Computer-Derived Nuclear Features Distinguish Malignant From Benign Breast Cytology

WILLIAM H. WOLBERG, MD, W. NICK STREET, MS, DENNIS M. HEISEY, PhD, AND OLVI L. MANGASARIAN, PhD

This article describes the use of computer-based analytical techniques to define nuclear size, shape, and texture features. These features are then used to distinguish between benign and malignant breast cytology. The benign and malignant cell samples used in this study were obtained by fine needle aspiration (FNA) from a consecutive series of 569 patients: 212 with cancer and 357 with fibrocystic breast masses. Regions of FNA preparations to be analyzed were converted by a video camera to computer files that were displayed on a computer monitor. Nuclei to be analyzed were roughly outlined by an operator using a mouse. Next, the computer generated a "snake" that precisely enclosed each designated nucleus. The computer calculated 10 features for each nucleus. The ability to correctly classify samples as benign or malignant on the basis of these features was determined by inductive machine learning and logistic regression. Cross-validation was used to test the validity of the predicted diagnosis. The logistic regression cross validated classification accuracy was 96.2% and the inductive machine learning cross-validated classification accuracy was 97.5%. Our computerized system provides a probability that a sample is malignant. Should this probability fall between 30% and 70%, the sample is considered "suspicious," in the same way a visually graded FNA may be termed suspicious. All of the 128 consecutive cases obtained since the introduction of this system were correctly diagnosed, but nine benign aspirates fell into the suspicious category. Fifty-seven FNAs were obtained that had been visually diagnosed elsewhere by others as "suspicious." Eleven (19.3%) were similarly classified as suspicious by the computer, but 84.8% of the remaining samples were correctly diagnosed. The methods described in this article will provide the basis for computerized systems to diagnose breast cytology. HUM PATHOL 26:792–796. Copyright © 1995 by W.B. Saunders Company

Key words: **breast cancer, image processing, cytology, diagnosis, inductive machine learning.**

Abbreviations: **FNA, fine needle aspiration; MSM-T, Multisurface Method–Tree.**

Progress over the past 30 years in computer analysis of microscope images has made possible highly accurate quantitative and objective feature assessment for diagnostic decision making.[1] We have identified computer-derived, quantitative digital features that accurately classify breast epithelial cells as benign or malignant. Classification is accomplished by inductive machine learning by a computer program based on the accumulated experience from previously diagnosed cases. The program can then be used to generalize to diagnose new cases that may differ from those previously encountered.

MATERIALS AND METHODS

Patients and Aspirate

The benign and malignant cell samples used in this study were obtained by fine needle aspiration (FNA) from a consecutive series of 569 patients: 212 with cancer and 357 with fibrocystic breast masses.

To prepare an FNA, a small drop of viscous fluid is aspirated from breast masses by making multiple passes with a 23-gauge needle while negative pressure is being applied to an attached syringe. The aspirated material is expressed onto a silane-prepared glass slide and the aspirate is spread when a similar slide is applied face to face, and the slides are separated with a horizontal motion. Preparations are immediately fixed in 95% ethanol, stained with hematoxylin-eosin, and processed.

In our analysis, aspirates were classified as cancer based on surgical biopsy and histological confirmation. No data concerning breast histology was available in five patients who did not have definitive breast surgery because extensive distant metastases were present at diagnosis and in one patient who had definitive surgery elsewhere. Among the remaining 206 patients, 10 had in situ and 196 had infiltrating cancers: 148 ductal, 18 not otherwise specified, 12 lobular, four medullary, three tubular, three comedo, two mucinous, and one each of signet ring, cribriform, sebaceous, papillary, inflammatory, and pseudosarcomatous carcinoma. Cytologically diagnosed benign breast masses were confirmed either by biopsy or by follow-up for a year. Fifty-one cytologically benign lesions were surgically excised: 35 fibroadenomas, 12 fibrocystic disease, and one each of fat necrosis, stromal fibrosis, adenolipoma, and atypical lobular hyperplasia. After a year, nonbiopsied masses were considered to be benign if they had not enlarged.[2]

One hundred twenty-eight consecutive specimens (94 benign and 34 malignant), obtained at this institution since the introduction of the system and 57 specimens visually diagnosed as suspicious at another institution were obtained and diagnosed with the trained algorithm.

