Computers and Biomedical Research **33,** 296–313 (2000)
doi:10.1006/cbmr.2000.1546, available online at http://www.idealibrary.com on IDEAL®

The Reliability Issue of Computer-Aided Breast Cancer Diagnosis[1]

Boris Kovalerchuk,* Evangelos Triantaphyllou,†,[2] James F. Ruiz,‡ Vetle I. Torvik,† and Evgeni Vityaev§

**Department of Computer Science, 400E 8th Avenue, Central Washington University, Ellensburg, Washington 98926; †Department of Industrial Engineering, 3128 CEBA Building, Louisiana State University, Baton Rouge, Louisiana 70803-6409; ‡Department of Radiology, Woman's Hospital, Baton Rouge, Louisiana 70895; and §Department of Mathematics, Novosibirsk University, Novosibirsk, Russia*

Received January 12, 2000

This paper introduces a number of reliability criteria for computer-aided diagnostic systems for breast cancer. These criteria are then used to analyze some published neural network systems. It is also shown that the property of monotonicity for the data is rather natural in this medical domain, and it has the potential to significantly improve the reliability of breast cancer diagnosis while maintaining a general representation power. A central part of this paper is devoted to the representation/narrow vicinity hypothesis, upon which existing computer-aided diagnostic methods heavily rely. The paper also develops a framework for determining the validity of this hypothesis. The same framework can be used to construct a diagnostic procedure with improved reliability.

Key Words: neural networks; machine learning; discriminant analysis; data monotonicity; computer-aided diagnostic systems; reliability of diagnosis; representation/narrow vicinity hypothesis.

1. INTRODUCTION

Breast cancer is the most common cancer in women in the United States, with an estimated 182,000 cases in 1995 (*1*). The most effective tool in the battle against breast cancer is screening mammography. However, several retrospective analyses have found diagnostic error rates ranging from 20 to 43% (*2, 3*). Also, of the

[1] The first two authors gratefully recognize the partial support from the Office of Naval Research (ONR) Grant N00014-95-1-0639. The first, second, and fifth authors also recognize the partial support from the CAST and COBASE programs at the National Research Council (NRC). The second and fourth authors recognize the partial support from ONR Grant N0001-97-1-0632.

[2] To whom correspondence should be addressed. Web: http://www.imse.lsu.edu/vangelis.

0010-4809/00 $35.00

breast biopsies performed due to suspicious mammograms, 70–89% will be found benign (*4*). Elmore *et al.* in (*5*), studied the variability of radiologists' interpretation of a set of mammograms. They observed an average intraobserver variability of approximately 8% in addition to a 19% interobserver variability for the diagnosis of cancer, for which the variability in management recommendations was 25%. They also found that 9 out of 10 radiologists recognized fewer than 3% of the mammograms which they screened 5 months prior, while 1 out of 10 claimed to have recognized about 25% of the cases (*5*). These startling statistics and other discussions on computer-aided diagnosis (CAD) (*6–8*) clearly demonstrate the need for (and the possible magnitude of) improvements in the reliability of breast cancer diagnosis.

Today, with the proliferation of powerful computer technology, a great effort is directed toward developing computerized methods that can assist radiologists in breast cancer diagnosis. Currently, such methods include neural networks, nearest neighbor methods, discriminant analysis, cluster analysis, decision trees, and linear programming based methods (see, for instance, (*9–16*)). These methods extract general rules which are based on a sample of specific cases. Thus, the better the available data represent the underlying rules, the more accurate the predictions based on the extracted rules can become. Therefore, these methods rely on obtaining representative samples.

Often the available training data are insufficient to achieve a desirable prediction accuracy. In other words, the available knowledge is often insufficient to make confident recommendations. According to Jonson (*17*), the use of Bayesian models in medical diagnosis can be controversial, if not unethical, because the fundamental requirement of strict randomness rarely occurs, and it can rarely be tested with the available training data. This critical issue is elaborated on in Sections 2 and 3 of this paper.

Monotonicity of the data is a frequent property that has not been adequately utilized by traditional approaches. This property has the potential to significantly improve the reliability of breast cancer diagnosis. The monotonicity approach described in this paper does not assume a particular model and in this sense maintains a general representation power. However, it should be stated at this point that if the existence of an appropriate parametric model (as described by Duda and Hart in (*18*)) can be established, then its application may lead to a higher degree of confidence than the use of the monotonicity approach described in this paper.

This paper is organized as follows. In Section 2 we introduce some reliability criteria of computer-aided breast cancer diagnosis. The same section also uses these criteria to analyze the reliability of some published diagnostic results which are based on neural networks. Section 3 is devoted to the representation/narrow vicinity hypothesis. Section 4 presents the results of the validation of this hypothesis on 11 mammographic and related clinical features. The last section summarizes the main results of this study and formulates some directions for future research.

2. RELIABILITY CRITERIA

The validity and accuracy (reliability) of a computer-aided diagnostic system should be reasonably high for clinical applications. To explore the issue of reliability, assume that we have the 11 binary (0 or 1 value) diagnostic features described in Appendix I. This example is a rather simple one since it assumes only 11 features which are binary valued and that the entire setting is deterministic (i.e., a given case always belongs to the same class). However, this illustrative setting is still sufficient to provide the main motivation of the key concepts described in this paper. By using the previous 11 features, each medical case can be expressed as a combination of binary values defined on these features. For example, the ordered sequence (01111100011) describes the case with "0" value for the 1st, 7th, 8th and 9th features and with "1" value for the rest of them. Furthermore, by considering the definitions in Appendix I, the above binary vector means that:

The number of calcifications/cm^2 is small (value 0);
the volume (in cm^3) is small (value 1);
the total number of calcifications is large (value 1);
the irregularity in the shape of individual calcifications is marked (value 1);
the variation in the shape of calcifications is marked (value 1);
the variation in size of the calcifications is marked (value 1);
the variation in density of the calcifications is mild (value 0);
the density of the calcifications is mild (value 0);
ductal orientation is not present (value 0);
comparison with previous exam is "pro cancer/biopsy" (value 1);
associated findings are "pro cancer/biopsy" (value 1).

Please note that the grade "small" was deliberately coded differently for the number of calcifications/cm^2 and the volume (in cm^3). This step allowed us to take advantage of the monotonicity property (as described later), which is of critical importance to the effectiveness of our method.

