Medical Image Analysis 18 (2014) 241–252

Contents lists available at ScienceDirect

Medical Image Analysis

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

Learning from unbalanced data: A cascade-based approach for detecting clustered microcalcifications

A. Bria [a,*], N. Karssemeijer [b], F. Tortorella [a]

[a] Department of Electrical and Information Engineering, University of Cassino and L.M., Via Di Biasio 43, 03043 Cassino (FR), Italy
[b] Diagnostic Image Analysis Group, Radboud University Nijmegen Medical Centre, P.O. Box 9102, 6500 HC Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 26 June 2013
Received in revised form 18 October 2013
Accepted 31 October 2013
Available online 12 November 2013

Keywords:
Computer aided detection
Unbalanced data
Clustered microcalcifications
Mammography

ABSTRACT

Finding abnormalities in diagnostic images is a difficult task even for expert radiologists because the normal tissue locations largely outnumber those with suspicious signs which may thus be missed or incorrectly interpreted. For the same reason the design of a Computer-Aided Detection (CADe) system is very complex because the large predominance of normal samples in the training data may hamper the ability of the classifier to recognize the abnormalities on the images. In this paper we present a novel approach for computer-aided detection which faces the class imbalance with a cascade of boosting classifiers where each node is trained by a learning algorithm based on ranking instead of classification error. Such approach is used to design a system (*CasCADe*) for the automated detection of clustered microcalcifications (μCs), which is a severely unbalanced classification problem because of the vast majority of image locations where no μC is present. The proposed approach was evaluated with a dataset of 1599 full-field digital mammograms from 560 cases and compared favorably with the Hologic R2CAD ImageChecker, one of the most widespread commercial CADe systems. In particular, at the same lesion sensitivity of R2CAD (90%) on biopsy proven malignant cases, CasCADe and R2CAD detected 0.13 and 0.21 false positives per image (FPpi), respectively (p-value = 0.09), whereas at the same FPpi of R2CAD (0.21), CasCADe and R2CAD detected 93% and 90% of true lesions respectively (p-value = 0.11) thus showing that CasCADe can compete with high-end CADe commercial systems.

1. Introduction

Clustered microcalcifications (μCs) are one of the most important early indicators of breast cancer since they appear in 30–50% of cases diagnosed by mammographic screenings (Kopans, 2007). However, interpreting screening mammograms is a big challenge even for an expert radiologist since the low prevalence makes finding abnormalities difficult. Birdwell (2009) points out several subjective factors that may lead to a lack of perception or to mistakes in interpretation. Among the established methods to improve radiologist performance, it has been reported that having more than one radiologist or a Computer-Aided Detection (CADe) system improves the detection of cancer in mammograms (Karssemeijer et al., 2009; Eadie et al., 2012). To this end, several commercial CADe systems are nowadays available and their use is widespread among radiologists. However, even though CADe systems show a sensitivity similar to radiologists (Cole et al., 2012), there are still a few hundred false positives for every true positive in a screening setting, which is about two orders of magnitude higher than what the radiologists achieve (Karssemeijer et al., 2009) and this potentially limits the benefit that a CADe system can provide. For this reason, the design of CADe systems for clustered μCs is still an open research field as shown by the recent literature (El Naqa et al., 2002; Wei et al., 2005; Tang et al., 2009; Zhang et al., 2009; Oliver et al., 2010; Jing et al., 2011).

Among the proposed approaches, methods based on supervised learning techniques have received the largest share of research since they can yield powerful binary classifiers able to determine whether a μC is present (positive) or not (negative) at a pixel location. However, such methods have to face two major problems. First, the huge number of pixels to be analyzed (e.g., about 9 million in a digital mammogram) coupled with high-complexity classifiers may cause a computational burden not easy to sustain, especially when hundreds or thousands of images have to be processed. Second, the vast majority of image locations where no μC is present makes detection a severely unbalanced classification problem, where the negative class is several orders of magnitude bigger than the positive class. Generally speaking, binary classifiers trained on highly unbalanced data sets tend to be overwhelmed by the majority class, thus misclassifying the samples belonging to the minority class. This problem is also known as class imbalance and in recent years it has received considerable attention in

* Corresponding author. Tel.: +39 3480339824.
E-mail addresses: a.bria@unicas.it (A. Bria), n.karssemeijer@rad.umcn.nl (N. Karssemeijer), tortorella@unicas.it (F. Tortorella).

1361-8415/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.media.2013.10.014

the machine learning community (e.g., Barandela et al., 2003; Guo et al., 2008) and subsequently in the medical image analysis field (Li et al., 2010). Several approaches focused on μC detection (e.g., Wei et al., 2005; Zhang et al., 2009; Oliver et al., 2010; Marrocco et al., 2010) address class imbalance by randomly selecting a limited set of negative samples so as to obtain approximately the same size for the two classes. Nevertheless, there is no guarantee that the selected subset is actually representative of all the possible negative samples. A different solution is proposed in El Naqa et al. (2002) in which a Support Vector Machine (SVM) is employed with a *Successive Enhancement Learning* (SEL) scheme where the SVM is initially trained with a balanced training set containing a limited number of negative samples. The training is then restarted iteratively by incorporating another N misclassified negative samples from all the available training images. The retraining step is repeated until no more changes are observed in support vectors. In this way the total number of training samples is kept small and balanced at each retraining round, but the final classifier could be very complex since it could contain a very large number of support vectors, thus making the detection phase computationally intense.

In this paper we present *CasCADe*, a multistage system for the automatic detection of clustered μCs on full-field digital mammograms (FFDM), specifically designed to handle efficiently and effectively the computational complexity and the high class imbalance. Even though aimed at the μC detection problem, the approach proposed in this work could be more generally applicable in medical image analysis and especially in other unbalanced problems such as the automated detection of lung nodules in CT (e.g., vanGinneken10 et al., 2010), chest lymph nodes in CT (e.g., Barbu et al., 2012; Feulner et al., 2013), colon polyps in CT colonography (e.g., Van Ravesteijn et al., 2010), and retinal microaneurysms in Digital Color Fundus Photographs (e.g., Niemeijer et al., 2010). The rationale of our approach is to employ an ensemble of ranking-based boosting classifiers connected in series with increasing complexity and specificity like in the cascade face detector proposed by Viola and Jones (2001). The choice of boosting-based classifiers is particularly fitting for unbalanced problems as demonstrated by Galar et al. (2011) who empirically compared the most significant published approaches and showed that for two-class unbalanced problems the best results were obtained by using random undersampling techniques coupled with bagging or boosting ensembles. In our approach each classifier stage is trained with only a part of the negative samples, thus distributing the complexity of the whole problem among the classifiers and alleviating class imbalance at the same time. Nevertheless, the residual imbalance present at each node could produce unsatisfactory results if the learning algorithm used in the node is based on the optimization of a performance measure (such as the empirical error) highly affected by the class distribution skew. This is the case of AdaBoost, the learning algorithm employed in the approach of Viola and Jones. The same authors observe in a successive paper (Viola and Jones, 2002) that AdaBoost minimizes a quantity related to the classification error (and not the number of false negatives) and thus propose a variant aimed at moderating the effects of the class imbalance by introducing an asymmetric weight updating mechanism of the samples in the training set.

The novelty of our approach firstly lies in handling the class imbalance in each node through a boosting algorithm designed to maximize the Area under the ROC curve (AUC). The reason for this choice is that AUC is equivalent to the probability of correct pairwise ranking and thus provides a measure of the predictive ability of the classifier which is robust and insensitive to the class skew (Huang and Ling, 2005). To this end we adopted a reformulation of RankBoost for bipartite ranking problems (Freund et al., 2003), suitably modified to be embedded in a cascade structure. Another difference from the approach of Viola and Jones (2001) is that our cascade-based detector is used not only for μC localization, but also for accurately estimating the outline of a μC, which has been proven to play an important role for the automated differentiation between true positive and false positive detected μCs (Veldkamp and Karssemeijer, 1996). Indeed, after grouping μCs into clusters, we classify them into "abnormal" (true positive) and "normal" (false positive) clustered μCs, the latter including both benign clusters of μCs and erroneously detected clusters. The low prevalence of cancer within a mammographic screening cohort makes also this decision an unbalanced problem and thus we employ again a RankBoost classifier. To this end, we also propose a novel set of features especially aimed at capturing the topological relations between μCs.

The detection performance of CasCADe was evaluated on 1599 full-field digital mammograms from 560 cases obtained in routine screening and compared with the one of the most widespread commercial CADe systems, the Hologic R2CAD ImageChecker. To our knowledge, the scientific literature does not exhibit other CADe systems which have been compared with high-end commercial systems.

2. Method

The CasCADe system consists of a preprocessing stage, an initial detection stage and a classification stage in which the number of false positive detected clusters is reduced. A schematic overview of these stages is given in Fig. 1. Each of these stages is detailed in the following subsections.

