

Breast Cancer Diagnosis From Biopsy Images Using Generic Features and SVMs

Alexander Brook, Ran El-Yaniv, Eran Isler, Ron Kimmel, *Senior Member, IEEE*, Ron Meir,
and Dori Peleg

Abstract

A fully automatic method for breast cancer diagnosis based on microscopic biopsy images is presented. The method achieves high recognition rates by applying multi-class support vector machines on generic feature vectors that are based on level-set statistics of the images. We also consider the problem of classification with rejection and show preliminary results that point to the potential benefits.

Index Terms

breast cancer, biopsy, automatic diagnosis, SVM, image processing, feature extraction

I. INTRODUCTION

Recent years have witnessed a large increase of interest in automated and semi-automated breast cancer diagnosis [1], [2], [3], [4], [5], [6], [7], [8], [9]; see [10] for a recent review. In this paper we present an automatic classification method for breast cancer diagnosis based on microscopic biopsy images. In particular, we consider the problem of classifying a tissue specimen as either *healthy*, tumor *in situ*¹, or *invasive* carcinoma.

We propose a fully automatic classification method, using generic features and state-of-the-art statistical learning algorithms and methodologies. We experiment with a dataset that was previously used in [11], where a highly specialized morphology-based feature generation process was used. We show here that simple generic features lead to slightly superior (6.6% vs. 8.0%) error rate.

A desirable functionality of automated or semi-automated medical diagnosis systems is the option of ‘decision with rejection’ whereby the system generates decisions with confidence larger than some prescribed threshold and transfers the decision on cases with lower confidence to a human expert. In this work we also examine a simple method for decision with rejection and demonstrate its viability.

Figure 1 presents three sample images of healthy tissue, tumor in situ and invasive carcinoma. Note that the images are not clear-cut, and contain much information of no diagnostic significance. These were the individual samples in our work; they were not subdivided or cropped in any way. For more information on the dataset, see Section IV.

Corresponding author: Alexander Brook, abrook@technion.ac.il; Author names appear in alphabetical order.

¹From Latin “in place”, localized.

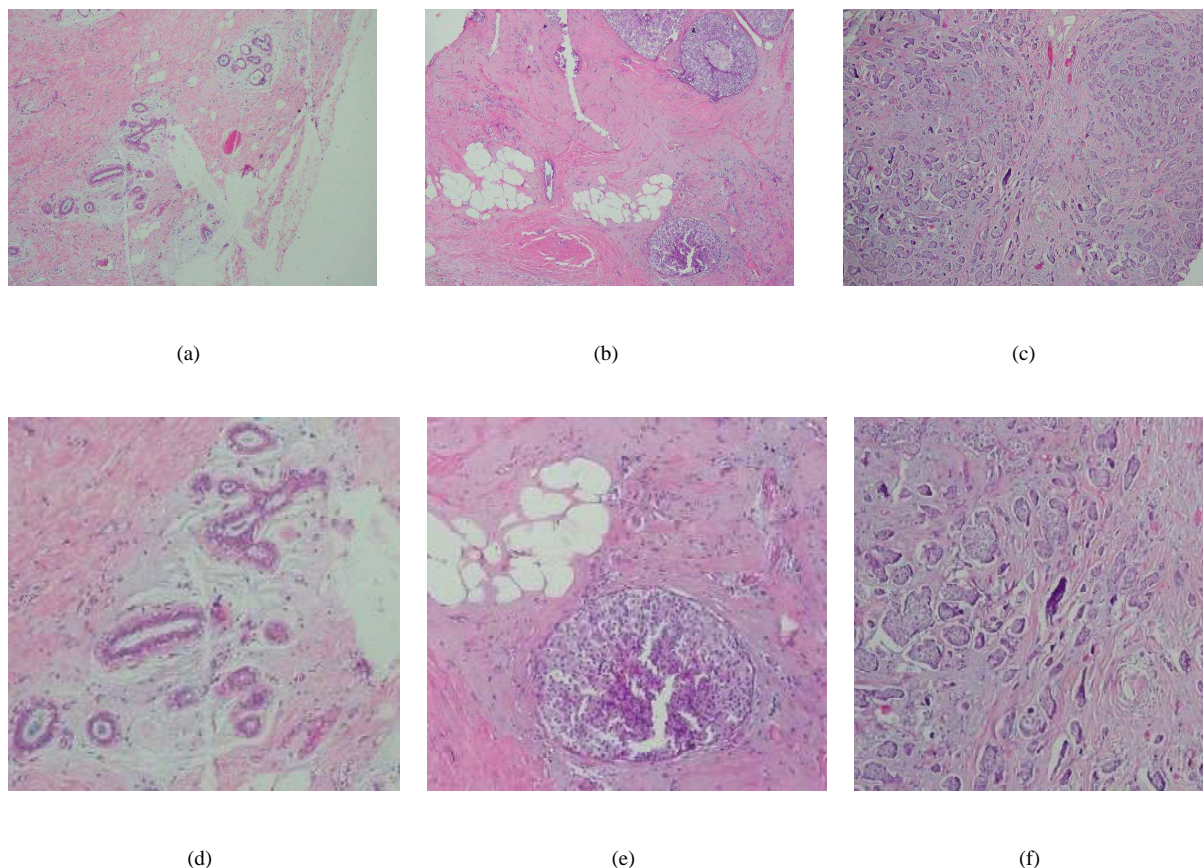


Fig. 1. Typical image instances. (a) normal: normal breast tissue, with ducts and finer structures; (b) carcinoma in situ: tumor confined to a well-defined small region, usually a duct; (c) invasive: breast tissue completely replaced by the tumor; (d-f) are enlarged parts of the images in (a-c), selected to clearly depict examples of structures specific to each of the three cases.