Next, a computer-aided diagnostic system which is based on the previous 11 binary features should be able to categorize new cases represented by binary vectors. Each such case is assumed to be either in the "highly suspicious for malignancy" class or in the "not highly suspicious for malignancy" class, and only one of them. That is, in mathematical terms a computer-aided diagnostic (CAD) system operates as a discriminant function, say $f(x_1, x_2, \ldots, x_n)$, which is defined in the space of n features denoted as $x_1, x_2, \ldots, x_n$. In order to help fix ideas, assume that a discriminant function for the current illustrative example was constructed from a sample of 80 training cases (each one of which is either highly suspicious for malignancy or not highly suspicious for malignancy). It should be noted here that many, if not the majority, of published studies consider sample sizes of about 80 cases each (*7*). Next, suppose that the function f discriminates the entire state space (which in this example is of size $2^{11} = 2{,}048$) by categorizing 78% (i.e., 1,597) of the cases as suggestive of cancer and the remaining 22% (i.e., 451) as negative for cancer.

Some key definitions are summarized next. The *state space* expresses all possible

combinations of features. Note that the concept of state space is different than that of *population*. In fact, the population is a subset of the state space, because the actual population may not exhibit all the feature combinations. That is, some cases may not occur in reality and can be eliminated from consideration. As a result, the *sample set* (i.e., *the training data*) consists of elements (binary vectors in our illustrative example) drawn, with replacement, from the population rather than from the state space. Another result is that a sample may not represent the population and the state space equally. For instance, if a population includes 1,843 unique cases (i.e., 90% of the state space of 2,048 cases), then a sample of 80 vectors covers at most 3.9% of the state space and 4.3% of the population.

Therefore, in this illustrative scenario, it is assumed that 80 different cases were used to represent 2,048 cases. At this point one may wish to ask the question, "Is the function, which was inferred using a training sample of no more than 4.3% of the population, sufficiently reliable to recommend surgery for new patients?" While this function can be an interesting one, its statistical significance is questionable for a reliable diagnosis of cancer. If one considers multivalued instead of binary features, then a sample of 80 (which is a common sample size in published studies) becomes a minuscule portion of the entire state space. The statistical weakness becomes even more dramatic if one considers more features.

Next we define some key parameters for dealing with the reliability issue. Let us denote the *number of unique cases in the sample* as S, the *size of the state space* as N, and the *population size* as P. Obviously, the following relationship is always true: $N \geq P \geq S$. It should be observed that the sample size may not be the same as the number of training cases, since the sample size is the number of unique cases. For instance, patient 1 and patient 2 may correspond to the same combination, say (01111100011). Therefore the size of the sample set may be smaller than the number of cases in the sample. For example, the 15,000 mammograms of breasts without malignancy (unpublished data provided to us by the Woman's Hospital of Baton Rouge, LA, 1995) can be represented by fewer than 300 combinations of 11 features. Thus, the number of cases here is about 50 times greater than S.

Next we define the *index of potential reliability* by the ratio S/N and the *index of actual reliability* by the ratio S/P. In practice, it is very difficult to accurately estimate the size of the population, and consequently the index of actual reliability S/P. Obviously, if one has a sample which covers the entire population, then the index of actual reliability is equal to 1. On the other hand, if one has a proper subset of the population, then the size of the population cannot be determined directly. Note that it is possible to have different levels of reliability for different diagnostic classes within the same training set. In order to demonstrate this, we next compute the indices of potential reliability for the previous 11 features. Suppose that there are $N_1 = 1{,}600$ "highly suspicious for malignancy" vectors in a state space (which is of size $2^{11} = 2{,}048$). Then the remaining $N_0 = 448$ $(= 2{,}048 - 1{,}600)$ cases correspond to "not highly suspicious for malignancy" vectors. Next, suppose that there are $S_1 = 50$ unique "highly suspicious for malignancy" cases in one training set. Similarly, let the class "not highly suspicious for

malignancy" have $S_0 = 400$ unique cases in this training set. Then the indices of potential reliability for the respective groups are $S_1/N_1 = 0.03125$ (= 50/1,600) and $S_0/N_0 = 0.89286$ (= 400/448), respectively.

In the light of the previous reliability indices, we next consider the neural network (NN) approach described in (*13*). These authors constructed two different feed forward neural networks (NNs) which contained two layers of processing elements (PEs). Both NNs contained 43 input units, each one corresponding to an extracted radiographic feature, and a single PE in the output layer, representing the diagnosis (which was 0 for benign and 1 for malignancy). The two NNs differ merely in the number of PEs in the hidden layer; the first one used 10 while the second one used only 5 PEs. For each NN independently, they trained the PEs by back propagating their errors. For the training process, a set of 133 cases was selected from a mammography atlas. In addition, 60 other cases (of which 26 were malignant and 34 were benign) were randomly selected to evaluate the accuracy of the trained neural network. An experienced mammographer extracted the 43 features from each case and rated each feature on a scale from 0 to 10.

We can compute the potential reliability, as expressed by the sample/state space ratio, for the training data. The state space was defined on 43 features, each using 11 grades. Thus, this state space corresponds to a total number of 11^{43} different cases. Therefore, $S/N = 133/11^{43} = 2.21 \times 10^{-43}$. This means that the available sample is $2.21 \times 10^{-41}\%$, a minuscule fraction of the total possible number of different vectors in the state space. Note that for a particular number of training cases, the reliability index depends on the size of the feature set. As a result, 133 cases may be an insufficient number of cases for the previous state space while, say, 32 cases could be sufficient for a reliable diagnosis in a smaller state space. Suppose that one has only 5 binary diagnostic features. Then, the state space consists of 32 ($= 2^5$) combinations of these features. If all the 32 training cases represent unique vectors, then the size of the sample is 32 and the sample/state space ratio is equal to 1 (= 32/32), which is much better than the previous value of $2.21 \times 10^{-41}\%$. This example illustrates that the relative number of cases (i.e., the indices of reliability) is crucial, while a large number of cases may not be as valuable.