2.1. Preprocessing stage: quantum noise equalization

In FFDM the dominant source of noise is quantum noise that is caused by fluctuations in photon fluence at the detector. These fluctuations can be described by a Poisson distribution with standard deviation $\sqrt{\lambda}$, where λ is the average number of detected photons (Beutel et al., 2000; Schie and Karssemeijer, 2008). Since in an FFDM system a linear relationship exists between gray level and exposure, quantum noise standard deviation σ_q can be estimated by (e.g., McLoughlin et al., 2004; Schie and Karssemeijer, 2008):

$$\sigma_q(y) = c\sqrt{y} \tag{1}$$

where c is a noise level parameter to be estimated and y is the pixel intensity.

Actually noise properties vary across the image and thus c should be estimated locally as, for example, proposed in Schie and Karssemeijer (2008). However, the same study reports that the improvement in μC detection performance obtained with a nonuniform noise model is quite limited and thus we adopted an uniform noise model in which c is constant across the image. On this basis, in order to rescale pixel intensities to a scale with uniform noise level, we consider a scale transform $y' = T(y)$ that satisfies the following differential equation:

$$dT(y) = \frac{\alpha}{\sigma_q(y)}dy \tag{2}$$

where α is the noise level on the transformed scale and the factor $\sigma_q(y)^{-1}$ eliminates the dependency of the differential dy on the noise variation. Coupling Eq. (2) with Eq. (1) and with border conditions, we obtain the following Cauchy problem:

$$\begin{cases} \dfrac{dT(y)}{dy} = \dfrac{\alpha}{c\sqrt{y}} \\ T(0) = 0 \\ T(y_{max}) = T_{max} \end{cases} \tag{3}$$

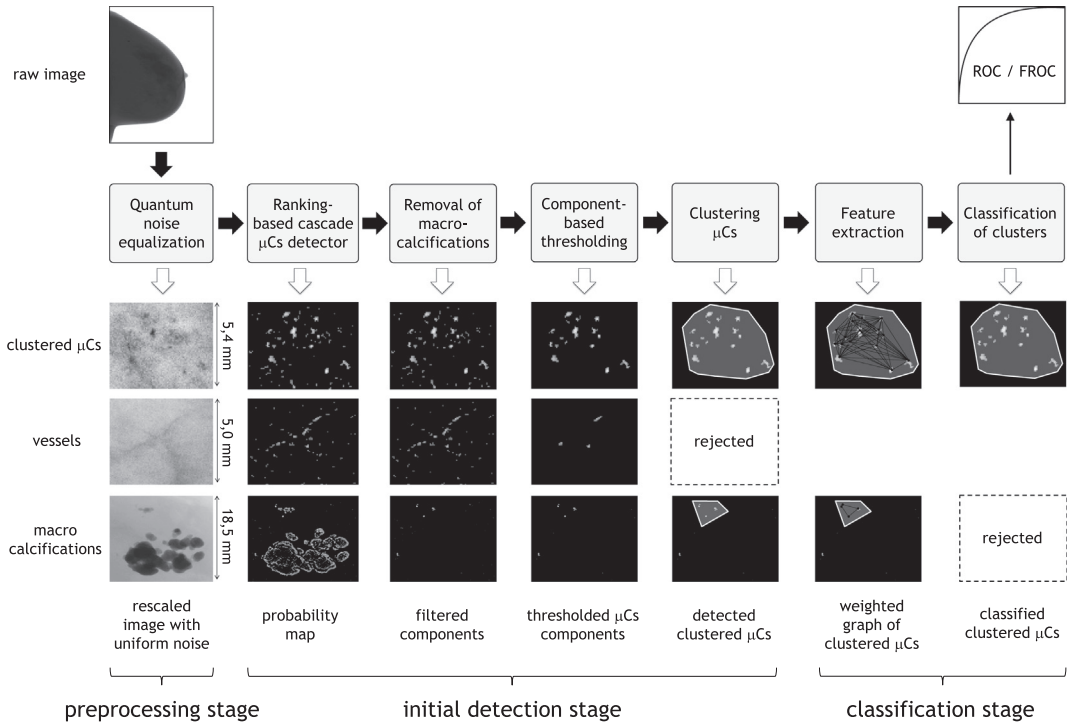

Fig. 1. A schematic overview of the CasCADe system. The substeps of each stage are reported on the top row. The output of each step is shown on three patches of different sizes: (i) a cluster of μCs; (ii) blood vessels; and (iii) a group of macrocalcifications. The images of the patches have been modified to improve their readability.

that has unique solution

$$T(y) = \frac{T_{max}}{\sqrt{y_{max}}}\sqrt{y} \tag{4}$$

where y_{max} is the maximum intensity value of the original scale and T_{max} the maximum intensity value of the transformed scale. In our case we fixed $y_{max} = 4095$ since our raw images were 12 bits/pixel and we chose $T_{max} = 65,535$ to store the rescaled image in 16 bits/pixel.

2.2. Initial detection stage

2.2.1. Ranking-based cascade μC detection

The proposed μC detection approach relies on a two-class ranking-based cascade classifier which classifies each pixel of the mammogram as positive (belonging to a μC) or negative (not belonging to any μC) using a subwindow of size $M \times M$ centered on it, in the following referred to as sample and denoted by $\mathbf{x}$. Each pixel classified as positive is assigned a score, thus obtaining a probability map whose peculiar characteristics are described in Section 2.2.2. In the following paragraphs we describe the features employed and both the detection and learning phases of the proposed μC detection method.

2.2.1.1. Feature set. Three groups of Haar-like features are used. For the first group the value of each feature is calculated as the difference between the sum of pixels belonging to adjacent rectangular regions, aimed at capturing edge and elongated patterns (see Fig. 2a). For the second group, the value is calculated in a similar way, but the support regions are two concentric rectangles, so being more suitable for the granule-like shape of μCs (see Fig. 2b). The third group is constituted by the 45°-rotated version of the features of the first two groups (see Fig. 2c). The features of the first two groups are evaluated very quickly thanks to the *integral image* (Viola and Jones, 2001), while for the rotated features a particular representation of the image is introduced, similar to the *integral image*, but suited for the calculation of tilted rectangle areas (Lienhart et al., 2003). All features are scaled and translated separately across all possible combinations on the subwindow (see Fig. 3), so obtaining tens of thousands of features. The task of selecting the best features is part of the training phase in each node classifier.

2.2.1.2. Detection phase: distributing the discriminative power. The underlying idea is to employ a sequence of node classifiers with different discriminative power (see Fig. 4). A given sample passes to the next node if the current node classifies it as a μC, otherwise it is rejected. The majority of samples belonging to easily detectable background tissue are discarded by the early nodes, while the most likely-μC samples go through the entire cascade. As a result, the detection rate D and false positive rate F of a cascade composed by n nodes is given by

$$D = \prod_{i=1}^{n} d_i \quad F = \prod_{i=1}^{n} f_i \tag{5}$$

where d_i and f_i are the detection rate and false positive rate of the ith node, respectively. Such approach, previously proposed in other fields (Viola and Jones, 2001), allows us to address the learning task in a more effective way. In fact, while it is hard for a monolithic classifier to ensure both a good sensitivity and a good specificity, the cascade provides a high constant sensitivity and a growing specificity through the nodes obtained by connecting more simpler classifiers with high sensitivity and sufficient specificity. As an example, to build a detector having $D = 0.990$ and $F = 0.001$ it would be sufficient to build 6 node classifiers, each with $d_i = 0.999$ and $f_i = 0.3$. In this way, the first nodes of the cascade have to face a simpler task (rejecting the most distinguishable background tissue regions), while the last ones are specialized to discriminate between actual μCs and the background tissue configurations most resembling a μC. This should reduce the number of false positives produced by

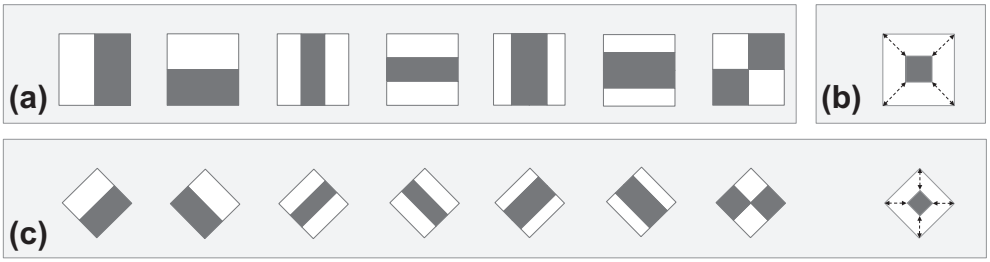

Fig. 2. The Haar-like feature groups used by the proposed cascade of classifiers. (a) Some examples of the first group. (b) An example of the second group. (c) Some examples of the third group.

