Computers in Biology and Medicine 120 (2020) 103735

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Addressing class imbalance in deep learning for small lesion detection on medical images☆

Alessandro Bria [a,*], Claudio Marrocco [a], Francesco Tortorella [b]

[a] Department of Electrical and Information Engineering, University of Cassino and Southern Latium, Cassino, FR 03043, Italy
[b] Department of Information and Electrical Engineering and Applied Mathematics, University of Salerno, Fisciano, SA 84084, Italy

ARTICLE INFO

Keywords:
Deep learning
Class imbalance
Lesion detection
Microaneurysms
Microcalcifications

ABSTRACT

Deep learning methods utilizing Convolutional Neural Networks (CNNs) have led to dramatic advances in automated understanding of medical images. However, in many medical image classification tasks, lesions occupy only a few pixels of the image. This results in a significant class imbalance between lesion and background. From recent literature, it is known that class imbalance may negatively affect the performance of CNN classification. However, very few research exists in the context of lesion detection. In this work, we propose a two-stage deep learning framework able to deal with the high class imbalance encountered during training of small lesion detectors. First, we train a deep cascade (DC) of long sequences of decision trees with an algorithm designed to handle unbalanced data that also drastically reduces the number of background samples reaching the final stage. The remaining samples are fed to a CNN, whose training benefits from both rebalance and hard mining done by the DC. We evaluated DC-CNN on two severely unbalanced classification problems: microcalcification detection and microaneurysm detection. In both cases, DC-CNN outperformed the CNNs trained with commonly used methods for addressing class imbalance such as oversampling, undersampling, hard mining, cost sensitive learning, and one-class classification. The DC-CNN was also ~10x faster than CNN at test time.

1. Introduction

Deep learning methods, and in particular Convolutional Neural Networks (CNNs) [1], have recently surpassed human capability on longstanding challenges ranging from recognizing the objects in an image [2], to mastering the game of Go [3]. In contrast to traditional machine learning techniques that use handcrafted features, CNNs automatically learn increasingly higher-level features from the data, while performing dimensionality reduction. These pivotal developments have led to marked advances also in medical image analysis, reaching expert-level performance in some areas [4–8]. However, there are a number of open challenges in this field, for example in the area of lesion detection. This task consists of the localization of small lesions in the full image space. It is a crucial step for diagnosis and requires intensive effort by clinicians. Typically, CNNs are applied for pixel classification, then post processing is used to obtain lesion candidates [9]. This approach raises three major issues:

1. *High class imbalance*. Since lesions occupy a few pixels of the image, there is a significant imbalance between lesion and background class. Recent literature indicates that class imbalance can have a detrimental effect on CNN classification performance [10,11].
2. *Uneven sample informativeness*. Regular patterns that characterize normal tissues lead to a high correlation between background samples. Treating uniformly the background samples makes the learning process unnecessarily time-consuming and prevents it to focus on the challenging samples [12].
3. *Inefficient processing*. At test time, pixel-wise classification results in orders of magnitude of redundant calculation, and makes the processing of large images undesirably inefficient [13].

Various solutions have been proposed in the literature to face these challenges, as will be presented in Section 2. However, very limited research is available in the area of lesion detection [9]. In addition, these problems are exacerbated when the lesions are very small, like microaneurisms in retinal images and microcalcifications in mammograms that have typical diameters of less than 20 pixels [14,15]. In such

☆ This work was supported by MIUR, Italy (Minister for Education, University and Research, Law 232/216, Department of Excellence). The authors gratefully acknowledge the support of NVIDIA Corporation for the donation of the Titan X Pascal GPUs used for this research.

* Corresponding author.
E-mail addresses: a.bria@unicas.it (A. Bria), c.marrocco@unicas.it (C. Marrocco), ftortorella@unisa.it (F. Tortorella).

https://doi.org/10.1016/j.compbiomed.2020.103735
Received 12 February 2020; Received in revised form 24 March 2020; Accepted 26 March 2020
Available online 1 April 2020

cases, the class imbalance may exceed 1:10 000 [16], which is orders of magnitude higher than in other applications [17].

In this paper, we provide two main contributions. First, we give an experimental evaluation of the existing methods applied on small lesion detection, especially for what concerns how to handle the high class imbalance. Second, we propose a two-stage deep learning approach that deals with all the aforementioned issues. The first stage is a deep cascade (DC) of long sequences of one-level decision trees (decision stumps), trained with a boosting algorithm designed to handle unbalanced data. It is constructed to discard the majority of background samples, while leaving the lesion samples practically unchanged. The remaining samples are fed to a CNN, which learns by focusing on the most informative background configurations identified by the DC, and deals with the residual imbalance by using simple data augmentations like flipping and rotations. At test phase, the DC is applied pixel-wise with an algorithm that eliminates redundant feature computations, allowing subsecond per-image processing time. Then, the CNN is applied on the positive findings of the DC, which constitute only a small fraction of the pixels of the original image. As a result, the combined DC-CNN approach is also a computationally efficient solution for pixel-wise deep learning classification.

This work extends our previous conference submission [18] in several directions: (i) the number of DC stages, that was fixed a priori, is automatically determined based on the target class imbalance verified on a validation set; (ii) the number of decision stumps in each DC stage, that was fixed to a maximum of 2000, is replaced by a weak stopping rule that checks whether the false positive rate has improved in the last 100 iterations; (iii) at test phase, an algorithm for sharing feature computations across the pixels of the image is described to achieve efficient pixel-wise classification; (iv) the method is tested on two publicly available datasets relative to two kinds of heavily unbalanced medical classification problems, microcalcification detection and microaneurysm detection; and (v) we provide an extensive experimental comparison of different existing methods for addressing class imbalance with CNNs.

The rest of the paper is organized as follows. In Section 2, we review the literature on class imbalance and related issues in the field of deep learning in medical imaging, and identify the methods that can be applied for small lesion detection. Section 3 describes the data sets used in this work. The proposed method is explained in Section 4, and experiments are detailed in Section 5. Results are shown in Section 6 and discussed in Section 7. Section 8 concludes the proposed work.

2. Related work

2.1. General methods

Methods for addressing class imbalance in deep learning can be divided into two categories: (i) data level methods that operate on the training set; and (ii) classifier level methods that adjust training or inference algorithms. In the following, we provide a brief overview of these methods. For a more comprehensive review, the reader can refer to [11].

Data level methods. One of the most successful data level methods is oversampling [19]. Its basic version consists of replicating randomly selected samples from the minority class. An alternative is to generate new samples from the existing data, a process known as data augmentation. This can be done in the image space by using geometrical and intensity transformations [20], or in the feature space to create artificial samples [21]. As opposed to oversampling, undersampling consists of randomly removing samples from the majority class. A variant of this method is to more carefully select the samples to be removed, for instance by applying hard mining [22], or to combine different classifiers trained on different subsets of the overrepresented class [23].

Classifier level methods. A commonly used classifier level method is cost sensitive learning [24]. It consists of assigning different costs to misclassification of samples from different classes, thereby mitigating the effect of class imbalance. This method can be implemented at training time by, for example, modifying the loss function with different error penalties [25], or by introducing weights inversely proportional to the class counts [26], or at test time by scaling or thresholding the outputs [27]. Another option is to turn classification into a novelty detection problem by training autoencoders to perform autoassociative mapping on one class [28]. At test time, the classification is performed based on the reconstruction error, which is expected to be higher on the unseen classes.

2.2. Methods in medical imaging

Few papers directly address the issues related to class imbalance in the area of deep learning for medical imaging. At the data level, a widely used approach is to apply simple data augmentation techniques (e.g. flipping, rotation) to the underrepresented classes [11]. Synthetic data augmentation with Generative Adversarial Networks (GANs) was also investigated [31], and compared favorably with classical data augmentation methods. At the classifier level, a common solution is to use custom loss functions that give more weight to the instances of the minority class [32,45,46], or the Dice loss proposed in the seminal V-Net paper [29] that is insensitive to class imbalance but requires pixel-level groundtruths. Alternatively, class imbalance can be eliminated by training convolutional autoencoders on the background class and thresholding the reconstruction error at test time to perform classification [43]. Other methods address class imbalance and uneven sample informativeness at the same time. For example, [41] applied a hard mining strategy by selecting the most difficult background samples with a CNN, in order to train another CNN with a less unbalanced dataset and more informative samples. A similar approach was proposed in [12], where the hard mining is performed at each epoch by the CNN under training. As to inefficient processing at test time, Fully Convolutional Networks (FCNs) can drastically reduce the calculations needed for pixel-wise classification [47] and thus are widely used especially in the area of lesion segmentation [13,29,44,46]. However, they require pixel-level groundtruths that may not be available when lesions are small and numerous. For instance, the vast majority of the individual microcalcifications annotated in the public mammogram dataset INbreast [48] as well in the private mammogram datasets used in the literature [16,42,49,50] have only center locations, which prevents from using FCN-based approaches like U-net [45].

In Table 1 we provide an overview of existing methods which readers can use to quickly assess the field. Although all the major issues related to class imbalance have been tackled individually, there is very few research addressing them as a whole, especially in the area of small lesion detection (last four rows of the table).

2.3. Methods compared in this study

In total, we examine seven methods to address class imbalance in deep learning for small lesion detection:

1. oversampling of the lesion class by replication
2. oversampling of the lesion class by data augmentation
3. undersampling of the background class
4. combined oversampling and undersampling
5. hard mining (2-phase training) on the background class
6. one-class learning on the background class
7. cost-sensitive learning (weighted cross entropy loss)

These selected methods are representative of the available approaches previously discussed. Those that do not appear in the above list have been discarded since they are hardly applicable to very small and low-contrasted lesions like microcalcifications and microaneurysms (GANs)

Table 1
Overview of papers addressing class-imbalance and related issues in deep learning for medical imaging.

