

● *Original Contribution*

DEVELOPMENT OF A SUPPORT VECTOR MACHINE-BASED IMAGE ANALYSIS SYSTEM FOR ASSESSING THE THYROID NODULE MALIGNANCY RISK ON ULTRASOUND

STAVROS TSANTIS,* DIONISIS CAVOURAS,[†] IOANNIS KALATZIS,[†] NIKOS PILIOURAS,[†]
NIKOS DIMITROPOULOS,[‡] and GEORGE NIKIFORIDIS*

*Department of Medical Physics, School of Medicine, University of Patras, Rio Patras, Greece; [†]Department of Medical Instrumentation Technology, Technological Educational Institution of Athens, Athens, Greece; and

[‡]Medical Imaging Department, EUROMEDICA Medical Center, Athens, Greece

(Received 21 April 2005; revised 4 July 2005; in final form 14 July 2005)

Abstract—An SVM-based image analysis system was developed for assessing the malignancy risk of thyroid nodules. Ultrasound images of 120 cytology confirmed thyroid nodules (78 low-risk and 42 high-risk of malignancy) were manually segmented by a physician using a custom developed software in C++. From each nodule, 40 textural features were automatically calculated and were used with the SVM algorithm in the design of the image analysis system. Highest classification accuracy was 96.7%, misdiagnosing two high-risk and two low-risk thyroid nodules. The proposed system may be of value to physicians as a second opinion tool for avoiding unnecessary invasive procedures. (E-mail: tsantis@med.upatras.gr) © 2005 World Federation for Ultrasound in Medicine & Biology.

Key Words: Ultrasound, Thyroid lesion discrimination, Support vectors machine, Classification.

INTRODUCTION

The thyroid gland plays an important role in the control of human metabolism, secreting vital hormones such as thyroxine (T4) and triiodothyronine (T3) (van Herle et al. 1982). Thyroid nodules are swells that appear in the thyroid gland and can be due to growth of thyroid cells or a collection of fluid known as a cyst. They can become large enough to press on nearby structures in the neck, they can overproduce thyroid hormone (hyperthyroidism) or they may be indicative of thyroid cancer (van Herle et al. 1982). Various techniques have been introduced for the detection and evaluation of thyroid nodules, such as physical examination, cytologic examination, scintigraphy, ultrasonography, magnetic resonance and computed tomography (Blum 1990).

High-resolution thyroid ultrasonography (US) is exceptionally sensitive in locating the size and number of thyroid nodules (Stanley 1996). The sonographic findings of the thyroid nodule are often employed as criteria in assessing the risk factor of malignancy and are crucial in patient management, *i.e.*, whether to recommend sur-

gical operation or not (Stanley 1996). Such criteria include echogenicity, absence of halo, calcifications, irregular margins and intra-nodular vascular patterns or spots (Koike et al. 2001). However, estimation of the risk factor involves the subjective evaluation of US images by the physician and, thus, it depends upon the experience of the examiner. Previous US studies (Gimondo et al. 1993; Marquee et al. 2000; McCaffrey 2000; Papini et al. 2002; Peccin et al. 2002; Watters et al. 1992) on thyroid nodules have reported different diagnostic accuracies in predicting malignancy based on the visual analysis of US images. It is evident that a quantitative assessment of the thyroid nodule's risk factor may be of value in avoiding unnecessary invasive intervention.

Previous studies on quantitative methods for estimating the risk associated with thyroid gland disease mainly concern evaluation of parameters from the gray-level histogram of the thyroid gland US image (Hirning et al. 1989; Mailloux et al. 1986), several textural features from gray-tone spatial-dependence matrices (Müller et al. 1989) and the application of discriminant analysis (Hirning et al. 1989; Müller et al. 1989). In those studies, discrimination accuracies between benign and malignant thyroid nodular lesions were 83.9% (Müller et

Address correspondence to: Stavros Tsantis, Dimitriou Xelioti 10, 13341, Ano Liosia, Athens, Greece. E-mail: tsantis@med.upatras.gr

al. 1989) and 85% (Hirning et al. 1989). The fact that echogenicity and the existence of different structures inside the thyroid nodule have been indicated as important factors leading to thyroid malignancy (Gimondo et al. 1993; Papini et al. 2002; Peccin et al. 2002; Watters et al. 1992), combined with the lack of recent quantitative studies in assessing the nature of thyroid nodules, necessitates research to continue employing (1) US thyroid images of modern high resolution scanners and (2) robust computer-based pattern recognition methods using state-of-the-art classification algorithms, for increasing the classification accuracy of objective methods and thus assisting physicians in the preoperative management of patients.