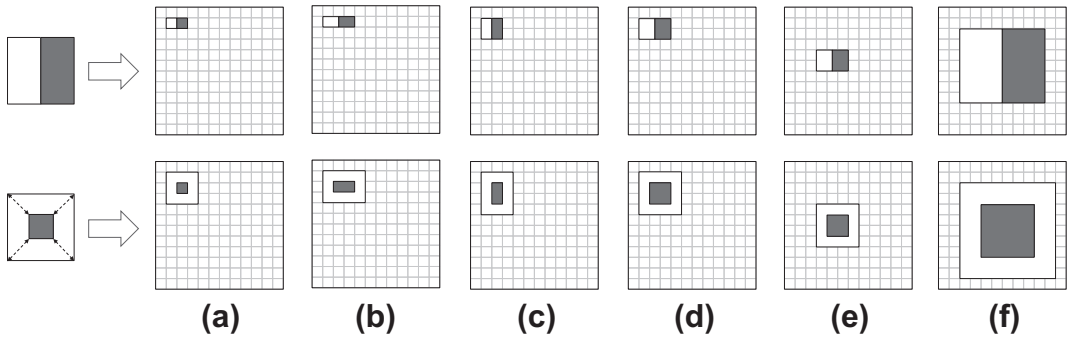

Fig. 3. Some examples of scaling and translation operations applied on two features on a 12×12 subwindow. (a) base feature of width w and height h. (b) scaling to $(w+1) \times h$. (c) scaling to $w \times (h+1)$. (d) scaling to $(w+1) \times (h+1)$. (e) scaling to $(w+1) \times (h+1)$ and translating by $(2,3)$. (f) scaling to $(w+6) \times (h+6)$ and translating by $(1,1)$.

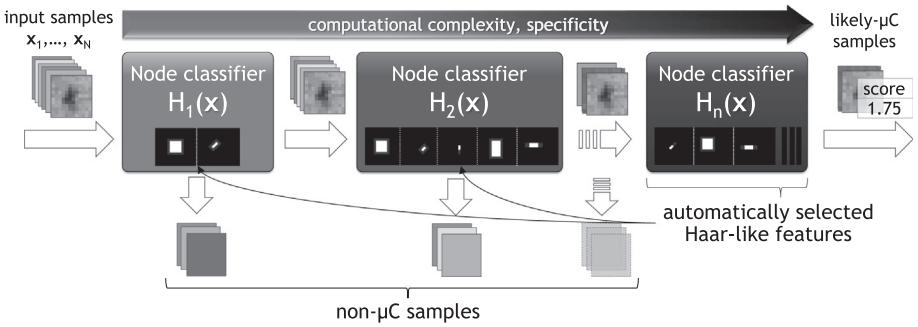

Fig. 4. The ranking-based cascade μC detector. A given sample $\mathbf{x}$ passes to the next node if the current node classifies it as μC, otherwise it is rejected. The majority of samples belonging to easily detectable background tissue are discarded by the early nodes, while the most likely-μC samples go through the entire cascade and receive a score by the last node classifier $H_n(\mathbf{x})$. Such scores are subsequently used to build the probability map.

the detector and concentrate the computational complexity of the system on the last classifiers of the cascade. At the end of the cascade, each not rejected pixel is assigned a score representing how confident the detector is in deciding for the presence of a μC (see Fig. 4). On the contrary, the background tissue pixels are gradually rejected by the nodes along the cascade and thus have a null value. As a result, the output of the cascade applied on a mammogram is a *probability map* in which the rejected pixels are null-labeled and the nonrejected pixels are labeled with their scores (see Fig. 5).

2.2.1.3. Learning phase: facing class imbalance. For building the cascade, positive samples are extracted from the μC locations in the training images, whereas negative samples are extracted from all the available locations in the training images where no μC is present (more details on the construction of the training set will be given in Section 3.2.1). Let us call $\mathcal{N}$ and $\mathcal{P}$ the negative and positive training sets, respectively. Since the occurrence of a μC is a rare event in a mammogram, the class imbalance ratio $|\mathcal{N}|/|\mathcal{P}|$ will be very high (typically greater than 10^3, see Section 3.2.1). Thus, it would be impractical to train a single classifier on such a skewed training set. The solution provided by the cascade is to part the complexity of the whole problem among the classifiers and alleviating the high class imbalance at the same time, as explained in the following.

The cascade is built incrementally by adding node classifiers $H_i(\mathbf{x})$ for $i = 1, 2, \ldots$ until the desired false positive rate F is reached or $\mathcal{N}$ is empty. For constructing the ith node classifier $H_i(\mathbf{x})$, we employ the set $\mathcal{P}_i \equiv \mathcal{P}$ and a set $\mathcal{N}_i$ of samples randomly extracted from $\mathcal{N}$ such that $|\mathcal{N}_i|/|\mathcal{P}_i| = k_r$ where k_r is the reduced class imbalance ratio, the same for all the nodes. The samples of $\mathcal{P}_i$ and $\mathcal{N}_i$ are equally distributed between a training set $\mathcal{T}_i$ and a validation set $\mathcal{V}_i$, that is used to set the decision threshold of the ith node classifier $H_i(\mathbf{x})$ so as to meet both the node learning goals of detection rate d_i and false positive rate f_i. When the ith node classifier $H_i(\mathbf{x})$ has been trained and a new node classifier $H_{i+1}(\mathbf{x})$ is required,

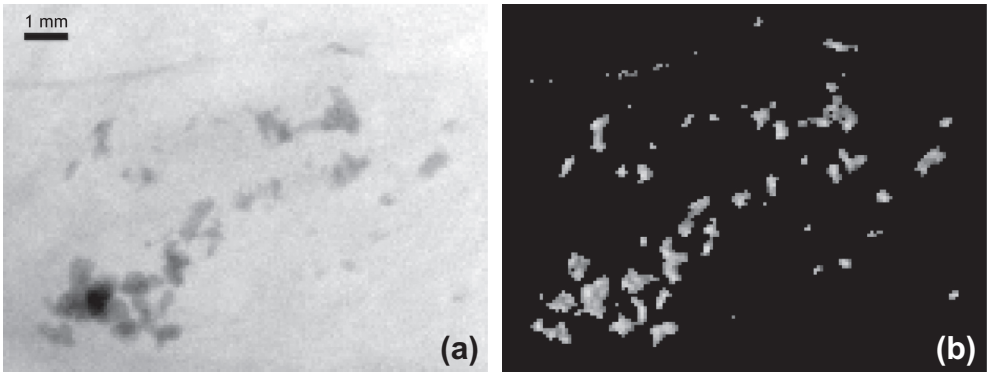

Fig. 5. A malignant cluster of μCs taken from a raw mammogram image (a) and the corresponding probability map obtained with the proposed ranking-based cascade μC detector (b): the pixels rejected by one node along the cascade are shown as black points. Image (a) has been modified to improve its readability.

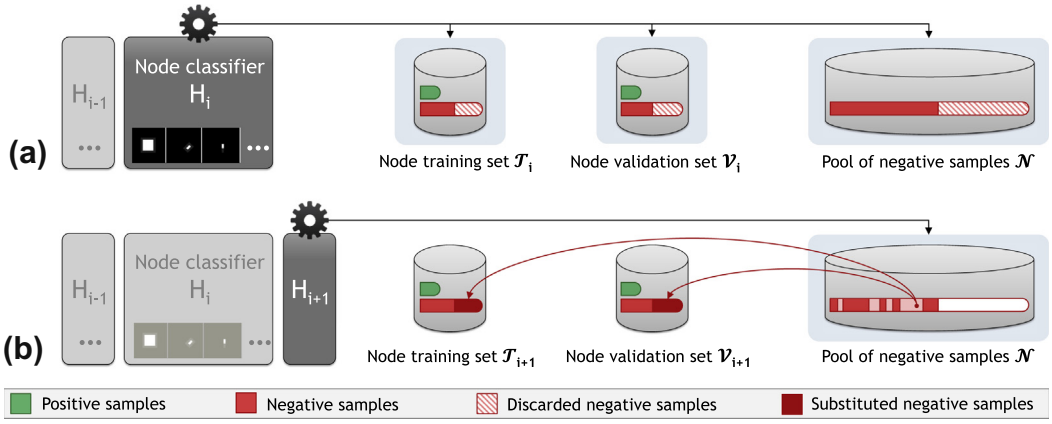

Fig. 6. The resampling strategy adopted by node classifiers throughout the cascade. (a) Once the ith node classifier $H_i(\mathbf{x})$ has been trained, it is applied to $\mathcal{T}_i, \mathcal{V}_i$ and $\mathcal{N}$ in order to remove all the correctly classified negative samples. (b) Subsequently, a new node classifier $H_{i+1}(\mathbf{x})$ is added which randomly samples from $\mathcal{N}$ as many negative samples as those discarded by the previous node.

the negative samples belonging to $\mathcal{T}_i, \mathcal{V}_i$ and $\mathcal{N}$ are reduced by discarding those rejected by $H_i(\mathbf{x})$ (see Fig. 6a). The remaining samples in $\mathcal{T}_i$ and $\mathcal{V}_i$ are used to form $\mathcal{T}_{i+1}$ and $\mathcal{V}_{i+1}$. Then, the class imbalance k_r is reestablished in $\mathcal{T}_{i+1}$ and $\mathcal{V}_{i+1}$ by randomly sampling from the remaining negative samples of $\mathcal{N}$ (see Fig. 6b).

