Artificial Intelligence In Medicine 103 (2020) 101749

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

A multi-context CNN ensemble for small lesion detection

B. Savelli[a,*], A. Bria[a], M. Molinara[a], C. Marrocco[a], F. Tortorella[b]

[a] Department of Electrical and Information Engineering, University of Cassino and L.M., Via G. Di Biasio 43, 03043 Cassino (FR), Italy
[b] Department of Electrical, Information Engineering and Applied Mathematics, University of Salerno, via Giovanni Paolo II 132, 84084 Fisciano (SA), Italy

ARTICLE INFO

Keywords:
Ensemble classifier
Deep learning
Convolutional neural networks
Computer-aided detection (CADe)
Mammograms
Ocular fundus images

ABSTRACT

In this paper, we propose a novel method for the detection of small lesions in digital medical images. Our approach is based on a multi-context ensemble of convolutional neural networks (CNNs), aiming at learning different levels of image spatial context and improving detection performance. The main innovation behind the proposed method is the use of multiple-depth CNNs, individually trained on image patches of different dimensions and then combined together. In this way, the final ensemble is able to find and locate abnormalities on the images by exploiting both the local features and the surrounding context of a lesion. Experiments were focused on two well-known medical detection problems that have been recently faced with CNNs: microcalcification detection on full-field digital mammograms and microaneurysm detection on ocular fundus images. To this end, we used two publicly available datasets, INbreast and E-ophtha. Statistically significantly better detection performance were obtained by the proposed ensemble with respect to other approaches in the literature, demonstrating its effectiveness in the detection of small abnormalities.

1. Introduction

Thanks to the recent progress in medical image modalities and in Machine Learning techniques, systems for Computer Aided Detection and Diagnosis (CADe, CADx) play nowadays an essential role in modern medicine and are integral part of the clinical workflow for the detection, diagnosis, and treatment of various diseases [1,2]. These systems help physicians in the tedious and challenging task of interpreting the invaluable source of information being held in medical images, preventing decisions to be affected by errors and improving the detection of subtle but important changes in anatomical structures and tissues, essential to timely treat diseases [3]. For CADe development, several approaches have been reported in the literature of the last few decades, ranging from conventional image analysis methodologies to Machine Learning techniques [4–11]. Deep Learning models, and in particular convolutional neural networks (CNNs), have recently acquired great popularity thanks to their remarkable performance in computer vision [12,13] and have proved to be powerful also in medical image analysis [14–18]. The reason behind this success is the capability of learning hierarchical feature representations directly from data, instead of using handcrafted features based on domain-specific knowledge. The typical CNN architecture for image processing consists of a series of layers of convolutional filters spaced with downsampling layers. Convolutional filters are applied to small patches of the input images (containing candidate lesion or background) and are able to build features with increasing relevance, from texture to higher order features like local and global shape. The output of the CNN is typically one or more values that represent the probability that an image patch contains a lesion or not.

In this context, patch dimensions play an important role, especially when the lesion is particularly small and similar to the surrounding tissue. If the patch is defined so as to strictly contain the lesion, it may be too small to produce a set of sufficiently discriminating representations. On the other hand, a larger patch would include much more background which can bias the detection system to focus on uninteresting details contained in the background part. As a consequence, the number of background patches erroneously detected as lesions may be high and limit the benefits that the CADe system can provide, even when deep learning techniques are applied [19,20].

A simple yet effective way, commonly used in Machine Learning, for boosting the performance of poor detection models is the so called “expert combination”: multiple detectors are trained by using different weight settings and/or different partitions of the same data and strategically combined to solve a particular detection problem [21–25]. The rationale is that differently trained networks can learn different representations of the training data and, in this way, can agree on

* Corresponding author.
E-mail addresses: b.savelli@unicas.it (B. Savelli), a.bria@unicas.it (A. Bria), m.molinara@unicas.it (M. Molinara), c.marrocco@unicas.it (C. Marrocco), ftortorella@unisa.it (F. Tortorella).

https://doi.org/10.1016/j.artmed.2019.101749
Received 28 April 2019; Received in revised form 23 October 2019; Accepted 27 October 2019

correct predictions and make their errors in different parts of the input space. When combined together, such diversity enforces the correct predictions and reduces the errors, minimizing the risk due to poor model selection. This approach is also useful in medical image analysis field, where ensembles of CNNs have been used to solve many medical image analysis tasks [26–28].

In this paper, we present an approach for the automated detection of small lesions in medical images, consisting of an ensemble of CNNs, each one specifically designed to learn a different view of the same lesion. Patches of different dimensions, centered at the same detection location, are extracted to separately train different CNNs, whose network architectures are tailored to the dimensions of the input samples. The idea is that, starting from image patches small enough to entirely contain the lesion to be detected, the size of the neighbourhood is progressively enlarged, and the depth of the network is increased at the same time. In this way, shallower networks become specialized in learning local image features, whereas deeper ones are well suited to learn patterns of the contextual background tissues. Once trained, the detectors are combined together to obtain a final ensemble that can effectively detect abnormalities with a substantial reduction of false positive regions (thanks to the diversity provided by the different spatial context learned by each network).

Recently, few other works have tried to add contextual information into the training phase. [29] proposed a two-pathway CNN architecture for brain tumour segmentation. Similarly, [30] employed a dual pathway architecture that processes 3-D input images at multiple scales simultaneously for accurate brain lesion segmentation. [31] proposed a context-sensitive DNN for microcalcification detection by merging, at training time, features coming from two different subnetworks.

Our approach stands out from these works since the networks are separately trained and the probability scores are merged at inference time, by allowing to focus on more different portions of the lesion background, without requiring a high computational burden and resulting in a more discriminating power. To evaluate the performance of the proposed approach, we considered two well-known medical detection problems that have also been recently addressed with CNNs [32–34]. In particular, we focused on microcalcification detection on digital mammograms and on microaneurysm detection on digital fundus images. In both cases, the task of accurately identifying lesions is a main challenge, due to the appearance of the lesions and to the heterogeneity of their contextual backgrounds.

The rest of the paper is organised as follows. We start with a brief overview of convolutional neural networks in Section 2, whereas Section 3 introduces the underlying concepts of the proposed method along with a detailed characterization of the proposed architecture. Section 4 reports the experimental analysis, followed by results in Section 5. Finally, Section 6 ends the paper with discussion and conclusions.

