

An automatic method for the identification and interpretation of clustered microcalcifications in mammograms

Ferdinand Schmidt†‡, Erich Sorantin†, Csaba Szepesvári§, Ewald Graif†,
Michael Becker†, Heinz Mayer† and Karin Hartwagner†

† Department of Radiology, Karl-Franzens-University, Graz, Austria

§ Research Group on Artificial Intelligence, 'Jozsef Attila' University, Szeged, Hungary

Received 10 August 1998, in final form 22 December 1998

Abstract. We investigated a method for a fully automatic identification and interpretation process for clustered microcalcifications in mammograms.

Mammographic films of 100 patients containing microcalcifications with known histology were digitized and preprocessed using standard techniques. Microcalcifications detected by an artificial neural network (ANN) were clustered and some cluster features served as the input of another ANN trained to differentiate between typical and atypical clusters, while others were fed into an ANN trained on typical clusters to evaluate these lesions.

The measured sensitivity for the detection of grouped microcalcifications was 0.98. For the task of differentiation between typical and atypical clusters an A_z value of 0.87 was computed, while for the diagnosis an A_z value of 0.87 with a sensitivity of 0.97 and a specificity of 0.47 was obtained.

The results show that a fully automatic computer system was developed for the identification and interpretation of clustered microcalcifications in mammograms with the ability to differentiate most benign lesions from malignant ones in an automatically selected subset of cases.

1. Introduction

In Europe and in the US every tenth woman will get breast cancer during her life. Unfortunately, the only known possibility for improving the poor life expectancy of these women is the early detection of cancer (Smith 1995). Today mammography is the only accepted screening modality available for early cancer detection. Obviously, considerable benefit could result from better means of mammographic discrimination between benign and malignant lesions, reducing anxiety, cosmetic problems and the cost of screening programmes. Clustered microcalcifications are one of the mammographic signs of early breast cancer, but cancer-related microcalcifications may sometimes be hard to differentiate from calcifications associated with benign diseases and other artefacts (Monsees 1995, Hogge *et al* 1995, De Paredes *et al* 1990). Radiologists have just started to use digital mammography in tumour-localization procedures, but full-view digital mammograms will be on the market in the near future. These will allow immediate computer-aided image processing techniques and automated methods of diagnosis to be applied to help reporting radiologists (Wu *et al* 1992, Schmidt *et al* 1995).

‡ Address for correspondence: Karl Franzens-University, Department of Radiology, Auenbruggerplatz 9, A-8036 Graz, LKH, Austria. E-mail address: ferdinand.schmidt@kfunigraz.ac.at

Several algorithms have already been developed for detecting breast calcifications (Wu *et al* 1992, Karssemeijer 1991, 1993, Kegelmayer and Allmen 1994) but only a few works are devoted to the computer-aided differentiation of benign and malignant lesions, i.e. to *interpretation*. These studies use either manual identification of calcifications along with computer-assisted feature extraction (Jiang *et al* 1996b) or human feature extraction as the input for their categorization systems (Baker *et al* 1995, Wu *et al* 1993). Therefore, these methods are labour-intensive and may be prone to human error.

Our aim was to develop a completely automated method for both the *identification* and *interpretation* of microcalcifications which could serve as a 'never tired second reader'. It is well known that double reading in mammography can improve the accuracy of radiological reports by as much as 5–15% and so an 'automated second reader' may be of a great value (Thurfjell *et al* 1994). It is not only the performance and the consistency of the CAD system that counts, but also that the computer system may interpret the films in a *different manner*. As a result the error of the human reader and the CAD system as a whole can be substantially reduced. To understand this, imagine that the computer system just imitates the human reader. This situation of course would not differ from the case when the human expert is working by himself. So a second objective of our study was to investigate whether the decision-making mechanism of the human experts differs substantially from that which was derived automatically using the applied statistical methods.

All the patients in our database had histological records so that we could ensure that our data are correlated with the truth. Unfortunately, this has the drawback that the number of available cases is limited. The current method could be extended to cases *without* histological data without much effort, for example by considering patients with at least three (or more) years of follow-up. Truth would be confirmed because of no change in the mammogram during time course instead of histology. It is important to note here that we are aware of the fact that the prototype developed so far solves only parts of the problems of computer-assisted mammography since at the moment the developed system cannot handle spiculated or circumscribed tumours without calcifications.

2. Materials and methods

2.1. Data acquisition

For two years the available mammographic films of all histologically confirmed cases with clustered microcalcification of patients having breast cancer or benign diseases were collected at the University Hospital of Graz. The database consisted of 272 films of 100 patients. We had at least two films for each patient, the films showing projections along different planes (the craniocaudal and mediolateral views). In most of the cases an additional lateral view was obtained. Histological examination of the specimens revealed malignancy in 54 out of the 100 patients while in the remaining 46 cases a benign condition was diagnosed. The 272 mammograms were digitized with the Pixelizer 6k digitizer (Medical Diagnostic Computing, Zeiss, Hannover, Germany) which is linear to the light transmitted by the film. A pixel size of 91.5 μm was used for image postprocessing; the bit depth was set to 15 bits.

