This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2758580, IEEE Transactions on Medical Imaging

Multi-instance Multi-label Learning for Multi-class Classification of Whole Slide Breast Histopathology Images

Caner Mercan, Selim Aksoy, *Senior Member, IEEE,* Ezgi Mercan, Linda G. Shapiro, *Fellow, IEEE,* Donald L. Weaver, Joann G. Elmore

***Abstract*—Digital pathology has entered a new era with the availability of whole slide scanners that create high-resolution images of full biopsy slides. Consequently, the uncertainty regarding the correspondence between the image areas and the diagnostic labels assigned by pathologists at the slide level, and the need for identifying regions that belong to multiple classes with different clinical significance have emerged as two new challenges. However, generalizability of the state-of-the-art algorithms, whose accuracies were reported on carefully selected regions of interest (ROI) for the binary benign versus cancer classification, to these multi-class learning and localization problems is currently unknown. This paper presents our potential solutions to these challenges by exploiting the viewing records of pathologists and their slide-level annotations in weakly supervised learning scenarios. First, we extract candidate ROIs from the logs of pathologists' image screenings based on different behaviors, such as zooming, panning, and fixation. Then, we model each slide with a bag of instances represented by the candidate ROIs and a set of class labels extracted from the pathology forms. Finally, we use four different multi-instance multi-label learning algorithms for both slide-level and ROI-level predictions of diagnostic categories in whole slide breast histopathology images. Slide-level evaluation using 5-class and 14-class settings showed average precision values up to 81% and 69%, respectively, under different weakly-labeled learning scenarios. ROI-level predictions showed that the classifier could successfully perform multi-class localization and classification within whole slide images that were selected to include the full range of challenging diagnostic categories.**

***Index Terms*—Digital pathology, breast histopathology, whole slide imaging, region of interest detection, weakly-labeled learning, multi-class classification**

Copyright © 2017 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.
Manuscript received July 3, 2017; accepted September 19, 2017. C. Mercan and S. Aksoy were supported in part by the Scientific and Technological Research Council of Turkey under Grant No. 113E602 and in part by the GEBIP Award from the Turkish Academy of Sciences. E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore were supported in part by the National Cancer Institute of the National Institutes of Health under Awards No. R01-CA172343 and No. R01-140560. The content is solely the responsibility of the authors and does not necessarily represent the views of the National Cancer Institute or the National Institutes of Health.

C. Mercan and S. Aksoy are with the Department of Computer Engineering, Bilkent University, Ankara, 06800, Turkey. Email: {caner.mercan,saksoy}@cs.bilkent.edu.tr.

E. Mercan and L. G. Shapiro are with the Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA. Email: {ezgi,shapiro}@cs.washington.edu.

D. L. Weaver is with the Department of Pathology, University of Vermont, Burlington, VT 05405, USA. Email: Donald.Weaver@vtmednet.org.

J. G. Elmore is with the Department of Medicine, University of Washington, Seattle, WA 98195, USA. Email: jelmore@u.washington.edu.

I. Introduction

Histopathological image analysis has shown great potential in supporting the diagnostic process for cancer by providing objective and repeatable measures for characterizing the tissue samples to reduce the observer variations in the diagnoses [1]. The typical approach for computing these measures is to use statistical classifiers that are built by employing supervised learning algorithms on data sets that involve carefully selected regions of interest (ROI) with diagnostic labels assigned by pathologists. Furthermore, performance evaluation of these methods has also been limited to the use of manually chosen image areas that correspond to isolated tissue structures with no ambiguity regarding their diagnoses. Unfortunately, the high accuracy rates obtained in studies that are built around these restricted training and test settings do not necessarily reflect the complexity of the decision process encountered in routine histopathological examinations.

Breast histopathology is one particular example with a continuum of histologic features that have different clinical significance. For example, proliferative changes such as usual ductal hyperplasia (UDH) are considered benign, and patients diagnosed with UDH do not undergo any additional procedures [2]. On the other hand, major clinical treatment thresholds exist between atypical ductal hyperplasia (ADH) and ductal carcinoma in situ (DCIS) that carry different risks of progressing into malignant invasive carcinoma [3]. In particular, when a biopsy that actually has ADH is overinterpreted as DCIS, a woman may undergo unnecessary surgery, radiation, and hormonal therapy [4]. These problems have become even more important because millions of breast biopsies are performed annually, and the inter-rater agreement has always been a known challenge. However, generalizability of the state-of-the-art image analysis algorithms with accuracies reported for the simplified setting of benign versus malignant cases is currently unknown for this finer-grained categorization problem.

In this paper, we propose to exploit the pathologists' viewing records of whole slide images and integrate them with the pathology reports for *weakly supervised learning* of fine-grained classifiers. Whole slide scanners that create high-resolution images with sizes reaching to $100{,}000 \times 100{,}000$ pixels by digitizing the entire glass slides at $40\times$ magnification have enabled the whole diagnostic process to be completed in digital format. Earlier studies that used whole slide images have focused on efficiency issues where classifiers previously

0278-0062 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2758580, IEEE Transactions on Medical Imaging

trained on labeled ROI were run on large images by using multi-resolution [5] or multi-field-of-view [6] frameworks. However, two new challenges emerging from the use of whole slide images still need to be solved. The first challenge is the uncertainty regarding the correspondence between the image areas and the diagnostic labels assigned by the pathologists at the slide level. In clinical practice, the diagnosis is typically recorded for the entire slide, and the local tissue characteristics that grabbed the attention of the pathologist and led to that particular diagnosis are not known. The second challenge is the need for simultaneous detection and classification of diagnostically relevant areas in whole slides; large images often contain multiple regions with different levels of significance for malignancy, and it is not known a priori which local cues should be classified together. Both the former challenge that is related to the learning problem and the latter challenge that corresponds to the localization problem necessitate the development of new algorithms for whole slide histopathology.

The proposed framework uses multi-instance multi-label learning to build both slide-level and ROI-level classifiers for breast histopathology. Multi-instance learning (MIL) differs from traditional learning scenarios by use of the concept of bags, where each training bag contains several instances of positive and negative examples for the associated bag-level class label. A positive bag is assumed to contain at least one positive instance, whereas all instances in a negative bag are treated as negative examples, but the labels of the individual instances are not known during training. Multi-label learning (MLL) involves the scenarios in which each training example is associated with more than one label, as it can be possible to describe a sample in multiple ways. Multi-instance multi-label learning (MIMLL) corresponds to the combined case where each training sample is represented by a bag of multiple instances, and the bag is assigned multiple class labels.

The use of multi-instance and multi-label learning algorithms has been quite rare in the field of histopathological image analysis. Dundar et al. [7] presented one of the first applications of MIL for breast histopathology by designing a large margin classifier for binary discrimination of benign cases from actionable (ADH+DCIS) ones by using whole slides with manually identified ROIs. Xu et al. [8] used boosting-based MIL for binary classification of images as benign or cancer. They also used multi-label support vector machines for multi-class classification of colon cancer [9]. Cosatto et al. [10] studied binary classification in the multi-instance framework for diagnosis of gastric cancer. Kandemir and Hamprecht [11] used square patches as instances for multi-instance classification of tissue images as healthy or cancer. Most of the related studies in the literature consider only either the MIL or the MLL scenario. Most also study only the binary classification of images as cancer versus non-cancer. In this paper, we present experimental results on the categorization of breast histopathology images into 5 and 14 classes.

The main contributions of this paper are twofold. First, we study the MIMLL scenario in the context of whole slide image analysis. In our scenario, a bag corresponds to a digitized breast biopsy slide, the instances correspond to candidate ROIs in the slide, and the class labels correspond to the diagnoses associated with the slide. The candidate ROIs are identified by using a rule-based analysis of recorded actions of pathologists while they were interpreting the slides. The class labels are extracted from the forms that the pathologists filled out according to what they saw during their interpretation of the image. The second contribution is an extensive evaluation of the performances of four MIMLL algorithms on multi-class prediction of both the slide-level (bag-level) and the ROI-level (instance-level) labels for novel slides and simultaneous localization and classification of diagnostically relevant regions in whole slide images. The quantitative evaluation uses multiple performance criteria computed for classification scenarios involving 5 and 14 diagnostic classes and different combinations of viewing records from multiple pathologists. To the best of our knowledge, this is the first study that uses the MIMLL framework for learning and classification tasks involving such a comprehensive distribution of challenging diagnostic classes in histopathological image analysis. The rest of the paper is organized as follows. Section II introduces the data set, Section III describes the methodology, Section IV presents the experiments, and Section V gives the conclusions. An earlier version of this work was presented in [12].