This paper is organized as follows. We begin in Section II with a description of the feature generation process. In Section III we explain the learning procedure and performance assessment protocol. The results are then presented and characterized in Section IV, and a comparison to previous work is provided in Section V. We summarize our conclusions and some directions for future research in Section VI.

II. FEATURE GENERATION

The ultimate goal of any automatic classification procedure is achieving low average risk rates. Clearly, this was also our goal in this research. However, in the present context we also aimed at achieving high accuracy with simple, generic features that are easy to understand, fast to compute and potentially useful for other related problems. With these goals in mind, we settled on simple statistics of gray-level images. Our feature generation method consists of the following three stages:

- 1) Grayscale conversion;
- 2) Level-set formation;
- 3) Computation of connected component statistics.

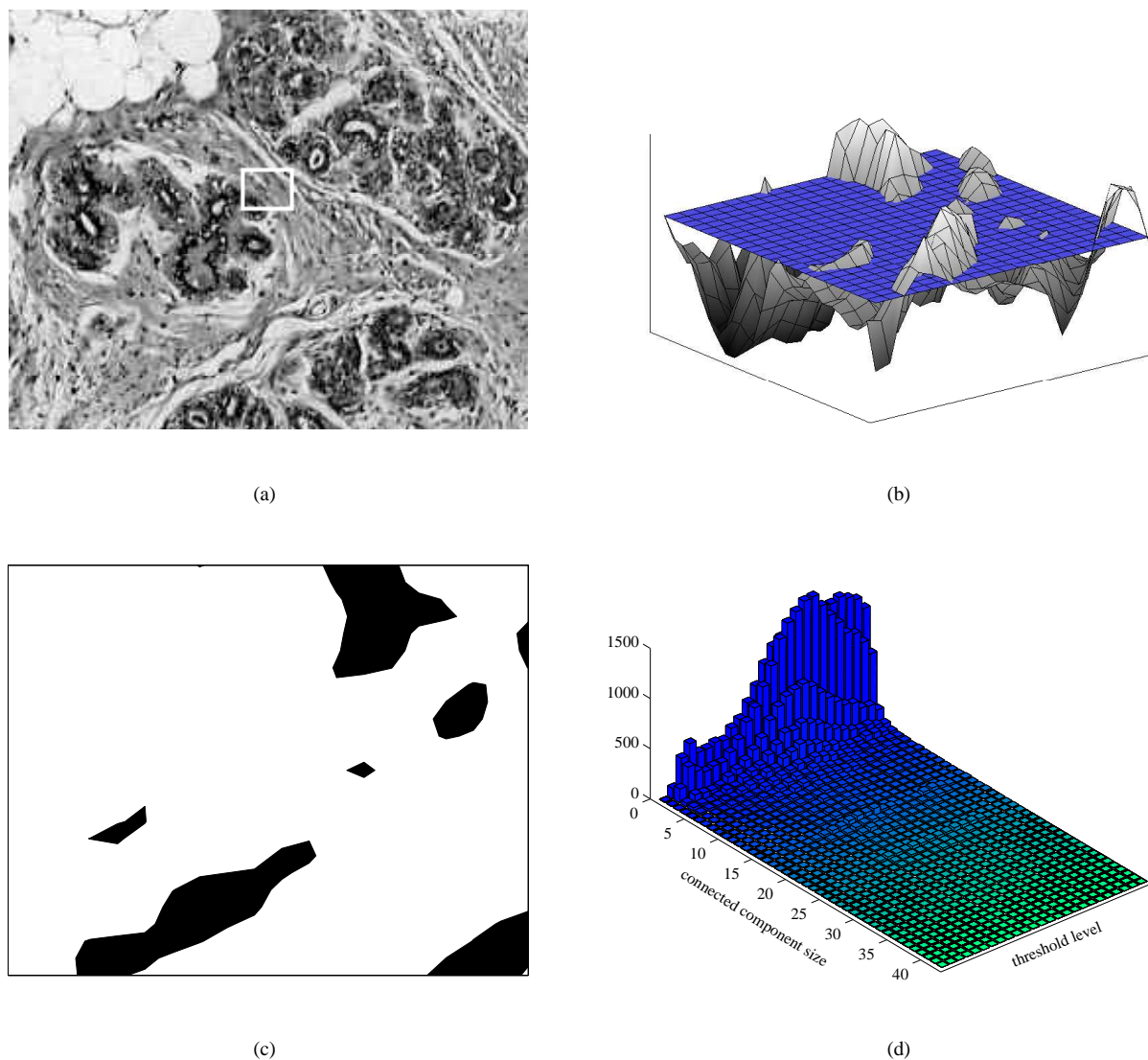


Fig. 2. Feature generation. (a) gray-level image; the white frame indicates the part used in (b) and (c); (b) a small portion of (a) viewed as a function of two variables; we are interested in the pixels where the brightness is above a certain threshold given by the horizontal hyperplane; (c) the corresponding binary image; (d) the histogram of sizes of connected components at several threshold levels.

We proceed with a description of each stage. The images were first converted from color to grayscale. Principal component analysis (PCA) was performed on the RGB values of the pixels of each image, and all the colors were then projected onto the principal axis. The principal axis was always very close to the gray axis.² The resulting grayscale images were then all brought to the same intensity range by stretching (while clipping the top and bottom 1% of the pixels).

