Multi-instance multi-label learning for whole slide breast histopathology

Caner Mercan[a], Ezgi Mercan[b], Selim Aksoy[a], Linda G. Shapiro[b], Donald L. Weaver[c], and Joann G. Elmore[d]

[a]Dept. of Computer Engineering, Bilkent University, Ankara, 06800, Turkey
[b]Dept. of Computer Science & Engineering, University of Washington, Seattle, WA 98195, USA
[c]Dept. of Pathology, University of Vermont, Burlington, VT 05405, USA
[d]Dept. of Medicine, University of Washington, Seattle, WA 98195, USA

ABSTRACT

Digitization of full biopsy slides using the whole slide imaging technology has provided new opportunities for understanding the diagnostic process of pathologists and developing more accurate computer aided diagnosis systems. However, the whole slide images also provide two new challenges to image analysis algorithms. The first one is the need for simultaneous localization and classification of malignant areas in these large images, as different parts of the image may have different levels of diagnostic relevance. The second challenge is the uncertainty regarding the correspondence between the particular image areas and the diagnostic labels typically provided by the pathologists at the slide level. In this paper, we exploit a data set that consists of recorded actions of pathologists while they were interpreting whole slide images of breast biopsies to find solutions to these challenges. First, we extract candidate regions of interest (ROI) from the logs of pathologists' image screenings based on different actions corresponding to zoom events, panning motions, and fixations. Then, we model these ROIs using color and texture features. Next, we represent each slide as a bag of instances corresponding to the collection of candidate ROIs and a set of slide-level labels extracted from the forms that the pathologists filled out according to what they saw during their screenings. Finally, we build classifiers using five different multi-instance multi-label learning algorithms, and evaluate their performances under different learning and validation scenarios involving various combinations of data from three expert pathologists. Experiments that compared the slide-level predictions of the classifiers with the reference data showed average precision values up to 62% when the training and validation data came from the same individual pathologist's viewing logs, and an average precision of 64% was obtained when the candidate ROIs and the labels from all pathologists were combined for each slide.

Keywords: Digital pathology, computer aided diagnosis, breast histopathology, whole slide imaging, region of interest analysis, image classification, multi-instance multi-label learning

1. INTRODUCTION

The diagnosis for cancer is made through a microscopic examination of a tissue sample by highly-trained pathologists. Histopathological image analysis systems have a great potential in aiding this process by relieving the workload of the pathologists through filtering out obviously benign areas, and by providing an objective quantification of the tissue content to reduce the intra- and inter-observer variations in the diagnoses.[1]

The most common approach to histopathological image analysis is to apply supervised learning techniques to manually selected regions of interest that are represented using color, texture, or structural features. The pre-processing stage that corresponds to data collection involves a time-consuming process that yields samples containing hand-picked cases with relatively less amount of diagnostic complexity. For such cases, application of supervised classification methods can be acceptable when manual selection of the image areas results in isolated tissue structures that have no ambiguity regarding their diagnoses.

Send correspondence to S.A.: E-mail: saksoy@cs.bilkent.edu.tr, Telephone: +90 (312) 2903405

A more realistic scenario is the analysis of high-resolution images that are acquired using whole slide image scanners. These virtual slides with sizes around 100,000 × 100,000 pixels at 40× magnification enable the whole diagnostic process to be done in digital form. However, they also provide two new challenges to image analysis algorithms. The first is the need for simultaneous localization and classification of malignant areas in these large images, as different parts of the image may have different levels of diagnostic relevance. The second challenge is the uncertainty regarding the correspondence between the image areas and the diagnostic labels provided by the pathologists. The labels are typically given at the slide level and the particular image regions that lead to these diagnoses are often not recorded. Both of these challenges limit the applicability of most methods in the extensive literature on supervised classification of histological images, and necessitate the development of new techniques for whole slide image analysis.

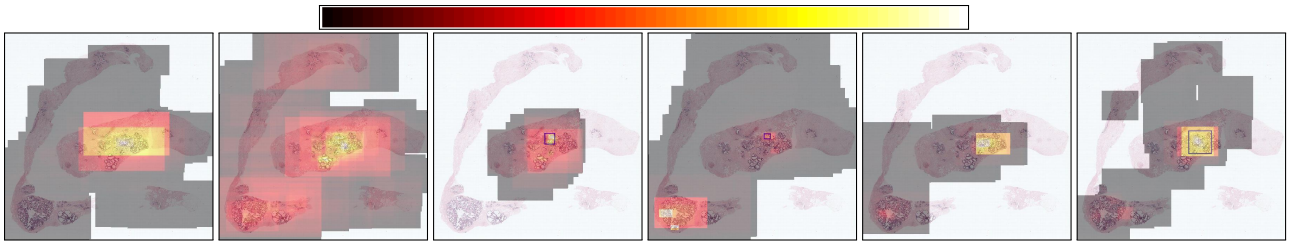
The goal of this paper is to evaluate the use of coarsely-grained diagnostic annotations by multi-instance multi-label learning algorithms to support weakly supervised learning scenarios for building classifiers for categorizing whole slide breast histopathology images. Multi-instance learning (MIL) uses training sets that consist of bags, each containing several instances that are either positive or negative examples for the class of interest. Only bag-level labels are given and the instance-level labels are unknown during training. Multi-label learning (MLL) corresponds to the scenario where each training sample is assigned more than one label. Multi-instance multi-label learning (MIMLL) combines MIL and MLL where each training sample is represented by multiple instances and is associated by multiple class labels. The MIMLL framework is quite new in the field of histopathological image analysis. Dundar et al.[2] studied the breast histopathology image classification problem using MIL. Xu et al.[3] used MIL for cancer image classification and segmentation. The same team also performed multi-class classification of colon cancer using MLL[4] and used deep learning to extract image features for binary classification of slides using MIL.[5] Cosatto et al.[6] applied binary classification using the multi-instance framework for diagnosis of gastric cancer. Most of the related work consider only either the MIL or the MLL scenario. Most also consider only the binary classification of images as cancer versus non-cancer samples.

