Computerized Classification of Intraductal Breast Lesions Using Histopathological Images

M. Murat Dundar*, *Member, IEEE*, Sunil Badve, Gokhan Bilgin, Vikas Raykar, Rohit Jain, Olcay Sertel, and Metin N. Gurcan, *Senior Member, IEEE*

***Abstract*—In the diagnosis of preinvasive breast cancer, some of the intraductal proliferations pose a special challenge. The continuum of intraductal breast lesions includes the usual ductal hyperplasia (UDH), atypical ductal hyperplasia (ADH), and ductal carcinoma in situ (DCIS). The current standard of care is to perform percutaneous needle biopsies for diagnosis of palpable and image-detected breast abnormalities. UDH is considered benign and patients diagnosed UDH undergo routine follow-up, whereas ADH and DCIS are considered *actionable* and patients diagnosed with these two subtypes get additional surgical procedures. About 250 000 new cases of intraductal breast lesions are diagnosed every year. A conservative estimate would suggest that at least 50% of these patients are needlessly undergoing unnecessary surgeries. Thus, improvement in the diagnostic reproducibility and accuracy is critically important for effective clinical management of these patients. In this study, a prototype system for automatically classifying breast microscopic tissues to distinguish between UDH and actionable subtypes (ADH and DCIS) is introduced. This system automatically evaluates digitized slides of tissues for certain cytological criteria and classifies the tissues based on the quantitative features derived from the images. The system is trained using a total of 327 regions of interest (ROIs) collected across 62 patient cases and tested with a sequestered set of 149 ROIs collected across 33 patient cases. An overall accuracy of 87.9% is achieved on the entire test data. The test accuracy of 84.6% is obtained with borderline cases (26 of the 33 test cases) only, when compared against the diagnostic accuracies of nine pathologists on the same set (81.2% average), indicates that the system is highly competitive with the expert pathologists as a stand-alone diagnostic tool and has a great potential in improving diagnostic accuracy and reproducibility when used as a "second reader" in conjunction with the pathologists.**

***Index Terms*—Cell segmentation, computer-aided diagnosis, histopathological image analysis, intraductal breast lesions, multiple instance learning.**

Manuscript received April 30, 2010; revised November 3, 2010 and December 29, 2010; accepted December 30, 2010. Date of publication February 4, 2011; date of current version June 17, 2011. This work was supported by the National Cancer Institute under Award R01CA134451. *Asterisk indicates corresponding author.*

*M. M. Dundar is with the Department of Computer and Information Science, Indiana University–Purdue University, Indianapolis, IN 46202 USA (e-mail: dundar@cs.iupui.edu).

S. Badve is with the Department of Pathology, Indiana University, Indianapolis, IN 46202 USA (e-mail: sbadve@iupui.edu).

G. Bilgin is with the Department of Computer and Information Science, Indiana University–Purdue University, Indianapolis, IN 46202 USA (e-mail: gbilgin@cs.iupui.edu).

V. Raykar is with the Siemens Healthcare Malvern, PA 19355 USA (e-mail: vikas.raykar@siemens.com).

R. Jain is with the Department of Pathology, Indiana University, Indianapolis, IN 46202 USA (e-mail: rkjain@umail.iu.edu).

O. Sertel and M. N. Gurcan are with the Biomedical Informatics Department, The Ohio State University, Columbus, OH 43210 USA (e-mail: osertel@bmi.osu.edu; metin.gurcan@osumc.edu).

Digital Object Identifier 10.1109/TBME.2011.2110648

I. Introduction

A. Background

THE continuum of intraductal breast lesions, which encompasses the usual ductal hyperplasia (UDH), atypical ductal hyperplasia (ADH), and ductal carcinoma in situ (DCIS), are a group of cytologically and architecturally diverse proliferations, typically originating from the terminal duct-lobular unit and confined to the mammary duct lobular system [1]. These lesions are highly significant as they are associated with an increased risk of subsequent development of invasive breast carcinoma, albeit in greatly differing magnitudes. Clinical follow-up studies indicate that UDH, ADH, and DCIS are associated with 1.5, 4–5, and 8–10 times of increased risk, respectively, compared to the reference population for invasive carcinoma [2]. Data from recent immunophenotypic and molecular genetic studies support the notion that both ADH and all forms of DCIS represent intraepithelial neoplasias characterized by morphological changes that result from clonal alterations in genes and thus carry a risk of variable magnitudes for invasion and metastasis [1]. On the other hand, there is currently no evidence to classify UDH as a precursor lesion.

The total number of new cases of intraductal lesions diagnosed each year in U.S. is predicted to be about 250 000. The current standard of care is to perform percutaneous needle biopsies for diagnosis of palpable and image-detected breast abnormalities. Patients diagnosed UDH are advised to undergo routine follow-up, while those with ADH and DCIS are operated by excisional biopsy followed by possible other surgical and therapeutic procedures. Thus, depending on the results of the percutaneous biopsy, the management of patients diagnosed UDH and ADH/DCIS may significantly vary.

