Analytical Cellular Pathology 16 (1998) 63–82
IOS Press

A review of caveats in statistical nuclear image analysis

Helene Schulerud[a,c], Gunner B. Kristensen[b], Knut Liestøl[a], Liljana Vlatkovic[c], Albrecht Reith[c], Fritz Albregtsen[a] and Hàvard E. Danielsen[c,*]
[a]*Image Processing Laboratory, Department of Informatics, University of Oslo, Oslo, Norway*
[b]*Department of Gynecological Oncology and* [c]*Department of Pathology, The Norwegian Radium Hospital, Ullernchausseen 70, Oslo, N-0310 Norway*

Received 12 May 1997
Revised 13 November 1997
Accepted 2 January 1998

Abstract. A large body of the published literature in nuclear image analysis do not evaluate their findings on an independent data set. Hence, if several features are evaluated on a limited data set over-optimistic results are easily achieved. In order to find features that separate different outcome classes of interest, statistical evaluation of the nuclear features must be performed. Furthermore, to classify an unknown sample using image analysis, a classification rule must be designed and evaluated. Unfortunately, statistical evaluation methods used in the literature of nuclear image analysis are often inappropriate. The present article discusses some of the difficulties in statistical evaluation of nuclear image analysis, and a study of cervical cancer is presented in order to illustrate the problems. In conclusion, some of the most severe errors in nuclear image analysis occur in analysis of a large feature set, including few patients, without confirming the results on an independent data set. To select features, Bonferroni correction for multiple test is recommended, together with a standard feature set selection method. Furthermore, we consider that the minimum requirement of performing statistical evaluation in nuclear image analysis is confirmation of the results on an independent data set. We suggest that a consensus of how to perform evaluation of diagnostic and prognostic features is necessary, in order to develop reliable tools for clinical use, based on nuclear image analysis.

Keywords: Nuclear image analysis, statistical evaluation, precancerous cells, cancerous cells, testing features, sampling methods, classification methods

1. Introduction

Automated image analysis in cytology is a field in rapid development. Over the past 10 years there has been an increasing number of papers dealing with diagnostic, prognostic and therapeutic issues, using nuclear image analysis. There are several reasons for this development for which the concern of quality control in cytopathology is the major one.

In the field of *diagnosis*, there has been an interest to improve the early detection of tumours by building devices for automatic image analysis of cells. Early detection is still the driving force by which one hopes to achieve cancer prevention, either through removal of a precancerous lesion or through an early treatment of a localized cancer. This type of prevention is particularly true for cervical cancer, where several studies have demonstrated significant reduction in cancer incidence as well as mortality after starting screening programs [2,23]. Various diagnostic approaches are in use,

of which histological evaluation and morphologic grading are the most important. However, these methods are based on subjective evaluations which often result in considerable inter- and intra-observer variation [52,62,63,107]. Therefore, image analysis methods have been developed in an attempt to obtain more objective diagnosis. Diagnostic systems using image analysis have been applied to many different organs, including the uterine cervix [41,49,80,90,102,131,133,137], the ovaries [30,66], the breasts [9,26,115,119], the liver [37,53,67], the thyroid gland [18,31,68,76,110], the lungs [71], the kidneys [124], the brain [27], the pancreas [105], the colon [89], the oral cavity [114], the skin [38,74] and lymphoid tissue [91]. Nuclear image analysis methods are still in an investigative phase, and although promising achievements have been made, their importance for diagnosis in clinical use has still to be shown.

Prognosis is another important goal for the use of image analysis in cytology. In pathology, it is well known that the biological behavior of tumours varies even when the tumours appear similar with respect to histology and cytology. In spite of such similarities some tumours prove more aggressive than others [138]. Patients with non-aggressive tumours require little or no clinical intervention besides surgical removal, while patients with aggressive tumours require additional treatment. For prognostic evaluation, clinicians and pathologists rely on clinical staging, histological evaluation, morphologic grading and the more recently developed DNA ploidy assessment techniques. However, the prognostic information provided by these methods is insufficient. The lack of reliable methods which completely discriminate aggressive and non-aggressive tumours may lead to patients undergoing either over- or undertreatment. Therefore, discrimination between patients with aggressive and non-aggressive tumours is of great importance and would make an individually tailored treatment possible. For this reason, efforts have been made in the recent years to develop cell image analysis systems as objective prognostic tools. Image analysis has been used in grading and prognostic evaluation for cancers originating from the uterine cervix [42,83,93,108,134], the uterine corpus [40], the ovaries [34,79,85], the breasts [3,4,25,36,61,64,69,71,95,122], the prostate gland [16,51,57,86,88,97,118,127, 129,136], the bladder [78,84,99,117,126,135], the kidneys [48,100], the thyroid gland [75], the lungs [50,70] and the rectum [87,128].

Besides the applications of image analysis for diagnostic and prognostic purposes, monitoring and evaluating the effect of radiotherapy and chemotherapy [19,24,29,96,109,116,121,130] have also been attempted.

Since nuclear image analysis methods are still in an investigative phase, various data acquisition, feature selection methods and evaluation methods need to be analyzed. Unfortunately, inappropriate statistical evaluation methods are often used [56]. This combination of various approaches and partly inappropriate statistical methods easily lead to confusing results. Thus, there is a need for standardization in the use of statistical evaluation methods. The aim of the present work is to discuss some of the caveats in statistical methods applied in nuclear image analysis. To illustrate some of these problems, results from a study of prognostic factors in cervical cancer are presented. Most of the aspects discussed in this paper are equally relevant for image analysis of histological data.

2. Nuclear features and feature selection

2.1. Standard features

A variety of different features can be measured in order to describe nuclei. The most commonly used types of features in nuclear image cytometry are morphometric, densitometric, and textural features.

- *Morphometric features* describe the shape and size of nuclei and their internal structures, such as form, size and number of nucleoli and chromatin granules.
- *Densitometric features* are based on the gray level density distribution of nuclei or nucleoli. If nuclei are stained by a stoichiometric DNA stain (e.g., Feulgen), the optical density corresponds to the relative amount of nuclear DNA.
- *Textural features* reflect the internal structure and measure features like granularity and regularity of the chromatin structure, i.e., they measure aspects of the organization of the DNA.

Textural features can be assessed from various matrices which describe the chromatin gray level distribution. Groups of texture features can be assessed from, e.g., the gray level co-occurrence matrix (GLCM) (also called Markovian features) [20,43,112,125], the gray level run-length matrix (GLRLM) [17,39] or the gray level entropy matrix (GLEM) [136]. The matrices can be measured with different parameters according to the number of gray levels and size of the texture element. An additional group of texture features are the fractal features, which measure the complexity of the chromatin structure [60,81]. Consequently, a very large number of features are available to describe the nuclei.

2.2. Single feature evaluation

In cell image analysis, we often want to test whether a feature is significantly different for the distinct outcome classes under consideration, e.g., patients with and without relapse. Moreover, test statistics can be used to rank and compare various features. Testing can be done by either parametric or non-parametric tests, the first depending on distributional assumptions (e.g., assuming that the feature has a Gaussian distribution) while the second does not. Both non-parametric and parametric tests assume independent samples. This assumption may be violated in nuclear image analysis if cell measures are used as independent samples. Cells taken from the same patient are in most circumstances not independent, and falsely assumed independence may lead to strongly misleading results (see Sections 5 and 6).

As mentioned earlier, a large variety of features are available to describe the characteristics of a nucleus. When several features are tested, wrong conclusions may be drawn from incorrect use of statistical tables designed for single rather than multiple comparisons [56]. The Bonferroni inequality may be used to correct for multiple tests. Suppose that we have a given number of features (m), and test for each feature the hypothesis stating that the feature is equally distributed in the outcome classes: m tests are then performed. Different types of tests may be used, e.g., comparison of means or variances of two independent groups of observations. For each test a p-value is found. The Bonferroni inequality ensures that if each hypothesis is rejected when the p-value is less than the given significance level (α) divided by the number of features ($p_i \leqslant \alpha/m$), then the probability of rejecting at least one hypothesis when all are true is no greater than α. That is, the probability of falsely concluding that one or more features differ between the outcome classes, when all the features are equally distributed in the outcome classes, is less than the significance level α.

A sharper Bonferroni bound for multiple comparisons has been proposed by Hochberg [46]. In this sharper test, the p-values are ranked from highest to lowest, and each value is multiplied by its rank in the list. If the adjusted p-value is less than the selected significance level α, the null hypothesis (that both groups are equal) is rejected. All remaining null hypotheses with smaller or equal p-values are also rejected. This sharper Bonferroni bound results in easier rejection of the null hypothesis and consequently more features are found significantly different in the outcome classes than for the standard Bonferroni bound.

