Article 25fa pilot End User Agreement

This publication is distributed under the terms of Article 25fa of the Dutch Copyright Act (Auteurswet) with explicit consent by the author. Dutch law entitles the maker of a short scientific work funded either wholly or partially by Dutch public funds to make that work publicly available for no consideration following a reasonable period of time after the work was first published, provided that clear reference is made to the source of the first publication of the work.

This publication is distributed under The Association of Universities in the Netherlands (VSNU) ‘Article 25fa implementation’ pilot project. In this pilot research outputs of researchers employed by Dutch Universities that comply with the legal requirements of Article 25fa of the Dutch Copyright Act are distributed online and free of cost or other barriers in institutional repositories. Research outputs are distributed six months after their first online publication in the original published version and with proper attribution to the source of the original publication.

You are permitted to download and use the publication for personal purposes. Please note that you are not allowed to share this article on other platforms, but can link to it. All rights remain with the author(s) and/or copyrights owner(s) of this work. Any use of the publication or parts of it other than authorised under this licence or copyright law is prohibited. Neither Radboud University nor the authors of this publication are liable for any damage resulting from your (re)use of this publication.

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please contact the Library through email: copyright@ubn.ru.nl, or send a letter to:

University Library
Radboud University
Copyright Information Point
PO Box 9100
6500 HA Nijmegen

You will be contacted as soon as possible.
Improving mass candidate detection in mammograms via feature maxima propagation and local feature selection

Jaime Melendez, a) Clara I. Sánchez, Bram van Ginneken, and Nico Karssemeijer
Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine,
Radboud University Medical Center, Nijmegen, 6525 GA, The Netherlands

(Received 5 February 2014; revised 10 June 2014; accepted for publication 15 June 2014; published 11 July 2014)

Purpose: Mass candidate detection is a crucial component of multistep computer-aided detection (CAD) systems. It is usually performed by combining several local features by means of a classifier. When these features are processed on a per-image-location basis (e.g., for each pixel), mismatching problems may arise while constructing feature vectors for classification, which is especially true when the behavior expected from the evaluated features is a peaked response due to the presence of a mass. In this study, two of these problems, consisting of maxima misalignment and differences of maxima spread, are identified and two solutions are proposed.

Methods: The first proposed method, feature maxima propagation, reproduces feature maxima through their neighboring locations. The second method, local feature selection, combines different subsets of features for different feature vectors associated with image locations. Both methods are applied independently and together.

Results: The proposed methods are included in a mammogram-based CAD system intended for mass detection in screening. Experiments are carried out with a database of 382 digital cases. Sensitivity is assessed at two sets of operating points. The first one is the interval of 3.5–15 false positives per image (FPs/image), which is typical for mass candidate detection. The second one is 1 FP/image, which allows to estimate the quality of the mass candidate detector’s output for use in subsequent steps of the CAD system. The best results are obtained when the proposed methods are applied together. In that case, the mean sensitivity in the interval of 3.5–15 FPs/image significantly increases from 0.926 to 0.958 (p < 0.0002). At the lower rate of 1 FP/image, the mean sensitivity improves from 0.628 to 0.734 (p < 0.0002).

Conclusions: Given the improved detection performance, the authors believe that the strategies proposed in this paper can render mass candidate detection approaches based on image location classification more robust to feature discrepancies and prove advantageous not only at the candidate detection level, but also at subsequent steps of a CAD system. © 2014 American Association of Physicists in Medicine. [http://dx.doi.org/10.1118/1.4885995]

Key words: computer-aided detection (CAD), mass candidate detection, feature maxima propagation, local feature selection

1. INTRODUCTION

It is well known that assessment of screening mammography is a complex task. Thus, in order to assist radiologists during breast cancer screening, computer-aided detection (CAD) systems are being developed. In the majority of cases, the sought abnormalities are either microcalcifications or masses, being the latter the most difficult to detect due to their large variation in size and shape and their poor image contrast. Therefore, considerable effort is devoted to improving CAD systems for detecting malignant masses as is the case of the current study.

CAD systems dealing with mass detection usually comprise multiple steps.1–3 In the initial step, a set of mass candidate locations is determined based on local image features. These features are typically processed by a classifier, which assigns a mass likelihood score to each image location. Then, those locations associated with a high score are identified as mass candidates and segmented. By adjusting the threshold applied to the classifier’s output, the sensitivity of the initial step can be varied. It is common to set this sensitivity to a relatively high value in such a way that most potential abnormalities are detected. However, due to misinterpretation of normal structures, the number of false-positive detections is relatively high as well. Therefore, in the following steps, the candidate regions are further processed, usually by considering a more elaborated set of features. Since the performance of the subsequent steps strongly depends on the quality of the candidate regions, the initial detection step may be regarded as the key in the complete system. Furthermore, the masses missed during this step cannot be recovered later. Ideally, the initial detection step should reach its maximum sensitivity with the minimum number of false positives. In this way, the subsequent steps could focus on a more specific problem and achieve better solutions, as they would not be influenced by irrelevant patterns that could be filtered during the initial step.