We then formed the *level sets* of these images (see Figure 2). These may be viewed as binary images with black corresponding to pixels with gray level which is above a given threshold. For images with pixel levels

²This means that the results with any standard color-to-grayscale conversion should be very similar.

between 0 and 255, we used threshold values separated by steps of 10, thus there were 25 threshold levels.

Finally, for each resulting binary image we computed a histogram of 42 bins corresponding to connected components' sizes. The bins were not uniform and were selected empirically to provide a reasonable tradeoff between resolution and the number of bins; mostly, we have used bins just large enough to prevent multiple empty bins. The bin boundaries are the only parameters of the feature extraction stage. See Figure 2(d) for an example of the resulting histogram.

The complete MATLAB code for feature generation from a gray-level image is

```
k=[1 2:2:30 40:10:100 150 200:100:1000 2000:1000:10000 inf];
% Create bin boundaries.
for j=10:10:250
    sizes = regionprops(bwlabel(image<j),'Area');
    % Here bwlabel labels the connected components, and regionprops calculates their areas.
    feature = histc([sizes.Area],k);
    % Compute the counts for all the bins...
    features(j/10,:) = feature(1:end-1);
    % ...and add them to the features vector.
end
```

III. LEARNING ALGORITHM

Following feature generation, the pattern recognition task is a three-class classification problem where each instance is represented by a vector of 1050 features.³ Due to the relatively large number of features and their correlative nature, we initially examined possibilities of feature selection. In particular, several state-of-the-art feature selection algorithms, both linear [12], [13], [14] and non-linear [15], were tested and compared to a support vector machine (SVM) [16], applied on all the features. These initial experiments indicated that feature selection only decreases the performance and we decided to apply the SVM on all the features.

A. Multiclass Problem Decomposition

The standard SVM (e.g. [16]) is applicable only to binary classification problems. Multi-class problems, such as the one addressed here, are handled by decomposing the multi-class problem into several binary subproblems and aggregating the results of the binary classifiers according to some scheme. Popular decomposition methods are *one-against-all*, *all-pairs*, and *error correcting output coding (ECOC)* [17], [18], [19].

For the ECOC framework we used [20]. Each of the k given classes is assigned a unique vector (called a *codeword*) of length ℓ over $\{1, 0, -1\}$. This collection of k codewords forms a $k \times \ell$ *coding matrix* M , whose ℓ columns define ℓ binary partitions of the k classes. The entries $M(i, j) = 1$ and $M(i, j) = -1$ signify that for classifier f_j , the class of pattern x_i is 1 and -1 , respectively. The zero entries $M(i, j) = 0$ signify that classifier

³In fact, 47 features were constant and therefore were ignored.

f_j ignores pattern x_i . In (2) below we show two examples of coding matrices corresponding to well-known decomposition methods for multi-class problems.

Given a training set $S = \{(x_i, y_i)\}$, ℓ binary classifiers are trained. The j th classifier f_j is assigned a unique binary partition defined by the j th column of M and is trained using a training set $\{(x_i, M(i, j))\}$, where zero entries (i.e., $M(i, j) = 0$) are ignored.

After the learning process is complete, whenever an unseen point x is given, it is classified by all binary classifiers. This results in a vector $\bar{f}(x) = (f_1(x), \dots, f_\ell(x))$ with $f_j(x)$ being the output of the j th classifier. The point x is assigned to the class whose matrix row is closest to $\bar{f}(x)$. This class assignment mechanism is called *decoding*. In the basic ECOC scheme [17], [18], a Hamming-based decoding is used where the distance between $\bar{f}(x)$ and the rows of the matrix is computed using the Hamming distance. Another possibility is to use an exponential distance function. The exponential distance of a pattern x from the code associated with class i is defined as

$$d_i(x) = \sum_{j=1}^{\ell} e^{-M(i,j)f_j(x)}. \quad (1)$$

We performed initial experiments with several multi-class decomposition schemes. Specifically, we tested one-against-all, all-pairs and ECOC applied with coding matrices corresponding to these two methods. These matrices are

$$M = \begin{pmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix} \quad M = \begin{pmatrix} 1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & -1 \end{pmatrix} \quad (2)$$

one-against-all all-pairs

ECOC implementations with both the Hamming and exponential distances were considered. The best results were obtained with ECOC when M was the one-against-all matrix and the decoding was performed with the exponential function.

In all our experiments we used a SVM [16] with the Radial Basis Function (RBF) kernel.⁴ In this setting there are two hyper-parameters which need to be optimized. The first is C , which is the tradeoff between the margin term and the training error penalty, and the second is σ , the width of the RBF kernel. Since we are decomposing our ternary problem in to three binary classification problems there are three binary classifiers whose hyper-parameters need to be determined. Therefore, overall we have six hyper-parameters to select. The optimization protocol we used is detailed next.

B. Performance Evaluation and Optimization Protocol

Based on the training set $S = \{(x_i, y_i)\}_{i=1}^m$ our goal is to estimate the performance (error rate) of our multi-category classification algorithm. Clearly, this error rate is critically dependent on the optimization routine used for determining the hyper-parameters C and σ of each of the (three) binary classifiers involved. Our protocol is based on standard n -fold cross-validation (see, e.g., Sec. 9.6.2 in [21]). The data is divided randomly into n equal parts (in our case, 5 parts), and in each fold one of the parts serves in turn as the test set, while the

⁴The RBF kernel classifier is the sign of $f(x) = \sum_{i=1}^n \alpha_i y_i K(x_i, x)$, where $K(x_i, x) = \exp\left(-\frac{\|x_i - x\|_2^2}{2\sigma^2}\right)$.

union of the other $n - 1$ parts is the training set. In particular, there can be no overlap between training and test sets.