2.3. Feature set selection

The number of samples needed to design a classifier should increase with the number of features included in the classification rule [32,54]. This limitation is related to what Bellman [10] called "the curse of dimensionality". This is also related to a phenomenon called the "peaking phenomenon" which states that for a finite training sample size, the correct classification rate may initially increase with the addition of new features, attain a maximum and then begin to decrease [104]. Therefore, only a limited number of features should be included in a classification rule when the number of samples is limited. The goal of feature selection is to find the subset of features which best characterizes the differences between the patient groups analyzed.

The number of features used in the classification rule is an important issue in the design of a pattern recognition system, and the problem has been discussed in many papers [54,55,59,98,104]. Raudys and Jain [104] give recommendations for practical use of statistical pattern recognition. The recommended number of features depends on the number of samples available and the classifier used. Given certain assumptions, they recommend that for a linear classifier the number of independent samples in each class should be at least 5 times the number of features. The linear classifier is a simple classifier where the boundaries between the outcome classes in the feature space are linear. For more complex classifiers, a higher ratio between features and samples is necessary.

There are several methods for selecting a feature set. For a given quality criterion, *exhaustive search* (testing all possible combinations) finds the optimal feature set. However, the number of possible sets grows exponentially with the number of features, and thus makes the method impractical even for a moderate number of features. A suboptimal approach is the selection of *the best single features* according to some quality criterion. However, the individually best features are often correlated and thus may give a clearly suboptimal discrimination [21,32,59]. *Sequential forward* or *backward* and *stepwise forward–backward selection* are other suboptimal approaches [32]. Sequential forward selection [132] adds one new feature to the current set of selected features in each step. Sequential backward selection [82] starts with all the possible features and discards one at the time. The main drawback with forward and backward selection is that when a feature is selected or removed this decision cannot be changed. This is called the nesting problem. Stepwise forward–backward selection [120] combines a forward and backward selection strategy, and thus overcomes the nesting problem. Stepwise forward–backward selection is a special case of "plus l-take away r", where a given number of features (l) are included and another given number of features (r) are excluded in each step. A more recently developed method is the *floating search* method, which is a generalization of the "plus l-take away r" method, where the number of features added and removed changes in each step [103,139].

Another group of feature extraction methods performs a transformation in the feature space and designs new features as (linear) combinations of the old ones, e.g., *principal component analysis* (PCA) [7]. The main drawbacks with transformation methods as feature selection methods are that the new features may be difficult to interpret and that all features used in the transformation must be measured for each new sample to be classified.

3. Classification methods

The aim of a classification system is to classify an unknown sample into one of the predefined outcome classes. Further, if the unknown sample differs from all the defined outcome classes, it may be classified as unknown (outlier), or if it is intermediate between classes, it may be assigned

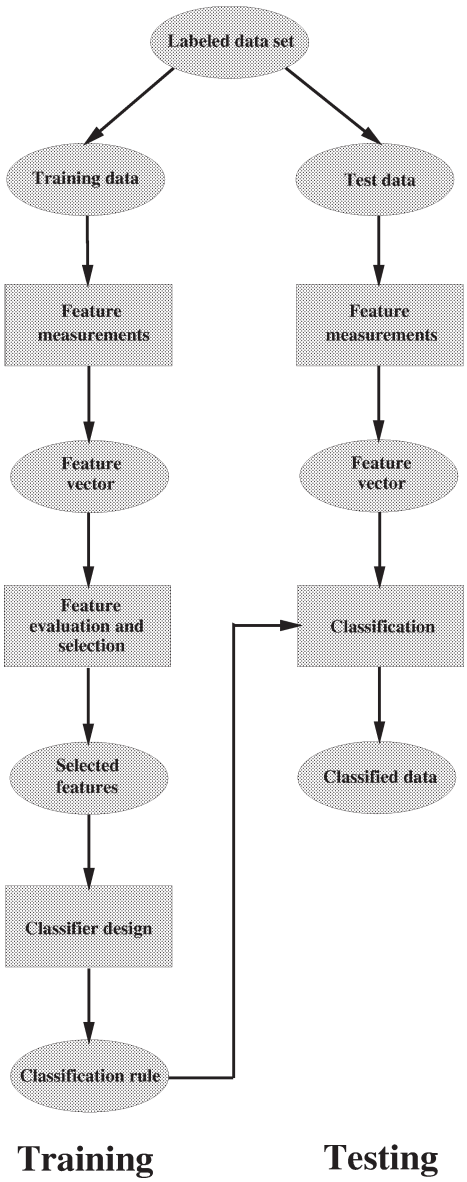


Fig. 1. Classification scheme with separate training and test data.

to a "doubt" class. For example, based on prognosis a tumour may be classified as aggressive, non-aggressive or uncertain.

In order to design a classifier, a data set including samples with known outcome classes has to be available. Construction of a classifier normally consists of two phases. In the first phase (the training phase), data with known outcome classes are used to design the classifier. In the second phase (the testing phase), the performance of the classification rule is estimated, see Fig. 1.

There are different types of classification systems based on a statistical or a structural framework [33]. The present work is focused on the statistical classification methods denoted Bayesian classification and regression based classification.

3.1. Bayesian classification

In Bayesian classification, the feature values of each outcome class are assumed to arise from underlying probability density functions. These functions, often assumed to be Gaussian distributions, must be estimated from training samples, where the true outcome is known. Furthermore, it is assumed that the *a priori* probability of new samples arising from each outcome class is known (or can be estimated). For given feature values of a new sample, the well-known Bayes formula may be used to determine the *a posteriori* probabilities that the new sample belongs to each outcome class. The Bayes formula may be given as

$$P(A_i|S) = \frac{p(S|A_i)P(A_i)}{\sum_j p(S|A_j)P(A_j)},$$

where S is the feature values of a new sample, A_i is one of the predefined classes, $P(A_i)$ is the *a priori* probability for class A_i, $p(S|A_i)$ is the probability density for S given that the sample belongs to class A_i, and $P(A_i|S)$ is the *a posteriori* probability that the new sample (S) belongs to class A_i. If we have two outcome classes A_1 and A_2, the *a posteriori* probability for each class ($P(A_1|S)$, $P(A_2|S) = 1 - P(A_1|S)$) is assessed. In Bayesian minimum error rate classification, the new sample is classified to the class with highest *a posteriori* probability [35]. If we assume Gaussian distributed density functions with common covariance matrix, the Bayesian classification rule becomes a linear discriminant function, that is, the boundaries between the outcome classes in the feature space are linear [5]. This latter classification rule is also called minimum distance classification [35].

3.2. Regression

Regression analysis is the analysis of the relationship between one variable (dependent or response variable) and another set of variables (independent or predictor variables). Regression analysis may also be used in feature selection, but we focus here on its use in classification.

While Bayesian classification involves separate assessments of the *a priori* probabilities for the outcome classes, this is not the case for regression methods. For that reason, regression methods depend on training sets that are (reasonably) representative with respect to the fraction of cases belonging to each outcome class.

Two regression methods, *logistic regression* [47] and *Cox regression* [22], are of specific interest. In its original form, logistic regression was designed for binary outcome variables (e.g., relapse/no-relapse), but extensions to several ordered or unordered categories are available [47]. For simplicity, only the binary case, e.g., relapse or no-relapse, will be considered. The regression procedure then estimates the probability of relapse as a function of the values of the feature variables. These feature variables are supposed to have a linear effect on a function called the *logit*:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k,$$

where p is the probability of relapse, x_i are the feature variables, b_i are the regression coefficients and k is the number of variables. A classification is obtained by setting thresholds for p, e.g., by defining two thresholds one may separate the patients into the three groups: probable relapse, uncertain and relapse unlikely.

The linear effect of the predictors may be replaced by more complex functions. In so-called additive models the terms $b_i x_i$ are replaced by general (but smooth) functions $f(x_i)$ [44], while neural nets allow complicated interactions between the variables [106]. The disadvantage of complex models is the requirement of a large sample size for obtaining reliable estimates of the model parameters.

Both logistic regression and the Bayes scheme described above require the patients in the training and test sets to be divided into well-defined groups, e.g., relapse, no-relapse. The time aspect of disease progression is thus neglected. If the time of relapse is thought to carry significant information, one can use a regression model adapted to survival data. The most well-known is the Cox model [22]. This model makes no assumption about the distribution of survival times. However, the Cox model also assumes an essentially linear combination of predictors. As for the logistic case, this assumption may be relaxed [77].

A method which is related to the regression methods is the *canonical discriminant analysis* [41,111]. It is a dimension-reduction technique which also is related to principal component analysis and canonical analysis [73]. The derived canonical variable is of the form

$$C = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n,$$

where x_i is feature number i, b_i is the corresponding canonical weight and n is the number of features. The canonical variable summarizes between-class variation in the same way as principal components summarizes total variation.