Features used for mass detection can be related to gray-level intensities, local texture, or morphological measures.4 In addition, it is common to take the distribution of spicules associated with masses into account. Spiculation is an
effective indicator of malignancy in masses, as more than half of the screening-detected cancerous masses show this pattern. However, the remaining lesions may only be detectable by the mass itself. Thus, in order to deal with the different cases, it is also common to extract gradient orientation information. The orientation of gradient vectors can provide an idea of the location of a mass’ center. Since the shape of a mass is approximately rounded, an image pixel located close to the center is expected to have most of its surrounding gradient vectors pointing to itself. As can be seen from the above, both spiculation and gradient orientation features seem to naturally encompass with each other. Therefore, including those features in a CAD system can be regarded as an important strategy for breast cancer detection.

Several approaches reported in the literature implemented spiculation and gradient features for mass detection. For instance, Kobatake et al. applied the Iris filter, which measures the convergence of gradient vectors at a given pixel, in order to enhance rounded opacities. Later, they extracted morphological line-skeletons for spicule enhancement. Karssemeijer and te Brake developed a method based on analysis of local gradient concentration and line orientation after image filtering with a set of first- and second-order directional Gaussian derivatives. In related work, Karssemeijer introduced an efficient approach to feature computation as a continuous function of spatial scale. Zwiggelaar et al. modeled the center of a mass and the surrounding pattern of linear structures using a directional recursive median filter and a multiscale directional line detector, respectively. Liu et al. detected spiculated lesions by means of a top-down multisolution scheme based on linear phase nonseparable wavelets. Sampat and Bovik applied the Radon transform for spicule enhancement and then detected the convergence location with a set of radial spiculation filters. Wei et al. computed the average gradient direction over neighboring concentric annular regions around a pixel to identify regions of interest. Sakellaropoulos et al. extracted intensity and gradient orientation features after wavelet decomposition in order to detect masses in dense parenchyma. Mencattini and Salmeri evaluated the eigenvalues of the Hessian matrix of the gradient image in order to locate “wells,” which can be associated with masses. Other related approaches can be found in the review by Oliver et al.

Most of the aforementioned features are designed to attain their maximum response at the center of a mass. Therefore, the task of labeling image locations as normal or abnormal corresponds to segregating the peaks resulting in the feature domain. This is not a simple task, thus, in order to achieve the best performance, the common approach is to combine several features by means of elaborate classification techniques. However, due to the large variability of the processed images, the theoretical properties expected from these features are not always observed and problems during classification may arise. In this study, we identified two of those problems:

1. The maxima among several features that characterize a given mass may not occur at the same image location. Since feature vectors for classification at this point are constructed on a per-image-location basis, these maxima may never be part of the same feature vector, which is undesirable, as it dims the discrimination power of the combined features.

2. Although the ideal output of these features is a clear well-defined peak at the center of a mass, in practice, lesions may be far from the ideal model for which a particular feature was designed and show distorted or scattered patterns, or simply appear at a different scale. Consequently, the resulting output would not be clearly concentrated at a given location but span several neighbors instead. If several features are combined, differences in peak spread are likely to occur. Depending on how large these differences are and the proportion of discrepant data, the coherence of the feature vectors may be compromised and the classifier may not optimally determine the correct relations among features.

Figure 1 illustrates an example of both the discussed problems. In this figure, a mammographic image with a malignant

![FIG. 1. A mammographic image with a malignant mass (left) and two of its associated feature images obtained by means of the spiculation-based (middle) and gradient orientation-based (right) feature extractors developed by Karssemeijer and te Brake (Refs. 8–10). Features were computed according to an 8 × 8 pixel grid. Lighter gray levels indicate higher feature values. The region where the mass is located is shown magnified in order to better appreciate the differences in peak location and spread yielded by the two features.](image-url)
mass (left) and two of its associated feature images (middle and right) are shown. From the magnified crops of the feature images, a displacement of two locations can be observed between the maxima of the peaks characterizing the sought lesion. Moreover, the peak corresponding to the first feature (middle) is substantially wider than the one corresponding to the second feature (right). Thus, even if they were aligned (with respect to their maxima), a mismatch would still exist in the following sense: while the location corresponding to the maximum in the narrow peak would be paired to the maximum of the wide peak, the neighboring locations around the center of the narrow peak, which have much lower values, would still be paired to high values in the wide peak. As a result, it would be difficult to establish a coherent pattern consisting of high responses for masses.

