Computers in Biology and Medicine 131 (2021) 104253

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

A multi-resolution model for histopathology image classification and localization with multiple instance learning

Jiayun Li [a,b], Wenyuan Li [a,b], Anthony Sisk [c], Huihui Ye [c], W. Dean Wallace [d], William Speier [a], Corey W. Arnold [a,b,c,*]

[a] Computational Diagnostics Lab, UCLA, 924 Westwood Blvd Suite 600, Los Angeles, CA, 90024, USA
[b] Department of Radiology, UCLA, 924 Westwood Blvd Suite 600, Los Angeles, CA, 90024, USA
[c] Department of Pathology & Laboratory Medicine, UCLA, 10833 Le Conte Ave, Los Angeles, CA, 90095, USA
[d] Department of Pathology, USC, 2011 Zonal Avenue, Los Angeles, CA, 90033, USA

ARTICLE INFO

Keywords:
Whole slide images
Multiple instance learning
Convolutional neural network
Image classification prostate cancer

ABSTRACT

Large numbers of histopathological images have been digitized into high resolution whole slide images, opening opportunities in developing computational image analysis tools to reduce pathologists' workload and potentially improve inter- and intra-observer agreement. Most previous work on whole slide image analysis has focused on classification or segmentation of small pre-selected regions-of-interest, which requires fine-grained annotation and is non-trivial to extend for large-scale whole slide analysis. In this paper, we proposed a multi-resolution multiple instance learning model that leverages saliency maps to detect suspicious regions for fine-grained grade prediction. Instead of relying on expensive region- or pixel-level annotations, our model can be trained end-to-end with only slide-level labels. The model is developed on a large-scale prostate biopsy dataset containing 20,229 slides from 830 patients. The model achieved 92.7% accuracy, 81.8% Cohen's Kappa for benign, low grade (*i.e.* Grade group 1) and high grade (*i.e.* Grade group $\geq$ 2) prediction, an area under the receiver operating characteristic curve (AUROC) of 98.2% and an average precision (AP) of 97.4% for differentiating malignant and benign slides. The model obtained an AUROC of 99.4% and an AP of 99.8% for cancer detection on an external dataset.

1. Introduction

Prostate cancer accounts for nearly 20% of new cancer diagnosed in men, and is the most prevalent and second deadliest cancer in men in the United States [49]. Active surveillance (AS) is an important management option for patients with clinically localized low-to intermediate-risk prostate cancer [57]. Prostate biopsy, which plays an essential role in treatment planning, is performed repeatedly during the course of the AS. Each biopsy can result in several tissue slides that are examined and, if cancer is present, assigned Gleason scores (GS) by pathologists based on the Gleason grading system. The GS is determined by two most predominant Gleason patterns that range from 1 (G1), closely resembling normal glands and carrying the lowest risk for dissemination, to 5 (G5), representing undifferentiated carcinoma and exhibiting the highest risk for dissemination. A recent study proposed to revise the Gleason grading system with 5 Gleason Grade groups (GGs) to reduce the over-treatment of low-grade prostate cancer [17]: GG 1 (GS $\leq$ 6), GG 2 (G3 + G4), GG 3 (G4 + G3), GG 4 (GS = 8) and GG 5 (GS $\geq$ 9). Patients with intermediate-to high-risk localized prostate cancer (GG $\geq$ 2) may be intervened with radiotherapy and radical prostatectomy, with or without hormonal therapy.

Currently, the diagnosis of prostate cancer relies on pathologists to examine multiple levels of biopsy cores at the scanning magnification, and identify suspicious regions for high power examination and immunohistochemistry if necessary. This process can be tedious and time-consuming. More importantly, some patterns, *e.g.* ill-defined G4 versus tangentially sectioned G3, are prone to inter- and intra-observer variability. Therefore, the current clinical practice can be improved by computer aided diagnosis tools (CAD) that can function as primary screening, to localize suspicious regions, and be utilized as a second reader for Gleason grading. Deep learning-based CAD models have been developed and demonstrated promising performance in many medical imaging fields [8,9,13,48,54]. However, the enormous size of whole slide images (WSI), the variability in tissue appearances, and the

* Corresponding author. Computational Diagnostics Lab, UCLA, 924 Westwood Blvd Suite 600, Los Angeles, CA, 90024, USA.
E-mail addresses: jiayunli@g.ucla.edu (J. Li), cwarnold@ucla.edu (C.W. Arnold).

https://doi.org/10.1016/j.compbiomed.2021.104253
Received 18 November 2020; Received in revised form 31 January 2021; Accepted 3 February 2021
Available online 10 February 2021

artifacts incurred during staining and scanning impose many unique challenges in developing such CAD tools.

1.1. Related work

Classification of small homogeneous regions of interest (ROIs) pre-selected by pathologists has been the main focus of most early work in WSI image analysis [16,19,40]. Though these methods have achieved good results, they cannot be easily extended to handle regions with heterogeneous tissue types because they require a set of manually selected tiles with the same cancer grade, which is non-trivial to obtain. Some work has addressed this challenge by developing segmentation models that can provide pixel-wise predictions for tiles with various tissue contents [20,24,28–30]. However, these models still analyzed tiles instead of the entire slide.

With an increasing number of scanned slides and computing power, research in WSI has been shifting to slide-level analysis [38,39,41]. For example, in a recent work by Nagpal *et-al.* [39], they developed a two-stage model on a dataset containing 752 biopsy slides and achieved 71.7% Cohen's Kappa for predicting benign, GG1, GG2, GG3, and GG4-5. However, the model relied on a large amount of expensive fine-grained annotations.

While these papers demonstrated promising performance in slide-level predictions [32,38,39], they required a large number of expensive pixel or patch-level manual annotations for training. Bulten *et-al.* utilized a semi-automated labeling technique for prostate biopsy slide classification [4,5]. Specifically, the authors used a pre-trained tissue segmentation network to identify tissue areas, within which cancerous regions were localized by a pre-trained tumor detection network. Non-epithelial areas were excluded from identified cancerous regions with an epithelium detection model. Detected epithelial areas from slides with a single Gleason pattern inherited slide-level labels and formed their initial training set for a U-Net model. Slide-level predictions were determined by percentage of Gleason patterns obtained from the segmentation network. However, this framework was built upon three pre-trained preprocessing modules, each of which still required pixel-wise annotations.

Multiple instance leaning (MIL) [2,15] has been utilized to address weakly-supervised learning challenges in tumor detection [35,36,44], segmentation [25,61], and classification [21,23,37,53,59,62]. Most MIL models fall roughly into two general categories [1,7,12]: instance-base and bag-based methods. Bag-based methods usually demonstrate better performance for tasks where global (*i.e.,* bag-level) predictions are more important. Nevertheless, they suffer from a lack of interpretability, since instance predictions are often unavailable [23]. Ilse *et-al.* developed an attention-based MIL model that can visualize the relative contribution of instances for final prediction through a trainable attention module without sacrificing bag-level prediction performances [23]. The model was utilized to identify epithelial and malignant patches within small tiles extracted from WSI for colon cancer and breast cancer datasets, respectively. However, they did not address the challenge of classifying much larger and more heterogeneous WSIs. Moreover, they only utilized attention maps for visualization.

Few recent works have utilized MIL for whole slide classification [6, 33,59,60]. Campanella *et-al.* employed an instance-based approach to discriminate between malignant and benign prostate WSIs [6]. They considered the top k tiles with the highest probabilities from positive slides after applying the CNN model as pseudo positive training samples, which were updated in each training epoch. In the second stage, they investigated aggregation functions to produce a final slide-level prediction. The model achieved promising performance on three different types of large-scale clinical datasets. However, the more difficult problem of Gleason grading was not investigated in the paper.

