REIHE INFORMATIK
13/95

Neural Classifier Systems for Histopathologic Diagnosis

R. Stotzka[1], R. Männer[1,3], P.H. Bartels[2]
[1] Lehrstuhl für Informatik V, Universität Mannheim, Mannheim, Germany
[2] Optical Sciences Center, University of Arizona, Tucson, AZ
[3] Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, Heidelberg, Germany

A modified version of this technical report has been accepted for publication by 'Analytical and Quantitative Cytology and Histology', June 1995

Abstract

Neural network and statistical classification methods were applied to derive an objective grading for moderately and poorly differentiated lesions, based on characteristics of the nuclear placement patterns. Using a multilayer network after abbreviated training as a feature extractor followed by a quadratic Bayesian classifier allowed grade assignment agreeing with visual diagnostic consensus in 96% of fields from the training set of 500 fields, and a 77% of 130 fields of a test set.

1 Introduction

Prostate cancer is now the most commonly diagnosed cancer in men and cause of death from cancer among men in the Western world. Localized prostate cancer encompasses a wide spectrum of disease, with highly variable biological behavior and response to therapy. Prognosis thus varies widely. Grading and staging are the standard methods of prognostication. Numerous grading schemes have evolved [6, 8, 13, 12, 14, 21, 26, 22, 24, 25, 27, 33, 34]. The most widely used schemes are the systems of Gleason and Mostofi [13, 22], in practice many pathologists apply a three grade system combining criteria used both by Mostofi and Boecking [6, 22]. In the US, the Gleason system, which relies primarily on histologic features, has found widespread acceptance. It correlates well with other measures, such as ploidy [**?**], or PSA serum levels [28]. The Gleason system rests on a two tier assessment, of a primary grade, and an assessment of the predominant secondary grade, combining the two into a "Gleason sum". The Gleason system has been found to lead to variable scores between the different diagnosticians, for this, the second tier assessment may be responsible. As Mostofi points out [24], even in Gleasons's hands the Gleason grading system did not exeed a reproducibility of 80%. Poor reproducibility between diagnosticians is a major problem for the grading of prostatic lesions. A critical review of attempts to improve diagnostic and prognostic capabilities has been given by Mostofi [23, 24, 25].

The point is made that the sole reliance on histologic structure may be a major cause for diagnostic and prognostic lack of consensus among pathologists. In fact, the National Prostate Cancer Project Task Force, as reported by Murphy and Whitmore [27], recently recommended that future studies should consider nuclear and cytologic characteristics. Mostofi presents strong arguments for the inclusion of nuclear features for the grading of prostatic lesions [24, 25]. Bibbo et al. reported high reproducibility in primary grade assignments, for medium power microscopic fields, in a Gleason scoring scheme augmented by a set of nuclear grading features [5].

A number of research efforts are aimed at the development of an objective grading system for prostatic lesions, so the development of an interactive image comparison workstation with computer graphic enhancement of diagnosis clues and decision support by a Bayesian Interference Network [5] at the University of Chicago, the development of a diagnostic decision support system also based on an interference network for the classification of premalignant prostatic PIN lesions at the University of Ancona by Montironi et al. [1], and the image analytic studies by Irinopoulos and Rigaut to identify cases with poor prognosis, based on nuclear morphology [2].

For small, low grade lesions the outcome is highly predictable. This is also the case for widely metastatic high grade tumors. However, for most prostate lesions visual grading has been of limited value.

For an objective, automated procedure one has to keep in mind that lesions spanning the range from low grade to high grade form a continuum, characterized by increasing

loss of tissue differentiation, and increasing nuclear anaplasia. The decision maker is faced with the need to assign a given lesion to a fuzzy set—such as the fuzzy set of "poorly differentiated" lesions—and there exists a fairly wide zone of transition from moderately differentiated lesions to poorly differentiated lesions. The defining of discretely labelled diagnostic categories for a continuous process of progression should not lead to an expectation that a decision procedure now should "correctly" classify lesion. Also, certain histologic and nuclear features do not change at the same points along a grading scale, so that it is not readily possible to identify single "grade indicating" features. Nevertheless, if a monotonic, multivariate characterization could be found that might allow an objective determination of the grade for a given lesion, and if rules could be defined how the section or biopsy material should be systematically sampled to derive such a grading, then both reproducibility of grading, correlation to clinical outcome, and consistency of patient management might be attained.

Loss of tissue differentiation can be measured by assessing the nuclear placement pattern. In this study an effort is made to explore the utility of neural networks for assessing nuclear placement patterns. The advantage of this approach would be the potential for a high speed implementation, as the networks would be directly applied to the image domain. This would make extensive sampling feasible. Also, one would not have to pre–define discriminating features of the nuclear placement patterns, suitable for the intended mapping, as the network would learn these features by itself.

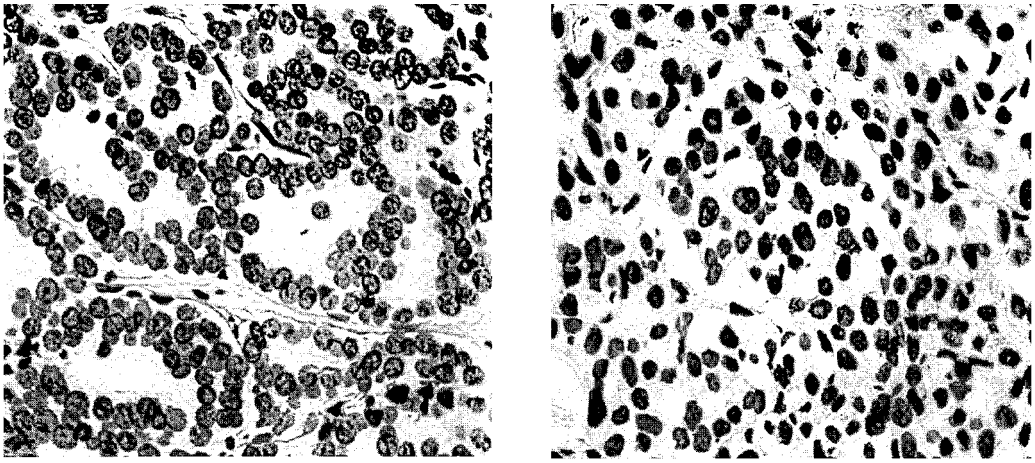

Figure 1: *Microscopic image of a moderately differentiated prostate lesion (left) and a poorly differentiated prostate lesion (right), $\times 400$.*

On the other hand, the well established usefulness of traditional classification methods is fully recognized. The second objective of this study was to examine the utility of such classification techniques based on features extracted from the imagery.

2 Materials and Methods

The clinical materials consisted of whole mount sections obtained from the National Registry of Pathology, AFIP, Washington, D.C., and histological sections from the University of Arizona, Dept. of Pathology. Sections had been cut to 5 micron, and stained by H&E.

Images were recorded under a 20:1 objective, N.A. 0.75, and a SONY MED 1234 three color CCD camera. For the study of nuclear placement patterns the nuclei of the digitized 512×512 images were segmented and their position represented in 64×64 binary pixel arrays by single pixels. Fig. 1 shows a field of a moderately and poorly differentiated lesion respectively, Fig. 2 shows the derived representations used to assess the nuclear placement patterns.

3 Classifier Design

In this section we give a brief discussion of the design of a classification system for moderately differentiated tissues (m) and poorly differentiated tissues (p), i.e., for the two classes ω_m and ω_p. Although most of the material presented here is widely known it is included for completeness.

A sample is described by an observation vector X of dimension n. The elements of X consist of the original data or of measured features.

3.1 Classification Strategy: Bayes Classifier

To decide whether an observation vector X belongs to ω_m or ω_p the Bayes classifier [11] can be used. Here the decision rule which maps X onto ω_m (resp. ω_p) is based on the probabilities $q_m(X)$ that the observation vector X belongs to class ω_m (resp. $q_p(X)$ for ω_p). If

$$q_m(X) > q_p(X)$$

X is considered as belonging to class ω_m. The a–posteriori probability $q_i(X),\ i \in \{m, p\}$ can be calculated according to the rule of Bayes using the frequency of occurrence P_i of a member of class i, the conditional density function $p_i(X)$, and the mixture density function $p(X)$ of both classes as

$$q_i(X) = \frac{P_i p_i(X)}{p(X)}\,.$$

The decision rule for ω_m can now be expressed as

$$P_m p_m(X) > P_p p_p(X)\,.$$

Thus the likelihood ratio $l(X)$ is defined as

$$l(X) = \frac{p_m(X)}{p_p(X)}$$

and the decision rule for ω_m becomes

$$\frac{p_m(X)}{p_p(X)} > \frac{P_p}{P_m}$$

The discriminant function $h(X)$ is defined as

$$h(X) = -\ln(l(X)) = -\ln(p_m(X)) + \ln(p_p(X))$$

and a test sample is thus assigned to class ω_m if

$$h(X) < \ln(\frac{P_m}{P_p})\,.$$

This comparison of probabilities is called the Bayes test for minimum error. The decision rule is theoretically optimal in minimizing the error probability.