2.2. Overview of the decision system

The decision system performs the following steps: (a) image preprocessing, which identifies potential microcalcifications; (b) filtering out true microcalcifications; (c) grouping the microcalcifications into clusters; (d) filtering out clusters that may result in an unreliable

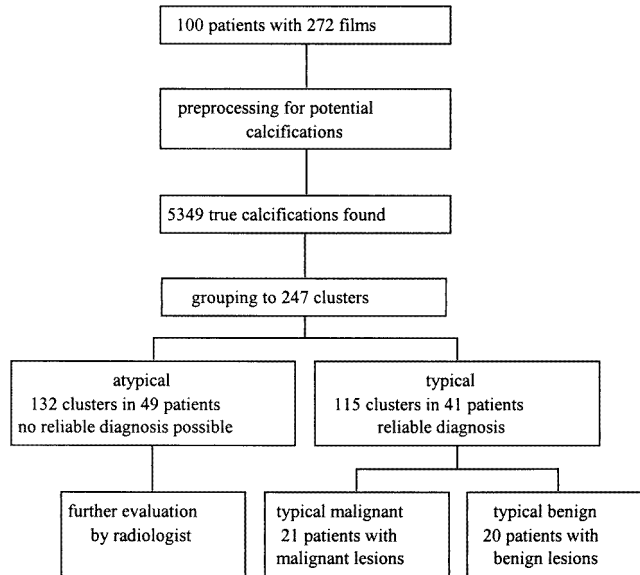


Figure 1. Flow chart describing the different steps for the CAD system.

diagnosis; (e) making a diagnosis for the remaining clusters; and finally (f) for the patients. The first two steps together are called the identification process, while the rest is called the interpretation process. Now we will describe these steps in detail (figure 1).

2.3. Detection

In the preprocessing phase of the detection a regional background correction method is first applied around the pixels. This is done by fitting a two-dimensional polynomial function of degree three to the image over a 27×27 area around the given pixel whereby the image is considered as a 2D surface over the region by treating the grey levels as the respective surface heights. The fitted surface is then subtracted from the original image (figure 2). Additionally, a local contrast image is computed from the original one using a balanced kernel, i.e. the original image was contrast enhanced with a 9×9 kernel (the centre value was set to 80, all other values were set to -1). Both the high-pass filtered image and the background corrected image were thresholded with the 98.5 percentile and saved as bi-level images. After multiplication of both bi-level images, the zeros represented the background, and the ones pixels of potential microcalcifications. Both the background-corrected and the local contrast images are combined in order to identify pixels that belong to individual microcalcification candidates. Connected pixels are grouped to form objects.

We will call artefacts and other bright objects (e.g. crossing of septal lines, calcified vessels) false positives. For filtering out false positive microcalcifications 13 features, shown in table 1, have been computed: these are the object contrast-value, a descriptive statistic (minimum, maximum, average, variance) of the grey-level distribution and the same for the line-feature distribution of the object and for the distribution of edge-values measured at the border of the object. The object contrast value, which is calculated on the original image, is defined as the difference between the average grey level over the area of the object and that of over a two pixel enlarged area around the same object. The aim of the line features associated

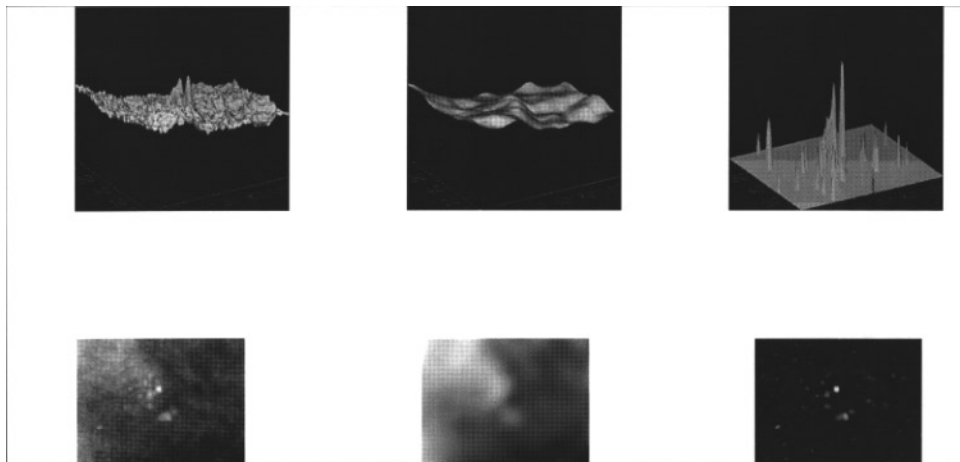


Figure 2. Background correction. In the upper row a 3D representation of the grey values within a kernel of size 27×27 are shown as surfaces, while in the lower one the corresponding image is shown. In the left column the original area, in the middle one the fitted polynomial, while in the right one the image obtained by the subtraction of the fitted polynomial from the original image is depicted.