We apply the MIMLL scenario to the whole slide image analysis problem where the instances correspond to the regions of interest (ROI) in a large slide and the labels correspond to the expert's annotations in the pathology report of this slide. The aim is to predict the slide-level (bag-level) labels in new images and simultaneously localize and classify diagnostically relevant regions. We use a data set that consists of whole slide images of breast biopsies as well as recorded actions of pathologists while they were interpreting these slides. First, we extract candidate ROIs from the logs of image screenings based on different actions corresponding to zoom events, panning motions, and fixations. Then, we model these ROIs using color and texture features, and represent each slide as a bag of ROIs where slide-level labels are extracted from the forms that the pathologists filled out according to what they saw during their screenings. Finally, we evaluate the performances of five different MIMLL algorithms on multi-class classification of these slides. We consider different learning and validation scenarios involving different combinations of viewing logs from multiple pathologists. The following sections summarize the data set (Section 2), the methodology (Section 3), and the results of the experiments (Section 4).

2. DATA SET

The data used in this paper were collected in the scope of an NIH-sponsored project titled "Digital Pathology, Accuracy, Viewing Behavior and Image Characterization (digiPATH)" that aims to evaluate the accuracy of pathologists' interpretation of digital images vs. glass slides. There are 240 haematoxylin and eosin (H&E) stained breast biopsy slides with different diagnostic categories ranging from benign to cancer. They were scanned at 40× magnification, resulting in an average image size of 100,000 × 64,000 pixels. Each image was interpreted independently by, on the average, 22 pathologists including three expert pathologists who are internationally recognized for research and education on diagnostic breast pathology.

With the aim of analyzing the viewing behavior of pathologists to identify visual scanning patterns associated with diagnostic accuracy and efficiency, data collection also involved detailed tracking of pathologists' interpretation of digital slides using a web-based software tool. The software allowed browsing of high-resolution digital images and recorded the changes on the screen as viewing logs. At the end of a tracking session, each participant was asked to provide a diagnosis by selecting one or more of the 12 classes (Non-proliferative changes only, Fibroadenoma, Intraductal papilloma without atypia, Usual ductal hyperplasia, Columnar cell hyperplasia,



Atypical (Atypical-DuctalHyperplasia) | **Atypical** (Atypical-DuctalHyperplasia) | **Atypical** (Columnar-CellHyperplasia, FlatEpithelialAtypia, AtypicalDuctalHyper-plasia) | **DCIS** (UsualDuctal-Hyperplasia, ColumnarCellHyper-plasia, DuctalCarcinomaIn-Situ) | **Atypical** (UsualDuc-talHyperplasia, SclerosingAdenosis, AtypicalDuctalHyper-plasia) | **DCIS** (AtypicalDuc-talHyperplasia, DuctalCarcinomaIn-Situ)

Figure 1. Viewing behavior of six different pathologists on a whole slide image with a size of 75568 × 74896 pixels. The time spent by each pathologist on different image areas are 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.

Sclerosing adenosis, Complex sclerosing lesion, Flat epithelial atypia, Atypical ductal hyperplasia, Intraductal papilloma with atypia, Ductal carcinoma in situ, Invasive carcinoma) from a pathology form to indicate what she had seen during her screening of the slide. Figure 1 illustrates the data for an example slide.

In the data set used in this paper, the 12 classes were mapped into a more general set of five classes (Non-proliferative changes only, Proliferative changes, Atypical ductal hyperplasia, Ductal carcinoma in situ, Invasive cancer). (The mapping results for the examples in Figure 1 are shown in bold.) Furthermore, we only used the viewing logs from the three expert pathologists, because they were the only ones that evaluated all of the 240 slides. The experiments in Section 4 present results based on individual and combined expert data as well as the consensus labels given to each slide at the end of the consensus meetings by the three experts.

3. METHODOLOGY

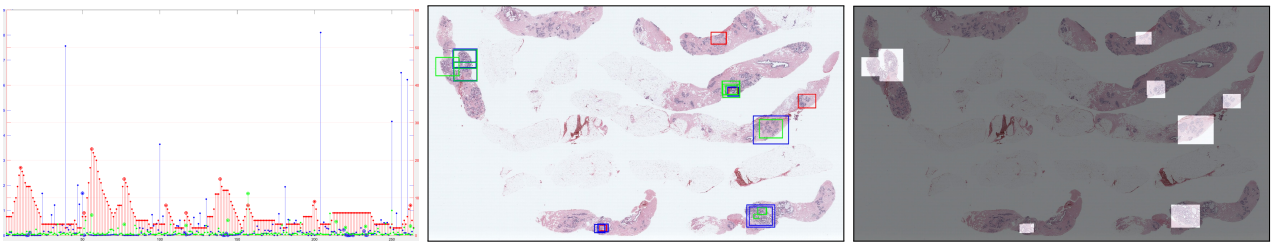
3.1 Identification of Candidate ROIs

The first step is the identification of the candidate ROIs from the pathologists' viewing logs. We used a procedure similar to the one described by Mercan et al.[7] The viewing log for each slide for each pathologist contained four entries per second. Each entry, called a viewport, contained the image coordinates with respect to the top-left pixel on the screen, the screen size, and the zoom level. We calculated the displacement between two log entries as the number of pixels between the centers of two consecutive viewports. Each entry also contained a time stamp that we used to calculate the duration a pathologist viewed a particular rectangular area.

