JMI Journal of Medical Imaging

RESEARCH PAPER

Patch relevance estimation and multilabel augmentation for weakly supervised histopathology image classification

Bulut Aygunes[a], **Ramazan Gokberk Cinbis**[b,*] **and Selim Aksoy**[a,*]
[a]Bilkent University, Department of Computer Engineering, Ankara, Turkey
[b]Middle East Technical University, Department of Computer Engineering, Ankara, Turkey

ABSTRACT. **Purpose:** Weakly supervised learning (WSL) is widely used for histopathological image analysis by modeling images as sets of fixed-size patches and utilizing image-level diagnoses as weak labels. However, in multiclass classification scenarios, patches corresponding to a wide spectrum of diagnostic categories can co-exist in a single image, complicating the learning process. We aim to address label uncertainty in such multiclass settings.

Approach: We propose a two-branch architecture and a complementary training strategy to improve patch-based WSL. One branch estimates patch-level class likelihoods, whereas the other predicts per-class patch relevance weights. These outputs are combined into image-level class predictions via a relevance-weighted sum of per-patch class likelihoods. To further improve performance, we introduce a multilabel augmentation strategy that forms new training samples by combining patch sets and labels from pairs of images, resulting in multilabel samples that enrich the training set by increasing the chance of having more patches that are relevant to the augmented label sets.

Results: We evaluate our method on two challenging multiclass breast histopathology datasets for region of interest classification. The proposed architecture and training strategy outperform conventional weakly supervised methods, demonstrating improved classification accuracy and robustness, particularly in underrepresented classes.

Conclusions: The proposed architecture effectively models the complex relationship between image-level labels and patch-level content in multiclass histopathological image analysis. Combined with the image-level multilabel augmentation strategy, it improves learning under label uncertainty. These contributions hold potential for more accurate and scalable diagnostic support systems in digital pathology.

Keywords: digital pathology; breast histopathology; weakly supervised learning; region of interest classification; multilabel data augmentation

Paper 25129SSR received Apr. 23, 2025; revised Oct. 22, 2025; accepted Nov. 12, 2025; published Dec. 5, 2025.

1 Introduction

Histopathological image analysis serves as an important tool for helping pathologists with the diagnostic process for cancer. Advances in scanning technology have made whole slide images

*Address all correspondence to Selim Aksoy, saksoy@cs.bilkent.edu.tr; Ramazan Gokberk Cinbis, gcinbis@ceng.metu.edu.tr

(WSIs) that are obtained by digitizing biopsy slides at high magnification the main source of data for image analysis. The most common setting for WSI analysis is binary classification as benign versus cancer. However, WSIs often contain many regions of interest (ROIs) that can belong to different diagnostic categories and can carry different levels of relevance for the slide-level diagnosis.

Breast histopathology provides an example scenario where a multiclass analysis becomes crucial as breast cancer patients can face a variety of clinical actions, such as surgery, radiation, or hormonal therapy, depending on the diagnosis made by the pathologists for the biopsy samples.[1] Different types of proliferations in the tissue structures carry different risks of progressing into malignancy; thus, the accuracy of the diagnosis in a fine-grained multiclass setting becomes critical. Furthermore, the pathologists do not have any restrictions on the ROI size when they evaluate the slides, and they can select and study the regions at any size and magnification deemed suitable.[2] Therefore, multiclass classification of arbitrarily sized ROIs continues to be an important problem that serves as a necessary step in the diagnosis of cancer.

The main challenges in histopathological image classification include class imbalance and label uncertainty. The first challenge is typically handled with data augmentation techniques such as random rotations, flips, scaling, and staining variations to artificially increase the number of samples.[3] However, these image-domain augmentations mainly aim to increase the number of patches, where the increased diversity may be useful for tasks such as patch classification, but their effectiveness is not sufficiently explored for slide-level learning scenarios.[4]

The second challenge is often studied in the weakly supervised learning (WSL) framework that relies on image-level labels without explicit correspondence between local image regions and these labels.[5] In the most commonly used WSL setup, multiple instance learning (MIL), each image is treated as a bag, fixed-sized patches sampled from the image constitute the instances in the bag, and aggregation of instance-level representations is used to make a bag-level decision. In the classical definition of MIL that corresponds to the binary classification scenario, learning is done using positively and negatively labeled bags where each positive bag is assumed to contain at least one positive instance, whereas all instances in a negative bag are treated as negative.

WSL has already become a well-established approach for the analysis of histopathological images. However, although the aforementioned MIL assumption for positive and negative bags holds for the binary (e.g., cancer versus benign) setting, the relationship between the individual patches and the image-level label can become more complex for the multiclass classification scenario. First, patches can belong to a wider spectrum of distinct diagnostic categories. For the example case of breast histopathology, the continuum of histologic features such as usual ductal hyperplasia, atypical ductal hyperplasia, ductal carcinoma *in situ*, and invasive carcinoma all have different clinical significance.[6] Furthermore, individual patches that belong to different parts of this spectrum can co-exist within the same image neighborhood. However, the image-level labelings are determined by considering the most severe diagnosis observed within the image,[1] and thus, patches from multiple categories can be sampled and grouped under a single label corresponding to that most severe category. Consequently, the image-level diagnosis that only serves as a weak label for the constituent patches introduces label noise, making learning more challenging, especially when the sample sizes are small. Moreover, the inherent uncertainty in the ROI detection process, both with manual annotation using rectangular regions and with automatic segmentation methods,[7] further increases the mismatch between the patches and the categories used during the learning process.

In this paper, we propose a new architecture along with a training strategy that is capable of handling the challenges of multiclass classification of arbitrarily sized ROIs with label uncertainty in the learning process (We define ROIs as image regions that are identified to be diagnostically relevant and are labeled with the most severe category that exists in the region. The low number of slides in our multiclass dataset limits our focus to ROI-level analysis but all of the proposed methods can readily work for WSI-level analysis as well.). The first contribution is a two-branch architecture for WSL with patch relevance estimation. A patch can be highly informative, containing parts of important structures relevant for the reference diagnosis, or may be carrying clues of structures belonging to other categories, or may even be just uninformative by belonging to background or connective tissue. In the proposed architecture that aims to learn the relevance of the individual patches with the ROI-level label, one of the branches estimates

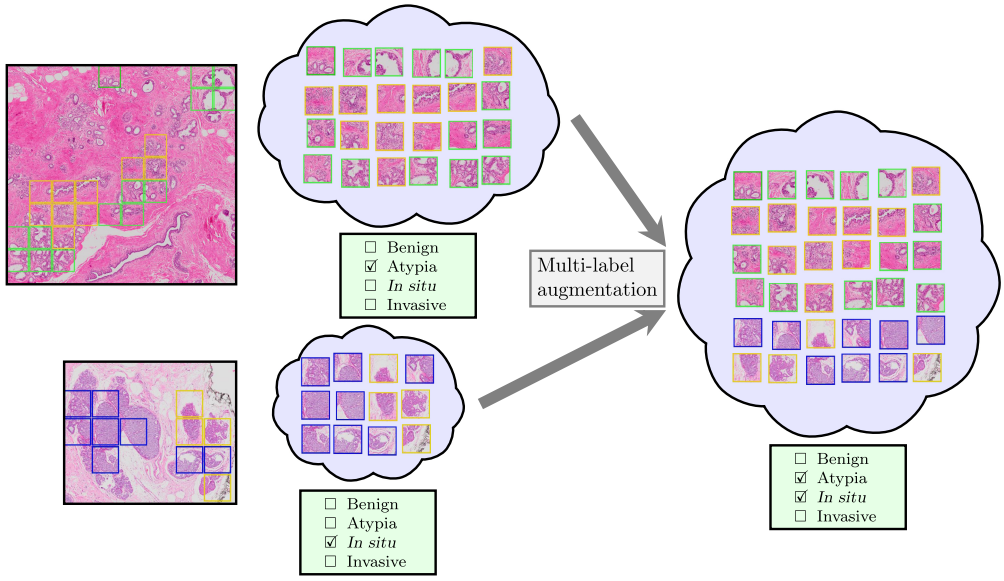

Fig. 1 Illustration of the multilabel augmentation strategy. Individual patches that belong to different parts of a wide spectrum of diagnostic categories can co-exist within the same ROI. However, the ROI is typically annotated with the single most severe category observed. For example, the ROI at the top has patches belonging to both atypical ductal hyperplasia (marked with yellow boundaries) and flat epithelial hyperplasia (benign, marked with green boundaries), whereas the ROI below has patches belonging to both ductal carcinoma *in situ* (marked with blue boundaries) and atypical ductal hyperplasia (marked with yellow boundaries). The proposed multilabel augmentation strategy forms new samples (on the right) as unions of single-label sets of patches with an increased likelihood of improved correspondence between individual patches and the multilabel augmented annotations. Note that the colored markings of the individual patches are used only for illustration purposes as the learning algorithm sees only the image-level annotations.

the patch-level likelihood of each class, whereas the other branch estimates the per-class patch relevance weight. Then, their outputs are combined to obtain the ROI-level class predictions as the relevance-weighted sum of per-patch class likelihoods.