However, to find the best classifier for two classes ω_p and ω_m the statistical properties of each class are needed. These are given by the conditional density functions $p_p(X)$ and $p_m(X)$. To estimate these functions requires sufficiently many training data. If the dimensionality of the observation vector X is high it is in most cases not possible to provide enough training data. The available data are sparsely scattered in the configuration space and reliable statistical distributions cannot be obtained.

In order to simplify the determination of the unknown statistics of the observation vectors, assumptions have to be made for the density functions. These lead to different classification schemes like the linear, quadratic, polynomial, k–nearest–neighbor classifier, etc.. All these classification algorithms tend to minimize the Bayes error.

Quadratic classification algorithm

Very often it can be assumed that the measured observation vectors are normally distributed. Then the $p_i(X)$'s are determined by the expectation vectors M_i and the covariance matrices Σ_i

$$p_i(X) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(X-M_i)^T\Sigma_i^{-1}(X-M_i)}\,.$$

If the frequencies of occurrence P_i are equal for both classes, the decision rule for ω_m becomes

$$h(X) = \frac{1}{2}(X-M_m)^T\Sigma_m^{-1}(X-M_m) - \frac{1}{2}(X-M_p)^T\Sigma_p^{-1}(X-M_p) + \frac{1}{2}\ln\frac{|\Sigma_m|}{|\Sigma_p|} < 0\,.$$

The decision boundary becomes a quadratic boundary in the n–dimensional space. This boundary separates the two classes optimally in the Bayesian sense. This classification algorithm often shows good classification results even if not all features or elements of the observation vector are exactly normally distributed.

3.2 Feature Extraction

In most cases it does not make sense to use original data as supplied by a data recording system directly for classification. As an example the binary images that represent the nuclear placement pattern used to grade prostate tissues contain 64×64 pixels. Thus there are $2^{4096} \approx 10^{1229}$ possible feature vectors so that the corresponding configuration space can never be explored for an estimate of the conditional density functions or the statistical parameters.

Instead the original data have to be preprocessed, i.e., features have to be extracted that still contain all information needed for the classification task but lack irrelevant information. In this way the dimensionality of the original data is reduced to the number of features used. Often some preprocessing steps are obvious. In our case, e.g., it is irrelevant for the classification if a microscopic image is shifted or rotated. Without extracting translation and rotation invariant features every translation and rotation of an image is considered as an independent sample. More often, however, expert knowledge on the classification problem has to be used in feature selection.

In most classification tasks the selection and extraction of good features is the most important and sometimes also the most demanding work. The quality of the optimal classifier expressed as the Bayes error depends only on the choice of the feature vector.

3.3 Principal component analysis

For efficiency reasons, a small set of features should be found that represent the samples accurately. But mostly all that can be done is to identify a larger number of features that only *could* be important for the classification. If chosen in this way some information represented by the features may be redundant, and some components of the feature space may contain no information at all. But even if this is not the case some features are less important than others and could be dropped without noticeably diminishing the classification ability.

In principal component analysis (PCA) the information on the feature space is examined and expressed in terms of the variances of the samples in several directions in the space. Directions with small variances correspond to features that are very similar for all presented training patterns. These features can be eliminated. The usual procedure is the following:

1. Calculation of the mixture covariance matrix.

2. Determination of the eigenvectors and eigenvalues of the covariance matrix.

3. Ordering of the eigenvectors according to the eigenvalues λ_i in descendant manner.

4. Elimination of eigenvectors with the smallest eigenvalues. These are directions in the feature space with the smallest variances.

5. Mapping of the feature vector X into the space of the remaining eigenvectors. Thus the dimension is reduced with a minimal loss of information.

3.4 Neural Networks as Classifiers

The motivation for using artificial neural networks to solve complex tasks comes from the information processing power of biological brains. These are capable to perform extremely difficult pattern recognition and classification tasks with noisy and incomplete data. One example is the recognition of partially hidden faces. Brains do such tasks by orders of magnitude faster than today's supercomputers. It is thus worthwhile to investigate brain–like information processing structures for tasks as the classification of prostate tissues.

Extremely simplified the human brain consists of about 10^{10} processing elements called neurons. They are connected by a network of about 10^{13} synapses. Every synapse transfers the activity of a neuron to another one according to the synaptic weight which expresses the relative importance of the influence of the source neuron to the destination neuron. Every neuron sums the arriving weighted inputs and answers with a signal to the connected neurons if the sum exceeds a certain threshold. The complex and parallel dynamics of the network effect the huge computing power of a brain. Two attributes influence this dynamics.

- External inputs are given by sensory perception and stimulate the neuron's activities. Starting from a certain neuron state this state changes on a fast time scale as determined by the synaptic weights.

- In addition the weights of the synapses change on a slow time scale depending on the current activities of the neurons. This process is called learning.

Thus the neurons represent the processing elements of the brain, the synapses the memory.

Artificial neural nets imitate only the most basic features of real neural networks. Moreover only very small networks of typically up to 10^3 neurons can today be set up in hardware or can be simulated. Two learning schemes can be distinguished. In supervised learning the learning samples and the desired network response are known and simultaneously presented to the network. The output of the network is compared with the desired output, and the synaptic weights are adjusted to diminish the difference. In unsupervised learning only the input patterns are presented and the network finds output classes by self–organization.

Supervised Learning Classifier Systems

If an input pattern and the corresponding output pattern is presented the neural network should change its internal state so that it answers with the corresponding output if a similar input pattern is presented after learning.

The simplest model of an artificial neural network is the perceptron proposed by Rosenblatt [30]. It consists of a layer of input neurons S_i which are activated by the input pattern, and a single output neuron o. It has synaptic links of weight w_i to all input neurons. The output neuron sums the weighted input and produces an output using a nonlinear function, e.g., a threshold or a sigmoid function $f(x)$

$$o = f(\sum_{i=1}^{n} w_i S_i)$$

$$f(x) = \frac{1}{1+e^{-x}} .$$

The most common (supervised) learning rule is the delta rule [30]. It calculates the difference between the network output and the desired output and adjusts the weights whenever a learning sample is presented. This has to be done iteratively.

Unfortunately the perceptron is capable to distinguish only linearly separable classes [20], i.e., classes that have hyperplanes as the decision boundary in the configuration space. To perform more complex classification tasks the architecture of the perceptron has been extended to the multilayer perceptron [32]. The multilayer perceptron consists of two or more layers of perceptrons. A proposed iterative learning rule for this architecture is error backpropagation [32]. This learning rule starts with an arbitrary decision boundary and tends to minimize the classification error in a successive process. Every time a new pattern is presented the output of a network is calculated and the classification error estimated. The error is then propagated backward through the network to adjust the weights for all layers. This process is called learning cycle. When all training patterns have been presented, an epoch is complete. Learning in a multilayer perceptron is very slow. Even a small problem like XOR [32] needs approximately 1000 epochs until the function was learned.

In principle the multilayer perceptron is able to learn every classification task [17]. But for large networks and difficult classification tasks a multilayer perceptron requires an extremely high computational effort and an extremely large training set, i.e., extremely long learning time. Moreover the convergence of the learning algorithm to correct synaptic weights is not guaranteed. Depending on the initialization of the weights the algorithm may converge toward a local minimum and never reach the real optimum, the Bayes error.

Since the convergence of backpropagation learning is very slow many acceleration algorithms have been proposed in recent years [10, 37]. Hofmann [16] has shown that these techniques tend to work very well and much faster for small problems, but generally do not improve the convergence speed for complex classification tasks.