Image Preparation

The imaged area on the aspirate slides is visually selected for minimal nuclear overlap. Areas of apocrine metaplasia are avoided. The image for digital analysis is generated by a JVC TK-1070U (JVC, Elmwood Park, IL) color video camera mounted atop an Olympus (Lake Success, NY) microscope

From the Departments of Surgery, Human Oncology, and the Computer Sciences Department, University of Wisconsin, Madison, WI. Accepted for publication November 13, 1994.

Supported in part by Air Force Office of Scientific Research Grant AFOSR F49620-94-1-0036 and National Science Foundation Grant CCR-9322479.

Address correspondence and reprint requests to William H. Wolberg, MD, Department of Surgery, University of Wisconsin Clinical Sciences Center, 600 Highland Ave, Madison, WI 53792.

0046-8177/95/2607-0014$5.00/0

and the image is projected into the camera with a 63 X objective and a 2.5 X ocular. The image is captured by a Computer-Eyes/RT color framegrabber board (Digital Vision, Inc, Dedham MA) as a 640 × 400, 8-bit-per-pixel Targa file.

User Interface (Xcyt)

The first step in successfully analyzing the digital image is to specify the exact location of each cell nucleus. A computer graphics program called Xcyt was developed that allows the user to input the approximate location of a sufficient number of nuclei (10 to 20) to provide a representative sample. The program was developed using the X Window System and the Athena WidgetSet on a DECstation 3100 (Digital Equipment Corporation, Nashua, NH).

A mouse is used to trace a rough outline of cell nuclei on the computer monitor. From this rough outline, the actual boundary of the cell nucleus is located by an active contour model known as a "snake."[3,4] The "snake" is a deformable spline that seeks to minimize an energy function defined over the arc length of a curve. The energy function is defined in such a way that the snake, in the form of a closed curve, conforms itself to the boundary of a cell nucleus. The mathematic aspects of the snake calculations are described elsewhere.[5]

Nuclear Features

Once the nuclei to be analyzed have been identified by the operator and have been enclosed by the computer-generated snakes, the computer calculates 10 nuclear features for each nucleus.[5] These features are modeled such that higher values are typically associated with malignancy. Features were verified using idealized phantom cells.[6] Nuclear size is expressed by the radius and area features. Nuclear shape is expressed by smoothness, concavity, compactness, concave points, symmetry, and fractal dimension features. Both size and shape are expressed by the perimeter feature.[6] Nuclear texture is measured by finding the variance of the gray scale intensities in the component pixels. The mean value, worst (mean of the three largest values), and standard error of each feature are computed for each image, resulting in a total of 30 features.

Classification Procedure and Cross-Validation

In both our inductive machine learning and logistic regression analyses, the accuracy of correctly classifying samples as benign or malignant on the basis of digitally determined nuclear features is determined by 10-fold cross-validation.[7] The resulting estimate is unbiased and accurate in cases such as ours that have a large number of training samples. Initially, the data set is divided into 10 randomly selected, equal parts. One part is removed. A classifying algorithm is created from the nine remaining parts, and the accuracy of the classifier is tested on the 10th part. The 10th part is then returned, and the process is repeated until all parts have been tested.

Classifying Algorithm Developed by Inductive Machine Learning

Image processing produces a database consisting of one 30-dimensional point for each sample. The 30 dimensions consist of the mean, standard error, and worst for the 10 features. The classification procedure becomes one of pattern separation, specifically, that of determining how points can best be separated into benign and malignant sets. The classification procedure is a variant on the Multisurface Method,[8,9] known as Multisurface Method–Tree (MSM-T).[10,11] This method uses linear programming iteratively to place a series of separating planes in the feature space of the samples. If the benign and malignant sets can be separated by a single plane, the first plane will be placed between them. If the sets are not linearly separable, MSM-T constructs another plane that minimizes an average distance of misclassified points. Depending on the separation accuracy attained, the procedure is recursively repeated on the two regions generated by each plane until satisfactory separation is achieved (ie, each of the final regions contains mostly points of one category). The classifier thus obtained is then used as a decision tree to categorize new cases. MSM-T is similar to other decision tree methods such as CART[12] and C4.5,[13] but has been shown to be faster and more accurate on several real-world data sets.[10]

Generally, simpler classifiers perform better on new data than do more complex ones. Therefore, we minimize not only the number of separating planes but also the number of features used in constructing the planes.