Ref.	Application/Modality	Input size	Issues faced	Remarks
[29]	Prostate segmentation Magnetic Resonance Imaging	$128 \times 128 \times 64$	Class imbalance, inefficient processing	Proposes the Dice loss based on the Dice coefficient which is insensitive to class imbalance; uses FCNs to avoid redundant computations.
[30]	Glioblastoma cell segmentation Confocal Spinning Disk Microscopy	Not reported	Class imbalance	Uses focal cross-entropy loss to perform hard negative mining.
[31]	Liver lesion classification Computed Tomography	64×64	Class imbalance	Applies synthetic data augmentation to the lesion class utilizing Generative Adversarial Networks.
[32]	Eye diseases diagnosis Retroillumination photography	128×128	Class imbalance	Includes misclassification costs in the loss function to train a cost-sensitive residual CNN (CS-ResCNN).
[33]	Lung nodules classification Computed Tomography	64×64	Class imbalance	Augments underrepresented classes by oversampling patches from different view planes.
[34]	Lung nodules classification Computed Tomography	64×64	Class imbalance	Applies data augmentation (rescaling with random view selection).
[35]	Lung nodules segmentation Computed Tomography	$3 \times 35 \times 35$	Class imbalance	Uses a weighted sampling method to select only the challenging voxels.
[36]	Lung nodules detection Computed Tomography	128×128	Class imbalance, inefficient processing	Applies data augmentation (flipping and translation); the candidate detection is based on Faster R-CNN.
[37]	Brain lesion segmentation Magnetic Resonance Imaging	$193 \times 229 \times 193$	Class imbalance	Builds training batches with equal sampling probability for each class.
[38]	Breast/prostate cancer grading Histopathology Images	128×128	Class imbalance	Applies data augmentation (flipping and rotation).
[39]	Brain tumor segmentation Magnetic Resonance Imaging	$4 \times 33 \times 33$, $4 \times 65 \times 65$	Class imbalance	Trains the CNN on a balanced data set and then fine-tunes the output layers; applies data augmentation (flipping).
[40]	Brain tumor segmentation Magnetic Resonance Imaging	$4 \times 33 \times 33$	Class imbalance	Applies data augmentation (intensity transforms) which compared favorably with minority class undersampling.
[13]	Coronary artery calcium segmentation CT Angiography	$15 \times 15 \times 15$, $25 \times 25 \times 25$	Class imbalance, inefficient processing	Trains paired FCNs with equal sampling probability for each class.
[41]	Microcalcification detection Digital Mammography	13×13	Class imbalance, uneven sample informativeness	Trains a CNN with the most challenging background samples selected by another CNN; applies data augmentation (flipping and rotation).
[42]	Microcalcification detection Digital Mammography	$95 \times 95, 9 \times 9$	Class imbalance	Trains a multiscale CNNs with equal sampling probability for each class in the first half of training.
[43]	Microcalcification detection Digital Mammography	128×128	Class imbalance	Trains a convolutional autoencoder on the normal tissue class, and uses the reconstruction error at test time to discriminate between abnormal and normal tissue.
[44]	Microaneurysms detection Color Fundus Photography	32×32	Class imbalance, inefficient processing	Uses FCNs trained with Dice loss function; applies data augmentation (flipping and rotation).

or require pixel-level groundtruths (FCNs, U-Net, Dice loss) that are not available in our case (see Section 2.2) or have limited effectiveness with extreme class imbalance (Dice loss) [51].

3. Materials

3.1. Microcalcification dataset

We used the INbreast database [48] consisting of 115 cases and 410 full-field digital mammograms. The acquired images have dimensions ranging from 3328×4084 to 2560×3328 pixels, with a pixel size of 70 μm. A total of 6880 individual calcifications annotated by expert radiologists were available in 305 images, whereas the remaining 105 images did not contain any calcification annotation.

For our study, we extracted patches of size 14×14 pixels, corresponding to 1×1 mm, so that each microcalcification was fully contained in one patch [15]. For each annotated microcalcification center, we extracted one patch centered on it, obtaining 5759 positive samples. After having discarded all calcification regions, including annotated clusters, we extracted background patches from all the regions remaining in the images, totaling 74,315,102 negative samples. The resulting class imbalance was 1:12,904.

3.2. Microaneurysm dataset

We used the publicly available e-ophtha database [52] consisting of 381 digital color fundus images. The acquired images have dimensions ranging from 1440×960 to 2544×1696 with a pixel size of 7 μm. A total of 1306 microaneurysms manually annotated by expert ophthalmologists were available in 148 images, whereas the remaining 233 images did not contain any microaneurysm annotation.

For our study, we extracted the green channel from all the images since microaneurysms appear with high contrast in this channel [53,54]. Also in this case, we extracted patches of size 170×170 μm, corresponding to 24×24 pixels, which is large enough to fully contain individual microaneurysms that have size from 25 to 100 μm in diameter [55]. For each annotated microaneurysm, we extracted one patch centered on it, obtaining 1306 positive samples. After having discarded all microaneurysms regions, we extracted background patches from all the regions remaining in the images, totaling 15,722,707 negative samples. The resulting class imbalance was 1:12,039.

4. DC-CNN

The proposed method consists of a two-stage binary classifier, a high-sensitivity Deep Cascade (DC) followed by a Convolutional Neural Network (CNN).

The DC [56] was originally designed to handle high (e.g., 1:10^4) class imbalances at training time and to be very efficient (<1 s/Mpixel on CPU) at test time by early discarding the majority of easy detectable background samples. On the other hand, Convolutional Neural Networks have in some cases showed superior detection performance at the expense of a higher processing time [18] and difficulty in handling the high class imbalance during training [41].

The advantage of the combined DC-CNN approach is threefold: (i) at training time, the DC handles most of the class imbalance which is reduced by different orders of magnitude, so that the subsequent CNN can deal with the residual imbalance by using simple data augmentation techniques; (ii) the DC performs hard mining on the negative samples, so that the subsequent CNN can focus on the most challenging background configurations where it is able to outperform the DC; (iii) at test time, the DC greatly reduces the computation required to process the entire image, since the majority of background samples are rejected early in the process, and are not processed by the CNN.

The overall system is depicted in Fig. 1. An example result obtained on mammography and retinal images is shown in Fig. 2. Methodological details are given in the following subsections.

4.1. Deep cascade

The DC is a sequence of n classifiers $\{H_i(\mathbf{x})\}_{i=1,\dots,n}$ where each sample $\mathbf{x}$ passes to the next classifier only if the current one classifies it as positive according to a high-sensitivity decision threshold Θ_i. Each $H_i(\mathbf{x})$ is a linear combination of base learners $h_{i,t}(\mathbf{x}) \in \{0,1\}$ (0 for negative, 1 for positive) weighted by $\alpha_{i,t} \in \mathbb{R}$:

$$H_i(\mathbf{x}) = \sum_t \alpha_{i,t} h_{i,t}(\mathbf{x}). \tag{1}$$

A base learner $h_{i,t}(\mathbf{x})$ is a decision stump that compares a single Haar feature $\varphi_{i,t}(\mathbf{x})$ with a threshold $\theta_{i,t}$:

$$h_{i,t}(\mathbf{x}) = \begin{cases} 1 & \text{if } \varphi_{i,t}(\mathbf{x}) > \theta_{i,t} \\ 0 & \text{if } \varphi_{i,t}(\mathbf{x}) \le \theta_{i,t}. \end{cases} \tag{2}$$

At each training round t, the choice of $\varphi_{i,t}$, $\theta_{i,t}$ and $\alpha_{i,t}$ is made to maximize the Area Under the ROC Curve (AUC) of $H_i(\mathbf{x})$ on the training set. This objective function is insensitive to the class skew and thus is a good choice when learning from unbalanced data sets [57]. When a new decision stump is added, the decision threshold Θ_i is tuned on a validation set to achieve exactly the desired high sensitivity s_i, at the cost of a false positive rate f_i'. If $f_i' \le f_i$ with f_i being the desired false positive rate, no more decision stumps are added. As a consequence, the overall sensitivity S and false positive rate F of the cascade are:

$$S = \prod_{i=1}^{n} s_i \quad F = \prod_{i=1}^{n} f_i'. \tag{3}$$

For a more comprehensive description of the DC, including the complete pseudocode of the learning procedure, the reader might refer to [56,58,59]. In the following subsections, we detail the pivotal contributions that play an important role for the combined DC-CNN system.

4.1.1. Hard mining strategy

Training data are organized in a training set $\mathcal{T}$, a validation set $\mathcal{V}$, and a large pool of negative samples $\mathcal{N}$. When $H_i(\mathbf{x})$ is trained, it is evaluated on all these sets to discard the true negatives, i.e. 'easy' negative samples that the current classifier is able to recognize and that will not be used to train the subsequent classifiers in the cascade. $\mathcal{T}$ and $\mathcal{V}$ are then refilled by randomly sampling from $\mathcal{N}$. As a result, the negative samples on which the ith classifier is trained and validated become more 'hard' as the training proceeds with the addition of new classifiers to the cascade. When the cascade is trained, the last classifier $H_n(\mathbf{x})$ is applied on $\mathcal{T}$, $\mathcal{V}$, and $\mathcal{N}$ to identify the false positives. These samples together form the 'hard' negative sample set used to train the subsequent CNN.

4.1.2. Class imbalance reduction

From Eq. (3) it follows that with just $n = 10$ classifiers having $s_i = 0.999$ and $f_i' \le f_i = 0.5\ \forall i$, one could obtain an overall sensitivity of $S = s_i^{10} = 0.99$ and a false positive rate of $F \le f_i^{10} = 10^{-3}$. This alone would be enough for the purpose of combining the DC with the CNN: at training time, it would correspond to a reduction of class imbalance of about three orders of magnitude, whereas at test time it would allow to discard $1 - F = 99.9\%$ of the background pixels in the image.

However, three key observations must be made. First, because of the hard mining strategy implemented during training of the DC, it may not be possible to meet the learning goals s_i and f_i for the last classifiers that are trained with more 'hard' negative samples. In these cases, a stopping rule at the individual classifier level is needed (see Section 4.1.3).

Second, there exist a trade-off between f_i and s_i. The lower the f_i (i.e., the higher the specificity), the lower must be the sensitivity s_i to allow the ith classifier to meet both the learning goals during training. Thus, in practice, one would choose $s_i = 0.995$ or even $s_i = 0.990$. In such case, the overall sensitivity with $n = 10$ classifiers would be $S = 0.90$ meaning that 10% of the positive samples are not passed to the CNN. So n should be small enough to avoid missing too many positive samples early in the process.

Third, if the positive class contains few samples (e.g. hundreds as in the case of microaneurism detection), obtaining a perfectly rebalanced training set would mean to train the subsequent CNN also with a few samples of the negative class, which may be underrepresented and limit the final classification performance of the CNN.

For all the aforementioned reasons, we introduced an early stopping rule at the cascade level given by[1]:

$$CascadeStop = \left[\frac{N \times \prod_{\tau=1}^{n} f_\tau'}{P \times \prod_{\tau=1}^{i} s_\tau} \le CI \right] \tag{4}$$

where P and N are the number of all positive and negative samples in the training data, and CI is the desired class imbalance. Differently from [18] where n was fixed a priori based on the learning goal f_i, here n is automatically regulated based on the actual false positive rates f_i' achieved by the individual classifiers on the validation set and the desired class imbalance CI. We chose $CI = 10^2$ to have a sufficiently represented negative class for training the CNN. This choice also results in a shorter cascade compared to lower class imbalances, so that the overall sensitivity remains high.

4.1.3. Weak stopping rule

At the individual classifier level, and in particular for the last classifiers in the cascade that are trained with more 'hard' negative samples, achieving $f_i' \le f_i$ might be difficult and require a high number of base learners. For this reason, in our previous work [56] we established a maximum number of iterations $t_{max} = 2000$ high enough to minimize f_i' at the cost of longer training times and base learner sequences (hence the term 'deep' cascade). However, we experimentally observed that the minimum of f_i' usually occurred far from t_{max}, whereas in the other cases there was still room for improvement after t_{max} iterations. Thus, we replaced the hard stopping rule

$$ClassifierStop = [t \ge t_{max}] \tag{5}$$

with a weak stopping rule given by:

$$ClassifierStop = \left[t \ge 100 + \arg\min_t f_i'(t) \right]. \tag{6}$$

That is, we stop training the ith classifier if the false positive rate has not improved in the last 100 iterations. In such cases, we revert to the training round t for which f_i' was minimum. Thanks to this rule, less

[1] The notation [pr] (Iverson bracket) is defined to be 1 if predicate pr holds and 0 otherwise.