4. Estimation of the classification performance

After a classification rule is designed, estimation of the performance of the classification rule must be performed. This is a very important part of the classification scheme, since it shows how well the system will work on unknown samples in the future. However, in most practical situations, there is a limited set of known samples available, and the problem is how to use the given samples as training and test sets. More data in the training set will give a better design of the classifier, while more data in the test set will give a better estimate of the actual error rate.

4.1. Common training and test set

If all the data are used in the training phase, the best possible estimates of the parameters in the classifier are achieved. However, if the performance of the classifier is estimated by simply applying the rule on the training samples (resubstitution), the estimated error rate tends to be too optimistic [9,32], since the parameters are optimized on this particular data set. An alternative is to perform cross-validation. Here $N - n$ out of N samples are included in the design of the classifier while the remaining n samples are used to estimate the error rate. This procedure is repeated in a cyclic way so that all the samples are used once in the estimation of the error rate. Leave-one-out (also called jack-knife) is a special case of cross-validation with n equal to 1. If the focus is on one specific classification rule, cross-validation avoids the problems confronted by using resubstitution. Thus, to estimate the performance of the selected features in the training phase, cross-validation is recommended. However, cross-validation does not reduce the problem of multiple testing. If several different rules are tested (formally or informally) and the best result is reported, highly over-optimistic results will be obtained also when using cross-validation. Therefore, if cross-validation is to be used

in performance estimation of the classifier, without any tests on an independent data set, a strictly defined classification procedure is necessary and only one (or at most a few) feature combination should be tested. One should also note that screening for "good" features means testing several rules. Thus, using a common training and testing data set will normally not be feasible due to the high number of potentially useful features. Such a procedure should only be accepted as a part of a pilot project, where the only purpose would be to decide whether or not to carry out a full study.

4.2. Separate training and test sets

When an independent data set is available for testing, no strict procedure is necessary in the training phase, and several feature combinations can be evaluated. The strength of the separate training and testing sets approach is that if the selection procedure in the training phase results in over-fitting (a feature set or parameter values describing the samples in the training set rather than general properties of the classes), this will be clearly demonstrated in the test set.

If a classifier is designed and tested with an acceptable result on an independent test set, more accurate parameter estimates of the classifier may be obtained by including both the training and the test set in the parameter estimation for classification of new unknown samples.

The problem of how to split the data into training and test sets has been discussed in detail in the pattern recognition literature (for review, see [104]), but no clear answer exists. The problem depends on the data and the specific classification task [56]. However, these difficulties should not be used as an excuse for not testing. Without testing, the probability will be high that the results confuse rather than inform. A number of studies have been published on nuclear image analysis in diagnosis and prognosis of cancer without a well-defined test set [1,27,28,36–38,40,41,50,51,65,69,75,83,89,95,99, 112,115,122,127,134,137].

5. Specimen classification

In classification of a specimen, a number of factors must be taken into account. Firstly, a tumour specimen may contain normal cells of either stromal or epithelial origin. Secondly, within the tumour there may be a considerable heterogeneity [100,101,123], and small subpopulations of cells may determine the outcome for the patient. Some studies indicate that cells on the invasive border of the tumour may behave different biologically from cells in the more central part of the tumour, and that the former are more indicative for prognosis [13,14,92].

Discrimination of the tumour tissue into cancer cells and normal cells of either stromal or epithelial origin can be performed by a pathologist. However, how to perform sampling within the ltumour and to select cells in order to measure the relevant diagnostic or prognostic information is an unsolved problem. There are at least two different approaches to this dilemma. One approach assumes that most of the cells have some common features which differ in the distinct outcome classes, and distributions of the feature value between the outcome classes will be systematically shifted. Cells may then be randomly selected from the specimen, and classification of a patient can be performed by measuring the mean, median or standard deviation, etc., of the cell features of each patient. The number of cells taken from each patient must be sufficient to ensure good estimates of the distribution. An alternative is to classify each cell and let the classification results be utilized in a voting scheme for each patient. The patient may be classified into the same class as the majority of the cells.

If we do not assume that most of the cells differ in the outcome classes, another approach is to assume that a subpopulation of cells in each tumour is of specific diagnostic or prognostic value. A patient may then be classified into one of the classes, if a special type of cell is present. The last method resembles more how the pathologists work, but assumes that the special type of cells is sufficiently well defined. An alternative is to select cells randomly from the specimen and to create a histogram of the feature for all the cells from each patient. Characteristics of this distribution may then be used to classify the patients [11]. This approach is similar to DNA ploidy analysis, which discriminates subpopulations of cells with different ploidy by analyzing the DNA distribution of each patient from randomly selected cells. The number of cells taken from each patient must be sufficiently large so that "rare event" cells are represented in sufficient numbers.

A central question concerning classification of a specimen remains: What is the limiting factor considering the sample size and the number of features in the classification rule, the number of cells or the number of patients? There is unfortunately no general answer to this question, but clearly both numbers are important. However, if we assume a training set with 10 patients in each of two outcome classes and 2000 cells per patient, there is a total of 20,000 samples in each class. We would then have many samples per class, but they are certainly not independent for any feature selected. The number of patients, which are independent, is clearly too low. From a statistical point of view, methods that assume independent samples should generally assess sample size based on patients rather than cells, provided that the cells contain sufficient information to represent the patient (tumour or other entity).

A method which can be applied to nuclear image analysis, using both cell and patient information, is *nested variance analysis*. In nuclear image analysis, each class is represented by some patients and from each patient some cells are selected. This procedure is highly hierarchical and there is dependency between the different levels. Nested variance analysis [8,12] is designed to analyze such hierarchical structures. The total variance of the data can be separated into parts assignable to cell, patient and class level. If some specimens are taken from each tumour, the tumour heterogeneity may also be analyzed. Further, independent tests can be performed at each level so that significance tests for differences between the outcome classes and between the patients can be performed separately [8].

6. Early cervical cancer as an illustrative study of caveats

The material consists of 113 patients with early cervical cancer (FIGO stage IB), operated with radical hysterectomy with pelvic lymphadenectomy during 1987–1990 at the Norwegian Radium Hospital. Only tumours of squamous cell type were included in this study. A total of 22 patients had a relapse within a period of 5 years, while 91 patients showed no sign of relapse within the follow-up time. The patients with no relapse had a median follow-up time of 75 months, ranging from 60 to 107. For this study, one 50 μm section was cut from a paraffin block of each tumour. Before this section was cut, a 5 μm section was cut and stained with hematoxylin and eosin for verification of the tumour content by a senior pathologist. Monolayers of nuclei were prepared from the thick sections and Feulgen-stained as previously described [136] and according to well-established techniques [45]. The cells were studied under light microscopy and digitized at a final resolution of 0.254 μm per pixel.

A minimum of 100 epithelial tumour cells were randomly selected from each monolayer by a pathologist and the nuclei were segmented from the background using a global threshold. The variation in the illumination of the cell nuclei in the microscope was corrected by subtracting the background image from the cell nucleus image.

In this study morphometric nuclear features, morphometric chromatin features and texture features from the gray level co-occurrence matrix (GLCM), the gray level run-length matrix (GLRLM), the gray level entropy matrix (GLEM) as well as from the complexity curve [15,58] were analyzed. The morphometric and textural features are described in detail elsewhere [113]. The number of features analyzed was 40, but the texture features were measured with various parameter values. We measured the gray level co-occurrence matrix for different distances (1, 3, 5, 7) and with different number of gray levels (8, 16, 32, 64). For the gray level run length matrix we tested different number of gray levels (4, 8, 16, 32, 64) and for the gray level entropy matrix we tested different window sizes (3, 5, 7) for 8 and 64 gray levels. Thus the resulting number of initial features was 297.

Separate training and test sets were used in this study. A discriminant function was designed by using approximately half the data (randomly selected), and the classifier was tested on the remaining independent data. The training set consisted of 10 patients with relapse and 49 patients with no relapse, and the test set consisted of 12 patients with relapse and 42 patients with no relapse.

In Section 6.1, results are presented which were obtained utilizing the complete initial feature set in the feature selection procedure. In Section 6.2 the results obtained from analyzing less features (28) are described.

6.1. Selecting and testing features from a large feature set

6.1.1. Stepwise feature selection and classification

The first feature selection was performed by a stepwise forward–backward selection (plus 1-take away 1) (SAS procedure *stepdisc* [111]) with the probability to stay and enter $p = 0.15$. This procedure selects features using a quality criterion called Wilk's lambda. Wilk's lambda uses the ratio of the determinant of the within-class covariance matrix and the determinant of the total covariance matrix [6] to assess the significance of separation between the classes. The mean feature value of each lesion was applied in the feature selection and the classification procedure. The stepwise forward–backward feature selection method selected 17 out of 297 possible features as candidates for use in the classification. A Bayesian classification rule (SAS procedure *discrim* [111]) was applied assuming multinormally distributed feature values and equal *a priori* probability for each outcome class. Because of few samples in one of the classes, a common covariance matrix was used. The distribution of the selected features was controlled and the Gaussian assumption considered acceptable. To illustrate the classification results, Figs 2–4 show the canonical variable (see Section 3.2) of the feature combinations.