The second contribution is a training strategy inspired by mixup for ROI-level multilabel augmentation. The proposed method forms unions of pairs of ROIs together with their individual sets of patches and class labels to generate new training samples, as shown in Fig. 1. In breast histopathology, for instance, patches showing atypical ductal hyperplasia and ductal carcinoma *in situ* often appear in a close proximity within the same ROI, and benign hyperplasia may occur near more severe lesions.[6] By artificially creating multilabel ROIs, we aim to (i) increase the likelihood of having more patches relevant to the augmented label sets and (ii) expand the dataset with more ROIs labeled with underrepresented diagnoses to alleviate class imbalance. The resulting multilabel learning formulation promotes simultaneous recognition of multiple diagnostic categories within the augmented ROIs. The effectiveness of these two contributions is illustrated using two multiclass breast histopathology datasets in comparative experiments.

The paper is organized as follows. Section 2 discusses the related work. Section 3 introduces the dataset used in the experiments and describes the proposed methodology. Section 4 presents the experimental results. Finally, Sec. 5 gives the conclusions.

2 Related Work

Consistent with the aims of this paper, discussions of related work focus on both weakly supervised learning and data augmentation.

2.1 Weakly Supervised Learning

The generally studied setting in the recent literature has been to aggregate patch-level representations into image-level classification scores. First, a feature extractor module computes a feature

embedding vector for each patch. Then, an aggregation module pools the patch embeddings into an image-level representation. Finally, a classification module maps this representation to class labels. The most commonly used approach for aggregation has been attention-based pooling in an MIL framework. For example, Ilse et al.[8] utilized attention scores that are estimated using a neural network as weights in the weighted combination of instance representations into a bag-level representation in fixed-sized ROIs. The bag representation is further processed by a bag-level classifier in the traditional binary setting. Campanella et al.[5] employed a patch classifier that is trained using a standard MIL procedure to rank the patches in a slide and input the top 10 patches as an ordered sequence to a recurrent neural network for slide-level aggregation. We propose a generic feature representation for arbitrarily sized ROIs using weighted aggregation of the feature representations of fixed-sized patches sampled from these ROIs.[9] Both the patch-level feature representations and the weights are obtained from a convolutional network trained on patches sampled from ROIs in the training data. Lu et al.[10] presented a method named clustering-constrained-attention MIL that uses attention-based learning to identify regions of high diagnostic value. Multiple parallel attention branches are used to predict attention scores, and the slide-level representation is obtained as the average of all patches weighted by their respective scores. Furthermore, an additional supervised learning task tries to separate the most and least attended patches of each class into distinct clusters to refine the feature space. Li et al.[11] proposed a dual-stream MIL network where the first stream uses max-pooling to select the highest scored instance and the second stream computes attention scores for all patches based on their feature distances to this selected instance. The method also learns the input instance embeddings by self-supervised contrastive learning. Shao et al.[12] also emphasized the importance of correlation among the instances. The method uses two transformer layers to aggregate the instance representations. It also arranges all instances into a square form where convolutional kernels are used to aggregate the correlations. In our work, we adapt the WSL formulation for object detector training,[13] which was also used for multisource fine-grained object recognition.[14]

2.2 Image-Level Data Augmentation

The common assumption in image-level augmentation methods in the literature is that sampling instances (patches) from a bag (WSI or ROI) does not affect the overall diagnosis for the bag while reducing the computational and memory requirements for MIL methods. After decreasing and fixing the number of patches, recent augmentation methods use the mixup[15] strategy that constructs artificial training examples as linear interpolations of feature vectors and linear interpolations of their associated class labels. For example, Yang et al.[4] used k-means clustering to group patches into the same number of clusters in each WSI (bag) and mixed the cluster centroids from pairs of bags of the same class into a new bag of the same class. Zhang et al.[16] formed pseudo-bags by randomly sampling patches of a bag into several smaller bags. These pseudo-bags are assigned the same label as the parent bag and are used to train a new classifier. Gadermayr et al.[17] used the original mixup formulation by first sampling a fixed number of patches from each bag and interpolating the samples into new bags. They also generated synthetic descriptors from patches of the same bag for augmentation within that bag. Chen and Lu[18] used a network to assign scores to patches, ranked the patches using these scores, and selected a fixed number of patches from the top of the ranked list to interpolate the bags using the original mixup formulation. Keum et al.[19] summarized the patches into a fixed number of slots using pooling by attention so that the original mixup formulation can be used with bags containing the same number of slots and one representative vector from each slot. Liu et al.[20] clustered patches in a bag and sample patches from these clusters to form pseudo-bags for each WSI. Then, pseudo-bags from two WSIs are mixed by random selection to form a pseudo-WSI, and the class labels of these WSIs are interpolated to label this pseudo-WSI.

Most of the methods discussed above[5,8,11,16–19] focus on the binary classification setting. Our method tackles the multiclass classification problem by modeling the relevance of all patches for all possible classes. Furthermore, all of the aforementioned image-level augmentation approaches[4,16–20] require having the same number of patches in each bag via size alignment through sampling and grouping, which can lead to loss of important and diagnostically relevant but rare patches. Our method handles images with arbitrary numbers of patches and their unions in a multilabel learning process with no restriction on the image and patch formations.

3 Materials and Methods

In this section, first, we introduce the in-house dataset used in the experiments. Next, we provide a formal definition of the targeted histopathology ROI classification problem. Then, we explain our main approach that aims to distinguish the relevant patches within an ROI from the others through a patch relevance estimation (PatchRel) scheme. Finally, we describe our proposed ROI augmentation (ROIAugment) method, which helps to enrich the ROI classifier network training process.

3.1 Dataset

We use a dataset that contains 1376 ROIs extracted from 121 WSIs that were digitized from hematoxylin and eosin-stained specimens belonging to 99 different patients. The specimens were selected from the archives of the Department of Pathology at Hacettepe University based on their slide-level diagnoses (with approval from the Hacettepe University Non-invasive Clinical Research Ethics Board). The WSIs were acquired at 40× magnification using an Olympus slide scanner, resulting in an average image size of $170{,}000 \times 132{,}000$ pixels. The ROIs were annotated by experienced pathologists with no restriction on the sizes and shapes of the image masks. The resulting annotations were collected into four diagnostic classes: benign (including samples containing nonproliferative changes, apocrine metaplasia, usual ductal hyperplasia, columnar cell hyperplasia, flat epithelial hyperplasia, and intraductal papilloma without atypia), atypia (including samples containing atypical ductal hyperplasia, atypical lobular hyperplasia, and intraductal papilloma with atypia), *in situ* carcinoma (including both ductal carcinoma in situ and lobular carcinoma in situ), and invasive carcinoma. Table 1 shows the class-specific ROI size statistics and the associated high variation in the samples in this dataset.

The specimens also have a high variation in their staining as they were prepared at different times. We performed stain normalization by matching the histograms of the hematoxylin and eosin channels of each slide to the hematoxylin and eosin histograms of a slide chosen as the target from the dataset.[21] To obtain the histograms, we applied color deconvolution[22] to each slide using a unique stain matrix estimated for that slide. The hematoxylin stain vector was estimated using the pixels obtained from a nuclei mask extracted by a pre-trained network, and the eosin stain vector was estimated from the rest of the slide by eliminating the high luminosity regions.

We partitioned the dataset into fourfolds by stratified sampling while also making sure that slides from the same patient as well as ROIs from the same slide go to the same fold. Sampling was done using a genetic algorithm that rewards the splitting of the dataset into similar numbers of ROIs and slides in each fold. Two of the folds were randomly selected to form the training set, and the remaining two folds were used as validation and test sets, respectively. Table 2 shows the ROI- and slide-level class distributions and the associated imbalance in this challenging dataset.

Additional experiments are performed using the publicly available BReAst Carcinoma Subtyping (BRACS)[23] dataset, which is described in Sec. 4.3.

3.2 Problem Definition

The goal is to classify ROIs into one of the predefined classes, where each ROI is defined as an arbitrary-sized region on a slide imagery with a distinctive pattern of diagnostic interest. Although an ROI is expected to cover a region containing a particular pattern, they are typically

Table 1 ROI size statistics per diagnostic class in number of pixels at 10× magnification in the Hacettepe dataset. Rows show the average ROI size, the standard deviation of ROI sizes, and the ratio of the largest ROI size to the smallest one.