In this paper, we propose two strategies to tackle these problems. The first one operates at the feature level and consists of propagating (and hence aligning) the feature maxima through a given neighborhood. The second one operates at the classifier level and relies on the observation that the issues mentioned above do not always occur in all the available features. Therefore, exploiting the interaction of particular subsets of features given particular subsets of samples seems an appropriate choice. This is realized by means of a random forest classifier. In addition, we investigate the effectiveness of the proposed approaches by means of a CAD application consisting of mass candidate detection in screening mammography. The features used by the CAD system are based on the gradient orientation and spiculation measures derived by Karssemeijer and te Brake. These features have been used as the basis of successful CAD systems for mammography-based cancer detection. For instance, in a recent study by Hupse et al., it has been shown that a system implementing these features at the mass candidate detection step can achieve a comparable performance (76% vs 75% sensitivity at 0.21 false-positive marks per image) as a clinically available system (R2 ImageChecker, V1.4; Hologic). The methods proposed in this study further develop the aforementioned features and lead to a clear improvement of the basic mass candidate detector utilized by Hupse et al.

The remaining of the paper is organized as follows. Section 2 provides details about the mammogram data set used in this study. Section 3 elaborates on the proposed methods and describes the experimental setup. Section 4 shows our experimental results, which are then discussed in Sec. 5. Finally, Sec. 6 concludes the paper.

2. MATERIALS

The Dutch screening program targets women between 50 and 75 years, and invites them to participate every two years as a part of a free nationwide breast cancer screening service. Further information can be found in Ref. 18. The cases utilized in this study were acquired in the course of this program in the period 2003–2008 at both the Foundation of Population Screening Mid-West in Utrecht and the Radboud University Medical Center in Nijmegen. By following the Dutch guide-

3. METHODS

3.A. Mass candidate detection in mammography-based CAD

The CAD system utilized in this study closely follows the multistep paradigm described before. Considering its single-view mode, two main steps can be identified. In the first step, a set of basic features is computed and a classifier is utilized in order to determine a set of candidate regions. In the second step, these regions are further processed by computing a richer set of features and a new classification process is carried out in order to get the final malignancy score. More details can be found in related work. Since we are interested in the mass candidate detection problem, we focused on the first step of this system. This baseline mass candidate detector (or baseline detector in short) consists in turn of three steps: image preprocessing, feature extraction, and image location classification. Figure 2 shows a flowchart of this baseline detector with its three steps, as well as the improvements proposed in this paper (Secs. 3.B and 3.C). A description of each step is given below.

3.A.1. Image preprocessing

Prior to mass detection, raw input images are preprocessed. In the first step, a given image is downsampled to a resolution of 200 μm per pixel by means of bilinear interpolation. Then, the image is segmented into breast area, pectoral muscle (if it is an MLO view) and background. Background pixels are
TABLE I. Characteristics of the mammogram data set utilized in this study.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women age range</td>
<td>50–75 years (average 59.7 ± 7.3 years)</td>
</tr>
<tr>
<td>Number of cases</td>
<td>382</td>
</tr>
<tr>
<td>Number of cancer cases</td>
<td>191</td>
</tr>
<tr>
<td>Number of normal cases</td>
<td>191</td>
</tr>
<tr>
<td>Number of images</td>
<td>1324</td>
</tr>
<tr>
<td>Number of images in cancer cases</td>
<td>716 (336 CC, 380 MLO)</td>
</tr>
<tr>
<td>Number of images with masses</td>
<td>354 (164 CC, 190 MLO)</td>
</tr>
<tr>
<td>Number of images in normal cases</td>
<td>608 (228 CC, 380 MLO)</td>
</tr>
<tr>
<td>Number of lesions (considering all views)</td>
<td>374</td>
</tr>
<tr>
<td>Lesion diameter</td>
<td>Average 14.7 ± 6.1 mm</td>
</tr>
<tr>
<td>Mammographic equipment</td>
<td>Selenia digital mammography system (Hologic)</td>
</tr>
<tr>
<td></td>
<td>Senographe Essential ADS 41.0 (GE Medical Systems)</td>
</tr>
<tr>
<td>Gray value depth</td>
<td>14 bits</td>
</tr>
</tbody>
</table>

identified by using thresholding and a sequence of morphological operators. Subsequently, the pectoral muscle is segmented from the breast area by fitting a straight line to its boundary using a modified Hough transform. Since this boundary is typically slightly curved, an optimal path search considering the previous initial estimate is performed afterwards. In the next step, the raw image is converted to a standard representation by simulating the film-based mammogram acquisition procedure. However, instead of modeling the relation between intensity values and exposure with a nonlinear characteristic curve, a piece-wise linear model is applied. This approach follows more closely the linear relation between (logarithmic) exposure and tissue thickness established by Engeland et al., therefore, it provides a better image representation. The model parameters were determined in a set of independent experiments. After image conversion, a thickness equalization algorithm is applied to enhance the periphery of the breast. A similar algorithm is used in MLO images to equalize the intensity of the pectoral muscle in order to facilitate detection of masses located on or near the pectoral boundary.