We defined three actions in the viewport data. *Zoom peaks* are the points where the zoom level is higher than those of the previous and the next viewport logs. A zoom peak defines an area where a pathologist intentionally looked closer by zooming in. *Slow pannings* are the points where the zoom level is constant and displacement is small. They are intended for investigating a slightly larger and closer area without completely moving the viewport. *Fixations* are the points where the duration is longer than 2 seconds. A fixation captures the areas that a pathologist investigated longer. Since different pathologists can have very different viewing behavior, a combination of these three actions were necessary to identify the regions of interest where the union of all selected viewport rectangles were marked as a collection of the ROIs that were considered to be candidates for diagnostically important areas. Figure 2 illustrates the viewport logs and resulting ROIs for an example slide.

3.2 Feature Extraction

We used two widely exploited features, L*a*b* and local binary pattern (LBP) histograms, for representing the candidate ROIs. First, we used the CIE-L*a*b* color space to compute three separate color histograms and concatenated them into one. For the second set of features, we performed color deconvolution[8] to obtain a gray-scale image for the structures dyed with haematoxylin and another one for the structures dyed with eosin. Then, we computed the LBP histograms on these two images and concatenated the output vectors. Each histogram



(a) (b) (c)

Figure 2. ROI detection from the viewport logs. (a) Viewport log of a particular pathologist. The x-axis represents the log entry. The red, blue, and green bars represent the zoom level, displacement, and duration, respectively. The three types of selected actions are circled on the bars. (b) The rectangular regions visible on the pathologist's screen at the points selected from the viewport log 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) ROIs resulting from the union of the selected actions.

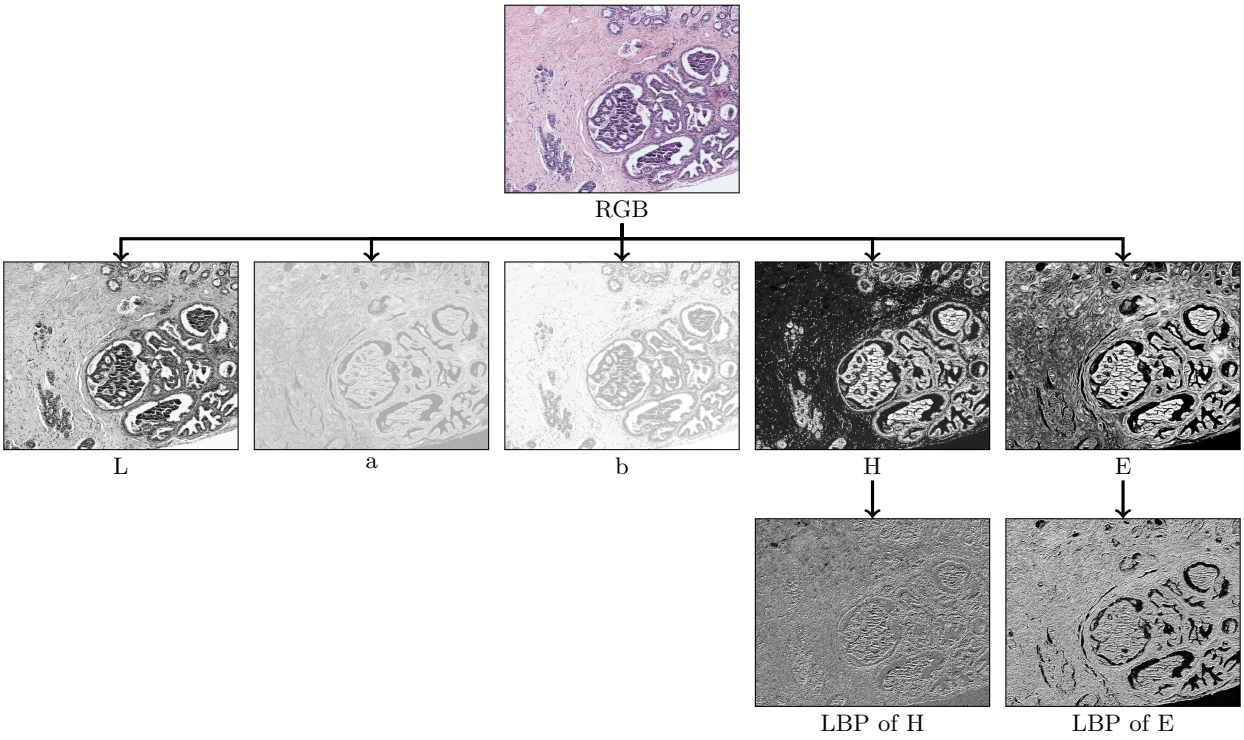


Figure 3. Feature extraction process for an example ROI. Contrast enhancement was performed for better visualization.

contained 64 bins, with a total feature vector length of 320 after concatenation. The features were kept simple as the focus was on learning. Figure 3 shows the feature extraction process for an example candidate ROI.

3.3 Learning and Classification

MIL, MLL, and MIMLL have been highly popular in the machine learning literature. As discussed earlier, many machine learning tasks involve samples that can be described by more than one instance of features and there can be more than one label associated with a sample. The learning problem that corresponds to the former is named MIL, and the latter is called MLL. MIMLL corresponds to the learning problems when both cases are present.

In the MIMLL framework, let a data set $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_M, Y_M)\}$ consist of a collection of samples where M denotes the number of samples. Each sample corresponds to a pair of a bag and a set of labels. A bag X_m contains a set of instances $\{\mathbf{x}_{m1}, \mathbf{x}_{m2}, \ldots, \mathbf{x}_{mn_m}\}$ where n_m is the number of instances in that bag, $\mathbf{x}_{mn} \in \mathbb{R}^d, n = 1, \ldots, n_m$, is the feature vector of the n'th instance of the m'th bag, and d is the number of features. A label set Y_m is composed of class labels $\{y_{m1}, y_{m2}, \ldots, y_{ml_m}\}$ where l_m is the number of labels in that set and $y_{ml} \in \{c_1, c_2, \ldots, c_L\}, l = 1, \ldots, l_m$, is one of L possible class labels. MIL can be considered as a special case of MIMLL where there is only one label $y_m \in \{c_1, c_2, \ldots, c_L\}$ associated with each bag X_m, resulting in a data set of pairs of bags and associated labels, $\{(X_1, y_1), (X_2, y_2), \ldots, (X_M, y_M)\}$. MLL is another special case in which an instance $\mathbf{x}_m \in \mathbb{R}^d$ is associated with a set of labels Y_m, resulting in a data set of pairs of instances and label sets, $\{(\mathbf{x}_1, Y_1), (\mathbf{x}_2, Y_2), \ldots, (\mathbf{x}_M, Y_M)\}$.

