Potential of a Standalone Computer-Aided Detection System for Breast Cancer Detection in Screening Mammography

Jaime Melendez, Clara I. Sánchez, Rianne Hupse,
Bram van Ginneken, and Nico Karssemeijer

Radboud University Nijmegen Medical Centre, Department of Radiology,
Geert Grooteplein Zuid 18, 6525 GA Nijmegen, The Netherlands
j.melendezrodriguez@rad.umcn.nl

Abstract. Current computer-aided detection (CAD) systems for mammography screening work as prompting devices that aim at drawing radiologists' attention to suspicious regions. In this paper, we investigate utilizing a CAD system based on a support vector machine classifier as a standalone tool for recalling additional abnormal cases missed at screening, while keeping the associated recall rate at low levels. We tested the system on a large database of 5800 cases containing abnormal instances (1%) corresponding to prior examinations missed at screening. The results showed that 26% of the missed cases could be detected with a low additional recall rate of 2%. Moreover, after extrapolating this result to a screening program, we determined that, with our system, 0.73 additional cancers per 20 additional recalls could be potentially detected. We also compared the proposed system with a regular CAD system intended for non-standalone operation. The performance of the proposed system was significantly better.

Keywords: Screening mammography, breast cancer, computer-aided detection, support vector machine.

1 Introduction

Breast cancer is one of the leading causes of death among women. Therefore, it is essential to detect the presence of any sign of this disease as early as possible. To accomplish this objective, screening programs have been deployed in several countries, being mammography the preferred examination method. However, it is known that screening mammography tends to be difficult for radiologists and screening errors cannot be avoided. For instance, previous work (e.g., [1,2,3]) has shown in a retrospective study that between 57% and 67% of the cancers detected at screening examination are already visible on a prior mammogram.

Taking into account these findings, the importance of developing tools that aid radiologists in their work, such as computer-aided detection (CAD) systems, becomes evident. In the last decades, CAD systems for screening mammography have been introduced in clinical practice and, for instance, in the United States,

A.D.A. Maidment, P.R. Bakic, and S. Gavenonis (Eds.): IWDM 2012, LNCS 7361, pp. 682–689, 2012.

they are nowadays applied on about three of four screening mammograms [4]. The aim of these systems is to prompt radiologists to any suspicious region on a mammogram, thus they are designed to achieve high sensitivity, at the expense of obtaining low specificity. In fact, they operate at a false positive rate that is at least an order of magnitude higher than that of radiologists.

In this paper, we investigate a rather different application of CAD that, instead of prompting suspicious regions, aims at detecting malignant cases potentially missed by screening radiologists. The idea is to run the CAD system on the set of not recalled cases in order to generate an additional set with the most suspicious exemplars and then send these exemplars back to radiologists for reconsideration. To operate at a low recall rate, the system is trained using data following the distribution encountered in screening setting, i.e., high prevalence of normal cases. Additionally, the CAD parameters are optimized to operate at a recall rate of 2%, which closely matches the numbers observed in some screening programs [5].

2 Materials and Methods

2.1 Image Database

A total of 18242 scanned film mammography images from Preventicon screening center (Utrecht, the Netherlands) have been used in our experiments. They correspond to 5800 patients and comprise 188 images from 58 prior exams with visible masses and architectural distortions that were not detected until a later screening round. The remaining 18054 images (5742 exams) correspond to normal cases with no sign of pathology. For both normal and abnormal exams, either two or four views have been included depending on their availability.

2.2 Overview of the CAD System

The developed CAD system consists of a pre-processing stage, an initial detection stage and an interpretation stage that aims at reducing the number of false positive detections.

During the pre-processing stage, mammograms are segmented into three regions: breast tissue, pectoral muscle and background. The image background is labeled by marking pixels with high exposure and low gradient values. This operation is followed by morphological transformations to remove labels and to fill small gaps. Subsequently, the pectoral muscle in the mediolateral oblique (MLO) views is segmented as a straight line using a method based on the Hough transform [6].