4 Results

To build a classification system we have investigated different methods. First the classification abilities of a neural network were tested on artificial data fields similar to the binary nuclear placement patterns. The chosen artificial classification problem was much simpler than the grading problem. Second the binary prostate data samples were processed and the classification abilities of statistical classification schemes and neural networks were measured. Eventually the different advantages of the neural networks and statistical classification methods were combined to build a hybrid classification system.

4.1 Neural Network Processing of Artificial Data Fields

4.1.1 Artificial data fields

4000 data fields were created to test the classification abilities of multilayer perceptrons on binary nuclear placement patterns. Each field consists of 30×30 pixels with 36 points placed on a regular grid. The position of each point is varied randomly according to a normal distribution. To create two classes only one parameter is varied: the standard deviations differ by about 40% (Figure 4).

Thus the two classes are linearly separable. We chose this kind of data to test the ability of neural networks to distinguish subtle differences in the statistics of random dot patterns as a simple model of the nuclear placement patterns of prostate nuclei in tumour tissues.

4.1.2 Multilayer perceptron classification of artificial data fields

To classify the artificial data fields by a multilayer perceptron the input pattern was presented as a 900–dimensional binary vector to the input layer. It turned out that this classification task is easy to learn—in contrast to the difficulties that a person encounters. We tested several multilayer perceptrons with a different number of hidden units. Only a perceptron is necessary to learn the training samples and to classify all 300 test samples correctly. This result is in accordance with the theory of Minsky and Papert [20] after which a simple perceptron is sufficient to learn linearly separable classification tasks.

Figure 5 shows the number of training epochs needed until the test set was recognized completely versus the size of the training set. Training with only a few (e.g. 10) training samples needs many epochs, but the the network is still able to generalize the difference of the variance of the displacement to perform the classification task.

4.2 Quadratic Classifier

Like the test problem of classifying artificial data, the given classification task, the automatic grading of prostate tumours, is difficult for human observers, even for well trained pathologists. In order to simplify the problem and to make it more comparable to the test problem, we did not aim at assigning a certain grade to an unknown sample but rather to distinguish two grades of prostate tumours,moderately and poorly differentiated ones. For that a quadratic classifier as described above was applied.

Applying the classifier to the original data was not possible. The size of the binary images of 64×64 results in an observation vector of length 4096. Only 250 training samples per class were available. These were by far not sufficient to estimate the statistical distribution of the data in the observation space. Thus feature extraction is required to reduce the dimensionality of this space, and the observation vector contains selected features.

4.2.1 Feature selection

The identification of useful features is based upon some knowledge on the given problem. Since the original observation space is so huge it is important to exploit all available a–priori knowledge, problem independent as well as problem dependent. In addition assumptions have to be made whose validity can only be tested. These assumptions have to be based on expert knowledge about differences between moderately and poorly differentiated tissue sections.

In our case the only problem independent knowledge available is the obvious fact that the correct assignment of a presented sample to one of the two grades cannot depend on the position and orientation in which the sample is presented to the system. Thus all features have to be invariant under translation and rotation. All other features relate to the spatial distribution of the pixels in the image, the positions of cell nuclei. To reduce the dimensionality of the observation space two approaches are possible. Some *global* features can be extracted that compute a few numbers from the whole image. However there are more *local* features which extract a value by assessing only onto a rather small area of the image. Hereby it is assumed that many clues for a correct assignment can be found in these small areas and that an appropriate combination of such clues leads to a more confident classification.

If local feature detectors are applied to all parts of an image in all orientations (if the detector is not itself rotational invariant), one or a few numbers are computed for every feature at every position and orientation (if not rotational invariant). In the simplest case the output of every feature detector is averaged, so that invariance against translation and rotation is obtained. One should note that even this primitive kind of preprocessing reduces the dimensionality of the observation space by a huge factor. There are, e.g., $2^{100} \approx 10^{30}$ feature detectors *in total* that look onto a field of 10×10 pixels. Assuming that only structures not larger than this are relevant for the classification, the $2^{4096} \approx 10^{1232}$–

dimensional observation space is reduced by about 1200 orders of magnitude without loss of information! Thus it is still difficult but not impossible to find the classification boundary in the reduced space. The dimensionality of the observation space can be reduced further if not *all possible* features detectors are used but only those that might be problem specific.

In Table 1 the most important features found are described. The discriminating ability of the selected feature using a quadratic Bayes classification algorithm is also given.

Feature class	**Number of parameter values**	**Accuracy on the training set (%)**	**Accuracy on the test set (%)**
Number of cell nuclei	1	63	60
Circle masks	9	66	59
Holes	8	66	63
Circle masks and holes	17	74	67
Linear aggregates	5	58	54
Local density distribution	2	67	66
Walsh basis functions	28	79	55
Fractal dimension	15	69	62
Euler relation	2	68	68
Minimum distances	2	55	54
Local co–occurrence	120	94	58
Local contrast	120	95	60
Neural features	63	96	77

Table 1: *Accuracy of feature classes averaged over different assemblies of training and test sets. The feature classes are explained in the text.*

In principle two different kinds of features can be distinguished: *structual features* describe information which correlates with histopathologic attributes, *texture and statistics describing features* extract information about statistical attributes and the "texture" of tissue structures.

Structural features

Number of cell nuclei: The number of cell nuclei in an image varies from 50 to 350. The cellularity of poorly differentiated tissues is higher than that of moderately differentiated ones. A moderately differentiated tissue section contains in the average 119 nuclei, a poorly differentiated one 150. This means that the number of cell nuclei in an image is an important attribute for the discrimination.

Circle masks: Prostate tissues consist of glands, roundish to elliptically shaped histologic structures lined by glandular cells. The proliferation of malignant glandular

nuclei disrupts the glandular structure, leading to a loss of tissue differentiation, ranging from well-differentiated lesions to poorly differentiated lesions. In moderately differentiated lesions the gland structure appears partially disrupted to several disturbed, in poorly differentiated lesions the glandular structure is practically no longer discernible (see Figure 1).

Thus, the detection of structures of circular to elliptical shape in the binary nuclear placement patterns in moderately and poorly differentiated lesions may provide a diagnostic discrimination criterion. The boundary of fragments of round structures is determind by convolving the binary pixel images with circle masks. Different masks with circle diameters from 1 to 9 pixels are used. The averages of the convolved images represent the resemblance of the processed histologic structures to circular structures as a function of diameter.

Holes: The lumen of a gland is free of nuclei in normal and well differentiated lesions, and even in moderately differentiated lesions few nuclei are located in the interior of the gland. Larger open areas free of nuclei are less frequently found in poorly differentiated lesions. Circular holes, not enclosing any nuclei are counted to measure this diagnostic clue. The diameter of the hole masks is varied from 2 to 9 pixels to detect structures of different sizes.

Linear aggregates of nuclei and isolated nuclei: In moderately differentiated lesions segments of glandular epithelium are preserved. Here, nuclei form linear aggregates. In poorly differentiated lesions isolated nuclei predominate. The number of linearly aggregated pixels representing cell nuclei in different orientations is searched by convolution of the binary images with 3×3 masks containing lines of different orientation each. The averages of the resulting images are used to measure the quantity existing lines.

To detect isolated nuclei the number of isolated pixels in the middle of an 3×3 area is counted.

Texture and statistics describing features

Local density distribution: The spatial density of nuclei differs within the tissue sections. In moderately differentiated sections the remaining glandular structures lead to density fluctuations within an image. In Figure 6 it is obvious that poorly differentiated sections are more homogeneous in the nuclear density.

Counting the nuclei within subareas of 10×10 pixels the average local density of the image and the local density variance are determined.

Walsh basis functions: Walsh basis functions can be used for oriented frequency filtering [7]. We use these functions as frequency filters (Figure 7) to examine periodical structures in the binary nuclear placement patterns. The binary images are convolved with a basis filter, and the resulting grey level image is averaged.

Fractal dimension: Structures within binary images are composed of the elementary structures (a) isolated points (dimension: 0), (b) lines (dimension: 1) and (c) areas (dimension: 2). The fractal dimension of a given structure is, roughly speaking, a measure for the similarity between this structure and these elementary structures. It is determined by looking at a given structure with different resolution. Resolutions of size 2, 4, 8,..., 64 are examined. In an image derived from the original one by lowering the resolution, a pixel was set if at least one pixel in the corresponding area of the original 64×64 image was set. The fractal dimension is calculated as the number of set pixels as a function of the resolution.