In this paper, we proposed a multi-resolution MIL-based (MRMIL) model for prostate biopsy WSI classification and weakly-supervised tumor region detection. Different from most existing studies, which rely on highly curated datasets with fine-grained manual annotations at pixel- or region-level, our model can be trained with only slide-level labels obtained from pathology reports. Similar to how WSIs are typically reviewed by pathologists, the proposed model scans through the entire slide to localize suspicious regions at a lower resolution (*i.e.*, at 5x), and then zooms in on the suspicious regions to make grade predictions (*i.e.*, at 10x). The main contribution of this paper can be summarized as the following:

1) We developed a novel MRMIL model, which can be trained with only slide-level labels, for prostate cancer WSI classification and detection. 2) We trained and validated our model on a large dataset, containing 13,145 slides from 661 patients, which were retrieved from clinical cases without manual curation. To have a better understanding of our model's performance, we also visualized the data representations learned by the model. 3) We tested on an independent test set consisting of 7114 biopsy slides from 169 patients and an external dataset. The model achieved 81.8% Cohen's Kappa (κ) and 92.7% accuracy (Acc) for classifying benign, low grade (*i.e.*, GG = 1), and high grade (*i.e.*, GG $\geq$ 2) slides. Additionally, We extended our best model for Gleason group prediction, and it obtained 71.1% κ and 86.8% quadratic κ.

2. Method

2.1. Problem definition

Due to the enormous size of WSIs, slides are usually divided into smaller tiles for analysis. However, different from works that utilized fine-grained manual annotations, our model is developed on then dataset with only slide-level labels (*i.e.,* We don't have labels for each tile, instead, we only have a slide-level label for a set of tiles.). Therefore, we formulate the WSI classification problem in the MIL framework. Specifically, a slide is considered as one bag. k tiles of size $N \times N$ extracted from the bag are denoted as instances within the bag, each of which may have different instance-level labels y_i, $i \in [1, k]$. During training, only the label for a set of instances (*i.e.*, bag-level) Y is available. Based on the MIL assumption, a positive bag should contain at least one positive instance, while a negative bag contains all negative instances [1,2,12,15] in a binary classification scenario. We build our system upon a bag-level MIL model with a parameterized attention module that aggregates instance features and forms the bag-level representation, instead of using a pre-defined function, such as maximum or mean pooling [23]. Fig. 1 shows the overview of our model.

2.2. Attention-based MIL with instance dropout

In the attention-based MIL model, a CNN is utilized to transform each instance into a d dimensional feature vector $\mathbf{v}_i \in \mathbb{R}^d$. A permutation invariant function $f(\cdot)$ can be applied to aggregate and project k instance-level feature vectors into a joint bag-level representation. We use a multilayer perceptron-based attention module as $f(\cdot)$ [23], which produces a combined bag-level feature vector $\mathbf{v}'$ and a set of attention values representing the relative contribution of each instance as defined in Eq (1).

$$\begin{aligned} \mathbf{v}' &= f(\mathbf{V}) = \sum_{i=1}^{k} \alpha_i \mathbf{v}_i \\ \alpha &= \mathrm{Softmax}\left[\mathbf{u}^T \tanh\left(\mathbf{W}\mathbf{V}^T\right)\right] \end{aligned} \tag{1}$$

where $\mathbf{V} \in \mathbb{R}^{k \times d}$ contains the feature vectors for k tiles, $\boldsymbol{u} \in \mathbb{R}^{h \times 1}$ and $\boldsymbol{W} \in \mathbb{R}^{h \times d}$ are parameters in the attention module, and h denotes the dimension of the hidden layer. The slide-level prediction can be obtained by applying a fully connected layer to the bag-level representations $\mathbf{v}'$. Both the CNN feature extractor and the attention-based aggregation function are differentiable and can be trained end-to-end using gradient descent. The attention module not only provides a

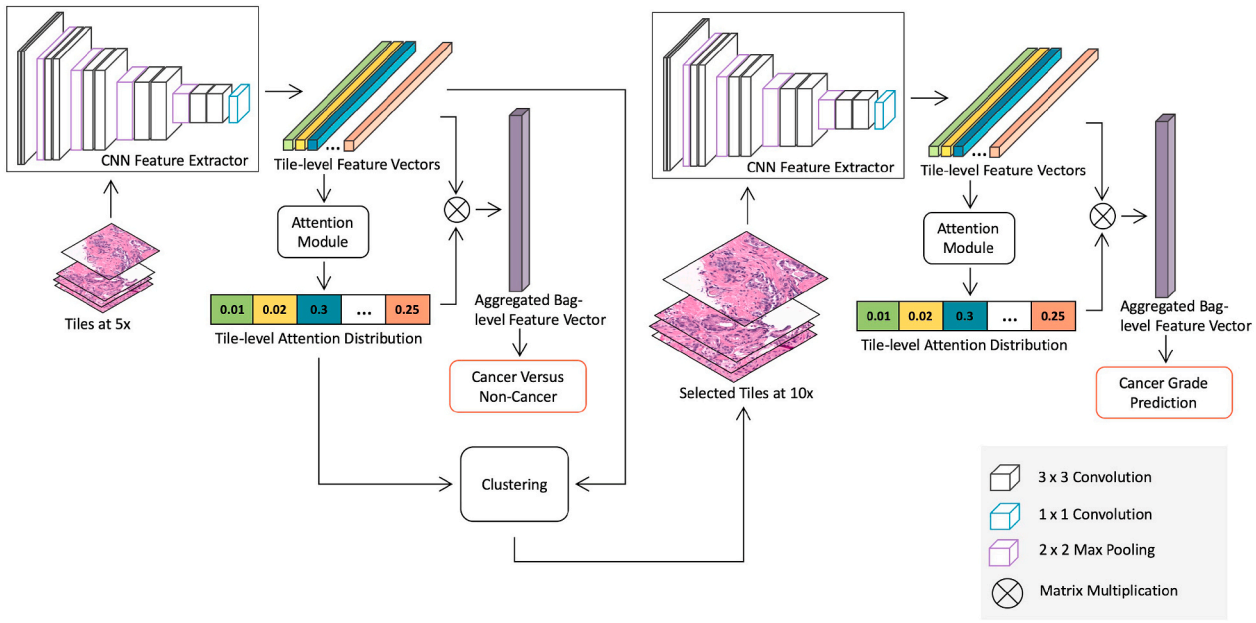

Fig. 1. Overview of the proposed whole slide image detection and classification model. The model consists of two stages: a cancer detection stage at a low magnification and a cancer classification stage at a higher magnification for suspicious regions. Both stages contains a CNN feature extractor, which is trained in the MIL framework with slide-level labels. Specifically, the detection stage model is trained with all tiles extracted from slides at 5x to differentiate between benign and malignant slides. The attention module in the detection stage model produces a saliency map, which represents relative importance of each tile for predicting slide-level labels. Then we use the K-means clustering method to group tiles into clusters based on tile-level features. The number of tiles selected from each cluster is determined by the mean of cluster attention values. Discriminative tiles identified by the detection stage model are then extracted at 10x and fed into the classification stage model for cancer grade classification.

more flexible way to incorporate information from instances, but also enables us to localize informative tiles. However, this framework encounters similar problems as other saliency detection models [22,51,63]. In particular, as pointed out in Ref. [23], instead of detecting the all informative regions, the learned attention map can be highly sparse with very few positive instances having large values. This issue may be caused by the underlying MIL assumption that only one positive instance needs to be detected for a positive bag. This can affect the performance of our classification stage model, which relies on informative tiles selected by the learned attention map. To encourage the model to select more relevant tiles, we used an instance dropout method similar to Refs. [51,52]. Specifically, instances are randomly dropped during the training, while all instances are used during model evaluation. To ensure the distribution of inputs for each node in the network remains the same during training and testing, pixel values of dropped instances are set to be the mean RGB value of the dataset [51,52]. This form of instance dropout can be considered a regularization method that prevents the network from relying on only a few instances for bag-level classification.

2.3. Attention-based tile selection

An intuitive approach to localize suspicious regions with learned attention maps is to use the top q percent of tiles with the highest attention weights. However, the percentage of cancerous regions can vary across different cases. Therefore, using a fixed q may cause over selection for slides with small suspicious regions and under selection for those with large suspicious regions. Moreover, this method relies on an attention map, which in this context is learned without explicit supervision at the pixel- or region-level. To address these challenges, we incorporate information embedded in instance-level representations by selecting informative tiles from clusters. Specifically, instance representations obtained from the MIL model are projected to a compact latent embedding space using principle component analysis (PCA). We then perform K-means clustering to group instances with similar semantic features based on their PCA transformed instance-level representations. The relevance of each cluster $\overline{\alpha}_s$ can be determined by the average attention weights of tiles within it as defined by $\overline{\alpha}_s = \frac{1}{m}\sum_{j=1}^{m}\alpha_j$. The intuition is that clusters that contain more relevant information for slide classification should have higher average attention weights. For example, in a cancer-positive slide, clusters consisting of cancerous glands should have higher attention weights compared to those with benign glands and stromal regions. Finally, we can determine the number of tiles to extract from each cluster based on the $\overline{\alpha}_s$ and the total number of tiles.