3.A.2. Feature extraction

As stated before, in this study we utilized the spiculation- and gradient-based feature extractors developed by Karssemeijer and te Brake. Spiculation characterization is based on detecting line-like patterns radiating from a central location. Line-based orientation estimates for the image pixels are obtained from the output of directional second-order Gaussian derivative operators. The orientation at which these operators have maximum response is selected. The distribution of these estimates is analyzed for each image location considering a circular neighborhood centered at that location. By imposing a selection criterion on the intensity of the filter output, a subset of pixels within this neighborhood is determined, representing potential sites of interest. To calculate the features, all pixels from this subset that are directed toward the center of the neighborhood are counted as a function of the direction in which they are located. Discrete directions are obtained by dividing the neighboring space into a number of radial bins. In order to achieve a multiscale representation, cumulative normalized counting is performed as a function of the radius. The first feature derived from this analysis, denoted as $l_1$, is related to the maximum number of pixels with directions pointing to the center of the neighborhood and gives a measure of the maximum spiculation. The radius at which this maximum occurs, denoted as $l_2$, is used as a second feature. Since pixels oriented toward a location can be found only in a few directions and, thus, it is not very likely that the site being evaluated belongs to the center of a spiculated lesion, a third feature, denoted as $l_3$, measuring in how many directions spiculation is
strong is defined. A similar procedure and reasoning is followed to compute the gradient-based features. In this case, however, first-order derivatives are utilized and only two features, denoted as \(g_1\) and \(g_2\), are defined. They are analogous to \(l_1\) and \(l_2\). The five features described so far are extracted for locations inside the tissue area sampled at 1.6 mm (every 8 pixels). This setting is a compromise in order to decrease the computational load while still being able to detect small masses.\(^7\) The resulting feature images are smoothed with a Gaussian kernel with \(\sigma = 1.6\) mm before classification.

### 3.A.3. Image location classification

Location classification is carried out by an ensemble of five neural networks. Each of these networks is a multilayer perceptron with randomly initialized weights and utilizes a different ratio \(\rho = 0.1, \ldots, 0.5\) of positive to negative training patterns. The remaining settings are shared among the networks and are configured as follows: there are five input neurons, five neurons in the single hidden layer and one neuron in the output layer. All neurons are configured with a sigmoid function. The standard backpropagation algorithm is used to train the networks to map abnormal patterns to a value close to one and normal patterns to a value close to zero. Training is carried out until \(10^6\) patterns are processed considering a learning rate of 0.005. During classification, averaging of the networks’ outputs results in an image whose pixel values represent the likelihood that a mass is present. This likelihood map is then slightly smoothed and its local maxima are determined. A local maximum is selected as a candidate location when its likelihood is above a certain threshold. This yields a list of locations that are of interest for further investigation. Before training and classification, features are normalized to zero mean and unit standard deviation using the training set.

### 3.B. Feature maxima propagation

As discussed in Sec. 1, misalignment and differences of spread among feature maxima may pose serious problems when performing image location classification. In order to alleviate these problems, we propose to propagate the maxima of the relevant features through a given neighborhood. This is accomplished by means of the maximum filtering algorithm described below:

1. For each feature image location \(F(i,j)\) that belongs to the tissue area, define a circular neighborhood \(S\) of radius \(R\) centered at location \(F(i,j)\). Members of this neighborhood are only those locations that belong to the tissue area as well.
2. Determine the maximum value of the neighborhood.
3. Assign this value to \(F'(i,j)\), which is the corresponding location on a temporal feature image.
4. Once all valid image locations have been explored, replace their original values \(F(i,j)\) with the temporal values \(F'(i,j)\).

Considering our particular application, the proposed algorithm is intended to be applied at the end of the feature extraction step described in Sec. 3.A.2 before Gaussian smoothing (see also Fig. 2). It is worth to mention that, although all the features with relevant peaks could be processed, such an approach is not necessary. In a pilot experiment, we verified that applying the maximum filter only to the gradient-based features \((g_1\) and \(g_2)\), so their maxima are aligned and propagated with respect to their spiculation-based counterparts \((l_1\) and \(l_2)\), yields comparable results as filtering all of them. Therefore, this approach was followed in the current study.

Figure 3 shows the same feature images corresponding to Fig. 1 once maximum filtering (and Gaussian smoothing) has been carried out. In this case, \(g_1\) (far right) was processed with a maximum filter configured with \(R = 3.2\) mm, while \(l_1\) (right) remained unprocessed. Comparing both feature images, it is possible to observe that the peak alignment is not perfect in the sense that the gap between the maxima still exists. However, the neighbors of the maxima, which now have very close values, are indeed aligned. In addition, the width of the peak in \(g_1\) better matches the width of the peak in \(l_1\), which leads to a better overlap and, thus, increases the chances of having

---

Fig. 3. The effect of applying the proposed maximum filtering algorithm to the gradient-based feature image shown in Fig. 1 (right). From left to right: original mammographic image, original spiculation-based feature image, original gradient-based feature image, smoothed spiculation-based feature image, and smoothed gradient-based feature image after maximum filtering. Note that although the maxima in the last two feature images are still misaligned, locations with very close values are now matched with each other.
correct representative values of both peaks in the same feature vectors. Although it can be argued that with a larger radius a better match in alignment and overlap could be obtained, such a setting may also increase the distortion in the feature image, which in turn may lead to “enhancement” of irrelevant locations that do not belong to a lesion pattern. Ideally, a compromise between enhancement and distortion should be determined. Thus, in this study, we also investigated the effect of different neighborhood radii on the detection performance (see Sec. 3.D).

