MammoInsight Computer Assisted Detection: Performance Study with Large Database

Johann Drexl, Peter Heinlein and Wilfried Schneider

Imagetool GmbH, 12489 Berlin
Email: drexl@imagetool.de

Abstract. We extend our previous publications by presenting results on a large number of mammograms digitized with a laser scanner and analyzed by our MammoInsight CAD-system for the detection of clustered microcalcifications. We measure sensitivities on a per breast basis and on a per cluster basis for our system and show how they interrelate. We compare the performance of our system to the state-of-the art in the research literature. Finally, we show that our CAD-system is able to obtain a high sensitivity per breast, justifying it's use in a clinical environment.

1 Introduction

Microcalcifications are tiny deposits of calcium embedded in the tissue of the female breast. They are an important early indicator of possible breast cancer. On mammograms, they appear as small spots of irregular shape and low contrast (diameter 0.13 mm - 1 mm), which makes them difficult to detect. The reliable detection of microcalcifications is therefore very important.

The reference standard for microcalcification detection is the system of Veldkamp and Karssemeijer [3]. They recently published results on a large database of mammograms.

There are two fundamentally different ways of evaluating the sensitivity of microcalcification detection algorithms. In a mammographic examination, two images are taken of each breast: the craniocaudal and the mediolateral view. Evaluating sensitivity on a per breast basis is common in the medical community [5], while evaluating on a per cluster basis is common in the image processing community [3]. Sensitivity per breast is also used in specifications of regulatory bodies like the FDA [7]. While sensitivity-per-breast indicates the potential use of the prompting to a radiologist, sensitivity-per-cluster is more detailed, thus preferred for optimizing and comparing CAD-algorithms.

2 Overview CAD for MammoInsight

We previously reported on a CAD-scheme for detecting clustered microcalcifications [1],[2]. We gave performance results for the Nijmegen-Database in [1], and for the MIAS-Database in [2]. There, we can detect all malignant clusters

at a false positive rate of 0.33 false positives/image [2]. Using this method, our diagnostic workstation MammoInsight is able to generate prompts for clustered microcalcifications.

Our CAD-scheme is based on a continuous wavelet decomposition of the original mammogram. The process starts with downsampling and the extraction of the breast tissue from the mammogram. All subsequent computations are restricted to this region. Next, noise equalization according to Veldkamp and Karssemeijer [3] is performed, to decouple the standard deviation of the wavelet coefficients from the digitized optical densities of the mammogram. We extract features that are designed to exploit the scale information encoded in wavelet coefficients. These include features for local contrast, size, edge strength, anisotropy and orientation. The classification is performed by support vector machine (SVM) classifiers in a two-stage-cascade. Training was done on a subset (20 images) of the Nijmegen images. The output is a list of findings, each with a score value, which allows to adjust the sensitivity of the detection process. Finally, we apply explicit postprocessing rules to the detection results: we remove isolated candidates, clusters with less than three findings, and findings with an area above a certain limit.

Our SVM classifiers employ discriminant functions which are polynomials of second order. We opted for the use of SVMs, because they have the following interesting property: Instead of minimizing a plain measure of error, SVMs minimize a measure of error penalized by a measure of classifier complexity [8]. This deals with the following challenges in our application:

1. Because of the high scale interdependency among the wavelet features, the covariance matrix of the feature vector is singular. This makes an application of classical statistical classifiers like Fisher's linear discriminant or the Mahalanobis classifier (which require inversion of covariance matrices) impossible.
2. We want to keep the number of training examples needed for training of the classifiers small. A penalty on the classifier complexity can bound the number of examples.

3 Material for evaluation

The largest public database for mammography by now is the Digital Database for Screening Mammography (DDSM), maintained by the University of Florida [4]. It contains images scanned on four different scanners. As we are primarily interested in using laser scanners as digitizer devices, we chose to select DDSM images from the Lumisys 200 scanner (50 microns resolution). The DDSM is organized in patients. Every patient consists of 4 images: left and right breast, and each in craniocaudal and mediolateral view. To construct our database, we retrieved all malignant calcification cases from the volumes cancer_01, _02, _05, _15 and _09.