B. Diagnostic Accuracy and Reproducibility

The pathology diagnoses are typically made according to a set of criteria defined by the World Health Organization (WHO), using formalin fixed paraffin embedded tissue specimens, which are stained with a mixture of hematoxylin/eosin (H&E). Of note, no single criterion is absolute. Thus, subjective assessment and weighing the relative importance of each criterion are performed to categorize the lesions. This, as several studies have clearly demonstrated, results in poor interobserver agreement, particularly when standardized criteria (and group training) are not used [3]. While the standardized criteria are generally easy to identify for most lesions, there are borderline cases where it becomes difficult to determine with absolute certainty whether a lesion belongs to one subtype or the other. The relative

weightage given by a pathologist to each of the criteria is difficult to assess and leads to a diversity of opinions among consulting pathologists. In our preliminary studies, we assessed the level of concordance among nine academic and community pathologists in classifying intraductal lesions [4]. The overall interobserver agreement among the nine pathologists for diagnosing 81 borderline lesions was fair with a Kappa value of 0.35. Among the 33 lesions of UDH, 27 lesions of ADH, and 11 lesions of DCIS classified by maximum agreement, complete agreement was achieved only for six UDH and three DCIS cases.

C. Proposed Approach

The proposed system is designed with the clinical management of patients in mind. Patients diagnosed UDH on percutaneous biopsy undergo routine follow-up, whereas those diagnosed ADH or DCIS, i.e., *actionable* subtypes, get excisional biopsy. Once identified as actionable, determining the true subtype of a lesion on percutaneous biopsy does not change initial patient management much, whereas misclassifying UDH as an actionable lesion or vice versa may have severe consequences. When a UDH case is misclassified as an actionable lesion, patients undergo unnecessary surgical operations, which may in turn cause additional complications and discomfort for the patient in addition to a considerable increase in cost. When an actionable lesion is misclassified as UDH, the patient gets undertreated. If the lesion later develops into invasive carcinoma, it may be too late for treatment by the time the patient is diagnosed cancer. To sum up, any improvement in the classification of UDH versus actionable subtypes on percutaneous biopsy will have two direct contributions on clinical patient management: 1) when UDH is more accurately distinguished from actionable cases on percutaneous biopsy, thousands of unnecessary excisional biopsies will get eliminated; and 2) when actionable cases are less often misclassified as UDH, possible undertreatment for thousands of patients will be prevented. Thus, the proposed system will make the most clinical impact when developed to address the binary classification of UDH versus actionable subtypes during the percutaneous biopsy stage.

Clinical impact aside, the binary classification approach is more feasible from the system training perspective as well. In order to train the system to perform multicategory classification, samples with reference standard from each subtype would be necessary. Since there is currently no known morphometric, immunohistochemical, or molecular features to distinguish ADH from low grade DCIS (LG-DCIS) [5], such a reference standard could not be established for ADH and certain types of DCIS. On the other hand, with the help of special immunostains most UDH can be identified from actionable subtypes. Thus, the reference standard required for the training of the classifier for classifying UDH versus actionable lesions can be established with a reasonable effort, whereas the same cannot be said for the reference standard required for multicategory classification.

The proposed system is developed using a dataset of 327 regions of interest (ROIs) obtained from 62 patient cases representing three different lesion subtypes. A clustering algorithm is implemented to identify regions of cells in the H&E-stained breast microscopic tissues. This was followed by a watershed-based segmentation algorithm, which identifies individual cells. The segmented cells are used to derive size, shape, and intensity-based features characterizing each ROI. These features along with the reference standard available at the slide level are used to train a binary classifier. The system is tested using an independent set of 33 cases with a total of 149 ROIs. The stand-alone diagnostic performance of the developed system is compared against nine expert pathologists on borderline cases. Methods implemented for data collection, clustering, segmentation, feature extraction, and classification are presented in Section II. Results are presented and discussed in Section III. Conclusion and future research directions are provided in Section IV.

II. Methods

A. Image Dataset

Patient cases in the study database are collected, on a retrospective basis, from the Clarian Pathology Lab, Indianapolis, IN, according to the approved Institutional Review Board protocol for this study. H&E-stained serial sections of each tissue specimen are examined by a surgical pathologist using the WHO published criteria to confirm the initial diagnosis associated with that particular archival tissue specimen. Cases evaluated this way are grouped into two categories by the pathologist: *well defined* and *borderline*. All cases based on the pathologist's own judgement that require careful analysis before reaching diagnosis are considered borderline irrespective of the initial diagnosis available in the pathology report. Well-defined cases are assigned to corresponding subtypes (UDH or DCIS) without further evaluation. For borderline cases, truthing is performed by a panel involving nine board-certified academic and community pathologists. Since most UDH is positive to high molecular weight keratin (HMWK), immunostained serial sections of each tissue specimen are obtained using ADH-5 multiplex immunohistochemistry staining, Biocare Medical LLC (Concord, CA). Lesions on each case are marked and the pathologists are instructed to only evaluate the marked area on the immunostained specimens as per their usual diagnostic criteria and assign diagnoses of UDH, ADH, and DCIS to each case. Reference standard, for each borderline case evaluated by the panel, is established by maximum agreement.

