Analytical Cellular Pathology 14 (1997) 31–40
IOS Press

Detection of malignancy associated changes in cervical cell nuclei using feed-forward neural networks

Roger A. Kemp*, Calum MacAulay, David Garner and Branko Palcic
BC Cancer Research Centre, Vancouver BC, V5Z1L3, Canada

Received 29 November 1996
Revised 14 February 1997
Accepted 19 March 1997

Abstract. Normal cells in the presence of a precancerous lesion undergo subtle changes of their DNA distribution when observed by visible microscopy. These changes have been termed Malignancy Associated Changes (MACs). Using statistical models such as neural networks and discriminant functions it is possible to design classifiers that can separate these objects from truly normal cells. The correct classification rate using feed-forward neural networks is compared to linear discriminant analysis when applied to detecting MACs. Classifiers were designed using 53 nuclear features calculated from images for each of 25,360 normal appearing cells taken from 344 slides diagnosed as normal or containing severe dysplasia. A linear discriminant function achieved a correct classification rate of 61.6% on the test data while neural networks scored as high as 72.5% on a cell-by-cell basis. The cell classifiers were applied to a library of 93,494 cells from 395 slides, and the results were jackknifed using a single slide feature. The discriminant function achieved a correct classification rate of 67.6% while the neural networks managed as high as 76.2%.

Keywords: MAC, neural networks, cervical screening, automated cytometry

1. Introduction

The majority of automated cervical screening systems rely on the detection of abnormal appearing cells [1,5,6,9,18,21]. In the past two decades researchers have begun to design systems to identify MACs among the cells appearing normal, which make up the majority of the cells in a cervical sample. The idea is that such a system may be able to detect a slide from a patient with a precancerous lesion even if no diagnostic cells are present or detected from the sample.

It has been observed that there are significant changes in the quantitative nuclear features of intermediate cells from cervical smears of different grades [2,4,11,19]. These nuclear changes have been used to design classifiers using linear discriminant functions (LDF) [12,14,15,20]. Nonlinear classifiers such as neural networks have already been successfully applied to identifying diagnostic cells in cervical smears [3,17]. It is proposed that neural network classifiers may also be better suited to the

*Corresponding author: Roger A. Kemp, Cancer Imaging, BC Cancer Research Centre, West 10th Avenue 601, Vancouver BC, V5Z1L3, Canada. Tel.: +1 604 877 6010; Fax: +1 604 875 6857.

task of identifying MAC cells from normal cells. Should there exist a nonlinear decision boundary between MACs and negative cells, or should either cell distribution be multimodal in feature space, neural network classifiers could outperform discriminant functions.

2. Materials and methods

The samples used in this study consisted of conventional spatula collected cervical cell smears of premenopausal patients of the British Columbia Cancer Agency. A total of 190 slides diagnosed as "normal" and 144 slides diagnosed as containing "severe dysplasia" were used to design the cell-by-cell classifiers. Each slide was independently diagnosed by three or four cytotechnologists. Only slides for which three of the diagnoses were identical were candidates for use in the study.

The slides were stained with a Feulgen–Thionin stain, a quantitative DNA stain. Cell nuclei were collected as 256 grey-scale images with an automated image cytometer which had a $20\times$ objective lens magnification and an effective pixel size of $0.34\ \mu\text{m} \times 0.34\ \mu\text{m}$. To prevent batch-to-batch variations in the staining intensity of the slides from significantly influencing the results, the slides were randomly selected from 35 different staining batches. Slides from each staining batch were included in both the training and testing data. This ensured that the training data would be representative of data normally scanned by an automated image cytometer during routine cervical screening.

The LDF and neural network [16] cell-by-cell classifiers were trained using intermediate cells taken from normal and severe dysplasia slides. The cells were all diploid, intermediate cells in good focus segmented correctly, as verified by human observers. In this work, the intermediate cells from normal slides will be referred to as negatives and the intermediate cells from severe dysplasia slides will be referred to as MACs. Due to computing power limitations, it was not practical to use all available cells to design the classifiers. Hence, it was necessary to restrict the number of cells selected from each slide for the cell classification study.