	Benign	Atypia	*In situ*	Invasive
Average	1966 *K*	473 *K*	4227 *K*	13528 *K*
Standard deviation	4207 *K*	687 *K*	8325 *K*	19263 *K*
Max–min ratio	2216.0	704.5	1989.1	762.5

Table 2 Class distribution of slides and ROIs in training, validation, and test sets in the Hacettepe dataset. Note that a slide can contain multiple ROIs corresponding to different diagnostic labels, resulting in a multilabel setting for each slide. Thus, the numbers of slides for each diagnostic class in the table do not sum up to the total number of slides for a given set. We focus on ROI-level classification in this paper.

		Benign	Atypia	*In situ*	Invasive	Total
Slide	Training set	42	20	22	15	53
	Validation set	18	11	10	7	23
	Test set	20	10	11	7	26
	Total	80	41	43	29	102
ROI	Training set	291	73	192	118	674
	Validation set	147	38	128	59	372
	Test set	144	35	95	56	330
	Total	582	146	415	233	1376

nonhomogeneous and therefore may contain irrelevant patches, in addition to those containing the specific pattern characterizing the most severe diagnosis in the ROI. Therefore, the problem is not suitable to be cast as a traditional texture classification problem. Similarly, global ROI shape and size characteristics are not typically informative. Therefore, it is, in principle, crucial to utilize distinctive patch-level patterns that may appear in arbitrary parts of an ROI while evaluating the ROI holistically in the classification process.

To learn to classify ROIs, we assume the availability of a training set of ROIs, where each ROI R_i corresponds to an image x_i. During training, we also have access to ROI-level annotations, where the i'th ROI is annotated with the label y_i indicating one of the C classes. In practice, ROIs may be generated via manual delineations by experts[2] or through algorithmic techniques such as semantic segmentation.[7] The approach described in this section aims to be agnostic to the ROI generation method. We only assume that the same ROI generation method is used at validation and test time for consistency across data splits.

3.3 Patch Relevance Estimation for ROI Classification

In our method, we represent each ROI as a set of 224×224 pixel patches sampled at regular intervals from the ROI, i.e., $R_i = \{x_{ij} \in \mathbb{R}^{3\times224\times224}\}_{j=1}^{|R_i|}$. This process results in a mixture of patches with some patches carrying differential information for the classification of the ROI and some patches being uninformative, confusing, or even misleading due to heterogeneity of the local cues. Given that each ROI is annotated by a single diagnostic label, which does not provide per-patch supervision, we aim to estimate the relevance of each patch during the ROI classification process.

To realize such a joint local–global classification model, we first define a patch encoder network ϕ that takes the 224×224 patches and maps them to $F \times 1$-dimensional vectors. To estimate the patch-level likelihood of each diagnostic class, we define a patch-classification layer

$$p(\bar{y}_{ij} = c|x_{ij}) = \frac{\exp \omega_c^T \phi(x_{ij})}{\sum_{c'=1}^{C} \exp \omega_{c'}^T \phi(x_{ij})}, \tag{1}$$

where $p(\bar{y}_{ij} = c|x_{ij})$ indicates the estimated c'th class likelihood for the j'th patch of the i'th ROI, and ω_c corresponds to the c'th class' classifier parameters.

To contextually accumulate the patch-level classification probabilities into ROI-level predictions, we adopt the differentiable weakly supervised modeling scheme originally proposed for the training of object detectors.[13] For this purpose, we define the per-class patch-relevance estimation layers β that estimate the relevance weight α_{ij}^c for the patch x_{ij} and the class c:

$$\alpha_{ij}^{c} = \frac{\exp \beta_c^T \phi(x_{ij})}{\sum_{j'=1}^{|R_i|} \exp \beta_c^T \phi(x_{ij'})}. \tag{2}$$

It can be observed that the estimated relevance weight α_{ij}^{c} is nonnegative and sums to one over the patch indices as a result of the softmax operation.

The final ROI-level class likelihoods are defined as the relevance-weighted sum of per-patch class likelihoods

$$p(y_i^c|R_i) = \sum_{j=1}^{|R_i|} \alpha_{ij}^{c} p(\bar{y}_{ij} = c|x_{ij}), \tag{3}$$

where y_i^c is the binary random variable corresponding to the class c. It can be observed that the definition of $p(y_i^c|R_i)$ guarantees it to be in the range [0, 1] for all classes. The training is carried out by minimizing the negative sum of per-class log likelihoods:

$$-\sum_{c=1}^{C} [y_i = c] \log p(y_i^c|R_i) + [y_i \neq c](1 - \log p(y_i^c|R_i)), \tag{4}$$

where $[\,]$ is the Iverson bracket function. We highlight the fact that training requires only ROI-level annotations, and the whole model is trained in an end-to-end manner. The corresponding loss function is also known as binary cross-entropy (BCE) loss, which can be used for both multiclass and multilabel classification problems. The following subsection clarifies how we leverage this flexibility of BCE for improved training.

3.4 Learning with Augmented ROIs

Although the framework presented so far requires only a single class label annotation per ROI and has a direct mechanism to estimate the patch relevances, training an accurate model is still difficult due to the practical limits on the number of training ROIs and the prevalence of irrelevant patches in them. To enhance the training process and help the model reduce the mis-assignments in $p(\bar{y}_{ij} = c|x_{ij})$ estimates, we propose an ROI-level multilabel augmentation scheme specifically targeting the main challenges in our target problem.

Our main idea is to combine patches from pairs of ROIs to create augmented ROI sets, where the model is expected to predict the classes of both patch sources and, therefore, improve its patch-level class likelihood estimates. More specifically, we define an augmented ROI R_a as the union of the patches from two randomly selected ROIs R_i and $R_{i'}$ with different class assignments y_i and $y_{i'}$ such that $R_a = R_i \cup R_{i'}$, where the ROI index a is notationally used for the temporarily generated augmented ROIs. The class estimates for the patches in an augmented ROI are individually computed just as in Eq. (1). The patch weighting mechanism, however, is updated such that the softmax runs over all patches in the augmented ROI

$$\alpha_{aj}^{c} = \frac{\exp \beta_c^T \phi(x_{aj})}{\sum_{j'=1}^{|R_a|} \exp \beta_c^T \phi(x_{aj'})}. \tag{5}$$

As a result, the ROI-level class likelihoods are now defined over all patches in the augmented ROI

$$p(y_a^c|R_a) = \sum_{j=1}^{|R_a|} \alpha_{aj}^{c} p(\bar{y}_{aj} = c|x_{aj}). \tag{6}$$

The training of the model is now driven by the combination of the labels y_i and $y_{i'}$ in the multilabel setting:

$$-\sum_{c=1}^{C} [c \in Y_a] \log p(y_a^c|R_a) + [c \notin Y_a](1 - \log p(y_a^c|R_a)), \tag{7}$$

where the augmented ROI label is the set $Y_a = \{y_i, y_{i'}\}$.

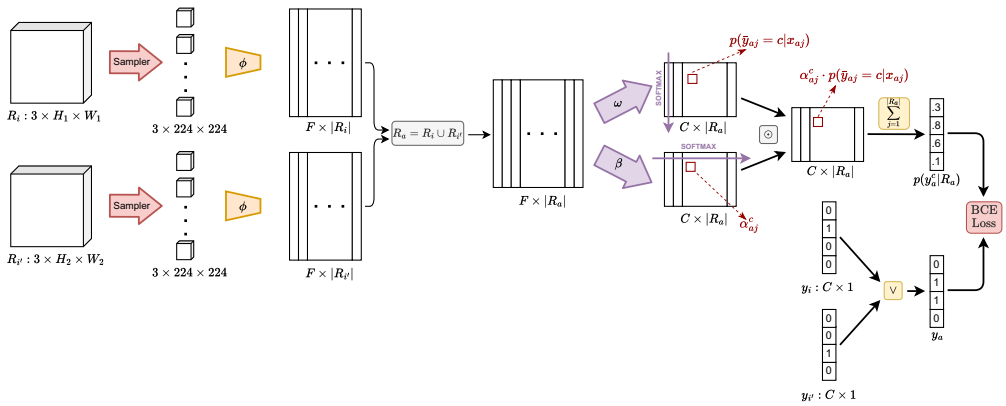

Fig. 2 Proposed patch-relevance estimation-based model and the improved ROI-augmentation-based training strategy. R_i and $R_{i'}$ denote the ROIs to be mixed, with their respective one-hot label vectors being y_i and $y_{i'}$. Patches sampled from each ROI are processed separately using the same patch encoder network ϕ, which results in matrices of size $F \times |R_i|$ and $F \times |R_{i'}|$, where F is the number of dimensions of the patch encoding vectors, and $|R_i|$ and $|R_{i'}|$ are the number of patches sampled from the first and the second ROI, respectively. These matrices are then concatenated into a larger matrix of size $F \times |R_a|$, where R_a refers to the union $R_i \cup R_{i'}$ of patches. The encodings are processed by the classification (ω) and relevance estimation (β) layers, which are then used to obtain per-patch class distributions $p(\bar{y}_{aj} = c|x_{aj})$ and the patch class relevance weights α^c_{aj} for all patches $x_{aj} \in R_a$ and all classes $c \in \{1 \dots C\}$. The weighted sums are obtained by summing the element-wise products of the rows, resulting in the vector of augmented ROI per-class likelihood estimates $p(y^c_a|R_a)$. This vector is used to calculate the total binary cross-entropy loss together with the two-hot reference label vector indicating $y_a = y_i \vee y_{i'}$.