TABLE I
DISTRIBUTION OF DIAGNOSTIC CLASSES AMONG THE 240 SLIDES.

(a) 14-class distribution

Class	# slides
Non-proliferative changes only	7
Fibroadenoma	16
Intraductal papilloma w/o atypia	11
Usual ductal hyperplasia	65
Columnar cell hyperplasia	89
Sclerosing adenosis	18
Complex sclerosing lesion	9
Flat epithelial atypia	37
Atypical ductal hyperplasia	69
Intraductal papilloma w/ atypia	15
Atypical lobular hyperplasia	18
Ductal carcinoma in situ	89
Lobular carcinoma in situ	7
Invasive carcinoma	22

(b) 5-class consensus distribution

Class	# slides
Non-proliferative changes only	13
Proliferative changes	63
Atypical ductal hyperplasia	66
Ductal carcinoma in situ	76
Invasive carcinoma	22

II. DATA SET

We used 240 haematoxylin and eosin (H&E) stained slides of breast biopsies that were selected from two registries that were associated with the Breast Cancer Surveillance Consortium [13]. Each slide belonged to an independent case from a different patient where a random stratified method was used to include cases that covered the full range of diagnostic categories from benign to invasive cancer. The class composition is given in Table I. The cases with atypical ductal hyperplasia and ductal carcinoma in situ were intentionally oversampled to gain statistical precision in the estimation of interpretive concordance for these diagnoses [4].

The selected slides were scanned at $40\times$ magnification, resulting in an average image size of $100{,}000 \times 64{,}000$ pixels. The cases were randomly assigned to one of four test sets, each including 60 cases with the same class frequency distribution, by using stratified sampling based on age, breast density, original reference diagnosis, and experts' difficulty rating of the case [13]. A total of 87 pathologists were recruited to evaluate the slides, and one of the four test sets was randomly

0278-0062 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2758580, IEEE Transactions on Medical Imaging

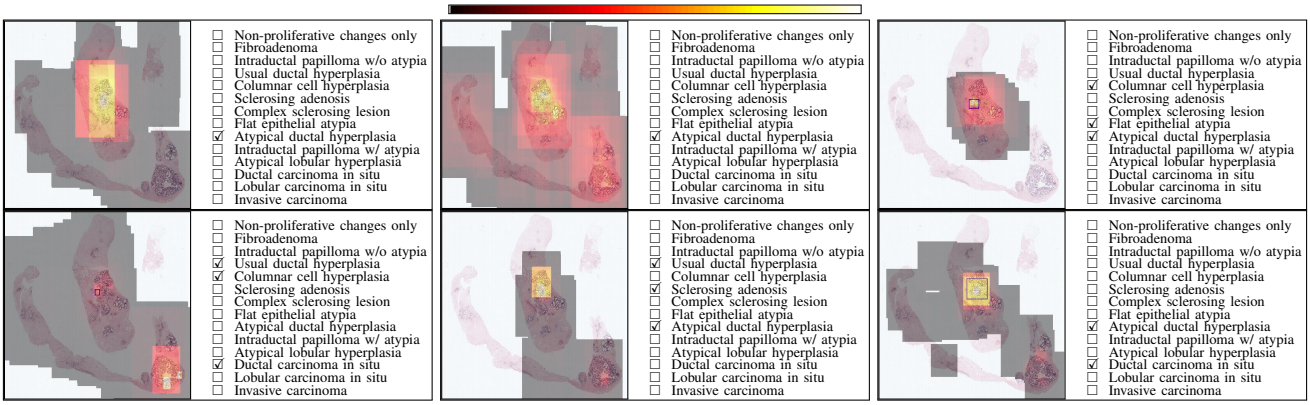


Fig. 1. Viewing behavior of six different pathologists on a whole slide image with a size of 74896×75568 pixels. The time spent by each pathologist on different image areas is illustrated using the heat map given above the images. The unmarked regions represent unviewed areas, and overlays from dark gray to red and yellow represent increasing cumulative viewing times. The diagnostic labels assigned by each pathologist to this image are also shown.

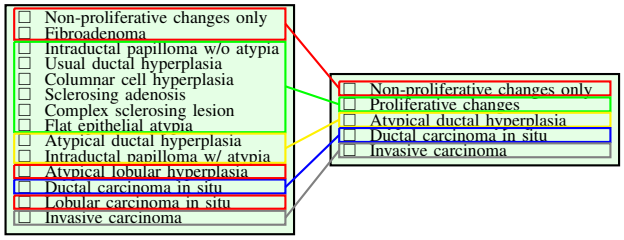


Fig. 2. Hierarchical mapping of 14 classes to 5. The mapping was designed by experienced pathologists [13]. The focus of data collection was to study ductal malignancies, so when only lobular carcinoma in situ or atypical lobular hyperplasia was present in a slide, it was put to the non-proliferative category.

assigned to each pathologist. Thus, each slide has, on average, independent interpretations from 22 pathologists. The data collection also involved tracking pathologists' actions while they were interpreting the slides using a web-based software tool that allowed seamless multi-resolution browsing of image data. The tracking software recorded the screen coordinates and mouse events at a frequency of four entries per second. At the end of the viewing session, each participant was also asked to provide a diagnosis by selecting one or more of the 14 classes on a pathology form to indicate what she/he had seen during her/his screening of the slide. Data for an example slide are illustrated in Figure 1. We also use a more general set of five classes with the mapping shown in Figure 2.

In addition, three experienced pathologists who are internationally recognized for research and education on diagnostic breast pathology evaluated every slide both independently and in consensus meetings where the result of the consensus meeting was accepted as the reference diagnosis for each slide. The difficulty of the classification problem studied here can be observed from the evaluation presented in [14] where the individual pathologists' concordance rates compared with the consensus-derived reference diagnosis was 82% for the union of non-proliferative and proliferative changes, 43% for ADH, 79% for DCIS, and 93% for invasive carcinoma. In our experiments, we only used the individual viewing logs and the diagnostic classifications from the three experienced pathologists for slide-level evaluation, because they were the only ones who evaluated all of the 240 slides. These pathologists' data also contained a bounding box around an example region that corresponded to the most representative and supporting ROI for the most severe diagnosis that was observed during their examination of that slide during consensus meetings. These consensus ROIs were used for ROI-level evaluation.

In summary, we used the three experienced pathologists' viewing logs, their individual assessments, and the consensus diagnoses for the four sets of 60 slides described above in a four-fold cross-validation setting so that the training and test slides always belonged to different patients. The study was approved by the institutional review boards at Bilkent University, University of Washington, and University of Vermont.

III. Methodology

A. Identification of candidate ROIs

The weakly supervised learning scenario studied in this paper used candidate ROIs that were extracted from the pathologists' viewing logs as potentially informative areas that may be important for the diagnosis of the whole slide. These candidate ROIs were identified among the viewports that were sampled from the viewing session of the pathologists and were represented by the coordinates of the image area viewed on the screen, the zoom level, and the time stamp.

Following the observation that different pathologists have different interpretive viewing behaviors [15], [16], we defined the following three actions: *zoom peak* is an entry that corresponds to an image area where the pathologist investigated closer by zooming in, and is defined as a local maximum in the zoom level; *slow panning* corresponds to image areas that are visited in consecutive viewports where the displacement (measured as the difference between the center pixels of two viewports) is small while the zoom level is constant; *fixation* corresponds to an area that is viewed for more than 2 seconds. The union of all viewports that belonged to one of these actions was selected as the set of candidate ROIs. Figure 3 illustrates the selection process for an example slide.

B. Feature extraction

The feature representation for each candidate ROI used the color histogram computed for each channel in the CIE-Lab space, texture histograms of local binary patterns computed

0278-0062 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2758580, IEEE Transactions on Medical Imaging

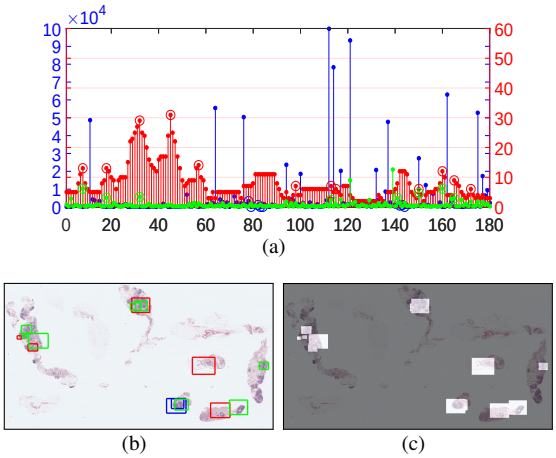


Fig. 3. ROI detection from the viewport logs. (a) Viewport log of a particular pathologist. The x-axis shows the log entry. The red, blue, and green bars represent the zoom level, displacement, and duration, respectively. (b) The rectangular regions visible on the pathologist's screen during the selected actions are drawn on the actual image. A *zoom peak* is a red circle in (a) and a red rectangle in (b), a *slow panning* is a blue circle in (a) and a blue rectangle in (b), a *fixation* is a green circle in (a) and a green rectangle in (b). (c) Candidate ROIs resulting from the union of the selected actions.