In the present study, a computer-based image analysis system employing the support vector machine (SVM) classifier (Borges 1998; Kecman 2001; Müller et al. 2001; Platt 1999) was developed for the automatic characterization of 120 verified thyroid nodules into two main classes, high-risk and low-risk for malignancy (Tomimori et al. 1999). The aim was the development of an objective second opinion tool for avoiding unnecessary thyroid nodule biopsies. For comparison reasons, the classical quadratic least squares minimum distance (QLSMD) (Ahmed and Rao 1975) and the quadratic Bayesian (QB) (Gonzalez and Woods 1992) classifiers were applied to the same data set. System evaluation was performed by means of the leave-one-out methods (LOO) (Theodoridis and Koutroubas 1999), and highest classification accuracies were determined by means of the exhaustive search method (Theodoridis and Koutroubas 1999). Besides the LOO method, the resubstitution method was also employed (all data were involved in the design and evaluation of the classifier) so as to find the upper and lower bounds of the classification error (Theodoridis and Koutroubas 1999).

MATERIALS AND METHODS

US image data acquisition

The study comprised 120 ultrasonic images displaying thyroid nodules of 120 patients. All US examinations were performed on an HDI-3000 ATL digital ultrasound system (Philips Ultrasound, Bothel, WA, USA) with a wide band (5 to 12 MHz) linear probe using various scanning methods such as longitudinal, transversal and sagittal cross sections of the thyroid gland. The data set was acquired in the time interval from October 2003 to September 2004. During ultrasound examinations the "SmallPartTest" Philips protocol was used. All protocol's settings remained constant throughout that period. Time gain compensation setting had a linear increasing gain compared to the depth. Magnification setting remains at 1:1 except rare cases of very small nodules

(<1.5 cm), which were not included in the present study. Dynamic range setting was relatively high (60 dB) to exploit the high capability of contemporary US systems to visualize US images with the maximum number of gray tones. Thermal and mechanical indexes were set at 0.2 and 0.9, respectively. Multiple focal lengths were set at 1, 2 and 3 cm simultaneously.

Each US image was digitized by connecting the video output of the ultrasound scanner to a Screen Machine II frame grabber using $768 \times 576 \times 8$ image resolution. Under real-time ultrasound guidance, all nodules with size above 1.5 cm underwent fine needle (23-gauge) aspiration biopsy. From various sites of the nodules, 6 to 10 specimens were taken and smears were placed in slides. Those slides were evaluated by two experienced observers. Excluding the rare cases of typical neoplasm of the thyroid, both observers graded the cases in two major categories: epithelial hyperplasia which can be characterized as high risk (42 cases) due to potential malignancy growth, thus leading to repeated ultrasound and cytology examinations of the patient, and in benign lesions (colloid nodules, 78 cases), in which the follow-up examinations can be performed in long time intervals. Retrospectively, the physician (N.D.) analyzed the US characteristics of the corresponding ultrasound images of the two classes given by the cytologists in accordance with Tomimori's grading. The low risk class mostly comprised iso-echoic or hyper-echoic solid nodules with or without cystic change and coarse calcification, while the majority of the high-risk class contained hypo-echoic solid nodules with regular borders or cystic nodules with solid components.

Data preprocessing

The boundary of each nodule was delineated by the physician employing an easy-to-use interactive software program implemented in C++ for the purposes of the present study (Fig. 1). Data processing was performed on a Pentium IV, 2.4 GHz computer. A number of textural features were automatically calculated from the segmented region of interest (ROI) of each thyroid nodule. Textural features are related to the gray-tone structure of the thyroid nodule as depicted on the ultrasound unit, and carry information relevant to the risk factor of malignancy. Four features were computed from the nodule's gray-tone histogram, 26 from the co-occurrence matrix (Haralick et al. 1973) and 10 from the run-length matrix (Galloway 1975).

Classification

For the purpose of the present study, three classifiers were designed, the support vector machine classifier and, for comparison reasons, the quadratic least squares minimum distance and the quadratic Bayesian



Fig. 1. Interface of the custom-made software system designed to read US images, segment by manual interaction thyroid nodules and extract specific features.

classifiers. In the design of a classifier, features play an important role that influences its final discriminatory performance. Ideally, all features at hand (40) should be employed, but since a number of them may be redundant due to mutual correlations (Theodoridis and Koutroubas 1999), an optimum number of them had to be selected to achieve highest classification accuracy. Choosing the best feature combination that will maximize the performance of the classifier is a necessary but time-consuming and computationally demanding procedure. The method we followed (exhaustive search) involved designing the classifier by means of every possible feature-combination (*i.e.*, 2, 3, 4 feature combinations) and all thyroid data available, each time testing the classifier's performance in correctly classifying the thyroid data, and finally selecting that feature combination that demonstrated the highest classification accuracy with the smallest number of textural features. The exhaustive search method was chosen instead of statistical methods such as F-statistics because the latter could result in unreliable error probability estimation (Theodoridis and Koutroubas 1999), because of the small size of the data set.