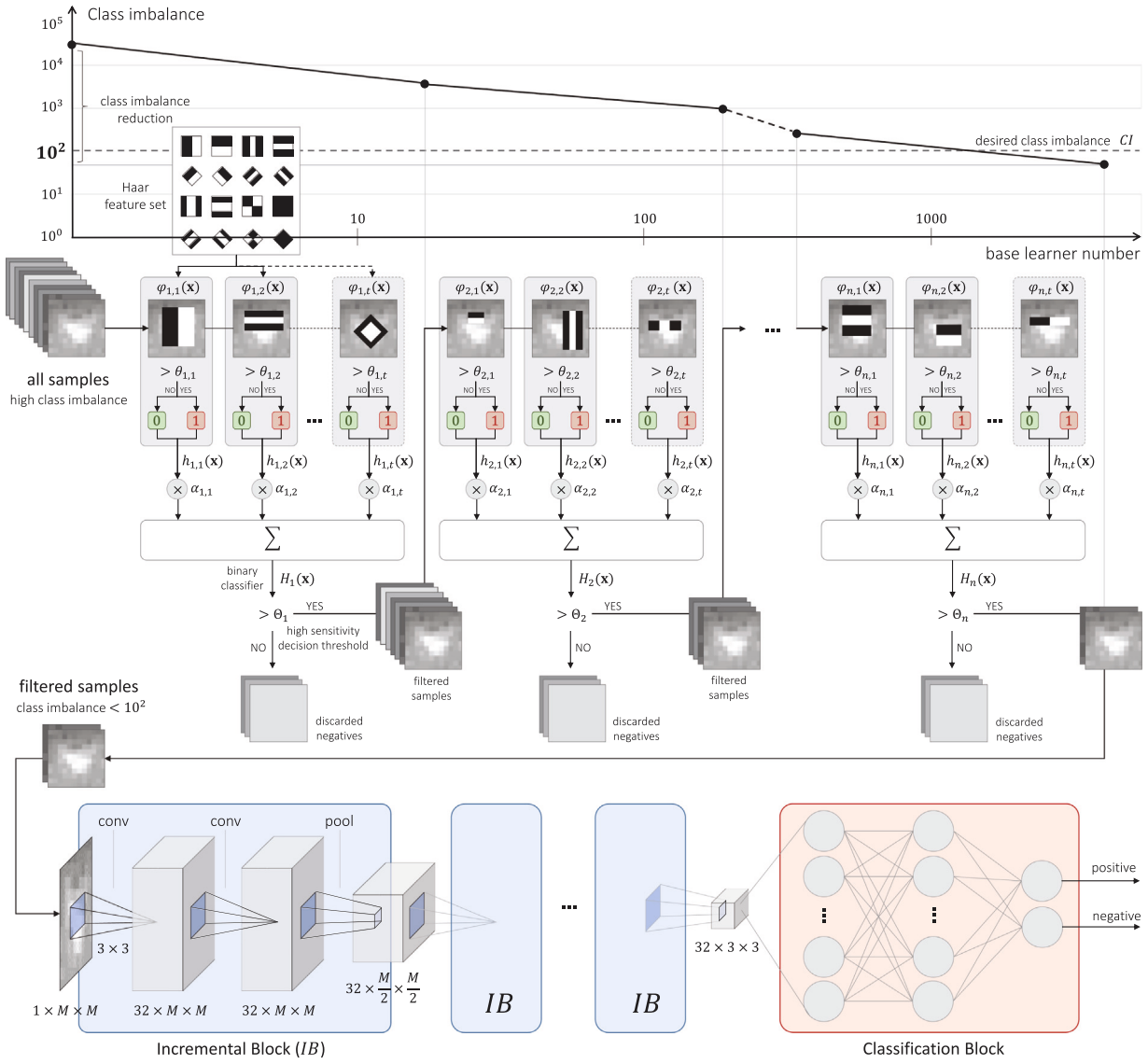

Fig. 1. DC-CNN system composed by Deep Cascade (top row) and Convolutional Neural Network (bottom row).

training computation is spent when the task becomes too difficult, and more when f_i' can still improve by adding more base learners.

4.1.4. Efficient pixel-wise classification

Pixel-wise classification at test time is based on a sliding window that scans the entire image. Thanks to the cascade approach, the majority of background windows are rejected early in the process. For a given window $\mathbf{x}$ that is being tested by the ith classifier, the computational bottleneck is the computation of the Haar features $\{\varphi_{i,t}(\mathbf{x})\}_{t=1,\dots}$ needed to evaluate $H_i(\mathbf{x})$. Following the implementation guidelines from [60], it is possible to calculate one Haar feature in $O(1)$ (constant) time by combining six to eight array references using the Summed-Area Table (SAT) defined as:

$$SAT(x,y) = \sum_{\substack{x' \le x \\ y' \le y}} I(x',y') \tag{7}$$

where $I(x,y)$ is the value of the pixel at (x,y). Due to the overlapping between sliding windows, a number of redundant calculations would be performed if the SAT is computed for each window separately. Instead, we calculate the SAT for the entire image, and correct the array references with the displacement of the window. In this way, the SAT is computed only once and it is shared among all feature computations for all the windows tested.

4.2. Convolutional neural network

For this study, we employed a CNN inspired by the VGGnet architecture [61]. The input of the CNN are the patches filtered by the DC described in the previous Section. Patches have size $M \times M$ that varies according to the lesion size (see Section 3). A schematic overview of the network architecture is provided in Fig. 1 (bottom row). It consists of a sequence of *incremental blocks* (see Table 2) followed by a *classification block* (see Table 3). An incremental block is defined by two convolutional layers and a max pooling layer, with ReLU as activation function. The number of incremental blocks is determined by the input patch size so as to have blobs of size 3×3 as input to the classification block. The classification block consists of three fully connected layers with dropout as regularization to prevent overfitting. The last fully connected layer has two output neurons with Softmax as activation function that generates a two-value probability vector associated to each prediction.

5. Experiments

We compared the microcalcification (MC) and microaneurysm (MA) detection performance of the proposed DC-CNN approach with the one

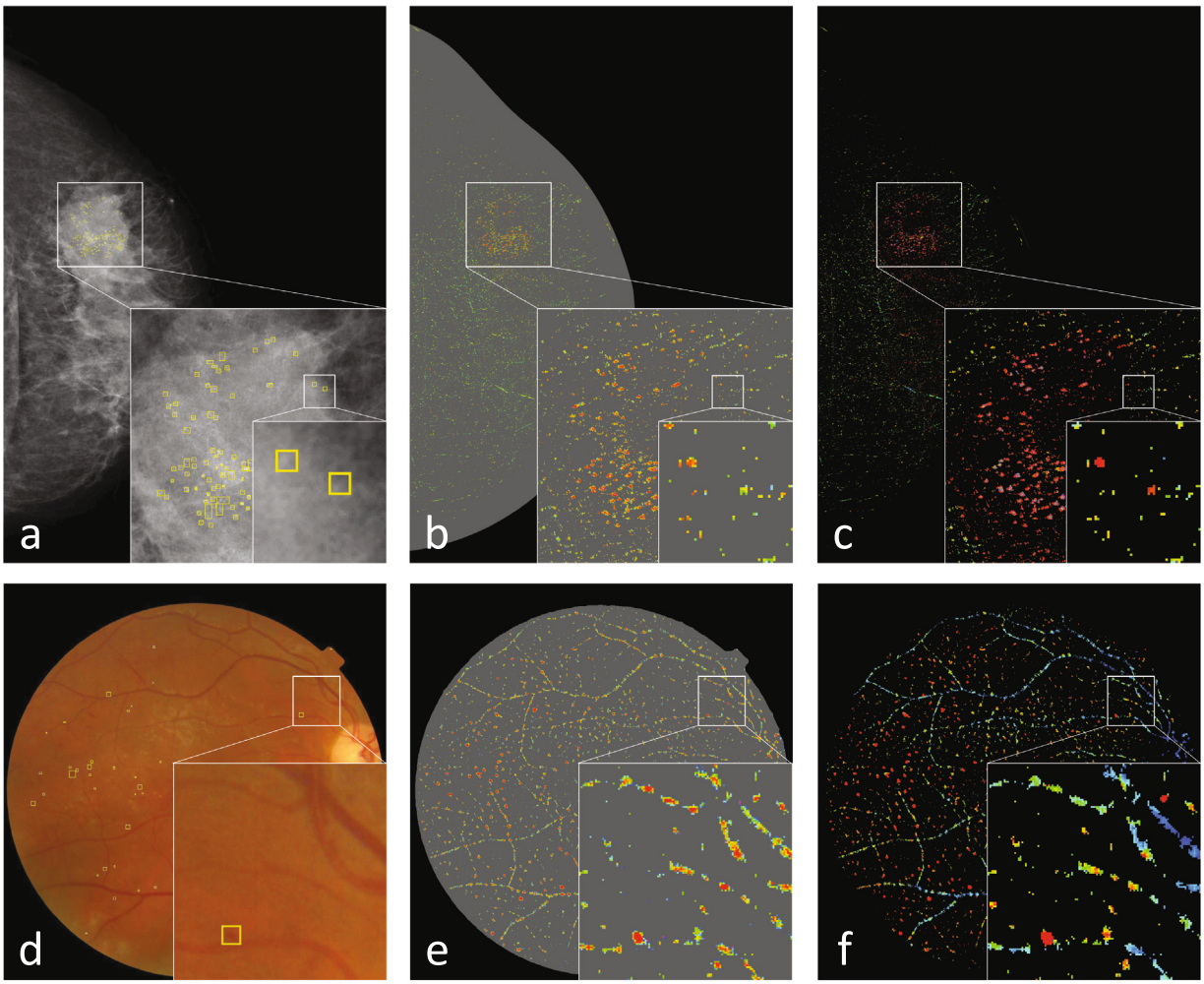

Fig. 2. Result of DC-CNN applied on a mammogram with suspicious microcalcifications (first row) and on a retinal image with microaneurisms (second row). First column: groundtruth annotations. Second and third columns: DC and DC-CNN color heatmap probability images (gray = rejected pixels, black = unprocessed pixels). On (b) and (e) it can be seen that DC rejects a vast majority of background pixels, that are not processed by the subsequent CNN. Moreover, the capability to discriminate lesions from normal tissue is higher in the DC-CNN output (see the bottom-right box in (c) and (f) compared to the one in (b) and (e)). Note that the final detection is obtained by applying the steps described in Section 5.5 to the DC-CNN outputs.

Table 2
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

Table 3
Details of the *classification block*.