Table 1. Thirteen features used for the detection of microcalcifications.

Minimum of grey level
Maximum of grey level
Mean of grey level
Standard deviation of grey level
Minimum of gradient at border points
Maximum of gradient at border points
Mean of gradient at border points
Standard deviation of gradient
Minimum of line features
Maximum of line features
Mean of line features
Standard deviation of line features
Mean of grey level of region points—mean of grey level of surrounding points

with the pixels is to detect if the object is elongated around the pixel. The line feature of a pixel is determined by the following procedure: firstly, the gradients around the pixel are calculated using the Sobel operators with a kernel of size 3×3 and then the gradients are transformed to angles which, in turn, are mapped to one of the 16 main directions. If the density of these discretized angles shows two peaks over a 9×9 area around the pixel the peaks corresponding to approximately opposite directions (e.g. 0° and 135° are considered as approximately opposite directions) then the product of the density values at those peaks is stored as the value of the line feature. If there are no such peaks then the corresponding line-feature value is set to zero. Edge values are calculated as the absolute value of the Sobel operators. After a suitable rescaling (in which features are normalized by linearly mapping the average feature value minus twice the standard deviation to -1 and the average plus twice the standard deviation to $+1$), the above features serve as the input of a two-layer feedforward ANN having two hidden neurons (Haykin 1994). The network is trained to differentiate between true microcalcifications and false positives.

2.4. Interpretation

(In the following, for brevity, we will write ‘microcalcification’ for detected objects, hoping that this causes no confusion.) After the microcalcifications are identified they are first rotated by the Hotelling transformation so that their main axes become parallel. In this way the shapes of the microcalcifications become comparable on the basis of their ‘shape indices’ from which any microcalcification has 16 associated with it: its extent along the eight main directions measured from the centre of gravity and the length (l), width (w), aspect ratio (l/w), area ($l \times w$) and eccentricity of the minimum enclosing rectangle, the latter being represented by the four numbers that describe the distances from the centre of gravity of the microcalcification to the border of its minimum enclosing rectangle. Also the area (a), perimeter (p) and circularity (p^2/a), and the above described statistics of the grey-level distribution within the microcalcification are obtained.

Microcalcifications are then grouped automatically to form clusters. Namely: two microcalcifications are defined as belonging to the same cluster if their distance apart is less than 1 cm. A total of 247 clusters have been identified in this way. The following cluster features are then calculated: the above detailed descriptive statistics for all the above mentioned microcalcification features for the microcalcifications within the cluster, the same descriptive statistics for the within-cluster intermicrocalcification distances, some shape parameters of the cluster such as the area (a), the perimeter (p), and the circularity (p^2/a), as well as the numbers of microcalcifications within the cluster (n), and the density of microcalcifications within the cluster (n/a). The shape parameters are calculated as the corresponding parameters of the convex hull of the cluster (see figure 3).

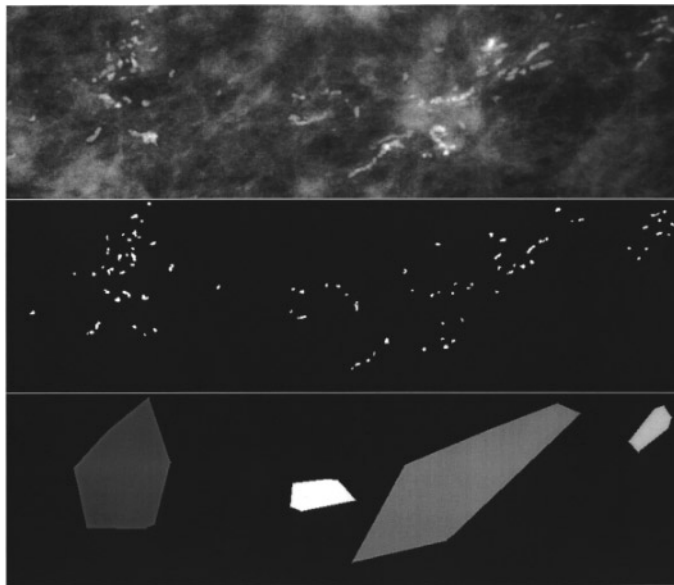


Figure 3. (top) The original image of a tumour area containing clustered microcalcifications. (middle) Calcifications detected by the CAD system. (bottom) The convex hull around the cluster.