Classification of the samples in the training data applying the 10 best features selected by the feature selection procedure resulted in an average correct classification rate (CCR) of 95% using resubstitution and 87% by cross-validation. Testing this classification rule on the independent test set gave an average CCR of only 60%, and only 3 out of 12 patients with relapse were correctly classified. Figure 2 shows that almost the same result was achieved when using less features. The estimated CCR increases with the number of features used in the classification rule on the training data, but testing shows that the rule is useless. Furthermore, the figure shows that using cross-validation also results in over-optimistic results. Even the case with only two features, which correspond to one-fifth of the training samples in the smallest group, shows an over-optimistic result.

One possible explanation for the difference between the classification results of the training and test data could be that the features selected discriminate the specific patients in the training data, and not the general classes of patients, i.e., relapse, no-relapse. Logistic regression was also performed on the same data, and similar results were obtained.

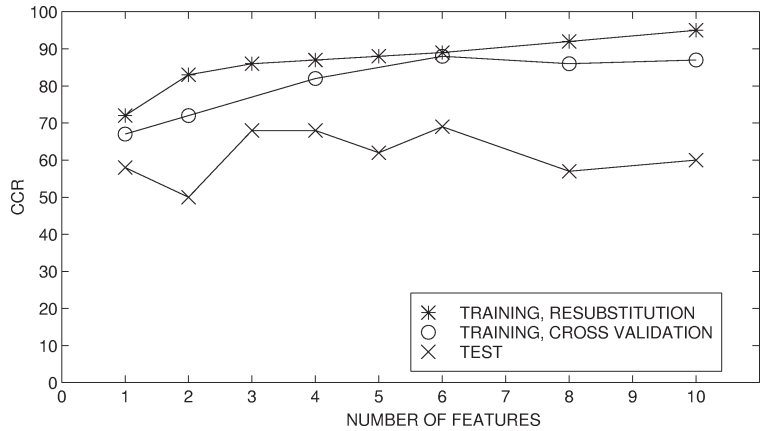


Fig. 2. Estimated classification result as a function of the number of features used in the classification rule selected from the initial feature set of 297 features. The resubstitution estimate from the training data is highly over-optimistic and increases with the number of features used. The cross-validation estimate is also highly over-optimistic and increases with the number of features used, up to a certain point. Testing the classification rules on the independent test data revealed CCRs of about 60%, which are useless results since only few patients with relapse were correctly classified. CCR, correct classification rate.

In many studies published on nuclear image analysis the ratio of the number of samples in the smallest group compared to the number of features in the classification rule is less than 5, without testing the classification rule on an independent data set [1,27,28,36–38,50,51,65,75,80,83,89,95,99, 112,127,134,137]. The danger of over-optimistic results is then very high, as illustrated by the results of this study.

6.1.2. Univariate testing and stepwise feature selection

We first tested if the distribution of each single feature differed between patients with and without relapse from the training data, using the mean feature value for each patient. Utilizing the Wilcoxon test, 45 features showed significant difference ($p < 0.05$). These 45 features were submitted to the stepwise forward–backward feature selection method and four features were selected. Classification with these four features gave an average CCR of 83% by cross-validation on the training data, while testing this classification rule on the independent test set gave an average CCR of only 47%.

We also tested if the distribution of each single feature differed between patients with relapse and patients with no-relapse from the training data, when we corrected for multiple tests. Using Wilcoxon tests with the Bonferroni correction for multiple tests, as proposed by Hochberg [46], none of the features showed significant differences ($p < 0.05$).

Several studies published on nuclear image analysis perform a preselection of features utilizing univariate testing. Since performing univariate testing means testing many hypotheses, separate training and testing is highly necessary. Several studies perform univariate testing followed by classification without testing the performance of the classifier on an independent data set [36,38,51,65,75,90,115, 127,128].

6.1.3. Nested variance analysis

In a nested analysis of variance, the outcome classes (relapse and no-relapse) were defined as the first level variable, with patients as a random variable nested with the classes. The method separates the total variance between cells into parts attributable to differences between cells in each patient, patients within each class and classes.

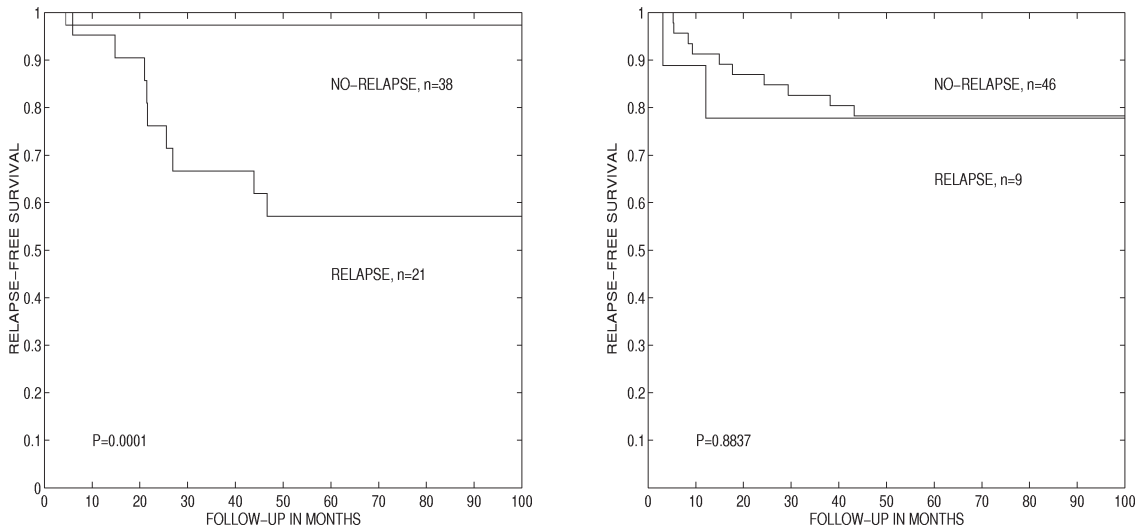
The analysis was carried out on the training set. When no correction was made for multiple hypotheses testing, 49 features showed significant differences between relapse and no-relapse ($p < 0.05$). Testing with the Bonferroni correction, none of the features showed significant differences between the classes. However, highly significant differences between patients, independent of the class to which they belonged, were found for almost all features. Hence, these features showed a significant difference between the individual patients, but no significant difference between the outcome classes (relapse, no-relapse).

6.1.4. Survival analysis

Survival analysis was performed on the two best features selected by stepwise forward–backward feature selection method from the 297 features. Relapse-free survival was analyzed from start of treatment to relapse or May 1996, using univariate and multivariate Cox regression. Survival analysis was performed independently on the training and the test set.

The univariate Cox regression showed that the difference in survival related to the two best features was highly significant ($p = 0.0003$ and $p = 0.013$) for the training data. Multivariate analysis, including also tumour size and vessel infiltration, showed that the most important features for survival were the two nuclear image features ($p = 0.0004$ and $p = 0.013$), with tumour size and vessel infiltration as the third and fourth best features ($p = 0.022$ and $p = 0.008$). Univariate and multivariate analyses were also performed on the test set. The two nuclear features were significant neither in the univariate nor in the multivariate analysis. Relapse-free survival related to the canonical variable of the two best nuclear features for the training and test set is shown in Fig. 3. Hence, the survival analysis shows the same result as found by the discriminant analysis and the logistic regression analysis.

As in discriminant analysis, some studies using survival analysis of nuclear image features analyze many features without testing the result on an independent data set [3,40,100].



(a) Training data (b) Test data

Fig. 3. Kaplan–Meier estimated survival curve of patients classified as “relapse probable” and “relapse unlikely” using a threshold on the canonical variable of the two best features from the discriminant analysis. The resubstitution classification result is presented for the training data. The canonical variable results in two significantly different survival curves in the training data, while there is almost no visual difference between the two survival curves in the test set.

6.2. Selecting and testing features from a smaller feature set

We have also evaluated the effect of first performing a more careful initial screening of the features. Analyzing a texture feature with different parameters, e.g., distance and number of gray levels, often results in highly correlated features. Therefore, we chose one set of parameters for each texture method from the initial feature set based on evaluation of the data. For the GLCM features, we used distance $d = 1$ and quantized the number of gray levels to 16, resulting in 12 features. For the GLRLM features, we quantized to 4 gray levels, resulting in 7 features. For the GLEM features, we quantized to 8 gray levels and used a window size equal to 3, resulting in 9 features. Consequently, a subset of 28 features were used.