2. Convolutional neural networks

In this work, the problem of detecting small lesions in medical images has been formulated in terms of a pixel-based two-class classification problem. To solve the classification task, we employed CNNs [35], a particular kind of deep neural networks well suited to work with images as they directly take in input 2D or 3D structures, preserving configuration information of the data. CNNs are based on three main architectural ideas: local receptive fields, weight sharing, and subsampling in the spatial domain. A typical CNN principally consists of three types of layers: (i) convolutional layers, (ii) sub-sampling layers, and (iii) output layers, that are arranged in a feed-forward structure [12]. Convolutional layers are responsible for detecting local features in all locations of the input images. To detect local structures, each node in a convolutional layer is connected to only a small subset of spatially connected neurons in the input image channels, called receptive field. Furthermore, to enable the search for the same local feature, connection weights are shared between all the nodes in the convolutional layers; each set of shared weights is called convolutional kernel. For each convolutional layer, a set of convolutional kernels $W = \{W_1, W_2, ..., W_n\}$ is convolved with the input image X, and biases $B = \{b_1, b_2, ..., b_n\}$ are added, so as to generate a new feature map X_i through an element-wise non-linear transform σ:

$$X_i = \sigma(W_i \, X + b_i) \qquad \forall \, i = 1, \, ..., n \tag{1}$$

This approach makes the network equivariant with respect to input translations and drastically reduces the number of parameters to be learned. Each sequence of convolutional layers is followed by max pooling layers, that are applied to reduce the size of feature maps by selecting the maximum value in local neighbourhoods. Like local connectivity, the pooling operation reduces the resolution w.r.t. previous layer and provides for translational invariance. At the end of the convolutional stream of the network, a number of consecutive fully connected layers is added, and the class distribution over the classes is generated by feeding them through an activation function.

The training procedure consists of an iterative propagation of samples through the network and modification of its weights, which are properly initialized [36]. CNNs are trained using the back-propagation algorithm by minimizing a given cost function with respect to the weights w. For a dataset D, the optimization objective is the average loss over all $|D|$ data instances:

$$L(w) = \frac{1}{|D|} \sum_{i=1}^{|D|} f_w(x(i)) \tag{2}$$

Since D can be very large, a stochastic approximation of this objective is used, where the cost over the entire training set is approximated with the cost over mini-batches of data. Drawing a mini-batch of $N << |D|$ instances the optimization function becomes:

$$L(w) \approx \frac{1}{|N|} \sum_{i=1}^{|N|} f_w(x(i)) \tag{3}$$

The stochastic gradient descent updates the weights w by a linear combination of the negative gradient ∇L_w and the previous weight update V_t according to the following formula:

$$V_{t+1} = \mu V_t - \alpha \nabla L(w_t) \tag{4}$$

where μ and α are hyperparameters chosen for the learning procedure. The coefficient α is the learning rate, controlling the size of the weight updates, whereas μ is the momentum, that indicates the contribution of the previous weight update in the current iteration. In order to prevent overfitting, some regularization technique are applied during the training procedure, being dropout [37] the most used. With dropout a subset of network units is drawn at random and temporarily "switched off" during training. When in this state, those units do not propagate signals when a sample is presented, nor participate in the process of error backpropagation.

3. Multi-context CNN ensemble

In this section, we present our multi-context CNN ensemble for the detection of small lesions in medical images. Specifically, the proposed ensemble consists of K different CNNs that are meant to focus on different spatial context of the images and thus to specialize both on local features and on contextual ones. To this end, each network of the ensemble is trained by using image patches of different size, aiming to capture the spatial context around the same detection location. Furthermore, according to the image patch dimensions, the K network architectures are set to different levels of depth, with the aim of using deeper, hence more discriminating, networks to manage larger image windows.

The size m of the smallest patches used in the ensemble is chosen to entirely contain a single lesion, and then it is progressively enlarged to include larger image portions, up to a dimension that is still

Table 1
Details of the *incremental block*.

Layer	Type	Output size	Kernel size	Stride	Padding
1	Convolutional	$32 \times m \times m$	3×3	1	1
2	ReLU	$32 \times m \times m$			
3	Convolutional	$32 \times m \times m$	3×3	1	1
4	ReLU	$32 \times m \times m$			
5	Max pooling	$32 \times \frac{m}{2} \times \frac{m}{2}$	2×2	2	1

representative of the context around the lesion. Similarly, the network architecture is set to a baseline configuration, and then its depth is increased as the image size grows. The baseline configuration is inspired by the VGG-Net [38], and it is defined as two blocks of two convolutional layers, interlaced by a ReLU activation function and followed by a max pooling layer. We named each of these blocks *incremental block*; the word incremental indicating they are added to the stack of layers in order to define deeper architectures. More details of the structure of an incremental block are given in Table 1. Following the design approach defined by the VGG-Net [38], small 3×3 kernels are used in each block, since they are faster to convolve with and contain less weights. For the same purpose of decreasing the amount of computations, data reduction layers need to be set to steadily decrease the spatial resolution of the input feature maps. Let s_{in} be the size of an image patch in input to a convolutional layer or a max pooling layer, we know that its output dimensions can be expressed as:

$$s_{\text{out}} = \frac{s_{\text{in}} + 2^{*}\text{pad} - \text{kernel}}{\text{stride}} + 1 \tag{5}$$

where *kernel* indicates the size of the filter, *pad* specifies the padding size, and *stride* the intervals at which the filter is applied. We set the stride of convolutional layers equal to 1, by fixing instead the stride of max pooling layers equal to 2 (see Table 1). As a result, the image patches are halved after each passage through an incremental block. As a consequence, we decided to progressively double the size of the input patches every time we added an incremental block to the baseline network architecture. To summarize, we can say that the proposed ensemble of CNNs consists of K different networks, each one trained on image patches of size $s = \{2^{i-1}m \times 2^{i-1}m\}$ and built with $d = i + 1$ incremental blocks, $\forall i = 1, 2, \ldots, K$.

