# The potential of feature selection by statistical techniques and the use of statistical classifiers in the discrimination of benign from malignant gastric lesions

TATIANA MONA MEGALOPOULOU[1], KONSTANTINOS KOUTROUMBAS[2], ABRAHAM POULIAKIS[1], GREGORY SIVOLAPENKO[3] and PETROS KARAKITSOS[4]

[1]Department of Histology and Embryology, Medical School of Athens, Athens University; [2]Institute for Space Applications and Remote Sensing, National Observatory of Athens; [3]Pharmacy School, Athens University; [4]Department of Cytopathology, University General Hospital ‘Attikon’, Athens, Greece

Received September 6, 2005; Accepted September 27, 2005

**Abstract.** The objective of this study was the investigation of the potential value of morphometry, feature selection and statistical classifiers techniques, such as neural networks, for the classification of benign from malignant gastric nuclei and cases. One hundred and twenty gastric smears, routinely processed and stained by Papanicolaou technique, were analyzed by a customized image analysis system. Data from half of the cases were selected to form the training set, while the remaining data formed the test set. A feature selection technique was applied in order to identify the most important nuclear features, which were used in a second stage by statistical classifiers to classify a nucleus as benign or malignant. Using the classifier results for the nuclear classification, a method to classify each individual patient was developed. The performance of the proposed method was validated through the test set. The technique described in this report produces significant results at the nuclear and patient level and promises to be a powerful assistance tool for everyday cytological laboratory routine.

## Introduction

The use of fiberoptic endoscopy has dramatically increased the accuracy of diagnostic gastroenterology. However, it is widely accepted that endoscopy alone is not sufficient to provide an accurate diagnosis of gastric lesions. The current practice of multiple biopsies as an additional examination significantly increases the overall accuracy. Moreover, the combined application of cytology and histology for diagnosis could increase the accuracy to >90%. However, the cytological approach of gastric brush smears has a second order diagnostic significance because of the difficulties in discriminating benign cells with severe regenerative alterations from well-differentiated cancer cells resulting in relatively low sensitivity (1).

In order to increase the diagnostic accuracy, nuclear morphometric and densitometric data have been used in combination with linear and non-linear classifiers. Moreover, in order to have an objective diagnosis both in histological sections and cytological smears, various combinations of morphometric data and classifiers have been applied with various results and on different organs (2,3). Especially for gastric lesions substantial efforts have been made including pure discriminant analysis (4) and various types of neural network classifiers (5,6). However, there is no report on the comparison of linear against non-linear classification techniques especially when the features are selected using statistical criteria.

## Materials and methods

One hundred and twenty gastric smears were examined during routine diagnosis at the Department of Histology and Embryology of the Medical School of Athens University (Athens, Greece). These were routinely processed and stained by Papanicolaou technique, and were analyzed by a customized image analysis system (4). The result of this procedure was a data set consisted of 13300 measured cell nuclei, each one being represented by a pre-specified set of morphometric measures, called features.

The study was carried out on brushing cytology smears taken during endoscopy from 120 cases with gastric lesions. The smears were routinely fixed in 96˚ ethanol for 30 min and stained using Papanicolaou technique. The cytological diagnosis was made at least by two cytopathologists with 10 years experience in gastric cytology, and further confirmed by the histological examination of biopsies and/or the surgical specimens. The correlation of cytological and histological findings is presented in Table I.

According to the cytological diagnosis, each cell was assigned to one of the following groups: a) ulcer, b) gastritis,

*Correspondence to*: Dr Petros Karakitsos, Department of Histology and Embryology, Iros Konstantopoulou 13, 16121 Kaisariani, Athens, Greece
E-mail: pkaraki@med.uoa.gr

*Key words*: discriminant analysis, morphometry, gastric lesions, quantitative cytology, linear classifier, Bayesian classifier, feed-forward neural network, combined neural network

Table I. Relation of 120 gastric cases according to histological and cytological findings.

| | Histological diagnosis | | |
|---|---|---|---|
| Cytological diagnosis | Ulcer | Gastritis | Cancer |
| Ulcer | 60 | | |
| Gastritis | | 25 | |
| Inflammatory dysplasia | 4 | 1 | |
| True dysplasia | | | 5 |
| Cancer | | | 25 |

Table II. Distribution of 13300 cells of sample size to the cytological diagnostic categories.

| Cytological diagnosis | Total sample size N (%) |
|---|---|
| Ulcer | 6550 (49.2) |
| Gastritis | 3150 (23.7) |
| Inflammatory dysplasia | 310 (2.3) |
| Cancer | 2920 (22.0) |
| True dysplasia | 370 (2.8) |
| Total | 13300 |

c) inflammatory dysplasia, d) cancer, and e) true dysplasia (the distribution of the assigned cells is shown in Table II). In our study, the golden standard for malignant cases is the result of the histological examination. From each case about 100 cells were measured. In typical cases the most representative cells were selected by the cytopathologists; from the remaining cases the most atypical cells were selected and used to assign a category according to the opinion of both cytopathologists.