TABLE II
SUMMARY OF THE FEATURES FOR EACH CANDIDATE ROI. NUCLEAR ARCHITECTURE FEATURES WERE DERIVED FROM THE VORONOI DIAGRAM (VD), DELAUNAY TRIANGULATION (DT), MINIMUM SPANNING TREE (MST), AND NEAREST NEIGHBOR (NN) STATISTICS OF NUCLEI CENTROIDS. THE NUMBER OF FEATURES IS GIVEN FOR EACH TYPE.

Type	Description
Lab (192)	64-bin histogram of the CIE-L channel 64-bin histogram of the CIE-a channel 64-bin histogram of the CIE-b channel
LBP (128)	64-bin histogram of the LBP codes of the H channel 64-bin histogram of the LBP codes of the E channel
VD (13)	Total area of polygons Polygon area: mean, std dev, min/max ratio, disorder Polygon perimeter: mean, std dev, min/max ratio, disorder Polygon chord length: mean, std dev, min/max ratio, disorder
DT (8)	Triangle side length: mean, std dev, min/max ratio, disorder Triangle area: mean, std dev, min/max ratio, disorder
MST (4)	Edge length: mean, std dev, min/max ratio, disorder
NN (25)	Nuclear density Distance to 3, 5, 7 nearest nuclei: mean, std dev, disorder # of nuclei in 10, 20, 30, 40, 50 μm radius: mean, std dev, disorder

for the haematoxylin and eosin channels estimated using a color deconvolution procedure [17], and architectural features [6] computed from the nucleus detection results of [18]. Table II provides the details of the resulting 370-dimensional feature vector. The use of deep features will be the focus of future work because it is not yet straightforward to model this kind of complex histopathological content by using convolutional structures with limited training data.

C. Learning

The granularity of the annotations available in the training data determines the amount of supervision that can be incorporated into the learning process. Among the most popular weakly labeled learning scenarios, multi-instance learning (MIL) involves samples where each sample is represented by a collection (bag) of instances with a single label for the collection, and multi-label learning (MLL) uses samples where each sample has a single instance that is described by more than one label. In this section, we define the multi-instance multi-label learning (MIMLL) framework that contains both cases. Figure 4 illustrates the different learning scenarios in the context of whole slide imaging.

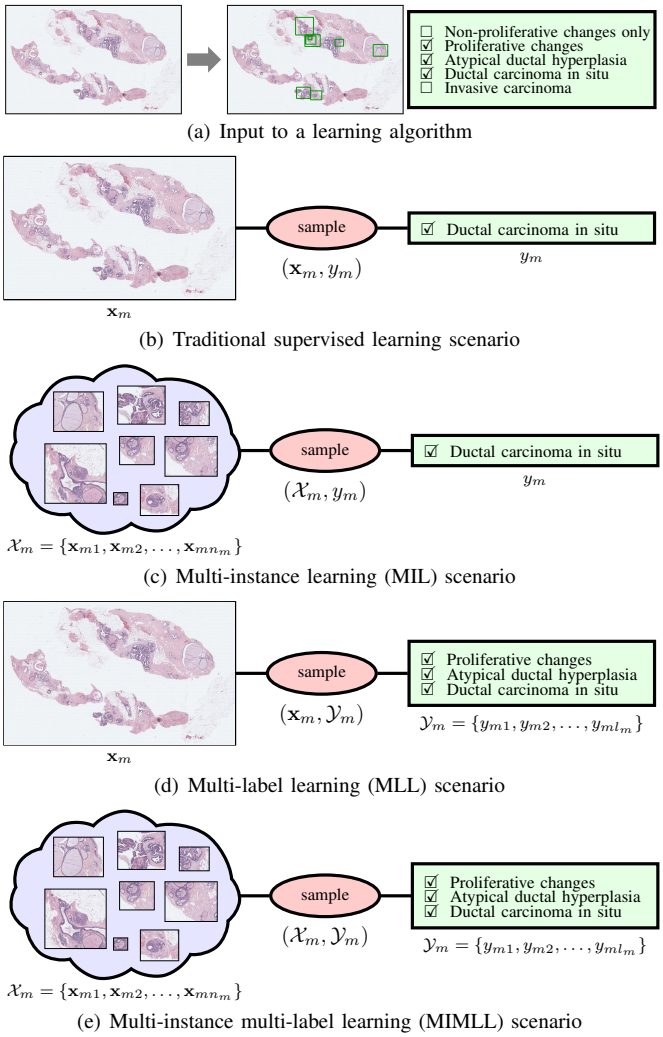


Fig. 4. Different learning scenarios in the context of whole slide breast histopathology. The input to a learning algorithm is the set of candidate ROIs obtained from the viewing logs of the pathologists and the diagnostic labels assigned to the whole slide. Different learning algorithms use these samples in different ways during training. The notation is defined in the text. The 5-class setting is shown, but we also use 14-class labels in the experiments.

Let $\{(\mathcal{X}_m, \mathcal{Y}_m)\}_{m=1}^{M}$ be a data set with M samples where each sample consists of a bag and an associated set of labels. The bag $\mathcal{X}_m$ contains a set of instances $\{\mathbf{x}_{mn}\}_{n=1}^{n_m}$ where $\mathbf{x}_{mn} \in \mathbb{R}^d$ is the feature vector of the n'th instance, and n_m is the total number of instances in that bag. The label set $\mathcal{Y}_m$ is composed of class labels $\{y_{ml}\}_{l=1}^{l_m}$ where $y_{ml} \in \{c_1, c_2, \ldots, c_L\}$ is one of L possible labels, and l_m is the total number of labels in that set. The traditional supervised learning problem is a special case of MIMLL where each sample has a single instance and a single label, resulting in the data set $\{(\mathbf{x}_m, y_m)\}_{m=1}^{M}$. MIL is also a special case of MIMLL where each bag has only one label, resulting in the data set $\{(\mathcal{X}_m, y_m)\}_{m=1}^{M}$. MLL is another special case where the single instance corresponding to a sample is associated with a set of labels, resulting in the data set $\{(\mathbf{x}_m, \mathcal{Y}_m)\}_{m=1}^{M}$.

In the following, we summarize four different approaches

0278-0062 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2758580, IEEE Transactions on Medical Imaging

adapted from the machine learning literature for the solution of the MIMLL problem studied in this paper.

1) MIMLSVMMI: A possible solution is to approximate the MIMLL problem as a multi-instance single label learning problem. Given an MIMLL data set with M samples, we can create a new MIL data set with $M \times \left(\sum_{m=1}^{M} l_m\right)$ samples where a sample $(\mathcal{X}_m, \mathcal{Y}_m)$ in the former is decomposed into a set of l_m bags as $\{(\mathcal{X}_m, y_{ml})\}_{l=1}^{l_m}$ in the latter by assuming that the labels are independent from each other. The resulting MIL problem is further reduced into a traditional supervised learning problem by assuming that each instance in a bag has an equal and independent contribution to the label of that bag, and is solved by using the MISVM algorithm [19].
2) MIMLSVM: An alternative is to decompose the MIMLL problem into a single-instance multi-label learning problem by embedding the bags in a new vector space. First, the bags are collected into a set $\{\mathcal{X}_m\}_{m=1}^{M}$, and the set is clustered using the k-medoids algorithm [20]. During clustering, the distance between two bags $\mathcal{X}_i = \{\mathbf{x}_{in}\}_{n=1}^{n_i}$ and $\mathcal{X}_j = \{\mathbf{x}_{jn}\}_{n=1}^{n_j}$ is computed by using the Hausdorff distance [21]:

$$h(\mathcal{X}_i, \mathcal{X}_j) = \max\Big\{ \max_{\mathbf{x}_i \in \mathcal{X}_i} \min_{\mathbf{x}_j \in \mathcal{X}_j} \|\mathbf{x}_i - \mathbf{x}_j\|, \max_{\mathbf{x}_j \in \mathcal{X}_j} \min_{\mathbf{x}_i \in \mathcal{X}_i} \|\mathbf{x}_j - \mathbf{x}_i\| \Big\}. \quad (1)$$