For the SVM-classifier employing the polynomial kernel of third degree, best feature combination comprised the mean gray-level value of the thyroid ROI's histogram and the sum variance from the co-occurrence matrix (Haralick *et al.* 1973).

The mean value was calculated by eqn 1:

$$\mu_g = \frac{1}{N_x N_y} \sum_{x=1}^{N_x} \sum_{y=1}^{N_y} I(x, y) \quad (1)$$

where $I(x, y)$ is the gray-tone value at coordinates (x, y) of the image matrix of dimensions $N_x \times N_y$.

The sum variance was determined by eqn 2:

$$SVA = \sum_{i=2}^{2N_g} \left[i - \sum_{i=2}^{2N_g} (ip_{x+y}(i)) \right]^2 p_{x+y}(i), \quad (2)$$

where N_g is the number of distinct gray levels in the image and $p_{x+y}(k)$ is given by the following relation:

$$p_{x+y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j), \quad k = 2, 3, \dots, 2N_g.$$

$i+j=k$

where $p(i, j)$ is the normalized entry of the co-occurrence matrix averaged over the four directions (0° , 45° , 90° , 135°) (Haralick *et al.* 1973).

For both the QLSMD and QB classifiers, highest classification accuracies were achieved by the feature combination of the mean gray-level value in eqn 1, the sum variance in eqn 2 and the run length nonuniformity from the run length matrix (Galloway 1972). Run length nonuniformity was determined by eqn 3.

$$\text{RNU} = \frac{\sum_{j=1}^{N_r} \left(\sum_{i=1}^{N_g} p(i, j) \right)^2}{\sum_{j=1}^{N_r} \sum_{i=1}^{N_g} p(i, j)} \quad (3)$$

where N_g is the number of distinct gray levels in the image, N_r is the number of different run lengths and $p(i, j)$ is the normalized entry of the run-length matrix, averaged over the four directions (0° , 45° , 90° , 135°), (Galloway 1972).

Support vector machine classifier

An SVM based classifier (see Appendix) is designed to work for two class problems and can be applied to linearly or nonlinearly separable data, with or without class data overlap (Burgess 1998). In the most difficult case of nonlinearly separable and overlapped data, which is often the case, data are first transformed from the input space to a higher dimensionality feature space, where classes are linearly separable. Then two parallel hyperplanes are determined with maximum distance between them and at the same time with minimum number of training points in the area between them (also called the *margin*). Finally, a third hyperplane through the middle of the margin is defined, which is the decision boundary of the two classes. The discriminant equation of the SVM classifier may thus be defined as in eqn 4:

$$g(x) = \text{sign} \left(\sum_{i=1}^{N_s} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b \right) \quad (4)$$

where α_i are weight parameters, $k(\mathbf{x}_i, \mathbf{x})$ is the kernel function employed for the data transformation into the linearly separable feature space, \mathbf{x}_i are the support vectors (*i.e.*, the training pattern vectors that have their corresponding weights $\alpha_i \neq 0$), N_s is the number of support vectors, \mathbf{x} is the input pattern vector, b is the bias or threshold and $y_i \in \{-1, +1\}$, depending on the class.

In the present work, the SVM classifier was designed employing various polynomial kernels up to the fourth degree and the radial basis function (RBF) kernel. These kernel functions are described by the following relations:

$$k_{\text{POLYNOMIAL}}(x_1, x_2) = ((x_1^T x_2) + 1)^d, \quad d = \text{degree} \quad (5)$$

$$k_{\text{RBF}}(x_1, x_2) = \exp \left(\frac{-\|x_1 - x_2\|^2}{2\sigma^2} \right),$$

$$\sigma = \text{standard deviation} \quad (6)$$

Quadratic least squares minimum distance classifier

The QLSMD classifier (Ahmed and Rao 1975) maps via a nonlinear transformation input the data set into a decision space where each class is clustered around a preselected point. The classification of a given test point is based on its minimum distance from each preselected point. For the QLSMD, the discriminant function for class i and for pattern vector \mathbf{x} is given by:

$$g_i(\mathbf{x}) = \sum_{j=1}^d \alpha_{ij} x_j^2 + \sum_{j=1}^{d-1} \sum_{k=i+1}^d \alpha_{ij} x_j x_k + \sum_{j=1}^d \alpha_{ij} x_j - b \quad (7)$$

where d is the number of features, α_{ij} are weight elements, b is a threshold parameter and x_j are the input vector feature elements.