Since the confidence of the performance assessment of classification algorithms decreases with the number of free parameters, a smaller number of features from the above are selected to be fed into the ANNs. The following automatic feature selection procedure formed the basis of

this selection: a patient-based leave-one-out test is performed for the best linear transformation of each feature, where the best linear transformation should be understood as the one which results in the minimum mean square error on the training examples. The test results are saved and afterwards an ROC curve (receiver operator curve) is created for every feature. Only those features whose ROC curve goes significantly above or below the diagonal going from the lower left to the upper right corner are included in the final analysis, i.e. those which has substantial positive or negative predictive power for some threshold values. This method can be argued to fit ANN training, but this fairly technical reasoning will be presented elsewhere.

First we tried to use ANNs directly to diagnose all images, but we found that the resulting decision system was of no practical use. After identifying the possible reasons for this failure (the analysis is partly presented in the discussion) we decided to extend the method to a two-step procedure, when in the first step the system filters out the images which are unsuitable for a reliable diagnosis (i.e. an automatic interpretation in which the result is not similar to chance—for these images no diagnosis should be made at all!), and a second step when for the remaining images a diagnosis is derived. Only those patients with all of their images found to be reliable for a diagnosis are included in the final step (e.g. a patient has benign calcifications if all clusters are rated by the CAD system as being benign). In order to provide sufficient information for the first step, the radiologists (FS, ES, KH) rated hand-marked clusters of microcalcifications on a consensus basis using similar concepts to the ones used in the BIRAD system (Reston 1995). Namely, clusters are categorized into five different groups: benign, probably benign, indeterminate, probably malignant and malignant. The hand-marking of clusters is made by enclosing the biopsied areas by rectangles, where special attention is paid to the fact that the marked areas should contain all the biopsied areas but only those. Clusters with labels 'benign' and 'malignant' form the group of 'typical' images, whereas the rest of the clusters form the 'atypical' group. 115 typical clusters has been identified in this way for 41 patients out of the total of 100 patients. The filtering process of the first step is based on this categorization; that is we wanted an ANN to learn to differentiate between typical and atypical clusters. Note that in this step only the manually marked regions are used so that we could assign labels to the automatically detected clusters in a meaningful way.

We determined a feature set for this task using the already described automatic feature selection procedure. The resulting set includes a total of 10 features summarized in table 2.

Table 2. Ten features used for the classification typical/atypical clusters.

Number of microcalcifications in cluster
Maximum of area
Maximum–minimum of area
Maximum of perimeter
Maximum–minimum of perimeter
Maximum of shape (shape = circularity)
Maximum–minimum of shape
Variance of distances between microcalcifications in cluster
Area of convex hull
Perimeter of convex hull

The rescaled features are used as the input of an ANN trained to differentiate between the typical and atypical lesions. Another ANN is trained to differentiate between benign and malignant clusters. In this latter case the results of the histological reports are used as the truth (note that the histological truth closely matched the rating of the radiologists) and the automatically selected set of features are summarized in table 3.

Table 3. Twelve features used for the classification benign/malignant clusters.

Number of microcalcifications in cluster
Maximum of area
Mean of area
Standard deviation of area
Maximum–minimum of area
Standard deviation of perimeter
Maximum distances between microcalcifications in cluster
Standard deviation of distances between microcalcifications in cluster
Variance of distances between microcalcifications in cluster
Minimum of first shape index
Area of convex hull
Perimeter of convex hull

The final decision is to say a patient has cancer if at least one of the clusters of the patient is rated as being malignant by the ANN.

2.5. Performance measurement

For comparison with others work the performance of the detection of *groups* of microcalcifications is measured, the groups being the manually marked ones. In agreement with some existing work (Karssemeijer 1993) an algorithm it is said to detect a *group* of microcalcifications if the ANN ‘detects’ at least two microcalcifications within the region. Estimation of the sensitivity and the fraction of false positives (i.e. the fraction of objects other than microcalcifications classified as being microcalcifications) followed the guideline of the above cited work.

Performances in the interpretation step are estimated using leave-one-out tests and the receiver operator curve (ROC) analysis, wherein the area under the ROC curve (the A_z value) is computed using a PC version of software distributed by the Rossmann Institute, University of Chicago, USA. The performance is estimated on a per patient basis since this is the value of primary interest.

The utilized hardware is a SunSparc 20 workstation (Sun Microsystems, Mount View, California) and the neurocomputer Synapse-1 (Siemens Nixdorf Advanced Technologies, Dresden, Germany) connected to a Sun workstation using an SBus adapter. Synapse-1 is programmed through a special C++ library (Neural Algorithms Programming Library 1.3.3, Siemens Nixdorf Advanced Technologies, Dresden, Germany). The image-processing algorithms are implemented with the help of the IDL 4.0 package (Interactive Data Language, Research Systems, Inc., Boulder, Colorado, USA). The ANNs are trained by an advanced version of the back-propagation algorithm (Vogl *et al* 1993).