In each train/test partition we perform an ‘‘internal’’ k -fold cross-validation to search for the best value of the hyper-parameter vector $\theta = (C^{(1)}, \sigma^{(1)}, C^{(2)}, \sigma^{(2)}, C^{(3)}, \sigma^{(3)})$. We perform this search over a grid of predetermined values. Specifically, let \mathcal{C} and Σ be predetermined sets of values of the hyper-parameters C and σ , respectively.⁵ In our implementation $|\mathcal{C}| = |\Sigma| = 10$ and therefore our hyper-parameter vector θ has 100 possible values for each individual binary classifier. For each fold (i.e., train/test partition) of the ‘external’ n -fold cross-validation we search for the best possible θ . Denoting by S' the training part of the fold, we search for θ using the ‘internal’ k -fold cross-validation as follows. For each possible value of θ , calculate its k -fold cross-validation error rate over S' . Then select the parameter θ that yields the lowest average error and use it to train the system for the external fold (with training set S').

For each of the k internal folds we must train $3 \times (|\mathcal{C}| \times |\Sigma|) = 300$ binary classifiers and test the performance of $(|\mathcal{C}| \times |\Sigma|)^3 = 10^6$ multi-class classifiers corresponding to all relevant hyper-parameter combinations. In our experiments we took $n = k = 5$ and therefore, overall our experiment included the generation (training) of 7500 binary classifiers and testing of 25×10^6 multi-class classifiers.

IV. EXPERIMENTS AND RESULTS

Dataset. We used a dataset consisting of 361 samples, of which 119 were classified by a pathologist as normal tissue, 102 as carcinoma in situ, and 140 as invasive ductal or lobular carcinoma. The samples were generated from slides of breast tissue biopsy, stained with hematoxylin and eosin. They were photographed using a Nikon Coolpix[®] 995 attached to a Nikon Eclipse[®] E600 at magnification of $\times 40$ to produce images with resolution of about 5μ per pixel. No calibration was made, and the camera was set to automatic exposure. The images were cropped to a region of interest of 760×570 pixels and compressed by lossy JPEG. The resulting images were again inspected by a pathologist to ensure that their quality was sufficient for diagnosis.

In order to enable future comparison by other researchers, we publish the entire raw dataset of images at <ftp://ftp.cs.technion.ac.il/pub/projects/medic-image/> as a service to the community.

Error rate. The average error rate we obtained is 6.6% with 0.8% standard error of the mean. The confusion matrix is given in Table I. Each row represents the probability (in percentage) of prediction given the true state. Note that each row sums up to 100%. For example, if the true diagnosis is carcinoma in situ, on average 4.7% of the time the classifier will diagnose these patients as normal (healthy).

ROC Curves. It is important to note that these results were derived by allocating an equal weight to each type of error. However, it is clearly the case that the clinical importance of different errors is not equal. False-positives (healthy classified as ill) are less dangerous than false-negatives (ill classified as healthy).

In order to present the trade-off between these errors, Receiver Operating Characteristic (ROC) curves were generated for the three binary problems⁶, as displayed in Figure 3. Each point (u, v) along this curve represents

⁵In our implementation \mathcal{C} consisted of 10 equally spaced numbers in $[1, 1000]$ and Σ consisted of 10 equally spaced numbers in $[100, 10000]$. These intervals were selected based on initial experimentation.

⁶ROC curves are only defined for binary problems.

True vs. Pred.	Normal	In situ	Invasive
Normal	92.6±1.1	6.5±1.2	0.9±0.9
In situ	4.7±1.3	93.4±1.4	1.9±1.2
Invasive	3.9±1.7	1.4±0.9	94.7±2.2

TABLE I

THE CONFUSION MATRIX; EACH ENTRY SPECIFIES THE MEAN AND THE STANDARD DEVIATION OF THE MEAN. FOR EXAMPLE, THE SECOND ENTRY OF THE FIRST ROW SHOWS WHICH PERCENTAGE OF NORMAL PATIENTS WILL BE DIAGNOSED AS HAVING CARCINOMA IN SITU.

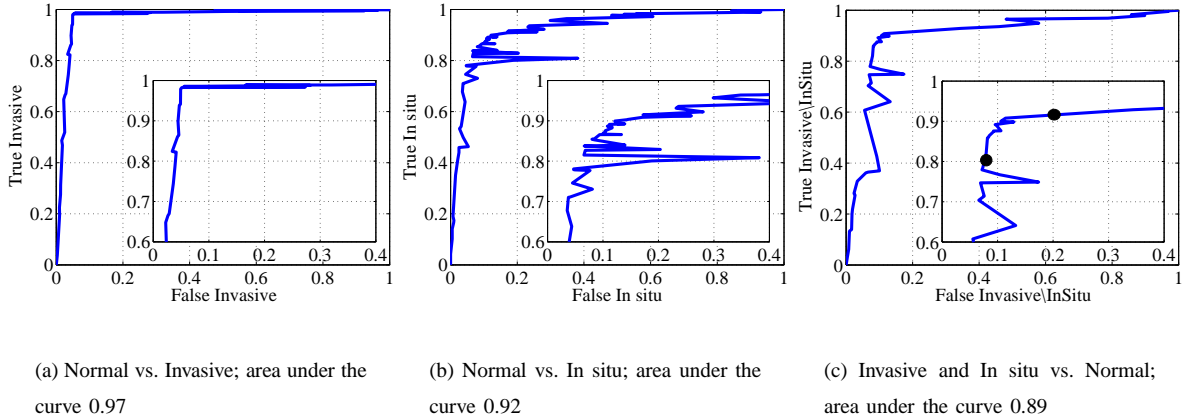


Fig. 3. The ROC curves for the three binary subproblems. Each sub-figure zooms into the top right part of the ROC curve appearing in the box $[0, 0.4] \times [0.6, 1]$

a confusion matrix where u is the estimated accuracy of correct classification to ‘Normal’ and v is the estimated accuracy of correct classification into the other, non-normal class (‘Invasive’ in (a); ‘In Situ’ in (b); and ‘Invasive’ or ‘In Situ’ in (c)).