A maximum of 75 normal intermediate cells were selected from each normal slide (negatives), and a maximum of 150 normal appearing intermediate cells were selected from each severe dysplasia slide (MACs). These limits were imposed to prevent a slide that had a large number of intermediate cells from being overrepresented in the design data. The human observer visual criteria used to select these two sets of cells were exactly the same. Intermediate cells make up a smaller fraction of the cells scanned from the severe dysplasia slides than scanned from normal slides and the available slide data set contained fewer severe dysplasia slides than normal slides. Thus, it was necessary to accept twice as many MACs, from the fewer positive slides, as negatives to provide adequate sized samples of each class of data.

Images of each object were examined at least five times in a hierarchical procedure to determine the cell type, condition and whether it was in good focus. Only those images of cell nuclei that satisfied these criteria were included in this study.

A total of 25,360 cells were used to design and test the cell-by-cell classifier, 12,680 from each class. One third of this data was used for training each classifier and the remainder was used for testing. Table 1 provides a breakdown of the number of cells of each type used in the study.

Each cell nucleus was represented by a vector of 53 features. The 53 features consisted of optical density, shape, Markovian and discrete texture features [8]. The value of each feature was normalized by converting it to a z-score by subtracting the feature mean and dividing by the feature standard deviation. The means and standard deviations were calculated for the training set population of 8456 cells.

Table 1
The breakdown of the number of cells of each type used to train and test the cell-by-cell classifiers

Cell type	Training	Testing	Total cells of type
Negative	4228	8452	12680
MAC	4228	8452	12680
Total	8456	16904	25360

The neural network classifiers in this study were feed-forward neural networks optimized on the training set of 8456 cells. Since the process of training a neural network amounts to a nonlinear optimization problem, the final network one obtains can depend on the initial weights used in the network [13]. Thus, for each neural network structure, 20 randomly selected initial weight configurations were used and the average classification rate of the 20 networks is used when discussing the performance of a particular network structure for cell-by-cell classification.

The neural networks each had 53 inputs, some number of hidden units, n, and one output. Between one and seven hidden units were used in the various neural networks. The networks were trained to return an output score of 0 for MACs and 1 for negatives. The Quickprop algorithm was used to train the neural networks [10]. Each network was trained for a sufficient number of iterations to allow the network weights to reach a local error minimum. For smaller networks, 1000 training iterations were adequate. For large networks, with seven hidden units, an upper limit of 20,000 iterations was used.

3. Results

3.1. Cell-by-cell results

Discriminant function analysis was applied to the training set of 8456 cells. The function was obtained using the 7M module of the BMDP statistical software package with F_{enter} and F_{remove} values of 4.0 and 3.996, respectively [7]. The resulting LDF, using 43 of the 53 features, correctly separated 62.1% of the training examples and 61.6% of the test examples. A quadratic discriminant function calculated for 53 features has 1432 variables to enter in the discriminant model. Since there are only 8456 training examples, there would be nearly one free parameter for every five examples. This has the potential problem of overfitting the data, so no quadratic discriminants were calculated for the MAC data.

One could attempt to select a subset of the features in attempt to find a quadratic discriminant function. A starting point might be to select some of the 43 LDF features with large F_{remove} values, but this does not necessarily yield the features which could generate a good quadratic classification function. The topic of how to select a good feature subset for use in any kind of classifier model deserves an in depth examination and is not addressed in this paper.

Neural networks with between one and seven hidden units were trained using the 8456 cell training set. Each network was then tested using the independent validation set containing 16,904 cells. Table 2 shows the average classification rate of 20 trials of each network structure on the training and test sets along with the best test set classification rate. The standard deviations of the training and test classification rate for networks trained using 20 sets of initial weights are given beside each result. The LDF classification rate is also shown for comparison.

The simplest neural networks, with no hidden units (or one hidden unit), form linear decision boundaries between the classes and are comparable to LDFs [13]. As expected, their classification

rate is almost identical to that of the LDF. These networks and the LDF achieve approximately 61% correct classification on the training and test sets.

There is a large performance improvement when two hidden units are used. The training set classification rate rises to 72.9% while the test set classification rate is 71.6%. The difference between the training and test set classification rates increases as more hidden units are used. The training set results improve marginally with each extra hidden unit, but the test set classification rate peaks when four hidden units are used. When five, six or seven hidden units are used the average test set results actually decline. These results are shown graphically in Fig. 1.