Therefore, the question which is naturally raised here is: "Can a relatively small number of training cases be considered reliable in order to assist in accurately diagnosing new (and thus unknown) cases?" Some neural network theoreticians, e.g., (*19*), suggest that the number of cases should be no less than 10 times the number of connections (i.e., the parameters needed to be estimated), to reliably train a neural network. Notice that this measure of reliability is similar to our index of potential reliability expressed as the ratio *S/N*. Gurney, in (*7*), has shown that this relatively weak requirement is not fulfilled in breast cancer CAD methods. For example, the largest neural net considered in (*13*) has 43 input units, 10 hidden layer PEs, and a single output PE. Thus, this network has 440 ($= 43 \times 10 + 10 \times 1$) connections, which is less than 532 ($= 4 \times 133$), where 133 is the number of cases used to estimate the weights for these connections.

Boone, in (*8*), disputed the 10:1 requirement on the number of cases versus the

number of connections. He compares the neural network with biological networks (e.g., radiologists) and argues that for biological networks the ratio is much worse, over 10^{10} times less, than the neural networks studies criticized by Gurney (*7*). Thus, Boone wonders: "Is there any reason that we should hold a computer to higher standards than a human?" Maybe not, but we should ask for both systems the question: "Is learning based on a small training subset sufficiently reliable to distinguish suspicious from nonsuspicious (of malignancy) cases given the vast diversity of mammographic images?"

Machine learning theory (see, for instance, (*20–24*)) addresses, among other issues, the problem of concept learning. It has been shown that there are relatively simple concepts that no algorithm is capable of learning in a reasonable amount of time (i.e., in polynomial time). The Probably Approximately Correct (PAC) learning theory, as introduced by Valiant in (*25*) (see also (*26, 27*)), provides a popular model of learnability. The central idea of the PAC model is that successful learning of an unknown target concept should entail obtaining, with high probability, a hypothesis that is a good approximation of the target concept (hence the term "probably approximately correct").

The machine learning literature provides a plethora of families of relatively simple concepts which cannot be learned reliably in this sense (please see also the previous references). Therefore, the question of reliability is among the most fundamental questions of scientific rigor and practical AI applicability to mammographic diagnosis. In this study we explore the two key questions: (*i*) "Are accessible and relatively small samples sufficiently representative for learning?" and (*ii*) "how can a broad range of mammographic features be evaluated?"

3. THE REPRESENTATION/NARROW VICINITY HYPOTHESIS

The problems of the sample/space ratio and the sample/population ratio reliability criteria are part of a general problem of many pattern recognition techniques. Pattern recognition techniques such as neural networks, methods based on logical formulae, decision trees, etc., generalize from prototypes (i.e., training sets). That is, these techniques propose methods that can discriminate new cases which were not among the prototypes (training cases). A common fundamental hypothesis, supporting small samples in pattern recognition, is the hypothesis that a small sample is representative of the entire population. This *representation hypothesis* is well stated by Miller, *et al.* in (*28*, p. 462) as follows: "The training data must still form a representative sample of the set of all possible inputs if the network is to perform correctly." The same authors also suggested that: "the principal problems which must be addressed when producing a complete network application are: collecting and classifying sufficient training and testing data, choosing a valid data presentation strategy and an appropriate network architecture." Thus, without confirming the representation hypothesis as it applies to mammography, pattern recognition with small samples may be of questionable reliability. In addressing this issue we study a restrictive version of the representation hypothesis, namely, the hypothesis of narrow vicinity (or the NV hypothesis):

"All real possible cases are in a *narrow vicinity* of an accessible small training sample."

If the NV hypothesis can be accepted, then we may generalize the training sample (of S vectors) to the actual population (of P vectors), but not to the remaining $(N - P)$ vectors, which do not represent feasible cases anyway. More formally, the NV hypothesis indicates that the P/N ratio is very small, i.e., the size of the actual population P is significantly less than the size of the state space N. If, for example, $S = 80$, $P = 160$, and $N = 2{,}048$, then the ratio P/N is equal to 0.078. Also, the index of the actual reliability S/P $(= 0.50)$ is significantly greater than the index of potential reliability S/N $(= 0.039)$. Therefore, the NV hypothesis provides the grounds to generalize from a small training subset. In Appendix II we describe some methods which can allow one to estimate the P/N, S/N, and S/P ratios without having the actual population for some typical mammographic and clinical features.

The problem of narrow vicinity is graphically illustrated in Fig. 1. This figure shows the areas (i.e., the narrow vicinities) surrounding the points that were used to train a hypothetical CAD system. The small ovals and rectangles represent test cases from the "not cancer" and "cancer" diagnostic classes, respectively. Next, suppose that the actual border line is the thick line near the "not cancer" training data and that linear discriminant analysis provided the dashed line. Then, the dotted rectangles will be misclassified by the estimated discriminant line. This illustrative example indicates that the extrapolation of training cases which are far from their narrow vicinities may lead to dramatically inaccurate conclusions.

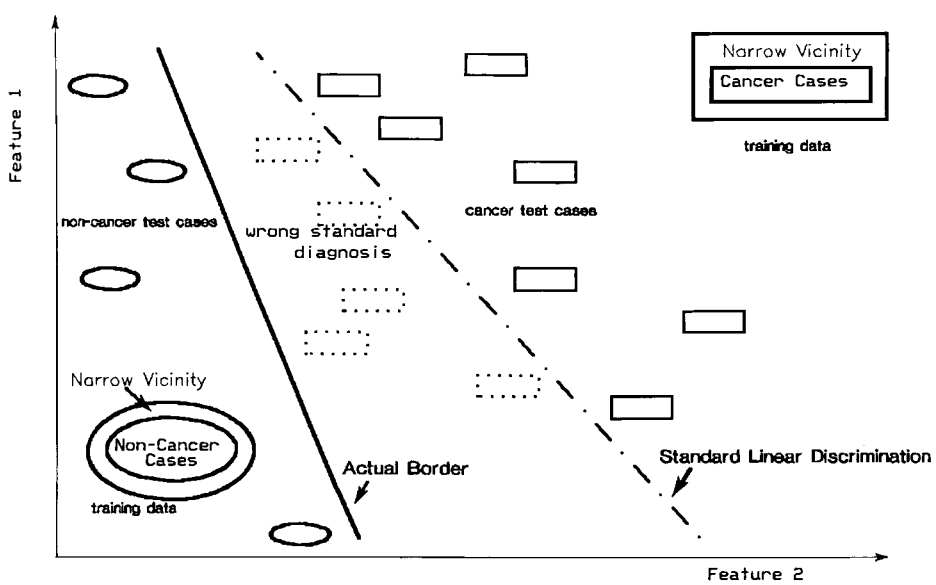

FIG. 1. Comparison of the actual and computed borders between diagnostic classes.