2.4. Multi-resolution WSI classification

Different from most medical imaging modalities, WSIs typically contain billions of pixels, which make them practically impossible to feed into GPU memory directly at full resolution. Though the size of WSIs is enormous, most regions typically do not contain relevant information for slide classification, such as stroma and benign glands. Pathologists tend to analyze the entire slide at a relatively low resolution, usually at 5x, to find suspicious regions and then switch to higher magnification in these areas to render a final diagnosis. Our proposed MRMIL model mimics this process, containing two stages as shown in Fig. 1. The detection stage model, which consists of an attention-based MIL with instance dropout, is trained with all tiles extracted at a lower magnification (*i.e.*, at 5x) to differentiate benign and malignant slides and generate attention maps. The attention-based clustering method is applied to select relevant tiles for the classification stage model. Selected tiles are extracted at the same location, but at a higher magnification (*i.e.* at 10x) and fed into the MIL network for cancer grade prediction.

3. Experiment

3.1. Dataset and data preprocessing

3.1.1. Dataset

Our dataset contains 20,229 slides from prostate needle biopsies from 830 patients pre- or post-diagnosis (IRB16-001361). Slides' labels extracted from their corresponding pathology reports. There are no additional fine-grained annotations at the pixel- or region-level for this dataset. Additionally, we did not rely on any pre-trained tissue, epithelium, or cancer segmentation networks, and did not perform extensive manual curation to exclude slides with artifacts such as air bubbles, pen markers, dust, *etc*. We randomly divided the dataset into 70% for training, 10% for validation, and 20% for testing, stratifying by patient-level GG determined by the highest GG in each patient's set of biopsy cores. This process produced a test set with 7114 slides from 169 patients and a validation set containing 3477 slides from 86 patients. From the rest of the dataset, we balanced sampled benign (BN), low grade (LG), and high grade (HG) slides. Table 1 shows more details on the breakdown of slides.

3.1.2. External dataset

We evaluated our models on a public prostate dataset, SICAPV1, collected by the Hospital Clínico Universitario de Valencia, which contains 512 × 512 tiles at 10x extracted from 79 slides of prostate needle biopsies with 50% overlapping [18]. 19 of these slides are benign, and the rest are malignant.

3.1.3. Data preprocessing

The majority of regions on WSIs are background. Thus, we converted slides downsampled at their lowest available magnification compressed in the .svs file into HSV color space and thresholded on the hue channel to produce tissue masks. Morphological operations such as dilation and erosion were used to fill in small gaps, remove isolated points, and further refine tissue masks. We then extracted tiles of size 256× 256 at 10x from the grid with 12.5% overlap. Tiles that contained less than 60% tissue were discarded from analysis. The number of tiles per slide ranges from 1 to 1,273, with an average of 275. To account for stain variability, we used a color transfer method [45] to normalize tiles extracted from the slide. The scanning objective was set at 20x (0.5 μm per pixel). We downsampled tiles to 5x for the detection stage model development. We divided the 512 × 512 tiles at 10x (as provided by the external dataset [18]) into 4 non-overlapping 256 × 256 sub-tiles, in order to match the input size of our models. The same stain normalization [45] was applied. Tiles with less than 60% tissue were also removed.

3.2. Implementation details

We used VGG11 with batch normalization (VGG11bn) [50] as the backbone for the feature extractor in the MRMIL model for both detection stage and classification stage. Sizes of input tiles for cancer detection and classification stages are 128 × 128 and 256 × 256, respectively. Thus, sizes of feature maps from the last convolutional of VGG11bn are 512 × 4 × 4 and 512 × 8 × 8, respectively. A 1× 1 convolutional layer was added after the last convolutional layer of VGG11bn to reduce dimensionality and generate instance-level feature maps for k tiles. Feature maps were flattened and fed into a fully connected layer with 256 nodes, followed by ReLU and dropout layers. This produced a $k \times 256$ instance embedding matrix, which was forwarded into the attention module. The attention part, which generated a $k \times n$ attention matrix for n prediction classes, consisted of two fully connected layers with dropout, tanh non-linear activations, and a softmax layer. Instance embeddings $k \times 256$ were multiplied with attention weights $k \times n$, resulting in a $n \times 256$ bag-level representation, which was flattened and input into the final classifier consisting of a fully connected layer. The probability of instance dropout was set to 0.5 for both model stages. Detailed model architectures were shown in Table 3 and 4 in the Appendix A.

The CNN feature extractor was initialized with weights learned from the ImageNet dataset [14]. After training the attention module and the classifier with the feature extractor frozen for three epochs, we trained the last three VGG blocks together with the attention module and classifier for 97 epochs. The initial learning rates for the feature extractor were set at 1×10^{-5} and 5×10^{-5} for the attention module and the classifier, respectively. The learning rate was decreased by a factor of 10 if the validation loss did not improve for the last 10 epochs. We used the Adam optimizer [26] and a batch size of one. Detection stage and classification stage models were trained separately using the same training hyperparameter (*e.g.,* learning rate, batch size and *etc*). Random flipping and rotation were utilized for data augmentation.

For clustering-based region selection, we projected $k \times 256$ instance embedding matrix to $k \times 32$ with PCA, and utilized K-Means clustering to group tiles. The number of clusters was set to be 3 to encourage tiles to be grouped into LG, HG and BN clusters. As shown in Appendix C, tile selection with k-Means is robust to different random initialization.

Hyperparameters were tuned on the validation set. We further extended our MRMIL model for GG prediction. The cross entropy loss weighted by reversed class frequency was utilized to address the class imbalance problem. Hyperparameters were selected using the validation set. Models were implemented in PyTorch 0.4.1 [42], and trained on an NVIDIA DGX-1.

3.3. Evaluation metrics

As our test dataset contained over 75% benign slides, accuracy (Acc) alone is biased metric for model evaluation. In addition, we used the AUROC and AP computed from ROC and precision and recall (PR) curves, respectively. For cancer grade classification, we measured the Cohen's Kappa (κ), $\kappa = \frac{p_o - p_e}{1 - p_e}$. p_o is the agreement between observers and p_e is the probability of agreement by chance. All metrics were computed using the scikit-learn 0.20.0 package [43].

3.4. Model visualization

In addition to quantitative evaluation metrics, interpretability is important in developing explainable machine learning tools, especially for medical applications. In order to have a better understanding of our model predictions, we performed t-Distributed Stochastic Neighbor Embedding (t-SNE) [34] of learned bag-level representations for both stage models. Specifically, for each slide we utilized the flattened $n \times 256$ feature vector before being forwarded to the final classification layer. The learning rate of t-SNE was set at 1.5×10^2, and the perplexity was set at 30.

The saliency map produced by the attention module in the MRMIL model only demonstrated the relative importance of each tile. To further localize discriminative regions within tiles, we utilized Gradient-weighted Class Activation Mapping (Grad-CAM) [47]. Concretely, given a trained MRMIL model and a target class c, we retrieved the top k tiles with the highest attention weights, which were fed to the model to

Table 1
Number of slides for each Grade group.

	Train	Validation	Test	Total
No. BN slides	3225	2579	5355	11,159
No. GG 1 slides	3224	412	807	4443
No. GG 2 slides	1966	307	587	2860
No. GG 3 slides	648	95	148	891
No. GG 4 slides	306	17	129	452
No. GG 5 slides	269	67	88	424
No. Patients	575	86	169	830

compute gradients and activations.