The total sample of 120 cases (13300 cells) was divided into two groups, the training group (60 cases corresponding to 6391 cells) and the test group from the remaining 50% of the cases (6909 cells). To preserve the data structure of the groups into the divided sets, 50% of cases from each diagnostic category (benign and malignant) were randomly selected (using simple random stratified sampling) to form the training set, the remaining cases were used to form the test set. Therefore for each set, 15 out of the 30 malignant cases and 45 out of 90 benign cases were extracted by using stratified random sampling (Table III). The benign diagnostic category was composed of cases classified as inflammatory dysplasia, ulcer or gastritis, while the malignant category involves cases diagnosed as cancer and true dysplasia. Cases classified by the cytological examination as inflammatory or true dysplasia were further confirmed by the histological findings.

*Nuclear features.* The measured features are grouped according to the physical characteristics of nuclei into two categories: geometric and densitometric. The geometric characteristics describe properties relative to the size and properties that provide information about the shape of nuclei. These are: area, circularity, major axis, minor axis, perimeter, form area, form perimeter, nuclear contour index, contour ratio, roundness factor, diameter. The densitometric features are relevant to the chromatin texture. From the various methods that have been proposed in the literature for the description of chromatin texture, four models were implemented, namely: histogram, differences histogram, run length matrix and co-occurrence matrix. The densitometric features employed in this study are: mean value, standard deviation and variance of histogram, short run, long run, grey level and distribution of run length matrix, maximum value, entropy and inertia of the co-occurrence matrix, mean value, variance, contrast and entropy of the differences histogram. For more details on the measured features and the image analysis system see refs. 4,7,8.

Table III. Distribution of cells in the training and test set.

| | Training set | Test set |
|---|---|---|
| Malignant | 1541 (15 cases) | 1749 (15 cases) |
| Benign | 4850 (45 cases) | 5160 (45 cases) |
| Total | 6391 (60 cases) | 6909 (60 cases) |

*Feature selection.* This step follows the extraction of the features from cell nuclei, here the most important characteristics are identified and thus the number of features for nuclear classification is reduced. This task was carried out by adopting an empirical transformation on the features (among several others) that fulfills the following three requirements (9): a) normality, which was tested by the skewness test, b) absence of collinearity and multicollinearity, which were tested by the Pearson correlation coefficients and the tolerance statistic respectively, c) homoscedasticity, i.e. the null hypothesis that the variance-covariance matrices of the two groups (benign and malignant) are equal, which was tested using the Box's M statistic. The above requirements are necessary for performing discriminant analysis.

The existence of collinearity and multicollinearity among some features and the application of discriminant analysis reduced the features representing a cell to 8 from 25. These are: form area, contour ratio, roundness factor, mean value of histogram, grey level of run length matrix, inertia of the co-occurrence matrix, variance and entropy of the differences histogram.

### *Classification*

*Classification per nucleus.* In this work four parametric classifiers and one non-parametric were applied. The parametric classifiers adopted in this work are: a) the linear classifier, b) Bayesian classifiers, where the estimation of the probability density functions of the two classes was carried out using the EM-algorithm, c) two layer feedforward neural network (2L FNN) classifiers, and d) combined two layer feedforward neural network classifiers generated by the Adaboost algorithm. The non-parametric classifier is the k-nearest neighbor algorithm (10).