In this paper, we study five different algorithms proposed for MIMLL for other domains within the context of simultaneous localization and classification of diagnostically relevant areas in whole slide breast histopathology images. These algorithms are briefly summarized as follows:

1. MimlBoost:[9] This algorithm solves the MIMLL problem in two steps. In the first step, each sample (X_m, Y_m) is decomposed into a set of L bags $\{[(X_m, c_1), \Psi(X_m, c_1)], [(X_m, c_2), \Psi(X_m, c_2)], \ldots, [(X_m, c_L), \Psi(X_m, c_L)]\}$, where $\Psi(X_m, c_l) = +1$ if $c_l \in Y_m$ and $\Psi(X_m, c_l) = -1$ otherwise, by assuming that the labels are independent from each other. The resulting transformation of the MIMLL problem with M samples into an MIL problem with $M \times L$ samples produces the data set $\{[(X_1, c_1), \Psi(X_1, c_1)], [(X_1, c_2), \Psi(X_1, c_2)], \ldots, [(X_1, c_L), \Psi(X_1, c_L)], \ldots, [(X_M, c_1), \Psi(X_M, c_1)], [(X_M, c_2), \Psi(X_M, c_2)], \ldots, [(X_M, c_L), \Psi(X_M, c_L)]\}$. In the second step, the MIL problem is solved using the MiBoosting algorithm[10] that assumes that all instances in a bag contribute independently in an equal way to the label of that bag. Finally, after learning a function that produces a binary output ($+1$ or -1) for an input pair (X, c) consisting of a new bag of instances X and a potential label c, the corresponding multi-label set for X is obtained as the collection of labels for which the output is $+1$.

2. MimlSvmMi: This algorithm is a variation of the MimlBoost algorithm. The first step is the same as the MimlBoost algorithm where the MIMLL problem is decomposed into a set of multi-instance single-label problems. Unlike the second step of MimlBoost, MimlSvmMi uses the MiSvm algorithm[11] to solve the resulting MIL problem.

3. MimlSvm:[9] This algorithm also solves the MIMLL problem in two steps. In the first step, the MIMLL problem is decomposed into a series of single-instance multi-label problems by assuming that the spatial distribution of each bag holds relevant information. Thus, the bags are collected into a set $\bigcup_{m=1}^{M} X_m$, and are clustered using the k-medoids algorithm. In each iteration of the algorithm, distances between the bags are computed using the Hausdorff distance.[12] The Hausdorff distance between two bags, $X_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in_i}\}$ and $X_j = \{\mathbf{x}_{j1}, \mathbf{x}_{j2}, \ldots, \mathbf{x}_{jn_j}\}$, is defined as

$$d_H(X_i, X_j) = \max\left\{ \max_{\mathbf{x}_i \in X_i} \min_{\mathbf{x}_j \in X_j} \|\mathbf{x}_i - \mathbf{x}_j\|, \max_{\mathbf{x}_j \in X_j} \min_{\mathbf{x}_i \in X_i} \|\mathbf{x}_j - \mathbf{x}_i\| \right\} \quad (1)$$

   where $\|\cdot\|$ denotes the Euclidean distance. The algorithm partitions the data space into K clusters, each of which is represented by its medoid, $M_k, k = 1, \ldots, K$. Consequently, a bag, X_m, is transformed into a vector $\mathbf{z}_m \in \mathbb{R}^K$ whose components are the Hausdorff distances between the bag and all K medoids, $\mathbf{z}_m = (d_H(X_m, M_1), d_H(X_m, M_2), \ldots, d_H(X_m, M_K)), m = 1, \ldots, M$. The multi-instance multi-label data set is then transformed into a single-instance multi-label data set, $\{(\mathbf{z}_1, Y_1), (\mathbf{z}_2, Y_2), \ldots, (\mathbf{z}_M, Y_M)\}$. In the second step, the resulting MLL problem is solved using the MlSvm algorithm.[13]

4. MimlNN: This algorithm is a variation of the MimlSvm algorithm. The first step is the same as in MimlSvm where the MIMLL problem is decomposed into a set of single-instance multi-label problems. In the second step, the MlSvm algorithm in MimlSvm is replaced with a two-layer neural network structure[14] to solve the resulting MLL problem.

5. M^{3}MIML:[15] The previous methods reduce the MIMLL problem into either a multi-instance single-label (the first two methods) or a single-instance multi-label (the last two methods) problem where relevant information between instances and labels could be destroyed during the decomposition process. Unlike these methods, M^{3}MIML tries to exploit any connection between instances and labels by defining a maximum margin method (M^3) for MIMLL by assuming a linear model for each class where the output for one class is set to be the maximum prediction of all the MIMLL examples' instances with respect to the corresponding linear model.

For each slide and each pathologist viewing that slide, the input to a particular learning procedure was a bag of candidate ROIs, each one being represented by a color-texture feature vector, and a set of labels assigned to that slide by the pathologist. Each of the algorithms listed above was used to train a multi-class classifier that can be used to assign labels to a new slide.