The overall approach is summarized in Fig. 2, presenting a computational summary of the mathematical framework presented above. When the ROI is assigned a single label, the training process inherently tries to identify the patches that are most relevant to this particular label. However, when two ROIs are combined by taking the union of their patches, the resulting set contains multiple histologic structures that can each be diagnostically relevant with the multilabel combination of their individual labels. For example, in Fig. 1, the ROI labeled as atypia includes both atypical ductal hyperplasia and flat epithelial hyperplasia, whereas the ROI labeled as *in situ* contains both ductal carcinoma in situ and atypical ductal hyperplasia. The atypia patches in the latter ROI do not necessarily contribute to the learning process in the original single-label formulation but are more actively involved when the augmented ROI has both atypia and *in situ* in its multilabel assignment. Therefore, a multilabel target that lists the classes present in the union better reflects the true semantic content of the augmented ROI because the augmented sample is simultaneously positive for multiple categories. Thus, our ROI-augmented training strategy enforces the joint recognition of multiple classes within each augmented sample, enabling the learning of more discriminative patch encodings and classification layers.

By contrast, the original mixup formulation[15] treats target labels as proportions of different categories, which is appropriate when the mixed sample is an interpolation of two mutually exclusive classes. However, in our setting, the ROI union does not create a fractional mixture of patch encodings; it creates a set of patches where each patch still fully belongs to a distinct but unknown category. Thus, the multilabel targets allow the model to learn that both categories may be present in the ROI and to localize them via patch-level reasoning in Eq. (1). Furthermore, union-based augmentation does not require the ROIs to have the same number of patches via size alignment through sampling and grouping such as the mixup-based approaches.[4,16–20]

Although this strategy may introduce label heterogeneity compared with the original single-label setting, the formulation aligns with the WSL framework that naturally tolerates uncertainty in patch and label correspondences. Our ROI-level aggregation in Eq. (6) reduces patch-level sensitivity to noisy labels because our model is trained to identify the relevant patches that explain an ROI-level label in Eq. (5). Furthermore, training using the BCE loss in Eq. (7) permits patches within an augmented ROI that are irrelevant to some labels to have low activation scores for those labels without penalizing the overall ROI-level objective. Finally, the proposed augmentation can

also be considered a method to synthetically increase the number of ROIs that contain patches from underrepresented categories and can mitigate the effects of class imbalance.

4 Experiments

We compare the proposed approach against a number of contemporary and strong baselines, as well as the ablated versions of our approach. Below, we first describe the baseline methods and the evaluation metrics and then give implementation details for our method and the baselines. Finally, we present and discuss our experimental findings.

4.1 Baseline Methods and Metrics

For the performance comparison of the proposed models in this section, we use the following methods as baselines:

4.1.1 *Patch-classifier-majority*

A patch classifier network consisting of a convolutional neural network (CNN) encoder and a fully connected classification layer is trained using the ROI labels as the labels for each patch of the ROI. After the patch classifier is trained, the inference is made by classifying each patch of an ROI independently and then taking the patch predictions' majority vote as the ROI-level prediction.

4.1.2 *Patch-classifier-mean-prob*

The same steps as the patch-classifier-majority are followed, except for the inference phase, where the patch probability scores are averaged to obtain an ROI-level probability score vector.

4.1.3 *ROI-classifier-penultimate-mean*

This baseline uses the same architecture as patch-classifier-majority, where a CNN encodes each patch into a feature representation. After that, the feature vectors of the patches of an ROI are averaged to obtain an ROI-level representation, which is then processed by a fully connected classification layer to come up with an ROI-level prediction.

4.1.4 *mMIL*

This baseline is the generalization of the max-pooling aggregation rule to the multiclass setting, as described by Lu et al.,[10] where a class score vector is obtained for each patch using a CNN and a classification layer. Then, the vector having the highest class score among all patches and classes is selected as the ROI-level score vector to be used in the loss calculation.

4.1.5 *MIL-attention*

Similar to ROI-classifier-penultimate-mean, each patch is encoded into a feature vector via an encoder CNN. These vectors are aggregated using a weighted summation to obtain an ROI-level feature representation, which is fed into a classification layer to get an ROI-level prediction. Weights of this summation operation are obtained via an attention layer with the following formulation:

$$a_k = \frac{\exp\{\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{h}_k^\top)\}}{\sum_{j=1}^{K} \exp\{\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{h}_j^\top)\}}, \tag{8}$$

where a_k is the attention weights, $\mathbf{h}_k$ is the patch feature vectors, $\tanh(\cdot)$ is the hyperbolic tangent activation function, and $\mathbf{w} \in \mathbb{R}^{M\times 1}$ and $\mathbf{V} \in \mathbb{R}^{M\times F}$ are learnable parameters.[8] Here, F is the number of dimensions of the patch feature representation vectors, and M is a hyperparameter denoting the number of hidden units in the attention layer.

4.1.6 *MIL-attention-gated*

This baseline applies the gated variant of the attention mechanism used in MIL-attention, which has the following form:

$$a_k = \frac{\exp\{\mathbf{w}^\top(\tanh(\mathbf{V}\mathbf{h}_k^\top) \odot \mathrm{sigm}(\mathbf{U}\mathbf{h}_k^\top))\}}{\sum_{j=1}^{K}\exp\{\mathbf{w}^\top(\tanh(\mathbf{V}\mathbf{h}_j^\top) \odot \mathrm{sigm}(\mathbf{U}\mathbf{h}_j^\top))\}}, \tag{9}$$

where $\mathbf{U} \in \mathbb{R}^{M\times F}$ represents additional learnable weights and $\mathrm{sigm}(\cdot)$ is the sigmoid activation function.[8]

4.1.7 *MIL-per-class-attention-gated*

This method is the extended version of MIL-attention-gated, which uses per-class attention branches, as described by Lu et al.,[10] which has the following form:

$$a_{kc} = \frac{\exp\{\mathbf{w}_\mathbf{c}^\top(\tanh(\mathbf{V}\mathbf{h}_k^\top) \odot \mathrm{sigm}(\mathbf{U}\mathbf{h}_k^\top))\}}{\sum_{j=1}^{K}\exp\{\mathbf{w}_\mathbf{c}^\top(\tanh(\mathbf{V}\mathbf{h}_j^\top) \odot \mathrm{sigm}(\mathbf{U}\mathbf{h}_j^\top))\}}, \tag{10}$$

where c is the class index and a_{kc} is the attention weight of the k'th patch for class c. The difference between Eqs. (9) and (10) is that each attention branch applies a separate linear transformation in the form of $\mathbf{w_c}$ in Eq. (10), after a shared attention backbone. A different weighted combination of patch feature representations is obtained for each branch, which results in a distinct ROI-level feature vector per branch. A separate linear transformation is applied to each per-branch ROI feature vector to transform the vector into a single number that represents the logit value for that particular class.

4.1.8 *CLAM*

Clustering-constrained attention multiple instance learning (CLAM) is a combination of MIL-per-class-attention-gated trained using standard cross-entropy loss and per-class binary patch clustering layers trained using multiclass support vector machine (SVM) loss, which is designed to help the model to better separate the positive and negative patches for each class from each other.[10] These two losses are combined as

$$\mathcal{L}_{\text{total}} = \lambda_{\text{CE}}\mathcal{L}_{\text{CE}} + \lambda_{\text{SVM}}\mathcal{L}_{\text{SVM}}, \tag{11}$$

where $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{SVM}}$ are the aforementioned cross-entropy and multiclass SVM losses and λ_{CE} and λ_{SVM} are hyperparameters that correspond to the weights of these loss terms, respectively.