Through this procedure 20 well-defined and 12 borderline DCIS, 24 ADH, and 39 UDH cases are identified. After regrouping ADH and DCIS cases under the *actionable* category, the study database contains 39 UDH and 56 actionable cases. Once the study database is constructed, a research associate, supervised by a pathologist, manually identified ROIs on each slide. These are generally regions that show proliferation of cells. Of the total of 95 cases available, 33 cases (about 30%) containing 149 ROIs are randomly selected by stratified sampling and sequestered as the test set. The remaining 62 cases containing 327 ROIs are used for training. H&E-stained serial section of each tissue specimen is fixed on a scanning bed and digitized using Aperio (Vista, CA), ScanScope digitizer at 40× magnification. The resulting whole-slide images each with size up to 5 GB are stored in the Scanscope Virtual Slide format. A

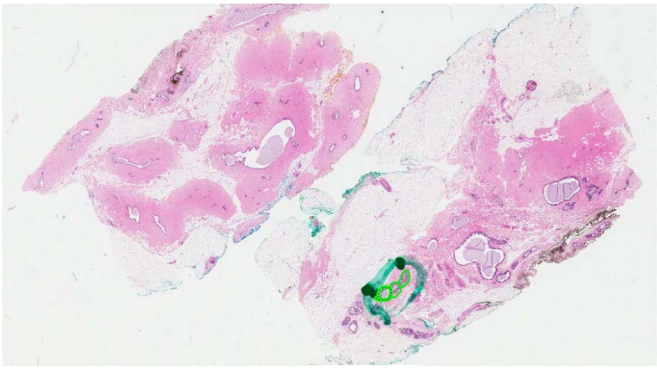


Fig. 1. Snapshot of the digitized scan of an H&E-stained specimen with three ROIs.

snapshot of a sample image is shown in Fig. 1 along with the ROIs identified on this slide.

B. Segmentation

H&E stain colors the basophilic structures consisting of nuclear and cytoplasmic regions with blue–purple hue, the protein-rich structures consisting of extracellular regions with hues of pink, and red blood cells and regions with necrosis with hues of red. Images of two sample ROIs, one for UDH and the other one for DCIS, are shown in Fig. 2(a) and (d).

Pathologists use cytological descriptors such as cell size, shape, composition, nuclear spacing, etc., evident in the H&E-stained tissue specimens for diagnosis. Therefore, cell segmentation would be the first step toward automated analysis of histopathological slides. This is implemented in two steps in this study. In the first step, cell regions are segmented by clustering the pixel data and in the second step segmented cell regions are further processed by a watershed-based segmentation algorithm to identify individual cells.

1) Cell Region Segmentation: The ROI images are first converted from the RGB color space to the La*b* color space. La*b* is a perceptually uniform color space, i.e., a change of the same amount in a color value produces a change of about the same visual importance. The La*b* color space also separates the luminance and the chrominance information such that L channel corresponds to illumination and a* and b* channels correspond to color-opponent dimensions.

Pixel data are modeled by a four-component Gaussian mixture model (GMM). One component is used for each of the following four cytological regions: cellular (nuclear and cytoplasmic), extracellular, regions with hues of red and illumina. The expectation maximization (EM) algorithm [6] is implemented using the a*, b* channels to estimate the parameters of the GMM model. The resulting mixture distribution is used to classify pixels into four categories. Those classified into the cellular component are further clustered in the L channel by dynamic thresholding [7] to eliminate blue–purple pixels with relatively less luminance. The remaining pixels are considered cell regions and images containing these regions are used in the next stage.

The EM algorithm is run only once with a dataset containing few million pixels, which are obtained by randomly sampling 10 000 pixels from each ROI image in the dataset. The same GMM model is used for segmenting all ROI images without rerunning EM for each image. Since each slide might contain several ROIs, estimating the distribution off-line and using it across all the ROI images saves significant online execution time. Fig. 2(b) and (e) shows the segmentation maps of the cell regions in sample ROI images associated with a DCIS and a UDH case, respectively.

2) Individual Cell Segmentation: Segmentation maps of cell regions obtained in the previous part are converted to gray-level images before they are used in this stage. Since most segmented regions contain multiple overlapping cells with cells only vaguely defined due to the presence of holes inside them, connected components in these images do not necessarily represent individual cells. These images are first preprocessed using hole filling and cleaning steps suggested in [8]. Overlapped cells result in blobs in the segmentation map. To separate these blobs properly so as to identify individual cells, we used a watershed algorithm based on immersion simulations [9]. In this approach, a gray-level image is considered a topographic relief where the gray level of a pixel is interpreted as its elevation. The water flows along a topographic relief following a certain descending path to eventually reach a catchment basin. Blobs in the image can be separated using this concept by identifying the limits of adjacent catchment basins and then separating them. The lines separating catchment basins are called *watersheds*. The implementation of this algorithm is available through the ImageJ platform [10]. This specific implementation uses the local maxima in the Euclidean distance map as seed points. These seed points are dilated as far as possible either until the edge of the particle or the edge of the region of another growing seed point is reached. We used the MATLAB (Natick, MA) interface to ImageJ to use this implementation alongside our own algorithms. Once all catchment basins are identified and separated with this approach, the region defined by a catchment basin is considered a cell region. These are easily identified in the resulting image by identifying connected components. Fig. 2(c) and (f) shows the segmentation maps of the individual cells in sample ROI images associated with a DCIS and a UDH case, respectively.