3. Results

From a total of 272 films of 100 patients, 247 clusters of microcalcifications containing 5349 single microcalcifications were found. Figure 4 shows the distribution of the ratings of microcalcifications in different groups.

For the identification of groups of microcalcifications sensitivities of 0.90, 0.98 and 1.00 were achieved at the respective false positive alarm rates of 1.3, 5.3 and 7.4 groups per image. For differentiation between typical and atypical clusters an A_z value of 0.80 was measured. For the corresponding patient-based differentiation an A_z value of 0.87 with a sensitivity of 0.97,

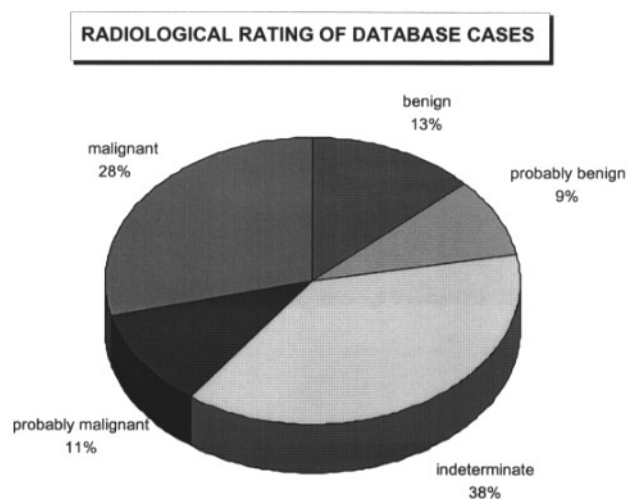


Figure 4. Distribution of the ratings of microcalcifications. The ratings were made by two human experts.

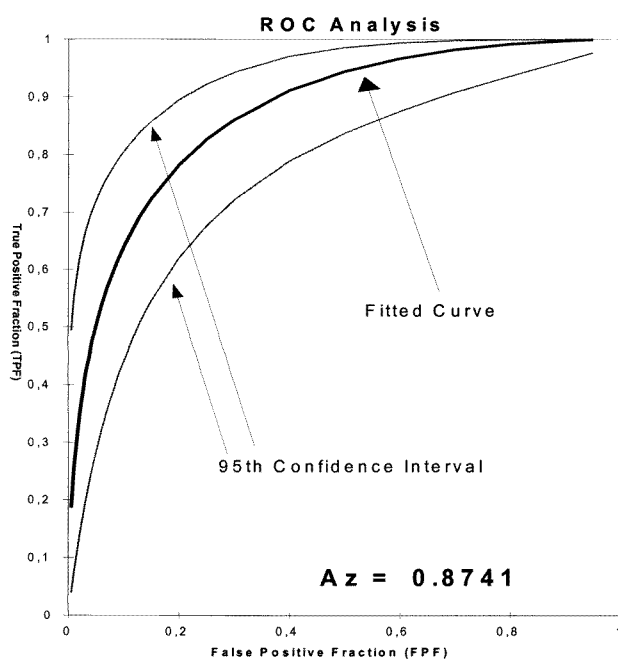


Figure 5. ROC curve obtained on a per patient basis for the differentiation between typical and atypical clusters.

specificity of 0.34 and accuracy of 0.70 (figure 5) was computed. The positive (negative) predictive values were 0.67 (respectively 0.91) in this case.

For typical clusters and a patient-based analysis the automated classification yielded an Az value of 0.87 with a sensitivity of 0.97, specificity of 0.47 and accuracy of 0.80 (figure 6). The positive and negative predictive values were 0.78 and 0.91 respectively in this case.

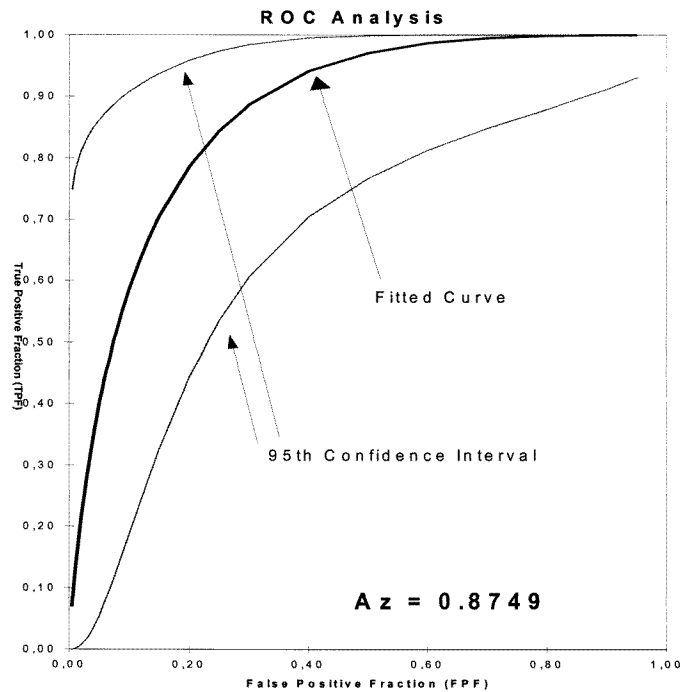


Figure 6. Malignant versus benign microcalcification clusters. The test is restricted to typical images.