Table 2
The average correct classification rate and standard deviations for 20 randomly initialized networks are shown when different numbers of hidden units are used. The training set classification rate improves as the number of hidden units is increased, but the test set classification rate peaks when four hiddens are used. The classification rate of the best network on the test data is given for each network configuration

Number of hidden units	Avg. training set classification rate	Avg. test set classification rate	Best test set classification rate
LDF	62.1%	61.6%	–
1	$61 \pm 3\%$	$61 \pm 3\%$	63.5%
2	$72.9 \pm 0.3\%$	$71.6 \pm 0.2\%$	71.9%
3	$73.4 \pm 1.0\%$	$71.8 \pm 0.2\%$	72.4%
4	$74.7 \pm 0.6\%$	$71.9 \pm 0.4\%$	72.5%
5	$75.7 \pm 0.4\%$	$71.6 \pm 0.3\%$	72.0%
6	$76.3 \pm 0.4\%$	$71.0 \pm 0.4\%$	71.6%
7	$76.6 \pm 0.7\%$	$71.1 \pm 0.4\%$	71.6%

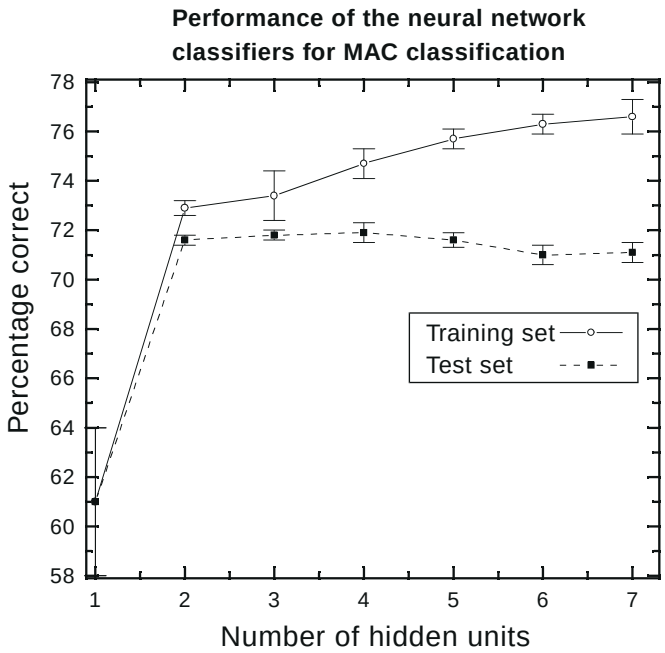

Fig. 1. The cell-by-cell classification rate of neural networks with different numbers of hidden units is shown for the training and test sets. Each result is the average of 20 runs using randomly initialized networks. The standard deviations are shown as error bars for each result.

The disparity between training and test set results grows as more hidden units are used. The significance of the difference between the training and test results can be seen when considering the standard deviation of each result. The error bars in the figure show the standard deviations of each average and are small compared to the difference between the two curves. Here, the neural networks can be seen to be overfitting the training data. A network with more than four hidden units and 53 input features has too many free parameters for this training set. The neural networks with five hidden units have 276 adjustable weights. This means that there are more than 30 training examples for each weight. Nonetheless, the average test set classification rate of the network declines when five hidden units are used.

With both the LDF and neural networks we can adjust the threshold used to assign cells to one class or another. If one plots the false negative versus false positive rates for each particular threshold value one can obtain the Receiver–Operator Characteristic (ROC) curve. Figure 2 shows the ROC curves for the LDF and neural networks with one, two and four hidden units. The ROC curves for the best networks containing three, five, six and seven hidden units nearly coincide with those of the two and four hidden-unit curves and are not shown. All the neural network ROC curves correspond to the networks whose classification rates are shown in the last column of Table 2.

The diagonal line in the figure represents the classification rate of a classifier which gives no discrimination (i.e., randomly assigning a class to each object). For example, calling all objects "MAC" would give a false positive rate of 100%, or calling all objects "negative" would give a false negative rate of 100%. The ROC curves for the better classifiers dip more deeply below this line, than do those for the poorer classifiers. The ROC curve for an ideal classifier would trace a path along the horizontal axis and up the vertical axis.