To describe the properties of nuclear placement pattern of prostate tissue sections the fractal dimension can be used to distinguish structures consisting of linear aggregates of nuclei or of single nuclei.

Euler relation: The Euler relation [35] determines for an image the number of connected areas minus the number of included holes. In the original 64×64 pixel images only a few connected areas exist. To measure structural information the resolution was reduced to 32×32 respectively 16×16.

Minimum distances: In some sections the remaining segments of glandular structures do not resemble round or elliptic structures. Rather, nuclei form linear aggregates of arbitrary shape. To detect the existence of such fragments of glandular epithelium we measure the average and the variance of the distance of cell nuclei to their nearest neighbors. The nearest neighbor of a point as a member of a linear aggregate is most likely also a member of this aggregate and the distances between both are very low. If a point is not member of a linear aggregate its nearest neighbor has a greater distance.

Local co–occurrence: We use the local co–occurrence and the local contrast to determine some statistical properties of the distribution of pixels within a neighborhood. The covariance between pixels in different distances Δx and Δy up to 11 pixels is measured:

$$cov(\Delta x, \Delta y) = E\{[I(x,y) - E\{I(x,y)\}][I(x+\Delta x, y+\Delta y) - E\{I(x,y)\}]\},$$

where $I(x,y)$ is the value of the pixel on position (x,y) and $E\{z\}$ the expectation value of z.

Local contrast: The local contrast is defined as the squared difference of two pixels at several distances [39]. Here distances of up to 11 pixels were used:

$$con(\Delta x, \Delta y) = E\{[I(x,y) - I(x+\Delta x, y+\Delta y)]^2\},$$

4.2.2 Classification results

To design a classifier system the best combination of features should be determined. Applying principal component analysis to all 400 features is not practical since all eigenvalues and eigenvectors of a 400×400 matrix have to be calculated.

We used random search to find good combinations of features. From all considered features 10 to 80 were randomly selected for the feature vector. A training set and a test set were also randomly selected from all available images. Principal component analysis was applied and the parameters of a quadratic Bayes classifier were calculated according to the formulas in section 3.1. The resulting classification success for the training set and the test set (a consequence of generalization) is shown in Figure 8 for some arbitrary feature combinations. Each point represents a quadratic classifier. The x–axis shows the accuracy on the training set; the y–axis the accuracy on the test set, i.e., the generalization success.

The discriminating abilities of a feature combination of high training and generalization success are shown in Figure 9. The classifier is a quadratic Bayes classifier using a promising set of features of Figure 8. We chose a classifier with a high accurancy on the training set and a high accurancy of the test set. Feature combinations with these discriminanting abilities are found in the upper right area of the cluster of classifiers in Figure 8.

A single point in Figure 9 represents a sample of class "moderately differentiated" (m) or class "poorly differentiated" (p). Each sample vector X is mapped onto the x–axis according to

$$\frac{1}{2}(X - M_m)^T \Sigma_m^{-1}(X - M_m)$$

and to the y–axis according to

$$\frac{1}{2}(X - M_p)^T \Sigma_p^{-1}(X - M_p) \ .$$

These terms are called the Malahanobis distances of a vector X to class m resp. p. It is obvious that both classes can be easily separated. The classification success of this classifier is 98.7% on the training set and 78.3% on the test set. The likelikood ratio distribution of the training samples concerning the described quadratic classifier is shown in Figure 10. The small overlap of both curves represents the remaining misclassification rate of 1.3% on the training set.

4.3 Neural Network Classifier

Because neural networks are capable to find their own features by self–organization, the original data of the binary images were chosen as input to the system. In principle the data processing steps in a multilayer perceptron are similar to those in a statistical classification system. The patterns are presented to the input layer and processed through the hidden units to the output unit. The number of hidden units is much smaller than the

number of input units to force the network to process only significant information for the discrimination process to the output neuron. This processing step can be regarded as a feature extraction by self–organization.

For two reasons it is not practical to present the full 64×64 input image to the neural network directly. First, the array is too large. The required computation time grows approximately with the square of the number of input units [16], because all connections must be modified during a learning step. Even a multilayer perceptron with 45×45 inputs uses several weeks computing time on a SUN 4. Second, the histologic structures whose preservation is indicative of the remaining differentiation may not fall into the imaged 512×512 field. No provisions are made to pre–position the subimages so that such structures would be fully covered when presented to the neural network. An adjustment would not be feasible anyway in a fully automated system. Such a system would have to rely on features which are independent of translation and rotation. Therefore, the following procedure has been used.

1. The position of a 45×45 subimage was randomly chosen. With respect to the size of the assumed typical structures observed in Figure 1 this size should be big enough. By choosing the location of the subimage randomly we attained translation invariance.

2. The subimage was randomly turned by 0, 90, 180, or 270 degrees to get a very rough approximation of rotation invariance.

3. The rotated subimage was presented to the network.

For the classification of a specific sample it is not sufficient to present only one subimage to the network. By chance the selected subimage may not contain typical structures of a class so that classification will fail. We therefore presented 20 randomly chosen and rotated subimages of a sample and we averaged over all outputs.

4.3.1 Multilayer perceptron

The classification abilities of multilayer perceptrons trained with an error backpropagation learning rule depend mainly on the network architecture, the initial weights and the learning parameters. In principle a multilayer perceptron consisting of one input, one hidden, and one output layer is sufficient to learn every binary mapping [29].

We tested such multilayer perceptrons with different numbers of hidden units and different learning parameters. The number of hidden units was varied from 10 to 100, and the learning rate from 0.1 to 1.0. We trained a multilayer perceptron until the network did not further improve its recognition abilities on the training set. Training of a multilayer perceptron with 35 hidden units required three weeks computing time on a SUN 4. Thus the training and test sets could not be varied and the classification results depend on the chosen sets.

The multilayer perceptron producing the best results consists of 2025 input units, 35 hidden units, and one output unit; it uses a learning rate of 0.4. Such a neural network is able to classify the training set with an accuracy of 82% and the test set with an accuracy of 65% (Figure 11).

4.3.2 Hybrid classification system

The advantage in using multilayer perceptrons as classification systems is the self–organizing feature extraction process performed by the neural network. The main disadvantages are the computational expenses, the difficult tuning of parameters, and the uncertain convergence. Figure 12 show a part of a trained multilayer perceptron and the synaptic connections of the hidden units to the input layer. It turns out that these units organize filters which are sensitive to different properties of the input patterns. It can be shown that the principal structure of these filters develops very early at the beginning of learning, but the network classifies the patterns wrongly because the synapses of the output unit are still badly developed. Learning in multilayer perceptrons generally tends to optimize the whole system at once. If a pattern is classified wrongly the learning algorithm modifies all synapses and not only those of the output unit. This is one reason for the slow convergence of the back–propagation learning rule.

In contrast the use of statistical methods as shown above is based on predefined features. If effective features are found they result in good classification systems. Regarding the classification of prostate tumours the costly part in designing a statistical classifier is the feature extraction.

It is possible to combine the advantages of multilayer perceptrons—the self–organizing feature extraction—and the outstanding discriminating abilities of statistical classification systems in a hybrid system. As a study we trained four multilayer perceptrons with different initial synaptic initializations and different number of hidden units. The training was aborted after $4-6$ days. This is approximatly $\frac{1}{4}$ of the training time that a multilayer perceptron normally requires for convergence. At this stage the neural network had not yet learned to discriminate the classes but it had already developed a set of more or less specific feature filters, one for every hidden unit. Important hidden units—and thus feature filters— were selected according to the strength of their synaptic link to the output unit. In order to use these feature filters for a statistical classifier, the strengths of the synapses from such a hidden unit to the input layer were interpreted as a 45×45 image. These images were used as convolution masks which were applied to the sample images. The average of the pixel values of the resulting 20×20 image is defined as a neural feature (Figure 13). We extracted the neural features of 63 different hidden units of our trained networks. To this feature vector, principal component analysis and quadratic Bayes classifiers were applied to determine the discriminating surface. The discriminating abilities of such a classifier using 63 features is shown in Figure 14.

Both classes (m and p) can be separated and the boundary between the clusters is

nearly as sharp as in the case were features have been defined "by hand". The classification success of this hybrid classifier is 96.0% on the training set and 77.3% on the test set. This example shows that the information filters developed by self–organization extract in principle the same information as the features described in chapter 4.2.1. Combining these features with neural features improves the generalization ability only by about 2%. Thus the Bayes classifier and the hybrid classification system—a combination of a neural network and statistical classifier—do not show significant differences in their classification results. The remaining generalization error of ≈22% could not be reduced further.