Each of the K networks ends with a *classification block*, i.e., with three fully connected layers intertwined with two dropout layers. At the end, a softmax function is applied to the two-output neurons to generate a two-value probability vector associated to each prediction. More details on the classification block are reported in Table 2. The K nets are individually trained and the output values Y_i, $\forall i = 1, \ldots, K$ of the K CNNs are merged together at inference time to aggregate the multi-level contextual information for the final classification. In particular, the probability values are averaged, resulting in a single probability vector $Y_{\text{en}} = \{Y_{\text{en},p}, Y_{\text{en},n}\}$ associated to each patch, stating the final decision about that sample:

$$Y_{en} = \{Y_{en,p}, Y_{en,n}\} = \{\sum_{i=1}^{K} \frac{Y_{i,p}}{K}, \sum_{i=1}^{K} \frac{Y_{i,n}}{K}\} \tag{6}$$

Table 2
Details of the *classification block*.

Layer	Type	Output size	Kernel size	Rate
1	Fully connected	256	1×1	
2	Dropout	256		0.5
3	Fully connected	256	1×1	
4	Dropout	256		0.5
5	Fully connected	2	1×1	

4. Experiments

We proved the effectiveness of the proposed approach on two well-known problems in medical image analysis: (i) the detection of microcalcifications on full field digital mammograms, and (ii) the detection of mycroaneurysms on digital ocular fundus images.

Microcalcifications (MCs) are one of the main symptoms of breast cancer and detecting them on mammograms is one of the most reliable way to identify the presence of breast cancer at an early stage [39]. MCs appear on mammograms (see some examples in Fig. 1(a–c)) as small granular bright spots of size between 0.1 mm and 1 mm, and they may occur alone or in clusters as a group of MCs closely distributed within a spatial region [40]. The task of accurately identifying individual MCs is very challenging due to their small dimensions and because of the inhomogeneity of the surrounding breast tissue. Furthermore, mammograms contain a variety of linear structures (e.g. vessels, ducts, etc.) that, together with MC-like noise patterns and artefacts, are very similar to MCs in size and shape, thus contributing to the occurrence of false positives in the MC-detection task [41].

Microaneurysms (MAs) are one of the first sign of diabetic retinopathy and the most common cause of blindness and vision loss in the working population of the western world, as stated by the World Health Organization in 2016. The screening programs use non mydriatic digital colour fundus cameras to acquire photographs of the retina (see some examples in Fig. 1(d–f)), and MAs detection represents a critical step in the process of early diagnosis and timely treatment of the disease MAs are described as isolated, small, round objects, of 10-100 µm of diameter, but sometimes they appear in combination with vessels. Retinal vessels, together with dot-hemorrhages and some other objects like the small and round spots resulting from the crossing of thin blood vessels, make MAs hard to distinguish

4.1. Datasets

4.1.1. MCs detection

We used the publicly available INbreast database [42]. This database was acquired from the Breast Centre of the university hospital of Porto, between April 2008 and July 2010, by using a MammoNovation Siemens full field digital acquisition system, equipped with a solid-state detector of amorphous selenium. The acquired images are matrices of 3328×4084 or 2560×3328 pixels, with a pixel-size of 70 µm and a 14-bit contrast resolution. The database has a total of 410 images, amounting to 115 cases, from which 90 cases are from women with both breasts, and 25 are from mastectomy patients. Several types of lesions such as masses, calcifications, and architectural distortions are included. Among the 410 images, calcifications can be found in 301 images, and a total of 6,880 individual calcifications have been identified. All mammograms were manually annotated and segmented by expert radiologists, and ground-truth data are provided.

In our experiments, all the images were used and image patches were extracted from the mammograms to train the CNNs. Each patch was labeled as positive or negative according to the information provided by the ground-truth. MC patches were extracted by centering the windows on the annotated MC centers, whereas background tissue patches were extracted from the remaining regions of the images with overlapping sliding windows. According to the multi-patch criterion, different subwindows of different size were extracted around the same center, by yielding 5628 positive samples and 26,887,769 negative ones. The resulting patches were used to train and test the proposed detection system.

4.1.2. MAs detection

E-ophtha is a public database of colour fundus images designed for scientific research in Diabetic Retinopathy [43]. It contains 233 healthy images and 148 images with microaneurysms or small hemorrhages which are manually annotated by expert ophthalmologists. The image