Our concerns about insufficient training data and the violation of the narrow vicinity hypothesis were confirmed for our actual data set. We used 156 actual cases, of which 77 were malignant and 79 were benign, provided to us by the Woman's Hospital in Baton Rouge, LA, in 1995. The cases were defined on the 11 attributes of clustered calcifications with the diagnostic classes "malignant" and "benign" as described in Appendix I. A raw version of this data set can be found in our web page, http://www.imse.lsu.edu/vangelis. We analyzed these data by using Fisher's linear discriminant analysis (*29–31*). By using linear discriminant analysis (LDA) one can estimate the line that minimizes the misclassification probability (given that this linear combination of the features follow a normal distribution and the classes have the same variance–covariance matrix). For the Woman's Hospital data, the line provided by LDA was able to correctly classify only 76% of the 156 cases. That is, a significant portion of the malignant cases were classified as benign, and vice versa. Note that the discriminant analysis framework is not capable of handling much more complex classification systems. For example, if the variance–covariance matrices are unequal, then the classification rule becomes quadratic and can lead to some strange results in dimensions higher than 2. This situation indicates the need for an entirely new framework of assumptions.

The classification patterns should be derived from the narrow vicinities of the available points. However, if one focuses only on the narrow vicinities of the available points, then there is the possibility of having too few data points, and thus the derived results may not be statistically significant on an 11-dimensional space. These observations are in direct agreement with the sample/population space (*S/P*) ratio problem discussed earlier. Furthermore, this brief analysis indicates the need for developing new inference approaches capable of dealing with the previous methodological weaknesses, such as wrongly extrapolating outside the observed points.

4. SOME COMPUTATIONAL RESULTS

Due to space limitation and the level of detail, we do not describe all the steps of the applied method, which is based on mathematical logic. A detailed description of the proposed method, and the specific steps, can be found in (*32–34*). This method of logical analysis is also briefly described in Appendix II. At first, let us note that the previous percentages of 78% and 22% of "cancer" and "not cancer" cases, respectively, are close to the actual percentages given in Section 2. About 80% of all possibilities in the state space indicate suspicion for cancer and recommendation for biopsy/short-term follow-up. A more detailed analysis has also shown that the borders of the biopsy/nonbiopsy regions are near the bottom of the state space (i.e., close to the vector containing all zeros).

We have found that our state space, as defined on the 11 binary diagnostic features, consists of 7.42% of possible cases (binary vectors) for which "biopsy/short term follow-up is not necessary," and 92.58% of the vectors for which "biopsy/short term follow-up is necessary." Similarly, this state space consists of

86.7% "highly suspicious for malignancy" cases and 13.3% "not highly suspicious for malignancy" cases (see also Table 1).

In order to understand the actual implications of the above issues one needs to consider the information derived by using actual historic cases. Suppose that one wishes to determine the above borders and percentages by using some sampled data S, which include cases of all examined patients at a hospital during a single year. At the Woman's Hospital of Baton Rouge, LA (unpublished data, 1995) there are 15,000 new cases with complete data each year. Approximately 0.2% of these patients have cancer and 98.8% have no cancer. Approximately 1.1% of these 15,000 women will undergo biopsy/short-term follow-up while the remaining 98.9% will receive routine follow-up.

The situation in the state space is almost the reverse of the real-life situation found in the Woman's Hospital experience, namely 0.2 and 99.8%. These numbers indicate that in a population of 15,000 mammograms we will have just 34 cases with cancer. Let us take this sample to discriminate 1,775 vectors representing suspicious findings (i.e., 86.7% of the total vectors) and the remaining 13.3% (i.e., 273 vectors) suggestive of benign lesions. Here the ratio sample/space (S/N) for cancer is equal to 34/1,775 = 0.019 (i.e., 1.9%) and for not cancer we have a ratio of 15,000/273 = 54.94 (i.e., 5,495%). Thus, we have a large surplus sample of patterns which are not representative of cancer and a very small sample representing highly suspicious findings indicating the presence of cancer (see also Table 1 and Figs. 2 and 3). Moreover, 1.9% is an upper estimate for the S/N ratio because different cases can be represented by the same combination of features.

The analysis presented in Figs. 2 and 3 shows that, in general, the narrow vicinity (NV) hypothesis is not valid for mammographic evaluation. Recall that this is exactly the hypothesis implicitly used by all traditional pattern recognition methods in breast cancer diagnosis! The diagnostic parameters which we used are typical for mammographic diagnosis (*13*).

The introduction of digital mammography has spawned a great deal of research

TABLE 1

Comparison of Sample and Class Sizes for Biopsy and Cancer (from Woman's Hospital in Baton Rouge, LA, Unpublished Data, 1995)

	Sample size	Class size in state space	Sample/space ratio
Total size	15,000	2,048	7.32
Cases with cancer	34	1,775	0.02
Percentage (%)	0.20	86.70	
Cases without cancer	14,966	273	54.82
Percentage (%)	99.80	13.30	
Number of biopsies	165	1,896	0.09
Percentage (%)	1.10	92.58	
Number of nonbiopsies	14,835	152	97.60
Percentage (%)	98.90	7.42	