Our computerized system provides a benign or malignant diagnosis together with a probability of malignancy determined by the distance the new point lies from the separating plane.[6] Should this probability fall between 30% and 70%, the sample is considered "suspicious," in the same way a visually graded FNA may be termed suspicious.

Classifying Algorithm Developed by Logistic Regression Analysis

The classifying algorithm developed by logistic regression analysis was performed with SAS[14] software (SAS Institute, Cary, NC). Other statistical analyses and graphics were performed with Systat[15,16] software (Evanston, IL).

RESULTS

Feature Analysis

Diagnostic features for 357 benign and 212 malignant samples are noted in Table 1. Values for area are expressed as square micra (μm^2), and for radius and perimeter as micra (μm). Values for remaining features are dimensionless. Independent samples *t*-test was not significant for differences between benign and malignant for mean of fractal dimension, standard error of texture, standard error of smoothness, standard error of symmetry, and for standard error of fractal dimension. The differences were $P < .001$ for all other features.

Logistic Regression Classification

A stepwise logistic regression selection process selected a model consisting of the variables standard error of the radius, worst radius, worst texture, and worst concave point. Standard error of the radius seemed to contribute little to the predictive ability of the model (as judged by sensitivity and specificity) and consequently was dropped. Nine benign and 12.4 malignant FNAs were misclassified when this model classified with 10-fold cross-validation repeated 100 times (Table 2).

TABLE 1. Diagnostic Features

	Benign ($N = 357$)	Malignant ($N = 212$)
Mean Radius (μm)	3.471 ± 0.508	5.004 ± 0.922
Mean area (μm^2)	37.80 ± 10.96	80.30 ± 30.36
Mean Perimeter (μm)	22.32 ± 3.372	33.06 ± 6.275
Mean Texture	17.92 ± 3.993	21.60 ± 3.786
Mean Smoothness	0.093 ± 0.013	0.103 ± 0.013
Mean Compactness	0.080 ± 0.034	0.146 ± 0.053
Mean Concavity	0.046 ± 0.044	0.163 ± 0.075
Mean Concave points	0.026 ± 0.016	0.089 ± 0.034
Mean Symmetry	0.174 ± 0.025	0.194 ± 0.028
Mean Fractal dimension*	0.063 ± 0.007	0.063 ± 0.008
SE Radius (μm)	0.081 ± 0.032	0.175 ± 0.100
SE Area (μm^2)	1.730 ± 0.723	6.008 ± 5.171
SE Perimeter (μm)	0.570 ± 0.220	1.247 ± 0.740
SE Texture*	1.222 ± 0.031	1.213 ± 0.483
SE Smoothness*	0.007 ± 0.000	0.007 ± 0.003
SE Compactness	0.022 ± 0.001	0.033 ± 0.018
SE Concavity	0.026 ± 0.002	0.042 ± 0.022
SE Concave points	0.010 ± 0.000	0.015 ± 0.006
SE Symmetry*	0.021 ± 0.000	0.021 ± 0.011
SE Fractal dimension*	0.004 ± 0.000	0.004 ± 0.002
Worst Radius (μm)	3.825 ± 0.567	6.049 ± 1.245
Worst Area (μm^2)	45.67 ± 13.37	116.7 ± 50.37
Worst Perimeter (μm)	24.88 ± 3.87	40.49 ± 8.488
Worst Texture	23.53 ± 5.487	29.24 ± 5.480
Worst Smoothness	0.125 ± 0.020	0.145 ± 0.022
Worst Compactness	0.184 ± 0.094	0.375 ± 0.170
Worst Concavity	0.167 ± 0.141	0.454 ± 0.182
Worst Concave point	0.075 ± 0.036	0.183 ± 0.046
Worst Symmetry	0.271 ± 0.042	0.325 ± 0.076
Worst Fractal dimension	0.080 ± 0.014	0.092 ± 0.021

Abbreviation: SE, standard error.