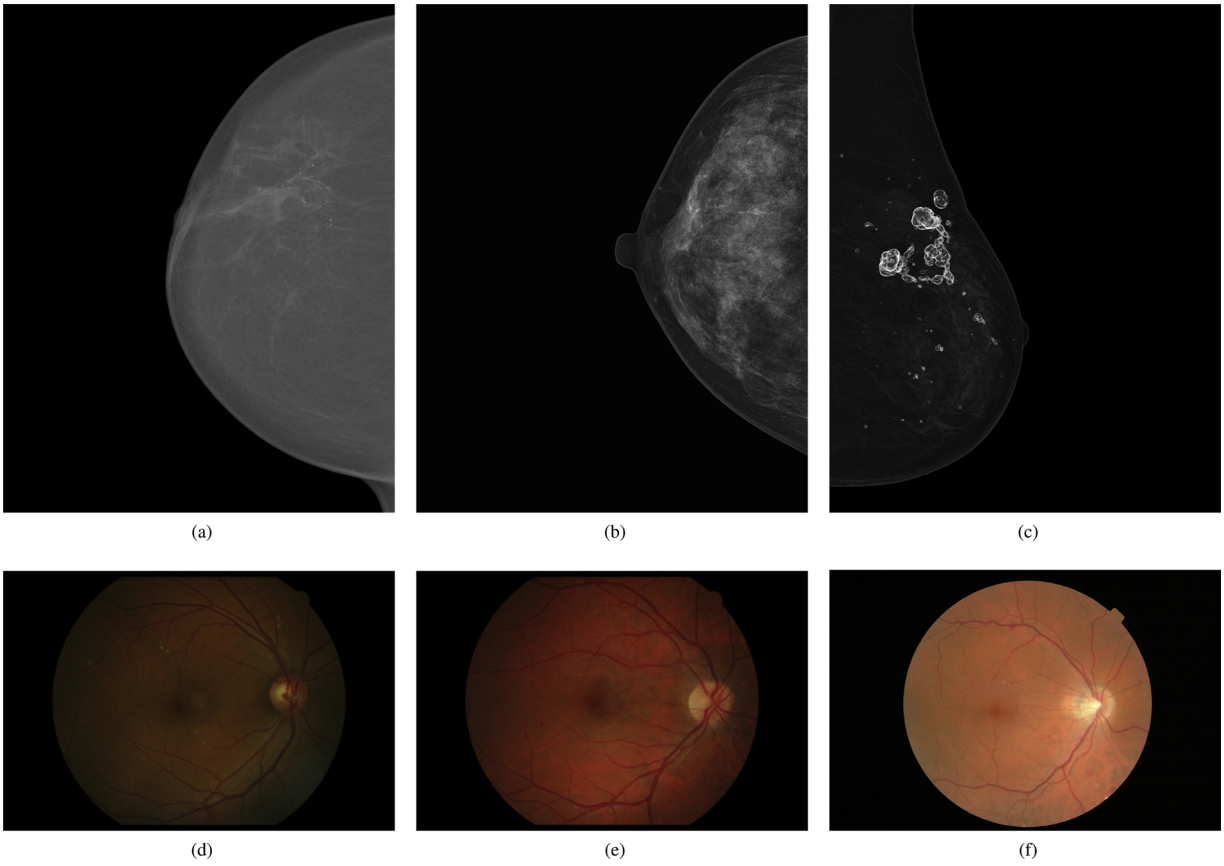

Fig. 1. Some examples of images from (a–c) INbreast and (d–f) E-ophtha respectively.

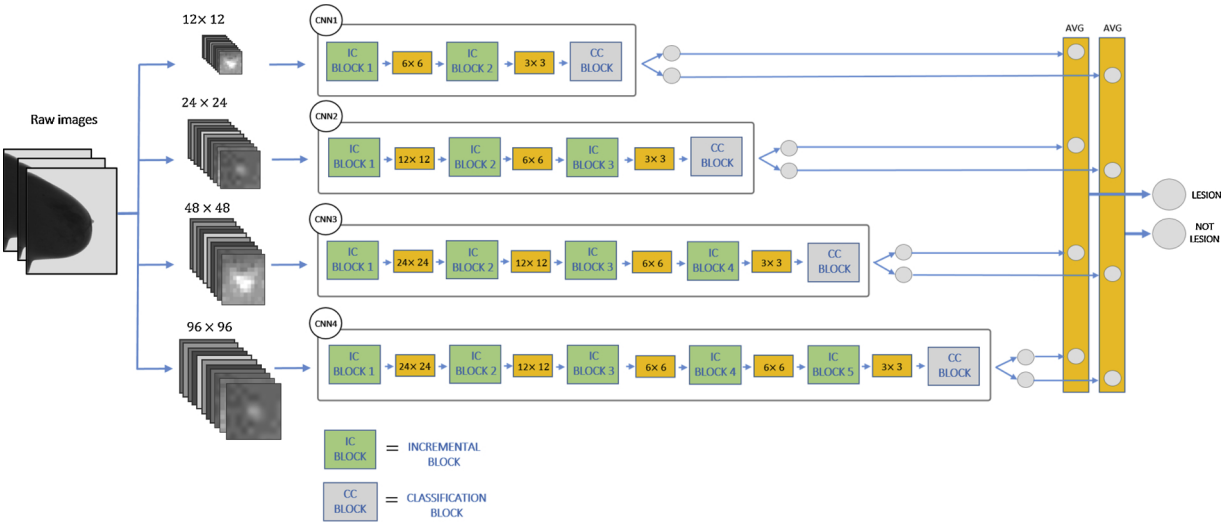

Fig. 2. Details of the proposed architecture.

matrices range from 1440 × 960 to 2544 × 1696 pixels with a 45° field of view (FOV), a pixel size of 7μm and JPEG format.

Retinal colour fundus images are RGB images, but experimental studies have observed that the blue channel is often characterized by low contrast and shadows and does not contain any information, whereas the red one is generally noisy or saturated. Therefore, we decided to extract the green channel and to work with grayscale images. Also in this case image patches were generated and associated to a class label according to the information provided by the ground-truth data. For the same location different image patches were extracted according to the multi-patch requirement, by yielding 1306 positive samples and 6,159,906 negative ones.

4.2. Architecture details

Our model consists of a multiple pathway of K specialized CNNs, each learning a different context extracted from an increasing area centered on the lesion. The choice of the size m is made in order to guarantee that input patches entirely contain at least the smallest lesions. Considering the size of MCs (0.1–1 mm) and MAs (10–100 μm) and the image spatial resolution, we found that in both tasks a patch size of $m = 12$ pixels was sufficient to cover the extent of the smallest

Table 3
Distribution of positive and negative samples before and after data augmentation for INbreast and E-ophtha datasets.

Dataset	Original		Augmented	
	pos	neg	pos	neg
INbreast	5628	26,887,769	26,887,769	26,877,769
E-ophtha	1306	6,159,906	6,159,906	6,159,906

lesions and to focus on their fine details. Then, we enlarged the input size for the other CNNs to capture larger lesions as well as their background context, by doubling the patch dimensions up to 96 pixels. Larger image portions were not considered since they were not representative of the background context of the lesions and to maintain a reasonable processing time (see Section 5 and Table 11).

In summary, the patch size ranges from 12×12 to 96×96, resulting in a final ensemble made up of $K = 4$ networks, the first ones more focused on learning details of the lesions and the others on learning background patterns.

The final architecture of the ensemble along with the dimension details of each CNNs are illustrated in Fig. 2.

4.3. Training parameters

According to the number of extracted patches, both MCs and MAs detection are heavily unbalanced classification problems. To avoid the classifiers being overwhelmed by the majority class and misclassify the samples of the minority class, we applied data augmentation, by restoring the balance between positive and negative samples(see Table 3). Thus, all the CNNs of the ensemble were trained on a perfectly balanced dataset. Augmentation of the positive class was performed by randomly flipping the patches horizontally and vertically and by randomly rotating the patches 90°, 180°, and 270°. Once generated, image patches were standardized by mean subtraction and normalization to unit variance [44].