To cope with the class imbalance in $\mathcal{T}_i$, the ith node classifier $H_i(\mathbf{x})$ is built so as to maximize the area under the ROC curve (AUC) achieved on $\mathcal{T}_i$, which we proved in Bria et al. (2012) to be a solution more robust and insensitive to the class skew with respect to other approaches like AsymBoost (Viola and Jones, 2002). To this end, we adopted for $H_i(\mathbf{x})$ a reformulation of RankBoost for bipartite ranking problems (Freund et al., 2003), suitably modified to be embedded in a cascade structure. It consists of a linear combination of weak classifiers $h_{i,\tau}(\mathbf{x}) \in \{0,1\}$ (0 for the negative class, 1 for the positive one) weighted by $\alpha_{i,\tau} \in \mathbb{R}$ and added in subsequent rounds $\tau = 1, \ldots, t$

$$H_{i,t}(\mathbf{x}) = \sum_{\tau=1}^{t} \alpha_{i,\tau} h_{i,\tau}(\mathbf{x}). \tag{6}$$

A weak classifier $h_{i,\tau}(\mathbf{x})$ consists of a simple *decision stump* given by

$$h_{i,\tau}(\mathbf{x}) = \begin{cases} 1 & \text{if } f_{i,\tau}(\mathbf{x}) > \theta_{i,\tau} \\ 0 & \text{if } f_{i,\tau}(\mathbf{x}) \leqslant \theta_{i,\tau} \end{cases} \tag{7}$$

where $f_{i,\tau}(\mathbf{x})$ is the feature selected at round τ and $\theta_{i,\tau}$ is the corresponding threshold.

To maximize the AUC achieved on $\mathcal{T}_i$, we consider all the sample pairs made by a positive sample $\mathbf{p} \in \mathcal{T}_i \cap \mathcal{P}_i$ and a negative sample $\mathbf{n} \in \mathcal{T}_i \cap \mathcal{N}_i$ (*crucial pairs*). Ideally, the maximum AUC (=1) is achieved by $H_{i,t}(\mathbf{x})$ when all the crucial pairs are correctly ranked by $H_{i,t}(\mathbf{x})$, i.e. when $H_{i,t}(\mathbf{n}) < H_{i,t}(\mathbf{p})\ \forall(\mathbf{p},\mathbf{n})$ (Hanley and McNeil, 1982). In practice, the AUC can be maximized by minimizing the number of misranked crucial pairs $R_{i,t}$ (*ranking loss*) defined as

$$R_{i,t} = \sum_{(\mathbf{p},\mathbf{n})} [[H_{i,t}(\mathbf{n}) \geqslant H_{i,t}(\mathbf{p})]] \tag{8}$$

with the notation $[[pr]]$ defined to be 1 if pr holds and 0 otherwise. A weight distribution $w_{i,t}(\mathbf{p},\mathbf{n})$ is maintained over the crucial pairs in $\mathcal{T}_i$ so that the misranked crucial pairs will be more influential in the following rounds. The weight update rule is given by

$$w_{i,t}(\mathbf{p},\mathbf{n}) = \frac{w_{i,t-1}(\mathbf{p},\mathbf{n}) \exp\left(\alpha_{i,t-1}\left(h_{i,t-1}(\mathbf{n}) - h_{i,t-1}(\mathbf{p})\right)\right)}{W_{i,t-1}} \tag{9}$$

where $W_{i,t-1}$ is a normalization factor so that $\sum_{\mathbf{p},\mathbf{n}} w_{i,t-1}(\mathbf{p},\mathbf{n}) = 1$. Assuming $\alpha_{i,t} > 0$, the update rule decreases the weight of crucial pairs in case of correct ranking (i.e., $h_{i,t}(\mathbf{p}) = 1$ and $h_{i,t}(\mathbf{n}) = 0$) and increases the weight otherwise.

Combining Eq. (8) with Eq. (9), the goal becomes to minimize the *weighted* number of misranked crucial pairs $R^w_{i,t}$ (*weighted ranking loss*) given by

$$R^w_{i,t} = \sum_{(\mathbf{p},\mathbf{n})} w_{i,t}(\mathbf{p},\mathbf{n}) [[H_{i,t}(\mathbf{n}) \geqslant H_{i,t}(\mathbf{p})]] \tag{10}$$

To this end, $h_{i,t}(\mathbf{x})$ (i.e. $f_{i,t}(\mathbf{x})$ and $\theta_{i,t}$) and $\alpha_{i,t}$ are chosen at each round t so as to minimize (10). Following Freund et al. (2003), for each $h_{i,t}(\mathbf{x})$ a suboptimal α is provided by

$$\alpha_{h_{i,t}} = \frac{1}{2}\ln\left(\frac{1+\sum_{\mathbf{p},\mathbf{n}} w_{i,t}(\mathbf{p},\mathbf{n})(h_{i,t}(\mathbf{n})-h_{i,t}(\mathbf{p}))}{1-\sum_{\mathbf{p},\mathbf{n}} w_{i,t}(\mathbf{p},\mathbf{n})(h_{i,t}(\mathbf{n})-h_{i,t}(\mathbf{p}))}\right) \tag{11}$$

and the weak classifier to be chosen at round t is

$$h_{i,t} = arg\,max_h \sum_{(\mathbf{p},\mathbf{n})} w_{i,t}(\mathbf{p},\mathbf{n})[[H_{i,t-1}(\mathbf{n})+\alpha_h h(\mathbf{n}) \geqslant H_{i,t-1}(\mathbf{p})+\alpha_h h(\mathbf{p})]] \tag{12}$$

In summary, at each round t, both the selection of the feature $f_{i,t}(\mathbf{x})$ and the evaluation of the coefficient $\alpha_{i,t}$ are made so as to maximize the correct pairwise ranking and thus the AUC (Cortes and Mohri, 2004). Further details on this stage along with the actual implementation can be found in Bria et al. (2012).

2.2.2. Removal of macrocalcifications

Once the cascade detector has been built, it is applied to each pixel in a mammogram. Each pixel $\mathbf{p}$ arriving to the last node classifier n is assigned a score equal to $H_n(\mathbf{p})$, while the rejected pixels are assigned a null score. The obtained scores are used to build a probability map having the same dimensions of the image under testing, with μCs appearing as connected components on a null-valued background (see Fig. 5). Differently from the original approach proposed by Viola and Jones (2001), this let us not only to localize individual μCs, but also to accurately estimate their outline. This has been proven to play an important role for discriminating between true positive and false positive detected μCs (Veldkamp and Karssemeijer, 1996).

However, the features employed by the node classifiers rely on a fixed size subwindow suited for the μC scale. As a consequence, objects larger than a μC such as macrocalcifications are fragmented into many components of different sizes (see Fig. 7b) which may result in a false positive detection in the subsequent steps, and therefore should be removed. To this end, we compute the minimum bounding rectangle enclosing each connected component and find the rectangles larger than 1 mm in both width and height. Since typical size of a μC is between 0.1 mm and 1 mm (Cheng et al., 2003), it is far likely that these rectangles do not enclose a μC, so we remove their corresponding components along with the smaller components whose bounding rectangles intersect with them (see Fig. 7c).

2.2.3. Component-based thresholding

Among the components not removed in the previous step, the detection process has to decide which of them should be regarded as a μC to be passed to the subsequent steps, according to desired sensitivity level of the μC detector. To this end, a component is labeled as a detected μC if the second highest value in its probability map is above a threshold corresponding to the desired sensitivity level. The choice of the second maximum makes the detector insensitive to the isolated peaks that might occur in the probability map.

2.2.4. Clustering of μCs

Starting from the centroids of the detected μCs, a weighted graph is generated by connecting all the pairs of μCs that are distant less than $d_{max} = 10$ mm from each other, with edge weights corresponding to the inter-μC distances. Pairs of μCs distant more than d_{max} were not considered since they did not convey any useful information about the nature of the cluster. Subsequently, we find the connected components of the graph, i.e. the subgraphs in which any two vertices (μCs) are connected to each other by a path (Diestel, 2010). Connected components having less than 3 μCs are discarded. The remaining components are the weighted graphs of clustered μCs detected by the first stage (see Fig. 8).