3.C. Local feature selection

Since maxima misalignment and differences in maxima spread do not always occur simultaneously in all the features, considering only the interactions of those features that do not show the aforementioned issues may lead to improved classification performance. Unfortunately, there is no way to know a priori which subsets of features are the most suitable for each particular case, as this condition may be different for a large number of image locations and their associated feature vectors. This makes the application of any feature selection method that estimates the relevance of features globally and only once per data set inappropriate. Therefore, in this study we investigate a particular classification technique, random forests, which incorporates a local feature selection mechanism as part of its learning and classification algorithms, and thus realizes (to a certain extent) the desired strategy that processes different subsets of features for different subsets of feature vectors. This classifier constitutes an alternative to the ensemble of neural networks described in Sec. 3.A.3 (see Fig. 2 as well).

A random forest is an ensemble of tree predictors that can deal with both classification and regression problems. For classification, the algorithm takes the input feature vector, classifies it with every tree in the forest and outputs the class label that receives the majority of votes. For regression, the output is the average of the outputs over all trees. We used regression trees in order to obtain continuous output. The following algorithm was utilized in order to construct each regression tree in the ensemble:

1. Let the number of cases in the training set be \( N \). Sample \( N \) cases at random with replacement from the original data and use these cases as a training set for the tree.
2. Let the number of features in the data set be \( M \). Define a number \( m \), \( m < M \), such that at each tree node, \( m \) features are randomly selected in order to determine the best split.
3. Grow the tree to the largest possible extent and do not prune it.

By carefully analyzing the previous algorithm, it is possible to identify a number of reasons why the random forest framework would be capable of achieving the desired local feature selection effect. First, the \( m \) randomly selected features are utilized at each tree node to determine the best split. Since the training set is recursively partitioned at each tree node, different subsets of feature vectors comprising different subsets of features are explored. This increases the chances that correctly constructed feature vectors are evaluated. Second, since there are usually several nodes (due to the lack of pruning) and several (independent) trees in a given forest, a wide variety of combinations are explored by the whole ensemble. By adding this to the previous point, diversity is further increased. Thus, we expect that the appropriate feature interactions can be learned with the appropriate feature vectors. Finally, in the same line as above, having different training sets for every tree in the ensemble also contributes (although to a lesser extent) to diversification.

Breiman\(^\text{16}\) showed that the generalization error of a random forest depends on the interplay between two parameters: the strength of each individual classifier and the correlation between any two individual classifiers. He also showed that increasing the strength of the trees decreases the forest error, while increasing the correlation increases this error. Moreover, reducing or increasing \( m \), respectively, reduces or increases both the strength and correlation, which means that an optimal range of \( m \) could be determined. This seems to be the only adjustable parameter to which random forests are somewhat sensitive. However, Breiman also pointed out that this optimal range is usually wide. Thus, taking into account that the number of available features in our application is already low, a fixed nonoptimized value of \( m \) was considered to suffice. Following Breiman’s approach, this value was set to the first integer less or equal than \( \sqrt{M} \).

3.D. Evaluation

Detection performance was evaluated using free-response receiver operating characteristic (FROC) analysis after five-fold cross-validation. During data splitting into folds, we took care that the images corresponding to the same case were always assigned to the same fold. Moreover, the ratio of abnormal to normal cases was roughly the same in each fold. Once the five folds were processed, the likelihood scores assigned to the resulting maxima in the five test sets were pooled together and lesion-based FROC was computed. Sensitivity was defined as the number of detected lesions divided by the total number of lesions. A lesion was considered detected when the local maximum reported by the system was inside the reference annotated region. If several local maxima could be associated with a given lesion, only the one with the highest score was counted.

In order to obtain a single performance measure, the mean sensitivity in a range of false positive rates on a logarithmic scale\(^\text{3}\) was computed as

\[
S(3.5, 15) = \frac{1}{\ln 15 - \ln 3.5} \int_{3.5}^{15} \frac{s(f)}{f} df, \tag{1}
\]

where \( f \) is the number of false positives per normal case image (FPs/image) and \( s(f) \) is the lesion sensitivity. Note that under a screening setting, only the normal cases are relevant when assessing the false-positive detections. The computed measure is proportional to the partial area under the FROC curve plotted on a logarithmic scale. Using a logarithmic scale avoids
that the measure is dominated by operating points at high false positive rates. The range of 3.5–15 FPs/image was selected as it largely covers the operating points reported in the literature for similar mass candidate detection systems.1,20,23–25

Statistical significance of the performance difference between pairs of evaluated approaches was determined by means of bootstrapping.26 Cases were sampled with replacement from the pooled cross-validation set 5000 times. Every bootstrap sample had the same number of cases as the original data set. For each new sample, two FROC curves were constructed using the likelihood scores yielded by the original data set. For each new sample, two FROC curves were constructed using the likelihood scores yielded by the two methods being compared. Then, the difference in mean sensitivity, ΔS, was computed. After resampling 5000 times, 5000 values of ΔS were obtained. p-values were defined as the fraction of ΔS values that were negative or zero. Since we performed three comparisons per experiment, we applied the Bonferroni correction to the statistical significance level. Performance differences were considered significant if \( p < 0.0167 \) (0.05/3).