Table IV. Nuclei classification results for the training and test sets.

| | Linear classifier | Bayesian classifier (50 normal distr. per class) | 2L FNN (nodes = 40) | Boosting with three 2L FNN (nodes = 40) | 5-nearest neighbor |
|---|---|---|---|---|---|
| Training set (%) | 87.98 | 96.42 | 97.01 | 97.83 | 95.35 |
| Test set (%) | 87.73 | 95.56 | 96.11 | 96.41 | 94.47 |

Table V. Confusion matrices and parameters of the classifiers for the patient classification.

| | Linear classifier | Bayesian classifier (40 normal distr. per class) | 2L FNN (nodes = 40) | Boosting with three 2L FNN (nodes = 40) | 5-nearest neighbor |
|---|---|---|---|---|---|
| Confusion matrix on the training set | 39 6 0<br>0 15 0<br>0 0 0 | 45 0 0<br>0 15 0<br>0 0 0 | 45 0 0<br>0 15 0<br>0 0 0 | 45 0 0<br>0 15 0<br>0 0 0 | 45 0 0<br>0 15 0<br>0 0 0 |
| Confusion matrix on the test set | 40 5 0<br>0 15 0<br>0 0 0 | 45 0 0<br>0 15 0<br>0 0 0 | 45 0 0<br>0 15 0<br>0 0 0 | 45 0 0<br>0 15 0<br>0 0 0 | 45 0 0<br>0 15 0<br>0 0 0 |
| $\theta_1$ (%) | 50.50 | 62.00 | 65.33 | 69.50 | 70.33 |
| $\theta_2$ (%) | 50.50 | 28.25 | 34.58 | 34.53 | 42.18 |
| $n_b^{min}$ for the training set (%) | 51.00 | 70.00 | 73.33 | 77.50 | 78.33 |
| $n_m^{max}$ for the training set (%) | 50.00 | 20.25 | 26.58 | 26.53 | 34.18 |
| $n_b^{min}$ for the test set (%) | 50.44 | 77.19 | 75.44 | 79.82 | 82.46 |
| $n_m^{max}$ for the test set (%) | 46.00 | 23.88 | 25.36 | 24.27 | 37.68 |

*Classification per case*. For the classification of a test case as benign or malignant two thresholds $\theta_1$ and $\theta_2$ ($\theta_1>\theta_2$) on the percentage of the benign nuclei of a case are established. If the percentage of the benign nuclei of the case at hand, $n_b$, is greater than $\theta_1$, i.e. $n_b>\theta_1$, then the case is characterized as benign. Else, if $n_b<\theta_2$, the case is characterized as malignant. Otherwise, no decision is made. The thresholds $\theta_1$ and $\theta_2$ are estimated via the following procedure: a) classify all the nuclei of the training set using a specific classifier; b) for each case whose nuclei are included in the training set, the percentage of the nuclei classified as benign, $p_b$, and the percentage of the nuclei classified as malignant $p_m$ are computed. If $p_b>p_m$, for this case, then the case is characterized as benign. Otherwise it is characterized as malignant; c) among all the cases of the training set characterized as malignant, determine the one with the maximum percentage of benign nuclei, say $n_m^{max}$. Also, among all the cases of the training set characterized as benign, determine the one with the minimum percentage of benign nuclei, say $n_b^{min}$; d) if $n_m^{max} + 8\% < n_b^{min} - 8\%$, then set $\theta_1 = n_b^{min} - 8\%$ and $\theta_2 = n_m^{max} + 8\%$. Otherwise, set $\theta_1 = \theta_2 = \left( n_m^{max} + n_b^{min}\right)/2$.

Note that since the nucleus classifiers used in the previous section achieve very high classification performance, and since for the malignant cases, only the most suspect for malignancy cells have been adopted, it is expected that for a malignant case a small percentage of the cells will be benign. Also, for a benign case a significant percentage of the nuclei will be benign. Therefore, it is expected that the condition $n_m^{max} + 8\% < n_b^{min} - 8\%$ is satisfied for most of the classifiers, which is verified by the experiments. In addition, it is expected for most of the classifiers, the undecided cases will be very rare. It is worth mentioning that the values of $\theta_1$ and $\theta_2$ depend on the specific classifier used. However, for most of the cases, the estimates for $\theta_1$ obtained by the use of most of the above classifiers are close to each other (the same holds for $\theta_2$).

## Results

The best results (success rates) obtained for each classifier for the nuclear classification are summarized in Table IV. From these results we see that all non-linear classifiers exhibit significantly better performance than the linear classifier. In addition, the 2L FNN and the boosting classifiers per-form slightly better than the Bayes and 5-nearest neighbor classifier.