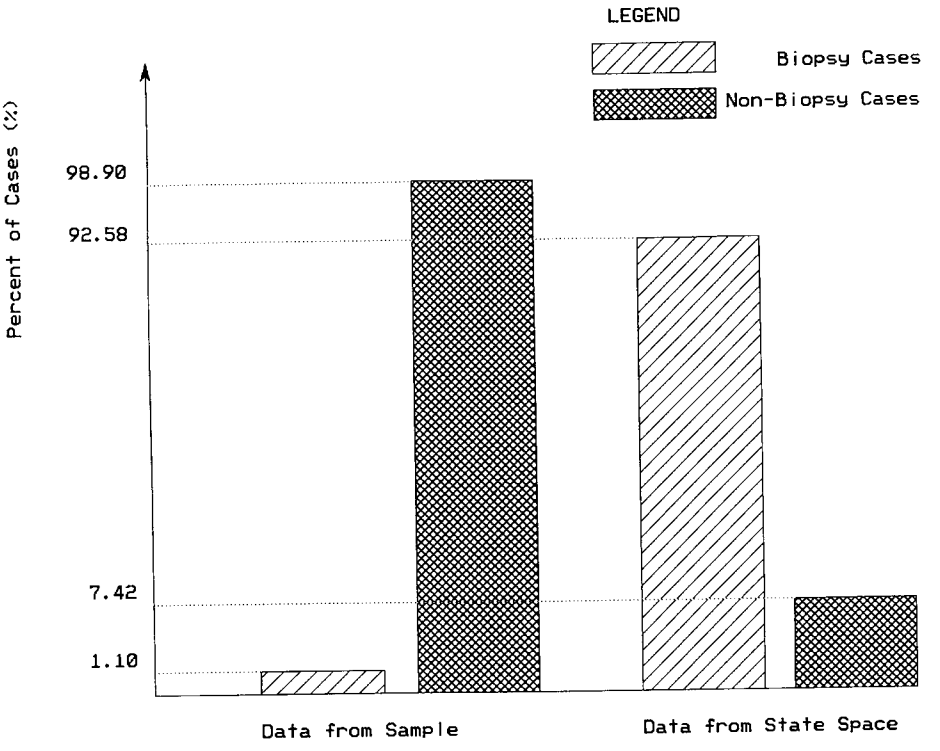

FIG. 2. Relations between biopsy class size and sample.

in artificial intelligence techniques applied to breast cancer diagnosis. These methods range from K-nearest neighbor rules (e.g., (*35*)) to the application of genetic algorithms (GAs) (e.g., (*36*)). However, the majority of these studies are concerned with the application of neural networks to extract features and classify tumors. For some recent developments in this particular area, the interested reader may want to consult the work reported in (*38–40*).

5. CONCLUDING REMARKS

This study shows that the development of reliable CAD methods for breast cancer diagnosis requires more attention on the problem of the selection of training and testing data and processing methods. Strictly speaking, all CAD methods are still very unreliable in spite of the apparent, and possibly fortuitous, high accuracy of cancer diagnosis reported in the literature. Our computations clearly show that a standard random selection of test cases (*13*) does not give a true picture of the accuracy/reliability of breast cancer diagnosis. The receiver operator characteristic (ROC)-based analysis (see, for instance, (*41–43, 12*)) used to evaluate the accuracy of diagnosis suffers from this weakness.

There are several approaches and methods which can be used to improve this

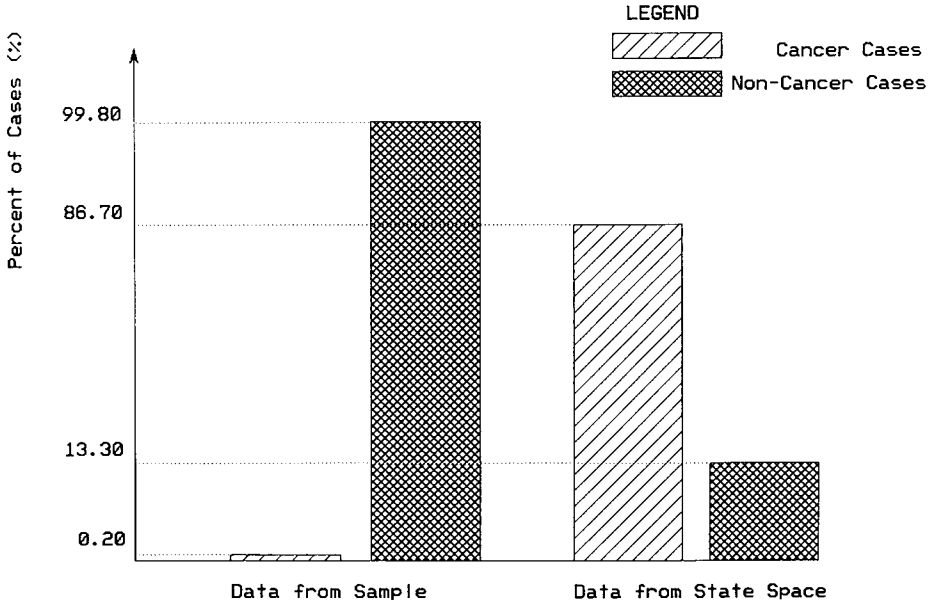

FIG. 3. Relations between cancer class size and sample.

situation (some of them were used in this study). Nevertheless, the low reliability of CADs should be recognized by the scientific community and their application should be reconsidered. It should be emphasized here that the problem is not only in the methods themselves. Not only they are implemented in situations where they are inappropriate but they also provide a false sense of security since the literature tends to inflate their reliability. Reliable diagnosis can be obtained if:

(*a*) research on the feature space has shown that the chosen training data actually represent the border between diagnostic classes and

(*b*) the mathematical method to be used can extract this border.

The main advantage of the methods which we used is that they allow one to identify and evaluate the reliability of CAD methods. Standard random selection of test cases often does not adequately represent the critical border points (see also Fig. 1). The proposed approach allows one to select test cases near the border of diagnostic classes, i.e., critical points for verifying those cases that are regarded as borderline benign/malignant cases. The last and most important point is that the applied method can improve gains in accuracy and construct reliable diagnostic (discriminant) functions. In summary, this paper focused on the following four main issues related to computerized breast cancer diagnosis:

(*i*) The state space of all possible features may be astronomically large, and as a result, the size of the population (which is a subset of the state space) may be of the same magnitude.

(*ii*) Most samples represent a tiny fraction of the possible population space.

Therefore, results obtained by traditional approaches, although might be correct on some test cases, are not statistically significant, unless the representation/narrow vicinity (NV) hypothesis is accepted.

(*iii*) The representation/narrow vicinity (NV) hypothesis may not always be valid.

(*iv*) Fortunately, real-life data may exhibit the monotonicity property. Approaches which explicitly use this critical property may alleviate some of the previous problems. The approach proposed by the authors in previous papers, which uses the monotonicity property, may offer an effective and efficient way to overcome these reliability problems.

Finally, it should be stated that we used the paradigm of breast cancer diagnosis because it is a socially and medically critical subject and because it possesses important characteristics that require a critical appraisal of the reliability issue in a real-life situation.

APPENDIX I: DEFINITIONS OF THE KEY FEATURES

The main study described in this paper was performed for the binary features presented below. We deliberately used nonspecific terms, such as "small," "large," "pro cancer," and "contra cancer," in order to allow us to further refine the language. An approach which uses nonbinary values and which is based on fuzzy logic is described in (*44*).