5 Conclusions

The objective assessment of tissue differentiation in prostatic lesions is a complex problem. Algorithms for automatic classification like the quadratic Bayes classifier or multilayer perceptrons are in principle able to perform the classification task. The advantage of the Bayes classification algorithm is its accuracy after a set of sufficiently discriminating features is determined. Its disadvantage is the time required to compute all feature values of a given binary nuclear placement pattern. Therefore this method seems to be inefficient to assess the grade of large tissue areas. On the other hand multilayer perceptrons as an example for a neural network classification system do not need expensive feature extraction by hand as a preprocessing step. Important features are developed by self–organization as connections between the layers of units during the training of the network. After training the resulting network could be easily implemented in hardware resulting in a very fast classification system. Unfortunately training takes 3 weeks of computing time, the convergence to an optimal discriminating surface is not guaranteed, and the resulting accurancy is worse than that of the Bayes classifer.

A hybrid classification scheme as a combination of a multilayer–perceptron and a quadratic classifier is proposed. It does not show significant differences in the classification abilities to the Bayes classifier and the preprocessed feature extraction. This may lead to the assumption that the classification tasks and the statistical properties of the binary image may be nearly optimally described by the features. The main advantages of the hybrid classifier are:

1. The training costs are reduced.
2. The feature filters develop by self-organization.
3. High accuracy.
4. The convolution and discriminant surface can easily be implemented in hardware for high speed histologic grading of large tissue fields.

The binary nuclear position images carry only information about the glandular structure of the tissue. By mapping the microscopic images to these images the information of

nuclear properties like size, structure, and form is lost. Adding these informations to the nuclear features may lead to better classification results.

In large tissue fields this improved classifier can be applied to assess the primary grade of local areas and to determine the distribution of the primary grades within the field. According to the statistics the secondary grade is computed and the "overall" grade is concluded. This automatic grading system may lead to more reliable results in clinical diagnosis.

6 Acknowledgments

This work has been supported by the NATO scientific division under grant No. CRG 910556. We acknowledge the many fruitful discussions with H. Horner and R. Kühn, both Institute of Theoretical Physics, University of Heidelberg, Heidelberg, Germany.

References

[1] Feb 1992. personal communication.

[2] Oct 1992. personal communication.

[3] Jan P.A. Baak. *Manual of Quanitative Pathology in Cancer Diagnosis and Prognosis.* Springer, Berlin,, 1991.

[4] Peter H. Bartels, M. Bibbo, A.R. Graham, S. Paplanus, R.L. Shoemaker, and D. Thompson. Image understanding system for diagnostic histopathology. *Analyt. Cellular Pathol.*, 1:195–214, 1989.

[5] M. Bibbo, P.H. Bartels, T. Pfeifer, D. Thompson, C. Minime, and H.G. Davidson. Belief networks for grading prostate lesions. *Analytical and Quantitative Cytology and Histology*, (15):124–135, 1993.

[6] A. Boecking and W. Auffermann. Cytologic grading of therapy induced tumor regression in prostate carcinomina: Proposal of a new system. *Diagn. Cytopathology*, 3:108–111, 1987.

[7] Gonzalez C. and Woods R.E. *Digital Image Processing.* Addison-Wesley, Reading, MA, 1992.

[8] T.D. Clark. Nuclear roundness factors, a quantitative approach to grading in prostate carcinomina, reliability of needle biopsy tissue and effect of tumor stage on usefulness. *Prostate*, (10):199–206, 1987.

[9] C. Di Loreto, B. Fitzpatrick, S. Underhill, D.H. Kim, H.E. Dytch, H. Galera-Davidson, and M. Bibbo. Correlation between visual clues, objective architectural features, and interobserver agreement in prostate cancer. *American Journal of Clinical Pathology*, 96,(1):70–75, July 1991.

[10] S.E. Fahlmann. An emperical study of learning speed in backpropagation networks. Technical Report CMU-CS-88-162, Carnegie Mellon University, Pittsburgh, PA, 1988.

[11] Keinosuke Fukunaga. *Statistical Pattern Recognition.* Academic Press, Harcourt Brace Jovanovich, Boston, 1990.

[12] D.F. Gleason. Classification of prostatic carcinomina. *Cancer Chemother. Rep.*, (60):125–128, 1966.

[13] D.F. Gleason. Gleason grading system. In D.G. Bostwick, editor, *Pathology of the Prostate*, pages 83–93. Churchill Livingstone, New York, 1990.

[14] D.F. Gleason and the Veterans Administration Cooperative Urologic Research Group. Histologic grading and clinical staging of prostatic carcinomina. In M. Tannembaum, editor, *Urologic Pathology: The Prostate*, pages 171–198, Philadelphia, 1977. Lea and Febinger.

[15] J. Hertz, A. Krogh, and R.G. Palmer. *Introduction to the Theory of Neural Computing*, volume 1 of *Santa Fe Institute, Studies in the Sciences of Complexity: Lecture Notes.* Addison-Westley Pub., Redwood City, 1991.

[16] A. Hofmann. Backpropagation Acceleratoren. Studienarbeit, Informatik V, University of Mannheim, 1992.

[17] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[18] T. Kohonen. *Self-Organization and Associative Memory.* Springer-Verlag, Berlin, 1989.

[19] Gary J. Miller. New developments in grading prostate cancer. *Seminars in Urology*, 8(1):9–18, February 1990.

[20] M.L. Minsky and S.A. Papert. *Perceptrons.* MIT Press, Cambridge, 1969.

[21] F.K. Mostofi. Problems in grading carcinomina of prostate. *Semin. Oncol.*, 3:161–169, 1976.

[22] F.K Mostofi. Grading of prostatic carcinomina: Current status. In A. Bruce and J. Trachtenberg, editors, *Adenocarcinomina of the Prostate*, pages 29–46. Springer, New York, 1987.

[23] F.K. Mostofi, C.J. David, and I.A. Sesterhenn. Pathology of carcinomina of the prostate. *Cancer*, (70):235–253, 1992.

[24] F.K. Mostofi, I. Sesterhenn, and C.J. Davis. Prostatic carcinomina: Problems in the interpretation of prostatic biopsies. *Hum. Pathol.*, (23):233–241, 1992.

[25] F.K. Mostofi, I. Sesterhenn, and C.J. Davis. A pathologist's view of prostatic carcinomina. Number 71 in Interational Histological Classification of Tumours, pages 906–932. World Health Organisation, Geneva, 1993.

[26] F.K. Mostofi, I. Sesterhenn, and L.H. Sobin. *Histologic Grading of Prostate Tumors.* International Histological Classification of Tumours. World Health Organisation, Geneva, 1980.

[27] G.P. Murphy and W.F. Whitmore. A report of the workshops on the current status of the histologic grading of prostate cancer. *Cancer*, (44):1490–1494, 1979.

[28] H. Potear, W. Welch, and D. Sacks. Gleason grade correlates with serum prostate specific antigen levels. Technical report, Brigham and Women's Hospital and Harvard Medical School, 1993.

[29] B.D. Ripley. Statistical aspects of neural networks. In E. Barndorff-Nielsen, D.R. Cox, J.L. Jensen, and W.S. Kendall, editors, *Networks and Chaos – Statistical and Probabilistic Aspects,*. Chapman and Hall, 1993.

[30] F. Rosenblatt. *Principles of Neurodynamics.* Spartan, New York, 1962.

[31] J. Rubner and P. Tavan. A self-organizing network for principal-component analysis. *Europhysics Letters*, 10:693–698, 1989.

[32] D.E. Rumelhard and J.L. McClelland. *Parallel Distributed Processing.* The MIT Press, Cambridge, MA, 1986.

[33] F.H. Schroeder, W.C. Hop, J.H. Bloom, and F.K. Mostofi. Grading of prostate cancer: Analysis of the prognostic significance of single characteristics. *Cancer*, (6):81–100, 1985.

[34] F.H. Schroeder, W.C. Hop, J.H. Bloom, and F.K. Mostofi. Grading of prostate cancer: Multivariant analysis of prognostic parameters. *Cancer*, (7):13–20, 1985.

[35] J. Serra. *Image Analysis and Mathematical Morphology.* Academic Press, 1982.

