SPANDIDOS PUBLICATIONS

# Discriminating benign from malignant thyroid lesions using artificial intelligence and statistical selection of morphometric features

BEATRIX COCHAND-PRIOLLET[1], KONSTANTINOS KOUTROUMBAS[2], TATIANA MONA MEGALOPOULOU[3], ABRAHAM POULIAKIS[3], GREGORY SIVOLAPENKO[4] and PETROS KARAKITSOS[5]

[1]Service Central d'Anatomie et de Cytologie Pathologiques, Hospital Lariboisiere, Paris, France; [2]Institute for Space Applications and Remote Sensing, National Observatory of Athens; [3]Department of Histology and Embryology, Medical School of Athens, Athens University, Athens; [4]Pharmacy School, Patras University, Patras; [5]Department of Cytopathology, University General Hospital 'Attikon', Athens, Greece

Received September 6, 2005; Accepted September 27, 2005

**Abstract.** The objective of this study was to perform a comparative investigation of the capability of various classifiers in discriminating benign from malignant thyroid lesions. Using May Grunvald-Giemsa-stained smears taken by fine needle aspiration (FNA) and a custom image analysis system, 25 nuclear features describing the size, shape and texture of the nuclei were measured in each case. A statistical pre-processing of features revealed that only 4 of the 25 features are important when discriminating benign from malignant thyroid lesions, which were transformed and fed to four classifiers for subsequent analysis. The cases were divided into one set used for the training of classifiers, a second set used as the test set, and the remaining cases with no clear classification formed an ambiguous test set. Classification was performed at the nuclear and patient level. The technique described in this study produced encouraging results and promises to be a helpful tool in the daily cytological laboratory routine.

## Introduction

The echotomographic investigation shows that about 30% of asymptomatic adults have thyroid nodules. However, carcinoma of the thyroid gland is diagnosed in 5-30% of all patients with nodular thyroid lesions referred for examination, and 4-40% of patients undergoing surgical intervention for thyroid pathology (1). No clinical symptoms or biologic and radiological findings allow the accurate discrimination between benign and malignant nodules. Fine needle aspiration (FNA) has gained wide acceptance in the investigation of thyroid lesions as it allows a dramatic decrease of unnecessary surgical excisions of thyroid gland, and is now recommended by most national and international guidelines as a fundamental part of the decision making process for all thyroid nodules measuring ≥1 cm. However, some diagnostic dilemmas in cytological and histological examination are also present (2). Automated cytology techniques such as DNA ploidy or morphometry have been applied to FNA and surgical material to obtain a safer diagnosis. All of these studies agree that the classification of nuclei is successful, but there are disagreements over its usefulness in classifying patient cases (3). This study investigated the potential role of morphometry combined with statistical classifiers in discriminating benign from malignant thyroid lesions in routinely prepared FNA smears.

## Materials and methods

This study was performed on 157 cases of FNA from an equal number of patients examined in Lariboisiere (Paris, France) and at the Department of Histology and Embryology of the Medical School of Athens University (Athens, Greece). The results of the cytological examination were confirmed by histology of the surgical specimens. The cytological smears were stained using the Giemsa technique and examined with a customized image analysis system described previously (4). The result of this procedure was a data set consisting of 7940 measured cell nuclei, each represented by a pre-specified set of morphometric measurements called features. These cell nuclei were extracted from 35 cases of goiter, 41 cases of benign oncocytomas, 38 cases of papillary carcinoma, 6 cases of malignant Hurthle, 2 cases of medullary carcinomas, 2 cases of follicular adenomas and 2 cases of follicular carcinomas. These 126 cases were divided into two sets, the training set which consists of 64

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

*Key words*: discriminant analysis, morphometry, thyroid lesions, quantitative cytology, linear classifier, Bayesian classifier, feedforward neural network, combined neural network

Table I. Success rates for the nuclear classifiers.

| | Linear classifier | 2L FNN (nodes = 50) | Boosting with three 2L FNN (nodes = 5) | Boosting with three 2L FNN (nodes = 30) | 25-nearest neighbor |
|---|---|---|---|---|---|
| Training set | 66.40% | 75.77% | 72.56% | 77.78% | 70.99% |
| Test set | 65.17% | 73.20% | 75.25% | 73.20% | 74.69% |
| Ambiguous set | 57.33% | 59.06% | 55.47% | 58.44% | 61.35% |

Table II. Confusion matrices and classifier characteristics for the patient classifiers.