C. Feature Extraction

This section discusses feature extraction methods implemented for quantitative characterization of ROIs in H&E-stained slides. Pathologists tend to rely heavily on morphological features such as cell size, shape, and nucleoli appearance. Three different sets of statistical features are computed for each ROI to model these histological descriptors. Perimeter is used for cell size, the ratio of major to minor axis of the best-fitting ellipse is used for cell shape, and the mean of the gray-level intensity is used for nucleoli appearance. For comparison, summaries of the set of features used by the pathologists for classifying intraductal breast lesions and those employed in the computerized system are reported in Tables I and II, respectively. There

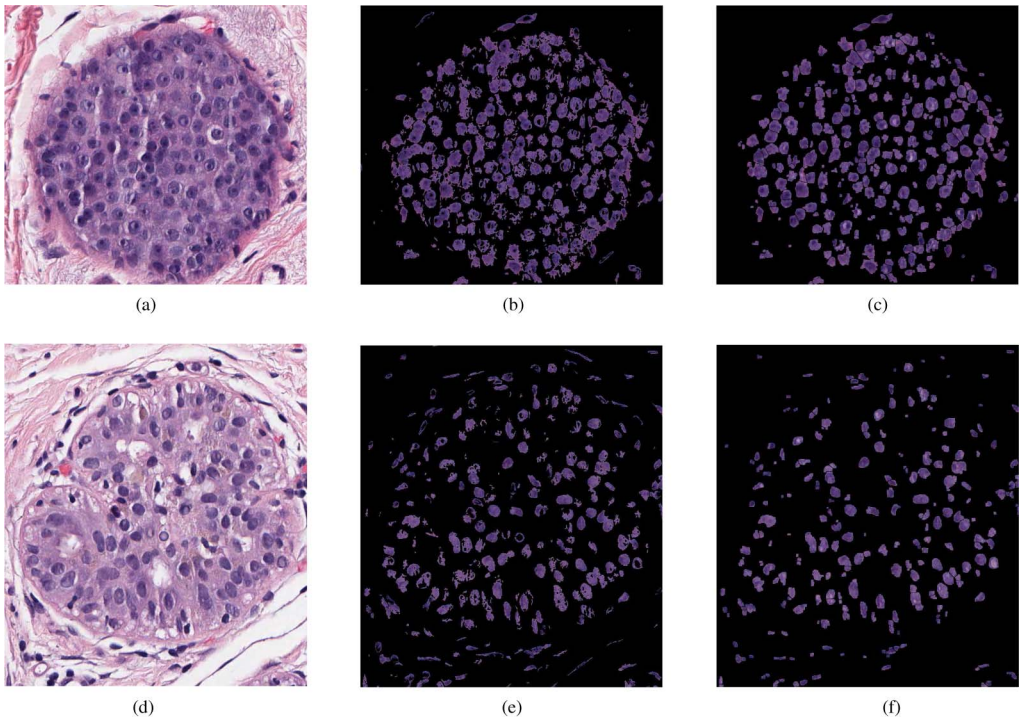


Fig. 2. Segmentation of the cells for two sample ROIs obtained from borderline UDH and DCIS cases, respectively. (a) DCIS—original image. (b) DCIS—segmented cell regions. (c) DCIS—segmented cells. (d) UDH—original Image. (e) UDH—segmented cell regions. (f) UDH—segmented cells.

TABLE I
QUALITATIVE FEATURES USED BY PATHOLOGISTS FOR DIAGNOSIS

Histological Features	Description
Cell Size	Small for UDH, small or medium-sized for ADH and low-grade DCIS, large for high-grade DCIS
Cell Shape	Ovoid and mixed for UDH, monotonous for ADH and low-grade DCIS, large and pleomorphic for high-grade DCIS
Nucleoli	Indistinct for UDH, single and small for ADH and low-grade DCIS, prominent and enlarged for high-grade DCIS

TABLE II
MORPHOLOGICAL AND PIXEL INTENSITY FEATURES CONSIDERED IN COMPUTERIZED DIAGNOSIS

Features	Description
Perimeter	The perimeter of each connected component is used to characterize cell size.
Ratio of major to minor axis	The ratio of major to minor axis for the best fitting ellipse corresponding to each connected component is used to characterize cell shape (ovoid vs. circular).
Mean of the gray level intensity	The mean of the gray level intensity for each connected component is used to characterize nucleoli appearance. Nucleoli, if prominent and/or large, appears darker than cytoplasmic regions in the cell.

are several hundred connected components present in each segmented image. Once the perimeter, the ratio of major to minor axis, and the mean of the gray-level intensity are computed for each connected component identified in an ROI, statistical features involving the mean, standard deviation, median, and mode are computed to obtain features at the ROI level. Thus, each ROI is characterized by a total of 12 features (3×4).