As to weight initialization and training parameters, all the CNNs of the ensemble were treated in the same way. For all weights in all the layers we used Xavier initialization [36], while each CNN was optimized to minimize the Softmax loss function by using backpropagation and Mini-Batch Stochastic Gradient Descent. The mini-batch size was of 32 samples and in each mini-batch positive and negative samples were balanced. The learning rate was set to the initial value of 10^{-3} and decreased during training by a factor of 10 every 6 epochs. The learning was stopped after 30 epochs. Momentum and weight decay were set respectively to 0.9 and 5×10^{-4}. The number of feature maps was set to 32, whereas dropout was performed with a probability of 0.5 indicating that, at each training stage, half of the units coming from the previous layer were ignored in the training of the successive layer. The proposed architecture was implemented with a modified version of the Caffe framework [45], and the experiments were conducted on a machine with 2 Intel Xeon e5-2609, 256 GB of RAM and 2 GPU NVIDIA Titan Xp.

5. Results

To evaluate the performance of the proposed ensemble, we applied an image-based 2-fold cross validation for all the experiments. In each cross validation step, each detector was trained on the 50% of the images and tested on the other 50%. When splitting the data into training and test sets, the patches belonging to the same image were assigned to the same set.

The detectors were evaluated in terms of Receiver Operating Characteristics (ROC) curve by plotting True Positive Rate (TPR) against False Positive Rate (FPR) for a series of thresholds on the detector output associated to each sample. It is worth remarking that the ROC curves were calculated using the image patches. The number of negative and positive patches tested are the same of the original dataset as reported in Table 3 (the two leftmost columns). Furthermore, the mean sensitivity of the ROC curve in the specificity range on a logarithmic scale was calculated and compared. The mean sensitivity [33] is defined as:

$$\bar{S}(a, b) = \frac{1}{\ln(b) - \ln(a)} \int_a^b \frac{s(f)}{f} \mathrm{d}f \tag{7}$$

where a and b are the lower and upper bound of the false positive fraction and $s(f)$ is the sensitivity at the false positive fraction f. The range $[a, b]$ in Eq. (7) was set to $[10^{-6}, 10^{-1}]$ corresponding to a wide range of operating points that are close to practical application requirements of CADe systems for both the problems under consideration [46].

Table 4
Results of mean MC and MA detection sensitivity $\bar{S}$ for standalone CNNs.

Method	Patch size	d	$\bar{S}_{MC}$	$\bar{S}_{MA}$
CNN1	12×12	2	76.30	70.11
CNN2	24×24	3	76.90	77.82
CNN3	48×48	4	77.45	76.29
CNN4	96×96	5	75.83	74.64

For the experimental evaluation, we firstly investigated the performance of the standalone CNNs, by varying the input patch size along with the network depth. In Table 4, the performance of the individually trained CNNs for growing values of patch size and network depth are reported, for MCs and MAs respectively. We can see that using larger patches with a deeper network is initially beneficial to improve detection performance for both MAs and MCs cases. The mean sensitivity increases from 76.30% of CNN1 to 77.45% of CNN3 for MCs, and from 70.11% of CNN1 to 77.82% of CNN2 for MAs. However, in both cases, increasing the size of the image window stops to be beneficial and performance decreases. The mean sensitivity reduced from 77.45% of CNN3 to 75.83% of CNN4 in the case of MCs and from 77.82% of CNN2 to 74.64% of CNN4 for MAs.

Furthermore, to understand how joint predictions of the individual pathways affects the performance, we also report in Table 5 the results obtained by combining the single CNNs. We can see that detection performance increases each time we add a new CNN to the ensemble, obtaining the best performance measure (indicated in bold) when all the networks are used. The proposed full architecture achieved a mean sensitivity of 83.54% and 81.62% respectively for MCs and MAs. It is worth noting that, even when a single CNN does not perform very well (as in the extreme cases of patch size 12 and 96) they still give a contribution when added to the ensemble.

Starting from this observation, we decided to train a CNN5 with patch size 192 and $d = 6$ in order to evaluate its contribution to the ensemble. Results are reported in Table 6. We can see that, for both MCs and MAs, the standalone CNN5 achieved lower performance than the CNN4, but in this case even when including CNN5 in the ensemble performance does not improve. This shows how adding background portions that are not representative of the lesion context does not give any significant contribution to the ensemble.

For the sake of completeness, we also investigated the effect on the proposed approach of different combination methods in addition to the mean rule. In particular, we combined the probability values of the standalone CNNs with the following rules [21]: (i) trimmed mean; (ii) maximum; (iii) minimum; and (iv) majority voting. Results are reported in Table 7 showing that the mean rule gave the best performance (in bold) in both cases.

To evaluate the performance of the proposed approach with respect to the literature, we compared our ensemble method with the deep network proposed by Wang et al. [31], a context-sensitive deep learning

Table 5
Results of mean MC and MA detection sensitivity $\bar{S}$ for combined CNNs.

Method	Patch size	d	$\bar{S}_{MC}$	$\bar{S}_{MA}$
CNN1 + CNN2	12 + 24	2 + 3	79.51	79.04
CNN1 + CNN2 + CNN3	12 + 24 + 48	2 + 3 + 4	81.39	81.12
CNN1 + CNN2 + CNN3 + CNN4	12 + 24 + 48 + 96	2 + 3 + 4 + 5	**83.54**	**81.62**

Table 6
Results of mean MC and MA detection sensitivity $\bar{S}$ when considering CNN5.

Method	Patch size	d	$\bar{S}_{MC}$	$\bar{S}_{MA}$
CNN5	192 × 192	6	74.24	72.77
CNN1 + CNN2 + CNN3 + CNN4 + CNN5	12 + 24 + 48 + 96 + 192	2 + 3 + 4 + 5 + 6	83.34	81.41

approach for MCs detection. To this end, we faithfully reproduced the network architecture and the training settings reported in [31] and evaluated its performance in terms of mean sensitivity both for MCs and MAs detection. For the sake of completeness, we also compared these two approaches with the best single CNNs, that are CNN3 for MCs and CNN2 for MAs. Statistical comparisons were performed by means of bootstrapping [47]. On the test set, average ROC curves were calculated over 1000 bootstraps and are reported in Fig. 3. In all test cases, the ROC curves of the proposed context-sensitive ensemble were notably higher in the FPR range of major interest with respect to those obtained from the other approaches.