NOTE. Values for area are expressed as square micra (μm^2), and for radius and perimeter as micra (μm). Values for remaining features are nondimensional. Independent samples *t*-test was not significant (indicated by * in the table) for differences between benign and malignant for mean of fractal dimension, standard error of texture, standard error of smoothness, standard error of symmetry, and for standard error of fractal dimension. The differences were $P < .001$ for all other features.

Inductive Machine Learning Classification

A computationally intensive search showed that worst Area, mean Texture and worst Smoothness gave the most accurate three-feature, single-plane classification separation. The location of these points in the three-dimensional space defined by the three classifying features is shown in Fig 1. The entire 10-fold cross validation process was done five times and the results reported occurred three of the five times. Seven benign and seven malignant FNAs were misclassified when MSM-T was used to classify with cross validation (Table 3).

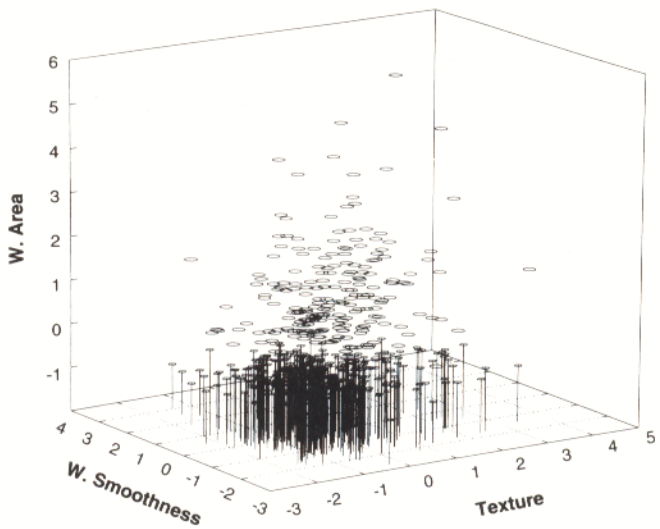

FIGURE 1. The location of 212 malignant ■ and 357 benign ○ samples in the three-dimensional space defined by the computer-generated features of mean Texture (Texture, x axis), worst Smoothness (W. Smoothness, y axis), and worst Area (W. Area, z axis). The feature values are standardized and the symbol sizes vary according to the perspective. MSM-T generates the plane that provides the best separation, in this case with 7 benign points lying on the side with 205 malignant points and 7 malignant points lying on the side with 350 benign points.

Prospective Analyses

All of the 128 consecutive cases obtained since the introduction of this system were correctly diagnosed, but nine benign aspirates were categorized as suspicious. Eleven (19.3%) of the 57 FNAs that were visually diagnosed elsewhere as "suspicious" were similarly classified by the computer, but 84.8% of the remaining samples were correctly diagnosed.[17]

DISCUSSION

Nuclear feature analysis is better performed on cytological FNA preparations than on the more commonly used histological tissue samples. FNA cells are preserved intact, whereas histological processing cuts cells at various planes. Selection of nuclei for analysis, as performed in our study, has been shown to be robust and operator independent.[18] A wide variety of histological tumor types was included in this study, not only "high grade" infiltrating ductal cancers that would have made classification easy. Our pathologists do not

TABLE 2. Logistic Regression Classification: Predicted Versus Confirmed Diagnosis

	Malignant, Confirmed	Benign, Confirmed	Total
Malignant, predicted	199.6 ± 0.65	9.0 ± 0.72	208
Benign, predicted	12.4 ± 0.65	348.0 ± 0.72	361
Total	212	357	569

TABLE 3. Inductive Machine Learning Classification: Predicted Versus Confirmed Diagnosis

	Malignant, Confirmed	Benign, Confirmed	Total
Malignant, predicted	205	7	212
Benign, predicted	7	350	357
Total	212	357	569

grade cancers because of difficulties in interobserver variability,[19] but the large standard error of nuclear Area across samples (Table 1) indicates that an assortment of tumor grades[20] were included in the study.