Quadratic Bayesian classifier

The Bayes decision theory develops a probabilistic approach to pattern recognition, based on the statistical nature of the generated features. The Bayes discriminant function (Gonzalez and Woods 1992) for class i and for pattern vector \mathbf{x} is given by:

$$g_i(\mathbf{x}) = \ln P_i - \frac{1}{2} \ln |\mathbf{C}_i| - \frac{1}{2} [(\mathbf{x} - \mathbf{m}_i)^T \mathbf{C}_i^{-1} (\mathbf{x} - \mathbf{m}_i)] \quad (8)$$

where P_i is the probability of occurrence of class i , \mathbf{m}_i is the mean feature vector of class i and \mathbf{C}_i is the covariance matrix of class i .

System performance evaluation

System evaluation was performed by means of the leave-one-out method. Accordingly, each classifier was designed employing its best feature combination determined in the classification paragraph and by all but one thyroid-ROI feature vector. The latter was presented to the input of the system to be classified as either high-risk

Table 1. Classification accuracies for various SVM kernels using the leave-one-out and re-substitution methods, for the "mean gray value-sum variance" best feature combination

SVM kernel	Classification accuracy		
	LOO ⁺ (%)	Resub.* (%)	N _{sv} **
Polynomial of 1 st degree	89.2	93.3	17
Polynomial of 2 nd degree	91.7	96.7	13
Polynomial of 3 rd degree	96.7	98.3	12
Polynomial of 4 th degree	94.2	99.2	15
RBF	94.2	97.5	15

⁺ Leave-one-out method.

* Re-substitution method.

** Number of support vectors employed using the re-substitution method.

Table 2. Truth table of the SVM classifier employing the 3rd degree polynomial kernel, and the “mean gray value–sum variance” best feature combination

Verified thyroid nodule classes	SVM classification (3 rd Degree polynomial kernel)		LOO ⁺ (resub.*) accuracy
	Low risk	High risk	
Low risk	76 (78)	2 (0)	97.4 (100)%
High risk	2 (2)	40 (40)	95.2 (95.2)%
Overall accuracy			96.7 (98.3)%

⁺ Leave-one-out method.

* Re-substitution method.

or low-risk. The process was repeated, each time leaving a different thyroid-ROI out, until all data had been processed. In this way, each classifier was evaluated by data that were not involved in its design. It is evident, however, that the classifiers had to be redesigned each time a thyroid-ROI was left out. This required a few hours of computer processing time.

RESULTS AND DISCUSSION

We have developed a quantitative method by means of an SVM based software classification system that employed a large number of textural features from US thyroid images for assessing the malignancy risk factor of thyroid nodules.

Table 1 shows the results obtained by the SVM classifier for different kernel functions. Results were

Table 3. Truth table of the QLSMD classifier employing the “mean gray value–sum variance–run length nonuniformity” best feature combination

Verified thyroid nodule classes	QLSMD classification		LOO ⁺ (resub.*) accuracy
	Low risk	High risk	
Low risk	74 (76)	4 (2)	94.9 (97.4)%
High risk	5 (2)	37 (40)	88.1 (95.2)%
Overall accuracy			92.5 (96.7)%

⁺ Leave-one-out method.

* Re-substitution method.

obtained by the leave-one-out method and by the re-substitution method. Maximum classification accuracy in distinguishing low-risk from high-risk thyroid nodules by the LOO method was 96.7% using the polynomial kernel of third degree. It’s worth noting that the maximum classification accuracy corresponds to the minimum number of support vectors involved in the design of the SVM-classifier. The number of support vectors among different kernel functions for the best feature combination ranged between 10% and 14.2% of the number of training points. The small number of support vectors is indicative of the SVM manageable class separability.

Table 2 gives a detailed account of the SVM-third degree polynomial kernel classification accuracies obtained by the LOO and re-substitution methods, employing the mean gray-level value and sum variance features