Although the authors studied the slide-level weakly supervised classification problem in the original implementation, we adapt their methodology to the ROI-level weakly supervised classification setting. While adapting the methodology to the problem we focus on, we need to make several changes to the original implementation. First, we use the penultimate layer features of an efficientnet-b0 pre-trained on ImageNet as patch-level features (the same architecture used in all experiments, explained in Sec. 4.2) and fine-tune its weights, whereas the original CLAM implementation utilizes intermediate layer features of an ImageNet pre-trained ResNet-50 without fine-tuning. We also remove the fully connected layer, which is applied to ResNet-50 features before the attention layers in the original implementation to make the method comparable to the other methods we experiment with. Finally, instead of choosing a fixed number of patches with the highest attention score and the same number of patches with the lowest attention score, we choose a ratio (b) of the highest-scoring patches and the same ratio of the lowest-scoring patches to calculate the multiclass SVM loss.

4.1.9 *CLAM-smooth*

This is a variant of CLAM that uses the smooth top-1 SVM loss instead of the multiclass SVM loss used in CLAM.[10]

4.1.10 *PatchRel variants*

In the experiments, we refer to the base version of the proposed framework (Sec. 3.3) as PatchRel, which stands for patch relevance estimation, and the proposed ROI augmentation technique as ROIAugment. In addition to the PatchRel and PatchRel + ROIAugment, we consider two additional variants as baselines for ROIAugment in the context of PatchRel. The first one is PatchRel + label-smoothing where label smoothing is applied to each ROI so that the label

encodings of the ROIs can slightly signal the presence of every class other than the reference label. The formulation for label smoothing can be described as

$$[y]_c = \begin{cases} 1-\epsilon & \text{if } c = c_{\text{ref}} \\ \epsilon/(C-1) & \text{otherwise} \end{cases}, \tag{12}$$

where $\epsilon = 0.2$ is the hyperparameter controlling the smoothing strength and C is the number of classes. The second version is PatchRel + random-label-add, which consists of converting the one-hot label encoding vector of an ROI into two-hot with a probability p_{flip} by flipping one of the zeros in the vector chosen randomly.

4.1.11 *Evaluation metrics*

When medical image classification datasets are characterized by class imbalance where some clinically important cases occur infrequently, commonly used metrics based on arithmetic averaging such as overall or macro-averaged accuracy may be dominated by the majority classes and may hide lower performance on rare or difficult categories such as atypia. Therefore, we adopt the geometric-mean (g-mean) as the main metric for both hyper-parameter selection (on the validation set) and experimental comparison (on the test set) by taking the geometric average of per-class recall values.[24] G-mean has the advantage of providing a fair and balanced evaluation over all diagnostic categories as it heavily penalizes the potential low per-class performances that might otherwise become invisible in arithmetic averaging. Thus, it helps picking models that do not overfit to the majority classes during hyper-parameter selection, and it promotes a consistent performance across all classes during experimental comparison. This makes it particularly suitable for medical imaging applications, where reliable recognition of underrepresented and clinically significant classes is of critical importance. We additionally provide per-class and macro-averaged precision, recall, and F-measure as performance scores. Macro-averaged F-measure is obtained by averaging per-class F-measure scores, where the F-measure score of a class is the harmonic mean of its precision and recall.

4.2 Implementation Details

The following paragraphs describe various important implementation details, which are used consistently for the proposed models and the baselines.

4.2.1 *End-to-end training, architecture*

We train all models in an end-to-end manner, unless specified otherwise. One of the challenges in the end-to-end training of the proposed methods is the large memory requirement for batch processing of the ROIs, which can cover large areas (see Table 1). We find that using 5× magnification with regularly sampled, 74 pixels-apart 224×224 pixel patches, combined with mixed-precision training, keeps the memory requirements manageable. We discard all ROIs with fewer than five patches. We use efficientnet-b0 architecture as the region encoder network ϕ, for its efficiency and low memory requirements,[25] yielding 1280-dimensional patch representations.

Another challenge is the variations in the number of patches across ROIs. When batches are fixed in terms of the number of sampled ROIs, the resulting number of patches can differ significantly between the iterations, depending on the sampled ROIs. This may cause training inefficiency due to under-utilizing compute resources, as well as out-of-memory problems. We address this problem using batches of fixed 512 patches, instead of a fixed number of ROIs. We also limit the number of patches loaded from a single ROI to 30 by randomly dropping patches of large ROIs. This trick not only reduces memory problems but also introduces a form of regularization, akin to the multiple instance augmentation technique proposed by Couture et al.[26]

4.2.2 *Image-domain augmentation, regularization*

We apply horizontal and vertical flips, axis-aligned rotations, coordinate jittering in the range $[-45, +45]$, and hue jittering in a randomized manner for image-domain data augmentation. We apply dropout to the penultimate layer, with a drop probability of 0.1. We also use stochastic depth[27] with a drop connect ratio of 0.25. We utilize weight decay with a rate of 10^{-4} on all parameters except the batch normalization and bias parameters, following Jia et al.[28]

4.2.3 *Class balancing*

To handle class imbalance, we oversample ROIs belonging to the atypia and *in situ* classes by rates of 3 and 2, respectively, chosen using the validation set. In the patch classification-based baselines, we oversample the minority ROIs in a way that leads to sampling approximately the same number of patches from each class. The ROI augmentation strategy also acts as a method to synthetically increase the number of samples for underrepresented classes in the corresponding experiments.

4.2.4 *Optimization and model selection*

We use stochastic gradient descent with a constant learning rate of 3×10^{-2} and a momentum of 0.75, starting from an ImageNet pretrained efficientnet-b0 model. We train all models for 40 epochs, with each epoch consisting of 500 batches. In all experiments, including the baselines, we first apply our oversampling methodology to obtain an oversampled version of the training dataset before proceeding with the training pipeline. We check the g-mean scores on the validation set after each epoch and use the model checkpoint with the highest validation score to report the test results. To make fair comparisons, all model hyperparameters are also chosen according to the validation g-mean score, using random search.

4.2.5 *Other hyper-parameters*

We introduce an additional hyperparameter for the PatchRel + ROIAugment model: ROI augmentation probability (p_{augment}), which denotes the probability to mix the patches of a given ROI with another ROI chosen randomly among the ROIs inside the same batch and belonging to other classes. No ROI augmentation is applied if there is no ROI inside the batch with a different label than the current ROI. When there is at least one ROI with a different label, we first choose a label randomly among the existing labels inside the batch, excluding the current label. Then, we choose the ROI with the number of patches closest to the number of patches of the current ROI among the ROIs with the chosen label. The motivation here is to construct sets of patches as balanced as possible to prevent certain ROIs from dominating the others. We use an ROI augmentation probability of $p_{\text{augment}} = 0.25$ in our experiments. For PatchRel + random-label-add, we use the same p_{augment} value as the random label addition probability (p_{flip}) for a fair comparison.

For the MIL-attention, MIL-attention-gated, MIL-per-class-attention-gated, CLAM, and CLAM-smooth baselines, we choose the number of hidden units in the attention layers (M) as 128. Among the hyperparameters specific to CLAM and CLAM-smooth, we set both the margin and the temperature parameters of the multiclass SVM loss in CLAM and the smooth top-1 SVM loss in CLAM-smooth to 1.0, similar to their original implementations. Cross-entropy and SVM loss weights λ_{CE} and λ_{SVM} are set to 0.7 and 0.3, respectively. The last hyperparameter b, the ratio of the highest-scoring patches and the lowest-scoring patches to be selected from each ROI for the SVM loss calculation, is selected as 0.1.

4.3 Results and Discussion

We present our experimental results on the Hacettepe dataset in Table 3. Our first observation is the superior performance of the patch-classifier-based approaches over ROI-classifier-penultimate-mean. An important difference between these approaches is the use of average pooling in the ROI classifier applied to the patch features before the classification layer. As the classification loss is computed on the ROI-level logits obtained after the pooling operation, the loss propagated through each patch becomes smaller for the ROIs with many patches. This results in more weight given to the patches of smaller ROIs during the training, which might not have a meaningful basis, potentially explaining the lower performance. It is also possible that the details of data (over)-sampling may (unintentionally) be more suitable for the patch-classifier models.

The second group of the table includes the MIL and attention-based approaches. Here, we first observe that mMIL is the worst-performing method among all baselines, which points out that max-pooling-based MIL formulation is inferior compared to the mean-pooling or attention-based formulations. We also observe that MIL-attention yields a low g-mean score, mostly due to its low recall on atypia, even though its gated variant MIL-attention-gated achieves comparable

Table 3 Experimental results for baselines and proposed methods trained end-to-end on the Hacettepe dataset. Per-class and macro-averaged precision, recall, and *F*-measure scores are listed in addition to the *g*-mean scores. For per-class scores, the abbreviations B, A, IS, and INV depict benign, atypia, *in situ*, and invasive classes, respectively.