These curves were generated as follows. The optimal values of the hyper-parameters $\{\sigma^{(1)}, \sigma^{(2)}, \sigma^{(3)}\}$ were fixed and the hyper-parameter C was separated into two distinct hyper-parameters C_+, C_- which penalize differently the training errors of the positively and negatively labeled patterns respectively. It is up to the user of the system to determine the desired point based on medical, legal and financial considerations.

Consider for example figure 3(c). In the zoomed sub-figure the two dots indicate two possible choices on the ROC curve: the left point will result in $(u, v) = (7.6\%, 80\%)$ and the right point in $(u, v) = (20\%, 91.6\%)$.

Decision with rejection. A desired feature in any computerized diagnosis system is *classification with rejection*, where we expect the system to only provide definitive diagnosis with a sufficiently high confidence. In cases where the system is not confident in the verdict it should abstain. When the system ‘rejects’ such a borderline pattern the pattern is relegated to an expert.

The rejection was performed by setting a threshold t . Given a training and test set, the training was performed with all the patterns of the training set. On the other hand, the error on the test set was calculated on all the

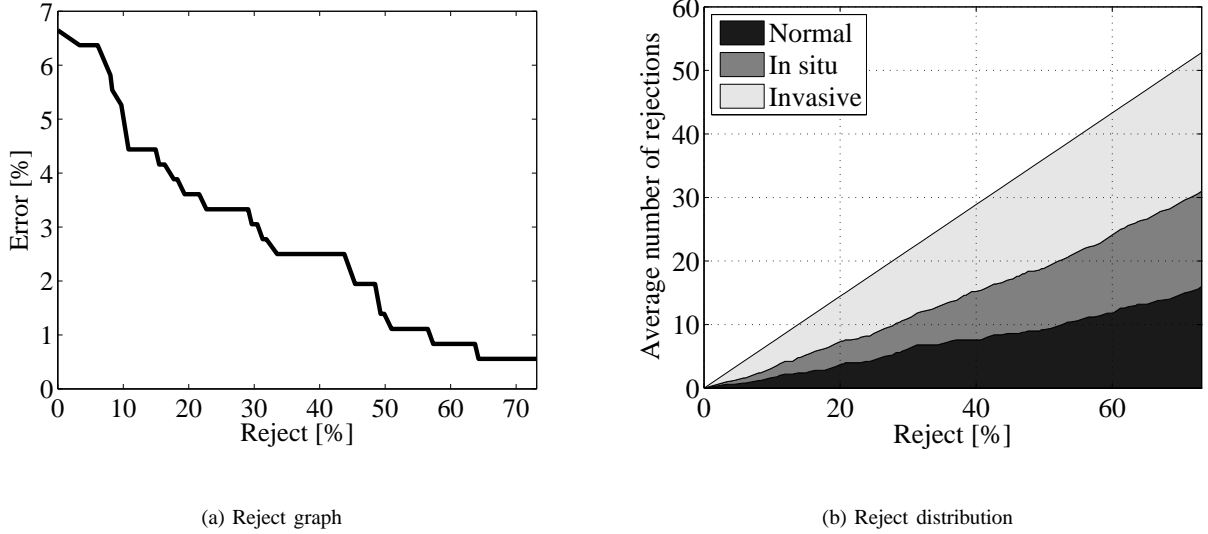


Fig. 4. The reject graph and the distribution of the rejected patterns; (a) shows the error rate achieved by rejecting a certain fraction of samples; (b) shows how many samples from each class are rejected..

patterns which were not rejected. Patterns were rejected if they met the following criterion:

- For each pattern x , the distance vector d , whose components are detailed in (1), is normalized so that $\|d\|_2 = 1$.
- Let $d_{[1]}, d_{[2]}, d_{[3]}$ denote the elements of the vector d sorted in non-descending order, namely $d_{[1]} \leq d_{[2]} \leq d_{[3]}$.
- If $d_{[2]} - d_{[1]} > t$, where t is the threshold, then the pattern is rejected.

The meaning of the criterion is that if the two smallest distances to codewords are close, according to threshold t , the pattern is rejected.

Indeed the improvement of accuracy is depicted in Figure 4(a). In this graph we see the trade-off between overall 5-fold cross-validation average error obtained versus the percent of test patterns that were rejected by the system. This trade-off curve shows that very low error rates can be obtained if we are willing to reject a fraction of the samples. For example, we can achieve an error rate close to 3.6% by relegating 20% of the harder decision tasks to a pathologist.

Figure 4(b) depicts the class distribution of the rejected patterns as a function of the rejection fraction. We see that this distribution is relatively uniform (as is the a-priori distribution of the classes in our dataset). Thus, there is no distinct propensity for rejecting a certain type of patterns.