We selected all image pairs containing calcifications in both views of the same breast, excluding images containing masses and excluding images consisting only of benign calcifications.

This resulted in 68 image pairs from pathologically proven malignant cases (136 images). With our selection scheme we acquired as many suitable images as possible from the DDSM.

All clusters in the DDSM have a degree of subtlety, which allows us to rate the degree of difficulty of our database. The distribution of subtleties is:

Subtlety	1	2	3	4	5
Number of clusters	5	24	56	34	46

where 1 is a very subtle cluster and 5 is a very easy to recognize cluster. Apart from subtlety 1, all subtleties are approximately equal distributed. Compared to the distribution of subtleties in all above mentioned five tapes, we notice that our subset has a higher degree of difficulty.

4 Evaluation method

The performance evaluation criteria employed in this study are based on the criteria of [3]. The DDSM comes along with a pathologically proven "gold standard". We generated the truth circles necessary for Karssemeijer's evaluation technique by drawing an enclosing circle about the cluster area in the DDSM gold standard. For sensitivity-by-cluster, a hit is scored when there are at least two prompts in a truth circle. Then the mean true positive fraction (TPF) due to [3] is computed.

For sensitivity-by-breast, we use the following scheme:

a hit is counted, when at least one truth circle was recognized in one or both views. Sensitivity-per-breast is defined [5]:

$$\text{TP/breast} := \frac{\text{Number of hits according to above definition}}{\text{Number of image pairs}} \tag{1}$$

Motivation for this metric: When a cancerous breast has at least one prompt, the breast is recognized as cancerous, even when not every cluster in both views is recognized by the system. At a near 100% rate it would be guaranteed that the CAD does not miss a cancerous breast. A performance metric like this is more appropriate for generating prompts in a diagnostic workstation than for example the TPF-metric. On the other hand, the TPF is more relevant when it comes to comparing and optimizing CAD-algorithms. Another advantage is the independence from different annotation styles. Some radiologists prefer to split a large cluster into several small cluster, while others prefer to annotate on a coarser scale. A scheme like the above is invariant to this [6].

The false positive rate is computed as in [3].

5 Results

We ran the CAD on our study data set and evaluated the generated promptings. Subsequently, we computed the TPF and the false positives per image. Veldkamp and Karssemeijer recently published results on a large database of 245 mammograms scanned on a Lumisys LS85 for their system [3]. Unfortunately, neither implementation nor database is accessible to the general public. Therefore, we decided to base the comparision on the bestperforming FROC-curve from [3]. In figure 1, we compare Veldkamps and Karssemeijers FROC curve with our result (we omitted points with more than 5 false positives per image from the curve). One notes that our system is better than the reference, although one