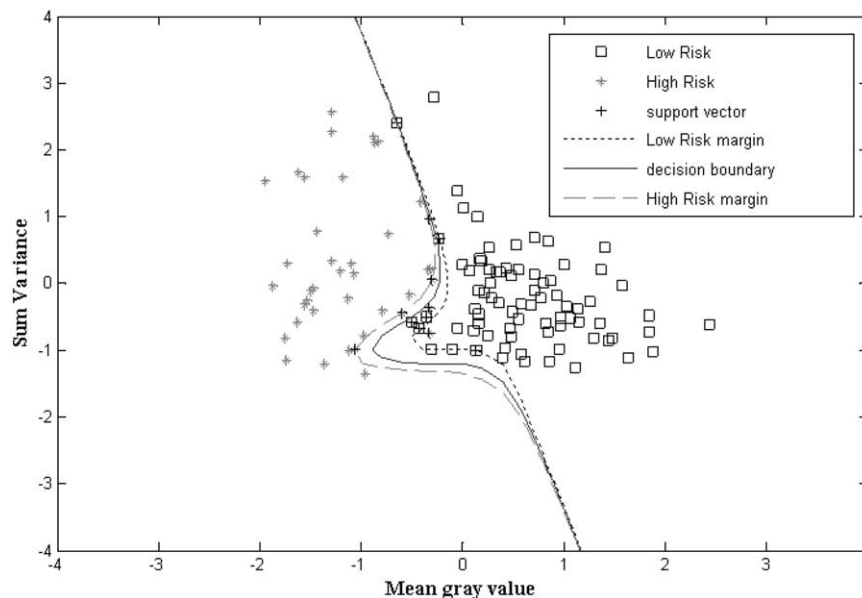


Fig. 2. “Sum variance” vs. “mean gray value” scatter diagram displaying the low-risk and high-risk thyroid nodule data points, the SVM classifier margins and the decision boundary employing the polynomial kernel of third degree.

Table 4. Truth table of the QB classifier employing the “mean gray value–sum variance–run length nonuniformity” best feature combination

Verified thyroid nodule classes	QB classification		LOO ⁺ (resub.)* accuracy
	Low risk	High risk	
Low risk	73 (73)	5 (5)	93.6 (93.6)%
High risk	4 (0)	38 (42)	90.5 (100.0)%
Overall accuracy			92.5 (95.8)%

⁺ Leave-one-out method.
^{*} Re-substitution method.

combination. Seventy-six of the low-risk thyroid nodules were correctly classified while two nodules were incorrectly assigned to the high-risk class, giving a classification accuracy of 97.4% by the LOO method. In the case of the high-risk thyroid nodules, 40 were assigned to the correct class while only two were wrongly classified to the low-risk class, scoring 95.2% class discrimination accuracy. Overall, the SVM achieved 96.7% precision in distinguishing correctly low-risk from high-risk thyroid nodules. Figure 2 shows a scatter diagram of the mean gray-level value against sum variance, the class margins and the decision boundary drawn by the SVM-third degree polynomial kernel classifier. The best features combination employed (mean gray-level value and sum variance) are related to the textural parameters visually evaluated by physicians in assessing the thyroid nodule’s

risk factor (Gimondo et al. 1993; Papini et al. 2002; Peccin et al. 2002; Watters et al. 1992). The mean gray-value is closely associated with the echogenicity of the nodule and the sum variance feature expresses useful spatial information inside the nodule linked to the existence of various structures within the nodule. Features of relative nature (upper 10% gray-level histogram distribution and entropy) have also been indicated in previous quantitative studies (Hirning et al. 1989) to play an important role in thyroid nodule malignancy assessment, scoring an overall of 85% classification accuracy. Similar accuracies (83.9%) were also obtained in another quantitative study employing discriminant analysis of thyroid nodules (Müller et al. 1989). The higher discriminatory precision achieved in the present study was most probably due to the improved resolution of the US images and to the nonlinear nature of the highly sophisticated SVM algorithm employed. This precision dropped when the classical QLSMD and QB classifiers were employed as shown in Tables 3 and 4, respectively.

Table 3 is the truth table giving the classification performance of the QLSMD classifier using the best feature combination found in the classification paragraph (mean gray-level value, sum variance, and run length nonuniformity). Seventy-four of the low-risk and 37 of the high-risk thyroid nodules were correctly classified using the LOO method, resulting in group classification accuracies of 94.9% and 88.1%, respectively, and overall precision of 92.5%. Similarly, Table 4 presents the results obtained by the QB classifier. Although the QB