		Precision					Recall					*F*-measure				
	G-mean	B	A	IS	INV	Avg.	B	A	IS	INV	Avg.	B	A	IS	INV	Avg.
Patch-classifier-majority	0.672	0.920	0.172	0.773	0.869	0.683	0.650	0.500	0.662	0.946	0.690	0.762	0.256	0.713	0.906	0.659
Patch-classifier-mean-prob	0.655	0.899	0.159	0.750	0.911	0.680	0.650	0.500	0.623	0.911	0.671	0.755	0.242	0.681	0.911	0.647
ROI-classifier-penultimate-mean	0.636	0.865	0.148	0.824	0.903	0.685	0.732	0.409	0.545	1.000	0.672	0.793	0.217	0.656	0.949	0.654
mMIL	0.545	0.843	0.099	0.688	0.867	0.624	0.610	0.364	0.429	0.929	0.583	0.708	0.155	0.528	0.897	0.572
MIL-attention	0.619	0.846	0.163	0.857	0.889	0.689	0.846	0.318	0.545	1.000	0.677	0.846	0.215	0.667	0.941	0.667
MIL-attention-gated	0.640	0.890	0.145	0.870	0.873	0.694	0.724	0.455	0.519	0.982	0.670	0.798	0.220	0.650	0.924	0.648
MIL-per-class-attention-gated	0.656	0.876	0.159	0.857	0.918	0.703	0.748	0.455	0.545	1.000	0.687	0.807	0.235	0.667	0.957	0.667
CLAM	0.660	0.876	0.179	0.870	0.889	0.703	0.805	0.455	0.519	1.000	0.695	0.839	0.256	0.650	0.941	0.672
CLAM-smooth	0.663	0.873	0.185	0.867	0.918	0.711	0.837	0.455	0.506	1.000	0.700	0.855	0.263	0.639	0.957	0.679
PatchRel	0.662	0.907	0.171	0.681	0.962	0.680	0.634	0.545	0.610	0.911	0.675	0.746	0.261	0.644	0.936	0.647
PatchRel + label-smoothing	0.609	0.829	0.133	0.830	0.981	0.693	0.789	0.364	0.506	0.946	0.651	0.808	0.195	0.629	0.964	0.649
PatchRel + random-label-add	0.654	0.860	0.175	0.812	0.949	0.699	0.797	0.455	0.506	1.000	0.689	0.827	0.253	0.624	0.974	0.670
PatchRel + ROIAugment	0.689	0.885	0.200	0.760	0.981	0.707	0.691	0.727	0.494	0.911	0.706	0.776	0.314	0.598	0.944	0.658

Table 4 Confusion matrix of patch-classifier-majority on the Hacettepe dataset.

	Predicted			
	Benign	Atypia	*In situ*	Invasive
Benign	80	35	6	2
Atypia	2	11	6	3
In situ	5	18	51	3
Invasive	0	0	3	53

results to other methods, showing the potential benefits of the gating mechanism in attention layers. A per-class attention formulation combined with the gating mechanism boosts the results further, as observed in MIL-per-class-attention-gated, CLAM, and CLAM-smooth results. Among these methods, we observe CLAM and CLAM-smooth to yield superior performances, with a small difference of 0.003 between their g-mean scores.

The proposed patch relevance estimation (PatchRel) approach's base version shows performance comparable to CLAM, CLAM-smooth, MIL-per-class-attention-gated, and patch-classifier-mean-prob, although it falls short of patch-classifier-majority, which surpasses all MIL and attention-based baselines as well. Although this result does not seem in favor of PatchRel at first, it might be hinting at the possibility of obtaining better results with PatchRel if more diverse batches can be used.

Next, we observe that PatchRel + ROIAugment surpasses all the baselines with a g-mean score of 0.689. Looking at the large performance difference between PatchRel + label-smoothing and PatchRel + ROIAugment, we can infer that the proposed multilabel ROI mixing strategy is effective beyond simple label smoothing in our weakly supervised learning with heterogeneous ROI contents. We also observe a clear performance drop in PatchRel with the addition of label smoothing, which points out that simple label smoothing can have a detrimental effect on the performance instead. Similarly, we observe a performance drop with PatchRel + random-label-add as well, suggesting the superiority of the proposed ROI augmentation methodology compared with introducing random label perturbations.

To compare the two best-performing models, patch-classifier-majority and PatchRel + ROIAugment, we look at the confusion matrices of the corresponding models obtained on the test set, given in Tables 4 and 5, respectively. Although both methods show similar performances for benign and invasive ROIs, their performances differ a lot for atypia and *in situ* classes. Specifically, we observe that patch-classifier-majority performs better on *in situ*, whereas PatchRel + ROIAugment obtains a higher score on atypia. One possible reason for this behavior is that the oversampling applied to the atypia class causes PatchRel + ROIAugment to mix more atypical ROIs with the other ones, resulting in a model more biased towards atypia, which results in decreased performance in other classes. Its lower performance on *in situ* can be explained from

Table 5 Confusion matrix of PatchRel + ROIAugment on the Hacettepe dataset.

	Predicted			
	Benign	Atypia	*In situ*	Invasive
Benign	85	29	8	1
Atypia	3	16	3	0
In situ	6	33	38	0
Invasive	2	2	1	51

a similar perspective as well. As *in situ* is the second most oversampled class after atypia, most of the mixture of ROIs are expected to be mixtures of atypia and *in situ*. This might introduce an additional bias toward atypia between these two classes, given that atypia is the most oversampled class.

4.3.1 *Statistical significance*

To assess the significance of the performance improvement obtained by PatchRel + ROIAugment, we perform a statistical significance analysis between each of the baselines and PatchRel + ROIAugment, using the predictions obtained on the test set. McNemar's test[29] is recommended for cases where there is a hold-out evaluation set with a limited amount of data.[30] However, due to McNemar's test being applicable only to 2×2 contingency tables, we use Bhapkar's test,[31] which can be used with tables of any size, hence allowing us to use 4×4 contingency tables where each column/row corresponds to one of the four classes we have. The highest p-value is obtained for mMIL with a value of 0.2465, and the second highest p-value is obtained for patch-classifier-mean-prob with a value of 0.0187, whereas all remaining p-values are observed to be smaller than 0.0025. These results indicate a significant difference between the performance of PatchRel + ROIAugment and the other baselines, except for mMIL where we observe a high p-value. This outcome, despite the magnitude of the accuracy difference between the two models, is due to the fact that their respective error distributions across the four classes are not statistically disparate enough to show marginal heterogeneity even though their disagreement ratio (the ratio of the number of disagreed samples to the total number of samples) in the contingency table is 34%.

4.3.2 *Incorporating graph neural networks*

As a final improvement, we investigate whether a contemporary graph neural network layer can improve the contextual assessment of ROIs in the context of the proposed PatchRel framework. For this purpose, we add graph attention network (GAT) layers[32] to PatchRel and PatchRel + ROIAugment. The GAT layer can be summarized as the application of a linear transformation to the vertex features, followed by a multihead attention mechanism consisting of a linear layer and a nonlinearity inside the first-order neighborhoods of each vertex to combine these transformed features into per-vertex aggregated features. Finally, the GAT layer applies a nonlinearity to the aggregated features to obtain the final outputs.

For GAT, we use four attention heads with each head consisting of 128 units, resulting in 512-dimensional vertex feature representations. The proximity threshold parameter is chosen as 200 pixels while constructing the adjacency matrices for all graph-based models. We investigate both 5× and 10× magnification settings. To make this combination possible, we switch from our end-to-end training scheme to a two-stage pipeline,[33] which first trains the patch encoder and then the rest of the model. The main reason for this change is that to construct a meaningful graph-based ROI representation to leverage the graph layers' capability of exploiting the spatial context, we need to have as many patches from an ROI inside the batch as possible. However, this is not possible in the end-to-end setting with batches of multiple ROIs, especially when larger ROIs are present inside the batch, due to memory limitations. We carry out the combination of two methods by replacing the fully connected layers in the classification (ω) and relevance estimation (β) branches with two GAT layers with a ReLU nonlinearity in between.