2.3. Classification stage

The classification of the detected clusters aims at discriminating between abnormal and non-abnormal clusters, the latter including both benign clusters of μCs and erroneously detected clusters. As previously stated, again this decision turns into an unbalanced problem and thus we employ again a RankBoost classifier. In this case, the classifier is trained on a set of true positive and false positive clusters found by the initial detection stage and works on a novel set of features especially aimed at capturing the topological relations between μCs. For each cluster detected, 35 individual μC features and 5 cluster features are extracted. The features employed are reported in Table 1 and can be grouped into the following categories: (i) *shape* features to describe μC heterogeneous morphology (e.g., round; oval; elongated) and size; (ii) *topological* features aimed at capturing topological relations between μCs from the weighted graphs associated to clusters; (iii) *score* features which use scores of μCs (see Fig. 5); (iv) *texture* features obtainable from both the original and preprocessed raw images, such as μC pixel values, contrast and difference between linear attenuation coefficients.

Individual μC features are computed for all the μCs belonging to a cluster. From these values, similarly to the work of Veldkamp and Karssemeijer (1999), four global features are computed to describe the cluster: (i) the mean value; (ii) the standard deviation; (iii) the minimum value; and (iv) the maximum value. As a result, we have a total number of $4 \times 35 + 5 = 145$ features employed to describe a cluster.

Feature descriptions and formulae to compute them are reported in Table 1. In the following paragraphs we give more details on the features that need further description.

2.3.1. μC normalized degree feature

In graph theory, the degree (denoted by δ) of a vertex of a graph is the number of edges incident to the vertex (Diestel, 2010). The distribution of this feature in a detected cluster describes its connectivity, which is expected to be higher for true positive clusters. We compute the measure $\delta(\mu_i)$ for each microcalcification μ_i in a

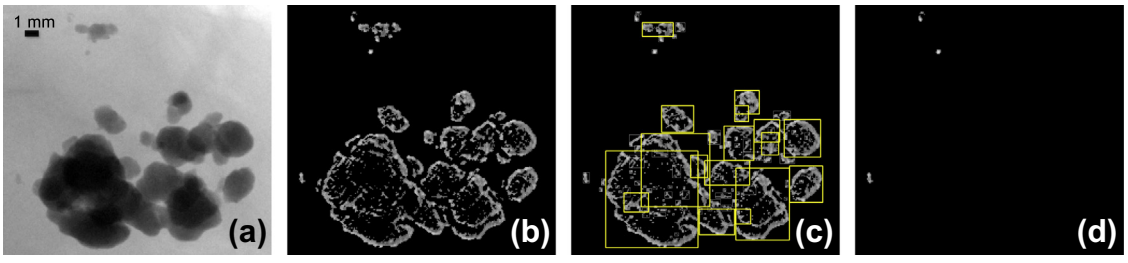

Fig. 7. An example of macrocalcifications removal from the probability map. After the probability map (b) has been obtained from the raw mammogram image (a), connected components and their minimum bounding rectangles are computed (c). Components whose bounding rectangles have both width and height greater than 1 mm are discarded along with the smaller components whose bounding boxes intersect with them (c and d). Image (a) has been modified to improve its readability.

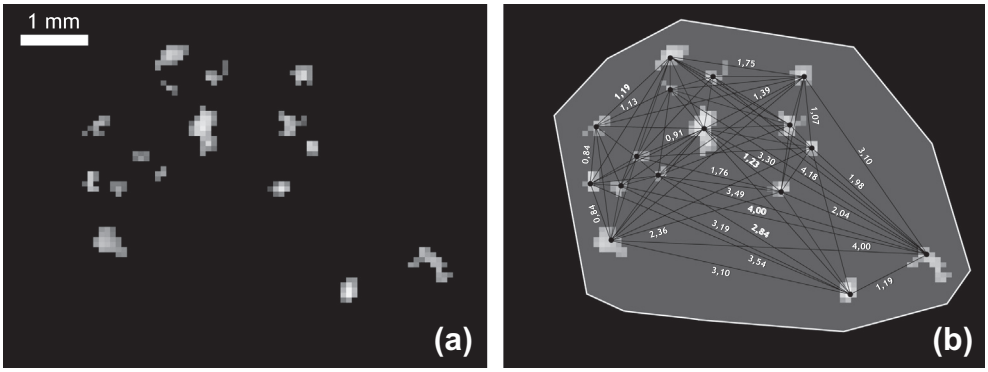

Fig. 8. A cluster of μCs (a) and the corresponding weighted graph (b) used for the extraction of topological cluster features. Vertexes correspond to the μCs and edges weights are the μC distances (in mm).

Table 1
Features used for the classification of clustered μCs.

Type	Category	Feature (symbol)	Description
Individual μC	Shape	Mc perimeter (p)	The number of pixel sides that touch a background pixel
		Mc area ($A_{\mu C}$)	The number of pixels belonging to the segmented μC
		Mc compactness	$\frac{p^2}{4\pi A_{\mu c}}$
		Mc eccentricity ($\varepsilon_{\mu C}$)	$\frac{I_{xx}+I_{yy}-\sqrt{(I_{xx}-I_{yy})^2+4I_{xy}^2}}{I_{xx}+I_{yy}+\sqrt{(I_{xx}-I_{yy})^2+4I_{xy}^2}}$ where I_{xx}, I_{xy} and I_{yy} are the moments of inertia
		Mc hu moments × 7	See Hu, 1963
		Mc thickness	Width of the best fitting rectangle
	Topological	Mc distance centroid	distance to the cluster centroid
		Mc distance closest	Distance to the closest μC
		Mc distance boundary	Distance to the skin-air boundary
		Mc degree	The number of edges incident to the μC
		Mc normalized degree	The sum of the normalized weights of the edges incident to the μC
	Probability	Mc probability × 4	Maximum, second maximum, mean and standard deviation of the probability map values of the μC
	Texture	Mc background × 2	Mean and standard deviation of background pixel values
		Mc foreground × 2	Mean and standard deviation of μC pixel values
		Mc contrast × 5	Maximum, mean, standard deviation, kurtosis and skewness of $C_i = \log(y_i) - \log(y_b)$ where y_i are μC pixel values and y_b are background pixel values, both obtained from the original raw mammogram image
		Mc attenuation × 5	Maximum, mean, standard deviation, kurtosis and skewness of $\Delta\mu = \mu_b - \mu_{\mu C} = \frac{C_i}{d_i}$ where d_i is μC thickness at location i
Cluster	Shape	Cls area (A_c)	Area of the cluster
		Cls eccentricity (ε_c)	Eccentricity of the cluster (see $\varepsilon_{\mu C}$)
	Topological	Cls number (n)	The number of μCs in the cluster
		Cls coverage	$\sum_{i=1}^{n} \frac{A_{\mu C_i}}{A_c}$
		Cls density	$\frac{2\lvert E\rvert}{n(n-1)}$ where E is the number of edges of the graph associated to the cluster

detected cluster on its associated weighted graph obtained in the clustering step (see Section 2.2.4). However, such measure does not take into account the inter-μC distances and then it is not able to accurately describe the cluster density nor the contribution given by each μC to form a dense cluster. For this reason, we introduce a further suitable measure defined as follows. Let be $\mathcal{D}(\mu_i)$ the set of microcalcifications connected to μ_i and $w(\mu_i, \mu_j)$ the weights of the edges (μ_i, μ_j) $\forall \mu_j \in \mathcal{D}(\mu_i)$, corresponding to the inter-μC distances. The *μC normalized degree* (denoted by δ_{norm}) of each microcalcification μ_i is:

$$\delta_{norm}(\mu_i) = \sum_{\mu_j \in \mathcal{D}(\mu_i)} 1 - \frac{w(\mu_i, \mu_j)}{d_{max}} \tag{13}$$

This feature aims at capturing three pieces of information: (i) the number of neighboring μCs; (ii) how close the neighboring μCs are to the current μC; (iii) the contribution given by the current μC to form a dense cluster.

2.3.2. μC contrast features

By applying the imaging model described by van Engeland et al. (2006) for digital mammograms it turns out that the exposure I_{mc} of an X-ray beam intersecting a μC can be approximated by

$$I_{mc} \approx I_0 \exp[-\mu_b(d - d_i^{mc}) - \mu_{mc} d_i^{mc}] \tag{14}$$

whereas for an X-ray beam I_b intersecting background tissue

$$I_b \approx I_0 \exp[-\mu_b d] \tag{15}$$

with I_0 the exposure of the incident X-ray beam, d the breast thickness at the current location, d_i^{mc} the μC thickness at location i, μ_{mc} and μ_b the effective linear attenuation coefficients for the μC and the background tissue, respectively. Since in an unprocessed FFDM pixel values are proportional to the total exposure I, we can replace exposure value I with pixel value y and, after some algebraic manipulation, we obtain:

$$\log(y_i^{mc}) - \log(y_b) = d_i^{mc}(\mu_b - \mu_{mc}) \tag{16}$$

where y_i^{mc} is the intensity of the pixel i in the μC and y_b is an estimate of the background tissue intensity around the μC. Since the expression in Eq. (16) is approximately independent of breast thickness and exposure level, it is a good candidate feature to describe μCs. Thus, we define the *μC contrast* C_i^{mc} for each pixel i in the μC as:

$$C_i^{mc} = \log(y_i^{mc}) - \log(y_b) \tag{17}$$

In particular, for each μC, we extract 5 measures from the distribution of C_i^{mc} on the pixels of the lesion: maximum, mean, standard deviation, kurtosis and skewness.