[36] Rainer Stotzka, Reinhard Männer, and Peter H. Bartels. Neural network grading of prostate tumors. In Omar Benhar, Carlo Bosio, Paolo Del Giudice, and Eugenio Tabet, editors, *Neural Networks: From Biology to High Energy Physics*, pages 323–333, 1991.

[37] S.B. Thrun, J. Bala, E. Bloedorn, I. Bratko, B. Cestnik, J. Cheng, K. De Jong, S. Dzeroski, S.E. Fahlman, D. Fisher, R. Hamann, K. Kaufman, S. Keller, I. Kononenko, J. Kreuziger, R.S. Michalski, T. Mitchell, P. Pachowicz, Y. Reich, H. Vafaie, W. Van de Welde, W. Wenzel, J. Wnek, and J. Zhang. The monk's problem: A performance comparison of different learning algorithms. Technical Report CMU-CS-91-197, Carnegie Mellon University, 1991.

[38] W.M. Weiss and C.A. Kulipowski. *Computer Systems that Learn.* Morgan Kauffmann Pub., San Mateo, CA, 1990.

[39] Chung-Ming Wu and Yung-Chang Chen. Statistical feature matrix for texture analysis. *Graphical Models and Image Processing*, 54(5):407–419, 1992.

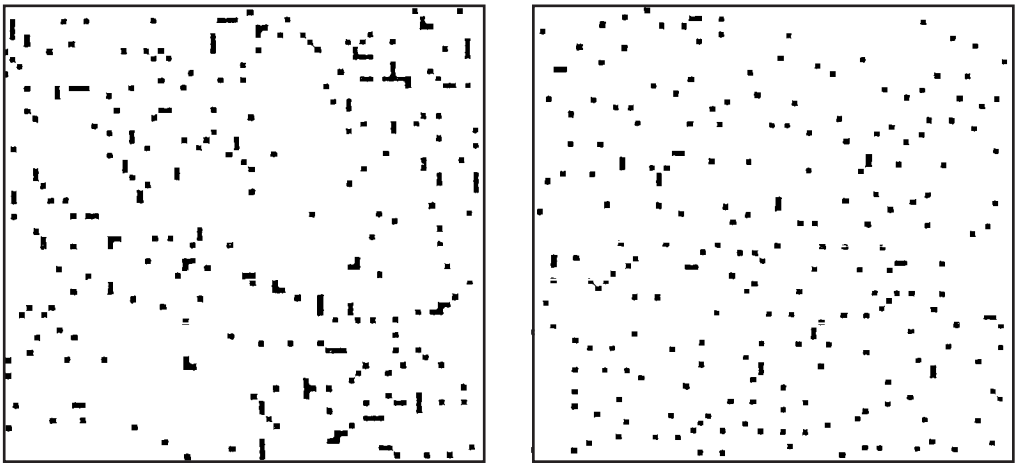

Figure 2: *Reduced and binary representation of nuclear placement for Fig. 1.*

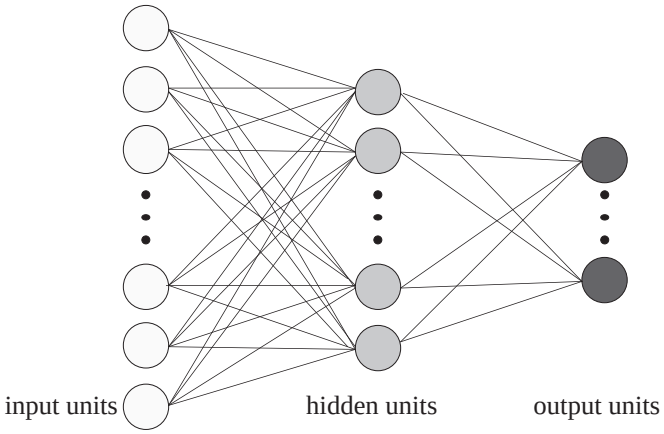

Figure 3: *Structure of a multilayer perceptron.*

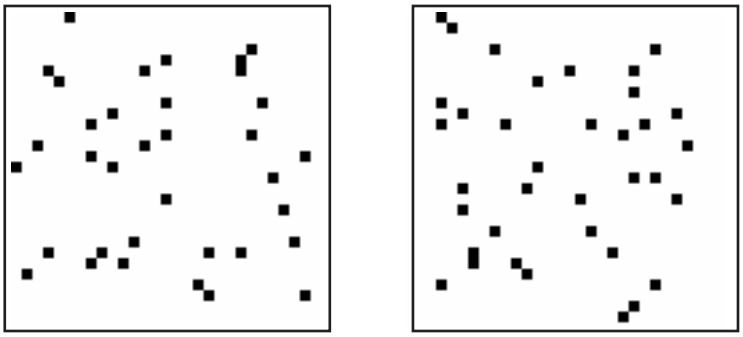

Figure 4: *Artificial data fields. The pixels of both images are put on a regular grid and displaced randomly. The variance of the displacement of the right image is 40% larger than of the left one. The human eye is not capable perceive and assess this difference.*

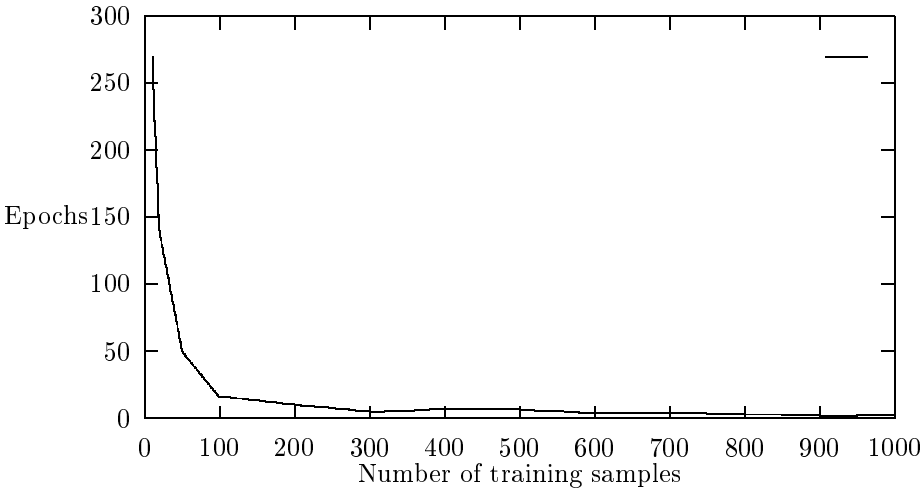

Figure 5: *Training duration vs. number of training samples.*

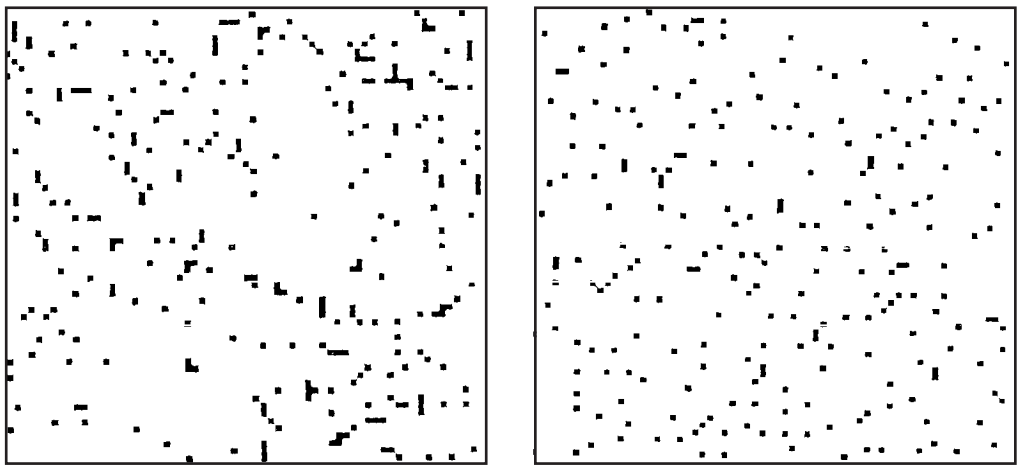

Figure 6: *Reduced and binary representation of nuclear placement for Fig. 1.*

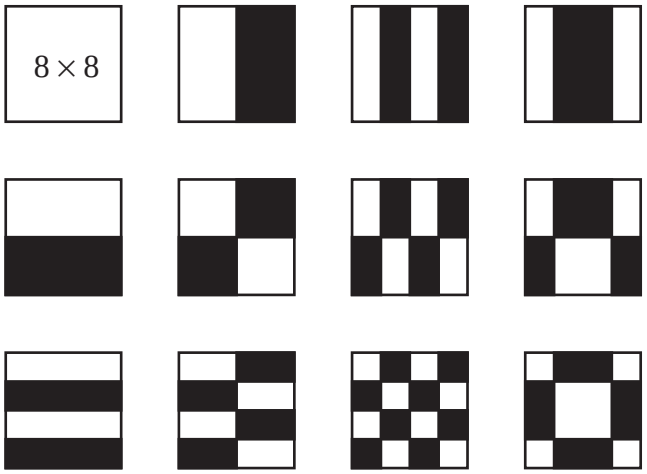

Figure 7: *The first 3 Walsh basis filters in vertical orientation, the first 4 in horizontal orientation and the corresponding mixture filters are shown. Each filter consists of an 8×8 pixel array. White and black denote +1 and −1, respectively.*