Layer	Type	Output size	Kernel size	Dropout 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	

of a standalone DC and of several CNNs trained with different literature methods to address class imbalance. The full list of methods along with the implementation details are provided in Table 4. In all cases except one (CAE), the CNN architecture was the same used in our approach so as to allow a fair comparison.

5.1. Training and test sets

All lesion detectors were trained and tested on the image patches extracted from the full images (see Section 3). To reproduce the behavior of a full Computer Aided Detection (CAD) scheme, the patches were extracted pixel-wise from all the regions of the image except the easily segmentable background (the region outside the breast on mammograms, and the region outside the field of view on retinal images).

Case-based 2-fold cross validation was applied in all the experiments. Within each cross validation repetition, we trained the lesion detector on 50% of the cases, and tested it on the other 50%. Image patches belonging to the same case were always assigned to the same set. A total of 96 individual training/test experiments were performed, resulting from 24 methods to compare (21 from the literature plus CNN-Baseline, DC and DC-CNN), 2 datasets (MCs and MAs), and 2-fold cross validation.

5.2. Training hyperparameters

DC detectors were trained with sensitivity $s_i = 0.995$ and false positive rate $f_i = 0.3$ for each classifier, whereas for DC-CNN we used $s_i = 0.990$ and $f_i = 0.3$. In DC-CNN, the cascade was composed by a varying number of classifiers automatically regulated by the desired

Table 4
Methods compared in this study to address class imbalance in deep learning for small lesion detection.

Abbreviation	Method	Implementation	Submethods
CNN-Baseline	Baseline method	At training time, no resampling/augmentation methods are applied. The CNN architecture is the same described in Section 4.2.	CNN-Baseline
CNN-Bal	Baseline method with balanced classes	At training time, the lesion class is randomly oversampled so as to match the same number of background samples (1:1 class imbalance) using different data augmentation techniques like replication, flipping and 90-degrees rotation, affine transforms, and warping.	CNN-Bal-Replicate CNN-Bal-FlipRotate CNN-Bal-Affine CNN-Bal-Warp
CNN-Over-x	Oversampling of the lesion class	At training time, the lesion class is randomly oversampled[a] so as to have a class imbalance of 1:x.	CNN-Over-1 ≡ CNN-Bal CNN-Over-10 CNN-Over-100 CNN-Over-1000
CNN-Under-x	Undersampling of the background class	At training time, the background class is randomly undersampled so as to have a class imbalance of 1:x.	CNN-Under-10 CNN-Under-100 CNN-Under-1000
CNN-Under-x-Bal	Combined under/oversampling	At training time, the background class is randomly undersampled so as to have a class imbalance of 1:x. Then, the lesion class is randomly oversampled[a] so as to have a perfectly balanced training set.	CNN-Under-10-Bal CNN-Under-100-Bal CNN-Under-1000-Bal
CNN-...-Hard-x	Hard mining on the background class/two-phase training	Following the implementation guidelines from [41], a first CNN is trained on a perfectly balanced training set obtained by randomly undersampling the background class and by randomly oversampling[a] the lesion class. Then, this CNN is applied to select the top ranked x% background samples from the entire training set. The CNN is re-trained using these 'hard' negative samples together with all positives augmented[a] to have 1:1 class imbalance.	CNN-Bal-Hard-1 CNN-Bal-Hard-10 CNN-Under-10-Bal-Hard-1 CNN-Under-10-Bal-Hard-10 CNN-Under-100-Bal-Hard-1 CNN-Under-100-Bal-Hard-10
CAE	One-class learning on the background class/anomaly detection	Following the implementation guidelines from [43], we trained a Convolutional Autoencoder (CAE) on the background class, and at test time we calculated the root mean square error (RMSE) between the original and reconstructed image patch. The encoder consists of successive strided convolutional layers with decreasing patch size and increasing feature maps. Conversely, the decoder consists of successive upsampling and convolutional layers with increasing patch size and decreasing feature maps.	CAE
CNN-CS	Cost-sensitive learning	At training time, the cost-sensitive weighted cross-entropy loss is used with weights inversely proportional to the class counts [26]. No resampling/augmentation methods are applied.	CNN-CS

[a]Using the default data augmentation method (flipping and rotation).

class imbalance $CI = 10^2$ (see Section 4.1.2). In DC, more classifiers were added until all samples from the pool set were used [56].

All CNNs were trained using Stochastic Gradient Descent and network weights were updated in batches of 32 patches using the back-propagation algorithm. The base learning rate was 10^{-3} and followed a step decay policy with reductions by a factor of 10 every 6 epochs, and the learning was stopped after 30 epochs. Momentum and weight decay were 0.9 and 5×10^{-4}, respectively. The loss function was binary Cross Entropy, except for CNN-CS (weighted Cross Entropy) and CAE (Mean Squared Error).

5.3. Implementation

DC-CNN was implemented in C++ using the OpenCV library [62] and integrated within the deep learning framework Caffe [63]. In the DC, we used the integral image representation available in OpenCV which allowed to calculate filter responses in constant time [59]. As to the CNN, we implemented in Caffe a custom Memory Data Layer input layer that shared in memory the images processed by the DC. In this way, there was no need to store the patches nor the filter responses since the patches filtered by the DC were directly passed to the CNN by reference. For all the experiments, we used a workstation equipped with four Intel Xeon E5-4610 v2, 256 GB of RAM, and one NVIDIA Titan X Pascal GPU.

5.4. Partial ROC analysis

For each detector, the Receiver Operating Characteristics (ROC) curve was obtained by calculating the True Positive Rate (TPR) against the False Positive Rate (FPR) as a function of the classification threshold applied to the detector output associated to each patch sample. However, due to the high class imbalance present in our application, only the left part of the ROC curve is informative. In such case, the partial area under the ROC curve (pAUC) is considered a more practically relevant summary index than the area under the entire ROC curve (AUC) [64], and is defined as:

$$pAUC(a,b) = \int_a^b ROC(f)df \tag{8}$$

where a and b are the lower and upper bound of the FPR range considered, and $ROC(f)$ is the ROC value at the false positive fraction f. Previous works on lesion detection suggest to normalize the pAUC in the logarithmic scale of FPR to avoid that the index is dominated by ROC points at high false positive rates [41,65–68]. The resulting index, called mean sensitivity $\overline{S}$, is defined as:

$$\overline{S}(a,b) = \frac{1}{ln(b) - ln(a)} \int_a^b \frac{ROC(f)}{f} df \tag{9}$$

where the logarithmic normalization is performed by the denominator f within the integral and by the multiplying coefficient outside the integral.

In our study, we chose the FPR interval $[10^{-6}, 10^{-1}]$ corresponding to a wide range of operating points that can be used in a full CAD system [68]. The remaining region of FPRs $\geq 10^{-1}$ would yield more than 1 every 10 background pixels classified as lesion, resulting in a number of false positive lesions easily exceeding thousands per image, which would be not relevant for practical applications.

5.5. FROC analysis

To assess the performance of the proposed detector for CAD applications, we obtained lesion candidates from the probability images resulting from pixel-wise classification of the input images (see Fig. 2). We applied the following minimal post-processing to obtain generalizable results: (i) binarization by global thresholding; (ii) connected components extraction; (iii) removal of objects smaller than 2 pixels in either horizontal or vertical direction; and (iv) morphological dilation with a disk structuring element of radius of 5 pixels to merge neighboring objects. Then, we obtained the lesion-based free receiver operating characteristic (FROC) curve by calculating the True Positive Fraction (TPF) of the detected lesions versus the average number of False Positives per Image (FPpI) as a function of the threshold applied during post-processing. A candidate lesion was considered as a true positive if its distance to the center of a groundtruth lesion was no larger than its radius.[2] All regions detected on healthy images were counted as false positives. We considered as healthy all the images not containing any annotated lesion (see Section 3 for more details).

To analyze and compare FROC curves, we chose the non-parametric approach suggested in [69] for evaluating CAD algorithms and used in the field of MC and MA detection [14,70]. The figure-of-merit is the partial area under the FROC curve to the left of FPpI $= \gamma$ calculated by trapezoidal integration and denoted as AUFC_γ. To obtain an index in the range $[0, 1]$, we normalized AUFC_γ by dividing with γ. Following [69], we chose γ as large as possible and higher than 5 FPpI. Specifically, we chose $\gamma = 10$ FPpI for MA detection and $\gamma = 50$ FPpI for MC detection[3] as typically done in the literature of the respective fields [14,71].

6. Results

ROC curves for the literature methods examined are reported in Figs. 3–4 for MC and MA detection along with the corresponding values of $\overline{S}$[4] that are summarized in Table 5. The best methods to address class imbalance were oversampling of the lesion class (CNN-Over-x) and its combination with hard mining on the background class (CNN-Bal-Hard-x). Specifically, the top 2 performances were CNN-Over-10 ($\overline{S} = 77.38$) and CNN-Bal-Hard-1 ($\overline{S} = 77.68$) for MC detection, and CNN-Over-10 ($\overline{S} = 79.13$) and CNN-Bal-Hard-10 ($\overline{S} = 78.37$) for MA detection.

The performance of DC-CNN was compared to the one of DC and of the two best performing aforementioned CNNs. For testing the significance of observed differences in figures-of-merit $\overline{S}$ and AUFC_γ, we applied the bootstrap method [72] that is widely adopted for statistical comparison of CAD performances [16,41,64–66,70]. Cases were sampled with replacement 1000 times, with each bootstrap containing the same number of cases as the original set. At each bootstrapping iteration, ROC and FROC curves were recalculated, and differences in figures-of-merit $\Delta\overline{S}$ and ΔAUFC_γ between DC-CNN and each of the three compared methods were evaluated. Finally, the obtained ROC and FROC curves were averaged along the sensitivity axis, and p-values were computed as the fraction of $\Delta\overline{S}$ and ΔAUFC_γ populations that were negative or zero. Performance differences were considered significant if $p < 0.017$ that resulted from choosing the significance level $\alpha = 0.05$ and applying the Bonferroni correction for multiple comparisons.[5] [73] Average ROC and FROC curves are shown in Figs. 5–6, whereas performance differences are reported in Table 6. In all test cases, results with DC-CNN were statistically significantly better than with the methods compared. On average, the improvements in $\overline{S}$ were 2.43 over the CNNs and 7.00 over the DC, whereas the improvements in AUFC_γ were 7.74 over the CNNs and 19.44 over the DC. Another interesting result from the FROC analysis is that for both MC and MA detection CNN-Over-10 was the best standalone CNN. Yet, it was overperformed by DC-CNN with an improvement in AUFC_γ of 4.99 and 5.57 for MC and MA detection, respectively.

Table 5
Mean lesion detection sensitivity $\overline{S}$ of the methods compared in this study for MC and MA detection. The best two values for each detection task are highlighted in bold.