Then, the set of bags is partitioned into K clusters, each of which is represented by its medoid $\mathcal{M}_k, k = 1, \ldots, K$, the object in each cluster whose average dissimilarity to all other objects in the cluster is minimal. Finally, the embedding of a bag $\mathcal{X}_m$ into a K-dimensional space is performed by computing a vector $\mathbf{z}_m \in \mathbb{R}^K$ whose components are the Hausdorff distances between the bag and the medoids as $\mathbf{z}_m = (h(\mathcal{X}_m, \mathcal{M}_1), h(\mathcal{X}_m, \mathcal{M}_2), \ldots, h(\mathcal{X}_m, \mathcal{M}_K))$ [22]. The resulting MLL problem for the data set $\{(\mathbf{z}_m, \mathcal{Y}_m)\}_{m=1}^{M}$ is further reduced into a binary supervised learning problem for each class by using all samples that have a particular label in their label set as positive examples and the rest of the samples as negative examples for that label, and is solved using the MLSVM algorithm [23].
3) MIMLNN: Similar to MIMLSVM, the initial MIMLL problem is decomposed into an MLL problem by vector space embedding. This algorithm differs in the last step in which the resulting MLL problem is solved by using a linear classifier whose weights are estimated by minimizing a sum-of-squares error function [24].
4) M^{3}MIML: This method is motivated by the observation that useful information between instances and labels could be lost during the transformation of the MIMLL problem into an MIL (the first method) or an MLL (the second and third methods) problem [25]. The M^{3}MIML algorithm uses a linear model for each label where the output for a bag for a particular label is the maximum discriminant value among all instances of that bag under the model for that label. During training, the margin of a sample for a label is defined as this maximum over all instances, the margin of the sample for the multi-label classifier is defined as the minimum margin over all labels, and a quadratic programming problem is solved to estimate the parameters of the linear model by maximizing the margin of the whole training set that is defined as the minimum of all samples' margins.

Each algorithm described in this section was used to learn a multi-class classifier for which each training sample was a whole slide that was modeled as a bag of candidate ROIs ($\mathcal{X}_m$), each ROI being represented by a feature vector ($\mathbf{x}_{mn}$), and a set of labels that were assigned to that slide ($\mathcal{Y}_m$). The resulting classifiers were used to predict labels for a new slide as described in the following section.

D. Classification

Classification was performed both at the slide level and at the ROI level. Both schemes involved the same training procedures described in Section III-C using the MIMLL algorithms.

1) Slide-level classification: Given a bag of ROIs, $\mathcal{X}$, for an unknown whole slide image, a classifier trained as in Section III-C assigned a set of labels, $\mathcal{Y}'$, for that image. In the experiments, the bag $\mathcal{X}$ corresponded to the set of candidate ROIs extracted from the pathologists' viewing logs as described in Section III-A. If no logs were available at test time, an ROI detector for identifying and localizing diagnostically relevant areas as described in [15] and [16] would be used. Automated ROI detection is an open problem because visual saliency (that can be modeled by well-known algorithms in computer vision) does not always correlate well with diagnostic saliency [26]. New solutions for ROI detection can directly be incorporated in our framework to identify the candidate ROIs.

2) ROI-level classification: In many previously published works, classification at the ROI level involves manually selected regions of interest. However, this cannot be easily generalized to the analysis of whole slide images that involve many local areas that can have very different diagnostic relevance and structural ambiguities which may lead to disagreements among pathologists regarding their class assignments.

In this paper, a sliding window approach for classification at the ROI level was employed. Each whole slide image was processed within sliding windows of 3600×3600 pixels with an overlap of 2400 pixels along both horizontal and vertical dimensions. The sizes of the sliding windows were determined based on our empirical observations in [15] and [16]. Each window was considered as an instance whose feature vector $\mathbf{x}$ was obtained as in Section III-B. The classifiers learned in the previous section then assigned a set of labels $\mathcal{Y}'$ and a confidence score for each class for each window independently. Because of the overlap, each final unique classification unit corresponded to a window of 1200×1200 pixels, whose classification scores for each class were obtained by taking the per-class maximum of the scores of all sliding windows that overlap with this 1200×1200 pixel region.

IV. EXPERIMENTAL RESULTS

A. Experimental setting

The parameters for the algorithms were set based on trials on a small part of the data, based on suggestions made in

0278-0062 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2758580, IEEE Transactions on Medical Imaging

TABLE III
SUMMARY STATISTICS (AVERAGE ± STANDARD DEVIATION) FOR THE NUMBER OF CANDIDATE ROIS EXTRACTED FROM THE VIEWING LOGS. THE STATISTICS ARE GIVEN FOR SUBSETS OF THE SLIDES FOR INDIVIDUAL DIAGNOSTIC CLASSES BASED ON THE CONSENSUS LABELS (NON-PROLIFERATIVE CHANGES ONLY (NP), PROLIFERATIVE CHANGES (P), ATYPICAL DUCTAL HYPERPLASIA (ADH), DUCTAL CARCINOMA IN SITU (DCIS), INVASIVE CARCINOMA (INV)) AS WELL AS THE WHOLE DATA SET. *All* CORRESPONDS TO THE UNION OF THREE PATHOLOGISTS' ROIS FOR A PARTICULAR SLIDE.

Pathologist	NP	P	ADH	DCIS	INV	Whole
E1	13.6923±14.2559	26.5079±18.7340	26.5000±18.3557	16.0000±13.1261	24.4091±9.1634	22.2917±16.7609
E2	22.6154±21.6354	58.2857±46.9895	49.2273±42.3741	31.6184±27.8136	25.9545±14.0254	42.4542±38.8228
E3	6.6923±7.1576	25.3333±22.5145	17.8636±16.4699	9.5132±9.1964	6.0455±6.4400	15.4917±16.9972
All	43.0000±32.9646	110.1270±74.2180	93.5909±63.5455	57.1316±40.8204	56.4091±21.3177	80.2375±61.0469

the cited papers. Three of the four algorithms (MIMLSVMMI, MIMLSVM, and M^{3}MIML) used support vector machines (SVM) as the base classifier. The scale parameter in the Gaussian kernel was set to 0.2 for all three methods. The number of clusters (K) in MIMLSVM and MIMLNN was set to 20% and 40%, respectively, of the number of training samples (bags), and the regularization parameter in the least-squares problem in MIMLNN was set to 1.

The three experienced pathologists whose viewing logs were used in the experiments are denoted as *E1*, *E2*, and *E3*. For each one, the set of candidate ROIs for each slide was obtained as in Section III-A, and the feature vector for each ROI was extracted as in Section III-B to form the bag of instances for that slide. The multi-label set was formed by using the labels assigned to the slide by that expert. Overall, a slide contained, on average, 1.77±0.66 labels for five classes and 2.66±1.29 labels for 14 classes when the label sets assigned by all experts were combined. Each slide also had a single consensus label that was assigned jointly by the three pathologists.

Table III summarizes the ROI statistics in the data set. There are some significant differences in the screening patterns of the pathologists; some spend more time on a slide and investigate a larger number of ROIs, whereas some make faster decisions by looking at a few key areas. It is important to note that the slides with consensus diagnoses of proliferative changes and atypical ductal hyperplasia attracted significantly longer views resulting in more ROIs for all pathologists. Studying the correlations between different viewing behaviors and diagnostic accuracy and efficiency is part of our future work.