After pre-processing, locations in the tissue area are sampled on a regular grid and, at each location, five features based on gradient and spiculation measures are computed to determine the presence of a potentially suspicious pattern [7,8]. These features are fed into an ensemble of five neural networks that are randomly initialized and trained on a small data set. In this way, a more powerful

classifier than with a single network is obtained. For each location at the grid, a likelihood score is computed by averaging the five network outputs. Together, these likelihood scores form a likelihood map. After smoothing this map, each local maximum that exceeds a threshold is selected as a candidate region and is segmented using the dynamic programming method described in [9].

In the interpretation stage, the segmented regions are classified into normal or malignant tissue by means of a soft margin support vector machine (SVM) configured with a radial basis function (RBF) kernel. For this stage, a new set of features measuring region contrast, location, linear texture, density, region size, compactness and contextual information is computed. In addition, the five gradient and spiculation features and the likelihood score computed in the initial detection stage are also used. Therefore, a total of 73 features is processed. Each of these features is normalized to have zero mean and unit standard deviation before classification.

2.3 CAD System Training

Training of a CAD system to operate at low recall rates involves a large number of normal cases (more than 4000 per fold considering the four-fold cross-validation evaluation scheme explained in Section 2.4). This is necessary for two reasons: first, to achieve the high specificity required for standalone operation and, second, to be able to accurately measure the performance of the system at that high-specificity operation point. In this work, this set of normal cases corresponds to a random sample of the whole population available in the complete Preventicon database and thus aims at modeling the actual distribution of the data.

Another key point during training is the optimization of the SVM classifier used in the interpretation stage. In this work, two of its parameters: the penalization parameter for the abnormal instances, C^+, and a coefficient related to the width of the RBF kernel, γ, have been determined using a grid search procedure.

The first parameter, C^+, derives from the formulation of the soft margin SVM, which aims at dealing with non-separable classification problems [10]. Since a hyperplane that perfectly separates the two analyzed classes may not always exist, the soft margin method determines a hyperplane that splits the classes as cleanly as possible, while allowing some of their instances to lie inside the margin or even be misclassified. These elements are penalized during optimization by a penalty parameter C by following the formulation shown below:

$$\begin{aligned} \min_{\mathbf{w},\xi,b} \quad & \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \right\} \\ \text{subject to} \quad & y_i(\mathbf{w}\cdot\mathbf{x}_i - b) \geq 1 - \xi_i, \\ & \xi_i > 0 \ , \end{aligned} \tag{1}$$

where $\mathbf{w}$ denotes the separating hyperplane. The larger the value of C, the larger the impact of those elements on the resulting model. Essentially, C can be

regarded as a tuning parameter and, in problems with highly imbalanced data, such as the one dealt with in this paper, separate parameters, C^+ and C^-, are used for abnormal and normal instances, respectively. Furthermore, in order to keep the tuning process tractable, one of them is usually kept constant and the other one is varied [11]. In this work, C^- has been set to one and C^+ has been searched over $C^+ \in \{1, 5, 10, 50\}$.

The second parameter, γ, derives from the fact that the sets to be discriminated are usually not linearly separable in the original space, thus an SVM often maps the input data into a higher (maybe infinite) dimensional space in which separation is expected to be easier. This mapping is achieved by means of a kernel function $K(\mathbf{x}, \mathbf{x}')$, which in our case corresponds to the RBF:

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{\gamma}) \ , \tag{2}$$

where γ is related to the kernel width and must also be tuned appropriately. In this work, the search grid for γ has been constructed by randomly sampling 1000 data points, computing their pairwise distances, deriving the 30-, 50- 70- and 90- percentile and averaging the results after ten trials.

The selected values for both C^+ and γ correspond to those yielding the highest sensitivity at 2% recall rate after classifying the training set using a three-fold cross-validation.