We examined the same feature selection methods as in the previous section (Section 6.1). However, using a more carefully selected initial feature set did not avoid over-optimistic results of the training data, as shown in Fig. 4. All the main points found for the larger feature set in the previous section were also found for the smaller feature set.

For illustrative purposes, we also tested if the distribution of each single feature differed when all the cells from each group were included, as if they were independent observations. Eighteen features showed significant differences between relapse and no-relapse at the cell level using significance level $p = 0.0001$. Even if the Bonferroni correction for multiple tests was used, the same 18 features gave significant differences. This obviously erroneous result is probably due to the fact that no correction was made for the dependency between the cells. A method which does compensate for this dependency is the nested variance analysis.

6.3. Summary of the classification results

When many features are evaluated, the risk of over-optimistic results on the training data is high and an independent test set is necessary to evaluate the performance of the classifier. In this study two sets of features were analyzed, both resulting in highly over-optimistic classification results on the training data, even when a limited number of features was used in the actual classifier. Both resubstitution

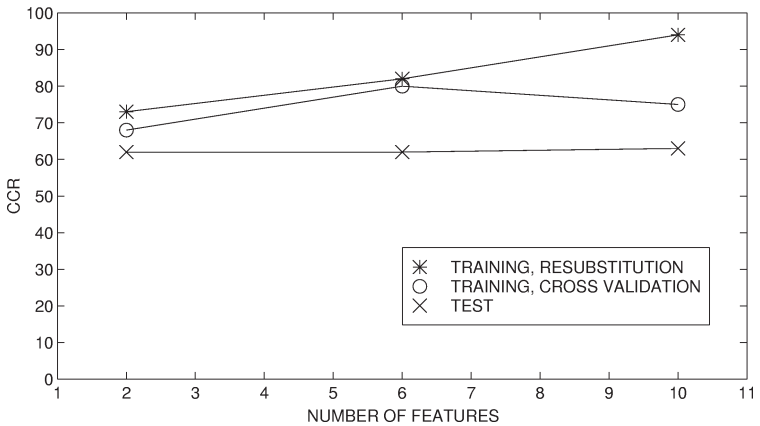


Fig. 4. Estimated classification result as a function of the number of features used in the classification rule selected from the smaller feature set of 28 features. The resubstitution estimate from the training data is highly over-optimistic and increases with the number of features used. The cross-validation estimate is also highly over-optimistic and increases with the number of features used. Testing the classification rules on the independent test data revealed CCRs of about 60%, which are useless results since only few patients with relapse were correctly classified. CCR, correct classification rate.

Table 1
Classification results for the different feature selection and classification methods used in the cervical study

Feature selection method	Classification method	Results	
		Training	Test
SFB	Bayes	+	−
SFB	Log. regr.	+	−
SFB	Cox regr.	+	−
WC + SFB	Bayes	+	−
WC + BON	*	−	−
NV	Bayes	+	−
NV + BON	*	−	−

SFB, stepwise forward–backward selection; WC, Wilcoxon; NV, nested variance analysis; BON, the sharper Bonferroni correction for multiple tests; *, not relevant; +, positive result; −, negative result.

and cross-validation estimates of the classification performance obtained from the training set were over-optimistic. Moreover, an increase in the number of features in the classification rule resulted in higher differences between estimated classification rate from the training and test set.

There were only small differences between the results obtained with different classification methods. Table 1 shows the different feature selection and classification methods used in the cervical study, together with the results from the training and test data. The only feature selection approach which did not find any significant features in the training set was the nested variance analysis with separate testing on the class and patient level, including the sharper Bonferroni correction for multiple tests. Thus, this method seems to give the most realistic result. The important part of the method is the Bonferroni correction of multiple tests. Utilizing a standard univariate test instead of nested variance analysis gave almost the same result, but as discussed earlier (Section 5), the nested analysis is expected to be more efficient. We used the sharper Bonferroni correction for multiple tests (see Section 2.2) in our study. No features were found to be significantly different in the two groups from the training data when we used this method. This result indicates that the sharper Bonferroni is a sufficient method to correct for multiple tests.

7. Conclusions

Computer based image analysis of precancerous and cancerous cells is still in an investigative phase, and different sampling and classification methods need to be analyzed. The use of common evaluation methods is clearly needed, in order to be able to compare results and select the significant findings. Defining common evaluation approaches is not a trivial task, but it should be possible to agree on some basic guidelines.

The classification performance depends completely on the features selected. Hence, the evaluation of features and the process of feature set selection are important parts of statistical image analysis, but is by no means easy. In particular, over-fitting is a serious problem. Over-fitting means that the selected features give a good characterization of the samples in the specific training data, but not the general classes. Over-fitting typically occurs if many features are analyzed, when the sample size is small or if the number of patients in the smallest group (compared to the number of features utilized in the classification rule) is low.

The choice of classification method was of less importance in our study. Similar observations were made by Bengtsson [11] and Palcic [94]. The selected feature combination, found by the stepwise

forward–backward method, discriminated well between the patients with relapse and non-relapse in the training data, but not in the independent test data. Kaplan–Meier survival curves showed the same result, as did the Cox regression analysis.

Performing a more careful feature selection using nested analysis of variance with Bonferroni correction for multiple testing showed that none of the features were significantly different for the outcome classes (relapse, no-relapse). Furthermore, the patient-to-patient variance was high and the features showed significant differences between the individual patients. Therefore, a successful classification of the outcome classes may erroneously be found if a test set is constructed from the same patients as in the training set, but with other cells [113].

Many studies have analyzed a large number of features, used a low sample-feature ratio, and have estimated the performance of the classification rule by resubstitution or cross-validation of the training data. Our example shows that over-optimistic results may easily be obtained if one or more of these approaches are used.

With respect to the methodological problems, interesting parallels may be drawn to the field of clinical drug trials. Twenty years ago, a large fraction of the trials was small, loosely co-ordinated and often methodologically doubtful. Through concerted action of drug agencies (particularly the FDA in the USA), ethical committees and journals (statistical reviewing), *de facto* standards have evolved on how clinical trials should be designed, conducted and reported. These standards have had a remarkable effect on the quality of such studies. Some of these standards are highly relevant also in studies based on nuclear image analysis:

- Acceptable experimental designs should be used. In drug trials, comparative randomized studies are mandatory if practically and ethically possible. In nuclear image analyses, separate training and testing stands out as a similarly important requirement.
- Proper statistical methods should be applied. In nuclear image analyses, correct handling of the relation between cells and patients is one of the central statistical requirements. Great caution must be exercised when applying statistical methods that assume independencies on data where independencies cannot be established.

The issues discussed in this work may hopefully serve as a background for a discussion on defining standard methods for statistical evaluation of nuclear image analysis. Standard methods for nuclear image analysis, taking the statistical problems into consideration, should be worked out in consensus groups. The authorities will not intervene as they did in the case of drug trials, thus journals and their refereeing practice will be crucial if such standards are to be established and followed-up in clinical practice.

Acknowledgement

The authors give special thanks to Ruth Punthervold for excellent technical assistance.

References

[1] F. Albregtsen, H. Schulerud and L. Yang, Texture classification of mouse liver cell nuclei using invariant moments of consistent regions, *Springer Lecture Notes in Comput. Sci.* **970** (1995), 496–502.
[2] G.H. Anderson et al., Organisation and results of the cervical cytology screening programme in British Columbia, 1955–1985, *Br. Med. J.* **296** (1988), 975–978.