B. Evaluation criteria

Quantitative evaluation was performed by comparing the labels predicted for a slide by an algorithm to the labels assigned by the pathologists. The four test sets described in Section II were used in a four-fold cross-validation setup where the training and test samples (slides) came from different patients. Given the test set that consisted of N samples $\{(\mathcal{X}_n, \mathcal{Y}_n)\}_{n=1}^{N}$ where $\mathcal{Y}_n$ was the set of reference labels for the n*'th* sample, let $f(\mathcal{X}_n)$ be a function that returns the set of labels predicted by an algorithm for $\mathcal{X}_n$ and $r(\mathcal{X}_n, y)$ be the rank of the label y among $f(\mathcal{X}_n)$ when the labels are sorted in descending order of confidence in prediction (the label with the highest confidence has a rank of 1). We computed the following five criteria that are commonly used in multi-label classification:

- $hammingLoss(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}|f(\mathcal{X}_n) \triangle \mathcal{Y}_n|$, where $\triangle$ is the symmetric distance between two sets. It is the fraction of wrong labels (i.e., false positives or false negatives) to the total number of labels.
- $rankingLoss(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{|\mathcal{Y}_n||\overline{\mathcal{Y}_n}|}|\{(y_1, y_2)|r(\mathcal{X}_n, y_1) \geq r(\mathcal{X}_n, y_2),\ (y_1, y_2) \in \mathcal{Y}_n \times \overline{\mathcal{Y}_n}\}|$, where $\overline{\mathcal{Y}_n}$ denotes the complement of the set $\mathcal{Y}_n$. It is the fraction of label pairs where a wrong label has a smaller (better) rank than a reference label.
- $one\text{-}error(f) = \frac{1}{N}\sum_{n=1}^{N}\mathbb{1}\big[\arg\min_{y\in\{c_1,c_2,\ldots,c_L\}} r(\mathcal{X}_n, y) \notin \mathcal{Y}_n\big]$, where $\mathbb{1}$ is an indicator function that is 1 when its argument is true, and 0 otherwise. It counts the number of samples for which the top-ranked label is not among the reference labels.
- $coverage(f) = \frac{1}{N}\sum_{n=1}^{N}\max_{y\in\mathcal{Y}_n} r(\mathcal{X}_n, y) - 1$. It is defined as how far one needs to go down the list of predicted labels to cover all reference labels.
- $averagePrecision(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{|\mathcal{Y}_n|}\sum_{y\in\mathcal{Y}_n}|\{y'| r(\mathcal{X}_n, y') \leq r(\mathcal{X}_n, y),\ y' \in \mathcal{Y}_n\}|/r(\mathcal{X}_n, y)$. It is the average fraction of correctly predicted labels that have a smaller (or equal) rank than a reference label.

TABLE IV
5-CLASS SLIDE-LEVEL CLASSIFICATION RESULTS OF THE EXPERIMENTS WHEN A PARTICULAR PATHOLOGIST'S DATA (CANDIDATE ROIS AND CLASS LABELS) WERE USED FOR TRAINING (ROWS) AND EACH INDIVIDUAL PATHOLOGIST'S DATA WERE USED FOR TESTING (COLUMNS). THE BEST RESULT FOR EACH COLUMN IS MARKED IN BOLD.

		E1	*E2*	*E3*
E1	MIMLSVMMI	0.7094 ± 0.0600	0.6253 ± 0.0584	0.6326 ± 0.0153
	MIMLSVM	0.7757 ± 0.0419	0.6577 ± 0.0453	0.6901 ± 0.0060
	MIMLNN	**0.7823 ± 0.0332**	0.6813 ± 0.0323	0.7113 ± 0.0215
	M^{3}MIML	0.7420 ± 0.0476	0.5922 ± 0.0450	0.6702 ± 0.0162
E2	MIMLSVMMI	0.6524 ± 0.0174	0.5956 ± 0.0197	0.5908 ± 0.0243
	MIMLSVM	0.7664 ± 0.0381	**0.6905 ± 0.0383**	0.6932 ± 0.0168
	MIMLNN	0.7565 ± 0.0296	0.6737 ± 0.0279	0.7117 ± 0.0396
	M^{3}MIML	0.7471 ± 0.0345	0.6073 ± 0.0604	0.6993 ± 0.0245
E3	MIMLSVMMI	0.6406 ± 0.0521	0.5599 ± 0.0278	0.5971 ± 0.0400
	MIMLSVM	0.7570 ± 0.0239	0.6569 ± 0.0363	**0.7322 ± 0.0083**
	MIMLNN	0.7657 ± 0.0199	0.6705 ± 0.0175	0.7233 ± 0.0135
	M^{3}MIML	0.7449 ± 0.0505	0.6102 ± 0.0357	0.6745 ± 0.0119

To illustrate the evaluation criteria, consider a classification problem involving the labels $\{A, B, C, D, E\}$. Let a bag $\mathcal{X}$ have the reference labels $\mathcal{Y} = \{A, B, D\}$, and an algorithm predict $f(\mathcal{X}) = \{B, E, A\}$ in descending order of confidence. Hamming loss is $2/5 = 0.4$ (because D is a false negative and E is a false positive), ranking loss is $2/6 = 0.33$ (because (A, E) and (D, E) are wrongly ranked pairs), one-error is 0, coverage is 3 (assuming that D comes after A in the order of confidence), and average precision is $(2/3 + 1 + 3/4)/3 = 0.806$. Smaller values for the first four criteria and a larger value for the last one indicate better performance.

0278-0062 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2758580, IEEE Transactions on Medical Imaging

TABLE V
5-CLASS SLIDE-LEVEL CLASSIFICATION RESULTS OF THE EXPERIMENTS WHEN THE UNION OF THREE PATHOLOGISTS' DATA (CANDIDATE ROIS AND CLASS LABELS) WERE USED FOR TRAINING (ROWS). TEST LABELS CONSISTED OF THE UNION OF PATHOLOGISTS' INDIVIDUAL LABELS AS WELL AS THEIR CONSENSUS LABELS IN TWO SEPARATE EXPERIMENTS. THE EVALUATION CRITERIA ARE: HAMMING LOSS (HL), RANKING LOSS (RL), ONE-ERROR (OE), COVERAGE (COV), AND AVERAGE PRECISION (AP). THE BEST RESULT FOR EACH SETTING IS MARKED IN BOLD.

	Test data: $E1 \cup E2 \cup E3$				
	HL	RL	OE	COV	AP
MIMLSVMMI	0.3367 ± 0.0122	0.3361 ± 0.0197	0.4125 ± 0.0551	2.0542 ± 0.0798	0.7058 ± 0.0190
MIMLSVM	0.2675 ± 0.0164	0.2045 ± 0.0222	0.2958 ± 0.0438	1.6917 ± 0.0967	0.7790 ± 0.0228
MIMLNN	$\mathbf{0.2375 \pm 0.0189}$	$\mathbf{0.1771 \pm 0.0194}$	$\mathbf{0.2708 \pm 0.0498}$	$\mathbf{1.5583 \pm 0.0096}$	$\mathbf{0.8068 \pm 0.0262}$
M^{3}MIML	0.2842 ± 0.0152	0.2611 ± 0.0488	0.3250 ± 0.0518	1.9500 ± 0.1790	0.7301 ± 0.0374
	Test data: *Consensus*				
	HL	RL	OE	COV	AP
MIMLSVMMI	0.3042 ± 0.0117	0.3528 ± 0.0096	0.5167 ± 0.0593	1.7333 ± 0.1667	0.6518 ± 0.0250
MIMLSVM	0.2783 ± 0.0197	0.2295 ± 0.0351	0.4250 ± 0.1221	1.3958 ± 0.0774	0.7161 ± 0.0624
MIMLNN	$\mathbf{0.2567 \pm 0.0255}$	$\mathbf{0.2049 \pm 0.0421}$	$\mathbf{0.4125 \pm 0.1181}$	$\mathbf{1.2792 \pm 0.1031}$	$\mathbf{0.7377 \pm 0.0577}$
M^{3}MIML	0.2650 ± 0.0244	0.2792 ± 0.0812	0.4583 ± 0.1251	1.5833 ± 0.2289	0.6802 ± 0.0864

C. Slide-level classification results

The quantitative results given in this section show the average and standard deviation of the corresponding criteria computed using cross-validation. For each fold, the number of training samples, M, is 180, and the number of independent test samples, N, is 60.

1) 5-class classification results: Two experiments were performed to study scenarios involving different pathologists. The goal of the first experiment was to see how well a classifier built by using only a particular pathologist's viewing records (candidate ROIs and class labels) on the training slides could predict the class labels assigned by individual pathologists to the test slides. Table IV shows the average precision values for the experiments repeated using the data for each of the three pathologists separately. The results showed that MIMLNN and MIMLSVM performed the best, followed by M^{3}MIML, with MIMLSVMMI having the worst performance. An expected result (illustrated by the columns of Table IV) was that the classifier that performed the best on the test data labeled by a particular pathologist was the one that was learned from the training data of the same pathologist (different slides but labeled by the same person). Among the three pathologists, the first one had the largest average number of labels assigned to the slides (1.55 labels compared to 1.20 for the second and 1.26 for the third), that probably boosted the average precision values of the classifiers on the test data of the first pathologist.