Three experiments were carried out in order to assess the performance of the proposed approaches. Table II summarizes the system configurations utilized in each of them. More details are given below.

In the first experiment, we compared the detection performance of the spiculation- and gradient-based features with and without the maximum filtering algorithm described in Sec. 3.B. We also assessed the effect of using different neighborhood radii with the maximum filter. The explored values of \( R \) were 1.6, 3.2, and 6.4 mm. Classification was carried out by the ensemble of neural networks and the methodology described in Sec. 3.A.3.

In the second experiment, we compared the detection performance of the random forest classifier described in Sec. 3.C with that of the ensemble of neural networks used by the baseline detector. In order to have further insights from this experiment, a gentle adaboost classifier27 with regression stumps as weak learners was also included. We selected to compare with this technique for several reasons. First, it constitutes a classifier ensemble similar to the random forest investigated in this study. Second, it is considered a state-of-the-art classifier. Third, and most important, it results in a sort of feature selection algorithm that evaluates features according to differently weighted training sets when selecting the best classifiers for the ensemble. Consequently, it represents a point in between what we criticized in Sec. 3.C about traditional feature selection approaches and what is expected from the feature selection capabilities of random forests. All the compared techniques processed the same baseline features as described in Sec. 3.A.2 and followed the same classification methodology as described in Sec. 3.A.3. The number of trees utilized by the random forest was 100. This parameter value was the one that yielded the highest mean sensitivity during parameter optimization, although by a very narrow margin, which is in accordance with the observation that, after a certain point, adding more trees to the ensemble will not significantly increase or decrease the prediction performance. We evaluated several values: 50, 100, 200, and 500 in order to select the best number of trees. A similar approach was followed to select the number of regression stumps for the gentle adaboost classifier. The final setup and the reported results correspond to 100 stumps.

In the final experiment, we combined both the maximum-filtered features and the random forest classifier, and compared the resulting detection performance with the one obtained by applying each of the approaches independently (as in the previous two experiments). The baseline detector was included as well. The radius of the neighborhood used by the maximum filtering algorithm was set to \( R = 1.6 \text{ mm} \), as it yielded the best performance (Sec. 5 elaborates on this point).

### 4. RESULTS

The results of the performed comparisons are listed in Tables III–V. In these tables, the second column shows the mean sensitivity obtained by the approach given in the first column, the third column shows the approach with which the current one is compared and, for each comparison, the \( p \)-value is given in the fourth column. Significant differences are
TABLE III. Mean sensitivity ($S$) achieved by the mass candidate detection system provided with features with and without maximum filtering (baseline features).

<table>
<thead>
<tr>
<th>Feature set</th>
<th>$S$</th>
<th>Compared with</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline features</td>
<td>0.926</td>
<td>. . . . . .</td>
<td>. . . . .</td>
</tr>
<tr>
<td>Maximum filtering ($R = 1.6$)</td>
<td>0.953</td>
<td>Baseline features</td>
<td>&lt;0.0002*</td>
</tr>
<tr>
<td>Maximum filtering ($R = 3.2$)</td>
<td>0.953</td>
<td>Baseline features</td>
<td>0.0022*</td>
</tr>
<tr>
<td>Maximum filtering ($R = 6.4$)</td>
<td>0.940</td>
<td>Baseline features</td>
<td>0.1156</td>
</tr>
</tbody>
</table>

indicated with a star (*). The FROC curves yielded by the compared approaches are shown in Figs. 4–6.

5. DISCUSSION

The results in Table III indicate that the proposed method for feature maxima propagation yielded a significant improvement over the unprocessed features utilized as the baseline in this study. For instance, using a neighborhood with a radius of 1.6 mm, the maximum filtering algorithm increased the mean sensitivity from 0.926 to 0.953 in the range of 3.5–15 FPs/image. This increase in sensitivity corresponds to a substantial decrease of 36.5% in the number of missed masses during candidate detection, which we expect will lead to a substantial improvement of the overall sensitivity when subsequent refinement steps, such as false positive reduction, are added. Taking into account particular operating points such as 4 FPs/image, which can be considered a typical choice in order to maximize the performance of those subsequent steps while minimizing the candidate complexity, the improvement is even larger (0.882 vs 0.931) and accounts for a 41.5% reduction of missed masses.