The results of PatchRel + ROIAugment and GAT combination on the Hacettepe dataset is presented in Table 6. Note that PatchRel and PatchRel + ROIAugment results at 5× magnification in Table 6 differ from the results in Table 3 due to the differences in the training pipeline discussed above. Given that the performances of these methods are lower than the previous results, we can argue that the end-to-end training scheme is more beneficial for PatchRel and PatchRel + ROIAugment. Our second observation is that the GAT combinations with PatchRel and PatchRel + ROIAugment perform significantly higher at 5× than 10×. Overall, we conclude that the best classification performance among all our experimental results is achieved by the combination of PatchRel + ROIAugment and GAT at 5× magnification, with a g-mean score of 0.728.

Table 6 Effect of magnification and GAT layer on PatchRel and PatchRel + ROIAugment in two-stage training on the Hacettepe dataset. For per-class scores, the abbreviations B, A, IS, and INV depict benign, atypia, *in situ*, and invasive classes, respectively.

			Precision					Recall					*F*-measure				
	Mag.	G-mean	B	A	IS	INV	Avg.	B	A	IS	INV	Avg.	B	A	IS	INV	Avg.
PatchRel	5×	0.649	0.902	0.185	0.649	1.000	0.684	0.748	0.455	0.649	0.804	0.664	0.818	0.263	0.649	0.891	0.655
PatchRel + ROIAugment		0.679	0.897	0.208	0.691	1.000	0.699	0.780	0.500	0.610	0.893	0.696	0.835	0.293	0.648	0.943	0.680
PatchRel + GAT		0.720	0.888	0.267	0.778	0.907	0.710	0.772	0.545	0.727	0.875	0.730	0.826	0.358	0.752	0.891	0.707
PatchRel + ROIAugment + GAT		0.728	0.905	0.239	0.783	0.936	0.716	0.699	0.727	0.701	0.786	0.728	0.789	0.360	0.740	0.854	0.686
PatchRel	10×	0.514	0.824	0.116	0.612	0.824	0.594	0.724	0.364	0.532	0.500	0.530	0.771	0.176	0.569	0.622	0.535
PatchRel + ROIAugment		0.545	0.861	0.095	0.697	0.976	0.657	0.756	0.273	0.597	0.714	0.585	0.805	0.141	0.643	0.825	0.604
PatchRel + GAT		0.682	0.875	0.195	0.725	0.923	0.679	0.569	0.682	0.649	0.857	0.689	0.690	0.303	0.685	0.889	0.642
PatchRel + ROIAugment + GAT		0.684	0.903	0.200	0.743	0.900	0.687	0.683	0.591	0.675	0.804	0.688	0.778	0.299	0.707	0.849	0.658

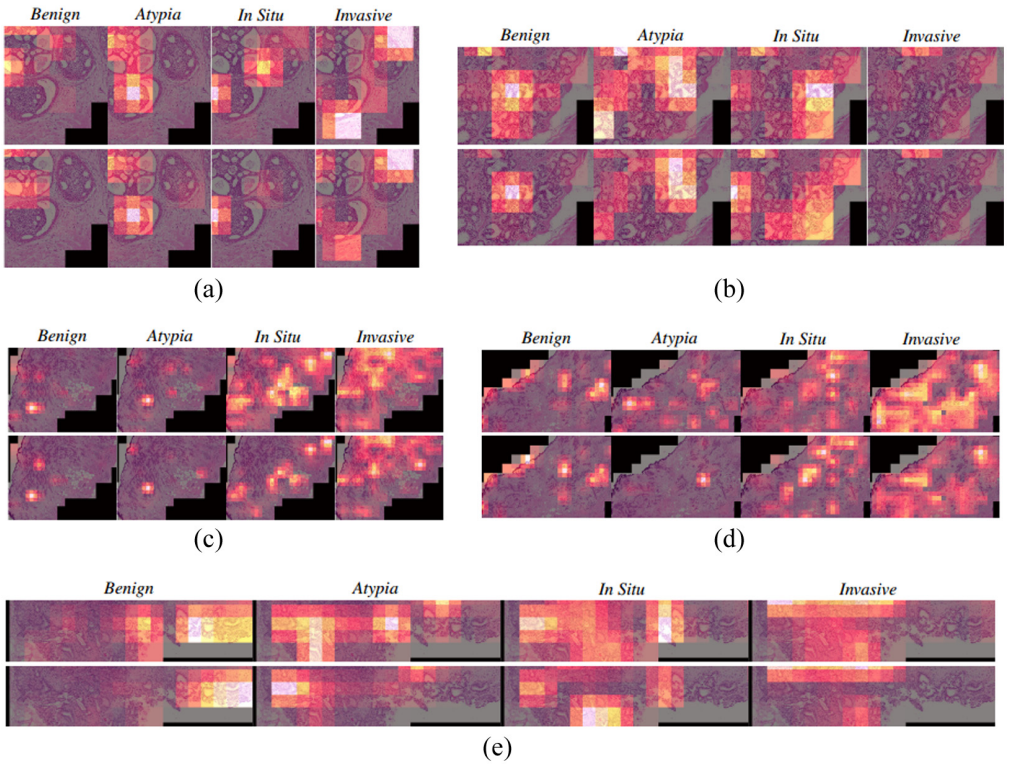

Fig. 3 Heatmaps of predicted relevance scores for the PatchRel and PatchRel + ROIAugment models for example ROIs. The hot color map is used for the score overlays. For each example ROI, the top row shows the scores for PatchRel and the bottom row shows the scores for PatchRel + ROIAugment. In each row, the images from left to right present the predictions for classes benign, atypia, *in situ*, and invasive. (a) PatchRel + ROIAugment correctly classifies as *in situ*, but PatchRel misclassifies as atypia. (b) PatchRel + ROIAugment correctly classifies as benign, but PatchRel misclassifies as atypia. (c) PatchRel + ROIAugment correctly classifies as invasive, but PatchRel misclassifies as *in situ*. (d) PatchRel + ROIAugment correctly classifies as invasive, but PatchRel misclassifies as *in situ*. (e) PatchRel + ROIAugment correctly classifies as benign, but PatchRel misclassifies as *in situ*.

4.3.3 *Qualitative evaluation*

We present heatmaps of predicted relevance scores in Eqs. (2) and (5) for the PatchRel and PatchRel + ROIAugment models, respectively, for example ROIs as evaluation in Fig. 3. These examples show that the PatchRel + ROIAugment model often attends the patches more selectively with better localization compared with the PatchRel model.

4.3.4 *Assessment of generalizability*

To assess the generalizability of the proposed approach on other datasets, we perform additional experiments using the publicly available BRACS[23] dataset that includes 547 hematoxylin and eosin-stained WSI scanned at 40× magnification. The dataset also provides annotations for 4539 ROIs extracted from 387 slides belonging to 151 patients. These ROIs are annotated according to three lesion types: benign, atypical, and malignant (combining *in situ* and invasive). We present experimental results using the test set that includes 242 benign, 162 atypical, and 166 malignant ROIs. The experiments use the ROIs at 10× magnification with the same hyperparameter settings for all models as the previous experiments. The models are trained using the training set with the checkpoint achieving the highest g-mean score on the validation set used to report the final results on the test set.

The experimental results are presented in Table 7. We observe that the relative ranking of all methods is similar to the previous experiments. The proposed PatchRel model achieves a g-mean score of 0.7396, which is higher than those of all baseline methods. Furthermore, the PatchRel + ROIAugment variant surpasses all methods with a g-mean score of 0.7442. Moreover, it has the

Table 7 Experimental results for baselines and proposed methods trained end-to-end on the BRACS dataset. Per-class and macro-averaged precision, recall, and *F*-measure scores are listed in addition to the *g*-mean scores. For per-class scores, the abbreviations B, A, and M depict benign, atypical, and malignant classes, respectively.