The goal of the second experiment was to evaluate the effect of diversifying the training data, where the instance set for each training slide corresponded to the union of all candidate ROIs of the three pathologists (the last row of Table III), and the label set was formed as the union of all three pathologists' labels for that slide. As test labels, we used the union of three pathologists' labels as one setting, and the consensus diagnosis as another setting for each test slide. Table V shows the resulting performance statistics. The highest average precision of 0.8068 was obtained when the test labels were formed from the union of all pathologists' data. The more difficult setting that tried to predict the consensus label for each test slide resulted in an average precision of 0.7377 with MIMLNN as the classifier. (The consensus label-based evaluation is harsher on wrong classifications than multi-label evaluation when at least some of the labels are predicted correctly.)

TABLE VI
14-CLASS SLIDE-LEVEL CLASSIFICATION RESULTS OF THE EXPERIMENTS WHEN A PARTICULAR PATHOLOGIST'S DATA (CANDIDATE ROIS AND CLASS LABELS) WERE USED FOR TRAINING (ROWS) AND EACH INDIVIDUAL PATHOLOGIST'S DATA WERE USED FOR TESTING (COLUMNS). THE BEST RESULT FOR EACH COLUMN IS MARKED IN BOLD.

		E1	*E2*	*E3*
E1	MIMLSVMMI	0.5154 ± 0.0399	0.4443 ± 0.0774	0.4509 ± 0.0460
	MIMLSVM	0.6485 ± 0.0124	0.4950 ± 0.0370	0.5051 ± 0.0406
	MIMLNN	$\mathbf{0.6787 \pm 0.0354}$	0.5243 ± 0.0258	$\mathbf{0.5534 \pm 0.0425}$
	M^{3}MIML	0.6019 ± 0.0237	0.3828 ± 0.0429	0.4342 ± 0.0573
E2	MIMLSVMMI	0.4864 ± 0.0683	0.4470 ± 0.0447	0.4415 ± 0.0210
	MIMLSVM	0.4953 ± 0.0637	0.5524 ± 0.0708	0.5035 ± 0.0402
	MIMLNN	0.5671 ± 0.0503	$\mathbf{0.5724 \pm 0.0451}$	0.5412 ± 0.0269
	M^{3}MIML	0.5363 ± 0.0685	0.5139 ± 0.0555	0.5011 ± 0.0692
E3	MIMLSVMMI	0.3988 ± 0.0587	0.3914 ± 0.0335	0.4196 ± 0.0353
	MIMLSVM	0.5455 ± 0.0339	0.5262 ± 0.0523	0.5387 ± 0.0289
	MIMLNN	0.5891 ± 0.0362	0.5194 ± 0.0335	0.5448 ± 0.0264
	M^{3}MIML	0.5837 ± 0.0757	0.5242 ± 0.0347	0.5414 ± 0.0171

2) 14-class classification results: We used the same experimental setup in Section IV-C1 for 14-class classification. Table VI shows the average precision values. MIMLNN and MIMLSVM, that both formulated the MIMLL problem by embedding the bags into a new vector space and reducing it to an MLL problem, consistently outperformed both MIMLSVMMI that transformed the MIMLL problem into an MIL problem by assuming independence of labels, and M^{3}MIML that used a more complex model that was more sensitive to the amount of training data. Due to similar reasons as in the previous section, the scores when the first pathologist's test data were used were higher than the scores on the test data of the second and third pathologists. Also similar to the 5-class classification results, a particular pathologist's test data were classified the best by the classifier learned from the same pathologist's training data with the exception of the third one's test data which were classified the best when the training data of the first one were used. However, the best classification performance of the third pathologist's classifier, 0.5448, was very close to the first one's classifier's best classification score of 0.5534. These experiments once again confirmed the difficulty of whole slide learning and classification by using slide-level information compared to the same by using manually selected, well-defined regions as commonly studied in the literature.

The second set of experiments followed the same procedure as in Section IV-C1 as well. Table VII presents the quantitative

0278-0062 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2758580, IEEE Transactions on Medical Imaging

TABLE VII
14-CLASS SLIDE-LEVEL CLASSIFICATION RESULTS OF THE EXPERIMENTS WHEN THE UNION OF THREE PATHOLOGISTS' DATA (CANDIDATE ROIS AND CLASS LABELS) WERE USED FOR TRAINING (ROWS). TEST LABELS CONSISTED OF THE UNION OF PATHOLOGISTS' INDIVIDUAL LABELS AS WELL AS THEIR CONSENSUS LABELS IN TWO SEPARATE EXPERIMENTS. THE EVALUATION CRITERIA ARE: HAMMING LOSS (HL), RANKING LOSS (RL), ONE-ERROR (OE), COVERAGE (COV), AND AVERAGE PRECISION (AP). THE BEST RESULT FOR EACH SETTING IS MARKED IN BOLD.

	Test data: $E1 \cup E2 \cup E3$				
	HL	RL	OE	COV	AP
MIMLSVMMI	0.2054 ± 0.0171	0.3138 ± 0.0406	0.5500 ± 0.0793	6.4750 ± 0.1917	0.5425 ± 0.0353
MIMLSVM	0.1646 ± 0.0134	0.1912 ± 0.0239	0.3542 ± 0.0786	5.5208 ± 0.7504	0.6432 ± 0.0293
MIMLNN	$\mathbf{0.1604 \pm 0.0103}$	$\mathbf{0.1591 \pm 0.0120}$	$\mathbf{0.3125 \pm 0.0685}$	$\mathbf{5.0042 \pm 0.4267}$	$\mathbf{0.6917 \pm 0.0307}$
M^{3}MIML	0.1732 ± 0.0041	0.2297 ± 0.0209	0.3833 ± 0.0491	6.1667 ± 0.4587	0.5661 ± 0.0324
	Test data: *Consensus*				
	HL	RL	OE	COV	AP
MIMLSVMMI	0.1792 ± 0.0186	0.3093 ± 0.0326	0.6625 ± 0.0658	5.4083 ± 0.6198	0.4864 ± 0.0323
MIMLSVM	0.1592 ± 0.0119	0.2181 ± 0.0211	0.5750 ± 0.1206	4.8417 ± 0.8742	0.5281 ± 0.0402
MIMLNN	$\mathbf{0.1557 \pm 0.0080}$	$\mathbf{0.1843 \pm 0.0089}$	$\mathbf{0.5208 \pm 0.1109}$	$\mathbf{4.2333 \pm 0.5418}$	$\mathbf{0.5855 \pm 0.0456}$
M^{3}MIML	0.1565 ± 0.0044	0.2618 ± 0.0165	0.6125 ± 0.1117	5.4833 ± 0.5307	0.4568 ± 0.0507

results. In agreement with the 5-class classification results, the best performance was achieved when the union of all pathologists' data were used for both training and testing, but with a drop in average precision from 0.8068 to 0.6917 for the more challenging 14-class setting. We would like to note that it was not straightforward to compare the 5-class and 14-class performances with respect to all evaluation criteria, as the number of labels in the respective test sets could often be different, and some performance criteria (e.g., coverage) were known to be more sensitive to the number of labels than others. The results obtained in this section will also be used as baselines in our future studies. Our future work will investigate the similarities and differences between the ROIs from different pathologists at the feature level, study the relationships between slide-level diagnoses and ROI-level predictions, and extend the experiments by using different scenarios that exploit data from additional pathologists.

D. ROI-level classification results

We followed the sliding window approach described in Section III-D2 to obtain confidence scores for all classes at each 1200×1200 pixel window of a whole slide image. The best performing classifier of the previous section, MIMLNN, was selected for training with the union of all candidate ROIs from the three pathologists and with the slide-level consensus labels. We used only the 5-class setting, since the consensus reference data used for performance evaluation at the ROI level had only 5-class information.

As mentioned in Section II, ADH and DCIS cases were oversampled during data set construction [4]. This made automatic learning of the minority classes NP and INV difficult even though they are relatively easier for humans. Therefore, we employed an upsampling approach for these two classes where a new bag was formed by sampling with replacement from the instances of a randomly selected bag until the number of training samples increased by twofold. The resulting set was used for weakly-labeled training of a multi-class classifier from slide-level information for ROI-level classification.

Since only the diagnostic label of the consensus ROI was known for each slide, only the 1200×1200 subwindows within that region were used for quantitative evaluation. We used the following protocol for predicting a label for this ROI by using its subwindows. First, we assigned the class that had the highest score as the diagnostic label of each subwindow. Then, we used a classification threshold on these scores to eliminate the ones that had low certainty. Finally, we picked the most severe diagnostic label among the remaining subwindows as the label of the corresponding ROI. If a slide-level grading is desired, the connected components formed by the subwindows that pass the classification threshold can be found, and the most severe diagnosis can be used as the diagnostic label of that slide. The components also provide clinically valuable information as one may want to localize all diagnostically relevant regions that may belong to different classes.