3.5. Model comparison

Handcrafted features. We converted input tiles at 10x into HSV color space and thresholded on the H channel to get tissue masks. Then we utilized the PyRadiomics package [58] to extract 90 features for each tile, including 16 first-order statistics, 23 Gy level co-occurrence matrix-based, 16 Gy level run length matrix-based, 16 Gy level size zone matrix-based, 5 neighbouring gray tone difference matrix-based, and 14 Gy level dependence matrix-based features. The maximum pooling was applied to aggregate tile-level features, which were fed into the final slide-level classifier. We experimented with Xgboost [11] and random forest (RF) [31] classifiers. Grid search with 3-fold cross validation was used to select hyperparameters for classifiers.

MIL model by Campanella *et-al.* [6]. We compared our model with the related recent work [6], which also trained slide classification models with only slide-level labels in the MIL framework. Different from our model, they utilized an instance-level MIL approach to train CNN-based feature extractors and aggregated tile-level information with an RNN. Specifically, a VGG11bn model pretrained on the ImageNet dataset was applied on tiles at 10x for cancer versus non-cancer classification. The top k tiles with highest probability within each slide were assumed to have the same label as the slide-level label. The k was set at 1. These tiles were then utilized to further optimize the model. This process was iterated until convergence. Then the RNN model was utilized to aggregate features from the top s tiles for final classification. The s was set at 10. We used the implementation provided by Ref. [6] and hyperparameters reported in the paper to re-train the model on our dataset.

MIL model by Tomczak *et-al.* [55] Tomczak *et-al.* investigated several instance-level MIL models, which utilized different permutation-invariant operators to combine patch-level predictions for histopathological image classification, with a Noisy-Or operator achieving the most promising results [55]. The model was developed to distinguish benign and malignant patches extracted from biopsy slides of breast cancer patients. In this experiment, we extended it for WSI classification. Since the Noisy-Or was mostly designed for binary classification, we mainly utilized it to classify benign and malignant slides.

Different aggregation methods. Instead of using the attention module to aggregate tile-level features to slide-level representations, we experimented with two aggregation methods to combine tile-level features for slide-level representations: maximum pooling and mean pooling aggregation. These models were trained using all tiles at 10x.

Single stage. To evaluate the effectiveness of the multi-resolution model in cancer grade prediction, we compared our model with a model trained with all extracted tiles at 5x only, referred as Single stage.

Blue ratio selection. Blue ratio (Br) image conversion, as defined in Eq (2), can accentuate the blue channel of a RGB image and thus highlight proliferate nuclei regions [10].

$$\mathrm{Br} = \frac{100 \times B}{1 + R + G} \times \frac{256}{1 + R + G + B} \tag{2}$$

where R, G, B are the red, green and blue channels in the original RGB image. Br conversion is one of the most commonly used approaches in previous studies to detect nuclei [10,46] and select informative regions from large-scale WSIs [3,27,56]. To evaluate the attention-based ROI detection, we replaced the first stage cancer detection model with the Br conversion to select the top $q = 25\%$ tiles with highest average Br values, referred to as *br selection* [3,27,56]. We performed the Br conversion using slides at 5x (the same magnification as our cancer detection stage model). Examples of Br selected tiles are shown in Fig. 7 in Appendix B.

Without instance dropout [23]. In this experiment, denoted as *w/o instance dropout*, we utilized the vanilla attention MIL model as developed in Ref. [23] and investigated whether instance dropout could improve the integrity of learned attention map and lead to better performance.

Attention-only selection. Instead of selecting informative clusters, we only utilized the attention map by choosing the top $q = 25\%$ tiles with the highest attention values as the input for the second stage model in the *att selection* experiment.

For fair comparison, we utilized VGG11bn as the feature extractor for all comparison experiments. Results of ablation experiments using other CNN architectures as feature extractors are shown in Table 5 in Appendix D.

4. Results

Fig. 2 shows both ROC and PR curves for the detection stage cancer models trained at 5x. The detection stage model in the MRMIL obtained an AUROC of 97.7% and an AP of 96.7% on our internal test set. On the external dataset, it achieved an AUROC of 99.4% and an AP of 99.8%. The model trained without using the instance dropout method yielded a slightly lower AUROC and AP on both internal and external datasets.

Since our dataset does not have fine-grained annotations, we visualized generated attention maps and compared them with pen markers annotated by pathologists during diagnosis. We masked out markers as mentioned in Section 3.1, thus they were not utilized for model training. Fig. 4 presents the comparison between attention maps learned from models with and without using instance dropout during training.

To further localize suspicious regions within a tile and better interpret model predictions, we applied Grad-CAM on the first detection stage MIL model as shown in Fig. 5. We generated Grad-CAM maps for not only true positives (TP), but also false positives (FP) to understand which parts of the tile led to false predictions. We selected two tiles with highest attention weights from each slide for visualization.

The MRMIL model projects input tiles to embedding vectors, which are aggregated and form slide-level representations. The t-SNE method enables high dimensional slide-level features to be visualized at a two dimensional space as demonstrated in Fig. 6. Fig. 6 (A) is the t-SNE plot for the detection stage model and (B) presents bag-level features produced by the classification stage model with selected high resolution tiles as inputs.

Table 2 shows model performances on BN, LG, HG classification. The proposed MRMIL outperformed all baseline models and achieved the highest Acc of 92.7% and κ of 81.8% as shown in the last row. Models with handcrafted features only obtained about 57% κ as demonstrated in row 3 and 4 in Table 2. As shown in row 5, the model by Campanella *et-al.* [6] got 4% lower κ compared with our MRMIL model. Models with simple mean and maximum pooling aggregation methods also achieved lower performance than the MRMIL model as reported in row 7 and 8. Row 9 to 12 demonstrated results on ablation study of the MRMIL model. The single stage attention MIL model trained at 5x achieved 76.3% κ. The br selection that relied on the Br image for tile selection only obtained an Acc of 90.8% and a κ of 76.0%. The w/o instance dropout model, got roughly 4% lower κ and 2% lower Acc compared with the MRMIL model. In addition, we combined LG and HG predictions from the classification model and computed the AUROC and AP for detecting cancerous slides. For instance, by zooming in on suspicious regions identified by the detection stage model, the MRMIL achieved an AUROC of 98.2% and an AP of 97.4%, both of which are higher than the detection stage only model. We present the confusion matrix for the MRMIL model on GG prediction in Fig. 3. The MRMIL model obtained an accuracy of 87.9%, a quadratic κ of 86.8%, and a κ of 71.1% for GG prediction.

5. Discussion

In this paper, we developed an MRMIL model, which addressed three challenges in predicting slide-level Gleason grades: 1) how to select

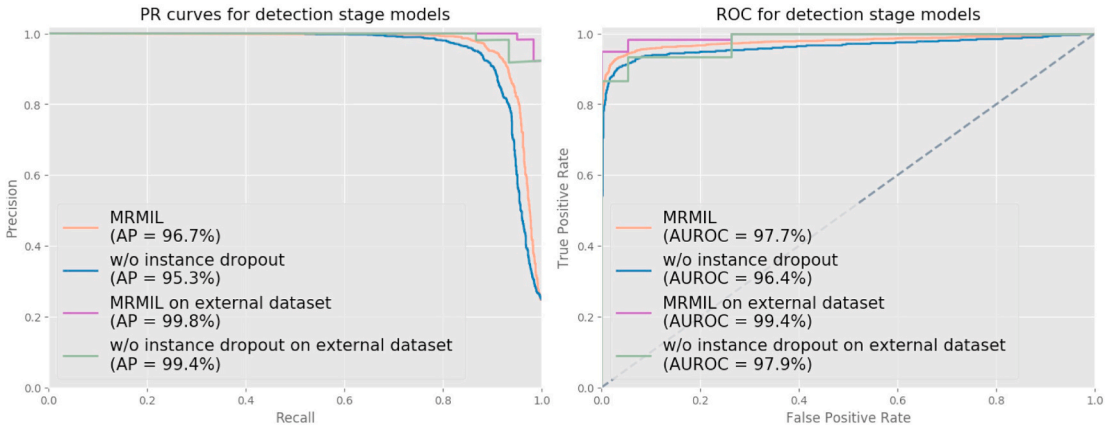

Fig. 2. ROC and PR curves for detection stage models on our test set and external dataset. In the detection stage, models were trained to distinguish malignant and benign slides with all tiles extracted from slides at 5x.