2.4 Evaluation Method

A four-fold cross validation scheme was used for the evaluation of the proposed CAD system. Image sets corresponding to individual cases have not been distributed among the cross-validation subsets. Moreover, the ratio of abnormal to normal cases has been roughly the same in each subset. Afterwards, a curve with case sensitivity values for different, increasing recall rates has been computed for each classifier. The partial area under this curve (PAUC) from 0 to 2% recall rate has been used as a performance measure. The statistical analysis has been carried out using the bootstrap method [12]. Cases were sampled with replacement from the compete cross-validation set 5000 times.

3 Results

The performance curve obtained for the developed system is shown in Fig. 1 (solid line). The mean PAUC from 0 to 2% recall rate is 3.39×10^{-3} (95% CI = 1.75×10^{-3} to 5.25×10^{-3}), while the mean case sensitivity at 2% recall rate is 0.264 (95% CI = 0.150 to 0.389), which indicates that 15 of the missed cases could be detected. Furthermore, at a lower recall rate of 1%, the mean case sensitivity is 0.202 (95% CI = 0.086 to 0.317) and thus 11 cases could still be detected. Some examples of detected cases are shown in Fig. 2.

We have extrapolated our results to the Dutch screening program carried out at Preventicon screening center. In a previous study, a detection rate of 0.49%

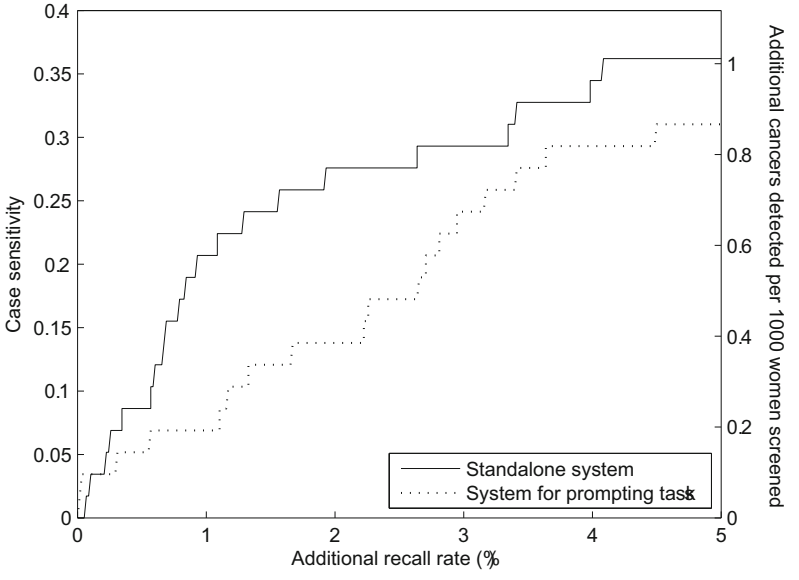

Fig. 1. Case-based performance curve for the standalone system proposed in this paper and the system intended for prompting tasks previously developed by our group [13]

for film-based mammography was obtained [5]. Assuming that 57% of these detected cancers should be visible on a prior mammogram as stated before [1,3], the additional number of detected cancers expected by applying our CAD system can be computed. They are associated with the right vertical axis in Fig. 1. Therefore, considering an additional recall rate of 2%, 0.73 additional detected cancers per 20 additional recalls are expected.

For the sake of comparison, a CAD system previously developed by our group and designed for prompting tasks has been evaluated [13]. This system comprises the same pre-processing and detection stages as the standalone system and processes the same features, but utilizes an ensemble of five neural networks in the interpretation stage. Training of these networks involves a stopping criterion based on a validation learning curve generated using an independent set of images (additional to the training set). In our experiments, these networks have been configured with 12 nodes in the hidden layer, a learning rate of 0.005 and a sampling ratio of 9:1 negative to positive instances presented during training. These settings correspond to the ones used in normal operation.

The performance curve for this system is also shown in Fig. 1 (dotted line). The mean PAUC from 0 to 2% recall rate is 1.61×10^{-3} (95% CI $= 0.46 \times 10^{-3}$ to 2.99×10^{-3}), while the mean case sensitivity at 2% recall rate is 0.147 (95% CI = 0.051 to 0.258). Comparing this performance with the one achieved by the standalone system, there is a clear and statistically significant advantage in favor of the latter ($p < 0.05$).