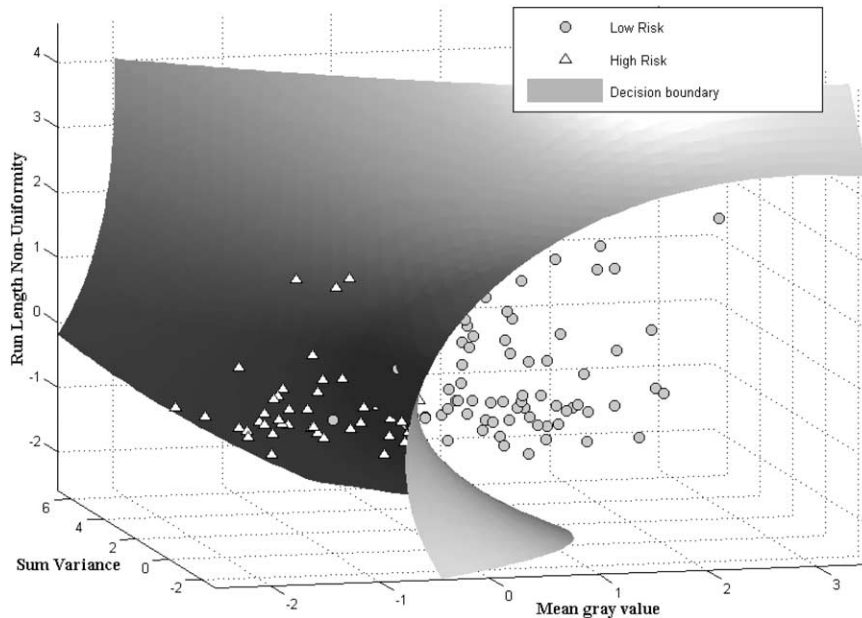


Fig. 3. “Sum variance”, “mean gray value” and “run length nonuniformity” scatter diagram, displaying the low-risk and high-risk thyroid nodule data points and the QLSMD classifier decision boundary.

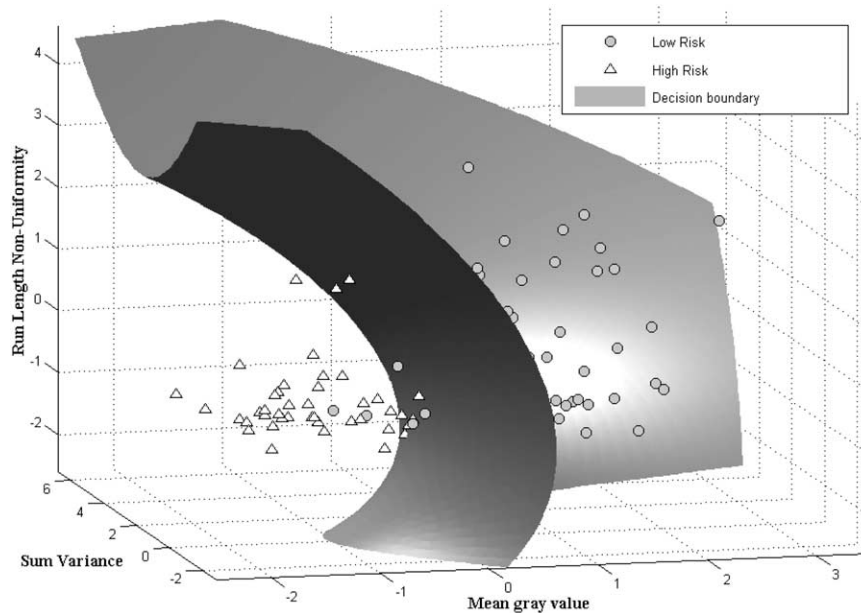


Fig. 4. “Sum variance”, “mean gray value” and “run length nonuniformity” scatter diagram, displaying the low-risk and high-risk thyroid nodule data points and the QB classifier decision boundary.

overall accuracy (92.5%) was similar to that obtained by the QLSMD classifier, the corresponding group accuracies differed to 93.6 and 90.5% for the low-risk and high-risk thyroid nodules, respectively. These differences are insignificant and may be attributed to differences in the nature of the algorithms. The third feature employed (run length nonuniformity) by both classifiers signifies the existence of structures of different sizes within the thyroid nodule, which is related to the optical criteria employed by physicians in assessing the nodule’s risk factor for malignancy (Gimondo *et al.* 1993; Papini *et al.* 2002; Peccin *et al.* 2002; Watters *et al.* 1992). Figures 3 and 4 show 3-D scatter diagrams of the mean gray-level value, sum variance and run length nonuniformity, as well the decision boundaries drawn by the QLSMD and QB classifiers respectively.

Comparing the SVM with the other two classical classifiers, it is evident that the latter had to employ an extra feature to enhance their performance; however, without reaching the SVM’s precision. This is indicative of the effectiveness of the SVM. The penalty, however, that had to be paid for employing the SVM algorithm was much higher processing time during classifier design (training). We have tackled the above problem by suitably distributing computer processing to different workstations and by using the resubstitution method to find well-behaved feature combinations with high classification accuracies and small numbers of support vectors (*i.e.*, leading to separable classes) before system evaluation by the LOO method.

In conclusion, an efficient classification system was designed, based on the SVM algorithm, for assessing the malignancy risk factor of thyroid nodules from their US images. This system could be a useful diagnostic tool by providing a second opinion to the physician and, thus, it may be of value to patient management in avoiding unnecessary invasive procedures.

Acknowledgments—This work was carried out for the computer-based system for the automatic diagnosis of thyroid nodule cancer project, cofunded 75% from the European Union and 25% from the Greek Government under the framework of the Education and Initial Vocational Training Program—Archimedes.