V. COMPARISON TO PREVIOUS WORK

Substantial efforts have been recently devoted to different aspects of automated breast cancer diagnosis from biopsy data. However, with the exception of [11], we are not aware of any work to which we can reasonably compare our work. A major obstacle for such comparisons is the diversity of image types and magnifications used. Second, many studies analyze biopsy breast cancer images, but do not attempt statistical learning and

classification of those images. A third obstacle is that rigorous performance assessment criteria are often not adhered to (e.g., overfitting is likely to occur). Finally, we note that datasets are often not made publicly available thus precluding the possibility for a fair comparison.⁷

Some of the papers mentioned below consider substantially different images. There are two sources of these differences. First, different studies used images stained using different methods. Besides the hematoxylin and eosin staining considered here, other researches used slides stained with Feulgen, Papanicolaou, and immuno-labeled for estrogen or progesterone receptors. Second, slides are viewed at different magnifications, varying from $\times 40$ in our work to $\times 1000$ in works dealing with the inner structure of nuclei.

Another class of work that cannot be quantitatively compared with our results is work that does not provide any classification results. Such research falls basically into two categories: work focusing on nuclei segmentation [1], [2], [22] and work that only shows statistically significant correlation between certain features and cancer grades [3], [4], [5]. See also the review [6].

In [7] hematoxylin and eosin stained images at the magnification of $\times 100$ are used to differentiate between mastopathy and carcinoma. The images are manually binarized and used to calculate features based in part on areas and perimeters of connected components, which is the main point of similarity between that paper and ours. The images are then classified by a neural network using these features. The authors report accuracies of up to 98%, but there are some methodological problems: the authors use only 40 examples, and it would seem that they train *and* test on the same 40 examples. It is likely that their results display a high degree of overfitting.

In the recent work [23] the authors consider hematoxylin and eosin stained images taken at $\times 10$ magnification. The images are analyzed using several levels of wavelet decomposition, and these features are used to classify images as belonging to one of three classes, as in our work. The classification is performed using either discriminant analysis or a neural network. The best results are shown using only 2 levels of Haar wavelet and lead to 87.78% correct classification, without cross-validation.

The work described in [8] is one of the earliest papers on automated breast cancer diagnosis (and prognosis). The authors developed a program that allows for manual segmentation of nuclei on $\times 900$ magnification images of Fine Needle Aspiration biopsies using Papanicolaou stain - a very different procedure from our images. Using a classifier developed by one of the authors, which is a combination of a linear classifier and a decision tree, they predict accuracy of 97.5% estimated by cross-validation. These are very impressive results, but we should stress again that their data is very different from ours, and the system is not fully automatic.

Finally, [9] is the work most similar to ours in several respects. The images are of biopsies, stained with hematoxylin and eosin and photographed at the $\times 200$ magnification. The images were segmented to delineate duct and lumen boundaries by “several interacting expert systems” with some human intervention at the initial stages of the segmentation process. The authors suggest several features, mostly based on the geometry (areas, lengths, mutual distances) of glands and lumens. The authors then calculate “patient scores” in an undisclosed

⁷As mentioned above we made our data publicly available at <ftp://ftp.cs.technion.ac.il/pub/projects/medic-image/>.

fashion, and based on that use a simple Bayesian classifier to distinguish between ductal hyperplasia and ductal carcinoma in situ—a problem which is more difficult than the classification problem we define here. A classification accuracy of 81% is reported.

We should again mention the survey [10], which has a much more extensive bibliography. Notice also, that our work does not fall neatly into the classification established in this survey: our features are somewhere in between morphological, fractal-based, and topological features, and SVMs are not even mentioned anywhere in that survey.

Finally, a previous paper by a subset of the present authors addressed the same classification problem [11]. The emphasis in [11] was on the establishment of a density based morphological feature extraction procedure, contrary to the more generic approach taken in this paper. The classification results established here are slightly superior to those presented in [11].

VI. DISCUSSION

We have shown in this paper that our features, in spite of their simplicity, perform at least on a par with highly complex and specialized features [11]. A possible explanation for this may be that these features provide an insight into the spatial organization of the objects that are visible on a breast biopsy slide.

In Figure 5 we can see that generally, connected components of images with different thresholds correspond to different levels of the hierarchical structure of the breast tissue.

We note that we constructed these features without any consultation with a pathologist. While a direct interaction with a pathologist can potentially improve the results, it is not clear a priori that features that perform well for a human specialist are necessarily the best for a generic learning algorithm. Since our features were not tailored to the specific problem of breast cancer diagnosis we expect that they will perform well in other problems that involve spatial organization in gray-level images.

When considering our results (6.6% error rate) we should also consider the question of a “gold standard”. The most reasonable option is a comparison with a human pathologist’s performance. While it is clear that humans are also error-prone, there are no definitive results on error rates in histopathology. The review [24] mentions results of several studies and audits, with 3.4%–4.0% rate of “serious diagnostic errors” and 1.1%–1.4% rate of errors that affected patient management. A misclassification in the framework of this work probably qualifies as the latter.

The authors of [25] report that pathology second opinions altered surgical therapy in 7.8% of cases. This is much higher than the rates mentioned earlier, probably because of greater disparity of expertise of the pathologists who wrote the primary report, and the pathologists specializing in breast cancer. However we choose to treat these reports, it is clear that the baseline for this problem is not zero error rate, but rather is above 1% in the best case, and the most reasonable guess is about 3.5%.

The goal of 1% error rate seems over-ambitious even for a human pathologist, let alone an automatic system. In our case, to achieve it we will need to reject about one-half of the samples, which is impractical. On the other hand, the rate of 3.5% is achievable with 20% rejection.