Additionally, the mean sensitivity was calculated for each bootstrap and p-values were computed for testing significance. The statistical significance level was chosen as $\alpha = 0.05$, but performance differences were considered statistically significant if $p < 0.025$ due to the Bonferroni correction[1] [48]. Comparative results are reported in Tables 8 and 9 and statistically significant peformance are indicated in bold. Results of the proposed architecture were statistically significantly better than the other considered approaches. The improvements in mean sensitivity were large with respect to both the context-sensitive approach of [31], +2.70% for MCs detection and +8.43% for MAs, and the best standalone CNN, +6.09% and +3.8% respectively for MCs and MAs, revealing to be significantly better in detecting lesions.

To assess the performance on the whole image, we calculated the lesion-based free receiver operating characteristic (FROC) curve that reports the true positive fraction of the detected lesions versus the average number of false positives per image (FPpI) when varying the decision threshold over the operating range. Being r the radius of a lesion in the ground truth, a detected region is considered as a TP if its distance to the centre of a true lesion is no larger than r; otherwise it is counted as an FP. To easily compare the different methods, the detection performance was summarized in a single score (FROC score) obtained by averaging the sensitivity values corresponding to the FPpI rates values of 1/8, 1/4, 1/2, 1, 2, 4, and 8, as described in [49]. Lesion-based FROC curves evaluated on the test set are shown in Fig. 4 for MCs and MAs detection and the relative FROC scores are reported in Table 10 (best values in bold). In all cases, the performance of the proposed ensemble is notably higher than the others for both tasks, proving the effectiveness of the proposed method also when applied on the whole image.

Finally, per-image processing times are reported in Table 11. As expected, the time needed for testing a single image increases with the input size, being strictly related to the network depth. The testing time of the proposed approach is evaluated as the sum of the processing time of the 4 standalone CNNs,[2] resulting to be lower than the time required by [31]. It is also worth noting that CNN5 required a per-image processing time much higher than the proposed approach. Thus, if included in the ensemble, it would make the processing time no longer in line with the requirements of a medical application.

Table 7
Results of MC and MA detection sensitivity $\bar{S}$ for combined CNNs according to different combination rules.

CNN1 + CNN2 + CNN3 + CNN4	$\bar{S}_{MC}$	$\bar{S}_{MA}$
Mean	**83.54**	**81.62**
Trimmed mean	81.92	80.57
Max	77.51	78.89
Min	81.27	78.12
Majority voting	81.25	79.30

6. Discussion and conclusions

In this paper, we proposed a novel and effective method for the detection of small lesions in digital medical images, as a result of an analysis of the limitations of the current methods proposed for similar applications.

First, we investigated the performance of CNNs when using larger image windows during the training phase together with deeper architecture. The obtained results indicate that using small patches, hence focusing only on the local image characteristic of a lesion, is not sufficient to obtain high detection performance. This is because, ignoring the context in which the lesions are in, the detector response is susceptible to all the lesion-like image patterns that lies in the background, affecting the overall performance of the classifier. However, even too large image patches are not sufficient to obtain high level performance, being the network not able to capture the fine details of the lesions and to recognize them in their broad spectrum of appearance. Moreover, deeper networks are more difficult to train, due to the vanishing signals and the internal covariate phenomenon.

To include both local and larger contextual information, we decided to combine at inference time the predictions coming from the individual networks, resulting in the proposed multi-context CNN ensemble. The ensemble combines the predictions of 4 different networks, each one with a different level of depth and processing at training time input patches with a different level of context-information. The devised approach achieved statistically significantly more accurate results in detecting small lesions when compared to standalone CNNs, and it additionally outperformed the context-sensitive approach proposed by [31] for similar tasks.

The obtained results proved the effectiveness of using different

[1] the significance level was obtained as α divided by the number of comparisons.

[2] This assumption refers to the worst case in which only one processing node is available and thus the 4 CNNs must be activated sequentially.

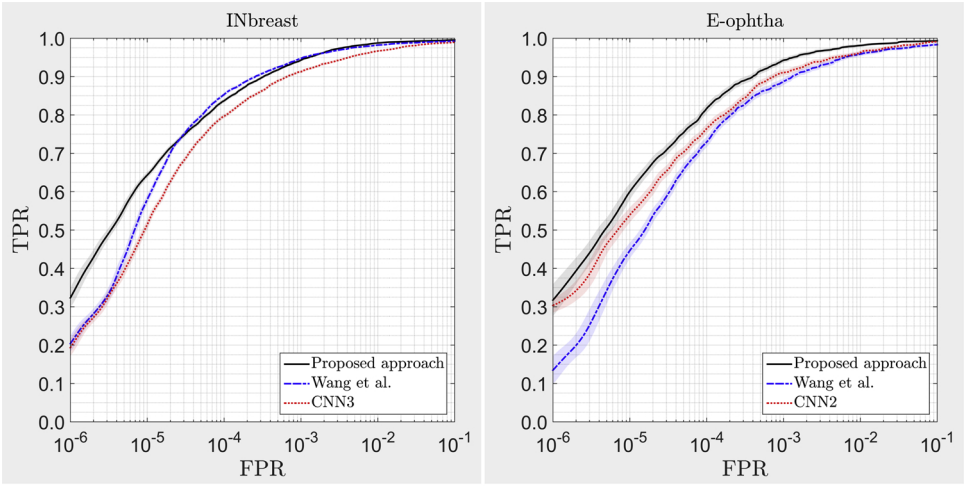

Fig. 3. Average ROC curves obtained from 1000 bootstrap iterations for (a) INbreast dataset and (b) E-ophtha. Confidence bands indicate 95% confidence intervals along the TPR axis.

Table 8
Comparative results of mean MC detection sensitivity $\bar{S}$.

Method	$\bar{S}$	Compared to	Difference	*p*-Value
CNN3	77.45	–	–	–
Wang et al.	80.84	–	–	–
Proposed approach	**83.54**	CNN3	+6.09	< **0.025**
		Wang et al.	+2.7	< **0.025**

Table 9
Comparative results of mean MA detection sensitivity $\bar{S}$.