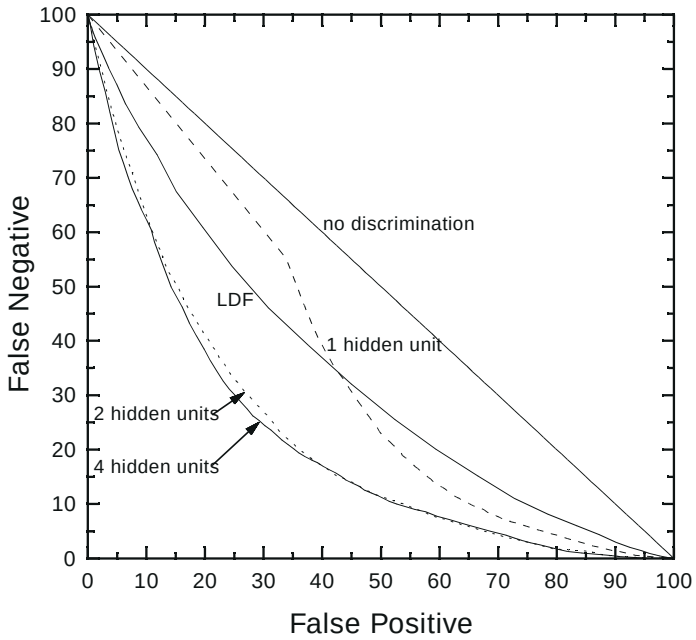

Fig. 2. The cell-by-cell ROC curves for the LDF and networks giving the best test set classification rate are shown. The ROC curves for the best networks using three, five, six and seven hidden units nearly coincide with those of the networks with two and four hidden units and are not shown.

Although the LDF and the best neural network with one hidden unit (63.5% classification rate) are both linear classifiers, their ROC curves are different. The neural network converged to a set of weights that gives it a different behavior than that of the LDF. It gives a lower false negative rate when the false positive rate is allowed to be large. This neural network is more suited to the task of identifying MACs in a situation where the penalty for accidentally classifying a MAC as "negative" must be avoided. The LDF, however, performs better in a situation where misclassifying negatives as "MAC" has a more severe penalty. Its ROC curve is lower in the region where the false positive rate is constrained to be small.

The difference between these two classifiers is apparent when the weights that connect the inputs to the hidden unit are treated as defining a vector. The angle between the perceptron vector and the discriminant function vector can be found from their inner product. The angle between the vector of the perceptron weights and the linear discriminant vector is 107°, which is 17° from orthogonal. Yet, both classifiers correctly classify more than 61% of the test data. This result demonstrates that there exist nearly orthogonal hyper-planes that separate significant portions of the two classes. This argues that a nonlinear decision surface should give better separation between MACs and negatives than any linear classifier.

3.2. Slide-by-slide results

It is the slide-by-slide classification rate of a classifier that is of clinical interest. In general, the cell-by-cell classifiers are used as a decision step in a larger classification system, used to diagnose the whole slide. For comparison purposes, the best MAC cell-by-cell classifiers (shown in the last column of Table 2) were applied to slide-by-slide classification. The classifiers were applied to all available cells on each slide.

The data used to train and test the slide-by-slide classifiers consisted of 93,494 cells taken from a total of 395 slides. All normal and severe dysplasia slides with at least 20 cells were included. There were 65,645 negative cells taken from 251 normal slides and 27,985 MAC cells taken from 144 severe dysplasia slides.

This set is a superset of the data used to design the cell-by-cell classifiers, introducing the possibility of a biasing error in the slide classification. This potential error is mitigated by the fact that the cells used to train the cell-by-cell classifiers comprised less than 10% of the total data. The statistical features used by slide classifiers are calculated based on all the available cells on the slide. Most slide datasets contain between 200 and 600 intermediate cells, up to 50 of which were used to design the cell classifiers. Their influence on the value of each slide feature is small and their effect on the results should be negligible.

Slide classification is typically done using the average classifier score or frequency of cells with score beyond a threshold for all the cells on the slide. For the neural networks, the frequency of cells (f_{MAC}) with a neural network score less than 0.25 (0 = MAC, while 1.0 = negative) was calculated for each slide. For the 251 normal slides, f_{MAC} averaged 0.16, while for the 144 severe dysplasia slides, f_{MAC} averaged 0.43. The exact value of f_{MAC} used for the decision boundary was jackknifed by removing each slide, in turn, from the data set and setting the threshold based on the remaining 394 slides. The jackknifed classification rate for f_{MAC} is given in Table 3, and the ROC curves for three of the classifiers listed in the table are shown in Fig. 3.