		Precision				Recall				*F*-measure			
	G-mean	B	A	M	Avg.	B	A	M	Avg.	B	A	M	Avg.
Patch-classifier-majority	0.7227	0.8273	0.6000	0.8034	0.7436	0.8527	0.5416	0.8173	0.7372	0.8398	0.5693	0.8103	0.7398
Patch-classifier-mean-prob	0.7347	0.8414	0.6250	0.7950	0.7538	0.8466	0.5555	0.8434	0.7485	0.8440	0.5882	0.8185	0.7502
ROI-classifier-penultimate-mean	0.7261	0.8313	0.6190	0.7933	0.7479	0.8466	0.5416	0.8347	0.7410	0.8389	0.5777	0.8135	0.7434
mMIL	0.6388	0.7034	0.4151	0.8356	0.6514	0.7830	0.5238	0.6354	0.6474	0.7411	0.4632	0.7219	0.6420
MIL-attention	0.6980	0.7714	0.4815	0.8118	0.6882	0.7642	0.6190	0.7188	0.7006	0.7678	0.5417	0.7624	0.6906
MIL-attention-gated	0.7060	0.8495	0.4667	0.7473	0.6878	0.7453	0.6667	0.7083	0.7068	0.7940	0.5490	0.7273	0.6901
MIL-per-class-attention-gated	0.7153	0.8257	0.5789	0.8144	0.7397	0.8491	0.5238	0.8229	0.7319	0.8372	0.5500	0.8187	0.7353
PatchRel	0.7396	0.8053	0.5952	0.8539	0.7515	0.8585	0.5952	0.7917	0.7485	0.8311	0.5952	0.8216	0.7493
PatchRel + ROIAugment	0.7442	0.8529	0.6410	0.7864	0.7601	0.8208	0.5952	0.8438	0.7532	0.8365	0.6173	0.8141	0.7560

highest macro-averaged precision, recall, and F-measure scores among all methods. We also observe that atypia receives the lowest scores among all classes for all methods. Similar to the experiments on the Hacettepe dataset, the proposed PatchRel + ROIAugment method achieves the highest performance on atypia compared with all other methods. Overall, the experiments show that the proposed two-branch framework for patch relevance estimation and the complementary ROI-level multilabel augmentation strategy effectively model the complex relationships between image-level labels and patch-level content in multiclass histopathological image analysis.

5 Conclusion

We tackled the problem of classifying ROIs of arbitrary sizes in a multiclass weakly supervised setting. We modeled the ROIs as sets of patches with the ROI-level reference diagnoses serving as weak labels. We hypothesized that patches carry different levels of information, with some patches being relevant to the diagnosis of the ROI, whereas others are uninformative or even misleading. To obtain an ROI-level decision from such patches, we utilized a two-branch architecture that aimed to predict the relevance of individual patches to the ROI-level label and combined these predictions as the relevance-weighted sum of per-patch class likelihoods. We also proposed an ROI-level augmentation-based training strategy that exploits multilabel ROIs to increase the likelihood of having more patches that are relevant to the augmented label sets to mitigate the effects of class imbalance and label uncertainty. Comparative experiments that used several MIL frameworks with both patch-level aggregation and attention-based feature-level aggregation methods as baselines confirmed the validity of our hypothesis and showed the effectiveness of the proposed methodology for multiclass classification of two challenging breast pathology datasets. In addition, incorporating contextual information by introducing graph attention layers into the patch relevance estimation model further improved the performance.

Disclosures

The authors declare that there are no financial interests, commercial affiliations, or other potential conflicts of interest that could have influenced the objectivity of this research or the writing of this paper.

Code and Data Availability

Implementations of the key technical parts, including patch-relevance estimation loss function and multilabel augmentation, will be made publicly available upon publication. The data utilized in this study were obtained from the Department of Pathology at Hacettepe University. Data are available from the authors upon request and with permission from Hacettepe University.

Acknowledgments

B. Aygunes and S. Aksoy were supported in part by the Scientific and Technological Research Council of Turkey under Grant No. 117E172. Parts of this paper appeared in the Master's thesis of B. Aygunes with the title "Weakly Supervised Approaches for Image Classification in Remote Sensing and Medical Image Analysis." The authors would also like to acknowledge Ms. Sude Önder from Bilkent University for the experimental results on the BRACS dataset.

References

1. J. G. Elmore et al., "Diagnostic concordance among pathologists interpreting breast biopsy specimens," *J. Am. Med. Assoc.* **313**(11), 1122–1132 (2015).
2. D. B. Nagarkar et al., "Region of interest identification and diagnostic agreement in breast pathology," *Mod. Pathol.* **29**(9), 1004–1011 (2016).
3. D. Tellez et al., "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," *Med. Image Anal.* **58**, 101544 (2019).
4. J. Yang et al., "ReMix: a general and efficient framework for multiple instance learning based whole slide image classification," *Lect. Notes Comput. Sci.* **13432**, 35–45 (2022).
5. G. Campanella et al., "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nat. Med.* **25**(8), 1301–1309 (2019).

6. C. Mercan et al., "Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images," *IEEE Trans. Med. Imaging* **37**, 316–325 (2018).
7. B. Gecer et al., "Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks," *Pattern Recognit.* **84**, 345–356 (2018).
8. M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Int. Conf. Mach. Learn.*, pp. 2127–2136 (2018).
9. C. Mercan et al., "Deep feature representations for variable-sized regions of interest in breast histopathology," *IEEE J. Biomed. Health. Inf.* **25**, 2041–2049 (2021).
10. M. Y. Lu et al., "Data efficient and weakly supervised computational pathology on whole slide images," *Nat. Biomed. Eng.* **5**(6), 555–570 (2021).
11. B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, pp. 14318–14328 (2021).
12. Z. Shao et al., "TransMIL: transformer based correlated multiple instance learning for whole slide image classification," in *Adv. Neural Inf. Process. Syst. 34*, pp. 2136–2147 (2021).
13. H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, pp. 2846–2854 (2016).
14. B. Aygunes, R. G. Cinbis, and S. Aksoy, "Weakly supervised instance attention for multisource fine-grained object recognition with an application to tree species classification," *ISPRS J. Photogramm. Remote Sens.* **176**, 262–274 (2021).
15. H. Zhang et al., "mixup: beyond empirical risk minimization," in *Int. Conf. Learn. Represent.* (2018).
16. H. Zhang et al., "DTFD-MIL: double-tier feature distillation multiple instance learning for histopathology whole slide image classification," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, pp. 18780–18790 (2022).
17. M. Gadermayr et al., "MixUp-MIL: novel data augmentation for multiple instance learning and a study on thyroid cancer diagnosis," *Lect. Notes Comput. Sci.* **14225**, 477–486 (2023).
18. Y.-C. Chen and C.-S. Lu, "RankMix: data augmentation for weakly supervised learning of classifying whole slide images with diverse sizes and imbalanced categories," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, pp. 23936–23945 (2023).
19. S. Keum et al., "Slot-mixup with subsampling: a simple regularization for WSI classification," arXiv.2311.17466 (2023).
20. P. Liu et al., "Pseudo-bag mixup augmentation for multiple instance learning-based whole slide image classification," *IEEE Trans. Med. Imaging* **43**(5), 1841–1852 (2024).
21. A. Basavanhally and A. Madabhushi, "EM-based segmentation-driven color standardization of digitized histopathology," *Proc. SPIE* **8676**, 86760G (2013).
22. A. Ruifrok and D. Johnston, "Quantification of histochemical staining by color deconvolution," *Anal. Quant. Cytol. Histol.* **23**(4), 291–299 (2001).
23. N. Brancati et al., "BRACS: a dataset for breast carcinoma subtyping in H&E histology images," *Database* **2022**, 1–10 (2022).
24. H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.* **21**, 1263–1284 (2009).
25. M. Tan and Q. Le, "Efficientnet: rethinking model scaling for convolutional neural networks," in *Int. Conf. Mach. Learn.*, pp. 6105–6114 (2019).
26. H. D. Couture et al., "Multiple instance learning for heterogeneous images: training a CNN for histopathology," *Lect. Notes Comput. Sci.* **11071**, 254–262 (2018).
27. G. Huang et al., "Deep networks with stochastic depth," in *Eur. Conf. Comput. Vision*, pp. 646–661 (2016).
28. X. Jia et al., "Highly scalable deep learning training system with mixed-precision: training ImageNet in four minutes," arXiv.1807.11205 (2018).
29. Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika* **12**(2), 153–157 (1947).
30. T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.* **10**(7), 1895–1923 (1998).
31. V. P. Bhapkar, "A note on the equivalence of two test criteria for hypotheses in categorical data," *J. Am. Stat. Assoc.* **61**(313), 228–235 (1966).
32. P. Velickovic et al., "Graph attention networks," in *Int. Conf. Learn. Represent.* (2018).
33. B. Aygunes et al., "Graph convolutional networks for region of interest classification in breast histopathology," *Proc. SPIE* **11320**, 113200K (2020).

Bulut Aygunes received his BS and MS degrees from Bilkent University. His research interests include applications of deep learning in digital pathology and remote sensing.

Ramazan Gokberk Cinbis received his PhD from the Université de Grenoble, France, in 2014. He is currently an associate professor at the Department of Computer Engineering at Middle East Technical University. His research interests include machine learning and computer vision, with a special interest in deep learning with incomplete weak supervision.

Selim Aksoy received his PhD from the University of Washington, Seattle, in 2001. He is currently a professor and the head of the Department of Computer Engineering at Bilkent University. His research interests include computer vision, pattern recognition, and machine learning with applications to medical imaging and remote sensing. He has been serving in the Program Committee of the Digital and Computational Pathology Conference as part of SPIE Medical Imaging.