Method	Submethod	$\overline{S}_{MC}$	$\overline{S}_{MA}$
CNN-Baseline	CNN-Baseline	n.a.	n.a.
CNN-Bal	CNN-Bal-Replicate	75.23	76.43
	CNN-Bal-FlipRotate	76.84	77.26
	CNN-Bal-Affine	52.60	72.27
	CNN-Bal-Warp	52.05	73.72
CNN-Over-x	CNN-Over-1 ≡ CNN-Bal	76.84	77.26
	CNN-Over-10	**77.38**	**79.13**
	CNN-Over-100	76.02	76.83
	CNN-Over-1000	65.09	32.32
CNN-Under-x	CNN-Under-10	61.46	33.33
	CNN-Under-100	65.95	53.16
	CNN-Under-1000	65.33	20.73
CNN-Under-x-Bal	CNN-Under-10-Bal	61.20	54.30
	CNN-Under-100-Bal	68.37	73.32
	CNN-Under-1000-Bal	75.22	76.76
CNN-...-Hard-x	CNN-Bal-Hard-1	**77.68**	64.07
	CNN-Bal-Hard-10	76.51	**78.37**
	CNN-Under-10-Bal-Hard-1	61.35	31.68
	CNN-Under-10-Bal-Hard-10	72.71	43.00
	CNN-Under-100-Bal-Hard-1	76.51	78.23
	CNN-Under-100-Bal-Hard-10	75.81	77.62
CAE	CAE	10.59	37.96
CNN-CS	CNN-CS	73.55	70.26

For completeness, in Fig. 7 we report the average per-image processing time for the two datasets tested with DC-CNN and CNN,[6] with and without CPU–GPU parallelization. In all cases, DC-CNN was ~10x faster than CNN. Remarkably, the CPU–GPU parallelized version of DC-CNN took only 1.2 s and 1.4 s to process a single mammogram and a single retinal image. The key contribution to this result comes from the DC that quickly rejects most of the background pixels in the image, leaving only few pixels (25 K and 58 K on average for breast and retinal images) for the subsequent higher-complexity CNN-based classification.

7. Discussion

7.1. Methods compared in this study

Several observations can be made on the CNN methods compared in this study by analyzing the results in Table 5 and the corresponding ROC curves in Figs. 3–4:

1. *High class imbalances had a detrimental effect on CNN performance*. This can be seen by looking at the performances of CNN-Over-100 and CNN-Over-1000. In CNN-Baseline where the class imbalance was at a maximum (~10^4), the experiment failed (the loss never decreased during training) despite a number of attempts that were made with different settings of the training parameters. In this case, when the Cross Entropy loss was replaced with its weighted version (CNN-CS), the training was successful but the overall performance ranked only 9th for both MC and MA detection.

[2] We used unit radius when the lesion outline or radius was not available (e.g. this happened for 4339 of 6880 MCs in INbreast).

[3] Here we refer to *individual* MC detection that can serve to MC *clusters* detection, with MC clusters typically consisting of tenths of individual MCs.

[4] For better readability, throughout the manuscript we report $\overline{S}$ and AUFC_γ values in the range $[0, 100]$ instead than in $[0, 1]$.

[5] The significance level was α divided by the number of comparisons (3).

[6] Our CNNs differ in training procedure but not in architecture, so the computational complexity at test time is the same for all of them.

Table 6
Comparative results of $\overline{S}$ and $AUFC_\gamma$ obtained from 1000 bootstrap iterations. Statistically significant differences (p-value < 0.017) are listed in bold.

	Method	$\overline{S}$	$AUFC_\gamma$	Compared to	$\Delta\overline{S}$	$\Delta AUFC_\gamma$
MC detection	DC	73.38	57.69	–	–	–
	CNN-Over-10	77.38	74.79	–	–	–
	CNN-Bal-Hard-1	77.70	72.20	–	–	–
	DC-CNN	80.19	79.78	DC	**+6.81**	**+22.09**
				CNN-Over-10	**+2.81**	**+4.99**
				CNN-Bal-Hard-1	**+2.49**	**+7.58**
MA detection	DC	73.80	51.97	–	–	–
	CNN-Over-10	79.16	63.19	–	–	–
	CNN-Bal-Hard-10	78.41	55.96	–	–	–
	DC-CNN	80.99	68.76	DC	**+7.19**	**+16.79**
				CNN-Over-10	**+1.83**	**+5.57**
				CNN-Bal-Hard-10	**+2.58**	**+12.80**

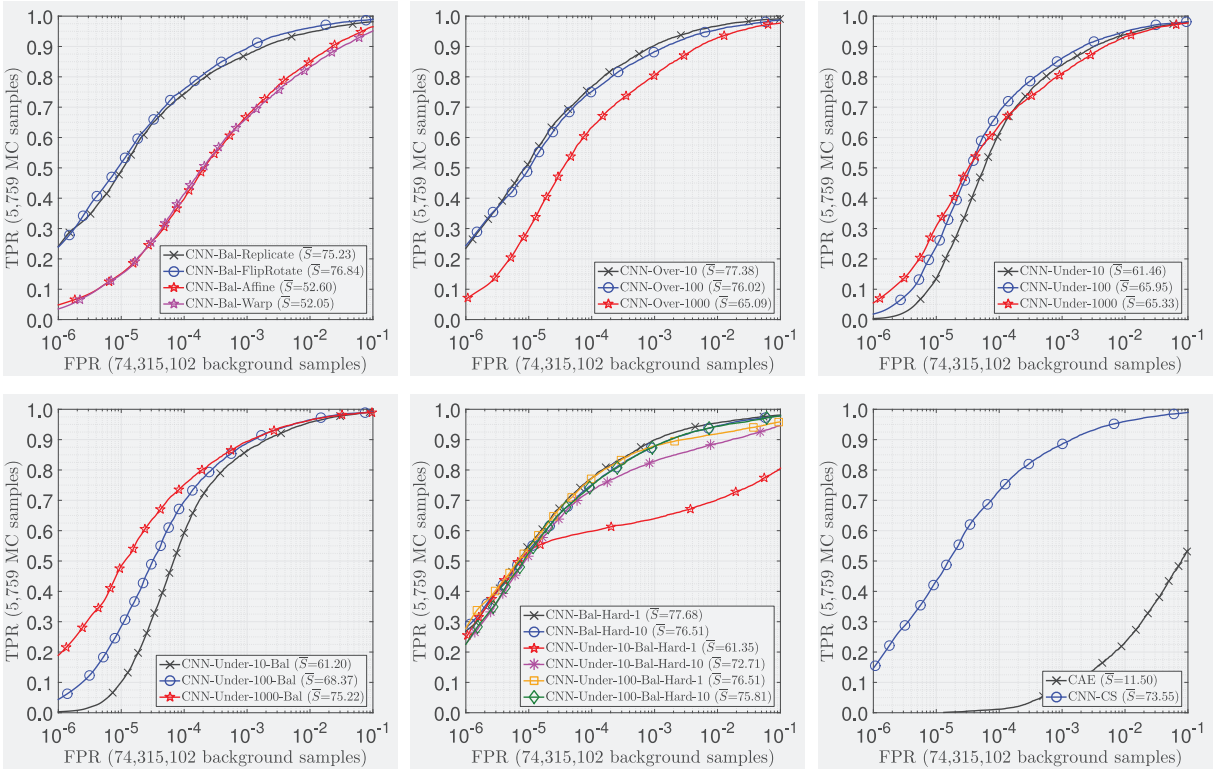

Fig. 3. ROC curves of MC detection with the methods compared in this study to address class imbalance for CNNs.

2. *The best methods to address class imbalance were oversampling of the lesion class and hard mining on the background class*. Specifically, the top 2 performances were CNN-Over-10 and CNN-Bal-Hard-1 for MC detection, and CNN-Over-10 and CNN-Bal-Hard-10 for MA detection. As to CNN-Over-10, it can be seen as a trade-off between the need to reduce the class imbalance and the need to avoid that data augmented samples dominate the original samples at training time (see next consideration). As to CNN-Bal-Hard-x, it is actually a combination of oversampling of the lesion class and hard mining on the background class, so it can benefit from both approaches by reducing the class imbalance and allowing the CNN to focus only on the top x% challenging background samples.
3. *The best data augmentation method was flipping and rotation*. Differently from affine and warping transforms that apply interpolation, this technique does not change the original pixel values but simply rearranges their position. Because of the high class imbalance, the vast majority of lesion samples in the balanced training set of CNN-Bal come from data augmentation. The more different these samples from the 'genuine' ones, the more biased is the CNN during training. This problem is exacerbated when the image patch is very small, as confirmed by the larger drop in performance when using affine and warping transforms.
4. *Undersampling the background class was not effective*. This can be seen by looking at the performances of CNN-Under-x, that were among the worst obtained. A possible explanation is that the background class actually consists of several subclasses (e.g. for mammograms: fat tissue, fibroglandular tissue, vasculature, muscle, etc.), and that a heavy undersampling (like CNN-Under-10-*) is unable to fully capture the heterogeneity of the background class and its subclasses. On the other hand, when the undersampling is less strong (e.g. CNN-Under-100-*) and is coupled with oversampling (CNN-Under-100-Bal) and hardmining (CNN-Under-100-Bal-Hard-x), it can be used to reduce the training time without sacrificing the detection performance.
5. *One-class learning on the background class was not effective*. We dedicated a considerable effort to this method, by experimenting with a number of variants of the CAE architecture proposed

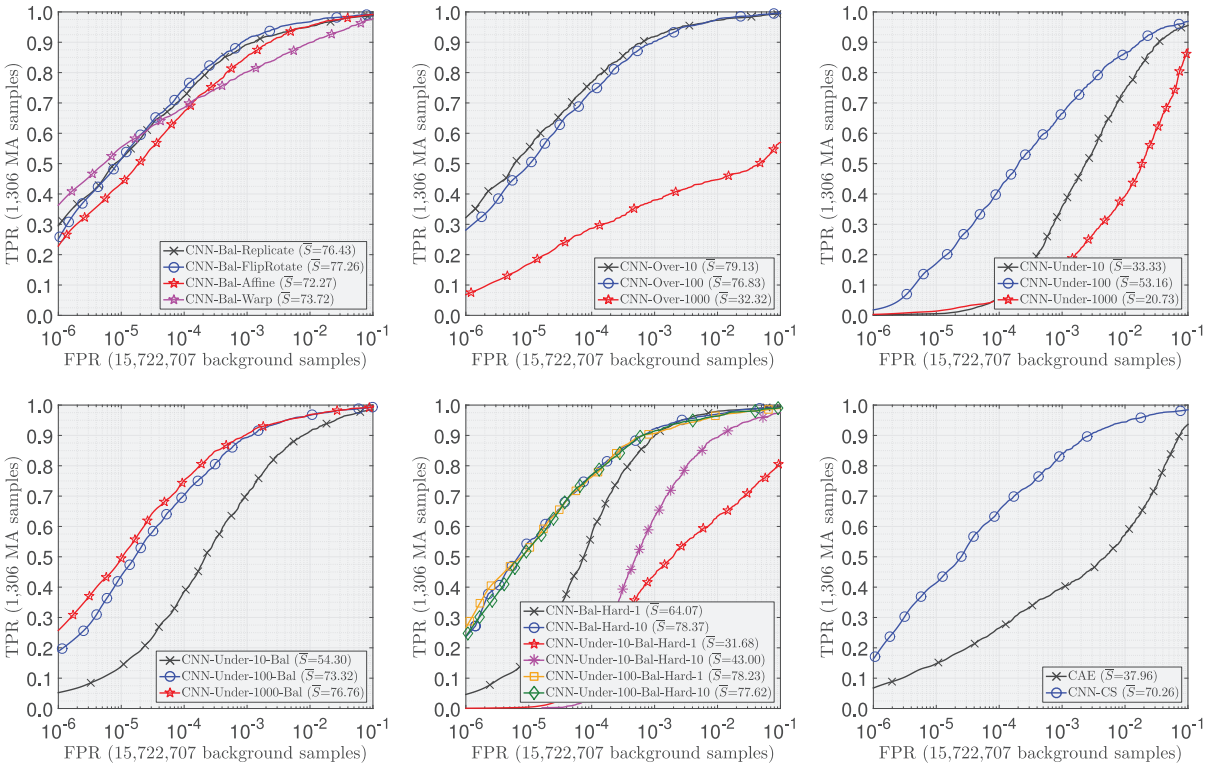

Fig. 4. ROC curves of MA with the methods compared in this study to address class imbalance for CNNs.