4. EXPERIMENTS

4.1 Experimental Setting

The parameters of the MIMLL algorithms described above were selected based on our observations as well as the suggestions provided in the respective papers. In particular, 25 boosting rounds were used in MIMLBOOST. MIMLSVM, MIMLSVMMI and M^{3}MIML depend on support vector machines (SVM) for classification. Across the three methods, the SVM parameters were kept the same. The SVM kernel was chosen as the Gaussian kernel with the scale parameter (γ) set to 0.2. In MIMLSVM, the number of clusters (K) was set to 20% of the size of the training set. A two layer neural network structure was used for MIMLNN with the number of clusters (K) set to 40% of the size of the training set and the regularization parameter used in matrix inversion (λ) set to 1.

4.2 Evaluation Criteria

In this paper, quantitative performance was measured by comparing the labels assigned to a slide by an algorithm with the labels given by one or more pathologists. We used five evaluation criteria named the Hamming loss, ranking loss, one-error, coverage, and average precision.[9] Given a test set $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)\}$ with N samples where Y_n is the set of reference (true) labels for the n'th sample, let $h(X_n)$ be a function that returns the predicted set of labels for X_n, $h(X_n, y)$ be a function that returns a value that indicates the confidence that y is in $h(X_n)$, and $r(X_n, y)$ be the rank of y in the sorted list of predicted labels with respect to $h(X_n, y)$ where the rank of the label with the highest confidence ($\arg\max_{y \in \{c_1, c_2, \ldots, c_L\}} h(X_n, y)$) is 1. The evaluation criteria can be computed using these functions as follows:

- $hammingLoss(h) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{L} |h(X_n) \triangle Y_n|$, where $\triangle$ is an operator used to compute the symmetric difference between two label sets. Hamming loss is the fraction of the wrong labels (false positives or false negatives) to the total number of labels. Hence, a smaller value indicates better performance.
- $rankingLoss(h) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{|Y_n||\overline{Y_n}|} |\{(y_1, y_2) | r(X_n, y_1) \geq r(X_n, y_2),\ (y_1, y_2) \in Y_n \times \overline{Y_n}\}|$, where $\overline{Y_n}$ denotes the complement of the set Y_n. Ranking loss counts the number of times the rank of a wrong label is above the rank of a reference label. Hence, a smaller value indicates better performance.
- $one\text{-}error(h) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\left[\arg\min_{y \in \{c_1, c_2, \ldots, c_L\}} r(X_n, y) \notin Y_n\right]$, where $\mathbb{1}$ is an indicator function that is 1 when its argument is true, and 0 otherwise. One-error counts the number of times when the top-ranked label is not one of the reference class labels. Like the criteria above, a smaller value indicates better performance.
- $coverage(h) = \frac{1}{N} \sum_{n=1}^{N} \max_{y \in Y_n} r(X_n, y) - 1$. Coverage is defined as the average number of labels to investigate on the ordered list of predicted numbers (by their ranks) to cover all reference labels. Since fewer number of visits on the list is desired, a smaller value indicates better performance.
- $averagePrecision(h) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{|Y_n|} \sum_{y \in Y_n} \frac{|\{y' | r(X_n, y') \leq r(X_n, y),\ y' \in Y_n\}|}{r(X_n, y)}$. Average precision is the average fraction of reference labels ranked above a particular label. Unlike the rest of the criteria, the performance is better as the average precision gets closer to 1.

Table 1. Summary statistics (average ± standard deviation) for the number of ROIs extracted from the pathologists' viewing logs for the 240 slides. The statistics are given for the whole data set as well as for subsets of the slides based on the individual diagnostic classes from the consensus labels (Non-proliferative changes only (NP), Proliferative changes (P), Atypical ductal hyperplasia (ADH), Ductal carcinoma in situ (DCIS), Invasive cancer (INV)). *All* corresponds to the union of all three experts' ROIs for a particular slide.

Expert	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

4.3 Results

As discussed in Section 2, we used the viewing logs of the three expert pathologists (denoted as *E1*, *E2*, and *E3*) in the experiments. The procedure described in Section 3.1 was used to obtain the candidate ROIs for each expert and each slide, and the features listed in Section 3.2 were extracted from each candidate ROI to construct the bags of instances. The labels assigned to each slide by each pathologist formed the multi-label sets. In addition to the labels assigned by individual pathologists, we also used the labels given to each slide at the end of the consensus meetings by the three experts. Table 1 gives some summary statistics regarding the number of ROIs in the data set. Even with data from only three pathologists, we could observe that there can be differences in the pathologists' screening patterns; some spend more time on a slide and investigate larger number of ROIs, whereas others may make faster decisions by looking at a few key areas. We also observed that the slides with consensus diagnoses of Proliferative changes and Atypical ductal hyperplasia required significantly longer views with more ROIs on the average for all pathologists. Studying the possibility of correlations between different viewing behaviors and diagnostic accuracy and efficiency is part of our future work.

We used four-fold cross-validation to compute the quantitative performance criteria defined above. The partitioning of the data into four subsets of 60 slides was created by an expert pathologist to resemble the distribution of the classes in the whole data set within each subset. All results presented below show the average and standard deviation of each performance criterion by using the four folds.

We conducted two experiments to study different learning and validation scenarios involving various combinations of viewing logs from multiple pathologists. In the first experiment, we used each particular expert's data (candidate ROIs and class labels) as training samples, trained classifiers using these samples, and compared the labels assigned by these classifiers to the ones given by the individual experts. The goal of this experiment was to see how well a classifier trained using the ROIs of one expert could predict the class labels of other slides diagnosed by herself and the other two experts. This experiment was repeated using the data for each of the three experts separately. Tables 2–4 give the resulting performance statistics computed using four-fold cross-validation. The results showed that MimlNN performed the best according to most of the settings, followed by MimlBoost that was also consistently ranked among the top two methods. On the other hand, MimlSvm and MimlSvmMi showed the worst performance according to most of the measures. When we focused on the results for the scenarios based on individual pathologists' data, we observed that the classifiers learned by using the first expert's training data performed the best on the validation data from the same expert. In other words, the classifiers could predict the class labels in the same pathologist's validation data better than the ones in the data from the other two experts. The classifiers built by using the third expert's training data showed a similar result by performing better on the validation data of the third expert on more cases when compared to the results for the validation data of the other two experts. However, the results from the classifiers trained by using the second expert's training data showed a different behavior; better results were obtained by using the other two experts' validation data. One reason for this might be the significantly larger number of ROIs extracted from the viewing logs of the second expert (as seen from Table 1) and potential similarities between these ROIs and the ones from the other two experts. When all results were compared, we could conclude that the classifiers learned by using the third expert's data resulted in higher average precision and lower Hamming loss, ranking loss, one-error and coverage values, indicating better performance compared to the other classifiers in our particular experimental settings.