To compare the neural network results with those of the LDF several different slide features were calculated based on the LDF score for the respective cells. Using the "frequency of cells with

Table 3
The slide-by-slide classification rates of the best cell classifiers are shown for the neural networks with different numbers of hidden units. These are the jackknife results for a set of 251 normal slides and 144 slides containing severe dysplasia

Number of hidden units	Slide-by-slide classification rate
LDF	67.3%
1	68.1%
2	75.7%
3	73.9%
4	75.7%
5	76.2%
6	73.8%
7	75.4%

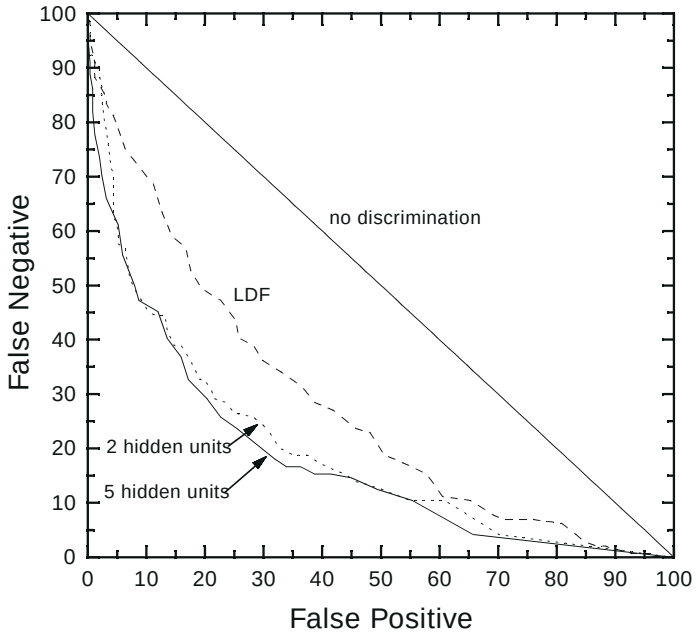

Fig. 3. The slide-by-slide ROC curves for the LDF and networks giving the best test set classification rate are shown. The ROC curves for the best networks using three, four, six and seven hidden units nearly coincide with those of the networks with two and five hidden units and are not shown.

score less than a particular cut-off", as was done with the neural networks, did not perform as well as simply calculating the average LDF score for the cells on each slide. The results for the best slide feature obtained for the LDF, average LDF score for each slide, are shown in Table 3 and Fig. 3.

Although we measured the frequency of cells that had a neural network score less than 0.25, this particular choice was somewhat arbitrary. One could have measured the frequency of cells that had a score less than 0.3 or 0.5, for example. In either case, the classification results that are within 2% of the values shown in Table 3.

There is significant improvement observed when using a neural network with two hidden units as opposed to a LDF for the first step of the classification procedure. There is at least a 5–8% increase in the slide-by-slide classification rate when using neural networks to classify cells. The networks with more than two hidden units offered no significant improvement over the network with two hidden units for slide-by-slide classification.

4. Discussion

In this study there was a significant improvement in cell-by-cell and slide-by-slide classification rates when neural networks were used in place of a LDF. As Table 2 shows, increasing the number of hidden units in the neural network beyond four overfitted the data causing the test set classification rate to decline. This occured even though there were 30 training examples per adjustable weight.

Part of the problem is that some of the features included in the neural network provide little or no discriminating value. Since a neural network is a nonlinear model of the data interaction it is difficult to predict *a priori* which features will prove useful to the network. Hence, one wants to include all possible discriminating features. Useless features, at best, increase the time it takes to train the network. At worst, they contribute to the overall prediction variance of the classifier.

The neural network with two hidden units performs in a manner similar to applying two successive discriminant function vectors to the MAC data. If we plot the projection of the training data along the two vectors defined by the two hidden units of the neural network we obtain the scatter plot shown in Fig. 4. This graph is a projection of the training data into a two-dimensional plane that best separates