The list of indirect diagnostic features, along with their meaning, are defined as follows.

x_1	Amount and volume of calcifications	(0-contra cancer/biopsy; 1-pro cancer/biopsy)

Note, that x_1 was considered to be a function $\psi(w_1, w_2, w_3)$ of the features w_1, w_2, w_3 defined as follows.

w_1	Number of calcifications/cm^2	(1-large, 0-small)
w_2	Volume, cm^3 (approximate)	(1-small, 0-large)
w_3	Total number of calcifications	(1-large, 0-small)

x_2	Shape and density of calcifications	(0-cancer/biopsy; 1-pro cancer/biopsy)

Note that we consider x_2 as a function $\psi(y_1, y_2, y_3, y_4, y_5)$ of y_1, y_2, y_3, y_4, y_5, which are determined as follows.

y_1	Irregularity in the shape of individual calcifications	(1-marked, 0-mild)
y_2	Variation in the shape of calcifications	(1-marked; 0-mild)
y_3	Variation in the size of calcifications	(1-marked; 0-mild)
y_4	Variation in the density of calcifications	(1-marked; 0-mild)
y_5	Density of the calcifications	(1-marked; 0-mild)

x_3	Ductal orientation	(0-not ductal; 1-ductal)
x_4	Comparison with previous exam	(0-contra cancer/biopsy; 1-pro cancer/biopsy)
x_5	Associated findings	(0-contra cancer/biopsy; 1-pro cancer/biopsy)

Thus, we used the state space that consisted of the 11 binary features w_1, w_2, w_3, y_1, y_2, y_3, y_4, y_5, x_3, x_4, x_5. Features x_1 and x_2 were used to construct a hierarchy of features, as described in Appendix II.

APPENDIX II: TECHNICAL PROCEDURES

The Interactive Approach

Let us consider how one can validate the narrow vicinity (NV) hypothesis when a small sample set is available. If one has a large sample set available, then one does not need the NV hypothesis. On the other hand, with a small sample, one does not have to directly validate this hypothesis. We developed a new methodology to overcome these difficulties. The main idea is to extend insufficient clinical cases with information from an experienced radiologist. Another approach is mentioned in (*28*, p.462): "*One obvious solution to the problem of restricted training and testing data is to create simulated data using either a computer based or physical model.*" We used experienced experts as a "*human*" model to generate new examples.

One can ask a radiologist to evaluate a particular case when a number of features take on a set of specific values. A typical query in our experiments had the following format:

"If feature 1 has value V_1, feature 2 has value V_2, . . . , feature n has value V_n, then should biopsy/short-term follow-up be recommended or not? Or, does the above setting of values correspond to a highly suspicious case or not?"

The above queries can be defined with artificially constructed vectors (as will be explained below) or with artificially generated new mammograms by modifying existing ones. In this way one may increase a sample size, but not as much as may be necessary. Roughly speaking, the technical weakness now is the same as before. That is, it is practically impossible to ask a radiologist to generate many thousands of artificial mammographic cases.

One can overcome these difficulties in two ways. First, if the features can be organized in a hierarchical manner, then a proper exploitation of this structure can lead to a significant reduction of the needed queries. Second, if the property of monotonicity, as explained below, is applicable, then the available data can be generalized to cover a larger training sample. The specific mathematical steps of how to achieve the above two goals are best described in (*33, 34*).

At this point it should be stated that the issue of monotonicity in Boolean functions has been studied extensively by Hansel (*45*). Hansel proposed what has become a famous theorem on the worst-case complexity of learning monotone Boolean functions. However, his theorem had not been translated into English until recently, when the authors discussed monotonicity in (*32–34, 46*). However, there

are numerous references to it in the non-English literature (Hansel wrote his paper in French). This theorem is one of the finest results of the long-term efforts in monotone Boolean functions that begun with Dedekind in 1897 (*47*).

The Hierarchical Approach

One can construct a hierarchy of *medically interpretable* features from a very generalized level to a less generalized level. For example, we considered the generalized binary feature "Shape and density of calcification" with grades (0-"contra cancer" and 1-"pro cancer") denoted by x_2. On the second level we considered the feature x_2 to be some function ψ of five other features, $y_1, y_2, \ldots, y_5$. That is, $x_1 = \psi(y_1, y_2, \ldots, y_5)$, where

y_1 is irregularity in the shape of the individual calcifications,
y_2 is variation in the shape of the calcifications,
y_3 is variation in the size of the calcifications,
y_4 is variation in the density of the calcifications,
y_5 is density of the calcifications.

For illustrative purposes we will consider the above features as being binary valued with grades (1) for "marked" and (0) for "minimal" or, equivalently, (1)-"pro cancer" and (0)-"contra cancer."

The Monotonicity Property

If we can identify regularities in advance, then it is possible to decrease the number of calls to a radiologist required to classify (diagnose) particular vectors (clinical cases). Monotonicity is one such regularity, and it may greatly reduce the number of diagnoses while maintaining a general hypothesis because many nonmonotone regularities can also be represented as a combination of several monotone regularities (*32–34*).

In order to clarify how the monotonicity property can be applied to the breast cancer diagnosis problem, consider the evaluation of calcifications in a mammogram. For simplicity and illustrative purposes assume that x_1 is the number and the volume occupied by calcifications, in a binary setting, as follows: (0-"contra cancer", 1-"pro cancer"). Similarly, let

x_2	{shape and the density of the calcifications}, with values:	0-"contra cancer", 1-"pro cancer",
x_3	{ductal orientation}, with values:	0-"contra cancer", 1-"pro cancer",
x_4	{comparison with previous examination}, with values:	0-"contra cancer", 1-"pro cancer",
x_5	{associated findings}, with values:	0-"contra cancer", 1-"pro cancer".

Given the above definitions we can represent clinical cases in terms of binary vectors with these five features as $(x_1, x_2, x_3, x_4, x_5)$. Next consider the two clinical cases which are represented by the two binary vectors (10100) and (10110). The vector

(10100) means that the number and the volume occupied by calcifications is *"pro cancer"* (e.g., $x_1 = 1$) and ductal orientation is "pro cancer" (e.g., $x_3 = 1$) for the first case. The vector (10110) shows an extra "pro cancer" feature for the second case; i.e., the comparison with the previous examination is "pro cancer" (i.e., $x_4 = 1$).