We compared our feature values with those reported in the literature. Morphologic alterations result from any fixation process and features, such as nuclear texture vary according to the nuclear stain used (eg, Feulgen versus hematoxylin). Therefore, feature values presented here pertain only to FNAs processed in the same manner as described herein. Separate algorithms will have to be developed if deviations are made in preparing or processing the FNAs (eg, air drying). Nevertheless, the perimeter measurements for our benign and malignant nuclei were the same as those reported by Hutchison et al.[21] They found the mean nuclear perimeter ± standard deviation for 169 benign samples to be 23.99 ± 2.97 micra compared with 22.32 ± 3.38 micra for our 357 benign samples. Similarly, their values for the perimeter of 168 malignant samples was 35.29 ± 5.98 micra compared with 33.06 ± 6.28 micra for our 212 malignant samples. Their values for the nuclear area of benign samples was 36.43 ± 8.89 square micra and ours was 37.80 ± 10.96, whereas their malignant sample area was 76.38 ± 23.03 and ours was 80.30 ± 30.36. Therefore, the size values for both benign and malignant samples seem to be the same despite methodological differences. However, the methodology used seems to make a difference in shape. "Shape" was defined by Hutchison et al[21] as perimeter$^2/4\pi$ area, which we termed compactness. They found the mean nuclear "shape" and standard deviation for 169 benign samples to be 1.28 ± 0.12 versus 1.071 ± 0.032 for our 357 benign samples. Similarly, their values for the "shape" of 168 malignant samples was 1.33 ± 0.15 versus 1.120 ± 0.049 for our 212 malignant samples. The difference between our compactness and the nuclear "shape" of Hutchison et al[21] is significant ($P < .001$) for both benign and malignant samples. Although the absolute differences are small, the statistical significance develops from the small variability in the measurements: 3% and 4%, respectively, for benign and malignant in our series, and 9% and 11% for that of Hutchison et al.[21]

A size, a texture, and a shape feature were selected to give the most accurate classification by both the logistic regression and by the inductive machine learning process. Worst Radius, worst Texture, and worst Concave points were used by the logistic regression model and worst Area, mean Texture, and worst Smoothness were used by the inductive machine learning model. The number of classifying features was kept to a minimum to avoid data overfitting. Pearson's correlation coefficient was so strong ($r = .983$) between worst Radius and worst area, and between worst Texture and mean Texture ($r = .912$) that these feature values seem virtually interchangeable. However, the correlation was weaker ($r = .546$) between worst Concave points and worst Smoothness. Classification based on digital feature analysis is robust; similar classification accuracy was obtained by logistic regression and inductive machine learning classification. The logistic regression cross validated classification accuracy is 96.2% and the inductive machine learning cross-validated classification accuracy is 97.5%. These results are considerably better than the 89% accuracy based on individual cell analysis achieved by Hutchison et al.[21] Our improved accuracy apparently was achieved through the use of shape and texture features not measured by Hutchison et al.[21]

The logistic regression classification sensitivity is 94.1% ± 0.3%, and the specificity is 97.5% ± 0.2%. The inductive machine learning classification sensitivity is 96.7%, and the specificity is 98.0%. These performance parameters were determined by cross-validation to test the validity of the predicted diagnosis. With our comprehensive assessment of shape features, we achieve better performance parameters than have previously been reported.[21] In fact, our performance parameters rival those obtained by visual diagnosis.[22] Giard and Hermans[22] emphasized the need for developing individual performance characteristics for persons doing FNA of breast masses because the accuracy achieved is operator dependent. The reported accuracy of visually diagnosed breast FNAs is more than 90%. The overall accuracy was 94.3% in a 62-series study with a total of 23,741 satisfactory breast FNAs.[6] Individually, the mean sensitivity for these series was 0.91 ± 0.07, and the mean specificity was 0.87 ± 0.18. The relatively high standard deviations indicate that the accuracy achieved in individual series varies considerably and reflects the subjectivity of the procedure. Additionally, it is unlikely that such accuracy is generally achieved because publication bias is toward publishing favorable results. The variation in accuracy is largely caused by the subjectivity that is inherent in visual interpretation. Visually assessed size, shape, and texture features that distinguish benign from malignant cells are now measured by computers. We believe that our methods provide a basis for highly accurate computerized diagnostic systems for breast cytology.