| | Linear classifier | 2L FNN (nodes = 50) | Boosting with three 2L FNN (nodes = 5) | Boosting with three 2L FNN (nodes = 30) | 25-nearest neighbor |
|---|---|---|---|---|---|
| Confusion matrix on the training set | 1 23<br>0 40 | 21 3<br>1 39 | 17 7<br>2 38 | 21 3<br>0 40 | 14 10<br>1 39 |
| Overall accuracy | 64% | 94% | 86% | 94% | 83% |
| Confusion matrix on the test set | 0 24<br>0 38 | 18 6<br>1 37 | 17 7<br>0 38 | 17 7<br>1 37 | 16 8<br>0 38 |
| Overall accuracy | 61% | 89% | 89% | 87% | 87% |
| Confusion matrix on the ambiguous set | 0 12<br>0 19 | 5 7<br>0 19 | 5 7<br>1 18 | 6 6<br>0 19 | 6 6<br>0 19 |
| Overall accuracy | 61% | 77% | 74% | 81% | 81% |
| $\theta$ | 53.57% | 50.49% | 49.71% | 49.87% | 50.64% |
| $n_b^{min}$ for the training set | 57.14% | 50.98% | 50.77% | 51.25% | 52.28% |
| $n_m^{max}$ for the training set | 50.00% | 50.00% | 48.65% | 48.48% | 50.00% |

cases and test set (n=62), using a simple random stratified sampling. The remaining 31 cases were deemed suspect in the cytological report, and cytopathologists were unable to assign them to one of the above categories (benign or malignant) during the routine diagnostic procedure. However, some evidence was available from the histological examination, and these cases therefore formed the ambiguous set, which was not used for classifier training, but instead for testing statistical classifiers considered in this study.

*Nuclear features and feature selection*. The set of 25 features measured for each nucleus (see ref. 4 for more details) were reduced, using a methodology described in our previous studies (4,5), to only 4 features: roundness factor, standard deviation of the histogram, maximum value of the co-occurrence matrix, and mean value of the differences histogram.

*Classification*

*Classification per nucleus*. In this case, three parametric classifiers and one non-parametric were utilized. Specifically the parametric classifiers adopted in this study were: a) the linear classifier, b) two layer feedforward neural network (2L FNN) classifiers and c) combined two layer feedforward neural network classifiers generated by the Adaboost algorithm. The non-parametric classifier was the k-nearest neighbor algorithm (see ref. 6).

*Classification per case*. For the classification of a test case as benign or malignant, a threshold $\theta$ on the percentage of the benign nuclei of a case was established. If the percentage of the benign nuclei of the case, $n_b$, is greater than $\theta$, i.e. $n_b > \theta$ then the case is characterized as benign. Otherwise, when $n_b < \theta$, the case is characterized as malignant. The threshold was estimated via the following procedure:

i) Classify all nuclei of the training set using a specific classifier.

ii) 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 it is characterized as benign. Otherwise, the case is characterized as malignant.

SPANDIDOS PUBLICATIONS

ng all cases of the training set characterized as , determine the one with the maximum percentage of benign nuclei, say $n_m^{max}$. Also, among all cases of the training set characterized as benign, determine the one with the minimum percentage of benign nuclei, say $n_b^{min}$.

iv) Set, $\theta = (n_m^{max} + n_b^{min})/2$.

Note that since for the malignant cases, only the most suspect for malignancy nuclei have been adopted, it is expected that a smaller percentage of the nuclei will be benign for a malignant case, compared to the percentage of benign nuclei of a benign case. Also, a significant percentage of the nuclei will be benign for a benign case. It is worth mentioning that the value of $\theta$ depends on the specific classifier used. However, for most cases, the estimates obtained for $\theta$ by using most of the above classifiers were similar to each other.

## Results

The best results (success rates) obtained for each classifier in the nuclear classification are summarized in Table I. As shown, all classifiers (except the linear one) exhibit relatively the same performance on the test set. In addition, they perform significantly better than the linear classifier. Also, there were no significant deviations in the performance of all classifiers in the ambiguous set. In these cases, the 25-NN classifier had a slightly better performance compared to the others.

The results of Table II were obtained by applying the case classification strategy for each of the classifiers. In the confusion matrices, the columns and rows correspond to benign and malignant cases, respectively; the rows indicate the diagnosis category, and columns indicate the category suggested by the classifiers. Also, $n_b^{min}$ is the minimum number of benign nuclei among all benign cases in the training set, and $n_m^{max}$ is the maximum number of benign nuclei among all malignant cases in the training set for each classifier. As expected, $n_b^{min}$ and $n_m^{max}$ are different among the tested classifiers according to their performance at the nuclear level. All classifiers (except the linear one) exhibit relatively similar performances both on the test and ambiguous sets. However, performance on the test set was better than that on the ambiguous set.