The results in Table III also give an insight into the effect of using increasing neighborhood radii during maxima propagation. While with the smallest radii of 1.6 and 3.2 mm the obtained mean sensitivities were the highest, using larger radii, such as 6.4 mm and above, significantly deteriorated the detection performance. A possible explanation is that, under those increasing radii, more normal image locations, that meant to have mismatched features due to their normal condition, were randomly matched with spurious peaks. As a consequence, they became as suspicious as true lesions for the classifier and resulted in false-positive detections. This brings into consideration the point stated in Sec. 3.B about selecting a compromise value for the neighborhood radius. Given our experimental results for this particular CAD application, a value of 1.6 mm is preferred, since it yielded virtually the same mean sensitivity as with 3.2 mm (with even a lower p-value when comparing with the baseline features) and is expected to introduce less distortion.

Considering our random forest-based method, similar trends as in the previous experiment were observed (see

![FROC curves for the mass candidate detection system provided with features with and without maximum filtering (baseline feature).](image)

FIG. 4. FROC curves for the mass candidate detection system provided with features with and without maximum filtering (baseline feature).

TABLE IV. Mean sensitivity ($S$) achieved by the mass candidate detection system when the ensemble of neural networks, the gentle adaboost classifier, and the random forest classifier are applied for image location classification.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>$S$</th>
<th>Compared with</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural networks (5 networks)</td>
<td>0.926</td>
<td>. . . . . .</td>
<td>. . . . .</td>
</tr>
<tr>
<td>Gentle adaboost (100 stumps)</td>
<td>0.934</td>
<td>Neural networks (5 networks)</td>
<td>0.0418</td>
</tr>
<tr>
<td>Random forest (100 trees)</td>
<td>0.948</td>
<td>Neural networks (5 networks)</td>
<td>0.0002*</td>
</tr>
</tbody>
</table>
<pre><code>                                                             | Gentle adaboost (100 stumps) | 0.0008*  |
</code></pre>

TABLE V. Mean sensitivity ($S$) achieved by the mass candidate detection system when the maximum-filtered features and the random forest classifier are applied independently or together.

<table>
<thead>
<tr>
<th>Detection with</th>
<th>$S$</th>
<th>Compared with</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural networks + maximum filtering ($R = 1.6$)</td>
<td>0.953</td>
<td>. . . . . .</td>
<td>. . . . .</td>
</tr>
<tr>
<td>Random forest + baseline features</td>
<td>0.948</td>
<td>. . . . . .</td>
<td>. . . . .</td>
</tr>
<tr>
<td>Random forest + maximum filtering ($R = 1.6$)</td>
<td>0.958</td>
<td>Neural networks + baseline features</td>
<td>&lt;0.0002*</td>
</tr>
</tbody>
</table>
<pre><code>                                                             | Neural networks + maximum filtering ($R = 1.6$) | 0.0748   |
                                                             | Random forest + baseline features | 0.0098*  |
</code></pre>
In this case, the mean sensitivity increased from 0.926 to 0.948, which corresponds to a 29.7% reduction of missed masses. Due to its relation with the proposed random forest classifier, the gentle adaboost classification method was evaluated as well. The results in Table IV indicate that this method also lead to improvement over the ensemble of neural networks, although it was not significant. This improvement can be explained by the weighting mechanism used during training. Since the weights associated with the feature vectors are modified by the algorithm, the training set used at each iteration can be regarded as a different set with a different prevalence for each feature vector, which results in a similar feature vector resampling approach as the one taken by random forests. Actually, when error estimates for the feature vectors cannot be introduced as part of the adaboost learning procedure, resampling of the original training set according to the assigned weights is suggested. Thus, under the conditions given by our problem, the feature selection process carried out by the stumps over the modified training sets is expected to give a certain improvement, as the influence of those vectors containing incorrectly constructed features could be minimized due to the varying weights. However, even though this results in a more refined approach than traditional feature selection, all the available features are still considered at once and, therefore, the obtained benefits are not maximized. This situation is corroborated by our results, which show a clear and statistically significant advantage in favor of the random forest classifier compared with the combination of gentle adaboost and regression stumps.

In the final experiment, we applied both the maximum filtering and random forest approaches. The resulting mean sensitivity increased to 0.958. While this improvement was not significant when comparing with the first proposed method, it was significant when comparing with the baseline detector and the random forest using the original unfiltered features. In the former case, the achieved improvement corresponds to a reduction of 43.2% in the number of missed masses, which is the highest in this study. Regarding the radius of the neighborhood used by the maximum filter, we observed again a similar although more pronounced trend than in the first experiment. It seems that the sampling mechanism of the random forest makes it more sensitive to the introduced distortion. Thus, by fusing the resulting outliers, the performance degraded sooner than with the ensemble of neural networks. As a consequence, the best option was to use a maximum filter with $R = 1.6$ mm.