TABLE VIII
CONFUSION MATRIX FOR ROI-LEVEL CLASSIFICATION.

		Predicted				
		NP	P	ADH	DCIS	INV
True	NP	0	5	3	3	0
	P	0	15	30	13	4
	ADH	3	20	32	10	1
	DCIS	0	5	22	42	6
	INV	0	0	2	10	10

TABLE IX
CLASS-SPECIFIC STATISTICS ON THE PERFORMANCE OF ROI-LEVEL CLASSIFICATION. THE NUMBER OF TRUE POSITIVES (TP), FALSE POSITIVES (FP), FALSE NEGATIVES (FN), AND TRUE NEGATIVES (TN) ARE GIVEN. PRECISION, RECALL (ALSO KNOWN AS TRUE POSITIVE RATE AND SENSITIVITY), FALSE POSITIVE RATE (FPR), AND SPECIFICITY (ALSO KNOWN AS TRUE NEGATIVE RATE) ARE ALSO SHOWN.

Class	TP	FP	FN	TN	Precision	Recall/ Sensitivity	FPR	Specificity
NP	0	3	11	222	0.00	0.00	0.01	0.99
P	15	30	47	144	0.33	0.24	0.17	0.83
ADH	32	57	34	113	0.36	0.48	0.34	0.66
DCIS	42	36	33	125	0.54	0.56	0.22	0.78
INV	10	11	12	203	0.48	0.45	0.05	0.95

We evaluated different parameter settings for the protocol described above. The best results were obtained when the classification threshold was 0.7. Tables VIII and IX summarize the classification results. Among the five classes, namely non-proliferative changes only (NP), proliferative changes without atypia (P), atypical ductal hyperplasia (ADH), ductal carcinoma in situ (DCIS), and invasive cancer (INV), we observed that the classifier could predict P, ADH, DCIS, and INV better than NP. In spite of the upsampling, most of the NP cases

0278-0062 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2758580, IEEE Transactions on Medical Imaging

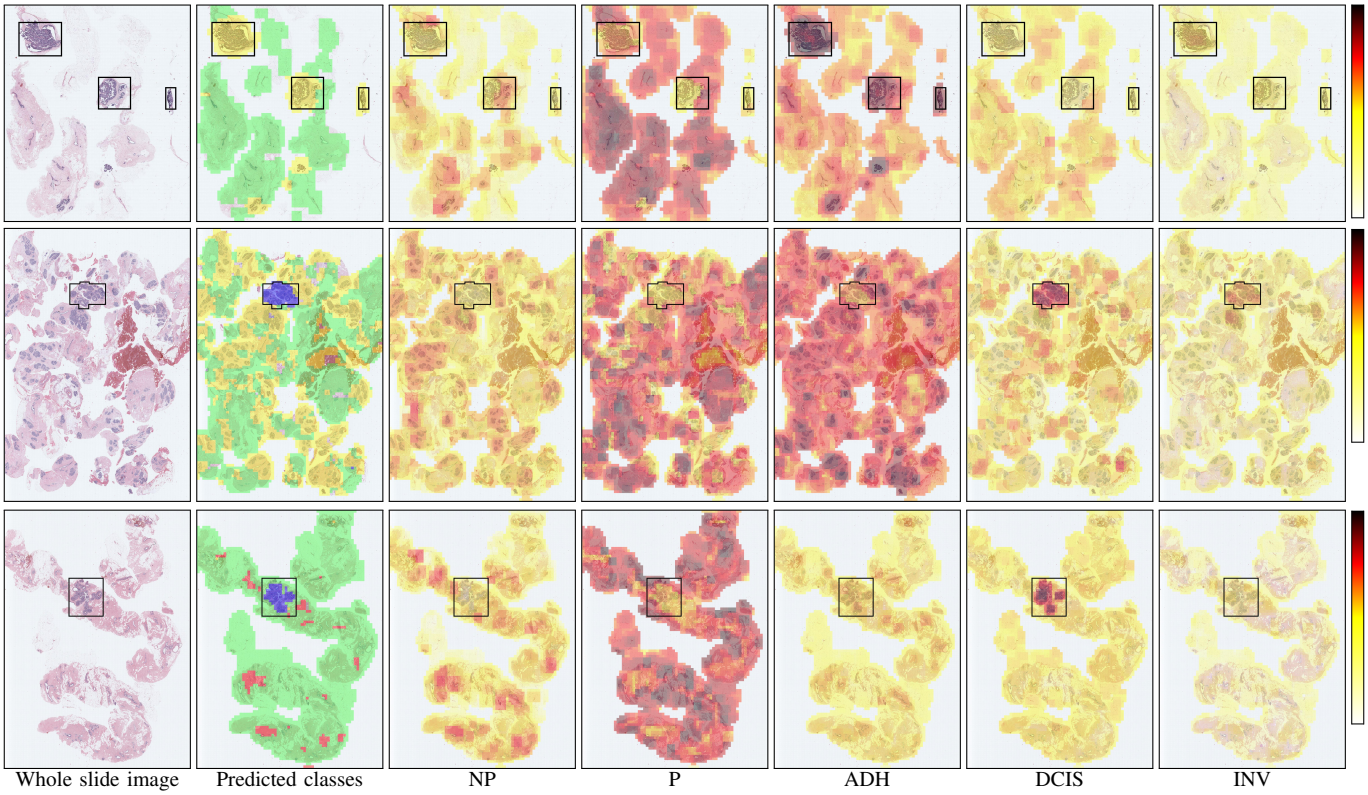


Fig. 5. Whole slide ROI-level classification examples. From left to right: original image; each 1200×1200 window is colored according to the class with the highest score (see Figure 2 for the colors of the classes); scores for individual classes using the color map show on the right. The consensus ROIs are shown using black rectangles. The consensus diagnosis for the case in the first row is atypical ductal hyperplasia, and the consensus diagnoses for the second and third rows are ductal carcinoma in situ.

were incorrectly labeled as P, followed by ADH and DCIS. Precision values indicated better performance for DCIS and INV, followed by ADH and P. Recall values for P indicated a large number of missed cases; most were misclassified as ADH and a comparatively smaller number were misclassified as DCIS. ADH and DCIS were more successfully captured, with DCIS having a relatively smaller false positive rate compared to ADH where the classifier incorrectly assigned a class label of ADH to a large number of cases associated with P and a smaller number of cases associated with DCIS. The classifier could detect 10 out of the 22 INV cases correctly and 10 of the misclassified 12 cases were labeled as DCIS, which was not an unexpected result given that most cases labeled as INV also included DCIS in their pathology reports.

Even though slide-level predictions achieved precision values up to 81%, ROI-level quantitative accuracy appeared to be lower than human performance. The main cause of the ROI-level predictions counted as errors was the difficulty of the multi-class classification problem by using weakly-labeled learning from pathologists' viewing records. For example, the multi-label training sets with INV, DCIS, or ADH as the most severe diagnosis often also included other classes, and the candidate ROIs that were included in the bags that corresponded to these multi-label sets covered diagnostically relevant regions that belonged to the full continuum (P, ADH, low-grade DCIS, high-grade DCIS, etc.) of histologic categories. Unfortunately, there is no comparable benchmark that studied these classes in the histopathological image analysis literature where discrimination of classes such as ADH and DCIS was intentionally ignored as being too difficult even in fully supervised settings and when manually annotated ROIs were used for training [7], [27]. These classes are also often the most difficult to differentiate even by experienced pathologists using structural cues, and this was particularly apparent for our data set, as well [4], [14]. The proposed classification setting was powerful enough to work with generic off-the-shelf features that were not specifically designed for breast pathology. Our future work includes the development of new feature representations that can model the structural changes used by humans in diagnosis and weakly labeled learning algorithms that further exploit the pathologists' records for the discrimination of these challenging classes.

Figure 5 presents ROI-level classification examples. In general, the multi-class classification within the whole slide and the localization of regions with different diagnostic relevance appeared to be more accurate compared to the numbers given in quantitative evaluation.