Table 2. Results of the experiments when the first expert's (*E1*) data (candidate ROIs and class labels) were used for training and each individual experts' data were used for validation. The evaluation criteria are: Hamming loss (HL), ranking loss (RL), one-error (OE), coverage (COV), and average precision (AP). The best result for each criterion is marked in bold.

	Validation data: *E1*				
	HL	RL	OE	COV	AP
MimlBoost	**0.2000 ± 0.0000**	0.3138 ± 0.0166	0.6528 ± 0.0278	1.2554 ± 0.0665	0.5921 ± 0.0186
MimlSvmMi	0.2017 ± 0.0033	0.3968 ± 0.0691	0.6612 ± 0.0326	1.5746 ± 0.2612	0.5613 ± 0.0352
MimlSvm	0.3047 ± 0.0302	0.4070 ± 0.0479	0.7617 ± 0.0755	1.6280 ± 0.1914	0.5079 ± 0.0534
MimlNN	**0.2000 ± 0.0000**	**0.2877 ± 0.0192**	**0.6319 ± 0.0177**	**1.1509 ± 0.0769**	**0.6121 ± 0.0186**
M^{3}Miml	0.2645 ± 0.0141	0.3160 ± 0.0209	0.6612 ± 0.0353	1.2639 ± 0.0837	0.5883 ± 0.0226
	Validation data: *E2*				
	HL	RL	OE	COV	AP
MimlBoost	**0.2000 ± 0.0000**	0.3286 ± 0.0270	**0.6312 ± 0.1020**	1.2636 ± 0.0165	0.5783 ± 0.0193
MimlSvmMi	0.2034 ± 0.0068	0.4805 ± 0.0617	0.7787 ± 0.0789	1.9052 ± 0.2631	0.4757 ± 0.0513
MimlSvm	0.2745 ± 0.0228	0.4165 ± 0.0314	0.6862 ± 0.0569	1.6660 ± 0.1257	0.5394 ± 0.0290
MimlNN	**0.2000 ± 0.0000**	**0.3119 ± 0.0348**	0.6616 ± 0.0901	**1.2476 ± 0.1393**	**0.5884 ± 0.0540**
M^{3}Miml	0.2778 ± 0.0166	0.3285 ± 0.0214	0.6946 ± 0.0416	1.3142 ± 0.0856	0.5684 ± 0.0219
	Validation data: *E3*				
	HL	RL	OE	COV	AP
MimlBoost	**0.2000 ± 0.0000**	0.3275 ± 0.0279	0.6696 ± 0.0399	1.2850 ± 0.0937	0.5806 ± 0.0257
MimlSvmMi	0.2008 ± 0.0017	0.4436 ± 0.0664	0.7449 ± 0.0665	1.7701 ± 0.2677	0.5053 ± 0.0497
MimlSvm	0.2812 ± 0.0192	0.3848 ± 0.0773	0.7029 ± 0.0481	1.5391 ± 0.3093	0.5467 ± 0.0484
MimlNN	**0.2000 ± 0.0000**	**0.2898 ± 0.0154**	**0.6484 ± 0.0267**	**1.1590 ± 0.0618**	**0.6039 ± 0.0119**
M^{3}Miml	0.2962 ± 0.0147	0.3609 ± 0.0159	0.7404 ± 0.0368	1.4434 ± 0.0637	0.5325 ± 0.0216

Table 3. Results of the experiments when the second expert's (*E2*) data (candidate ROIs and class labels) were used for training and each individual experts' data were used for validation. The evaluation criteria are: Hamming loss (HL), ranking loss (RL), one-error (OE), coverage (COV), and average precision (AP). The best result for each criterion is marked in bold.