4. Discussion

The results for the detection of microcalcifications show that the proposed scheme works well since it could detect 90% of all groups of microcalcifications on all images with a detection of a rather low number of false positive groups. Since typically there are two groups of microcalcifications per image, the results mean that every group is detected without the detection of any additional (false positive) group for almost all of the images. We can understand this pretty good result if we take into account that the feature set used in the detection process is the same as the ones used by Wu *et al* (1993) who also achieved good results ($A_z = 0.84$ for 43 features, $A_z = 0.89$ for 14 features). We also tried different network topologies and found (on the basis of the estimated performances) that the relatively simple two-layer two-hidden-node ANNs are already sufficiently good. As has been mentioned already, the confidence of the performance assessment decreases with increasing complexity (i.e. the number of free parameters, more precisely the VC dimension) of the applied decision method and so small nets with a performance equal to large ones should be preferred (Devroye 1996).

Recall that we had to break the interpretation scheme into two parts. In the beginning we tried to use the CAD system on the whole data set but the results were unsatisfactory (an A_z value of 0.64 was obtained on the best runs and also a closer examination of the learned weights of the ANN revealed counterintuitive results). Searching for the roots of this failure we found that (a) the films are hard to diagnose in the daily routine as is also suggested by the fact that all the patients had an open biopsy and (b) the normal routine of the radiologists would be to unconsciously differentiate between two different kinds of clusters of microcalcifications,

a cluster would appear immediately as being typically benign or malignant (e.g. teacup-type or y-shaped calcifications) to a radiologist or very difficult to diagnose. We called clusters of this latter type atypical. The existence of atypical clusters follows from the fact that the basic physics of the calcification processes underlying different diseases is largely the same at the beginning of the calcification *independently* of the causes of the calcification. Therefore there is no way to differentiate in the early stages, nor even in theory, between microcalcifications due to cancer and benign disease. This fact forced us to change the CAD system in such a way that it performs a two-step procedure, the first step being the differentiation of typical clusters from atypical ones and the second step being the classification of clusters recognized as being typical into the categories benign and malignant. It turned out that this approach is quite successful since the A_z value in the typical group increased to 0.87 and at the same time the differentiation between the typical and atypical groups worked well too ($A_z = 0.87$). Note that the A_z value of the ROC curve is not always an appropriate measure of performance because of its global nature, as was pointed out by Metz (1989). In our case for example it is more important not to miss any single cancer than to have some patient erroneously categorized as having a cancer, so the operating point on the ROC curve should be chosen as close to the 100% sensitivity as possible. (Note that since we filter out any patient who has any cluster ranked by the system as atypical, and since we call a patient from the remainder healthy only if all of her clusters are rated as being benign the CAD system itself tries to maximize the sensitivity of the diagnostic process.) A pointwise inspection of the empirical ROC curve then yields a sensitivity of 0.97 at the respective specificity of 0.47 for the final step, while a sensitivity of 0.97 at the respective specificity of 0.34 can be obtained for the differentiation between typical and atypical clusters. We prefer these pointwise measures to partial area indices suggested by Jiang *et al* (1996a) since these reflect better the actual working of the system. The percentage of patients processed in the final step is 41, which means that by using the system as a second reader at least 20 patients could have been saved from unnecessary open biopsy without essentially missing any cancer. In this respect, our results could be compared for example with that of the paper which reported that nine radiologists reached an A_z value of 0.77 for rating clustered microcalcifications in 77 mammographic films (Veldkamp and Karssemeijer 1996). It is worth mentioning that we tried different network topologies in all of the decision-making steps and we found that two-layer ANNs with two hidden neurons performed the best.