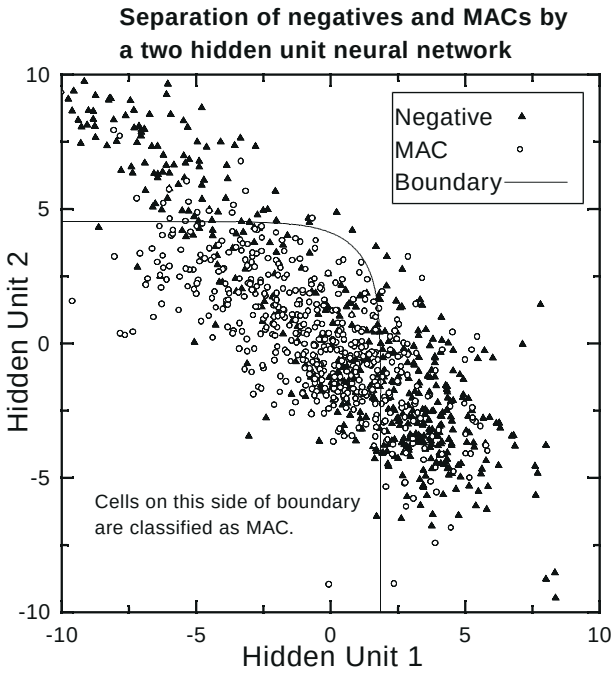

Fig. 4. The projection of the training data along the vectors that make up a two hidden unit neural network is shown. This network correctly classifies 71.9% of the test data. For clarity, only every eighth cell is shown in the plot. The decision bounday for this network is the curved line that separates cells in the lower left quadrant from the rest.

the negative cells from MACs. Each axis represents the contribution of each hidden unit towards separating the two data classes. This neural network gives the best classification rate of 71.9% correct classification of the test data. For clarity, only every eighth cell is plotted. Plotting other subsets of the data give graphs of similar appearance.

Figure 4 also shows the decision boundary between the two classes of data. It is a curved line that separates the MACs near the center of the graph from what appears to be two clusters of negative cells. This figure demonstrates the essential difference between the LDF and the two hidden unit network. The network has been trained such that it has found a two-dimensional plane in which the negative cells are primarily split into two clusters. The network is able to draw a curved decision boundary that separates these clusters from the MAC cells in the center of the graph. The performance of the nonlinear decision boundary to separate the two classes of objects demonstrates the advantage to training a neural network to classify the data.

It should be noted that the feed-forward neural network structures have been compared to a single linear discriminant function, a comparison that is bound to favor the neural network. In practice, one combines the LDF with judiciously selected thresholds or other LDFs in a stepwise fashion to perform the desired classification task. By doing this, one effectively constructs a nonlinear classifier from a sequence of linear classifiers. In this manner a difficult decision problem can be broken into a a sequence of more manageable smaller problems.

This paper does not suggest that one abandon this methodology in favor of attempting to find a single nonlinear classifier to do the whole task. Reducing a problem to smaller pieces allows us to design classifiers that are easier to train and should, in principle, lead to more robust classifiers. However, at a particular decision step in a tree, it may not be clear how to subdivide a problem further. Training a neural network to perform this decision task may provide a more powerful alternative than any linear, or quadratic, discriminant function.

Further, by examining the final form of the neural network we may discover how we can subdivide the decision problem further. Figure 4 shows the contributions of the two perceptron vectors of the two hidden unit network to separating the MACs and non-MACs. Further investigation may reveal merit in implementing this classifier as two successive linear classifiers. The magnitudes of the weights in the two perceptron vectors can demonstrate which features are useful in concert, or can even lead us to some derived features based on the vectors. The neural network has given us another way to look at our data, which may be the strongest argument in its favor.

5. Conclusions

This study has compared the correct classification rates of feed forward neural networks with that of a single LDF for detecting MACs versus truly normal cells. The neural networks correctly classified a 16,904 cell test set approximately 10% more accurately on a cell-by-cell basis than did the LDF. When more than four hidden units were used the neural networks began to overfit the training data, as evidenced by reduced test set classification rates. When the neural network and LDF classifiers were used to classify slides, the neural networks managed 6–8% higher correct sample classification rate based on the jackknifed results for a group of 395 slides.

Acknowledgements

The authors would like to acknowledge Deanna Haskins, Jagoda Korbelik and Paul Lam for their help in this study. This work was supported in part by the BC Cancer Agency, MRC of Canada

and Xillix Technologies Corp. The authors are grateful to Oncometrics Imaging Corp. who provided equipment and technical support for this study.