If a radiologist correctly diagnosed the first clinical case (10100) as malignant, then we can also conclude that the second clinical case (10110) should also be malignant. The latter case has all the "pro cancer" features of the first case plus an extra one (i.e., $x_4 = 1$). In a similar manner, if we know that (01010) is not considered suspicious for cancer, then the second case (00000) should not be considered suspicious for cancer. This is true because the second case has all the "contra cancer" characteristics of the former one in addition to other "contra cancer" characteristics. These examples roughly illustrate the property of monotonicity in Boolean functions and indicate how our algorithms explicitly exploit monotonicity. One can combine a hierarchical approach with monotonicity and generalize accordingly. In this way, weaknesses of the traditional pattern recognition methods can be alleviated.

Logical Discriminant Functions

In (*33, 34*) we show that by using the idea described above, the monotone Boolean discriminant functions for the features on the uppermost level of the hierarchy are as follows.

For the "biopsy/short term follow-up" subproblem,

$$f_1(x) = x_2x_4 \vee x_1x_2 \vee x_1x_4 \vee x_3 \vee x_5. \qquad [1]$$

Similarly, for the second subproblem (i.e., "highly suspicious for cancer") the extracted function was

$$f_2(x) = x_1x_2 \vee x_3 \vee (x_2 \vee x_1 \vee x_4)x_5. \qquad [2]$$

Regarding the second level of the hierarchy (which has 11 binary features), we interactively constructed the following functions (an interpretation of the features is presented in Appendix I):

$$x_1 = \varphi(w_1, w_2, w_3) = w_2 \vee w_1w_3, \qquad [3]$$

$$x_2 = \psi(y_1, y_2, y_3, y_4, y_5) = y_1 \vee y_2 \vee y_3y_4y_5. \qquad [4]$$

By combining the functions in [1]–[4], we obtained the formulae of all the 11 features for "biopsy/short-term follow-up",

$$f_1(x) = (y_2 \vee y_1 \vee y_3y_4y_5)x_4 \vee (w_2 \vee w_1w_3)(y_2 \vee y_1 \vee y_3y_4y_5) \vee (w_2 \vee w_1w_3)x_4 \vee x_3 \vee x_5, \qquad [5]$$

and for highly suspicious for cancer,

$$\begin{aligned} f_2(x) &= x_1x_2 \vee x_3 \vee (x_2 \vee x_1 \vee x_4)x_5 \\ &= (w_2 \vee w_1w_3)(y_1 \vee y_2 \vee y_3y_4y_5) \vee x_3 \\ &\quad \vee (y_1 \vee y_2 \vee y_3y_4y_5) \vee (w_2 \vee w_1w_3 \vee x_4)x_5. \end{aligned} \quad [6]$$

The benefit of having these functions is twofold. First, they express patterns as logical expressions (i.e., as decision rules) which can allow us to identify the *real border* between diagnostic classes. Second, they allowed us to compute the *size of the classes* presented in Table 1 and depicted in Figs. 2 and 3.

REFERENCES

1. Wingo, P. A., Tong, T., and Bolden, S. Cancer statistics. *Ca Cancer J. Clinicians* **45**(1), 8–30 (1995).
2. Bird, R. E., Wallace, T. W., and Yankaskas, B. C. Analysis of cancer missed at screening mammography. *Radiology* **184,** 613–617 (1992).
3. Burhenne, H. J., Burhenne, L. W., Goldberg, D., Hislop, T. G., Worth, A. J., Ribeck, P. M., and Kan, L. Interval breast cancer in screening mammography program of British Columbia: Analysis and calcification. *AJR* **162,** 1067–1071 (1994).
4. Hall, F. M., Storella, J. M., Silverstone, D. Z., and Wyshak, G. Nonpalpable breast lesions: Recommendations for biopsy based on suspicion of carcinoma at mammography. *Radiology* **167,** 353–358 (1988).
5. Elmore, J., Wells, M., Carol, M., Lee, H., Howard, D., and Feinstein, A. Variability in radiologists' interpretation of mammograms. *New England J. Med.* **331**(22), 1493–1499 (1994).
6. Kopans, D. The accuracy of mammographic interpretation (editorial). *New England J. Med.* **331**(22), 1521–1522 (1994).
7. Gurney, J. Neural networks at the crossroads: Caution ahead. *Radiology* **193**(1), 27–28 (1994).
8. Boone, J. Sidetracked at the crossroads. *Radiology* **193**(1), 28–30 (1994).
9. Gale, D., Roebuck, E., and Riley, E. Computer aids to mammographic diagnosis. *Br. J. Radiology* **60,** 887–891 (1987).
10. Getty, D., Pickett, R., D'Orsi, C., and Swets, J. Enhanced interpretation of diagnostic images. *Investigative Radiology* **23,** 240–252 (1988).
11. Swets, J., Getty, D., Pickett, R., D'Orsi, C., Seltzer, S., and McNeil, B. Enhancing and evaluating diagnostic accuracy. *Med. Decision Making* **11,** 9–18 (1991).
12. D'Orsi, C., Getty, D., Swets, J., Picket, R., Seltzer, S., and McNeil, B. Reading and decision aids for improved accuracy and standardization of mammographic diagnosis. *Radiology* **184,** 619–622 (1992).
13. Wu, Y., Giger, M., Doi, K., Vyborny, C. J., Schmidt, R., and Metz, C. Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer. *Radiology* **187**(1), 81–87 (1993).
14. Vyborny, C. J. Can computers help radiologists read mammograms? *Radiology* **191,** 315–317 (1994).
15. Vyborny, C. J., and Giger, M. Computer vision and artificial intelligence in mammography. *AJR* **162,** 699–708 (1994).
16. Mangasarian, O. L. Mathematical programming in neural networks. *ORSA J. Comput.* **5**(4), 349–360 (1993).
17. Johnosn, N. Everyday diagnostics—A critique of the Bayesian model. *Med. Hypotheses* **34**(4), 289–296 (1991).
18. Duda, R. O., and Hart, P. E. "Pattern Classification and Scene Analysis." Wiley, New York, 1973.
19. Baum, E. B., and Hausler, D. What size net gives valid generalizations? *Neural Computation* **1,** 151–160 (1989).