## Discussion

Most published studies reported a high detection sensitivity and lower specificity, giving a satisfying overall accuracy (2,3). However, FNA has two major disadvantages: 1) sampling is not always representative of the tumor; and 2) cases from follicular and Hürthle cell lesions cannot be correctly classified by cytologic criteria as benign or malignant (1,2).

Based on the statistical behavior of data from thyroid lesions taken by means of morphometry, we attempted to investigate the potential role of several classifiers in discriminating benign from malignant thyroid lesions in routinely prepared FNA smears. The results of the linear classifier indicate that the malignant thyroid nuclei are not linearly separable from benign thyroid nuclei. This fact could explain the difficulties encountered by cytopathologists during routine examination of thyroid lesions.

On the contrary, non-linear classifiers gave better classification results. However, the observed overlapping between categories, indicated by all classifiers reaching an upper classification level, is strong evidence that more features may be required to achieve better classification accuracy.

Despite the diagnostic procedure being more objective in the case of artificial intelligence, as well as in every diagnostic test, there are still false positive and false negative results. Specifically, at the clinical level, in the cases where there is coincidence of the cytopathologist and the classifier classification, the diagnostic result is 100% correct. In suspect cases (ambiguous set) where the diagnosis is a neoplasm (i.e. the nature of the neoplasm cannot be identified by cytological examination), the positive diagnosis of the classifier is correct. However, in cases where the classifier diagnosis is negative and the cytological diagnosis favors malignancy, we cannot draw any conclusions. Therefore, the combination of artificial intelligence and cytological diagnosis may decrease the number of unnecessary surgical excisions because the uncertainty of cytological diagnosis is reduced.

According to the classical morphological approach, confirmation of the malignancy is based on architectural rather than cytological features (8,9). However, the size of the nucleus and presence of intranucleus cytoplasmic inclusions have been considered important features for discriminating benign from malignant cells. Nevertheless, discriminating between follicular, benign and malignant lesions is difficult, and many cases are characterized as follicular neoplasms after performing a routine diagnostic procedure. In histological evaluation, only 30% of samples are really follicular carcinomas (1,10,11). Efforts have been made to improve the accuracy of morphologic diagnosis using digitized images analyzed by different techniques (12-15). Among these techniques, three have produced clinically significant results (16-18).

This study attempted to evaluate the potential of discriminant analysis combined with the linear or non-linear classifiers. The statistical preprocessing of features indicated that the most important features in discriminating benign from malignant cases appear to be nuclear shape (roundness factor), chromatin texture (maximum value of the co-occurrence matrix and mean value of the differences histogram) and the distribution of chromatin (standard deviation of the histogram). This issue requires further investigation by cytologists to evaluate if it can be applied in a daily routine or during training on thyroid FNA. In digitized images, the linear classifier choice gave poor results at the nuclear level and proved to be incapable of discrimination at the patient level. Non-linear classifiers gave better results. The most promising classifier proved to be the KNN classifier. At the patient level, the most important classifier (the combined two layer feedforward neural network classifier generated by the Adaboost with 14 2L FNN -5 nodes) had an overall accuracy of 88.71%, and a certainty of PVPR =100% in the diagnosis of malignancy.

Of the 7 missed malignant cases, 4 corresponded to papillary carcinomas, 2 to Hürtle cell carcinomas and 1 to follicular carcinoma. Therefore, the combined application of

the best classifier and morphological diagnosis does not affect the correct patient treatment. The fact that no histological benign lesions were misclassified could lead to the conclusion that application of this method may decrease the uncertainty of morphological diagnosis.

Histological examination of the 31 cytologically ambiguous cases revealed 7 papillary carcinomas, 1 follicular carcinoma, 4 follicular adenomas, 16 Goiter and 3 iodine effect. Of these cases, the combined two layer feedforward neural network classifier generated by the Adaboost (with 14 2L FNN -5 nodes) had overestimated 1 case of iodine effect in Graves disease and underestimated 2 cases of papillary carcinoma. In practical terms, combining the morphological diagnosis and classifier would have led to only 1 unnecessary surgical intervention.