[3] M. Aubele et al., Identification of a low-risk group of stage I breast cancer patients by cytometrically assessed DNA and nuclear texture parameters, *J. Pathol.* **177**(4) (1995), 377–384.
[4] J.P. Baak, The relative prognostic significance of nucleolar morphometry in invasive ductal breast cancer, *Histopathology* **9**(4) (1985), 437–444.
[5] P.H. Bartels, Numerical evaluation of cytologic data. IV. Discrimination and classification, *Anal. Quant. Cytol.* **2**(1) (1980), 19–24.
[6] P.H. Bartels, Numerical evaluation of cytologic data. VII. Multivariate significance tests, *Anal. Quant. Cytol.* **3**(1) (1981), 1–8.
[7] P.H. Bartels, Numerical evaluation of cytologic data. VIII. Computation of the principal components, *Anal. Quant. Cytol.* **3**(2) (1981), 83–90.
[8] P.H. Bartels, Numerical evaluation of cytologic data. XI. Nested designs in multivariate analysis of variance, *Anal. Quant. Cytol.* **4**(2) (1982), 81–94.
[9] R.L. Becker, U.V. Mikel and T.J. O'Leary, Morphometric distinction of sclerosing adenosis from tubular carcinoma of the breast, *Pathol. Res. Pract.* **188**(7) (1992), 847–851.
[10] R. Bellman, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, 1961.
[11] E. Bengtsson, The measuring of cell features, *Anal. Quant. Cytol. Histol.* **9**(3) (1987), 212–217.
[12] G.E.P. Box, W.G. Hunter and J.S. Hunter, *Statistics for Experimenters. An Introduction to Design, Data Analysis and Model Building*, 1st edn, Wiley, 1978.
[13] M. Bryne, H.S. Koppang, R. Lillen and A. Kjaerheim, Malignancy grading of the deep invasive margins of oral squamous cell carcinomas has high prognostic value, *J. Pathol.* **166** (1992), 375–381.
[14] M. Bryne, P.S. Thrane and E. Dabelsteen, Loss of expression of blood group antigen H is associated with cellular invasion and spread of oral squamous cell carcinomas, *Cancer* **67** (1991), 613–618.
[15] Y.Q. Chen, M.S. Nixon and D.W. Thomas, Statistical geometrical features for texture classification, *Pattern Recogn.* **28** (1995), 537–552.
[16] R. Christen et al., Chromatin texure features in hematoxilin and eosin-stained prostate tissue, *Anal. Quant. Cytol. Histol.* **15**(6) (1993), 383–388.
[17] A. Chu, C.M. Sehgal and J.F. Greenleaf, Use of gray value distribution of run lengths for texture analysis, *Pattern Recogn. Lett.* **11** (1990), 415–420.
[18] F. Collin, I. Salmon, I. Rahier, J.L. Pasteels, R. Heimann and R. Kiss, Quantitative nuclear cell image analyses of thyroid tumours from archival material, *Hum. Pathol.* **22**(2) (1991), 191–196.
[19] E. Colomb and P.M. Martin, Testing of a chemosensitivity screening method on sensitive and resistant breast tumoural epithelial cell lines, *Anal. Cell. Pathol.* **6**(2) (1994), 105–116.
[20] R.W. Conners and C.A. Harlow, A theoretical comparison of texture algorithms, *IEEE Trans. Pattern Anal. Mach. Intell.* **3** (1980), 204–222.
[21] T.M. Cover, The best two independent measurements are not the two best, *IEEE Trans. Syst. Man Cybern.* **4** (1974), 116–117.
[22] D.R. Cox, Regression models and life tables with discussion, *J. Roy. Statist. Soc.* **45** (1972), 150–166.
[23] D.W. Cramer, The role of cervical cytology in the declining morbidity and mortality of cervical cancer, *Cancer* **34** (1974), 2018–2027.
[24] F. Darro, A. Kruczynski, C. Etievant, J. Martinez, J.L. Pasteels and R. Kiss, Characterization of the differentiation of human colorectal cancer cell lines by means of Voronoi diagrams, *Cytometry* **14**(7) (1993), 783–792.
[25] A.E. Dawson, R.E. Austin and D.S. Weinberg, Nuclear grading of breast carcinoma by image analysis, *Path. Pat.* **95** (1991), S29–S37.
[26] A.E. Dawson, E.S. Cibas, J.W. Bacus and D.S. Weinberg, Chromatin texture measurement by Markovian analysis. Use of nuclear models to define and select texture features, *Anal. Quant. Cytol. Histol.* **15**(4) (1993), 227–235.
[27] C. Decaestecker et al., Identification of high versus lower risk clinical subgroups in a group of adult patients with supratentorial anaplastic astrocytomas, *J. Neuropathol. Exp. Neurol.* **54**(3) (1995), 371–384.
[28] L. Deligdisch and J. Gil, Characterization of ovarian dysplasia by interactive morphometry, *Cancer* **63** (1989), 748–755.
[29] L. Deligdisch, H. Kerner, C.J. Cohen, D. Dargent and J. Gil, Morphometric differentiation between responsive tumour cells and mesothelial hyperplasia in second-look operations for ovarian cancer, *Hum. Pathol.* **24**(2) (1993), 143–147.
[30] L. Deligdisch, C. Miranda, J. Barba and J. Gil, Ovarian dysplasia. Nuclear texture analysis, *Cancer* **72** (1993), 3253–3257.
[31] M. Deverell and J. Salisbury, Nuclear morphometry of primary B cell thyroid lymphoma, *Pathol. Res. Pract.* **188**(4) (1992), 500–503.
[32] P.A. Devijver and J. Kittler, *Pattern Recognition. A Statistical Approach*, 1st edn, Prentice-Hall International, London, 1982.
[33] P.A. Devijver and J. Kittler, *Pattern Recognition Theory and Applications*, Springer, 1986.

[34] C.W. Drescher, A. Flint, M.P. Hopkins and J.A. Roberts, Comparison of the pattern of metastatic spread of squamous cell cancer and adenocarcinoma of the utrine cervix, *Gynecol. Oncol.* **33** (1989), 340–343.
[35] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, 1st edn, Wiley-Interscience, 1973.
[36] J. Dufer, F. Liautaud-Roger, D. Barbarin and P. Coninx, Nucleus image analysis as a possible prognostic tool in grading breast cancer, *Biomed. Pharmacother.* **47**(4) (1993), 131–135.
[37] B.S. Erler, H.M. Truong, S.S. Kim, M.H. Huh, S.A. Geller and A.M. Marchevsky, A study of hepatocellular carcinoma using morphometric and densitometric image analysis, *Am. J. Clin. Pathol.* **100**(2) (1993), 151–157.
[38] M.G. Fleming and R.J. Friedman, Multiparametric image cytometry of nevi and melanomas, *Am. J. Dermatopathol.* **15**(2) (1993), 106–113.
[39] M.M. Galloway, Texture analysis using gray level run length, *Comput. Vis. Graph. Image Process.* **4** (1975), 108–114.
[40] J.P. Geisler, M.D. Michael, C. Wiemann, Z. Zhou, G.A. Miller and H.E. Geisler, Markov texture parameters as prognostic indicators in endometrial cancer, *Gynecol. Oncol.* **62** (1996), 174–180.
[41] T.L. Hall, K.R. Castleman and D.L. Rosenthal, Canonical analysis of cells in normal and abnormal cervical smears, *Anal. Quant. Cytol. Histol.* **10**(3) (1988), 161–165.
[42] A. Hanselaar, G.P. Vooijs, B.H. Mayall, M. Pahlpatz and A. Hof-Grootenboer, DNA changes in progressive cervical intraepithelial neoplasia, *Anal. Cell. Pathol.* **4** (1992), 315–324.
[43] R.M. Haralick, K. Shannmugam and I. Dinstein, Textural features for image classification, *IEEE Trans. Syst. Man Cybern.* **3** (1973), 610–621.
[44] T.J. Hastie and R.J. Tibshirrani, *Generalized Additive Models*, Chapman and Hall, 1990.
[45] D.W. Hedley, M.L. Friedlander and I.W. Taylor, Application of DNA flow-cytometry to paraffin-embedded archival material for the study of aneuploidy and its clinical significance, *Cytometry* **6** (1985), 327–333.
[46] Y. Hochberg, A sharper Bonferroni procedure for multiple tests of significance, *Biometrika* **75**(4) (1988), 800–802.
[47] D.W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, Wiley Series in Probability and Mathematical Statistics, 1989.
[48] K. Van den Houte, R. Kiss, C. de Prez, A. Verhest, J.L. Pasteels and R. Van Velthoven, Use of computerized cell image analysis to characterize cell nucleus populations from normal and neoplastic renal tissues, *Eur. Urol.* **19**(2) (1991), 155–164.
[49] 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, *Anal. Quant. Cytol. Histol.* **14**(4) (1992), 330–334.
[50] N. Ikeda et al., Use of high-resolution cytometry in predicting the biologic behavior of T1 adenocarcinoma of the lung, *Anal. Quant. Cytol. Histol.* **17**(1) (1995), 69–74.
[51] T. Irinopoulou, J.P. Rigaut and M.C. Benson, Toward objective prognostic grading of prostatic carcinoma using image analysis, *Anal. Quant. Cytol. Histol.* **15** (1993), 341–344.
[52] S.M. Ismail et al., Observer variation in histopathological diagnosis and grading of cervical intraepithelial neoplasia, *Br. Med. J.* **298** (1989), 707–710.
[53] R. Jagoe, C. Sowter and G. Slavin, Shape and texture analysis of liver cell nuclei in hepatomas by computer aided microscopy, *J. Clin. Pathol.* **37** (1984), 755–762.
[54] A.K. Jain and B. Chandrasekaran, Dimensionality and sample size considerations in pattern recognition practice, in: *Handbook of Statistics*, Vol. 2, Krishnaiah and Kanal, eds, 1st edn, North-Holland, Amsterdam, 1982.
[55] N.T. James, *Classification Algorithms*, Collins, London, 1985.
[56] N.T. James, Common statistical errors in morphometry, *Pathol. Res. Pract.* **185** (1989), 764–768.
[57] T. Jørgensen, K. Yogesan, F. Skjørten, O. Kaalhus, K.J. Tveter and H.E. Danielsen, Nuclear texture analysis: A new prognostic tool in metastatic prostate cancer, *Cytometry* **24** (1996), 277–283.
[58] S. Kamata, R.O. Eason and E. Kawaguchi, Complexity curves versus histogram and their application to image segmentation, in: *Proc. 7th Scandinavian Conference on Image Analysis*, Denmark, 1991, pp. 1070–1077.
[59] L. Kanal, Patterns in pattern recognition: 1968–1974, *IEEE Trans. Inf. Theory* **20** (1974), 697–722.
[60] J.M. Keller, S. Chen and R.M. Crownover, Texture description and segmentation through fractal geometry, *Comput. Vis. Graph. Image Process.* **45** (1989), 150–166.
[61] E.B. King et al., Characterization by image cytometry of duct epithelial proliferative disease of the breast, *Mod. Pathol.* **4**(3) (1991), 291–296.
[62] P. Klinkhamer, G. Vooijs and A. Haan, Intraobserver and interobserver variability in the diagnosis of epithelial abnormalities in cervical smears, *Acta Cytol.* **32** (1988), 794–800.
[63] P. Klinkhamer, G. Vooijs and A. Haan, Intraobserver and interobserver variability in the quality assessment of cervical smears, *Acta Cytol.* **33** (1989), 215–218.
[64] D. Komitowski, M.M. Hart and C.P. Janson, Chromatin organization and breast cancer prognosis, *Cancer* **72** (1993), 1239–1246.
[65] D. Komitowski and C. Janson, Quantitative features of chromatin structure in the prognosis of breast cancer, *Cancer* **65** (1990), 2725–2730.