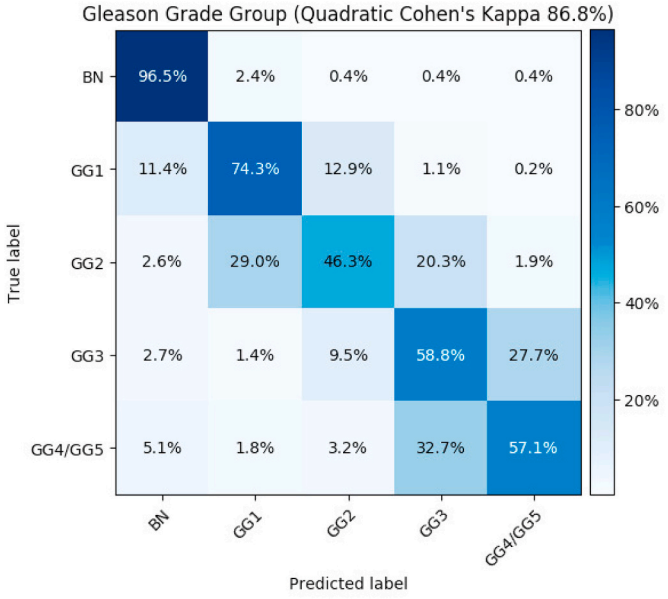

Fig. 3. Confusion matrix for Gleason grade group prediction.

regions of interest; 2) how to train a model with large-scale whole slide images, which are usually very large, contain heterogeneous contents and only labeled at the slide-level; and 3) how to effective combine tile-level information for slide-level classification.

Our detection stage model achieved promising results on both an internal test set and an external dataset, which demonstrates the generalizability of the model. One potential explanation for slightly better performances on external dataset is that our independent test set is relatively large (*i.e.* 7114 slides from 830 patients.) and is collected from clinical database without any data curation.

Handcrafted features-based models performed relatively well on differentiating benign and malignant slide with an AUC of 93.3%, however, they obtained much lower κ on the hard task of classifying LG, HG and BN slides. The model proposed by Campanella *et-al.* [6] first used an instance-based MIL approach, which considered tiles with highest probabilities as having the same label as the corresponding slide, and then utilized the RNN model to aggregate representations from top tiles for slide classification. In contrast, our model used a more flexible attention aggregation method that can detect discriminative tiles and combine tile-level features in the same time. The model [6] achieved comparable performance on detecting cancerous slides with 98.3% AUC and 97.3% AP. Yet, it showed inferior results on predicting LG, HG, and BN classes compared with the MRMIL model. Nagpal *et-al.* developed a two-stage model for Gleason grade prediction of prostate cancer biopsy slides [39]. Their first stage model, which was trained to provide tile-level Gleason pattern classification, was developed using 114 million labeled tiles from over 1000 slides of prostatectomies and

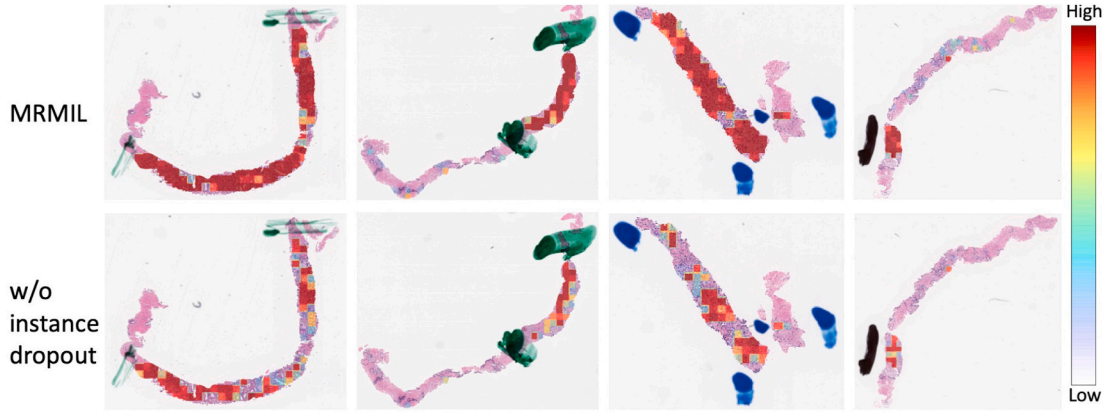

Fig. 4. WSIs overlaid with attention maps generated from the first stage cancer detection model. Pen marks as mentioned in Section 3.1 indicate cancerous regions. The first row shows attention maps from the model with instance dropout, while the second row is from the model without using instance dropout. Figures are best viewed in color.

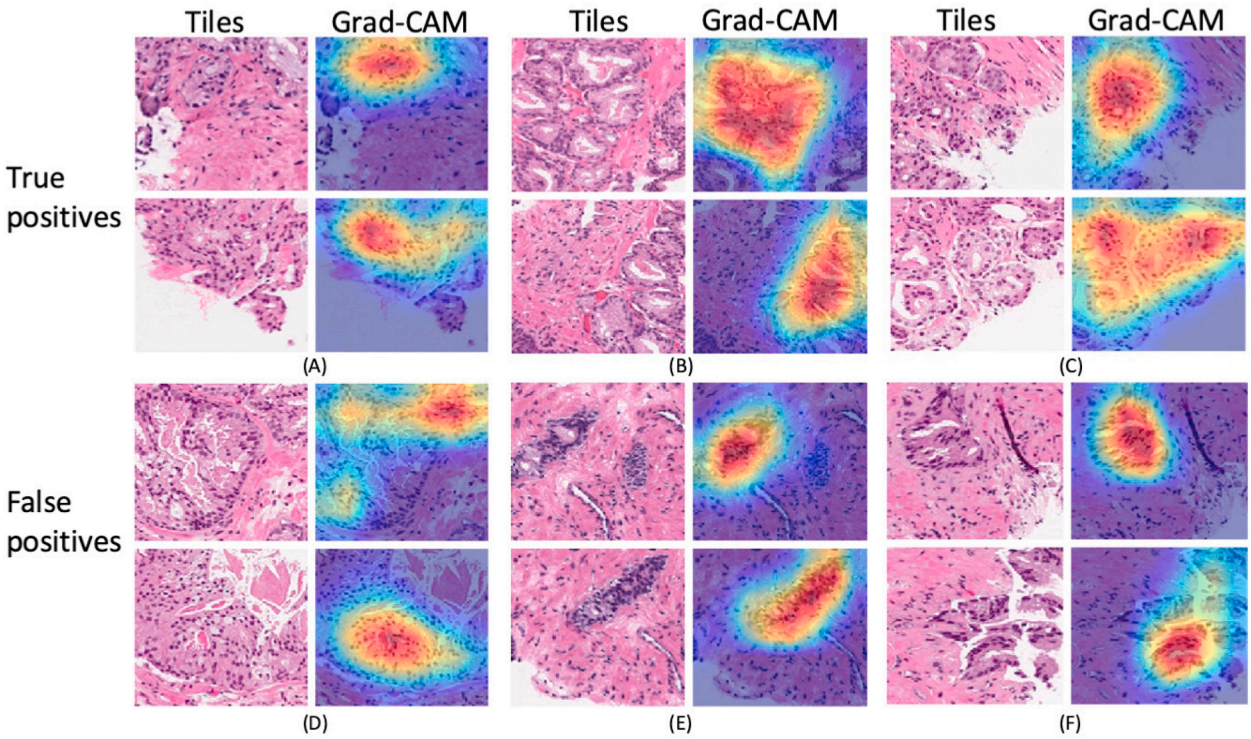

Fig. 5. Visualization of discriminative regions within tiles for TP and FP predictions. For each slide (A)–(F), we selected the top two tiles with the highest attention weights from the model, which were then forwarded to the model to generate activations and gradients for Grad-CAM.