Very interesting conclusions can be drawn by interpreting the results of the automatic feature selection procedure. At the beginning we started from a large feature set (we have implemented 72 features altogether) which included all the features that were used in other studies dealing with computer-aided mammography (Kegelmayer and Allmen 1994, Metz 1989), the idea being that the automatic feature selection procedure can select the best set of features autonomously. The selected set of features is shown in tables 1, 2 and 3. Although the correlation among some features is rather high (above 0.9), no proper subset of this feature set could approximate the results obtained with the help of the whole feature set. It is also worth noting that most of the features identified by the system are also 'used' by radiologists. Namely, the following rough rules applied by radiologists fit the found features (we list the features followed by an interpretation): (a) number of microcalcifications within the cluster: if this number is small (2 or 3) then the lesion is benign, while if it is big then the cluster is atypical (imagine that many microcalcifications are diffusely distributed over the whole breast); (b) area of microcalcifications: if the microcalcifications are all large then they are benign, and if they are small then the cluster is typical but can be either benign or malignant; (c) intermicrocalcification distances: if the distribution of the distances between the closest microcalcifications have big lags then the cluster is malignant and *vice versa*; (d) the radius of

the minimum enclosing circle: if the minimum of these radii is large (all microcalcifications are large) then the cluster is benign; (e) size of the cluster: if the cluster is very large (remember that the microcalcifications are diffusely distributed over the whole breast) then the cluster is atypical but if it must be classified then it is usually benign.

By the inspection of some tuned ANNs without a hidden layer one can observe that the number of microcalcifications and the intramicrocalcifications distances have the strongest positive correlations with the type of the cluster (typical is called positive in this case), while perimeter length of the cluster has the strongest negative correlation. This is in agreement with what we have said before, i.e. that very large clusters with microcalcifications scattered over the whole breast are atypical, since if the cluster is large but has a relatively small number of microcalcifications then the output of the ANN will be negative, meaning that the cluster is atypical and *vice versa*. Nevertheless note that single-layer ANNs did not perform very well on the tests so the above derived 'rules' should be handled with care. Also note that although the correlation (precisely the Pearson coefficients) between some of the features is as high as 0.9 or above, ANNs trained on uncorrelated (not strongly correlated) subsets resulted in substantially worse results. This underlines the importance that the features contribute to the network output in a nonlinear way. The same conclusions hold for the other two decision steps.

In the final decision step, the minimum of the radii of the enclosing circles and the variability of the sizes of the microcalcifications within the cluster (the first of these two quantities measures how elongated the microcalcifications are within the cluster, more precisely this quantity is big only if all the microcalcifications within the cluster are stretched) together with the size of the cluster seem to be the strongest indicators of malignancy. Other, weaker, indicators are the variability of the perimeters of the microcalcifications and the maximum of the intermicrocalcification distances (i.e. if all intermicrocalcification distances are small then the cluster is benign). On the other hand, the variance of intermicrocalcification distances along with the number and the average area of the microcalcifications within the cluster are strong predictors of benign disease.

The above comment, i.e. that these interpretations should be used with care, applies here, too. The drawn features should be understood only as sensitivity factors since these features by themselves are insufficient to make reliable diagnosis and this also holds for their linear combinations. Moreover, since we used a strongly biased subsample of the whole population, every one of our results should be understood as being valid only within the subset of patients which would be sent to have a biopsy. Nevertheless, these simplified rules seem to be in good agreement with the working methods of radiologists which reinforces that CAD systems can be built up in this way and used in the daily routine with success.

A very important question is how reliable will the system be in practice in the future. There are numerous ways of estimating the performance of the system, likewise the resubstitution estimate, the cross-validation estimate or the leave-one-out test. Of course, the confidence of the estimate is the critical point and this depends mainly on two parameters: the number of cases used for testing and the complexity of the decision algorithm. The confidence typically increases exponentially with the number of cases and scales linearly with the complexity of the decision algorithm (Vogl *et al* 1993). Unfortunately, the trade-off between the two is such that a few data points do not enable the development of reliable estimates, especially not in the case of the resubstitution and cross-validation estimates which typically require a minimum of thousands of cases assuming an algorithm of medium complexity (like ours). Since we are only interested in an upper bound on the average number of future misinterpretations committed by the system, negatively biased pessimistic estimates, like the leave-one-out test, come naturally in the picture. The leave-one-out test has the advantage that it gives quite a reliable upper bound on the performance even for medium sized data sets (Kearns and Ron 1998). This may

seem somewhat at odds with the fact that it has a big variance, but it is important to realize that the leave-one-out test is known to have a large variance; it is important to notice here that the origin of this is that the distribution of the error estimate of the leave-one-out test is negatively skewed and so the upper confidence interval, in which we are really interested in, is much smaller than the lower one. Unfortunately, even if we take these modifying factors into account, the parametric-free confidence bounds that are known at the time of writing are still too broad to be of practical value. It should be noted here that the current database is one of the largest databases ever used in computer-aided mammography and the acquisition of a substantially larger database would be difficult at the moment because of the limited number of patients operated on at each institution. Parametric methods provide an alternate way of developing estimates of confidence intervals and typically give much tighter results. In the presentation of results we have chosen the binormal assumption as our parametric assumption on the data which is commonly used in studies dealing with computer-aided mammography (Jiang *et al* 1996a, b, Veldkamp and Karssemeijer 1996). Although the obtained data can hardly be said to meet the binormal assumption (since we, as does everyone else using a trainable method, *tune* the decision methods to output extreme values), it is widely believed that the binormal assumption is robust and so can be applied. A more appropriate method, which would be important to develop in the future, would be to derive the same type of confidence bounds for a different parametric assumption, for example when we assume that the data are 'bi-Weibull'.