Method	$\bar{S}$	Compared to	Difference	*p*-value
CNN2	77.82	–	–	–
Wang et al.	73.19	–	–	–
Proposed approach	**81.62**	CNN2	+3.8	< **0.025**
		Wang et al.	+8.43	< **0.025**

pathways, where each path specializes in capturing information at different context levels so that the system is able to closely learn the global contextual features as well as the local detailed features. The local appearance of the lesions and their underlying characteristics were captured, with different specificity levels, by the first two pathways, while higher level features, such as the nature of the tissues of the lesions are inserted in, were learned by the deeper paths. As a result, we obtained a set of specialized and complementary detectors (based on different representations derived from the different contexts) whose combination led to a final system that is able to overcome the limitations of single-pathway networks, with a clear improvement of the discriminating power. Moreover, the reported results suggest that the approach of training the networks separately and averaging the outputs at inference time is effective to get over the optimization difficulties that might occur in the case of joint training. We think that, when the multiple pathways are simultaneously trained as in [31], the detector might find it difficult during the learning phase to come across the co-adaption between the local and the global pathways. Finally, we can observe that the very good performance obtained for both MCs and MAs detection showed that the proposed approach is not designed for a specific task, thus making it suitable for other detection tasks.

Starting from the last remark, our future work will be devoted to extend the proposed method to similar CADe problems as well as to explore other network architectures and combining methods. Furthermore, we believe that the improvements in detection performance obtained in this work can be transferred to full CAD schemes, including diagnosis modules which can determine the nature of the

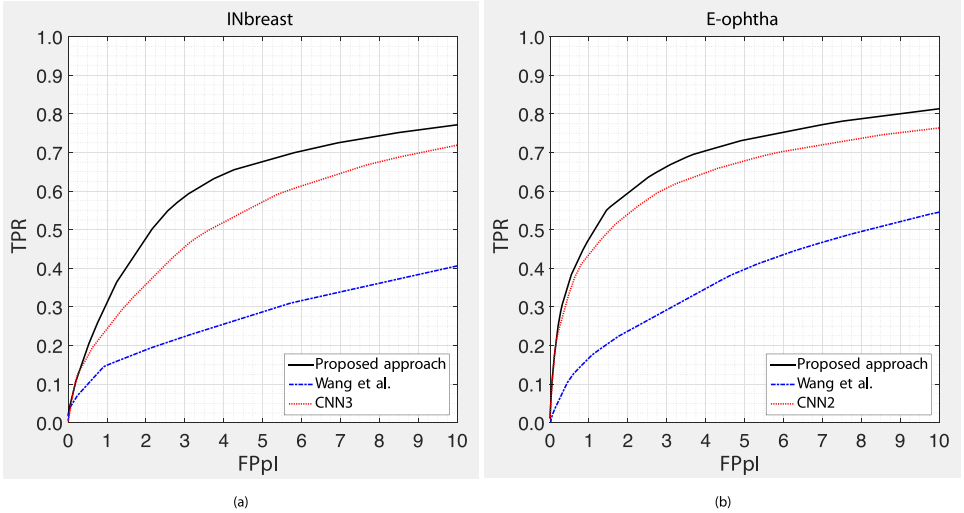

Fig. 4. FROC curves for (a) INbreast dataset and (b) E-ophtha.

Table 10
Comparative results of the FROC score and sensitivities at specific FPpI.

Dataset	Method	Sensitivity against FPpI							FROC score
		1/8	1/4	1/2	1	2	4	8	
INbreast	CNN3	0.0699	0.1260	0.1684	0.2378	0.3531	0.5148	0.6746	0.3064
	Wang et al.	0.0515	0.0713	0.0990	0.1487	0.1886	0.2552	0.3611	0.1679
	Proposed approach	0.0684	0.1186	0.1925	0.3025	0.4695	0.6433	0.7430	**0.3625**
E-ophtha	CNN2	0.1697	0.2487	0.3354	0.4288	0.5343	0.6455	0.7354	0.4425
	Wang et al.	0.0333	0.0557	0.1090	0.1659	0.2352	0.3350	0.4935	0.2039
	Proposed approach	0.1815	0.2602	0.3608	0.4734	0.5925	0.7016	0.7862	**0.4795**

Table 11
Results of MC and MA per-image processing time for the trained networks.

Method	t_{MC}	t_{MA}
CNN1	7 s	4 s
CNN2	21 s	10 s
CNN3	78 s	39 s
CNN4	280 s	150 s
CNN5	812 s	464 s
Wang. et al	822 s	266 s
Proposed approach	386 s	203 s

detected lesions. These systems could be implemented in a routine clinical setting, being very useful to the clinicians not only for detecting suspect cases, but also for assisting in the diagnostic decision as a second reading.

Declaration of interests

None declared.

Acknowledgment

The authors gratefully acknowledge the support of NVIDIA Corporation for the donation of the Titan Xp GPUs.