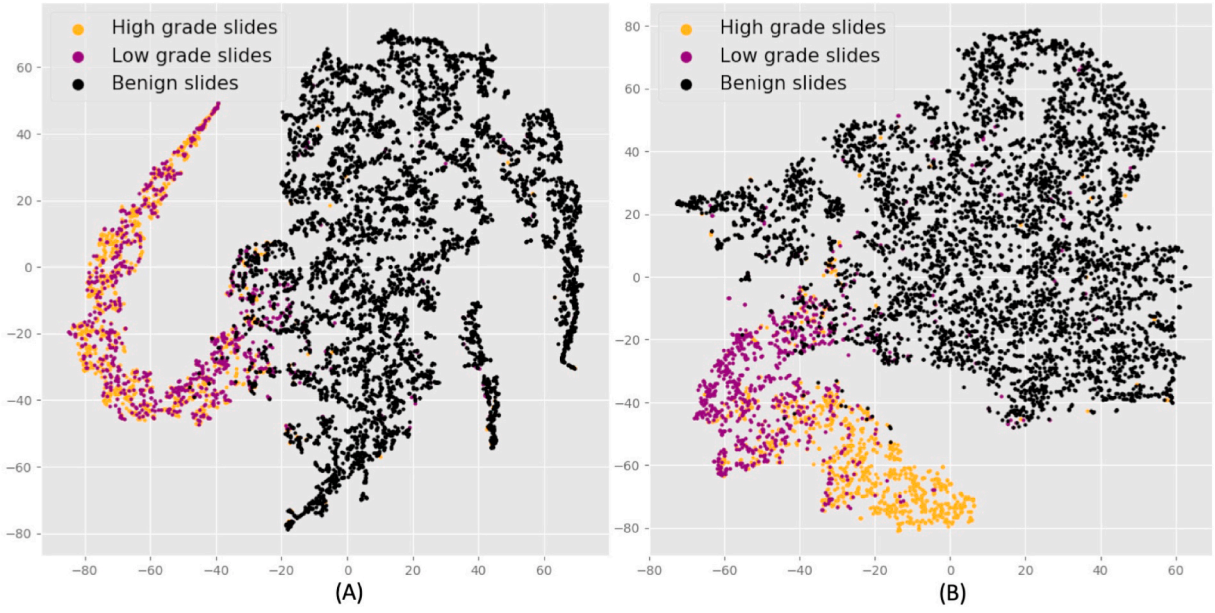

Fig. 6. t-SNE visualization of slide-level features. Black dots denote benign, purple dots indicate LG, and orange dots represent HG slides. (A) is the slide-level representations from the detection stage model. There is distinct separation between benign and cancerous slides. (B) shows the slide-level features from the classification stage model. We can see a better separation between LG and HG slides.

biopsies. The model obtained a κ of 71.7% on GG1, GG2, GG3 and GG4/5 prediction. Our model, which does not rely on fine-grained annotations and can be trained with only slide-level labels, achieved a comparable performance ($\kappa = 71.1\%$).

The quality of attention maps from the detection stage model is essential for selecting discriminative regions for the classification stage model. As shown in Fig. 4, attention maps learned with only weak (*i.e.* slide-level) labels are consistent with cancerous regions identified by pathologists during diagnosis. This demonstrates that our detection stage model not only achieves strong performance in classifying malignant versus benign slides, but also identifies suspicious regions for classification stage models. In addition, the generated attention maps can be integrated into a WSI viewer to potentially help pathologists more quickly localize relevant areas and reduce diagnostic time. Fig. 4 also shows that the original attention-based MIL model [23] (*i.e.* w/o instance dropout) only focuses on a few most discriminative tiles instead

Table 2
Model performance on BN, LG, and HG slides classification.

		BN, LG, HG Classification		Cancer Detection	
Experiment Name	Model Details	Cohen's Kappa (%)	Acc (%)	AUROC (%)	AP (%)
Handcrafted + RF	90 radiomics features + RF at 10x	57.0	81.5	93.1	83.9
Handcrafted + Xgboost	90 radiomics features + Xgboost at 10x	55.9	80.9	93.3	83.9
Campanella *et-al.* [6]	MIL + RNN at 10x	77.2	90.7	98.3	97.3
Tomczak *et-al.* [55]	Noisy-Or aggregation at 10x	n/a	n/a	97.4	96.0
Mean aggregation	Mean aggregation at 10x	77.1	90.8	97.9	96.6
Max aggregation	Max aggreagtion at 10x	79.5	91.9	97.4	96.3
Single stage	Single resolution MIL at 5x	76.3	90.5	97.4	95.8
Br selection [3,27,56]	Multi-resolution + Br	76.0	90.8	95.9	94.3
W/o instance dropout [23]	Multi-resolution + Att	77.3	91.0	97.3	96.0
Att selection	Multi-resolution + Att + instance dropout	80.7	92.4	**98.4**	**97.4**
MRMIL	Multi-resolution + Att + instance dropout + clusters	**81.8**	**92.7**	98.2	**97.4**

of entire suspicious regions. As reflected in Table 2, the w/o instance dropout model obtained a κ of 77.3%, which is about 4% lower than the one trained with instance dropout. Moreover, the performance of the model that relied on the Br image is inferior to the models that utilized attention maps. This demonstrates that areas with the most blue color may not be diagnostic relevant regions and that our attention module is able to extract high-level predictive representations rather than purely color features.

Grad-CAM visualization facilitates understanding of predictions from 'black-box' deep learning models, as shown in Fig. 5. For TP predictions in Fig. 5(A)–(C), our model captured the most relevant parts in the tile, though some cancerous regions were missed. For example, the first tile in (B) contains densely clustered cancerous glands, but the corresponding Grad-CAM only highlighted the most central area, and cancerous glands closer to the boundary were not detected. FP predictions are usually also hard cases for pathologists, with features that resemble prostate cancer. For example, regions highlighted by Grad-CAM in (E) contain benign glands with increased number of basal cells due to tangential tissue sectioning. (F) in Fig. 5 shows the seminal vesicle/ejaculatory duct tissue that form small outpouching glands with amphophilic cytoplasm, which mimic malignant glands. Our model was only trained to detect and grade acinar adenocarcinoma for prostate biopsies. Interestingly, as shown in (D), the model was able to identify intraductal carcinoma of the prostate gland (IDC-P), which is usually associated with high-stage invasive cancer and adverse prognosis.

From Fig. 6 (A), we can see that benign slide representations are clustered together on the right and malignant slides form a small cluster on the left. There is no distinct separation between features from LG and HG slides, since the objective of the detection stage model is to classify cancerous versus benign slides. Fig. 6 (B) shows that features of LG and HG slides generated from the classification stage model form their own distinct clusters, and representations from LG slides lie closer to benign slides in the embedding space.

To quantitatively evaluate our model performance, we performed experiments to understand the contribution of different model components, as summarized in Table 2. Using attention maps to select higher resolution tiles improved the κ of the one with br selection by 1%. Instance dropout further boosted the κ by over 3%. The final model MRMIL with all components achieved the highest κ for BN, LG, and HG classification, 98.2% AUROC for detecting malignant slides, and a quadratic κ of 86.8% for GG prediction, which is comparable to state-of-the-art models that require pre-trained segmentation networks [5].

6. Limitations and future work

In this section, we discuss limitations of this work and some potential directions for future research. In this work, we developed a two-stage model to first detect suspicious regions and then classify cancer grade with selected tiles at a higher magnification. The tile selection is determined by the detection stage model, and there is no mechanism to adaptively update selected tiles according to the loss from the classification stage model. In future work, a recurrent network or reinforcement learning can be incorporated to dynamically resample suspicious regions during training.

We only developed the model for acinar adenocarcinoma detection and classification for prostate biopsies. Other prognostically relevant histopathological types, such as ductal adenocarcinoma and IDC-P, need to be investigated in future studies.

In this study, we merely qualitatively evaluated our attention maps by visually inspecting learned maps for slides with pen markers. Though our model was able to identify similar regions as indicated by pen markers, quantitative evaluation with manual region-level annotations could provide a better metric for the attention module.

Additionally, besides achieving promising κ and Acc, a successful CAD tool should be able to facilitate clinical diagnosis. In future work, we will investigate different approaches to evaluate the effectiveness of our model as a CAD tool.

7. Conclusion

In this paper, we developed a novel MRMIL model that consists of a detection stage and a grade classification stage. The model can be trained with weak supervision from slide-level labels and localize cancerous regions. We provided visualization of saliency maps at both the slide- and tile-level, and learned representations to enhance model interpretability. The model was developed and evaluated on a dataset with over 20k prostate slides from 830 patients and an external dataset [18], and achieved promising performance. We believe that these types of models could have multiple applications in the clinic, including allowing pathologists to increase their efficiency, empowering more general pathologists to perform at the level of experts, and performing "second reads" of biopsy slides for quality assurance.