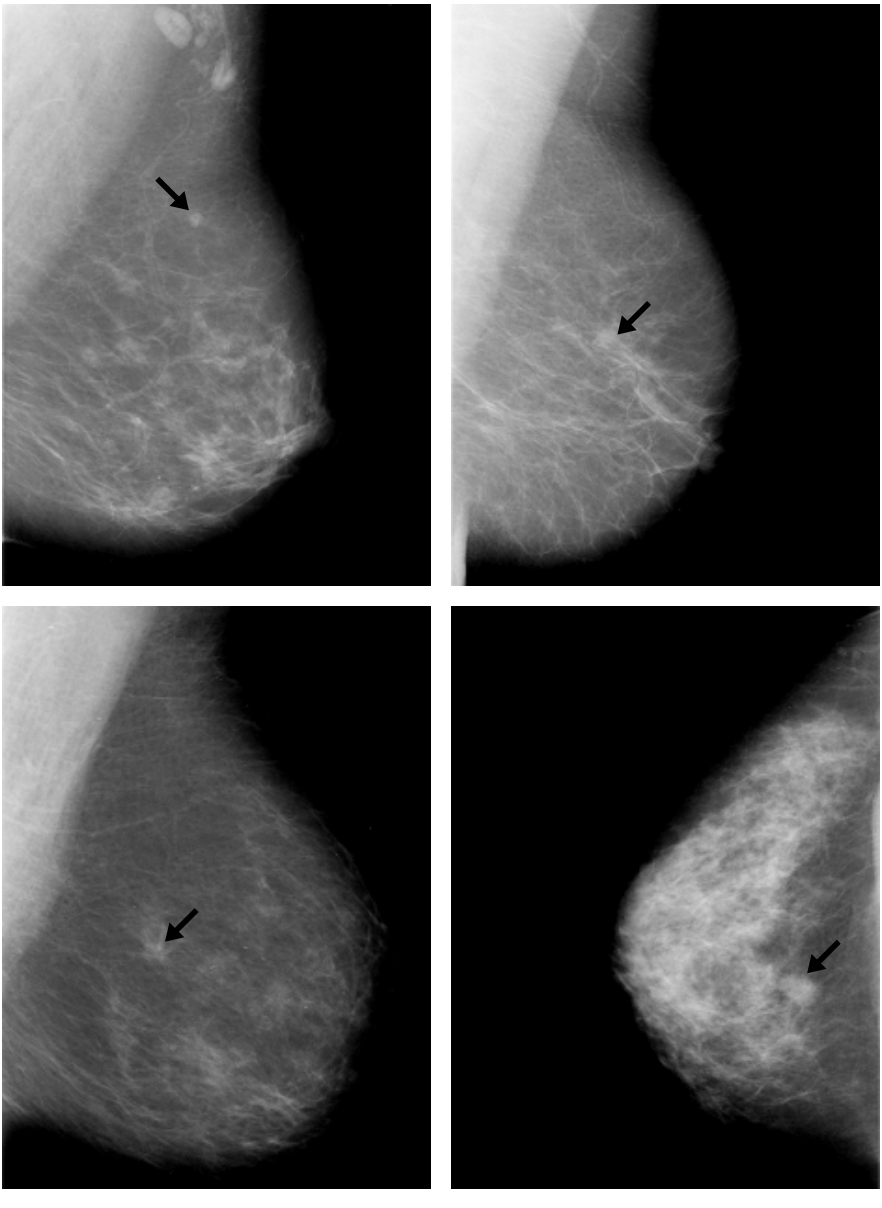

Fig. 2. Examples of malignant cases missed during screening that were detected by the proposed standalone system. The detected lesions are indicated with an arrow.