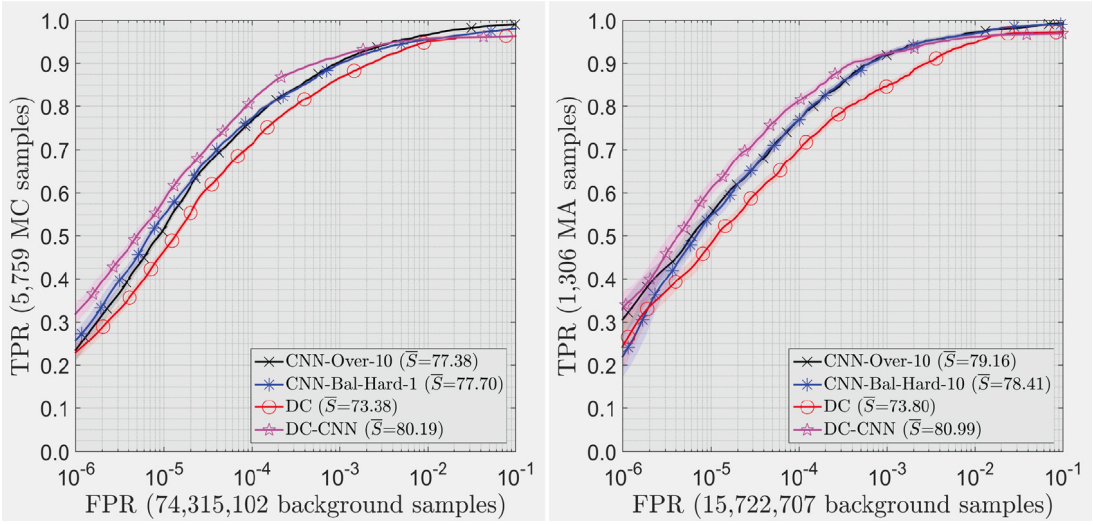

Fig. 5. Average ROC curves of MC detection (left) and MA detection (right) obtained from 1000 bootstrap iterations. Confidence bands (semi-transparent) indicate 95% confidence intervals along the TPR axis.

in [26]. The best experiments, that are the ones reported in Table 5, showed a good decay of the MSE loss and achieved an AUC of 0.82 and 0.87 for MC and MA detection, respectively, which are comparable to the AUC of 0.84 reported in [26] obtained on larger lesions (128 × 128 pixels). Despite this, CAE achieved the lowest $\overline{S}$ among the methods that we tested. We believe that in our case the lesions were too small for the reconstruction error to discriminate well between the two classes. This was especially true for MCs as confirmed by the lower $\overline{S}$ compared to MAs.

These observations, and in particular those in 1–3, confirm the findings from the literature regarding the effectiveness of oversampling [10,11] and hard mining [12,41].

7.2. DC-CNN performance

The improvements in performance of DC-CNN over the compared methods were remarkably large in all the test cases considered. From the ROC curves, it can be seen that this improvement comes mainly from FPRs < 10^{-4} that roughly correspond to the FPpI range of the respective FROC curves. As a consequence, the improvements in AUFC$_\gamma$ were higher than those in $\Delta\overline{S}$, proving the effectiveness of the proposed method for CAD applications. This low FPR/FPpI region is also where background samples become more challenging, so a key contribution comes from the second-stage of classification of the DC-CNN where the CNN is trained with hard negative samples that the DC cannot

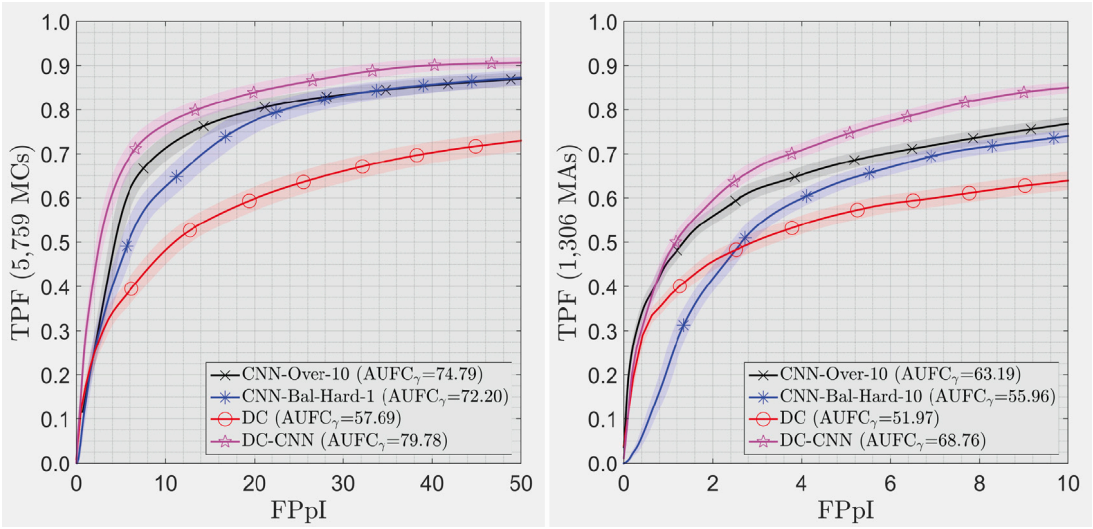

Fig. 6. Average FROC curves of MC detection (left) and MA detection (right) obtained from 1000 bootstrap iterations. Confidence bands (semi-transparent) indicate 95% confidence intervals along the TPF axis.

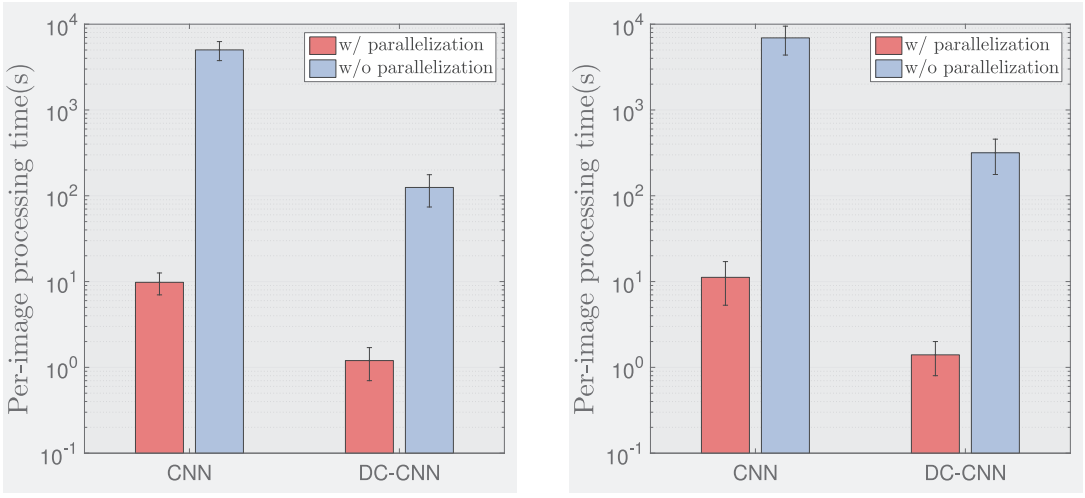

Fig. 7. Average per-image processing time for MC detection (left) and MA detection (right).

recognize. The FROC analysis also shows that the hard mining performed by the DC is more beneficial to the CNN than that performed by the CNN itself. In both MC and MA detection, CNN-Bal-Hard-x achieved the lowest AUFC$_\gamma$ among the CNN-based approaches, suggesting that the CNN was not effective as DC in selecting the most challenging background configurations.

7.3. Existing methods in the literature

7.3.1. Hessian-based candidate detection

Existing CAD methods and in particular those targeting small lesions like microcalcifications and microaneurysms often incorporate an initial candidate detection step based on the Hessian matrix [74–80]. These methods exploit the assumption that small lesions usually appear as circular-like regions with enough contrast with the background, hence they can be characterized by having strong derivatives along both orthogonal directions. The Hessian matrix is a square matrix of second-order partial derivatives. By calculating the Hessian matrix for each pixel neighborhood, it is possible to detect candidate regions where the two eigenvalues of the Hessian matrix λ_1 and λ_2 are high (contrasted object) and similar (circular shape). This approach is computationally efficient and is especially effective when lesions share similar characteristics with other structures in the image but have a different shape, like microaneurysms (circular) vs. blood vessels (linear) in retinal images.

While it would be theoretically possible to replace the first stage (DC) of our DC-CNN method with Hessian-based candidate detection, there are some drawbacks to this approach worth to be considered. First, it requires finding a trade-off between sensitivity and specificity by choosing appropriate thresholds for λ_1, λ_2, and λ_1/λ_2. In our DC, the overall sensitivity and specificity are automatically regulated based on the desired class imbalance (see Eq. (4)). Second, Hessian analysis may not be able to capture the heterogeneity of lesion shapes, such as microcalcifications that can be round, oval, linear, granular, or irregular. In contrast, DC does not make assumptions on lesion characteristics, but learns directly from data. As to the computational complexity of DC compared Hessian analysis, our DC is an adaptation of the original Cascade of Viola and Jones [60] designed for real-time face detection. This is the key factor that allowed us to obtain ~1 s per-image processing time.

7.3.2. Object detection methods

While our method is two-stage and focuses on very small lesions in medical images, several one-stage methods have been proposed in

recent literature in the more general field of object detection [81–84]. In particular, the state-of-the-art RetinaNet [84] consists of a Feature Pyramid Network (FPN) and on two subnetworks for classifying and regressing predefined bounding boxes called *anchor boxes*. This is an efficient and effective alternative to the sliding window approach and to two-stage methods that build on region proposals. However, the limited number and the size of anchor boxes (from 32^2 to 512^2 in Lin et al. [84]) might be an issue for very small objects that are only visible at the original image resolution mapped to the first layer of the FPN. Translated to the medical images domain, very small lesions like microcalcifications and microaneurysms that often have diameter less than 10 pixels (see Section 3 and [68]) may be difficult to detect with this method, and the multiscale approach of the FPN would be of limited benefit. In contrast, it could be useful when applied to detect larger lesions like mammography masses, as recently done in Jung et al. [85].