2.3.3. μC attenuation features

Another interesting feature would be given by the attenuation coefficient μ_{mc} which is expected to characterize the points belonging to a μC. Looking at Eq. (16), we can easily obtain an estimate of the difference $\Delta\mu_i$ between the linear attenuation coefficients of the μC at pixel i and the background tissue:

$$\Delta\mu_i = \mu_b - \mu_{mc} = \frac{\log(y_i^{mc}) - \log(y_b)}{d_i^{mc}} \tag{18}$$

The thickness d_i^{mc} of the μC at point i is estimated by means of the distance transform (Rosenfeld and Pfaltz, 1966) applied to the segmented μC. Also in this case we extract 5 measures from the distribution of $\Delta\mu_i$ on the pixels of the lesion: maximum, mean, standard deviation, kurtosis and skewness.

3. Performance evaluation

3.1. Mammogram datasets

Three datasets have been extracted from a database containing more than 40,000 cases obtained in routine screening in the Preventicon screening center in Utrecht (The Netherlands): (i) dataset *A*, consisting of 252 images from 129 abnormal cases (70 benign and 59 malignant) extracted by selecting all the available images with individual μCs manually labeled by an experienced reader (7758 in total); (ii) dataset *B*, consisting of 447 images from 242 abnormal cases (186 benign and 56 malignant) extracted by selecting all the available images with clusters of μCs manually labeled by an experienced reader (457 in total); and (iii) dataset *C*, consisting of 1152 images from 318 normal cases, i.e. patients that have not been recalled by radiologists. Malignant findings on the mammograms have been proven to be cancer through a biopsy. All the images were acquired with a Hologic Selenia full field direct digital mammography system with a resolution of 70 μm per pixel and 12-bit grayscale pixel depth.

3.2. Performance evaluation study

CasCADe's performance has been evaluated according to the following procedure: (i) the preprocessing stage was applied to all mammograms; (ii) positive training set $\mathcal{P}$ (containing μCs) and negative training set $\mathcal{N}$ (containing patches of background tissue) were extracted from dataset *A*; (iii) the ranking-based cascade μC detector was trained with $\mathcal{P}$ and $\mathcal{N}$; (iv) the cluster classification stage was trained and tested by means of 10-fold cross validation and bootstrapping on the set $B \cup C$. In order to characterize CasCADe's performance on biopsy proven malignant cases, the final step was repeated using the subset $B_1 \subset B$ containing only the malignant cases of *B*. Each step is detailed in the following paragraphs.

3.2.1. Extracting the training set for the ranking-based cascade μC detector

The training set was extracted from the 252 mammograms of dataset *A* as follows. For each μC location marked in a mammogram, a subwindow of $M \times M$ pixels centered on it was extracted; the vector formed by the subwindows extracted from all the training images was the positive training sample set $\mathcal{P}$ and contained 7758 samples. We initially chose $M = 14$ because μCs in mammograms may typically be embedded in a 1 mm-sided subwindow, corresponding to 14 pixels-sided subwindows on our mammograms. Our final choice was $M = 12$ because it exhibited the same detection performance obtained with higher values for *M*, but requiring a lower computational load. The same training images were used to extract negative samples from all the breast regions remaining after the removal of μCs-subwindows. From these regions, overlapping subwindows of 12×12 pixels were extracted to form the negative training sample set $\mathcal{N}$. Depending on the overlap amount, $|\mathcal{N}|$ ranged from 6, 221, 211 (no overlap) to 900, 875, 720 (maximum overlap) with class imbalance ratio $k = |\mathcal{N}|/|\mathcal{P}|$ ranging from about 10^3 to 10^5. We chose overlap equal to 50% thus obtaining $|\mathcal{N}| = 25,190,476$ and class imbalance ratio $k \approx 3 \times 10^3$. This allowed us to obtain a cascade detector composed by a not too high number of nodes, thus maintaining high the overall detection rate *D*.

3.2.1.1. Training the ranking-based cascade μC detector.

The cascade was built using $d_i = 0.99$ and $f_i = 0.30$, and trained with sample sets $\mathcal{P}$ and $\mathcal{N}$ extracted from dataset *A*. The reduced class imbalance ratio k_r was chosen equal to $k_r = 5$ for each node. The cascade so obtained was composed by 5 nodes employing respectively 2, 3, 5, 12, 40 features automatically selected from the 14,709 possible ones (see Fig. 9). The overall detection and false positive rate obtained by the cascade on the validation set were $D \approx 0.96$ and $F \approx 9 \times 10^{-4}$, respectively.

3.2.1.2. Performance evaluation with cross validation and bootstrapping.

To evaluate CasCADe's overall performance, we performed 10-fold cross validation on B (B_1) and C using clusters detected in the step previously described. In each cross validation step, a RankBoost classifier composed by 100 weak learners was trained on the clusters detected in 90% of the cases and tested on the clusters detected in the other 10%. When splitting the data into a training and test set, the images belonging to the same case were assigned to the same set. In order to remove the effect of case distributions in the cross validation dataset, we used the following bootstrapping procedure. Cases were sampled with replacement from the complete cross validation dataset 5000 times. Each new set of sampled cases contained the same number of cases as the original set. For each resampling, a lesion-based FROC curve and a case-based ROC curve were obtained for a series of thresholds on the abnormality score. Finally, the FROC and ROC curves resulting from the 5000 bootstrap samples were averaged on the sensitivity axis. For the lesion-based FROC curves, true and false positives were counted with a method partially inspired by Jing et al. (2011): (i) an annotated cluster *t* was hit by a CasCADe finding *f* if at least 2 μCs of *f* were within *t*; (ii) an annotated cluster *t* was a *true positive* if it was hit at least once by a CasCADe finding *f*; and (iii) a CasCADe finding *f* was a *false positive* if it did not hit any annotated cluster *t*. For the case-based ROC curves, true and false positives were counted as follows: (i) an abnormal case was a *true positive* if at least one CasCADe finding hit one of the annotated clusters in the two views and (ii) a normal case was a *false positive* if at least one CasCADe finding was present in any of the two views.

Fig. 9. The features automatically selected by the first 3 nodes of the trained ranking-based cascade μC detector. (a) The 2 features selected by the first node. (b) The 3 features selected by the second node. (c) The 5 features selected by the third node.

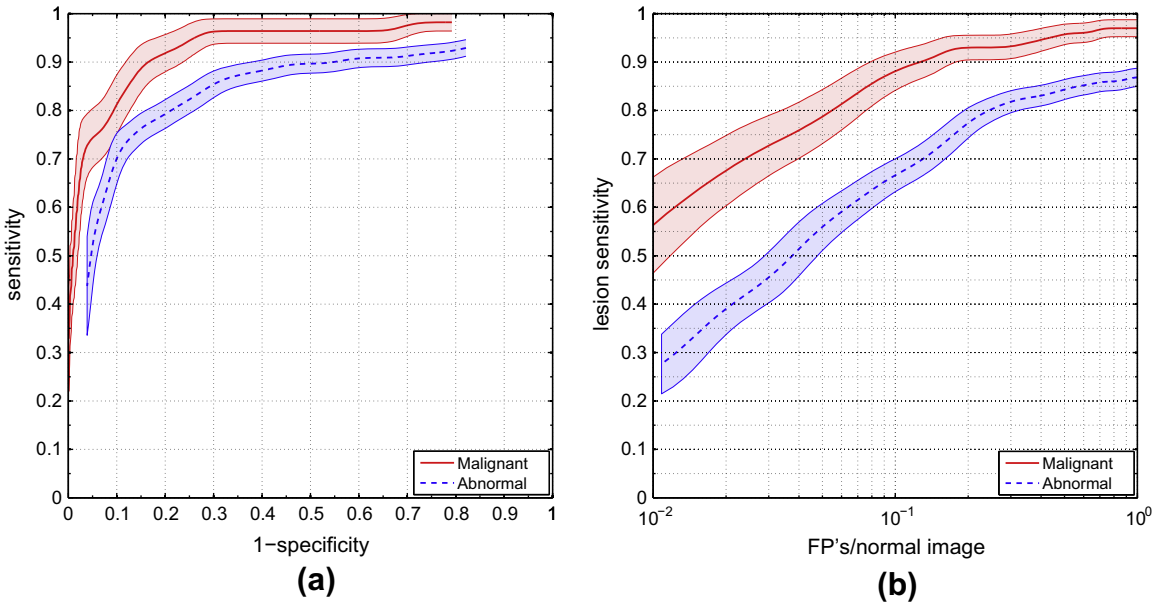

Fig. 10. Average case-based ROC curves (a) and lesion-based FROC curves and (b) of CasCADe obtained from 5000 bootstrap samples. Confidence bands indicate the standard deviation along the sensitivity axis.