[66] D. Komitowski, C. Janson, J. Szamaborski and B. Czernobilsky, Quantitative nuclear morphology in the diagnosis of ovarian tumours of low malignant potential (borderline), *Cancer* **64** (1989), 905–910.
[67] D. Komitowski and G. Zinser, Quantitative description of chromatin structure during neoplasia by the method of image processing, *Anal. Quant. Cytol. Histol.* **7** (1985), 178–182.
[68] A. Kriete et al., Computer analysis of chromatin arrangement and nuclear texture in follicular thyroid tumours, *Histochemistry* **78**(2) (1983), 227–230.
[69] K.D. Kunze, G. Haroske, V. Dimmer, W. Meyer and F. Theissig, Grading and prognosis of invasive ductal mammary carcinoma by nuclear image analysis in tissue sections, *Pathol. Res. Pract.* **185**(5) (1989), 689–693.
[70] M. Ladekarl, T.B. Hansen, R. Henrik-Nielsen, C. Mouritzen, U. Henriques and F.B.S. Rensen, Objective malignancy grading of squamous cell carcinoma of the lung, *Cancer* **76** (1995), 797–802.
[71] D. Larsimont et al., Relationship between computerized morphonuclear image analysis and histopathologic grading of breast cancer, *Anal. Quant. Cytol. Histol.* **11**(6) (1989), 433–439.
[72] D. Larsimont, R. Kiss, Y. de Launoit and M.R. Melamed, Characterization of the morphonuclear features and DNA ploidy of typical and atypical carcinoids and small cell carcinomas of the lung, *Am. J. Clin. Pathol.* **94**(4) (1990), 378–383.
[73] D.N. Lawley, Test of significance in canonical analysis, *Biometrika* **46** (1959), 59–66.
[74] G. Leitinger, L. Cerroni, H.P. Soyer, J. Smolle and H. Kerl, Morphometric diagnosis of melanocytic skin tumours, *Am. J. Dermatopathol.* **12**(5) (1990), 441–445.
[75] F. Liautaud-Roger, J. Dufer, M.J. Delisle and P. Coninx, Thyroid neoplasms. Can we do any better with quantitative cytology?, *Anal. Quant. Cytol. Histol.* **14**(5) (1992), 373–378.
[76] F. Liautaud-Roger, J. Dufer, M. Pluot, M.J. Delisle and P. Coninx, Contribution of quantitative cytology to the cytological diagnosis of thyroid neoplasms, *Anticancer. Res.* **9**(1) (1989), 231–234.
[77] K. Liestøl, P.K. Andersen and U. Andersen, Survival analysis and neural nets, *Stat. Med.* **13** (1994), 1189–1200.
[78] P. Lipponen, M. Eskelinen and J.K. Nordling, Classic prognostic factors, flow cytometric data, nuclear morphometric variables and mitotic indexes as predictors in transitional cell bladder cancer, *Anticancer Res.* **11** (1991), 911–916.
[79] C. Ludescher et al., Prognostic significance of tumour cell morphometry, histopathology, and clinical parameters in advanced ovarian carcinoma, *Int. J. Gynecol. Pathol.* **9**(4) (1990), 343–351.
[80] C. MacAulay and B. Palcic, Fractal texture features based on optical density surface area, *Anal. Quant. Cytol. Histol.* **12** (1990), 394–398.
[81] B.B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, New York, 1983.
[82] T. Marill and D.M. Green, On the effectiveness of receptors in recognition systems, *IEEE Trans. Inf. Theory* **9** (1963), 11–17.
[83] G. Mariuzzi et al., Cytometric evidence that cervical intraepithelial neoplasia I and II are dysplasias rather than true neoplasias, *Anal. Quant. Cytol. Histol.* **14** (1992), 137–147.
[84] U. de Meester, I. Young, J. Lindeman and H.V. der Linden, Towards a quantitative grading of bladder tumours, *Cytometry* **12** (1991), 602–613.
[85] B. Miller, A. Lynn and D. Horbelt, The prognostic value of image analysis in ovarian cancer, *Cancer* **67** (1991), 1318–1321.
[86] C. Minimo et al., Importance of different nuclear morphologic patterns in grading prostatic adenocarcinoma. An expanded model for computer graphic filters, *Anal. Quant. Cytol. Histol.* **16**(5) (1994), 307–314.
[87] B. Mitmaker, S. Kyzer, L. Begin and P. Gordon, The value of nuclear morphometry in the management of patients with colorectal polyps that contain invasive adenocarcinoma, *J. Surg. Oncol.* **51**(1) (1992), 42–46.
[88] J. Mohler et al., Prediction of prognosis in untreated stage A2 prostatic carcinoma, *Cancer* **69**(2) (1992), 511–519.
[89] A. Montag, P. Bartels, H. Dytch, E. Lerma-Puertas, F. Michelassi and M. Bibbo, Karyometric features in nuclei near colonic adenocarcinoma, *Anal. Quant. Cytol. Histol.* **13**(3) (1991), 159–167.
[90] A. Montag, P. Bartels, E. Lerma-Puertas, H. Dytch, S. Leelakusolvong and M. Bibbo, Karyometric marker features in tissue adjacent to *in situ* cervical carcinomas, *Anal. Quant. Cytol. Histol.* **11**(4) (1989), 275–280.
[91] A. Moragas, M. Garcia-Bonafe, I. de Torres and M. Sans, Textural analysis of lymphoid cells in serious effusions. A mathematical morphologic approach, *Anal. Quant. Cytol. Histol.* **15**(3) (1993), 165–170.
[92] E. Odell et al., The prognostic value of individual histologic grading parameters in small lingual squamous cell carcinomas. The importance of the pattern of invasion, *Cancer* **74** (1994), 789–794.
[93] V. Orille, A. Sampedro, J. Ferrer-Barriendos, N. Corral and A. Martinez, Quantitative pathology of the cervical intraepithelial neoplasia, *Eur. J. Gynecol. Oncol.* **14**(6) (1993), 491–500.
[94] B. Palcic, C. MacAulay, S. Shlien, W. Treurinet, H. Tezcan and G. Anderson, Comparison of three different methods for automated classification of cervical cells, *Anal. Cell. Pathol.* **4** (1992), 429–441.
[95] B. Palcic, B. Susnik, D. Garner and I. Olivotto, Quantitative evaluation of malignant potential of early breast cancer using high resolution image cytometry, *J. Cell Biochem.* (Suppl.) **17G** (1993), 107–113.