D. Classifier Training

Each slide contains multiple ROIs and a positive (actionable) diagnosis is confirmed when *at least one of the ROIs* in the slide is identified as positive. For a negative diagnosis (UDH), the pathologist has to rule out the possibility of *each and every ROI* being actionable. The objective here is to develop a classifier to optimize classification accuracy at the slide level. Traditional supervised training techniques which are trained to optimize classifier performance at the instance level yield suboptimal performance in this problem.

The problem of learning with multiple instances (MIL) was first defined in the context of a drug activity prediction application [11]. In our earlier work, we have developed an MIL approach based on the convex-hull idea and performed experimental studies on two different computer-assisted detection applications, which demonstrated that our approach significantly improves detection accuracy when compared to other MIL techniques proposed in the literature [12]. Since this approach was developed for problems where only positive samples are characterized by multiple instances, it was not directly applicable to the histopathological classification problem where negative cases also contain multiple instances. In a more recent work, we have proposed an extension of this approach by defining a pair of asymmetric loss functions for positive and negative samples in a large-margin framework and presented results on a dataset containing 40 histopathology slides (20 UDH and 20 DCIS) that showed competitive performance with the state of the art [13].

In what follows, we first provide a brief overview of the large margin principle [14] and then briefly discuss our earlier work on MIL in [13]. The words *sample* and *instance* refer to a pathology slide and ROI, respectively. As such, slide/sample and ROI/instance are used interchangeably throughout this text.

1) Margin Maximization With Single Instance Per Sample: Notation—When each sample is characterized by a single instance, the following notation holds. $D = \{x_i, y_i\}_{i=1}^{N}$ denotes a training dataset with N samples where $x_i \in \mathcal{X} = \Re^d$ is an instance (d-dimensional feature vector) characterizing sample i and $y_i \in \mathcal{Y} = \{\pm 1\}$ is the corresponding known label. In the large margin approach, the classifier function $f(x) = w \cdot x + w_0$ is optimized by solving the following optimization problem with respect to w and w_0:

$$\mathcal{J}(w, w_0) = \Phi(w) + C \sum_{i=1}^{N} (1 - y_i(w \cdot x_i + w_0))_+ \tag{1}$$

where $(\cdot)_+ = \max(0, \cdot)$ represents the hinge loss function, and C is the cost preassigned to the misclassification associated with x_i.

2) Margin Maximization With Multiple Instances Per Sample: Notation—In the MIL framework, we use the following notation:

$$D = \left\{ \left\{ x_i^j, y_i \right\}_{j=1}^{M^i} \right\}_{i=1}^{N} \tag{2}$$

where x_i^j is the feature vector characterizing the jth ROI in slide i, M^i is the number of ROIs in slide i, and $y_i \in \mathcal{Y} = \{\pm 1\}$ contains the label information, which is available at the slide level only.

The MIL extension of the problem in (1) can be obtained by defining different loss functions for positive and negative samples such that a positive sample is penalized only when all of its instances, i.e., all ROIs in that slide, are classified negative, whereas a negative sample is penalized when at least one of its instances is classified positive. Next, we define two new loss functions for positive and negative samples.

a) Loss function for negative samples: For negative cases, loss is incurred when at least one of the ROIs in a slide is classified as positive. Thus, we replace the hinge loss function used in (1) with its multivariable counterpart defined by

$$\max\left(0, e_i^1, \ldots, e_i^{M^i}\right) \tag{3}$$

where $e_i^j = 1 + (w \cdot x_i^j + w_o)$. This ensures that the loss incurred by a negative slide i is zero only if all ROIs in the slide are correctly classified as negative.

b) Loss function for positive cases: For positive cases, loss is incurred when all of the ROIs in a slide are classified as negative. In other words for correct diagnosis, it is sufficient to identify at least one ROI on a slide as positive. If a point that lies within the convex hull of all the instances of a sample is classified positive, this will indicate that at least one of the original ROIs is also classified positive. Let λ_i s.t. $0 \leq \lambda_i^j, e \cdot \lambda_i = 1$, be the vector containing the coefficients of the convex combination of all ROIs in slide i, and e be a vector of ones. Then, the feature vector characterizing slide i is defined by

$$\bar{x}_i = x_i^1 \lambda_i^1 + \cdots + x_i^{M_i} \lambda_i^{M_i} = X^i \cdot \lambda_i \tag{4}$$

where $X^i = [x_i^1 \ldots x_i^{M_i}]$ is the data matrix containing feature vectors of all ROIs within slide i. The loss function for a positive case in this new framework can be defined by

$$\left(1 - \left(w \cdot \left(X^i \cdot \lambda_i\right) + w_o\right)\right)_+ \tag{5}$$

which is a function of both convex-hull coefficients λ_i and classifier coefficients w. More detailed discussion on convex-hull characterization of samples with multiple instances can be found in our earlier work in [12]. With the new loss functions for positive and negative samples added to the problem, the large-margin formulation becomes

$$\begin{aligned} \min_{(w, w_0, \lambda^i)} \; & \Phi(w) + C_- \sum_{i \in \Omega^-} \max\left(0, e_i^1, \ldots, e_i^{M^i}\right) \\ & + C_+ \sum_{i \in \Omega^+} \left(1 - \left(w \cdot \left(X^i \cdot \lambda_i\right) + w_o\right)\right)_+ \\ \text{s.t. } & 0 \leq \lambda_i \\ & e \cdot \lambda_i = 1 \end{aligned} \tag{6}$$

where Ω^+ and Ω^- are the corresponding sets of indices for the positive and negative samples, respectively, and the two constraints are imposed to ensure that the feature vector characterizing a positive sample is always within the convex hull of its instances. This problem can be optimized by iterating between two convex subproblems in an alternating manner. The first subproblem optimizes for w and w_0, while λ_i are fixed. The second subproblem optimizes for λ_i while w and w_0 are fixed. Since both subproblems are convex, the objective function in (6) is guaranteed to decrease after each iteration. Thus, convergence can be established by defining termination criteria, based on the change in the value of the objective function between any two iterations.