5. Conclusion

An entirely automated CAD system prototype was developed for mammography which, according to our knowledge, is the first such system. We found that in the cases rated by human experts as being hard to diagnose the computer system could not infer a reliable diagnosis, and this called for a two-step procedure, which in the first step sorts out patients for whom a reliable diagnosis is not possible to be sent to a radiologist for further evaluation. For the rest of the patients the system proposes a diagnosis. It was found that some simplified rules extracted from the knowledge acquired by the system corresponded closely to that routinely used by radiologists. Therefore mimicking expert radiologists is useful. The estimate of the performance of the system predicts that the system, when used as a second reader, would be able to save 20% of the patients from open biopsy while essentially not missing any cancer.

References

- Baker J A, Korngruth P J, Lo J Y, Williford M E and Floyd C E Jr 1995 Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon *Radiology* **196** 817–20
- De Paredes E S, Abbit P L, Tabbarah S, Bickers M A and Smith D C 1990 Mammographic and histologic correlations of microcalcifications *Radiographics* **10** 577–89
- Devroye L, Györfi L and Lugosi G A 1996 *Probabilistic Theory of Pattern Recognition, Application of Mathematics: Stochastic Modelling and Applied Probability* (New York: Springer)
- Hogge J P, Robinson R E, Magnant C M and Zuurbier R A 1995 The mammographic spectrum of fat necrosis of the breast *Radiographics* **15** 1347–56
- Haykin S 1994 *Neural Networks, a Comprehensive Foundation* (New York: Macmillan)
- Iglewicz B 1983 Robust scale estimators and confidence intervals for location *Understanding Robust and Exploratory Data Analysis* ed D C Hoaglin, M Mosteller and J W Tukey (New York: Wiley)
- Jiang Y A, Metz C E and Nishikawa R M 1996a A receiver operating characteristic partial area index for highly sensitive diagnostic tests *Radiology* **201** 745–50
- Jiang Y A, Nishikawa R M, Wolverton D E, Metz C E *et al* 1996b Malignant and benign clustered microcalcifications: automated feature analysis and classification *Radiology* **198** 671–8

- Karssemeijer N 1991 A stochastic model for automated detection of calcifications in digital mammograms *Int. J. Inform. Process. Med. Imaging* **227**–38
- — 1993 Adaptive noise equalization and recognition of microcalcification clusters in mammograms *Int. J. Pattern Recog. Artificial Intell.* **7** 1357–76
- Kearns M and Ron D 1998 Algorithmic stability and sanity-check bounds for leave-one-out cross-validation *10th Ann. Conf. on Computational Learning Theory*
- Kegelmayer W P and Allmen M C 1994 Dense feature maps for detection of calcifications *Digital Mammography, Excerpta Medica* ed A G Gale, S M Astley, D R Dance and A Y Cairns pp 3–12
- Metz C 1989 Some practical issues of experimental design and data analysis in radiological ROC studies *Invest. Radiol.* **24** 234–45
- Monsees B S 1995 Evaluation of breast microcalcifications *Radiol. Clin. North Am.* **33** 1209–21
- Reston V A 1995 *ACR -BIRADS (American College of Radiology Breast Imaging—Reporting and Data System)*
- Schmidt R A, Wolverton D E and Vyborny C J 1995 Computer aided diagnosis in mammography *Syllabus: A Categorical Course in Breast Imaging RSNA* ed D B Kopans and E B Mendelson, pp 199–207
- Smith R A 1995 The epidemiology of breast cancer *Syllabus: A Categorical Course in Breast Imaging RSNA* ed D B Kopans and E B Mendelson, pp 7–20
- Thurfjell E L, Lernevall K A and Taube K A 1994 Benefit of independent double reading in a population-based mammography screening program *Radiology* **191** 241–4
- Veldkamp W and Karssemeijer N 1996 Automated classification of clustered microcalcifications in digital mammography *Digital Mammography, Excerpta Medica* pp 23–30
- Vogl T P, Mangis J K, Rigler A K, Zink W T and Alkon D L 1993 Accelerating the convergence of the back-propagation method *Biol. Cybernetics* **59** 257–63
- Wu Y, Doi K, Giger M and Nishikawa R 1992 Computerized detection of clustered microcalcifications in digital mammograms: applications of artificial neural networks *Med. Phys.* **19** 550–60
- Wu Y, Giger M L, Doi K, Vyborny C J, Schmidt R A and Metz C E 1993 Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer *Radiology* **187** 81–7