Table V contains the results obtained after applying the case classification strategy for each one of the classifiers. In the confusion matrices the columns and the rows correspond to the benign, malignant and 'do not know' case. In Table V we see that all (except the linear) classifiers classify correctly all the cases. It is also worth mentioning that the range ($\theta_1$, $\theta_2$) is large enough for all (but the linear) classifiers, which indicates high degree of confidence in the classification of each case.

## Discussion

Although it appears to increase the accuracy of the conventional diagnostic procedure of endoscopic biopsies, cytological examination of gastric lesions has not been widely accepted due to the difficulties that appear in the discrimination of the highly differentiated carcinomas from benign ulcers with increased regenerative alterations (1). This problem was present as well, in the morphological examination of the material used in this study. In order to improve the cytological diagnosis the potential of discriminant analysis combined with the linear classifier or non-linear classifiers were evaluated.

According to the findings the statistically most important features appear to be: form area, contour ratio, roundness factor, mean value of histogram, grey level of run length matrix, inertia of the co-occurrence matrix, variance of the differences histogram, entropy of the differences histogram. These features are related to the nuclear shape, the chromatin density and the chromatin texture, features that are important for the cytological diagnosis as well.

The validation of the classifiers in the cell nucleus level indicates that among the four non-linear classifiers the best results are provided by Boosting with three 2L FNN (nodes = 40). The difference in the overall accuracy between the linear classifier and the worst of the non-linear classifiers (5-nearest neighbor) indicates that the linear classifiers have statistically important reduced performance ($Z=-13.9$, $p<0.001$). These findings indicate that the clear discrimination between regenerative and well-differentiated adenocarcinomas is not possible, even though the nuclear characteristics are quantified. Usage of the nuclear classification results for patient classification via the proposed methodology indicated that the four non-linear classifiers do not produce false positive or negatives, even for the dysplasia cases that require further examination. In contrast, the linear classifier misclassified 5 out of 45 negative cases. Therefore, in an attempt to apply this methodology as a decision support system in order to decrease the uncertainty of the morphological diagnosis, it seems better to avoid employing the linear classifier.

The combination of morphometric nuclear characteristics, with statistical feature selection and statistical classification techniques, as described in this report, produces significant results at the nuclear and patient level and promises to be a powerful diagnostic assistance tool for the everyday cytological diagnostic routine.

## Acknowledgments

Part of this work was financially supported by the Special Account for Research Grants of the University of Athens KAE:70/31703 and by 01-Pliroforiki SA. Under the EC project VALUE/CCS with contract number CTT-471.

## References

1. Kasugai T and Kobayashi S: Evaluation of biopsy and cytology in the diagnosis of gastric cancer. Am J Gastrenterol 62: 199-203, 1974.
2. Boon ME, Kurver PHJ, Baak JPA, *et al*: The application of morphometry in gastric cytological diagnosis. Virchows Arch 393: 159-164, 1981.
3. Wolfe P, Murphy J, McGinley J, *et al*: Using nuclear morphometry to discriminate the tumorigenic potential of cells: a comparison of statistical methods. Cancer Epidemiol Biomarkers Prev 13: 976-988, 2004.
4. Karakitsos P, Megalopoulou TM, Pouliakis A, Tzivras M, Archimandritis A and Kyroudes A: The application of discriminant analysis and quantitative cytological examination of gastric lesions. Anal Quant Cytol Histol 26: 314-322, 2004.
5. Karakitsos P, Pouliakis A, Koutroumbas K, Botsoli-Stergiou E, Tzivras M, Archimandritis A and Ioakim-Liossi A: Neural network application in the discrimination of benign from malignant gastric cells. Anal Quant Cytol Histol 22: 63-69, 2000.
6. Karakitsos P, Botsoli-Stergiou E, Pouliakis A, Tzivras M, Archimandritis A, Liossi A and Kyrkou K: Comparative study of artificial neural networks in the discrimination of benign from malignant gastric cells. Anal Quant Cytol Histol 19: 145-152, 1997.
7. Pitas I: Digital Image Processing Algorithms. Prentice-Hall, London, 1993.
8. Sonka M, Hlavac V and Boyle R: Image Processing Analysis and Machine Vision. Chapman & Hall, 1994.
9. Hair JF Jr, Anderson RE, Tatham RL and Black WC: Multivariate Data Analysis. 5th edition. Prentice-Hall International, Inc., London, 1998.
10. Hastie T, Tibshirani R and Friedman J: The Elements of Statistical Learning. Springer, New York, 2001.