4 Discussion and Conclusion

In this paper, we have proposed a new application for a CAD system consisting of detecting suspicious cases missed by radiologists during screening, while keeping a low additional recall rate. We have developed a highly specific system intended for standalone operation by training an SVM classifier with a large set of normal cases and by optimizing its parameters at 2% recall rate. This system has been evaluated on a large image database that aims at approximating the typical setting observed in screening, which consists of a large number of normal cases (99% in our database) and a very small number of abnormal ones (1% in our database).

Moreover, to demonstrate the potential of the developed system, we have selected the set of abnormal cases in such a way that it corresponds to examinations missed by radiologists during prior screening rounds. The preliminary experimental results showed that, operating at a similar recall rate as screening radiologists, the developed system was able to detect 26% of these missed cases. Extrapolating this result to a mammography screening program, the obtained detection sensitivity corresponds to 0.73 additional detected cancers per 1000 women screened. As a consequence, several late screen-detected cancers could be detected earlier at the expense of a 2% additional recall rate. However, we hypothesize that, by referring the set of cases selected by the system back to radiologists for validation, the final number of false positive cases recalled for further examination could be considerably lowered. We are currently planning a study to assess this hypothesis.

As part of this work, we have also compared the proposed standalone system with a CAD system for prompting. The results showed that the performance of the standalone system is significantly better, which is mainly due to the specific optimization procedure followed during training, as well as the generalization capabilities of the SVM and its ability to deal with high-dimensional spaces.

References

1. van Dijck, J.A., Verbeek, A.L., Hendriks, J.H., Holland, R.: The Current Detectability of Breast Cancer in a Mammographic Screening Programme: A Review of the Prior Mammograms of Interval and Screen-Detected Cancers. Cancer 72, 1933–1938 (1993)
2. Warren Burhenne, L.J., et al.: Potential Contribution of Computer-Aided Detection to the Sensitivity of Screening Mammography. Radiology 215, 554–562 (2000)
3. Otten, J.D.M., et al.: Effect of Recall Rate on Earlier Screen Detection of Breast Cancers Based on the Dutch Performance Indicators. J. Natl. Cancer I. 97, 748–754 (2005)
4. Rao, V.M., Levin, D.C., Parker, L., Cavanaugh, B., Frangos, A.J., Sunshine, J.H.: How Widely is Computer-Aided Detection Used in Screening and Diagnostic Mammography? J. Am. Coll. Radiol. 7, 802–805 (2010)
5. Karssemeijer, N., et al.: Breast Cancer Screening Results 5 Years after Introduction of Digital Mammography in a Population-Based Screening Program. Radiology 253, 353–358 (2009)

6. Karssemeijer, N.: Automated Classification of Parenchymal Patterns in Mammograms. Phys. Med. Biol. 43, 365–378 (1998)
7. Karssemeijer, N., te Brake, G.M.: Detection of Stellate Distortions in Mammograms. IEEE T. Med. Imaging 15, 611–619 (1996)
8. te Brake, G.M., Karssemeijer, N.: Single and Multiscale Detection of Masses in Digital Mammograms. IEEE T. Med. Imaging 18, 628–639 (1999)
9. Timp, S., Karssemeijer, N.: A New 2D Segmentation Method Based on Dynamic Programming Applied to Computer Aided Detection in Mammography. Med. Phys. 31, 958–971 (2004)
10. Cortes, C., Vapnik, V.: Support-Vector Networks. Mach. Learn. 20, 273–297 (1995)
11. Lesniak, J., et al.: Computer Aided Detection of Breast Masses in Mammography using Support Vector Machine Classification. In: Proc. SPIE, vol. 7963 (2011)
12. Samuelson, F.W., Petrick, N., Paquerault, S.: Advantages and Examples of Resampling for CAD Evaluation. In: Proc. IEEE Int. Symp. Biomed. Imag., pp. 492–495 (2007)
13. Hupse, R., Karssemeijer, N.: Use of Normal Tissue Context in Computer-Aided Detection of Masses in Mammograms. IEEE T. Med. Imaging 28, 2033–2041 (2009)