REFERENCES

- Ahmed N, Rao KR. Orthogonal transforms for digital signal processing. New York: Springer-Verlag, 1975:225–258.
- Blum M. Evaluation of thyroid function: Sonography, computed tomography and magnetic resonance imaging in principles and practice of endocrinology and metabolism. 1990:289–293.
- Burges CJS. A tutorial on support vector machines for pattern recognition, data mining and knowledge. Discovery 1998;2:121–167.
- Galloway MM. Texture analysis using gray level run lengths. Comput Graphics Image Processing 1975;4:172–179.
- Gimondo P, Mirk P, Messina G, Pizzi G, Tomei A. The role of ultrasonography in thyroid disease. Minerva Medica 1993;84:671–680.
- Gonzalez RC, Woods RE. Digital image processing. New York: Addison-Wesley, 1992:588–590.
- Haralick RM, Shanmugam K, Dinstein I. Textural features for image analysis. IEEE Trans Syst Man Cybern 1973;6:610–621.
- van Herle AJ, Pick P, Ljung BME, Ashcraft MW, Solomona DH, Keeler EB. The thyroid nodule. Ann Intern Med 1982;96:221–232.
- Hirning T, Zuna I, Schlaps D, *et al.* Quantification and classification of echographics findings in the thyroid gland by computerized B-mode texture analysis. Eur J Radiology 1989;9:244–247.

- Kecman V. Learning and soft computing, support vector machines, neural networks, and fuzzy logic models. Cambridge: MIT Press, 2001:121–191.
- Koike E, Noguchi S, Yamashita H, et al. Ultrasonographic characteristics of thyroid nodules: Prediction of Malignancy. Arch Surgery 2001;136:334–337.
- Mailloux G, Bertrand M, Stampfler R, Ethier S. Computer analysis of echographic textures in Hashimoto disease of the thyroid. J Clin Ultrasound 1986;14:521–527.
- Marquee E, Benson CB, Frates MC, et al. Usefulness of ultrasonography in the management of nodular thyroid disease. Ann Int Med 2000;133:696–700.
- McCaffrey T. Evaluation of the thyroid nodule. Cancer Control 2000;7:223–228.
- Müller MJ, Lorenz D, Zuna I, Lorenz WJ, van Kaick G. The value of computer-assisted sonographic tissue characterization in focal lesions of the thyroid. Der Radiologe 1989;29:132–136.
- Müller K-R, Mika S, Ratsch G, Tsuda K, Schölkopf B. An introduction to kernel-based learning algorithms. IEEE Trans Neural Networks 2001;12:181–202.
- Papini E, Guglielmi R, Bianchini A, et al. Risk of malignancy in nonpalpable thyroid nodules: Predictive value of ultrasound and color-Doppler features. J Clin Endocrinol Metab 2002;87:1941–1946.
- Peccin S, de Castros JAS, Furlanetto TW, Furtado APA, Brasil BA, Czepielewski MA. Ultrasonography: Is it useful in the diagnosis of cancer in thyroid nodules? J Endocrinol Invest 2002;25:39–43.
- Platt JC. Sequential minimal optimization: A fast algorithm for training support vector machines. In: B. Schölkopf B, Burges CJC, Smola AJ, Eds. Advances in kernel methods—support vector learning. Cambridge: MIT Press, 1999:185–208.
- Stanley F. AACE clinical practice guidelines for the diagnosis and management of thyroid nodules. Endocrine Practice 1996;2:78–84.
- Theodoridis S, Koutroubas K. Pattern recognition. New York: Academic Press, 1999.
- Tomimori EK, Camargo RYA, Bisi H, Medeiros-Neto G. Combined ultrasonographic and cytological studies in the diagnosis of thyroid nodules. Biochimie 1999;81:447–452.
- Watters DA, Ahuja AT, Evans RM, et al. Role of ultrasound in the management of thyroid nodules. Am J Surg 1992;164:654–657.

APPENDIX

The support vector machines classifier

A classifier based on support vector machines (SVM) (Kecman 2001; Müller et al. 2001) is a general classifier that it can be applied to linearly as well to nonlinearly separable data, with or without overlap between the classes.

In the most general case of overlapped and nonlinearly separable data, the problem is (a) to transform the training patterns from the input space to a feature space with higher dimensionality ($\mathbf{x} \in \mathbf{R}^d \mapsto \Phi(\mathbf{x}) \in \mathbf{R}^h$) where the classes become linearly separable, and (b) find two parallel hyperplanes with maximum distance between them and at the same time with minimum number of training points in the area between them (also called the *margin*).

The separating hyperplanes in the transformed feature space are defined by eqn 1:

$$\mathbf{w} \cdot \Phi(\mathbf{x}) + b = \pm 1 \quad (1)$$

where +1 is referred to class 1, -1 is referred to class 2, \mathbf{x} if the pattern vector, \mathbf{w} is the normal vector to the hyperplanes, and b the bias or threshold which describes the distance of the decision hyperplane from the origin (that is equal to $b/\|\mathbf{w}\|$).