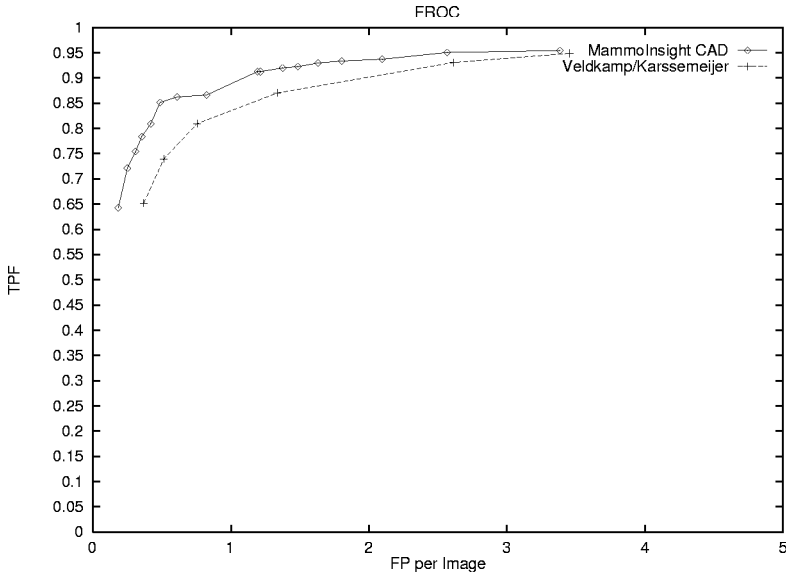

Fig. 1. MammoInsight CAD and Veldkamp/Karssemeijer

must of course be cautious when comparing the two FROC curves because the underlying databases differ. On the other hand:

- images obtained by the LS200 and the LS85 are very similiar, as they are both laser scanners from the same manufacturer.
- both databases use similar annotation styles.
- both databases are very large and contain a large number of microcalcification clusters, which dimishes random influences on the statistics.

We compute TPF and TP/breast for our system and compare them in table 1. Using the cut-off $\theta_2 = 30\%$, one sees that it's possible to reach a TP/breast of 97.1%. This corresponds to a TPF of 86.3%. Thus, we conclude that our CAD-scheme has a high sensitivity-per-breast for malign, clustered microcalcifications.

Table 1. Sensitivity for MammoInsight

Cut-off θ_2 [%]	10	20	30	40	50	60	70	80	90
TP/breast [%]	98.5	97.1	97.1	95.6	92.6	92.6	89.7	85.3	75.0
TPF [%]	91.3	86.6	86.3	85.1	81.0	78.4	75.4	72.1	64.3

The mean difference between the two performance figures is 11%. Thus, in studies giving only sensitivity per breast, it appears reasonable to conjecture that the sensitivity per cluster may be actually by ten percent lower than the sensitivity per breast.

6 Summary

We conclude that our CAD-system compares favorable to the existing standards for detection of clustered microcalcifications. We further conclude that our CAD-system has a high sensitivity-per-breast for malign, clustered microcalcifications. system is very important for the use in a clinical environment, because in order to use the device effectivly, users have to be confident about the device's promptings.

The diagnostic workstation MammoInsight is used in clinical routine in the clinic of Aschaffenburg, Germany, since June, 2002.

References

1. Heinlein P, Drexl J, Gössler A, et al.: An integrated approach to computer aided diagnosis in digital mammography, in: Lemke HU et al.: CARS 2001, Elsevier Science B.V. 2001, p.538-543
2. Schneider W, Heinlein P, Drexl J, et al.: Anwendungen der computerassistierten Diagnostik (CAD) und Tele-Mammographie bei Brustkrebs-Vorsorgeuntersuchungen, Röntgenpraxis 54, p. 192-198 (2002)
3. Veldkamp W, Karssemeijer N: Normalization of local contrast in mammograms, IEEE Trans Med Imaging, Vol. 19, No. 7, July 2000.
4. Bowyer K, Kopans D, Kegelmeyer W, et al.: The digital database for screening mammography, in: Doi K et. al: Digital mammography '96 , Elsevier Amsterdam 1996, p. 431-434
5. Roehrig J, Doi T, Hasegawa A, et al.: Clinical results with R2 Image Checker, in: Karssemeijer N: Digital mammography, Kluwer Academic Publ Dordrecht 1998, p. 395-400
6. Hume A, Thanisch P, Hartswood M, et al.: On the evaluation of microcalcificat ion algorithms, in: Doi K et. al: Digital mammography '96 , Elsevier Amsterdam 1996, p.273-276
7. U.S. Food and Drug Administration (FDA) application #P980058, approved, June 26, 1998
8. Platt J: Probabilistic Outputs for Support Vector Machines and Comparisions to Regularized Likelihood Methods, in: Smola A et. al.: Advances in Large Margin Classifiers, MIT Press 1999