Declaration of competing interest

Corey W. Arnold, Jiayun Li, William Speier, and Wenyuan Li are inventors for the submitted patent application: Application No. 62/852,625.

HuiHui Ye is a consultant for Janssen Pharmaceuticals.

Anthony Sisk: None Declared.

Dr. Wallace declares no relevant conflicts with regard to this research.

Acknowledgments

The authors would like to acknowledge support from the NIH/NCI R21CA220352.

A. Detailed Model Architecture

Table 3 shows the detailed architecture for the first stage cancer detection stage model, and Table 4 shows the architecture for the classification stage model. Two stage models were trained separately.

Table 3
Cancer detection stage model architecture.

Module	Layers	Number of filter	Filter size	Output size
Input		–	–	3 × 128 × 128
VGG11bn	Conv + BN + ReLU	64	3 × 3	64 × 128 × 128
	Max Pool	64	2 × 2	64 × 64 × 64
	Conv + BN + ReLU	128	3 × 3	128 × 64 × 64
	Max Pool	128	2 × 2	128 × 32 × 32
	Conv + BN + ReLU	256	3 × 3	256 × 32 × 32
	Conv + BN + ReLU	256	3 × 3	256 × 32 × 32
	Max Pool	256	2 × 2	256 × 16 × 16
	Conv + BN + ReLU	512	3 × 3	512 × 16 × 16
	Conv + BN + ReLU	512	3 × 3	512 × 16 × 16
	Max Pool	512	2 × 2	512 × 8 × 8
	Conv + BN + ReLU	512	3 × 3	512 × 8 × 8
	Conv + BN + ReLU	512	3 × 3	512 × 8 × 8
	Max Pool	512	2 × 2	512 × 4 × 4
Instance feature embedding	Conv	256	1 × 1	256 × 4 × 4
	FC + ReLU + Dropout	–	–	256
Attention module	FC + Tanh + Dropout	–	–	512
	FC	–	–	1
Classifier	FC	–	–	1

Table 4
Cancer classification stage model architecture.

Module	Layers	Number of filter	Filter size	Output size
Input		–	–	3 × 256 × 256
	Conv + BN + ReLU	64	3 × 3	64 × 256 × 256
	Max Pool	64	2 × 2	64 × 128 × 128
	Conv + BN + ReLU	128	3 × 3	128 × 128 × 128
	Max Pool	128	2 × 2	128 × 64 × 64
	Conv + BN + ReLU	256	3 × 3	256 × 64 × 64
	Conv + BN + ReLU	256	3 × 3	256 × 64 × 64
VGG11bn	Max Pool	256	2 × 2	256 × 32 × 32
	Conv + BN + ReLU	512	3 × 3	512 × 32 × 32
	Conv + BN + ReLU	512	3 × 3	512 × 32 × 32
	Max Pool	512	2 × 2	512 × 16 × 16
	Conv + BN + ReLU	512	3 × 3	512 × 16 × 16
	Conv + BN + ReLU	512	3 × 3	512 × 16 × 16
	Max Pool	512	2 × 2	512 × 8 × 8
Instance feature embedding	Conv	256	1 × 1	256 × 8 × 8
	FC + ReLU + Dropout	–	–	256
Attention module	FC + Tanh + Dropout	–	–	512
	FC	–	–	3
Classifier	FC	–	–	3

B. Blue Ratio Conversion

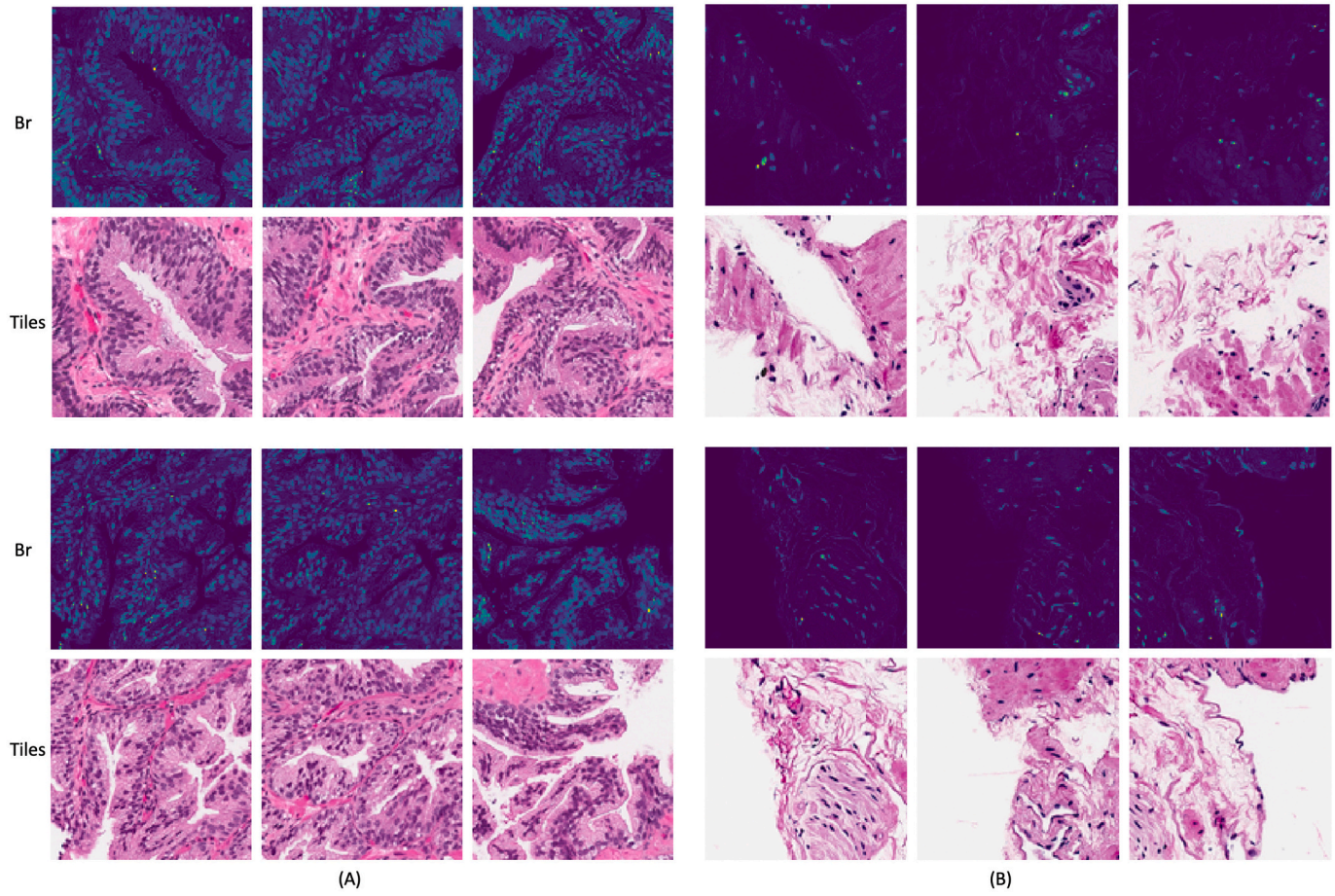

Fig. 7. Br conversion. We performed Br conversion for slides at 5x. The first two rows demonstrate tiles from a benign slide and the bottom two show ones from a malignant slide. (A) are 3 tiles with the highest average tile-level Br values, and (B) are ones with the lowest Br values. We can see that the br conversion is able to highlight regions with most nuclei.

C. K-Means Clustering

We used PCA to project n × 256 instance-level embedding vectors of n tiles to n × 32 (i.e., the number of components is set to be 32). For K-means clustering, the k was set to be 3 to encourage tiles to be grouped into benign, low-grade and high-grade clusters. Our attention clustering-based selection method was robust to different initializations. Specifically, we re-ran the K-means clustering with different random seeds for 10 times, and computed mean intersection over union (IoU) for selected tiles. Our method achieved a mean IoU of 97.65%.