The discriminant function is given by:

$$g(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \Phi(\mathbf{x}) + b) \quad (2)$$

The parameters \mathbf{w} and b are calculated as follows:

Let N training pattern vectors $\mathbf{x}_i \in \mathbf{R}^d$, $i = 1 \dots N$ (where d is the number of features) belonging to two classes identified by the label

$y_i \in \{-1, +1\}$. The conditions for the hyperplanes may take the following mathematical formulation:

(i) minimize the number of training pattern vectors that lie between the two hyperplanes, so:

$$y_i(\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b) + \xi_i \geq 1 \quad (3)$$

where $\xi_i \geq 0$, $i = 1 \dots N$ are real nonnegative slack-variables.

(ii) the distance between the two hyperplanes (which is equal to $2/\|\mathbf{w}\|$) must be maximized, so $\frac{1}{2}\|\mathbf{w}\|$ must be minimized.

The above conditions lead to minimizing $\frac{1}{2}\|\mathbf{w}\|^2 + \sum C\xi_i$ subject to (3), where C is a positive constant that reflects a trade-off between the classification errors and the size of the margin. Introducing Lagrangian multipliers α_i, β_i , $i = 1 \dots N$, the Lagrangian is given by:

$$L_P = \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i(y_i(\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b) + \xi_i - 1) - \sum_{i=1}^N \beta_i \xi_i \quad (4)$$

The problem is now to maximize L_P subject to $\frac{\partial L_P}{\partial b} = 0$, $\frac{\partial L_P}{\partial \mathbf{w}} = 0$ and $\frac{\partial L_P}{\partial \xi_i} = 0$ (with $\alpha_i, \beta_i \geq 0$). These constraints give respectively:

$$\sum_{i=1}^N \alpha_i y_i = 0 \quad (5)$$

$$\mathbf{w} = \sum_{i=1}^N \alpha_i y_i \Phi(\mathbf{x}_i) \quad (6)$$

and

$$C = \alpha_i + \beta_i \quad (7)$$

The equation 7, in combining with $\alpha_i, \beta_i \geq 0$, results that $0 \leq \alpha_i \leq C$.

Substituting eqns 5, 6 and 7 in relation with eqn 4, we take the dual variables Lagrangian L_D :

$$L_D = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \quad (8)$$

By use of a *kernel function*, that it can replace the inner product $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ in the higher dimensional feature space, the dual Lagrangian L_D can take the form of (9):

$$L_D = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \quad (9)$$

A function can be used as a kernel function if it satisfies the following *Mercer's condition*:

Any symmetric function $k(\mathbf{x}, \mathbf{y})$ in the input space is equivalent to an inner product in the feature space, if $\iint k(\mathbf{x}, \mathbf{y}) g(\mathbf{x}) g(\mathbf{y}) d\mathbf{x} d\mathbf{y} \geq 0$, for any function $g(\mathbf{x})$ for which $\int g^2(\mathbf{x}) d\mathbf{x} < \infty$.

Using eqns 2, 6 and 9, it may be seen that ξ_i and β_i have vanished, so the discriminant function of the SVM classifier may be written as:

$$g(\mathbf{x}) = \text{sign} \left(\sum_{i=1}^{N_s} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b \right) \quad (10)$$

where N_s is the number of pattern vectors (also called the *support vectors*) with nonzero α_i 's.

Combining the eqns 1, 6 and 9, the threshold b may be found as:

$$b = \frac{1}{N_S} \sum_{j=1}^{N_S} \left(y_j - \sum_{i=1}^N \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}_j) \right) \quad (11)$$

and the coefficients α_i are obtained by solving the “dual” problem, which is maximization of L_D (eqn 9) subject to eqn 5, with $0 \leq \alpha_i \leq C$.

Functions that are commonly used as kernels are:

i) The linear kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \mathbf{x}_j \quad (12a)$$

ii) The polynomial kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \mathbf{x}_j + \theta)^d \quad (12b)$$

where d is the degree of the polynomial and θ an offset parameter,

iii) The Gaussian radial basis kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j)}{2\sigma^2}\right) \quad (12c)$$

where σ is the standard deviation.

iv) The sigmoidal kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\kappa(\mathbf{x}_i \mathbf{x}_j) + \theta) \quad (12d)$$

where κ the gain and θ the offset.

v) The inverse multiquadric kernel:

$$k(\mathbf{x}, \mathbf{y}) = ((\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j) + c^2)^{-1/2} \quad (12e)$$

where c a nonnegative real number.