REFERENCES

1. Wied G, Bartels P, Bibbo M, et al: Image analysis in quantitative cytopathology and histopathology. HUMAN PATHOL 20:549-571, 1989

2. Layfield L, Chrischilles E, Cohen M, et al: The palpable breast nodule: A cost effective analysis of alternate diagnostic approaches. Cancer 72:1642-1651, 1993

3. Kass M, Witkin A, Terzopoulos D: Snakes: Active contour models. Proceedings of the First International Conference on Computer Vision. 1987, pp 259-269

4. Williams DJ, Shah M: A fast algorithm for active contours. In Proceedings of the Third International Conference on Computer Vision. Osaka, Japan, 1990, pp 592-595

5. Street WN, Wolberg WH, Mangasarian OL: Nuclear feature extraction for breast tumor diagnosis. Proceedings of the IS&T/SPIE International Symposium on Electronic Imaging 1905:861-870, 1993

6. Wolberg WH, Street WN, Mangasarian OL: Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. Cancer Lett 77:163-171, 1994

7. Stone M: Cross-validatory choice and assessment of statistical predictions. J R Stat Soc 36:111-147, 1974

8. Mangasarian OL: Multi-surface method of pattern separation. IEEE Trans Inform Theory IT-14:801-807, 1968

9. Mangasarian OL, Setiono R, Wolberg WH: Pattern recogni-

tion via linear programming: Theory and application to medical diagnosis, in Coleman TF, Li Y (eds): Large-Scale Numerical Optimization. Philadelphia, PA, SIAM, 1990, pp 22-30

10. Bennett KP: Decision tree construction via linear programming, in Evans M (ed): Proceedings of the Fourth Midwest Artificial Intelligence and Cognitive Science Society Conference, 1992, pp 97-101

11. Mangasarian OL: Mathematical programming in neural networks. ORSA J Computing 5:349-360, 1993

12. Breiman L, Friedman J, Olshen R, et al: Classification and Regression Trees. Pacific Grove, CA, Wadsworth, 1984

13. Quinlan JR: C4.5: Programs for Machine Learning. San Mateo, CA, Morgan Kaufmann, 1993

14. SAS/STAT User's Guide, Version 6 (ed 4). Cary, NC, SAS Institute, 1989

15. Wilkinson L, Hill MA, Welna JP, et al: SYSTAT for Windows: Statistics (ed 5). Evanston, IL, SYSTAT, 1992

16. Wilkinson L, Hill MA, Miceli S, et al: SYSTAT for Windows: Graphics (ed 5). Evanston, IL, SYSTAT, 1992

17. Wolberg WH, Teague MW, Street WN, et al: Computers improve the certainty of breast mass diagnosis by fine needle aspiration (FNA). Presented at the Society for Surgical Oncology meeting, Boston, MA, March 1995

18. Wolberg WH, Street WN, Mangasarian OL: Breast cytology diagnosis with digital image analysis. Anal Quant Cytol Histol 15:396-404, 1993

19. Gilchrist KW, Kalish L, Gould VE, et al: Interobserver reproducibility of histopathological features of stage II breast cancer. Breast Cancer Res Treat 5:3-10, 1985

20. VanDiest PJ, Risse EKJ, Schipper NW, et al: Comparison of light microscopic grading and morphometric features in cytological breast cancer specimens. Pathol Res Pract 185:612-616, 1989

21. Hutchinson ML, Isenstein LM, Zahniser DJ: High-resolution and contextual analysis for the diagnosis of fine needle aspirates of breast. Anal Quant Cytol Histol 13:351-355, 1991

22. Giard RWM, Hermans J: The value of aspiration cytologic examination of the breast. A statistical review of the medical literature. Cancer 69:2104-2110, 1992