4. Results and discussions

4.1. Detection performance

CasCADe's overall detection performance is summarized in Fig. 10 by average case-based ROC curves and lesion-based FROC curves. Case-based ROC curves plot the true positive fraction of abnormal cases (or malignant cases when substituting B with B_1) vs. the false positive fraction of normal cases. Lesion-based FROC curves plot the true positive fraction of abnormal lesions (or malignant lesions when substituting B with B_1) vs. the average number of false positive lesions per normal image.

The graphs show that the overall detection performance is higher when considering only malignant cases. A first reason is the different sensitivity of the initial detection stage that seems more effective on malignant lesions. Secondly, the features employed in the classification stage were more discriminative for malignant lesions (AUC = 0.97) than abnormal lesions (AUC = 0.93), which might include also benign lesions.

The ranking of the 10 most selected features in the classification stage is reported in Table 2. It suggests that topological features played an important role for both the classification tasks (i.e. malignant vs. normal, and abnormal vs. normal). In particular, the proposed *μC normalized degree* feature was one of the three most selected during training while in the testing phase it revealed to be the most discriminating feature. A possible reason is that among the lesions detected by CasCADe on the normal cases, most of them actually were not clusters of μCs, thus being easily recognized as negatives on the base of their topological properties. In addition, we ranked the false positives detected by CasCADe on the normal cases using their abnormality scores provided by the classification stage. All the first 50 false positive clusters were actually benign clusters (most of them with vascular calcifications), whereas all the last 50 ones were not clusters of μCs. This could be profitably used to recognize vascular calcifications which are important indicators of cardiovascular risk factors (Iribarren et al., 2004).

Finally, it is worth noting that the time necessary to completely process a mammogram was about 3 s, whereas the training times for the detection and the classification stage were about 10 min and 30 min, respectively, including the time needed for feature computation. Such results were obtained on a workstation equipped with 96 GB of RAM and 2 quad-core CPUs at 2.26 GHz. The whole system was implemented in C++.

4.2. Comparison with R2CAD

We compared the proposed CasCADe system with one of the most widespread commercial CADe systems, the Hologic R2CAD ImageChecker[1] on the detection of biopsy-proven malignant clusters of μCs by following a bootstrapping methodology suited for

[1] We used reports of R2CAD already present in datasets B_1 and C. All the reports were obtained with the following parameters: (i) algorithm version "R2 mammo calc 8.1"; (ii) configuration "[A] - high sensitivity", which was the highest sensitivity setting allowed by the system; (iii) "maximum inter-mc distance" equal to 10 mm, that is the same value we used for our detector (see Section 2.2.4).

Table 2
The first 10 most selected features in the cluster classification stage.

Datasets	Frequency (%)	Category	Feature
B vs. *C* (abnormal vs. normal)	6.73	Topological	Mean (mc normalized degree mean)
	5.84	Texture	Mean (mc background mean)
	3.86	Topological	Mean (mc distance closest)
	3.76	Texture	Mean (mc background stdev)
	3.56	Topological	Max (mc distance boundary)
	3.47	Topological	Min (mc distance closest)
	3.17	Shape	Mean (mc perimeter)
	2.87	Texture	Min (mc background stdev)
	2.87	Shape	Mean (mc thickness)
	2.77	Topological	Min (mc distance boundary)
B_1 vs. *C* (malignant vs. normal)	6.83	Topological	Mean (mc distance closest)
	4.65	Texture	Mean (mc background stdev)
	4.55	Topological	Mean (mc normalized degree mean)
	4.55	Topological	Stdev (mc distance boundary)
	4.16	Probability	Mean (mc probability max)
	3.37	Texture	Max (mc background mean)
	3.27	Topological	Min (mc distance closest)
	2.77	Topological	Mean (mc normalized degree stdev)
	2.77	Texture	Mean (mc contrast skewness)
	2.77	Texture	Min (mc attenuation stdev)

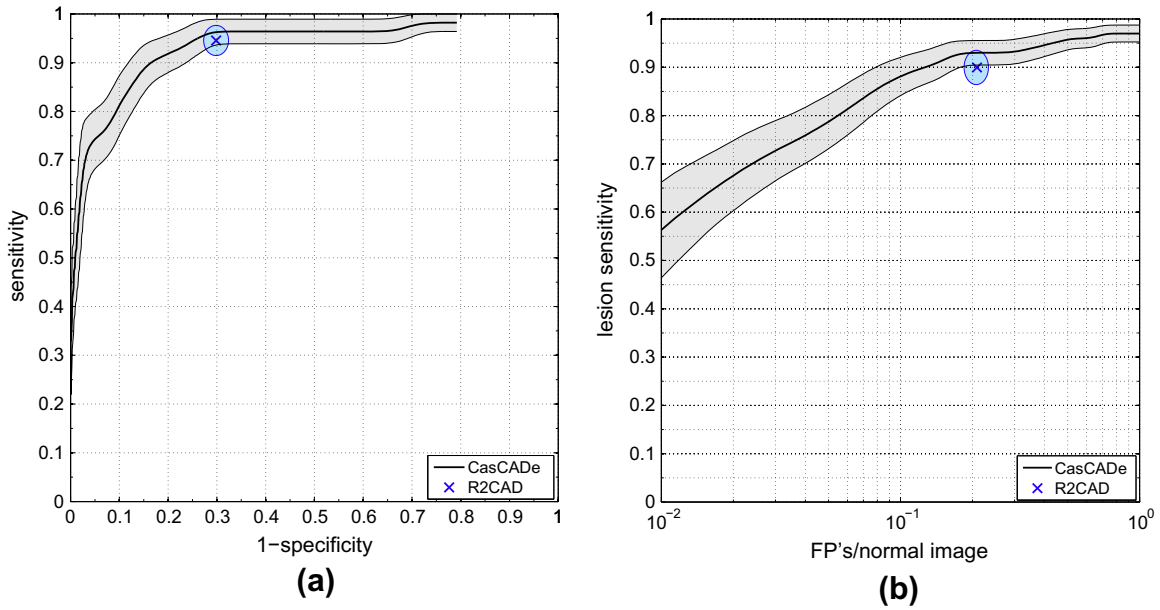

Fig. 11. Average case-based ROC curves (a) and lesion-based FROC curves and (b) of CasCADe and R2CAD obtained from 5000 bootstrap samples. For CasCADe, confidence bands indicate the standard deviation along the sensitivity axis. For R2CAD, the marks denote the average operating points and the ellipses indicate the standard deviation along both the axes.

the comparison between CAD systems (Samuelson and Petrick, 2006; Samuelson et al., 2007). Cases were sampled with replacement from $B_1 \cup C$ 5000 times. Each new set of sampled cases contained the same number of cases as the original set. For each resampling, case-based ROC curves and lesion-based FROC curves were constructed for both CasCADe and R2CAD according to the criterion described

Table 3
Performance differences between CasCADe and R2CAD.

Measure	Definition[a]	Value	*p*-Value
ΔTPFc	$TPFc^C - TPFc^R$, being TPFc the true positive fraction of malign cases	1.7%	0.2588
ΔSPEC	$SPEC^C - SPEC^R$, being SPEC the specificity of normal cases	1.8%	0.1764
ΔTPFl	$TPFl^C - TPFl^R$, being TPFl the true positive fraction of malign lesions	3.9%	0.1122
ΔFPpi	$FPpi^R - FPpi^C$, being FPpi the number of false positives per image	0.08	0.0888

[a] Measures of CasCADe and R2CAD are denoted by C and R superscripts, respectively.

in Section 3.2.1.2. For CasCADe, abnormality scores from the 10-fold cross validation step were used to construct ROC and FROC curves, whereas for R2CAD only one ROC point and one FROC point could be obtained, since no abnormality scores were present in the R2CAD reports. Subsequently, four performance differences were computed as reported in Table 3. Resampling 5000 times resulted in 5000 values for each performance difference. *P*-values were defined as the fraction of the corresponding performance difference that were negative or zero. Performance differences were considered significant if $p < 0.05$. For the sake of clarity we also report in Fig. 11 the average ROC and FROC curves of CasCADe together with the average operating points of R2CAD.

The results show that CasCADe compared favorably both in sensitivity and in specificity with R2CAD. In particular, at the same lesion sensitivity of R2CAD (90%) on biopsy proven malignant cases, the false positives per image (FPpi) decreased from 0.21 to 0.13 (p-value = 0.09), whereas at the same FPpi of R2CAD (0.21), the lesion sensitivity increased from 90% to 93% (p-value = 0.11). Even though the performance differences were not statistically significant, to our knowledge there are no other CADe systems presented in scientific literature providing evidence that they can compete with high-end commercial systems.

