Dermatol. Clin. 9 (4) (1992) 695–702.
[5] Q. Abbas, M. Emre Celebi, I.F. Garcia, W. Ahmad, Melanoma recognition framework based on expert definition of ABCD for dermoscopic images, Skin Res. Technol. 19 (1) (2013) 93–102.
[6] R. Kasmi, K. Mokrani, Classification of malignant melanoma and benign skin lesions: implementation of automatic ABCD rule, IET Image Process. 10 (6) (2016) 448–455.
[7] F. Nachbar, W. Stolz, T. Merkle, A.B. Cognetta, T. Vogt, M. Landthaler, G. Plewig, The ABCD rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions, J. Am. Acad. Dermatol. 30 (4) (1994) 551–559.
[8] G. Sforza, G. Castellano, S.K. Arika, R.W. LeAnder, R.J. Stanley, W.V. Stoecker, J.R. Hagerty, Using adaptive thresholding and skewness correction to detect gray areas in melanoma in situ images, IEEE Trans. Instrum. Meas. 61 (7) (2012) 1839–1847.
[9] M. Emre Celebi, H.A. Kingravi, H. Iyatomi, Y. Alp Aslandogan, W.V. Stoecker, R.H. Moss, J.M. Malters, J.M. Grichnik, A.A. Marghoob, H.S. Rabinovitz, S.W. Menzies, Border detection in dermoscopy images using statistical region merging, Skin Res. Technol. 14 (3) (2008) 347–353.
[10] M. Sadeghi, M. Razmara, T.K. Lee, M.S. Atkins, A novel method for detection of pigment network in dermoscopic images using graphs, Comput. Med. Imaging Graph. 35 (2) (2011) 137–143.
[11] L. Yu, H. Chen, Q. Dou, J. Qin, P.A. Heng, Automated melanoma recognition in dermoscopy images via very deep residual networks, IEEE Trans. Med. Imaging 36 (4) (2017) 994–1004.
[12] R. Kasmi, K. Mokrani, R.K. Rader, J.G. Cole, W.V. Stoecker, Biologically inspired skin lesion segmentation using a geodesic active contour technique, Skin Res. Technol. 22 (2) (2015) 208–222.
[13] M.E. Celebi, H. Kingravi, B. Uddin, H. Iyatomi, A. Aslandogan, W.V. Stoecker, R.H. Moss, A methodological approach to the classification of dermoscopy images, Comput. Med. Imaging Graph. 31 (6) (2007) 362–373.
[14] Y. Faziloglu, R.J. Stanley, R.H. Moss, W.V. Stoecker, R.P. McLean, Colour histogram analysis for melanoma discrimination in clinical images, Skin Res. Technol. 9 (2003) 147–155.
[15] W.V. Stoecker, K. Gupta, R.J. Stanley, R.H. Moss, B. Shrestha, Detection of asymmetric blotches (asymmetric structureless areas) in dermoscopy images of malignant melanoma using relative color, Skin Res. Technol. 11 (3) (2005) 179–184.
[16] A. Hajdu, B. Harangi, R. Besenczi, I. Lázár, G. Emri, L. Hajdu, R. Tijdeman, Measuring regularity of network patterns by grid approximations using the LLL algorithm, 23rd International Conference on Pattern Recognition (ICPR), Cancún, Mexico, 2016, pp. 1524–1529.
[17] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115–118.
[18] N. Codella, D. Gutman, M.E. Celebi, B. Helba, M.A. Marchetti, S. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, A. Halpern, Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC), 2017. Available from: arXiv preprint < arXiv:1710.05006 > .
[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 1–9.
[20] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inform. Process. Syst. (2012) 1097–1105.
[21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 770–778.
[22] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-scale Imagerecognition, 2014. Available from: arXiv preprint < arXiv:1409.1556 > .
[23] Y. Neuman, Computational Personality Analysis: Introduction, Practical Applications and Novel Directions, first ed., Springer Publishing Company, Incorporated, 2016.
[24] T. Mendonça, P.M. Ferreira, J.S. Marques, A.R.S. Marcal, J. Rozeira, PH2 - a dermoscopic image database for research and benchmarking, Proceeding of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2013, pp. 5437–5440.
[25] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.
[26] R.A. Jacobs, Methods for combining experts' probability assessments, Neural Comput. 7 (5) (1995) 867–888.

[27] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (3) (1998) 226–239.
[28] T.G. Dietterich, Ensemble methods in machine learning, Multiple Classif. Syst., LNCS 1857 (2008) 1–15.
[29] A.R. Conn, N.I.M. Gould, Ph.L. Toint, A globally convergent augmented lagrangian barrier algorithm for optimization with general inequality constraints and simple bounds, Math. Comput. 66 (217) (1997) 261–288.
[30] V. Granville, M. Krivanek, J.-P. Rasson, Simulated annealing: a proof of convergence, IEEE Trans. Pattern Anal. Mach. Intell. 16 (6) (1994) 652–656.
[31] R.T. Sousa, L.V. de Moraes, Araguaia Medical Vision Lab at ISIC 2017 Skin Lesion Classification Challenge, 2017. Available from: arXiv preprint < arXiv:1703.00856 > .
[32] ISIC 2017: Skin Lesion Analysis Towards Melanoma Detection - Part 3: Lesion Classification. < https://challenge.kitware.com/#phase/584b0afccad3a51cc66c8e38 > (accessed: 16.01.2018).
[33] B. Harangi, Skin Lesion Detection based on an Ensemble of Deep Convolutional Neural Network, 2017. Available from: arXiv preprint < arXiv:1705.03360 > .
[34] https://code.google.com/p/cuda-convnet/.
[35] http://cilvr.nyu.edu/doku.php?id=code:start.
[36] http://deeplearning.net/software/theano/.
[37] http://caffe.berkeleyvision.org.
[38] A. Vedaldi, K. Lenc, MatConvNet - convolutional neural networks for MATLAB, Proceeding of the ACM International Conference on Multimedia, (2015).