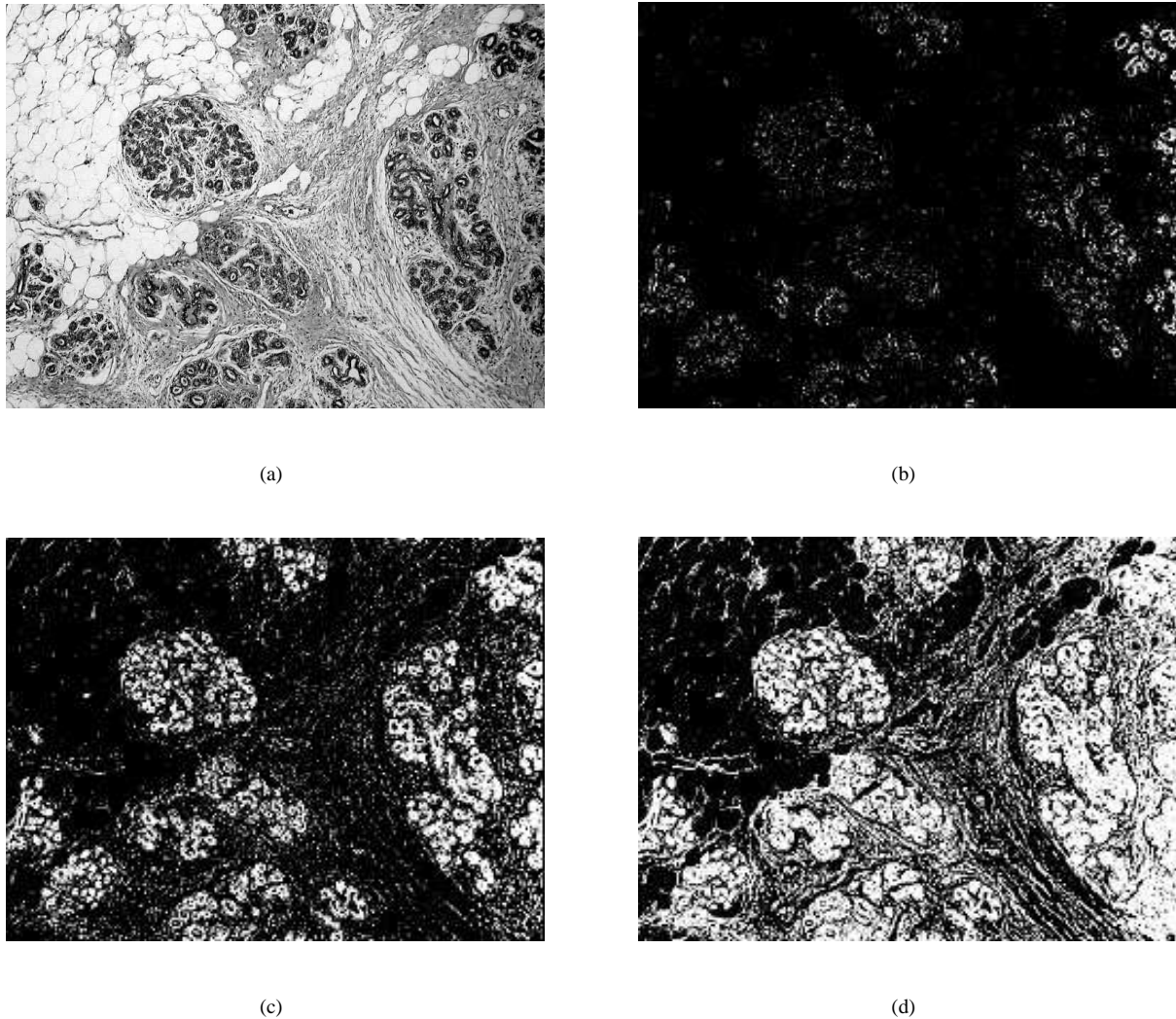


Fig. 5. Features capture spatial organization. (a) a portion of a gray-level image; (b), (c), (d) binary images corresponding to different threshold levels.

The system we present here cannot be used in a hospital setting at the current levels of performance, at least not as a primary method of diagnosis. However, since the system is fully automatic and works with very low resolution images, it can function as a high-speed preprocessor for a more complex (hence, higher performing) diagnostic system. From the relevant works we can see that noticeably better classification results can be obtained from magnifications $\times 400$ upwards. Our system thus processes 100 times less data than these systems and can be expected to be about 100 times faster, even without taking into account the need for human intervention in other systems.

Our work may be improved in several ways. First, it is likely that the addition of other spatial characteristics of the images may improve the representation. Other features we have considered and that may be useful for similar problems are the number of connected components at each brightness level and the histograms of perimeters of connected components. Our experiments with using perimeters for the features showed performance similar

to that of areas. Using perimeters and areas together did not improve the recognition rate.

We believe that significant improvements in accuracy can be obtained by considering an ensemble of classifiers each working on a different magnification level. The aggregation of results from these classifiers can be made using simple majority voting or by using more sophisticated ensemble techniques.

Our results on ‘classification with rejection’ indicate that this approach is plausible and effective. However, these results were obtained using a straightforward approach. It is likely that these results can be improved using more sophisticated optimization techniques. Also, the incorporation of rejection and misclassification costs into the SVM algorithm is less “natural” than in algorithms which produce probabilities. Therefore, the application of an algorithm such as kernel multinomial logistic regression may improve the performance.

Finally, we note that while the dataset we used is not considered small (compared to related work) it would be essential to test the system on larger sets before any attempt is made to use the system in a clinical setting.