V. Conclusions

We presented a study on multi-class classification of whole slide breast histopathology images. Contrary to the traditional fully supervised setup, where manually chosen image areas and their unambiguous class labels are used for learning, we considered a more realistic scenario involving weakly labeled whole slide images where only the slide-level labels were provided by the pathologists. The uncertainty regarding the

0278-0062 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2758580, IEEE Transactions on Medical Imaging

correspondences between the particular local details and the selected diagnoses at the slide level was modeled in a multi-instance multi-label learning framework, where the whole slide was treated as a bag, the candidate ROIs extracted from the screen coordinates as part of the viewing records of pathologists were used as the instances in this bag, and one or more diagnostic classes associated with the slide in the pathology form were used as the multi-label set.

Training and test data obtained through various combinations of three pathologists' recordings were used to evaluate the performances of four different multi-instance multi-label learning algorithms on classification of diagnostically relevant regions as well as whole slide images as belonging to 5 or 14 diagnostic categories. Quantitative evaluation of 5-class slide-level predictions resulted in average precision values up to 78% when individual pathologist's viewing records were used and 81% when the candidate ROIs and the class labels from all pathologists were combined for each slide. Additional experiments showed slightly lower performance for the more difficult 14-class setting. We also illustrated the use of classifiers trained using slide-level information for multi-class prediction of ROIs with different diagnostic relevance.

We would like to note that the 240 slides in our data set were selected to include the full range of cases, and with more cases of ADH and DCIS than in typical clinical practice, this image cohort is diagnostically more difficult. Additionally, the classifiers used were trained only using weakly labeled data at the slide level, where the number of training samples could be considered very small for such a multi-class setting. Given the difficulty and the novelty of the learning and classification problems in this paper, our results provide very valuable benchmarks for future studies on challenging multi-class whole slide classification tasks where collection of fully-supervised data sets is not possible.

REFERENCES

[1] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Reviews in Biomedical Eng.*, vol. 2, pp. 147–171, October 2009.

[2] R. K. Jain, R. Mehta, R. Dimitrov, L. G. Larsson, P. M. Musto, K. B. Hodges, T. M. Ulbright, E. M. Hattab, N. Agaram, M. T. Idrees, and S. Badve, "Atypical ductal hyperplasia: interobserver and intraobserver variability," *Modern Pathology*, vol. 24, pp. 917–923, 2011.

[3] K. H. Allison, M. H. Rendi, S. Peacock, T. Morgan, J. G. Elmore, and D. L. Weaver, "Histological features associated with diagnostic agreement in atypical ductal hyperplasia of the breast: Illustrative cases from the B-Path study," *Histopathology*, vol. 69, pp. 1028–1046, 2016.

[4] J. G. Elmore, G. M. Longton, P. A. Carney, B. M. Geller, T. Onega, A. N. A. Tosteson, H. D. Nelson, M. S. Pepe, K. H. Allison, S. J. Schnitt, F. P. O'Malley, and D. L. Weaver, "Diagnostic concordance among pathologists interpreting breast biopsy specimens," *Journal of American Medical Association*, vol. 313, no. 11, pp. 1122–1132, 2015.

[5] S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi, "A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies," *IEEE Trans. on Biomedical Eng.*, vol. 59, no. 5, pp. 1205–1218, May 2012.

[6] A. Basavanhally, S. Ganesan, M. Feldman, N. Shih, C. Mies, J. Tomaszewski, and A. Madabhushi, "Multi-field-of-view framework for distinguishing tumor grade in ER+ breast cancer from entire histopathology slides," *IEEE Trans. on Biomedical Eng.*, vol. 60, no. 8, pp. 2089–2099, August 2013.

[7] M. M. Dundar, S. Badve, G. Bilgin, V. Raykar, R. Jain, O. Sertel, and M. N. Gurcan, "Computerized classification of intraductal breast lesions using histopathological images," *IEEE Trans. on Biomedical Eng.*, vol. 58, no. 7, pp. 1977–1984, July 2011.

[8] Y. Xu, J.-Y. Zhu, E. Chang, and Z. Tu, "Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 964–971.

[9] Y. Xu, L. Jiao, S. Wang, J. Wei, Y. Fan, M. Lai, and E. I.-C. Chang, "Multi-label classification for colon cancer using histopathological images," *Microscopy Research and Technique*, vol. 76, no. 12, pp. 1266–1277, 2013.

[10] E. Cosatto, P.-F. Laquerre, C. Malon, H.-P. Graf, A. Saito, T. Kiyuna, A. Marugame, and K. Kamijo, "Automated gastric cancer diagnosis on H&E-stained sections; training a classifier on a large scale with multiple instance machine learning," in *SPIE Medical Imaging*, vol. 867605, 2013.

[11] M. Kandemir and F. A. Hamprecht, "Computer-aided diagnosis from weak supervision: A benchmarking study," *Computerized Medical Imaging and Graphics*, vol. 42, pp. 44–50, 2015.

[12] C. Mercan, E. Mercan, S. Aksoy, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Multi-instance multi-label learning for whole slide breast histopathology," in *SPIE Medical Imaging Symposium, Digital Pathology Conference*, San Diego, California, February 27–March 3, 2016.

[13] N. V. Oster, P. A. Carney, K. H. Allison, D. L. Weaver, L. M. Reisch, G. Longton, T. Onega, M. Pepe, B. M. Geller, H. D. Nelson, T. R. Ross, A. N. A. Tosteson, and J. G. Elmore, "Development of a diagnostic test set to assess agreement in breast pathology: practical application of the Guidelines for Reporting Reliability and Agreement Studies (GRRAS)," *BMC Women's Health*, vol. 13, no. 3, pp. 1–8, 2013.

[14] J. G. Elmore, G. M. Longton, M. S. Pepe, P. A. Carney, H. D. Nelson, K. H. Allison, B. M. Geller, T. Onega, A. N. A. Tosteson, E. Mercan, L. G. Shapiro, T. T. Brunye, T. R. Morgan, and D. L. Weaver, "A randomized study comparing digital imaging to traditional glass slide microscopy for breast biopsy and cancer diagnosis," *Journal of Pathology Informatics*, vol. 8, no. 1, pp. 1–12, 2017.

[15] E. Mercan, S. Aksoy, L. G. Shapiro, D. L. Weaver, T. Brunye, and J. G. Elmore, "Localization of diagnostically relevant regions of interest in whole slide images," in *International Conference on Pattern Recognition*, 2014, pp. 1179–1184.

[16] E. Mercan, S. Aksoy, L. G. Shapiro, D. L. Weaver, T. T. Brunye, and J. G. Elmore, "Localization of diagnostically relevant regions of interest in whole slide images: A comparative study," *Journal of Digital Imaging*, vol. 29, no. 4, pp. 496–506, August 2016.

[17] A. Ruifrok and D. Johnston, "Quantification of histochemical staining by color deconvolution," *Analytical and Quantitative Cytology and Histology*, vol. 23, no. 4, pp. 291–299, 2001.

[18] H. Xu, C. Lu, and M. Mandal, "An efficient technique for nuclei segmentation based on ellipse descriptor analysis and improved seed detection algorithm," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 5, pp. 1729–1741, September 2014.

[19] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in Neural Information Processing Systems*, 2002, pp. 561–568.

[20] L. Kaufman and P. J. Rousseeuw, "Clustering by means of medoids," in *Statistical Data Analysis Based on the L1-Norm and Related Methods*, Y. Dodge, Ed. North-Holland, 1987, pp. 405–416.

[21] G. Edgar, *Measure, Topology, and Fractal Geometry*. Springer Science & Business Media, 2007.

[22] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, "Multi-instance multi-label learning," *Artificial Intelligence*, vol. 176, no. 1, pp. 2291–2320, 2012.

[23] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[24] M.-L. Zhang and Z.-H. Zhou, "Multi-label learning by instance differentiation," in *AAAI Conference on Artificial Intelligence*, vol. 7, 2007, pp. 669–674.

[25] ——, "M3MIML: A maximum margin method for multi-instance multi-label learning," in *IEEE International Conference on Data Mining*, 2008, pp. 688–697.

[26] T. T. Brunye, P. A. Carney, K. H. Allison, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Eye movements as an index of pathologist visual expertise: A pilot study," *PLoS ONE*, vol. 9, no. 8, 2014.

[27] B. E. Bejnordi, M. Balkenhol, G. Litjens, R. Holland, P. Bult, N. Karssemeijer, and J. A. W. M. van der Laak, "Automated detection of DCIS in whole-slide h&e stained breast histopathology images," *IEEE Trans. on Medical Imaging*, vol. 35, no. 9, pp. 2141–2150, September 2016.

0278-0062 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.