III. Results

In this section, experiments are performed to validate the stand-alone performance of the developed system. To avoid numerical problems during optimization, it is a common preprocessing step to normalize each feature to between -1 and 1. The parameters of the classifiers are selected from a designated set of six different C_+ and C_- values by considering all possible pairs ($6 \times 6 = 36$ pairs) and selecting the pair that optimizes the leave-one-slide-out (LOSO) cross-validation performance of the classifier. LOSO cross-validation splits the training dataset into k folds, where k is equivalent to the number of slides. At each stage, one slide is left out as validation data, i.e., all ROIs for that slide are removed from the training data as validation data, and the classifier is trained with the ROIs of the remaining $k - 1$ slides and tested on the ROIs of the left-out slide. This process is repeated until all k slides are used for validation and the probabilities of all ROIs being positive are obtained. There is only one slide for each patient in our study database. Thus, each fold only contains data from a unique patient. In this regard, the LOSO approach used in this study is the same as leave-one-patient-out

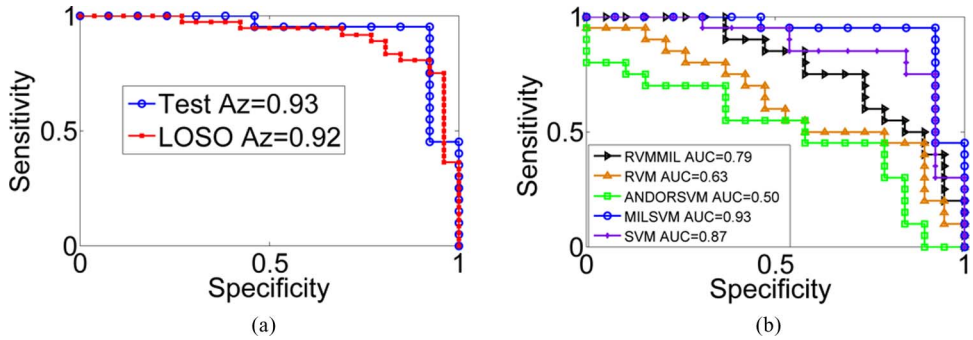


Fig. 3. (a) ROC curves obtained for the training (62 cases by LOSO) and test sets (33 cases) using the proposed approach (MILSVM). (b) ROC curves comparing MILSVM with other techniques from the literature on the test set.

cross validation. The classifier performance is measured by the area under the receiver operating characteristic (ROC) curve, or the so-called Az value. The different operating points along the ROC curve are obtained by comparing the classification results with the reference standard available for each case at varying thresholds, θ, i.e., $f(x) >= \theta$. Once the pair of classifier parameters that maximizes the Az value is determined, the classifier is trained with the entire training dataset to obtain the classifier function $f(x) = w \cdot x + w_0$. The two features with the largest weights in the optimized weight vector w are the mean of the mean intensity and the median of the ratio of the major to minor axis of the best-fitting ellipse. The first one characterizes nucleoli prominence and the second one cell shape.

Fig. 3(a) shows the ROC curve obtained by LOSO on the training data along with the ROC curve obtained on the test data. Az values of 0.92 and 0.93 are achieved, respectively, on the training and test datasets.

Next, the classifier defined by the classification boundary $f(x) = 0$ is evaluated for each ROI on the test data. ROIs with $f(x) >= 0$ are classified positive, whereas those with $f(x) < 0$ are classified negative. Here, x represents the feature vector characterizing the ROI. Classification at the slide level is rendered by classifying slides with at least one positive ROI as actionable and those where all ROIs are negative as UDH. An overall accuracy of 87.9% (29/33) is achieved by the computerized system. We also compared the accuracy of the system against the accuracies of the nine pathologists. These are the same group of nine pathologists used to establish the reference standard for each case in the study database. However, this step is different than truthing in that pathologists were provided immunostained serial sections of the specimens so that an accurate reference standard could be obtained for each case. This time pathologists are only provided H&E-stained serial sections for evaluation and are instructed to classify each case as UDH, ADH, and DCIS. Cases classified ADH or DCIS by the pathologists are grouped together under the actionable category. The study with the pathologists is performed using only the borderline cases in the test set (26 out of 33). The classification accuracies of the nine pathologists are presented in Table III.