Despite the fact that the operating point of mass candidate detectors is typically set at several FPs/image, as was done for evaluation purposes in this study, it is worth to consider the behavior of the combined proposed strategies at lower FP rates, such as 1 FP/image, where they achieved 73.4% sensitivity compared with the 62.8% of the baseline detector ($p < 0.0002$) (see Fig. 6). This is a key result if we take into account the performance of the whole system, since the likelihood assigned by the mass candidate detector, as well as measures derived from it, can be included as features in the subsequent steps. In fact, it has been shown in previous work that this kind of features is relevant for the final classification and, according to our observations, we may even state that the likelihood assigned by the mass candidate detector is probably the feature that most strongly influences the performance of the complete system. This is not surprising, as, from a feature selection perspective, a feature that on its own provides such a high sensitivity at an operating point that corresponds to the upper bound of the one set for the complete system (i.e., 0.05–1 FPs/image) is expected to be one of the most important.

Besides the performance measure values obtained from FROC analysis, it is interesting to observe the particularities of the masses that were previously missed by the baseline detector but were identified by applying the methods proposed in this study. Some examples, as well as crops of their original unprocessed feature images, are shown in Fig. 7. In all cases, the differences in peak location and spread among some...
Fig. 7. Examples of masses that were missed by the baseline detector but were identified by the methods proposed in this study (left) and crops of the original unprocessed feature images associated with these masses (right). The feature image crops in the top rows correspond to $l_1$ and $l_2$, while the ones in the bottom rows correspond to $g_1$ and $g_2$. The region where the mass should be detected is indicated with a square. The top two lesions belong to screening-detected cases, while the two at the bottom belong to prior cases. Note that the third lesion was located close to the image boundary.

Fig. 8. Examples of masses that were still missed despite the methods proposed in this study (left) and crops of the original unprocessed feature images associated with these lesions (right). The feature image crops in the top rows correspond to $l_1$ and $l_2$, while the ones in the bottom rows correspond to $g_1$ and $g_2$. The region where the mass should be detected is indicated with a square. The top lesion belongs to a prior case, while the one at the bottom belongs to a screening-detected case.

of the features (especially when comparing the spiculation-based features with the gradient-based ones) are evident. On the other hand, regarding the masses that were still missed despite the proposed approaches, we observed that, in most cases, many or all the features characterizing those masses did not exhibit the expected peaked response discussed along the paper. For instance, Fig. 8, first row, shows a lesion for which none of the evaluated features gave a strong well-defined response. Probably, the feature extractors were not sensitive enough given the subtlety of the lesion. A second, less severe case, where $l_1$ and $l_2$ did not yield relevant maxima and the one resulting from $g_2$ was not too strong, is shown in the next row. Even though the gradient-based features could detect this mass to a certain extent, the lack of evidence from their spiculation-based counterparts may have led the classifier to the wrong decision. From the previous examples, it becomes clear that the solution proposed in this study cannot deal with these kind of situations. Perhaps, the only way to solve these problems would be to modify the feature extractors themselves.

Although the current paper has focused on mass detection in screening mammography, the proposed methods are believed to be generalizable to a wide range of CAD applications, such as prostate cancer detection or lung nodule detection, as far as they follow a similar initial detection approach based on image location classification. Applying the maximum filtering algorithm to other tasks is straightforward, as it is based on direct operations on the obtained feature images. Moving from 2D to 3D analysis, for instance, only requires redefining the utilized neighborhood in terms of a volume instead of an area. Depending on the particular application, different neighborhood shapes, such as square or elliptical (or cubical or ellipsoidal in 3D) may also be explored. More importantly, recalling our experimental results regarding the size of the neighborhood, it may be necessary to properly adjust this parameter to match the classification strategy followed by the application. In the case of the
second proposed solution, probably nothing else than better tuning the number of randomly selected features may be advisable if the number of initial features is high. Determining the number of trees at which a stationary error rate is reached would be again a matter of grid search. Perhaps, an interesting line of future research would be to include the local selection mechanisms available in random forests within other types of classifiers besides trees. In a more general case, both of the proposed solutions can also be useful to compensate for mismatches between features that show peaked responses and features whose response is not so well-defined, as the latter can be viewed as an extreme case of the second identified problem.

6. CONCLUSION

In this paper, we identified two problems related to image location classification for mass candidate detection in CAD that may occur when combining multiple features that exhibit a peaked response at sites where relevant patterns are present. These problems consist of peak misalignment and differences of peak spread. As a solution, we proposed two methods that work at the feature and classifier level, respectively. The first method aims at propagating the maxima of a given feature image considering a circular neighborhood centered at each maximum location. The second method exploits the particular sampling and classification mechanisms of random forests in order to fuse the utilized features. To assess the effectiveness of the proposed methods, we selected a CAD application consisting of mass candidate detection in screening mammography. After independently applying the proposed methods, significant improvements on the detection performance at typical operating points were observed in both cases. The combination of the proposed methods further improved the obtained results. A similarly improved detection performance was observed at more specific operating points as well. Thus, we believe that the strategies proposed in this paper can render mass candidate detection approaches based on image location classification more robust to feature discrepancies and prove advantageous not only at the candidate detection level but also at later steps of a CAD system.

---

41Author to whom correspondence should be addressed. Electronic mail: jaime.melendezrodriguez@radboudumc.nl; Telephone: +31 24 3652272; Fax: +31 24 3540866.