	Validation data: *E1*				
	HL	RL	OE	COV	AP
MimlBoost	**0.2000 ± 0.0000**	0.3784 ± 0.0410	**0.5863 ± 0.1306**	1.3178 ± 0.0565	0.5747 ± 0.0205
MimlSvmMi	0.2058 ± 0.0096	0.4281 ± 0.0822	0.7194 ± 0.1093	1.6747 ± 0.2885	0.5265 ± 0.0693
MimlSvm	0.2861 ± 0.0155	0.4267 ± 0.0282	0.7153 ± 0.0386	1.7067 ± 0.1128	0.5210 ± 0.0189
MimlNN	**0.2000 ± 0.0000**	**0.3158 ± 0.0208**	0.6316 ± 0.0458	**1.2632 ± 0.0832**	**0.5978 ± 0.0290**
M^{3}Miml	0.2678 ± 0.0111	0.3284 ± 0.0190	0.6695 ± 0.0278	1.3135 ± 0.0761	0.5815 ± 0.0197
	Validation data: *E2*				
	HL	RL	OE	COV	AP
MimlBoost	**0.2000 ± 0.0000**	**0.3231 ± 0.0303**	**0.6357 ± 0.0558**	**1.2715 ± 0.0965**	**0.5972 ± 0.0359**
MimlSvmMi	0.2025 ± 0.0051	0.4917 ± 0.0199	0.7953 ± 0.0635	1.9375 ± 0.0832	0.4657 ± 0.0272
MimlSvm	0.2811 ± 0.0219	0.4236 ± 0.0434	0.7028 ± 0.0547	1.6944 ± 0.1734	0.5321 ± 0.0370
MimlNN	**0.2000 ± 0.0000**	0.3338 ± 0.0264	0.6989 ± 0.0414	1.3352 ± 0.1057	0.5666 ± 0.0287
M^{3}Miml	0.2778 ± 0.0117	0.3473 ± 0.0135	0.6945 ± 0.0294	1.3893 ± 0.0539	0.5605 ± 0.0122
	Validation data: *E3*				
	HL	RL	OE	COV	AP
MimlBoost	**0.2000 ± 0.0000**	0.3650 ± 0.0561	**0.5562 ± 0.1357**	1.3050 ± 0.0873	0.5816 ± 0.0230
MimlSvmMi	0.2008 ± 0.0017	0.4699 ± 0.0452	0.7407 ± 0.0309	1.8752 ± 0.1748	0.4974 ± 0.0273
MimlSvm	0.2660 ± 0.0216	0.3680 ± 0.0568	0.6651 ± 0.0541	1.4720 ± 0.2273	0.5675 ± 0.0469
MimlNN	**0.2000 ± 0.0000**	**0.3075 ± 0.0070**	0.6483 ± 0.0613	**1.2301 ± 0.0280**	**0.5976 ± 0.0250**
M^{3}Miml	0.2977 ± 0.0290	0.3743 ± 0.0288	0.7444 ± 0.0726	1.4974 ± 0.1151	0.5259 ± 0.0406

Table 4. Results of the experiments when the third expert's (*E3*) data (candidate ROIs and class labels) were used for training and each individual experts' data were used for validation. The evaluation criteria are: Hamming loss (HL), ranking loss (RL), one-error (OE), coverage (COV), and average precision (AP). The best result for each criterion is marked in bold.

	Validation data: *E1*				
	HL	RL	OE	COV	AP
MimlBoost	**0.2000 ± 0.0000**	0.3267 ± 0.0330	0.6502 ± 0.0674	1.2528 ± 0.0353	0.5795 ± 0.0047
MimlSvmMi	0.2034 ± 0.0068	0.4087 ± 0.0919	0.6955 ± 0.0967	1.6350 ± 0.3675	0.5412 ± 0.0738
MimlSvm	0.2851 ± 0.0135	0.3994 ± 0.0306	0.7128 ± 0.0338	1.5977 ± 0.1224	0.5317 ± 0.0273
MimlNN	0.2017 ± 0.0044	**0.2977 ± 0.0191**	**0.6457 ± 0.0279**	**1.1907 ± 0.0762**	**0.6010 ± 0.0138**
M^{3}Miml	0.2785 ± 0.0088	0.3175 ± 0.0049	0.6963 ± 0.0219	1.2698 ± 0.0198	0.5724 ± 0.0080
	Validation data: *E2*				
	HL	RL	OE	COV	AP
MimlBoost	0.2000 ± 0.0000	0.3260 ± 0.0440	**0.5736 ± 0.0294**	**1.1853 ± 0.0914**	**0.6077 ± 0.0319**
MimlSvmMi	**0.1992 ± 0.0017**	0.4439 ± 0.0259	0.7252 ± 0.0633	1.7714 ± 0.1034	0.5091 ± 0.0306
MimlSvm	0.2989 ± 0.0217	0.4708 ± 0.0652	0.7472 ± 0.0542	1.8831 ± 0.2610	0.4908 ± 0.0517
MimlNN	0.2017 ± 0.0019	**0.3026 ± 0.0237**	0.6540 ± 0.0215	1.2105 ± 0.0947	0.5970 ± 0.0174
M^{3}Miml	0.2700 ± 0.0084	0.3153 ± 0.0156	0.6751 ± 0.0209	1.2611 ± 0.0624	0.5814 ± 0.0146
	Validation data: *E3*				
	HL	RL	OE	COV	AP
MimlBoost	**0.2000 ± 0.0000**	0.3425 ± 0.0531	**0.5232 ± 0.1130**	1.2272 ± 0.0839	0.5900 ± 0.0319
MimlSvmMi	0.2008 ± 0.0017	0.4335 ± 0.0672	0.7381 ± 0.0538	1.7342 ± 0.2688	0.5108 ± 0.0387
MimlSvm	0.2754 ± 0.0213	0.3536 ± 0.0407	0.6884 ± 0.0532	1.4145 ± 0.1628	0.5620 ± 0.0345
MimlNN	0.2017 ± 0.0034	**0.2754 ± 0.0183**	0.6161 ± 0.0453	**1.1017 ± 0.0730**	**0.6225 ± 0.0263**
M^{3}Miml	0.2920 ± 0.0098	0.3374 ± 0.0099	0.7299 ± 0.0244	1.3498 ± 0.0394	0.5484 ± 0.0109

Table 5. Results of the experiments when the union of all experts' data (candidate ROIs and class labels) were used for training. Validation labels consisted of the union of experts' 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 criterion is marked in bold.