We believe that the accuracy of 87.9% achieved on all test cases and of 84.6% achieved on borderline cases by the computerized system is quite promising. The accuracy on borderline cases is better than the accuracy of the four pathologists and is in par with the accuracy achieved by three of the remaining five pathologists. When we take a closer look at the four cases misclassified by the computerized system, we see that three of them are UDH and one is actionable. However, these are cases for which pathologists did not have complete agreement either. One case is classified UDH by five pathologists and actionable by the remaining four. The other case is classified actionable by six pathologists and UDH by the remaining three. The other two cases are classified actionable by seven pathologists and UDH by the remaining two.

TABLE III
CLASSIFICATION ACCURACIES (%) OF THE NINE PATHOLOGISTS ON BORDERLINE CASES

	1	2	3	4	5	6	7	8	9	Avg.
Pathologists:	84.6	84.6	57.7	84.6	92.3	77.0	96.2	77.0	77.0	81.2

Finally, we compared the performance of the proposed MIL approach (MILSVM) with two other MIL techniques from the literature, namely the MIL versions of the relevance vector machine (MILRVM) [16] and AND-OR SVM [16]. Two traditional supervised training techniques, namely RVM [17] and linear SVM [15], are also included in this comparative analysis. All classifiers are tuned by LOSO using the training set and are validated with the test set. The ROC curves obtained for the five classifiers are shown in Fig. 3(b) along with the Az values achieved by each. The proposed approach achieves an Az value of 0.93. The second largest $Az = 0.87$ is achieved by SVM. Significance analysis between the two ROC curves using the approach in [18] yields a p-value of 0.048, which indicates that the improvement is statistically significant.

IV. CONCLUSION

In this study, a system for computerized analysis of breast microscopic tissues is developed for improved classification of intraductal breast lesions. The system is developed with 62 cases and tested on 33 cases. An overall accuracy of 87.9% is achieved on the entire test set involving seven well-defined and 26 borderline cases. An accuracy of 84.6% is recorded on borderline cases. This was slightly higher than the average accuracy of nine board-certified pathologists (81.2%) evaluated on the same set. We believe that the stand-alone classification performance of

the developed system is highly encouraging. However, before this system can be deployed in a clinical setting as a "second reader," its incremental value has to be assessed. Before such a study can be effectively carried out, the computerized system has to be improved such that its output is presented in a way that could easily be interpreted by the pathologists. This will involve developing intermediate models to map image features onto descriptors pathologists use for classification. Apart from making the output of the system more interpretable, we will also increase the size of the study database for more extensive testing of the developed system. Once these two goals are achieved, a reader study involving a panel of pathologists with varying levels of experience will be conducted to evaluate the incremental value of the computerized system as a second reader.

ACKNOWLEDGMENT

The project described was supported in part by Award Number R01CA134451 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute, or the National Institutes of Health.

REFERENCES

[1] F. A. Tavassoli and P. Devilee, *World Health Organization: Tumours of the Breast and Female Genital Organs (IARC WHO Classification of Tumours)*. Cedex, France: IARCPress-WHO, 2003.

[2] P. L. Fitzgibbons, D. L. Page, D. Weaver, A. D. Thor, D. C. Allred, G. M. Clark, S. G. Ruby, F. O'Malley, J. F. Simpson, J. L. Connolly, D. F. Hayes, S. B. Edge, A. Lichter, and S. J. Schnitt, "Prognostic factors in breast cancer," *Arch. Pathol. Lab. Med.*, vol. 124, no. 7, pp. 966–978, 2000.

[3] J. Rozai, "Borderline epithelial lesions of the breast," *Amer. J. Surg. Pathol.*, vol. 15, no. 3, pp. 209–221, 1991.

[4] R. Jain, R. Mehta, R. Dmitrov, L. Larsson, P. Musto, K. Hodges, T. Ulbright, E. Hattab, N. Agaram, M. Idrees, and S. Badve, "A typical ductal hyperplasia at 25 years-interobserver and intraobserver variability," *Mod. Pathol.*, vol. 23, no. 1, suppl. 1, pp. 53A:abstr. 229, 2010.

[5] D. Page, W. Dupont, L. Rogers, and M. Rados, "Borderline epithelial lesions of the breast," *Amer. J. Surg. Pathol.*, vol. 15, pp. 209–221, 1991.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm.," *J. R. Statist. Soc. Series B (Methodol.)*, vol. 39, no. 1, pp. 1–38, 1977. [Online]. Available: http://dx.doi.org/10.2307/2984875

[7] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[8] S. Eddins, "Cell segmentation," (2006). [Online]. Available: http://blogs.mathworks.com/steve/2006/06/02/cell-segmentation/

[9] L. Vincent and P. Soille, "Watersheds in digital spaces: An efficient algorithm based on immersion simulations," *IEEE Pattern Anal. Mach. Intell.*, vol. 13, no. 6, pp. 583–98, Jun. 1991.

[10] W. Rasband. (1997–2009). ImageJ, U.S. National Institutes of Health Bethesda, MD, [Online]. Available: http://rsb.info.nih.gov/ij/

[11] T. G. Dietterich, R. H. Lathrop, and T. L. Perez, "Solving the multiple-instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, no. 1–2, pp. 31–71, 1997.