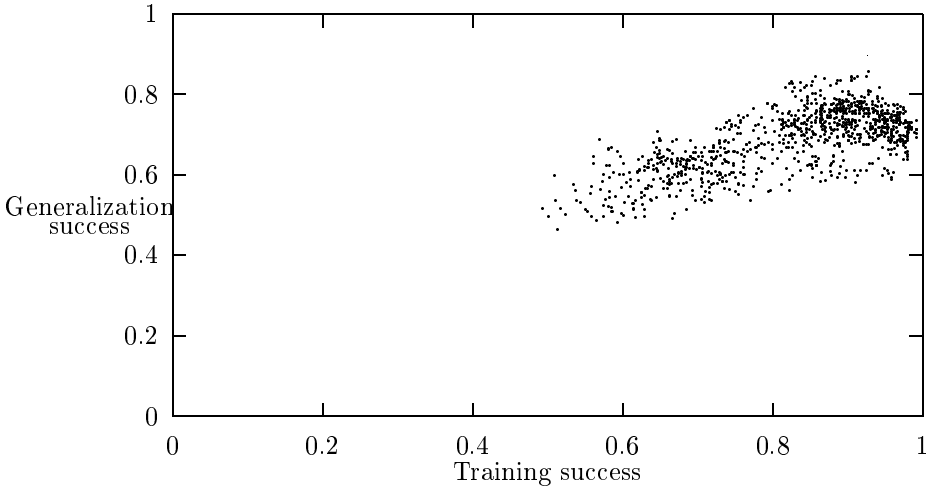

Figure 8: *Classification success of 1000 quadratic Bayes classifiers using arbitrary feature combinations.*

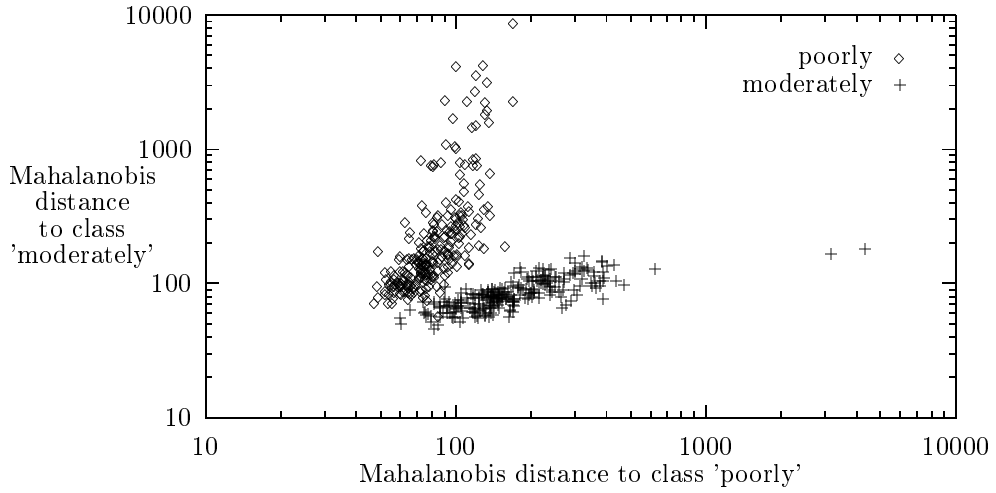

Figure 9: *Discriminating abilities of a given set of features found in Figure 7. The mapping is described in the text.*

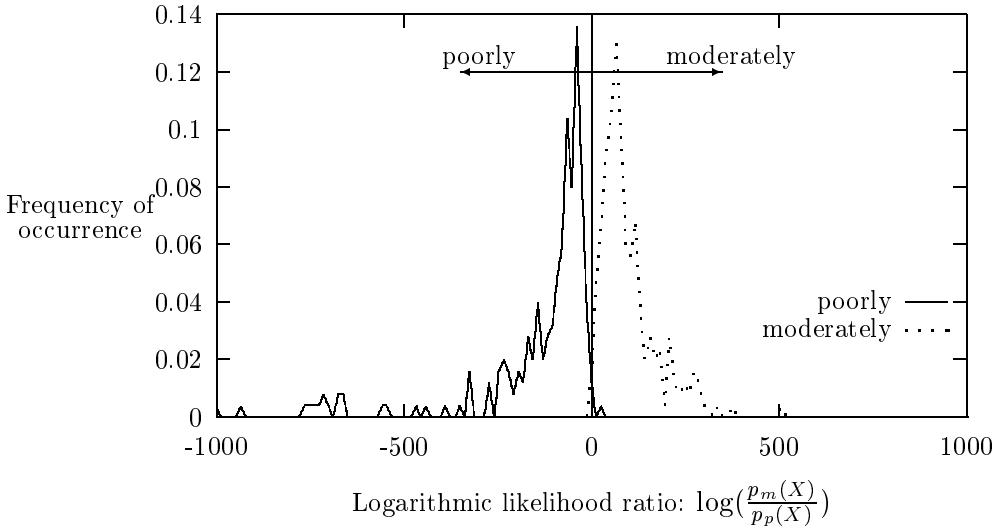

Figure 10: *Frequency distribution of the logarithmic likelihood ratio of poorly differentiated (full line) and moderately differentiated tissue samples (dotted line). The classifier shown in Fig. 6 categorizes a sample as poorly if the log. likelihood ratio is smaller as 0, and as moderately else. The certainty of a decision grows with the absolute value of the log. likelihood ratio of a sample. The overlap of both curves in the middle of the graph represent samples of the training set which cannot be classified correctly.*