Finally, one should note that the results obtained at the case level were better than those obtained at the nuclear level in terms of percentages. In addition, the results at the nuclear level differ significantly from those concerning other organs, such as the stomach (4). This indicates that the task of determining the best possible classifiers to assist the cytologist should be carried out for each organ separately.

**Acknowledgements**

This work was financially supported in part by the Special Account for Research Grants of the University of Athens (KAE:70/31703).

**References**

1. Kojic Katovic S, Halbauer M and Tomic-Brzac H: Importance of FNAC in the detection of tumours within multinodular goitre of the thyroid. Cytopathology 15: 206-211, 2004.
2. Cochand-Priollet B, Guillausseau PJ, Chagnon S, Hoang C, Guillauseau-Sholer CL, Chanson PH, Dahn H, Warnet A, Tran PBH and Valleur P: The diagnostic value of fine needle aspiration biopsy under ultra-sonography in non-functional thyroid nodules: a prospective study comparing cytologic and histologic findings. Am J Med 97: 152-157, 1994.
3. Ambros RA, Trost RC, Campbell AY and Lambert WC: Prognostic value of morphometry in papillary thyroid carcinoma. Hum Pathol 20: 215-218, 1989.
4. Karakitsos P, Megalopoulou TM, Pouliakis M, 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. Hair JF Jr, Anderson RE, Tatham RL and Black WC: Multivariate Data Analysis. 5th edition. Prentice-Hall International, Inc., London, 1998.
6. Hastie T, Tibshirani R and Friedman J: The Elements of Statistical Learning. Springer, New York, 2001.
7. Megalopoulou TM, Koutroumbas K, Pouliakis A, Sivolapenko G and Karakitsos P: The potential of feature selection by statistical techniques and the use of statistical classifiers in the discrimination of benign from malignant gastric lesions. Oncol Rep (In press).
8. El Hag IA and Kollur SM: Benign follicular thyroid lesions versus follicular variant of papillary carcinoma: differentiation by architectural pattern. Cytopathology 15: 200-205, 2004.
9. Wu HH, Jones JN, Grzybicki DM and Elsheikh TM: Sensitive cytologic criteria for the identification of follicular variant of papillary thyroid carcinoma in fine-needle aspiration biopsy. Diagn Cytopathol 29: 262-266, 2003.
10. Liebeskind A, Sikora AG, Komisar A, Slavit D and Fried K: Rates of malignancy in incidentally discovered thyroid nodules evaluated with sonography and fine-needle aspiration. J Ultrasound Med 24: 629-634, 2005.
11. Ko HM, Jhu IK, Yang SH, Lee JH, Nam JH, Juhng S and Choi C: Clinicopathologic analysis of fine needle aspiration cytology of the thyroid. A review of 1,613 cases and correlation with histopathologic diagnoses. Acta Cytol 47: 727-732, 2003.
12. Galeradavidson H, Bartels PH, Fernandez-Rodriquez A and Dytch HE: Karyometric marker features in fine needle aspirates of invasive follicular carcinoma of the thyroid. Anal Quant Cytol Histol 12: 35-41, 1990.
13. Ferrerroca O, Ballesterguardia E and Martinrodriguez JA: Morphometric, densitometric and flow cytometric criteria for the automated classification of thyroid lesions. Anal Quant Cytol Histol 12: 48-55, 1990.
14. Bibbo M, Bartels PH, Galeradavidson H, Dytch HE and Wied GL: Markers for malignancy in the nuclear texture of histologically normal tissue from patients with thyroid tumors. Anal Quant Cytol Histol 8: 168-176, 1986.
15. Bibbo M, Bartels PH, Salguero M, Dytch HE, Lermapuertas E and Galera-Davidson H: Karyometric marker features in fine needle aspirates of microinvasive follicular carcinoma of the thyroid. Anal Quant Cytol Histol 12: 42-47, 1990.
16. Harms H, Hofman M and Ruschenburg I: Fine needle aspiration of the thyroid – can an image processing system improve differentiation? Anal Quant Cytol Histol 24: 147-153, 2002.
17. Karakitsos P, Cochand-Priollet B, Pouliakis A, Guillausseau PJ and Liossi AI: Learning vector quantizer in the investigation of thyroid lesions. Anal Quant Cytol Histol 21: 201-208, 1999.
18. Karakitsos P, Cochand-Priollet B, Guillausseau P-J and Pouliakis A: Potential of the back propagation neural network in the morphologic examination of thyroid lesions. Anal Quant Cytol Histol 18: 494-500, 1996.