5. Conclusions

In this work we described CasCADe, a computer-aided detection system of clustered μCs on FFDM, that is based on an ensemble of ranking-based boosting classifiers structured as a cascade. Both the cascade structure and the choice of ranking instead of classification error as learning objective let us face effectively the high class imbalance between μC- and nonμC-patches. An additional stage of classification employing a novel set of features especially aimed at capturing the topological properties of the detected clusters drastically reduced the number of false positives while preserving the detected abnormalities. The proposed approach was evaluated with a dataset of 1599 full-field digital mammograms from 560 cases. Compared with Hologic R2CAD ImageChecker, one of the most widespread commercial CADe systems, CasCADe achieved an increase of about 3.9% in the malignant lesion sensitivity (p-value = 0.11) and a decrease from 0.21 to 0.13 in the number of false positive lesions per image (p-value = 0.09), thus showing that it can compete with high-end CADe commercial systems.

References

Barandela, R., Sánchez, J.S., García, V., Rangel, V., 2003. Strategies for learning in class imbalance problems. Pattern Recognition 36, 849–851.

Barbu, A., Suehling, M., Xu, X., Liu, D., Zhou, S., Comaniciu, D., 2012. Automatic detection and segmentation of lymph nodes from CT data. IEEE Transactions on Medical Imaging 31, 240–250.

Beutel, J., Kundel, H.L., VanMetter, R.L., 2000. Handbook of Medical Imaging: Physics and Psychophysics, vol. 1. S.P.I.E., Bellingham, WA.

Birdwell, R.L., 2009. The preponderance of evidence supports computer-aided detection for screening mammography. Radiology 253, 9–16.

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

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

Cole, E., Zhang, Z., Marques, H., Nishikawa, R., Hendrick, R., Yaffe, M., Padungchaichote, W., Kuzmiak, C., Chayakulkheeree, J., Conant, E., Fajardo, L., Baum, J., Gatsonis, C., Pisano, E., 2012. Assessing the stand-alone sensitivity of computer-aided detection with cancer cases from the digital mammographic imaging screening trial. AJR American Journal of Roentgenology 199, 392–401.

Cortes, C., Mohri, M., 2004. AUC optimization vs error rate minimization. Advances in Neural Information Processing Systems, 16.

Diestel, R., 2010. Graph Theory, . Graduate Texts in Mathematics, forth ed., vol. 173. Springer-Verlag.

Eadie, L.H., Taylor, P., Gibson, A.P., 2012. A systematic review of computer-assisted diagnosis in diagnostic cancer imaging. European Journal of Radiology 81, e70–e76.

El Naqa, I., Yang, Y, W.M.N., Galatsanos, N.P., Nishikawa, R.M., 2002. A support vector machine approach for detection of microcalcifications. IEEE Transactions on Medical Imaging 21, 1552–1563.

Feulner, J., Zhou, S.K., Hammon, M., Hornegger, J., Comaniciu, D., 2013. Lymph node detection and segmentation in chest CT data using discriminative learning and a spatial prior. Medical Image Analysis 17, 254–270.

Freund, Y., Iyer, R., Schapire, R.E., Singer, Y., 2003. An efficient boosting algorithm for combining preferences. Journal of Machine Learning Research 4, 933–969.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F., 2011. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, Cybernetics C: Applied Review 42, 463–484.

Guo, X., Yin, Y., Dong, C., Yang, G., Zhou, G., 2008. On the class imbalance problem. In: 4th International Conference on Natural Computation, vol. 4, pp. 192–201.

Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36.

Hu, M.K., 1963. Visual pattern recognition by moment invariants. IRE Transactions on Information Theory 8, 179–187.

Huang, J., Ling, C., 2005. Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on Knowledge and Data Engineering 17, 299–310.

Iribarren, C., Go, A.S., Tolstykh, I., Sidney, S., Johnston, S.C., Spring, D.B., 2004. Breast vascular calcification and risk of coronary heart disease, stroke, and heart failure. Journal of Women's Health 13, 381–389.

Jing, H., Yang, Y., Nishikawa, R.M., 2011. Detection of clustered microcalcifications using spatial point process modeling. Physics in Medical Biology 56, 1–17.

Karssemeijer, N., Bluekens, A.M., Beijerinck, D., Deurenberg, J.J., Beekman, M., Visser, R., van Engen, R., Bartels-Kortland, A., Broeders, M.J., 2009. Breast cancer screening results 5 years after introduction of digital mammography in a population-based screening program. Radiology 253, 353–358.

Kopans, D.B., 2007. Breast Imaging. Lippincott Williams & Wilkins, Hagerstown, Maryland, USA.

Lienhart, R., Kuranov, E., Pisarevsky, V., 2003. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In: DAGM 25th Pattern Recognition Symposium, pp. 297–304.

Li, D.C., Liu, C.W., Hu, S.C., 2010. A learning method for the class imbalance problem with medical data sets. Computers in Biology and Medicine 40, 509–518.

Marrocco, C., Molinara, M., D'Elia, C., Tortorella, F., 2010. A computer-aided detection system for clustered microcalcifications. Artificial Intelligence in Medicine 50, 23–32.

McLoughlin, K.J., Bones, P.J., Karssemeijer, N., 2004. Noise equalization for detection of microcalcification clusters in direct digital mammogram images. IEEE Transactions on Medical Imaging 23, 313–320.

Niemeijer, M., van Ginneken, B., Cree, M., Mizutani, A., Quellec, G., Sanchez, C., Zhang, B., Hornero, R., Lamard, M., Muramatsu, C., Wu, X., Cazuguel, G., You, J., Mayo, A., Li, Q., Hatanaka, Y., Cochener, B., Roux, C., Karray, F., Garcia, M., Fujita, H., Abramoff, M., 2010. Retinopathy online challenge: Automatic detection of microaneurysms in digital color fundus photographs. IEEE Transactions on Medical Imaging 29, 185–195.

Oliver, A., Torrent, A., Tortajada, M., Lladó, M.P., Tortajada, L., Sentís, M., Freixenet, J., 2010. A boosting based approach for automatic micro-calcification detection. In: Proceedings of International Workshop on Digital Mammography, LNCS, vol. 6136, pp. 251–258.

Rosenfeld, A., Pfaltz, J., 1966. Sequential operations in digital picture processing. Journal of the ACM 13, 471–494.

Samuelson, F.W., Petrick, N., 2006. Comparing image detection algorithms using resampling. In: IEEE International Symposium on Biomedical Imaging, pp. 1312–1315.

Samuelson, F.W., Petrick, N., Paquerault, S., 2007. Advantages and examples of resampling for CAD evaluation. In: IEEE International Symposium on Biomedical Imaging, pp. 492–495.

Schie, G., Karssemeijer, N., 2008. Detection of microcalcifications using a nonuniform noise model. In: Proceedings of the 9th International Workshop on Digital Mammography, pp. 378–384.

Tang, J., Rangayyan, R.M., Xu, J., El Naqa, I., Yang, Y., 2009. Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. IEEE Transactions on Information Technology in Biomedical 13, 236–251.

van Engeland, S., Snoeren, P.R., Huisman, H., Boetes, C., Karssemeijer, N., 2006. Volumetric breast density estimation from full-field digital mammograms. IEEE Transactions on Medical Imaging 26, 273–282.

van Ginneken, B., III, S.G.A., de Hoop, B., van Amelsvoort-van de Vorst, S., Duindam, T., Niemeijer, M., Murphy, K., Schilham, A., Retico, A., Fantacci, M.E., Camarlinghi, N., Bagagli, F., Gori, I., Hara, T., Fujita, H., Gargano, G., Bellotti, R., Tangaro, S., Bolaños, L., Carlo, F.D., Cerello, P., Cheran, S.C., Torres, E.L., Prokop, M., 2010. Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study. Medical Image Analysis 14, 707–722.

Van Ravesteijn, V., van Wijk, C., Vos, F., Truyen, R., Peters, J., Stoker, J., van Vliet, L., 2010. Computer-aided detection of polyps in CT colonography using logistic regression. IEEE Transactions on Medical Imaging 29, 120–131.

Veldkamp, W., Karssemeijer, N., 1996. Influence of segmentation on classification of microcalcifications in digital mammography. In: Proceedings of the 18th Annual International Conference of the IEEE, Engineering in Medicine and Biology Society, Bridging Disciplines for Biomedicine, pp. 1171–1172.

Veldkamp, W.J.H., Karssemeijer, N., 1999. An improved method for detection of microcalcification clusters in digital mammograms. Proceedings of the SPIE Medical Imaging: Image Processing, 3661.

Viola, P., Jones, M., 2001. Robust real-time object detection. International Journal of Computer Vision 57, 137–154.

Viola, P., Jones, M., 2002. Fast and robust classification using asymmetric adaboost and a detector cascade. Advances in Neural Information Processing System 16, 1311–1318.

Wei, L., Yang, Y., Nishikawa, R.M., Wernick, M.N., Edwards, A., 2005. Relevance vector machine for automatic detection of clustered microcalcifications. IEEE Transactions on Medical Imaging 24, 1278–1285.

Zhang, X., Gao, X., Wang, M., 2009. MCs detection approach using bagging and boosting based twin support vector machine. In: IEEE International Conference on Systems, Man and Cybernetics, pp. 5000–5505.