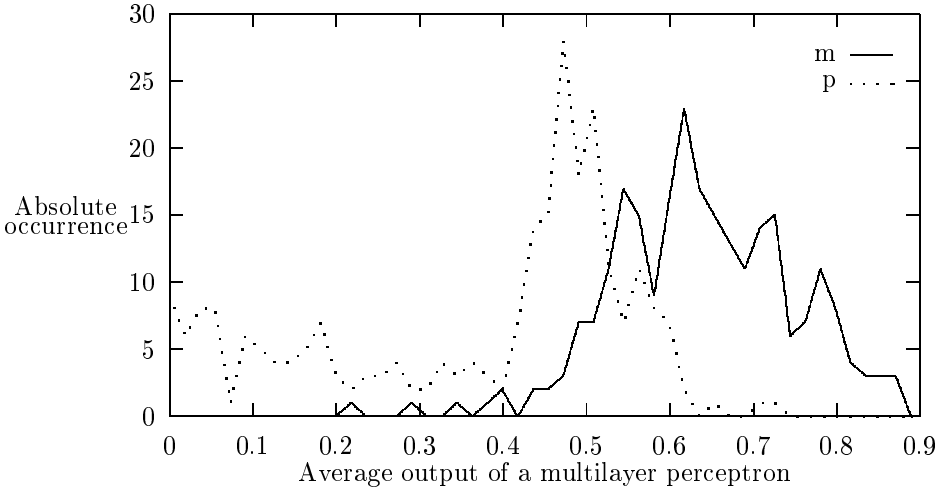

Figure 11: *Classwise output of a multilayer perceptron on the training set.*

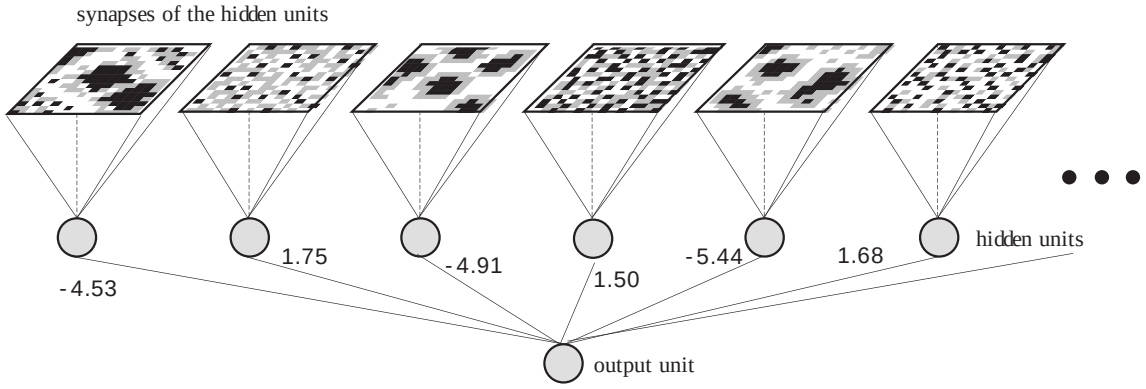

Figure 12: *Feature extraction using a multilayer perceptron. The stucture of the synaptic fields is shown schematically. A white pixel represents a high synaptic value, a black one a low (negative) value. The pattern shown in the synaptic fields force the corresponding hidden to maximum activation. Only 6 synaptic fields of a total set of 35 hidden units are shown.*

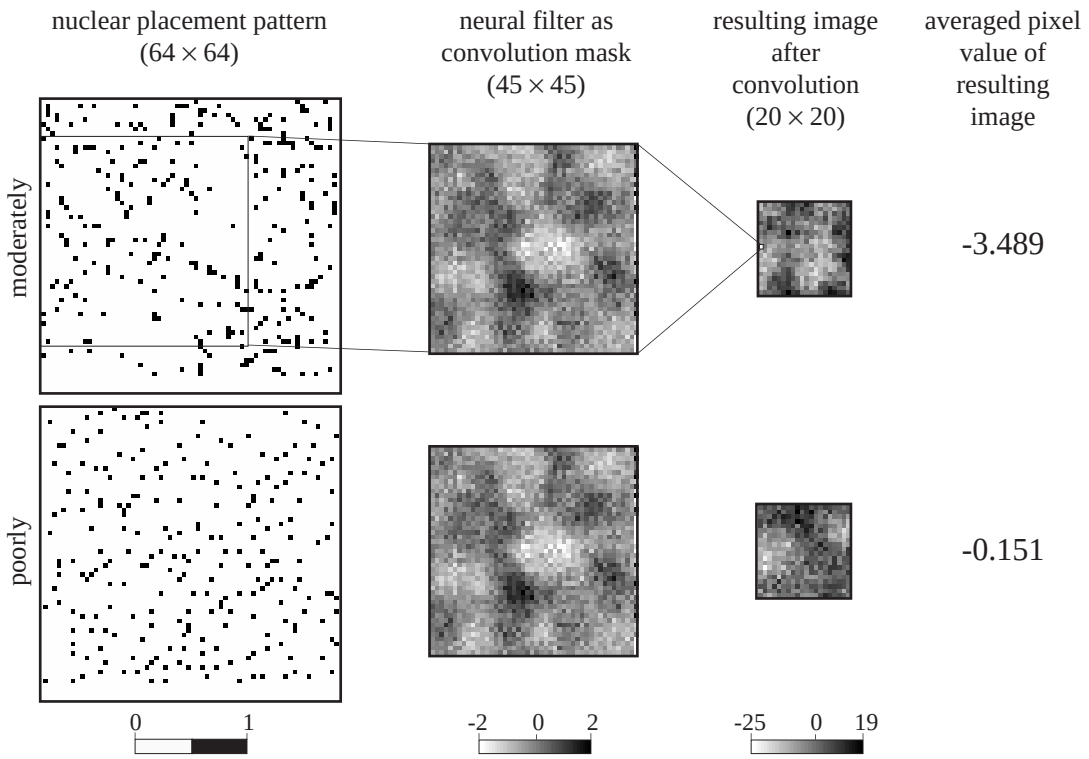

Figure 13: *The synaptic fields of a single hidden unit of a multilayer perceptron are interpreted as convolution masks. Applied to a binary nuclear placement pattern a convolved image is produced. Then it is averaged and used as a feature for the discriminant analysis. The upper half of the figure shows this procedure for a sample of class "moderatly differentiated". The highest absolute response in the resulting image (a single pixel) and the corresponding position in the original image are marked. The lower half of the figure shows the highest absolute response and the corresponding position for a sample of class "poorly differentiated".*

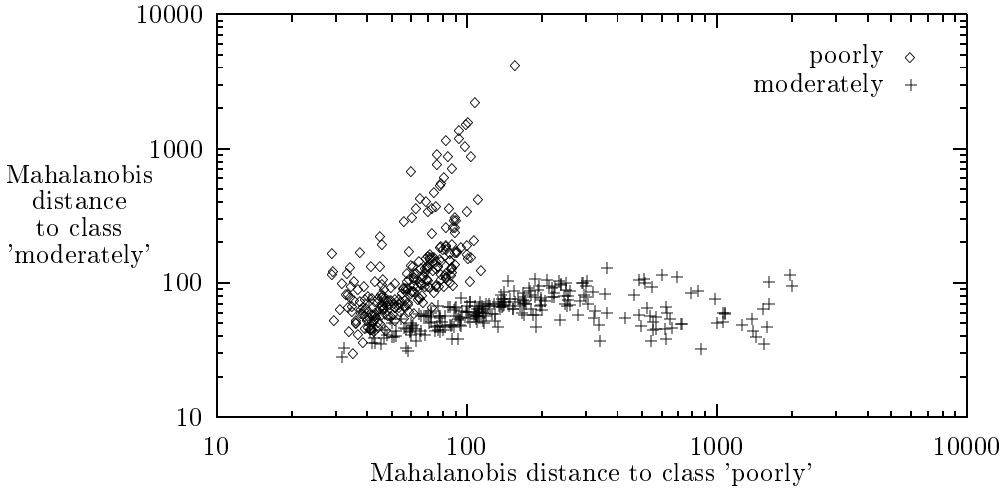

Figure 14: *Discriminating abilities of a quadratic Bayes classifier that uses "neural" features developed by multilayer perceptrons (compare Fig. 6 and explanations there in the text).*

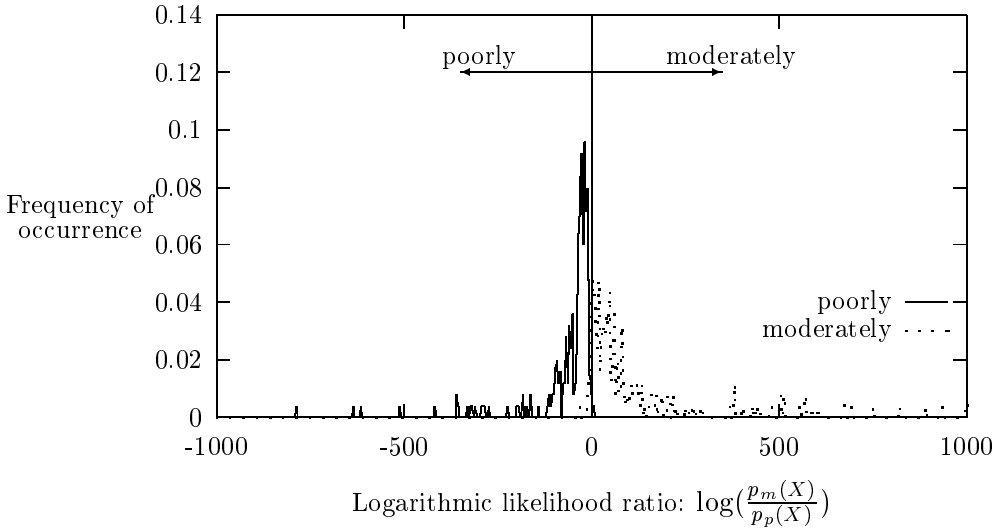

Figure 15: *The distribution of the logarithmic likelihood ratio of the tissue samples classified by the hybrid classifier of Figure 10.*