References

[1] Shiraishi J, Li Q, Appelbaum D, Doi K. Computer-aided diagnosis and artificial intelligence in clinical imaging. Seminars in nuclear medicine, vol. 41 2011:449–62.
[2] Suzuki K, Zhou L, Wang Q. Machine learning in medical imaging. Pattern Recognit 2017;63:465–7.
[3] Eadie LH, Taylor P, Gibson AP. A systematic review of computer-assisted diagnosis in diagnostic cancer imaging. Eur J Radiol 2012;81(1):e70–6.
[4] Papadopoulos A, Fotiadis DI, Likas A. An automatic microcalcification detection system based on a hybrid neural network classifier. Artif Intell Med 2002;25(2):149–67.
[5] D'Elia C, Marrocco C, Molinara M, Tortorella F. Detection of clusters of microcalcifications in mammograms: a multi classifier approach. Int. symp. on computer-based med. syst. 2008. p. 572–7.
[6] Marrocco C, Molinara M, D'Elia C, Tortorella F. A computer-aided detection system for clustered microcalcifications. Artif Intell Med 2010;50(1):23–32.
[7] Putzu L, Caocci G, Di Ruberto C. Leucocyte classification for leukaemia detection using image processing techniques. Artif Intell Med 2014;62(3):179–91.
[8] Pereira C, Veiga D, Mahdjoub J, Guessoum Z, Gonçalves L, Ferreira M, et al. Using a multi-agent system approach for microaneurysm detection in fundus images. Artif Intell Med 2014;60(3):179–88.
[9] Zhou F-Y, Jin L-P, Dong J. Premature ventricular contraction detection combining deep neural networks and rules inference. Artif Intell Med 2017;79:42–51.
[10] Höfener H, Homeyer A, Weiss N, Molin J, Lundström CF, Hahn HK. Deep learning nuclei detection: a simple approach can deliver state-of-the-art results. Comput Med Imag Graph 2018;70:43–52.
[11] Saha M, Chakraborty C, Racoceanu D. Efficient deep learning model for mitosis detection using breast histopathology images. Comput Med Imag Graph 2018;64:29–40.
[12] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.
[13] He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. Proc. of the IEEE int. conf. on computer vision 2015:1026–34.
[14] Litjens G, Kooi T, Bejnordi BE, Setio A, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
[15] Suzuki K. Overview of deep learning in medical imaging. Radiol Phys Technol 2017;10(3):257–73.
[16] Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19:221–48.
[17] Bernal J, Kushibar K, Asfaw DS, Valverde S, Oliver A, Martí R, et al. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. Artif Intell Med 2018;95(3):64–81.
[18] Arifoglu D, Bouchachia A. Detection of abnormal behaviour for dementia sufferers using convolutional neural networks. Artif Intell Med 2019;94:88–95.
[19] Shi Z, Hao H, Zhao M, Feng Y, He L, Wang Y, et al. A deep CNN based transfer learning method for false positive reduction. Multimed Tools Appl 2018:1–17.
[20] Li C, Zhu G, Wu X, Wang Y. False-positive reduction on lung nodules detection in chest radiographs by ensemble of convolutional neural networks. IEEE Access 2018;6:16060–7.
[21] Kuncheva LI. Combining pattern classifiers: methods and algorithms. 2nd ed. John Wiley & Sons; 2014.
[22] Marrocco C, Molinara M, Tortorella F. Exploiting AUC for optimal linear combinations of dichotomizers. Pattern Recognit Lett 2006;27(8):900–7.
[23] Ricamato MT, Marrocco C, Tortorella F. MCS-based balancing techniques for skewed classes: an empirical comparison. 19th int. conf. on patt. rec. 2008. p. 1–4.
[24] De Stefano C, Fontanella F, Marrocco C, Scotto di Freca A. A hybrid evolutionary algorithm for bayesian networks learning: an application to classifier combination. PART 1 LNCS, vol. 6024 2010:221–30.
[25] De Stefano C, Folino G, Fontanella F, Scotto Di Freca A. Using bayesian networks for selecting classifiers in GP ensembles. Inf Sci 2014;258:200–16.
[26] Li H, Jiang G, Zhang J, Wang R, Wang Z, Zheng W-S, et al. Fully convolutional network ensembles for white matter hyperintensities segmentation in mr images. NeuroImage 2018;183:650–65.
[27] Benou A, Veksler R, Friedman A, Raviv TR. Ensemble of expert deep neural networks for spatio-temporal denoising of contrast-enhanced mri sequences. Med Image Anal 2017;42:145–59.
[28] Zilly J, Buhmann JM, Mahapatra D. Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation. Comput Med Imag Graph 2017;55:28–41.
[29] Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, et al. Brain tumor segmentation with deep neural networks. Med Image Anal 2017;35:18–31.
[30] Kamnitsas K, Ledig C, Newcombe VF, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Med Image Anal 2017;36:61–78.
[31] Wang J, Yang Y. A context-sensitive deep learning approach for microcalcification detection in mammograms. Pattern Recognit 2018;78:12–22.
[32] Chudzik P, Majumdar S, Calivá F, Al-Diri B, Hunter A. Microaneurysm detection using fully convolutional neural networks. Comput Methods Programs Biomed 2018;158:185–92.
[33] Mordang J-J, Janssen T, Bria A, Kooi T, Gubern-Mérida A, Karssemeijer N. Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks. International workshop on digital mammography 2016:35–42.
[34] Bria A, Marrocco C, Molinara M, Savelli B, Mordang J-J, Karssemeijer N, et al. Improving the automated detection of calcifications by combining deep cascades and deep convolutional nets. Proc. SPIE, vol. 10718 2018:10718. 10718-8.
[35] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998;86(11):2278–324.
[36] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. Proc. of int. conf. art. int. and stat. 2010:249–56.
[37] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(1):1929–58.
[38] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014arXiv preprint arXiv:1409.1556.
[39] American Cancer Society. Cancer facts & figures, 2016. 2016.
[40] Stomper PC, Geradts J, Edge SB, Levine EG. Mammographic predictors of the presence and size of invasive carcinomas associated with malignant microcalcification lesions without a mass. Am J Roentgenol 2003;181(6):1679–84.
[41] Wang J, Yang Y, Nishikawa RM. Reduction of false positive detection in clustered microcalcifications. 20th int. conf. on image processing. 2013. p. 1433–7.
[42] Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. Inbreast: toward a full-field digital mammographic database. Acad Radiol

2012;19(2):236–48.
[43] Decencière E, Cazuguel G, Zhang X, Thibault G, Klein J-C, Meyer F, et al. Teleophta: machine learning and image processing methods for teleophthalmology. IRBM 2013;34(2):196–203.
[44] LeCun YA, Bottou L, Orr GB, Müller K-R. Efficient backprop. Neural networks: tricks of the trade. Springer; 2012. p. 9–48.
[45] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, et al. Caffe: convolutional architecture for fast feature embedding. Proc. of the 22nd int. conf. on multimedia 2014:675–8.
[46] Bria A, Marrocco C, Molinara M, Tortorella F. An effective learning strategy for cascaded object detection. Inf Sci 2016;340–341:17–26.
[47] Samuelson F, Petrick N. Comparing image detection algorithms using resampling. Int. symp. biomed. imag. 2006. p. 1312–5.
[48] Dunn OJ. Multiple comparisons among means. J Am Stat Assoc 1961;56(293):52–64.
[49] Niemeijer M, Van Ginneken B, Cree MJ, et al. Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. IEEE Trans Med Imag 2009;29(1):185–95.