[12] G. Fung, M. Dundar, B. Krishnapuram, and R. B. Rao, "Multiple instance learning for computer aided diagnosis," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, eds. Cambridge, MA: MIT Press, 2007, pp. 425–432.

[13] M. M. Dundar, S. Badve, V. Raykar, R. K. Jain, O. Sertel, and M. N. Gurcan, "A multiple instance learning approach toward optimal classification of pathology slides: A case study: Intraductal breast lesions," in *Proc. 20th Int. Conf. Pattern Recognit.*, Istanbul, Turkey, Aug. 23–26, 2010, pp. 2732–2735.

[14] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[15] V. C. Raykar, B. Krishnapuram, J. Bi, M. Dundar, and R. B. Rao, "Bayesian multiple instance learning: Automatic feature selection and inductive transfer," in *ICML 2008: Proc. 25th Int. Conf. Mach. Learning*. New York: ACM Press, 2008, pp. 808–815.

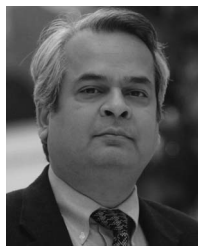
[16] D. Wu, J. Bi, and K. Boyer, "A min–max framework of cascaded classifier with multiple instance learning for computer aided diagnosis," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR 2009)*, Miami, FL, Jun. 20–25,, pp. 1359–1366.

[17] M. E. Tipping, "The relevance vector machine," in *Advances in Neural Information Processing Systems 12*, S. Solla, T. Leen, and K.-R. Muller, Eds. Cambridge, MA: MIT Press, 2000, pp. 652–658.

[18] J. A. Hanley and B. J. McNeil, "A method of comparing the areas under receiver operating characteristic curves derived from the same cases," *Radiology*, vol. 148, pp. 839–843, 1983.

M. Murat Dundar (M'06) received the B.S. degree from Bogazici University, Istanbul, Turkey, in 1997, and the M.S. and Ph.D. degrees from Purdue University, West Lafayette, IN in 1999 and 2003, respectively, all in electrical engineering.

He is currently an Assistant Professor in the Department of Computer and Information Science at IUPUI, Indianapolis, IN. His research interests include statistical pattern recognition and computational learning with applications to computer aided diagnosis, biosensors, and remote sensing.



Sunil Badve received the MD degree from Bombay University, Bombay, India, in 1986, completed his residency at the Albert Einstein College of Medicine, Bronx, NY, in 1998, and his fellowship at Yale University School of Medicine, New Haven, CT, 1999.

He is currently a Professor in the Department of Pathology and Laboratory Medicine, with additional appointment in the Department of Internal Medicine at Indiana University, Indianapolis. He serves as the Director of Translational Genomics Core at the Indiana University Cancer Center. His main research and clinical expertise is in breast cancer. He is the breast Pathologist for Eastern Cooperative Oncology group, where he serves as the Pathology Chair for several breast cancer clinical trials including the TAILORx clinical trial based on the Oncotype Dx assay.

Gokhan Bilgin received the B.S., M.S., and Ph.D. degrees from the Electronics and Telecommunications Department, Yildiz Technical University, Istanbul, Turkey.

He is currently a Postdoctoral Researcher in the Department of Computer and Information Science at IUPUI, Indianapolis, IN. His research interests include image and signal processing and machine learning with applications to biomedical engineering and remote sensing.

Vikas Raykar received the B.S. degree from the National Institute of Technology, Trichy, IN, and the M.S. and Ph.D. degrees from the University of Maryland, College Park.

He is currently works as a scientist in the CAD R&D Division for Siemens Healthcare, USA. His research interests include machine learning with an emphasis on fast computational methods.

Rohit Jain received the Bachelor of Medicine and Bachelor of Surgery (MBBS) degrees from Topiwala National Medical College, Mumbai, India and received the Master's degree in public health from Indiana University, Indianapolis, in May 2010.

He is currently assisting Dr. Sunil Badve, a breast cancer pathologists, on several of his research projects. His research interests include translational cancer research.

Olcay Sertel received the B.Sc. degree from Yildiz Technical University, Istanbul, Turkey in 2004 and the M.Sc. degree from Yeditepe University, Istanbul, Turkey in 2006, both in computer engineering. He is currently working toward the Ph.D. degree at the Department of Electrical and Computer Engineering, Ohio State University, Columbus.

He is also a Graduate Research Associate in the Department of Biomedical Informatics, Ohio State University.

Metin N. Gurcan (S'90–M'99–SM'06) received the B.Sc. and Ph.D. degrees in electrical and electronics engineering from Bilkent University, Ankara, Turkey, and the M.Sc. degree in digital systems engineering from the University of Manchester Institute of Science and Technology, Manchester, U.K.

He is currently an Associate Professor of Biomedical Informatics at the Ohio State University, Columbus. His research interests include image analysis and understanding, computer vision with applications to medicine.

Dr. Gurcan is the recipient of the British Foreign and Commonwealth Organization Award, Children's Neuroblastoma Cancer Foundation Young Investigator Award, and National Cancer Institute's caBIG Embodying the Vision Award. He is a Senior Member of the SPIE and RSNA.