ACKNOWLEDGMENTS

We thank Dr. Roman Goldenberg for the permission to use his software. This research was supported by the Ministry of Science infrastructural grant No. 01-01-01499.

REFERENCES

- [1] L. Latson, B. Sebek, and K. Powell, “Automated cell nuclear segmentation in color images of hematoxylin and eosin-stained breast biopsy,” *Anal. Quant. Cytol. Histol.*, vol. 26, no. 5, pp. 321–331, 2003.
- [2] H. Wu, J. Barba, and J. Gil, “Iterative thresholding for segmentation of cells from noisy images,” *J. Microsc.*, vol. 197, no. 3, pp. 296–304, 2000.
- [3] G. Klorin and R. Keren, “Ploidy and nuclear area as a predictive factor of histologic grade in primary breast cancer,” *Anal. Quant. Cytol. Histol.*, vol. 25, no. 5, pp. 277–280, 2003.
- [4] A. Ruiz, S. Almenar, M. Cerdá, J. Hidalgo, A. Puchades, and A. Llombart-Bosch, “Ductal carcinoma in situ of the breast: a comparative analysis of histology, nuclear area, ploidy, and neovascularization provides differentiation between low- and high-grade tumors,” *Breast J.*, vol. 8, no. 3, pp. 139–144, 2002.
- [5] A. Einstein, H. Wu, and J. Gil, “Self-affinity and lacunarity of chromatin texture in benign and malignant breast epithelial cell nuclei,” *Phys. Rev. Lett.*, vol. 80, no. 2, pp. 397–400, 1998.
- [6] J. Gil, H. Wu, and B. Wang, “Image analysis and morphometry in the diagnosis of breast cancer,” *Microscopy Research and Technique*, 2002.
- [7] T. Mattfeldt, H. Gottfried, V. Schmidt, and H. Kestler, “Classification of spatial textures in benign and cancerous glandular tissues by stereology and stochastic geometry using artificial neural networks,” *J. Microscopy*, vol. 198, no. 2, pp. 143–158, 2000.
- [8] O. Mangasarian, W. Street, and W. Wolberg, “Breast cancer diagnosis and prognosis via linear programming,” *Operations Research*, vol. 43, no. 4, pp. 570–575, 1995.
- [9] N. Anderson, P. Hamilton, P. Bartels, D. Thompson, R. Montironi, and J. Sloan, “Computerized scene segmentation for the discrimination of architectural features in ductal proliferative lesions of the breast,” *J. Pathology*, vol. 181, no. 4, pp. 374–380, 1997.
- [10] Ç. Demir and B. Yener, “Automated cancer diagnosis based on histopathological images: a systematic survey,” Rensselaer Polytechnic Institute, CS, Tech. Rep. TR-05-09, 2005.
- [11] I. Zingman, R. Meir, and R. El-Yaniv, “Size-density spectra and their application to image classification,” Electrical Engineering Dept., Technion, Tech. Rep. CCIT-566, 2005, <http://www.ee.technion.ac.il/~rmeir/ZingmanMeirElYaniv.pdf>.
- [12] G. Fung and O. L. Mangasarian, “Data selection for support vector machines classifiers,” in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, pp. 64–70.
- [13] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, “Use of the zero norm with linear models and kernel methods,” *The Journal of Machine Learning Research*, vol. 3, pp. 1439–1461, March 2003.

- [14] D. Peleg and R. Meir, "A feature selection algorithm based on the global minimization of a generalization error bound," in *Advances in Neural Information Processing Systems*, 2004, pp. 1065–1072.
- [15] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2002.
- [16] N. Cristianini and J. Shawe-Taylor, *An introduction to Support Vector Machines and other Kernel-based learning methods*, 2nd ed. John Wiley & Sons, 2001.
- [17] T. Sejnowski and C. Rosenberg, "Parallel networks that learn to pronounce English text," *Journal of Complex Systems*, vol. 1, no. 1, pp. 145–168, 1987.
- [18] T. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *J. of Artificial Intelligence Research (JAIR)*, vol. 2, pp. 263–286, 1995.
- [19] N. García-Pedrajas and D. Ortiz-Boyer, "Improving multiclass pattern recognition by the combination of two strategies," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 1001–1006, 2006.
- [20] E. Allwein, R. Schapire, and Y. Singer, "Reducing multiclass to binary: a unifying approach for margin classifiers," *J. of Machine Learning Research (JMLR)*, vol. 1, pp. 113–141, 2000.
- [21] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. Cambridge University Press, 2000.
- [22] F. Schnorrenberg, C. Pattichis, K. Kyriacou, and C. Schizas, "Computer-aided detection of breast cancer nuclei," *IEEE Trans. Inform. Technol. Biomed.*, vol. 1, no. 2, pp. 128–140, 1997.
- [23] H.-G. Hwang, H.-J. Choi, B.-I. Lee, H.-K. Yoon, S.-H. Nam, and H.-K. Choi, "Multi-resolution wavelet-transformed image analysis of histological sections of breast carcinomas," *Cellular Oncology*, vol. 27, pp. 237–244, 2005.
- [24] A. D. Ramsay, "Errors in histopathology reporting: detection and avoidance," *Histopathology*, vol. 34, pp. 481–490, 1999.
- [25] V. L. Staradub, K. A. Messenger, N. Hao, E. L. Wiley, and M. Morrow, "Changes in breast cancer therapy because of pathology second opinions," *Annals of Surgical Oncology*, vol. 9, pp. 982–987, 2002.