20. Schapire, R. "The Design and Analysis of Efficient Learning Algorithms." MIT Press, Boston, MA, 1992.
21. "Machine Learning '95, The 12th International Conference on Machine Learning, Tahoe City, California." Kaufmann, Los Altos, CA 1995.
22. Lavrac, N., and Wrobel, S., Eds. "Machine Learning: ECML-95, 8th European Conference on Machine Learning." Springer, Berlin, 1995.
23. "Computational Learning Theory, *Proceedings of the 8th Annual Conference on Computational Learning Theory.*" Assoc. Comput. Mach., Santa Cruz, CA, 1995.
24. "Computational Learning Theory, Second European Conference, EuroCOLT'95." Springer, Berlin, 1995.
25. Valiant, L. G. A theory of the learnable. *Comm. ACM* **27**(11), 1134–1142 (1984).
26. Angluin, D. Queries and concept learning. *Machine Learning* **2,** 319–342 (1988).
27. Haussler, D., and Warmuth, M. The probably approximately correct (PAC) and other learning models. *In* "Foundations of Knowledge Acquisition: Machine Learning" (A. L. Meyrowitz and S. Chipman, Eds.), pp. 291–312. Kluwer Academic, Norwell, MA, 1993.
28. Miller, A., Blott, B., and Hames, T. Review of neural network applications in medical imaging and signal processing. *Med. Biol. Eng. Comput.* **30,** 449–464 (1992).
29. Fisher, R. A. The statistical utilization of multiple measurements. *Ann. Eugenics* **8,** 376–386 (1938).
30. Fisher, R. A. The use of multiple measurements in taxonomic problems. *Ann. Eugenics* **7,** 179–188 (1936).
31. Johnson, R. A., and Wichern, D. W. "Applied Multivariate Statistical Analysis," 3rd ed. Prentice Hall, Upper Saddle River, NJ, 1992.
32. Kovalerchuk, B., Triantaphyllou, E., and Vityaev, E. Monotone Boolean functions learning technique integrated with user interaction. *In* "Proceedings, Workshop on Learning from Examples vs. Programming by Demonstration, 12th International Conference on Machine Learning, Tahoe City, California," pp. 41–48, 1995.
33. Kovalerchuk, B., Triantaphyllou, E., and Ruiz, J. F. Monotonicity and logical analysis of data: A mechanism for evaluation of mammographic and clinical data. *In* "Computer Applications to Assist Radiology, Denver, CO, Symposia Foundation, June 6–9," pp. 191–196, 1996.
34. Kovalerchuk, B., Triantaphyllou, E., Deshpande, A. S., and Vityaev, E. Interactive learning of monotone Boolean functions. *Information Sci.* **94**(1–4), 87–118 (1996).
35. Hojjatoleslami, A., and Kittler, J. Detection of clusters of micro-calsifications using a K-nearest neighbor rule with locally optimum distance metric. *In* "Digital Mammography '96, Proceedings of the 3rd International Workshop on Digital Mammography Chicago, IL, June 9–12" (K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt, Eds.), Chicago, IL, June 9–12, pp. 267–272, 1996.
36. Sahiner, B., Chan, H.-P., Petrick, N., Helvie, M. A., Goodsitt, M. M., and Adler, D. D. Classification of mass and normal tissue: Feature selection using a genetic algorithm. *In* "Digital Mammography '96, Proceedings of the 3rd International Workshop on Digital Mammography, Chicago, IL, 9–12 June" (K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt, Eds.), pp. 379–384, 1996.
37. Rosen, D., Martin, B., Monheit, M., Wolff, G., and Stanton, M. Bayesian neural network to detect micro-calsifications in digitized mammograms. *In* "Digital Mammography '96, Proceedings of the 3rd International Workshop on Digital Mammography, Chicago, IL, 9–12 June" (K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt, Eds.), pp. 277–282.
38. Chan, H. P., Lo, S. C. B., Sahiner, B., Lam, K. L., and Helvie, M.A. Computer-aided detection of mammographic micro-calsifications: Pattern recognition with an artificial neural network. *Med. Phys.* **22**(10), 1555–1567 (1995).
39. Floyed Carey, Jr., E., Lo, J. Y., Yun, A. J., Sullivan, D. C., and Kornguth, P. J. Prediction of breast cancer malignancy using an artificial neural network. *Cancer* **74**(11), 2944–2948 (1994).
40. Zhang, W., Doi, K., Giger, M., Wu, Y., Nishikawa, R. M., and Metz, C. Computerized detection of clustered microclassifications in digital mammograms using a shift-invariant artificial neural network. *Med. Phys.* **21**(4), 517–524 (1994).
41. Kegelmeyer, W., Pruneda, J., Bourland, P., Hills, A., Riggs, M., and Nipper, M. Computer-aided mammographic screening for spiculated lesion. *Radiology* **191**(2), 331–337 (1994).
42. Jiang, Y., Nishikawa, R.M., Metz, C. E., Wolverton, D. E., Schmidt, R. A., Papaioannou, J., and

Doi, K. A computer-aided diagnostic scheme for classification of malignant and benign clustered microclassifications in mammograms. *In* "Digital Mammography '96, Proceedings of the 3rd International Workshop on Digital Mammography, Chicago, IL, June 9–12" (K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt, Eds.), pp. 219–224, 1996.

43. Huo, Z., Giger, M. L., Vyborny, C. J., Wolverton, D. E., Schmidt, R. A., and Doi, K. Computer-aided diagnosis: Automated classification of mammographic mass lesions. *In* "Digital Mammography '96, Proceedings of the 3rd International Workshop on Digital Mammography, Chicago, IL, June 9–12" (K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt, Eds.), pp. 207–211, 1996.

44. Kovalerchuk, B. Triantaphyllou, E., Ruiz, J. F., and Clayton, J. Fuzzy logic in computer-aided breast cancer diagnosis: Analysis of lobulation. *Artificial Intelligence Med.* **11,** 75–85 (1997).

45. Hansel, G. Sur le nombre des fonctions Boolenes monotones den variables. *C.R. Acad. Sci. Paris* **262**(20), 1088–1090 (1966). [In French]

46. Triantaphyllou, E., Kovalerchuk, B., and Deshpande, A. S. Some recent developments in logical analysis. *In* "Interfaces in Computer Science and Operations Research" (R. Barr, R. Helgason, and J. Kennington, Eds.), pp. 215–236, 1996. Kluwer Academic, Dondrecht/Norwell, MA, 1996.

47. Dedekind, R. Ueber Zerlegungen von Zahlen durch ihre grossten gemeinsamen Teiler. *Festschrift Hoch. Braunschweig* **II,** 103–148 (1897). [In German]