As to popular two-stage methods for object detection like Faster R-CNN [86] and Mask R-CNN [87], they build on a Region Proposal Network (RPN) that slides over the feature map outputted by a convolutional network like VGG-16 [61]. Since the input feature map is much smaller than the original image, this approach is computationally efficient, but again it might not work well with very small lesions that occupy only a few pixels, whereas it is effective for detecting larger lesions like mammography masses [88–91].

There are other aspects that should be taken into account when considering the aforementioned object detection methods in the context of small lesion detection. First, they are inherently designed to trade off detection accuracy with computational complexity, which is of great importance for real-time computer vision, but it is not a strict requirement in medical image analysis. Second, small lesions often do not have precise outlines, as opposed to objects in real images that can be easily contoured. This may lead to a high degree of uncertainty in medical image annotations, in some cases even hindering manual segmentation, for example microcalcifications (InBreast [48] and other private mammogram datasets [16,42,49,50] have mostly center annotations). This impedes using methods that require pixel-level groundtruths, like Mask R-CNN or, more in general, Fully Convolutional Networks (FCNs) or FNC-based approaches like U-net [45]. Third, objects in real images appear at multiple scales, thus object detection methods usually incorporate mechanisms like image or feature pyramids but these are not beneficial for small lesion detection.

7.4. Limitations and future work

Although our method achieves very promising results in small lesion detection, there are still several limitations to be considered. First, our method may not effectively deal with larger lesions, such as mammography masses and retinal hemorrhages. This is mainly due by the first stage (DC) that is based on Haar features. The larger the patch size, the exponentially higher the number of Haar feature instances, which makes feature selection less effective and computationally heavier. Other features, like Histogram of Oriented Gradients (HOG), could serve better in this context. Second, our method is two-stage, with the first stage (DC) independent from the second (CNN). This requires a major implementation effort to avoid storing DC filter responses and/or filtered patches (see Section 5.3), and a careful choice of the DC parameters to optimize the subsequent CNN training. An alternative and methodologically interesting solution which we will investigate in future work would be an *online* one-stage approach, with DC performing hard negative mining during CNN training. Third, our method can serve for lesion localization, but it does not produce lesion segmentations. Some applications, for example mammography mass detection, greatly benefit from obtaining lesion contours that can help discriminating benign from malignant findings. In such cases, pixel-level groundtruths are often available, and U-net based solutions that can also deal with class imbalance should be preferred [92]. However, by looking at the probability images in Fig. 2, we observe that DC performs a rough segmentation of lesions. This could be the seed of a region growing approach (e.g. watershed segmentation within a fixed patch window centered at groundtruth locations) and then used to train a U-net for full segmentation, which is another possible direction of our future work.

8. Conclusions

In this work, we have focused on the strategies to address class imbalance in CNNs for small lesion detection. Having tested a wide range of methods, we found that oversampling of the lesion class and hard mining on the background class during training were the most successful approaches. We also proposed a two-stage deep learning scheme where all class imbalance related issues are handled by an ad hoc designed cascade of decision trees, so that the subsequent CNN can focus on discriminating the lesions from the most challenging background samples. Our method overperformed all other tested methods in terms of mean lesion sensitivity, and was $\sim 10\times$ faster at inference time. All detectors were trained and tested on patches from all the regions of the images. This allowed a direct comparison of the classification performance that was not influenced by the postprocessing usually required to obtain lesion candidates. To test the effectiveness of the proposed approach when applied on the whole image, we performed a FROC analysis and compared it with the best approaches previously tested. Also in this case, we obtained significant improvements in lesion detection performance. For these reasons, we believe that the findings of this work could be transferred to a full CAD scheme that includes ad-hoc lesion postprocessing and/or false positive reduction stages, similarly to what has been done in [56].

Declaration of competing interest

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

References

[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.
[2] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.
[3] D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Mastering the game of go with deep neural networks and tree search, nature 529 (2016) 484.
[4] V. Gulshan, L. Peng, M. Coram, M.C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al., Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, JAMA 316 (2016) 2402–2410.
[5] 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 (2017) 115–118.
[6] M. Ghafoorian, N. Karssemeijer, T. Heskes, I.W. Uden, C.I. Sanchez, G. Litjens, F.E. Leeuw, B. Ginneken, E. Marchiori, B. Platel, Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities, Sci. Rep. 7 (2017) 5110.
[7] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al., Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017, arXiv preprint arXiv:1711.05225.
[8] M. Sinclair, C. Baumgartner, J. Matthew, W. Bai, J. Martinez, Y. Li, S. Smith, C. Knight, B. Kainz, J. Hajnal, A. King, D. Rueckert, Human-level performance on automatic head biometrics in fetal ultrasound using fully convolutional neural networks, in: Conference Proceedings :. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference EMBC 2018, 2018.
[9] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. van derLaak, B. Van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.

[10] S. Li, W. Song, H. Qin, A. Hao, Deep variance network: An iterative, improved CNN framework for unbalanced training datasets, Pattern Recognit. 81 (2018) 294–308.

[11] M. Buda, A. Maki, M.A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, Neural Netw. 106 (2018) 249–259.

[12] M.J.J.P. van Grinsven, B. van Ginneken, C.B. Hoyng, T. Theelen, C.I. Sánchez, Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images, IEEE Trans. Med. Imaging 35 (2016) 1273–1284.

[13] J.M. Wolterink, T. Leiner, B.D. de Vos, R.W. van Hamersvelt, M.A. Viergever, I. Išgum, Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks, Med. Image Anal. 34 (2016) 123–136.

[14] B. Dashtbozorg, J. Zhang, F. Huang, B.M. ter Haar Romeny, Retinal microaneurysms detection using local convergence index features, IEEE Trans. Image Process. 27 (2018) 3300–3315.

[15] H. Cheng, X. Cai, X. Chen, L. Hu, X. Lou, Computer-aided detection and classification of microcalcifications in mammograms: a survey, Pattern Recognit. 36 (2003) 2967–2991.

[16] A. Bria, N. Karssemeijer, F. Tortorella, Learning from unbalanced data: A cascade-based approach for detecting clustered microcalcifications, Med. Image Anal. 18 (2014) 241–252.

[17] A. Carass, S. Roy, A. Jog, J. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C. Sudre, et al., Longitudinal multiple sclerosis lesion segmentation: Resource and challenge, Neuroimage 148 (2017) 77–102.

[18] A. Bria, C. Marrocco, M. Molinara, B. Savelli, J.J. Mordang, et al., Improving the automated detection of calcifications by combining deep cascades and deep convolutional nets, in: Imaging (IWBI 2018) 1071808, 2018, p. 6.

[19] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, G. Bing, Learning from class-imbalanced data: Review of methods and applications, Expert Syst. Appl. 73 (2017) 220–239.

[20] S.C. Wong, A. Gatt, V. Stamatescu, M.D. McDonnell, Understanding data augmentation for classification: When to warp? in: 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2016, pp. 1–6.

[21] S. Ando, C.Y. Huang, Deep over-sampling framework for classifying imbalanced data, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2017, pp. 770–785.

[22] A. Shrivastava, A. Gupta, R.B. Girshick, Training region-based object detectors with online hard example mining, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 761–769.

[23] M.T. Ricamato, C. Marrocco, F. Tortorella, Mcs-based balancing techniques for skewed classes: An empirical comparison, in: 2008 19th International Conference on Pattern Recognition, 2008, pp. 1–4.

[24] C. Elkan, The foundations of cost-sensitive learning, in: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, 2001, pp. 973–978.

[25] Y.S. Resheff, A. Mandelbom, D. Weinshall, Controlling imbalanced error in deep learning with the log bilinear loss, in: First International Workshop on Learning with Imbalanced Domains: Theory and Applications, 2017, pp. 141–151.

[26] Y.S. Aurelio, G.M. de Almeida, C.L. de Castro, A.P. Braga, Learning from imbalanced data sets with weighted cross-entropy function, Neural Process. Lett. (2019).

[27] Z.H. Zhou, X.Y. Liu, Training cost-sensitive neural networks with methods addressing the class imbalance problem, IEEE Trans. Knowl. Data Eng. 18 (2006) 63–77.

[28] H. Sohn, K. Worden, C.R. Farrar, Novelty detection using auto-associative neural network, in: Symposium on Identification of Mechanical Systems: International Mechanical Engineering Congress and Exposition, New York, NY, 2001, pp. 573–580.

[29] F. Milletari, N. Navab, S.A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 3D Vision (3DV), 2016 Fourth International Conference on, IEEE, 2016, pp. 565–571.

[30] T. Wollmann, M. Gunkel, I. Chung, H. Erfle, K. Rippe, K. Rohr, GRUU-Net: Integrated convolutional and gated recurrent neural network for cell segmentation, Med. Image Anal. 56 (2019) 68–79.

[31] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, H. Greenspan, GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification, 2018, arXiv preprint arXiv:1803.01229.

[32] J. Jiang, X. Liu, K. Zhang, E. Long, L. Wang, W. Li, L. Liu, S. Wang, M. Zhu, J. Cui, et al., Automatic diagnosis of imbalanced ophthalmic images using a cost-sensitive deep convolutional neural network, Biomed. Eng. Online 16 (132) (2017).

[33] F. Ciompi, K. Chung, S.J. Van Riel, A.A.A. Setio, P.K. Gerke, C. Jacobs, E.T. Scholten, C. Schaefer-Prokop, M.M. Wille, A. Marchianò, et al., Towards automatic pulmonary nodule management in lung cancer screening with deep learning, Sci. Rep. 7 (46479) (2017).

[34] X. Liu, F. Hou, H. Qin, A. Hao, Multi-view multi-scale CNNs for lung nodule type classification from ct images, Pattern Recognit. 77 (2018) 262–275.

[35] S. Wang, M. Zhou, Z. Liu, Z. Liu, D. Gu, Y. Zang, D. Dong, O. Gevaert, J. Tian, Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation, Med. Image Anal. 40 (2017) 172–183.

[36] H. Xie, D. Yang, N. Sun, Z. Chen, Y. Zhang, Automated pulmonary nodule detection in CT images using deep convolutional neural networks, Pattern Recognit. 85 (2019) 109–119.

[37] K. Kamnitsas, C. Ledig, V.F. Newcombe, J.P. Simpson, A.D. Kane, D.K. Menon, D. Rueckert, B. Glocker, Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, Med. Image Anal. 36 (2017) 61–78.

[38] G. Litjens, C.I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-Van De Kaa, P. Bult, B. Van Ginneken, J. Van Der Laak, Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis, Sci. Rep. 6 (2016) 26286.

[39] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.M. Jodoin, H. Larochelle, Brain tumor segmentation with deep neural networks, Med. Image Anal. 35 (2017) 18–31.