	Validation data: *E1* ∪ *E2* ∪ *E3*				
	HL	RL	OE	COV	AP
MimlBoost	0.2633 ± 0.0027	0.4104 ± 0.0438	**0.4958 ± 0.0685**	1.5708 ± 0.1125	0.6205 ± 0.0315
MimlSvmMi	0.2608 ± 0.0032	0.4990 ± 0.1115	0.6917 ± 0.1524	2.2375 ± 0.3134	0.5170 ± 0.0832
MimlSvm	0.2633 ± 0.0027	0.3472 ± 0.0393	0.5500 ± 0.0624	1.7583 ± 0.1309	0.6276 ± 0.0398
MimlNN	**0.2558 ± 0.0074**	**0.2931 ± 0.0134**	0.5667 ± 0.0333	**1.5083 ± 0.0645**	**0.6391 ± 0.0153**
M^{3}Miml	0.2633 ± 0.0027	0.3312 ± 0.0062	0.6042 ± 0.0344	1.6792 ± 0.0160	0.6047 ± 0.0118
	Validation data: *Consensus*				
	HL	RL	OE	COV	AP
MimlBoost	**0.2000 ± 0.0000**	0.4354 ± 0.0380	**0.5667 ± 0.0680**	1.3250 ± 0.0908	0.5638 ± 0.0289
MimlSvmMi	0.2008 ± 0.0032	0.4760 ± 0.1150	0.7458 ± 0.1243	1.8167 ± 0.3774	0.4926 ± 0.0883
MimlSvm	0.2633 ± 0.0176	0.3573 ± 0.0361	0.6583 ± 0.0441	1.4292 ± 0.1443	0.5785 ± 0.0304
MimlNN	0.2108 ± 0.0140	**0.3042 ± 0.0285**	0.6542 ± 0.0644	**1.2167 ± 0.1139**	**0.5940 ± 0.0371**
M^{3}Miml	0.2900 ± 0.0159	0.3469 ± 0.0239	0.7250 ± 0.0397	1.3875 ± 0.0956	0.5456 ± 0.0219

In the second experiment, the feature data for each slide were obtained by taking the union of all candidate ROIs from all three experts for that slide. We also combined the label sets assigned by all experts during their screening sessions as the training label data. Validation labels were similarly obtained as the union of the experts' class labels for each slide. We also compared the predictions by the resulting classifiers to the consensus diagnoses as the validation labels in a separate experiment. Table 5 gives the resulting performance statistics computed using four-fold cross-validation. The results indicated a better performance in most of the criteria, including

the highest average precision among all settings, when the union of all experts' class labels was used as the validation data for each slide. As expected, better performances were observed when the training and validation data were obtained according to the same protocol, compared to the results where consensus diagnoses were used as validation labels. We also would like to note that consensus diagnoses contained a single label for each slide, whereas union of individual experts' labels could correspond to a multi-label setting during both training and validation, and some performance criterion (e.g., coverage) were more sensitive to the number of labels than others. In future work, we 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.

5. CONCLUSIONS

This paper presented a study on the use of pathologists' viewing records of digital biopsy slides for generating data sets to train classifiers that can predict the diagnostic categories of whole slide breast histopathology images. Contrary to classical supervised learning applications where the classifiers are built and tested on manually selected regions of interest corresponding to isolated tissue structures with no ambiguity regarding their diagnoses, the proposed framework used weakly learning methods that have been designed to tackle the uncertainty regarding the correspondence between the particular image areas and the diagnostic labels provided by the pathologists at the slide level. We represented each slide as a bag of candidate ROIs extracted from the viewport logs of pathologists' image screenings together with a list of class labels extracted from the pathology forms. The learning stage exploited five different multi-instance multi-label learning algorithms where the multi-instance component corresponded to the bags of ROIs and the multi-label component corresponded to the slide-level annotations. The experiments studied different learning and validation scenarios involving various combinations of data from three expert pathologists. Quantitative performance measures that compared the slide-level predictions of the classifiers with the validation data showed that average precision values up to 62% were obtained when the training and validation data came from the same individual pathologist's viewing logs, whereas average precision increased to 64% when the candidate ROIs and the labels from all pathologists were combined for each slide.

ACKNOWLEDGMENTS

Caner Mercan and Selim Aksoy were supported in part by the Scientific and Technological Research Council of Turkey under Grant No. 113E602. Ezgi Mercan, Linda G. Shapiro, Donald L. Weaver, and Joann G. Elmore were supported in part by the National Cancer Institute of the National Institutes of Health under Award No. R01-CA172343. 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. Selim Aksoy was also supported in part by the GEBIP Award from the Turkish Academy of Sciences.

REFERENCES

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

[2] Dundar, M. M., Badve, S., Raykar, V. C., Jain, R. K., Sertel, O., and Gurcan, M. N., "A multiple instance learning approach toward optimal classification of pathology slides," in [*International Conference on Pattern Recognition*], 2732–2735 (2010).

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

[4] Xu, Y., Jiao, L., Wang, S., Wei, J., Fan, Y., Lai, M., and Chang, E. I.-C., "Multi-label classification for colon cancer using histopathological images," *Microscopy Research and Technique* **76**(12), 1266–1277 (2013).

[5] Xu, Y., Mo, T., Feng, Q., Zhong, P., Lai, M., and Chang, E. I.-C., "Deep learning of feature representation with multiple instance learning for medical image analysis," in [*IEEE International Conference on Acoustics, Speech and Signal Processing*], 1626–1630 (2014).

[6] Cosatto, E., Laquerre, P.-F., Malon, C., Graf, H.-P., Saito, A., Kiyuna, T., Marugame, A., and Kamijo, K., "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*], **867605** (2013).

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

[8] Ruifrok, A. and Johnston, D., "Quantification of histochemical staining by color deconvolution," *Analytical and Quantitative Cytology and Histology* **23**(4), 291–299 (2001).

[9] Zhou, Z.-H., Zhang, M.-L., Huang, S.-J., and Li, Y.-F., "Multi-instance multi-label learning," *Artificial Intelligence* **176**(1), 2291–2320 (2012).

[10] Xu, X. and Frank, E., "Logistic regression and boosting for labeled bags of instances," in [*Advances in Knowledge Discovery and Data Mining*], 272–281 (2004).

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

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

[13] Boutell, M. R., Luo, J., Shen, X., and Brown, C. M., "Learning multi-label scene classification," *Pattern Recognition* **37**(9), 1757–1771 (2004).

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

[15] Zhang, M.-L. and Zhou, Z.-H., "M3MIML: A maximum margin method for multi-instance multi-label learning," in [*IEEE International Conference on Data Mining*], 688–697 (2008).