[96] O. Pauwels and R. Kiss, Monitoring of chemotherapy-induced morphonuclear modifications by means of digital cell-image analysis, *J. Cancer Res. Clin. Oncol.* **119**(9) (1993), 533–540.
[97] M. Petein et al., Morphonuclear relationship between prostatic intraepithelial neoplasia and cancer as assessed by digital cell image analysis, *Am. J. Clin. Pathol.* **96** (1991), 628–634.
[98] J. Piper, Variability and bias in experimentally measured classifier error rates, *Pat. Rec. Lett.* **13** (1992), 685–692.
[99] H. Van der Poel et al., Morphometry, densitometry and pattern analysis of plastic-embedded histologic material from urothelial cell carcinoma of the bladder, *Anal. Quant. Cytol. Histol.* **13**(5) (1991), 307–315.
[100] H. Van der Poel et al., Prognostic value of karyometric and clinical characteristics in renal cell carcinoma. Quantitative assessment of tumour heterogeneity, *Cancer* **72** (1993), 2667–2674.
[101] H. Van der Poel, P. Mulders, T. Aalders, G. Oosterhof, F. Debruyne and J. Schalken, Karyometric analysis of intra-tumour heterogeneity in prostate adenocarcinoma, *Anal. Cell. Pathol.* **7** (1994), 153–170.
[102] N. Pressman, Markovian analysis of cervical cell images, *Histochem. Cytochem.* **24** (1976), 138–144.
[103] P. Pudil, J. Novovicova and J. Kittler, Floating search methods in feature selection, *Pat. Rec. Lett.* **15** (1994), 1119–1125.
[104] S. Raudys and A. Jain, Small sample size effects in statistical pattern recognition: recommendations for practitioners, *IEEE Trans. Pattern Anal. Mach. Intell.* **13** (1991), 252–264.
[105] F. Rickaert et al., Computerized morphonuclear characteristics and DNA content of adenocarcinoma of the pancreas, chronic pancreatitis and normal tissues: relationship with histopathologic grading, *Hum. Pathol.* **23**(11) (1992), 1210–1215.
[106] B. Ripley, Statistical aspects of neural networks, in: *Network and Chaos – Statistical and Probabilistic Aspects*, Barndorff-Nielsen, Jensen and Kendall, eds, Chapman and Hall, 1993.
[107] A. Robertson et al., Observer variability in histopathological reporting of cervical biopsy specimens, *J. Clin. Pathol.* **42** (1989), 231–238.
[108] D. Rosenthal and S. Suffin, Predictive value of digitized cell images for the prognosis of cervical neoplasia, *Monogr. Clin. Cytol.* **9** (1984), 163–180.
[109] M. Sadi and E. Barrack, Image analysis of androgen receptor immunostaining in metastatic prostate cancer, *Cancer* **71** (1993), 2574–2580.
[110] I. Salmon et al., Comparison of morphonuclear features in normal, benign and neoplastic thyroid tissue by digital cell image analysis, *Anal. Quant. Cytol. Histol.* **14**(1) (1992), 47–54.
[111] SAS, *STAT User Guide*, 4th edn, SAS Institute, 1992.
[112] H. Schulerud, J. Carstensen and H. Danielsen, Multiresolution texture analysis of four classes of mice liver cells using different cell cluster representation, *Proc. 9th Scand. Conf. Image Analysis* **1** (1995), 121–129.
[113] H. Schulerud, G.B. Kristensen, L. Vlatkovic, F. Albregtsen, K. Liestøl and H.E. Danielsen, Prognosis of cervical cancer using image analysis of cell nuclei, *Proc. 10th Scand. Conf. Image Analysis* **2** (1997), 651–658.
[114] E. Schulte, U. Joos, M. Kasper and H. Eckert, Cytological detection of epithelial dysplasia in the oral mucosa using Feulgen-DNA-image cytometry, *Diag. Cytopathol.* **7**(4) (1991), 436–441.
[115] D. Seigneurin, J. Louis and M.C. Villoud, The value of DNA image cytometry for the cytological diagnosis of well-differentiated breast carcinomas and benign lesions, *Anal. Cell. Pathol.* **7**(2) (1994), 115–125.
[116] J. Shaeffer, J.A. Tegeler, D.A. Kuban, C.B. Philput and A.M. el Mahdi, Nuclear roundness factor and local failure from definitive radiation therapy for prostatic carcinoma, *Int. J. Rad. Oncol. Biol. Phys.* **24**(3) (1992), 431–434.
[117] C. Sowter, G. Slavin, G. Sowter, D. Rosen and W. Hendry, Morphometry of bladder carcinoma, morphometry and grading complement each other, *Anal. Cell. Pathol.* **3** (1991), 1–9.
[118] P.J. Spaander, D.J. Ruiter, J. Hermans, H.J. de Voogt, J.A. Brussee and M.E. Boon, The implications of subjective recognition of malignant cells in aspirations for the grading of prostatic cancer using cell image analysis, *Anal. Quant. Cytol.* **4**(2) (1982), 123–127.
[119] D. Spina et al., Novel, contrast gradient-oriented, automated chromatin texture analysis. I. Feasibility study on nuclei from benign and malignant breast epithelial cell lines in fine needle aspirates, *Virchows Arch.* **62**(2) (1992), 119–124.
[120] S. Stearns, On selecting features or pattern classifiers, in: *Proc. 3rd Int. Conf. Pattern Recognition*, 1976, pp. 71–75.
[121] H. Stich, B. Palcic, R. Sankaranarayanan, B. Mathew and N.M. Krishnan, Quantitation of chromatin patterns by image analysis as a predictive tool in chemopreventive trials with vitamin A, *IARC Sci. Publ.* **104** (1990), 151–163.
[122] B. Susnik, N. Poulin, D. Phillips, J. LeRiche and B. Palcic, Comparison of DNA measurement performed by flow and image cytometry of embedded breast tissue sections, *Anal. Quant. Cytol. Histol.* **17**(3) (1995), 163–171.
[123] P. Tosi et al., Heterogeneous subgroups among malignant diffuse small B cell lymphomas. A combined nucleometric and immunocytologic study, *Lab. Invest.* **62**(2) (1990), 202–212.
[124] P.D. Unger, C.W. Watson, Z. Liu and J. Gil, Morphometric analysis of neoplastic renal aspirates and benign renal tissue, *Anal. Quant. Cytol. Histol.* **15**(1) (1993), 61–66.
[125] M. Unser, Sum and difference histogram for texture classification, *IEEE Trans. Pattern Anal. Mach. Intell.* **8** (1986), 118–125.

[126] J. Vasko, Prognosis in bladder cancer. A study of cytometric, morphometric and immunohistochemical techniques, *Scand. J. Urol. Nephrol.* (Suppl.) **160** (1994), 1–73.
[127] R. Veltri et al., Quantitative nuclear morphometry, Markovian texture descriptors, and DNA content captured on a CAS-200 image analysis system, combined with PCNA and HER-2/neu immunohistochemistry for prediction of prostate cancer progression, *J. Cell Biochem.* (Suppl.) **19** (1994), 249–258.
[128] A. Verhest et al., Characterization of human colorectal mucosa, polyps and cancer by means of computerized morphonuclear image analyses, *Cancer* **65** (1990), 2047–2054.
[129] N. Wang, B. Stenkvist and B. Tribukait, Morphometry of nuclei of the normal and malignant prostate in relation to DNA ploidy, *Anal. Quant. Cytol. Histol.* **14**(3) (1992), 210–216.
[130] A. Weger et al., The value of morphometry to predict chemotherapy response in advanced ovarian cancer, *Pathol. Res. Pract.* **185**(5) (1989), 676–679.
[131] N. Wheeler, S.C. Suffin, T.L. Hall and D.L. Rosenthal, Prediction of cervical neoplasia diagnosis groups. Discriminant analysis on digitized cell images, *Anal. Quant. Cytol. Histol.* **9**(2) (1987), 169–181.
[132] A. Whitney, A direct method of nonparametric measurement selection, *IEEE Trans. Comput.* **20** (1971), 1100–1103.
[133] G.L. Wied, P.H. Bartels, H.E. Dytch, F.T. Pishotta, K. Yamauchi and M. Bibbo, Diagnostic marker features in dysplastic cells from the uterine cervix, *Acta Cytol.* **26**(4) (1982), 475–483.
[134] G. Wolf, M. Beil and H. Guski, Chromatin structure analysis based on a hierarchic texture model, *Anal. Quant. Cytol. Histol.* **17**(1) (1995), 25–34.
[135] Q. Yang et al., Morphometric diagnosis of bladder tumour, *Oncology* **48**(3) (1991), 188–193.
[136] K. Yogesan, T. Jørgensen, F. Albregtsen, K.J. Tveter and H.E. Danielsen, Entropy-based texture analysis of chromatin structure in advanced prostate cancer, *Cytometry* **24** (1996), 268–276.
[137] 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.
[138] R. Zaino et al., Histopathologic predictors of the behavior of surgically treated stage 1B squamous cell carcinoma of the cervix, *Cancer* **69** (1992), 1750–1758.
[139] D. Zongker and A. Jain, Algorithms for feature selection: An evaluation, *Proc. ICPR* **B** (1996), 18–22.