References

[1] G.F. Bahr, P.H. Bartels, H.E. Dytch, L.G. Koss and G.L. Wied, Image analysis and its applications to cytology, in: *Diagnostic Cytology and its Histopathological Bases*, L.G. Koss, ed., J.B. Lippincott Company, Vol. 2, 1992, pp. 1572–1612.
[2] M. Bibbo, A.G. Montag, E. Lerma-Puertas, H.E. Dytch, S. Leelakusolvong and P.H. Bartels, Karyometric marker features in tissue adjacent to invasive cervical carcinomas, *Analytical and Quantitative Cytology and Histology* **11** (1989), 281–285.
[3] M.E. Boon and L.P. Kok, Neural network processing can provide means to catch errors that slip through human screening of Pap smears, *Diagnostic Cytopathology* **9** (1993), 411–416.
[4] G. Burger, U. Jutting and K. Rodenacker, Changes in benign cell populations in cases of cervical cancer and its precursors, *Analytical and Quantitative Cytology* **3** (1981), 261–271.
[5] T.J. Colgan, S.F. Patten and J.S.J. Lee, A clinical trial of the AutoPap 300QC system for quality control of cervicovaginal cytology in the clinical lab, *Acta Cytologica* **39** (1995), 1191–1198.
[6] R.P. De Cresce and M.S. Lifshitz, PAPNET cytological screening system, *Laboratory Medicine* **22** (1991), 276–280.
[7] W.J. Dixon, *BMDP Statistical Software Manual*, Vol. 1, University of California Press, 1992.
[8] A. Doudkine, C. MacAulay, N. Poulin and B. Palcic, Nuclear texture measurements in image cytometry, *Pathologica* **87** (1995), 286–299.
[9] A.M.J. van Driel-Kulker and J.J. Ploem-Zaaijer, Image cytometry in automated cervical screening, *Analytical Cellular Pathology* **1** (1989), 63–77.
[10] S.E. Fahlmann, An empirical study of learning speed in back-propagation networks, Carnegie Melon University Tech. Report CMU-CS-88-162, 1988.
[11] M. Guillaud, A. Doudkine, D. Garner, C. MacAulay and B. Palcic, Malignancy associated changes in cervical smears: systematic changes in cytometric features with grade of dysplasia, *Analytical Cellular Pathology* **9** (1995), 191–204.
[12] G. Haroske, S. Bergander, R. Konig and W. Meyer, Application of malignancy-associated chanes of the cervical epithelium in a hierarchic classification concept, *Analytical and Cellular Pathology* **2** (1990), 189–198.
[13] J. Hertz, A. Krogh and R.G. Palmer, *Introduction to the Theory of Neural Computation*, Sante Fe Institute in the Sciences of Complexity, Addison-Wesley, 1991.
[14] M.L. Hutchinson, L.M. Isenstein, J.J. Martin and D.J. Zahniser, Measurement of subvisual changes in cervical squamous metaplastic cells for detecting abnormality, *Analytical and Quantitative Cytology and Histology* **14** (1992), 330–334.
[15] L.M. Isenstein, D.J. Zahniser and M.L. Hutchinson, Combined malignancy associated change and contextual analysis for computerized classification of cervical cell monolayers, *Analytical Cellular Pathology* **9** (1995), 83–93.
[16] R.A. Kemp, C. MacAulay and B. Palcic, Opening the black box: The relationship between neural networks and linear discriminant functions, *Analytical Cellular Patholgy*, this issue.
[17] S.J. McKenna, I.W. Ricketts, A.Y. Cairns and K.A. Hussein, A comparison of neural network architectures for cervical cell classification, *Third International Conference on Artificial Neural Networks*, IEEE Conference Publication 372, 1993, pp. 105–109.
[18] B. Stenkvist and G. Strande, Analysis of machine selected cells with an image analyser system in normal and abnormal cervical specimens, *Analytical Cellular Pathology* **2** (1989), 1–13.
[19] G.L. Wied, P.H. Bartels, M. Bibbo and J.J. Sychra, Cytomorphometric markers for uterine cancer in intermediate cells, *Analytical and Quantitative Cytology* **2** (1990), 257–263.
[20] D.J. Zahniser, K.L. Wong, J.F. Brenner, H.G. Ball, G.L. Garcia and M.L. Hutchinson, Contextual analysis and intermediate cell markers enhance high-resolution cell image analysis for automated cervical smear diagnosis, *Cytometry* **12** (1991), 10–14.
[21] D.J. Zahniser, P.S. Oud, M.C.T. Raaijmakers, G.P. Vooys and R.T. van de Walle, BioPEPR: A system for the automatic prescreening of cervical smears, *Journal of Histochemistry and Cytochemistry* **27** (1979), 635–641.