[40] S. Pereira, A. Pinto, V. Alves, C.A. Silva, Brain tumor segmentation using convolutional neural networks in MRI images, IEEE Trans. Med. Imaging 35 (2016) 1240–1251.

[41] J.J. Mordang, T. Janssen, A. Bria, T. Kooi, A. Gubern-Mérida, N. Karssemeijer, Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks, in: International Workshop on Digital Mammography, Springer, 2016, pp. 35–42.

[42] J. Wang, Y. Yang, A context-sensitive deep learning approach for microcalcification detection in mammograms, Pattern Recognit. 78 (2018) 12–22.

[43] Q. Wei, Y. Ren, R. Hou, B. Shi, J.Y. Lo, L. Carin, Anomaly detection for medical images based on a one-class classification, in: Medical Imaging 2018: Computer-Aided Diagnosis, International Society for Optics and Photonics, 2018, p. 105751M.

[44] P. Chudzik, S. Majumdar, F. Calivá, B. Al-Diri, A. Hunter, Microaneurysm detection using fully convolutional neural networks, Comput. Methods Programs Biomed. 158 (2018) 185–192.

[45] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241.

[46] T. Brosch, L.Y. Tang, Y. Yoo, D.K. Li, A. Traboulsee, R. Tam, Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation, IEEE Trans. Med. Imaging 35 (2016) 1229–1239.

[47] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 640–651.

[48] I.C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M.J. Cardoso, J.S. Cardoso, INbreast: toward a full-field digital mammographic database, Acad. Radiol. 19 (2012) 236–248.

[49] R.K. Samala, H.P. Chan, L.M. Hadjiiski, K. Cha, M.A. Helvie, Deep-learning convolution neural network for computer-aided detection of microcalcifications in digital breast tomosynthesis, in: G.D. Tourassi, S.G.A. III (Eds.), Medical Imaging 2016: Computer-Aided Diagnosis, International Society for Optics and Photonics. SPIE, 2016, pp. 234–240.

[50] M.V. Sainz de Cea, R.M. Nishikawa, Y. Yang, Estimating the accuracy level among individual detections in clustered microcalcifications, IEEE Trans. Med. Imaging 36 (2017) 1162–1171.

[51] M.H. Hesamian, W. Jia, X. He, P. Kennedy, Deep learning techniques for medical image segmentation: Achievements and challenges, J. Digit. Imaging 32 (2019) 582–596.

[52] E. Decencière, G. Cazuguel, X. Zhang, G. Thibault, J.C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno, D. Elie, P. Massin, Z. Viktor, A. Erginay, B. Laš, A. Chabouis, TeleOphta: Machine learning and image processing methods for teleophthalmology, IRBM 34 (2013) 196–203, Special issue : ANR TECSAN : Technologies for Health and Autonomy.

[53] B. Antal, A. Hajdu, Improving microaneurysm detection using an optimally selected subset of candidate extractors and preprocessing methods, Pattern Recognit. 45 (2012) 264–270.

[54] M.U. Akram, S. Khalid, S.A. Khan, Identification and classification of microaneurysms for early detection of diabetic retinopathy, Pattern Recognit. 46 (2013) 107–116.

[55] H.E. Wiley, F.L. Ferris, Chapter 47 - nonproliferative diabetic retinopathy and diabetic macular edema, in: S.J. Ryan, S.R. Sadda, D.R. Hinton, A.P. Schachat, S.R. Sadda, C. Wilkinson, P. Wiedemann, A.P. Schachat (Eds.), Retina (Fifth Edition), fifth ed., W.B. Saunders, London, 2013, pp. 940–968.

[56] A. Bria, C. Marrocco, N. Karssemeijer, M. Molinara, F. Tortorella, Deep cascade classifiers to detect clusters of microcalcifications, in: International Workshop on Digital Mammography, Springer, 2016, pp. 415–422.

[57] J. Huang, C. Ling, Using AUC and accuracy in evaluating learning algorithms, IEEE Trans. Knowl. Data Eng. 17 (2005) 299–310.

[58] A. Bria, C. Marrocco, M. Molinara, F. Tortorella, A ranking-based cascade approach for unbalanced data, in: Pattern Recognition (ICPR), 2012 21st International Conference on, IEEE, 2012, pp. 3439–3442.

[59] A. Bria, C. Marrocco, M. Molinara, F. Tortorella, An effective learning strategy for cascaded object detection, Inform. Sci. 340 (2016b) 17–26.
[60] P. Viola, M. Jones, Robust real-time object detection, Int. J. Comput. Vis 57 (2001) 137–154.
[61] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[62] G. Bradski, The opencv library, Dr. Dobb's J. Softw. Tools (2000).
[63] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, 2014, arXiv preprint arXiv:1408.5093.
[64] H. Ma, A.I. Bandos, H.E. Rockette, D. Gur, On use of partial area under the roc curve for evaluation of diagnostic performance, Stat. Med. 32 (2013) 3449–3458.
[65] R. Hupse, N. Karssemeijer, Use of normal tissue context in computer-aided detection of masses in mammograms, IEEE Trans. Med. Imaging 28 (2009) 2033–2041.
[66] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C.I. Sánchez, R. Mann, A. den Heeten, N. Karssemeijer, Large scale deep learning for computer aided detection of mammographic lesions, Med. Image Anal. 35 (2017) 303–312.
[67] A. Bria, C. Marrocco, L.R. Borges, M. Molinara, A. Marchesi, J. Mordang, N. Karssemeijer, F. Tortorella, Improving the automated detection of calcifications using adaptive variance stabilization, IEEE Trans. Med. Imaging 37 (2018) 1857–1864.
[68] B. Savelli, A. Bria, M. Molinara, C. Marrocco, F. Tortorella, A multi-context cnn ensemble for small lesion detection, Artif. Intell. Med. 103 (2020) 101749.
[69] D.P. Chakraborty, Validation and statistical power comparison of methods for analyzing free-response observer performance studies, Acad. Radiol. 15 (2008) 1554–1566.
[70] J. Wang, Y. Yang, A hierarchical learning approach for detection of clustered microcalcifications in mammograms, in: 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 804–808.
[71] Z. Lu, G. Carneiro, N. Dhungel, A.P. Bradley, Automated detection of individual micro-calcifications from mammograms using a multi-stage cascade approach, 2016, arXiv preprint arXiv:1610.02251.
[72] F.W. Samuelson, N. Petrick, Comparing image detection algorithms using resampling, in: IEEE Int. Symp. Biomed. Imag., 2006, pp. 1312–1315,.
[73] O.J. Dunn, Multiple comparisons among means, J. Amer. Statist. Assoc. 56 (1961) 52–64.
[74] B. Thangaraju, I. Vennila, G. Chinnasamy, Detection of microcalcification clusters using hessian matrix and foveal segmentation method on multiscale analysis in digital mammograms, J. Digit. Imaging 25 (2012) 607–619.
[75] M. Muthuvel, B. Thangaraju, G. Chinnasamy, Microcalcification cluster detection using multiscale products based hessian matrix via the tsallis thresholding scheme, Pattern Recognit. Lett. 94 (2017) 127–133.
[76] Y. Wang, H. Zhao, H. Li, X. Pan, Y. Kang, An integrated detection method of clustered microcalcifications in mammography based on multiscale hessian matrix, in: 2012 IEEE International Conference on Information and Automation, 2012, pp. 106–110.
[77] S.S. Rubini, A. Kunthavai, Diabetic retinopathy detection based on eigenvalues of the hessian matrix, Procedia Comput. Sci. 47 (2015) 311–318, Graph Algorithms, High Performance Implementations and Its Applications (ICGHIA 2014).
[78] T. Inoue, Y. Hatanaka, S. Okumura, C. Muramatsu, H. Fujita, Automated microaneurysm detection method based on eigenvalue analysis using hessian matrix in retinal fundus images, in: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2013, pp. 5873–5876.
[79] K. Adal, S. Ali, D. Sidibé, T. Karnowski, E. Chaum, F. Mériaudeau, Automated detection of microaneurysms using robust blob descriptors, in: C.L. Novak, S. Aylward (Eds.), Medical Imaging 2013: Computer-Aided Diagnosis, International Society for Optics and Photonics. SPIE, 2013, pp. 158–164.
[80] M.M. Álvarez Cervera, M.F. Escalante Paredes, R. Nava Martínez, C. Castillo Ortiz, N. Ramírez Hernández, Development of a detection system microaneurysms in color fundus images, in: 2016 13th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE), 2016, pp. 1–5.
[81] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: Integrated recognition, localization and detection using convolutional networks, 2013, arXiv preprint arXiv:1312.6229.
[82] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: European Conference on Computer Vision, Springer, 2016, pp. 21–37.
[83] J. Redmon, A. Farhadi, Yolo9000: better, faster, stronger, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7263–7271.
[84] T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[85] H. Jung, B. Kim, I. Lee, M. Yoo, J. Lee, S. Ham, O. Woo, J. Kang, Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network, PLOS ONE 13 (2018) 1–16.
[86] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.
[87] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.
[88] D. Ribli, A. Horváth, Z. Unger, P. Pollner, I. Csabai, Detecting and classifying lesions in mammograms with deep learning, Sci. Rep. 8 (2018) 1–7.
[89] R. Reiazi, R. Paydar, A.A. Ardakani, M. Etedadialiabadi, Mammography lesion detection using faster r-cnn detector, 2018.
[90] J.Y. Chiao, K.Y. Chen, K.Y.K. Liao, P.H. Hsieh, G. Zhang, T.C. Huang, Detection and classification the breast tumors using mask r-cnn on sonograms, Medicine 98 (2019).
[91] H. Min, D. Wilson, Y. Huang, S. Kelly, S. Crozier, A.P. Bradley, S.S. Chandra, Fully automatic computer-aided mass detection and segmentation via pseudo-color mammograms and mask r-cnn, 2019, arXiv preprint arXiv:1906.12118.
[92] S. Li, M. Dong, G. Du, X. Mu, Attention dense-u-net for automatic breast mass segmentation in digital mammogram, IEEE Access 7 (2019) 59037–59047.

Alessandro Bria is currently an assistant professor at the Department of Electrical and Information Engineering of the University of Cassino and Southern Lazio. His research interests include biomedical image understanding and in particular Computer Aided Diagnosis systems, imbalanced classification in deep learning, large-scale bioimage assisted visualization and annotation, machine learning for brain computer interfaces. He has authored over 40 scientific papers in journals and international conference proceedings.

Claudio Marrocco is currently an assistant professor at the Department of Electrical and Information Engineering of the University of Cassino and Southern Lazio. His research activities include Pattern Recognition and Artificial Vision and, in particular, to the implementation of reliable classification systems, the study of deep learning and cost-sensitive classification systems, and the automatic interpretation of biomedical images. He authored more than 50 scientific papers in books, journals, and international conference proceedings.

Francesco Tortorella is a Full Professor of Computer Engineering at the University of Salerno, Italy. He has authored over 90 research papers in international journals and conference proceedings. His current research interests include classification techniques, statistical learning, medical image analysis and interpretation. Prof. Tortorella is a member of the IAPR and a senior member of the IEEE.