Table 5
Model performance on BN, LG, and HG slides classification

MRMIL with different backbones	BN, LG, HG classification		Cancer detection	
	Cohen's Kappa (%)	Acc (%)	AUROC (%)	AP (%)
VGG11bn	81.8	92.7	98.2	97.4
VGG13bn	79.9	92.0	97.8	96.9
ResNet34	78.7	91.6	96.9	95.3

D. Different CNN Architectures for the Feature Extractor

We performed experiments to evaluate our MRMIL model performances with different network backbones. Experiments were performed by replacing the feature extract in both model stages with different CNN architectures. As shown in Table 5, the VGG11bn achieved the best performance. Our model performances were affected around 2% by using different backbones. For example, VGG13bn obtained a κ of 79.9%, which was 1.9% lower than the VGG11bn architecture. We are going to investigate and incorporate other powerful CNN feature extractors into our framework to further improve model performances in the future work.

References

[1] J. Amores, Multiple instance classification: review, taxonomy and comparative study, Artif. Intell. 201 (2013) 81–105.

[2] S. Andrews, I. Tsochantaridis, T. Hofmann, Support vector machines for multiple-instance learning, in: Advances in Neural Information Processing Systems, 2003, pp. 577–584.

[3] E. Arvaniti, M. Claassen, Coupling Weak and Strong Supervision for Classification of Prostate Cancer Histopathology Images, 2018 arXiv preprint arXiv:1811.07013.

[4] W. Bulten, H. Pinckaers, H. van Boven, R. Vink, T. de Bel, B. van Ginneken, J. van der Laak, C. Hulsbergen-van de Kaa, G. Litjens, Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study, Lancet Oncol. 21 (2020) 233–241.

[5] W. Bulten, H. Pinckaers, H. van Boven, R. Vink, T. de Bel, B. van Ginneken, J. van der Laak, C.H.v. de Kaa, G. Litjens, Automated Gleason Grading of Prostate Biopsies Using Deep Learning, 2019 arXiv preprint arXiv:1907.07980.

[6] G. Campanella, M.G. Hanna, L. Geneslaw, A. Miraflor, V.W.K. Silva, K.J. Busam, E. Brogi, V.E. Reuter, D.S. Klimstra, T.J. Fuchs, Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nat. Med. 25 (2019) 1301–1309.

[7] M.A. Carbonneau, V. Cheplygina, E. Granger, G. Gagnon, Multiple instance learning: a survey of problem characteristics and applications, Pattern Recogn. 77 (2018) 329–353.

[8] A. Chaddad, M.J. Kucharczyk, T. Niazi, Multimodal radiomic features for the predicting gleason score of prostate cancer, Cancers 10 (2018) 249.

[9] A. Chaddad, T. Niazi, S. Probst, F. Bladou, M. Anidjar, B. Bahoric, Predicting gleason score of prostate cancer patients using radiomic analysis, Front. oncol. 8 (2018) 630.

[10] H. Chang, L.A. Loss, B. Parvin, Nuclear segmentation in h&e sections via multi-reference graph cut (mrgc), in: International Symposium Biomedical Imaging, 2012.

[11] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[12] V. Cheplygina, M. de Bruijne, J.P. Pluim, Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis, Med. Image Anal. 54 (2019) 280–296.

[13] G. Currie, K.E. Hawk, E. Rohren, A. Vial, R. Klein, Machine learning and deep learning in medical imaging: intelligent imaging, J. Med. Imag. Radiat. Sci. 50 (2019) 477–487.

[14] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.

[15] T.G. Dietterich, R.H. Lathrop, T. Lozano-Pérez, Solving the multiple instance problem with axis-parallel rectangles, Artif. Intell. 89 (1997) 31–71.

[16] S. Doyle, M. Feldman, J. Tomaszewski, A. Madabhushi, A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies, IEEE Trans. Biomed. Eng. 59 (2010) 1205–1218.

[17] J.I. Epstein, M.J. Zelefsky, D.D. Sjoberg, J.B. Nelson, L. Egevad, C. Magi-Galluzzi, A.J. Vickers, A.V. Parwani, V.E. Reuter, S.W. Fine, et al., A contemporary prostate cancer grading system: a validated alternative to the gleason score, Eur. Urol. 69 (2016) 428–435.

[18] Á.E. Esteban, M. López-Pérez, A. Colomer, M.A. Sales, R. Molina, V. Naranjo, A new optical density granulometry-based descriptor for the classification of prostate histological images using shallow and deep Gaussian processes, Comput. Methods Progr. Biomed. 178 (2019) 303–317.

[19] R. Farjam, H. Soltanian-Zadeh, K. Jafari-Khouzani, R.A. Zoroofi, An image analysis approach for automatic malignancy determination of prostate pathological images, Cytometry Part B: Clinical Cytometry: The Journal of the International Society for Analytical Cytology 72 (2007) 227–240.

[20] A. Gertych, N. Ing, Z. Ma, T.J. Fuchs, S. Salman, S. Mohanty, S. Bhele, A. Velásquez-Vacca, M.B. Amin, B.S. Knudsen, Machine learning approaches to analyze histological images of tissues from radical prostatectomies, Comput. Med. Imag. Graph. 46 (2015) 197–208.

[21] L. Hou, D. Samaras, T.M. Kurc, Y. Gao, J.E. Davis, J.H. Saltz, Patch-based convolutional neural network for whole slide tissue image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2424–2433.

[22] Q. Hou, P. Jiang, Y. Wei, M.M. Cheng, Self-erasing network for integral object attention, in: Advances in Neural Information Processing Systems, 2018, pp. 549–559.

[23] M. Ilse, J.M. Tomczak, M. Welling, Attention-based Deep Multiple Instance Learning, 2018 arXiv preprint arXiv:1802.04712.

[24] N. Ing, Z. Ma, J. Li, H. Salemi, C. Arnold, B.S. Knudsen, A. Gertych, Semantic segmentation for prostate cancer grading by convolutional neural networks, in: Medical Imaging 2018: Digital Pathology, International Society for Optics and Photonics, 2018, 105811B.

[25] Z. Jia, X. Huang, I. Eric, C. Chang, Y. Xu, Constrained deep weak supervision for histopathology image segmentation, IEEE Trans. Med. Imag. 36 (2017) 2376–2388.

[26] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, 2014 arXiv preprint arXiv:1412.6980.

[27] P. Lawson, J. Schupbach, B.T. Fasy, J.W. Sheppard, Persistent homology for the automatic classification of prostate cancer aggressiveness in histopathology images, in: Medical Imaging 2019: Digital Pathology, International Society for Optics and Photonics, 2019, 109560G.

[28] J. Li, K.V. Sarma, K.C. Ho, A. Gertych, B.S. Knudsen, C.W. Arnold, A multi-scale u-net for semantic segmentation of histological images from radical prostatectomies, in: AMIA Annual Symposium Proceedings, American Medical Informatics Association, 2017, p. 1140.

[29] J. Li, W. Speier, K.C. Ho, K.V. Sarma, A. Gertych, B.S. Knudsen, C.W. Arnold, An em-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies, Comput. Med. Imag. Graph. 69 (2018) 125–133.

[30] W. Li, J. Li, K.V. Sarma, K.C. Ho, S. Shen, B.S. Knudsen, A. Gertych, C.W. Arnold, Path r-cnn for prostate cancer diagnosis and gleason grading of histological images, IEEE Trans. Med. Imag. 38 (2018) 945–954.

[31] A. Liaw, M. Wiener, et al., Classification and regression by randomforest, R. News 2 (2002) 18–22.

[32] G. Litjens, C.I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-Van De Kaa, P. Bult, B. Van Ginneken, J. Van Der Laak, Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis, Sci. Rep. 6 (2016), 26286.

[33] M.Y. Lu, D.F. Williamson, T.Y. Chen, R.J. Chen, M. Barbieri, F. Mahmood, Data Efficient and Weakly Supervised Computational Pathology on Whole Slide Images, 2020 arXiv preprint arXiv:2004.09666.

[34] L.v.d. Maaten, G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res. 9 (2008) 2579–2605.

[35] J. Melendez, B. van Ginneken, P. Maduskar, R.H. Philipsen, H. Ayles, C.I. Sánchez, On combining multiple-instance learning and active learning for computer-aided detection of tuberculosis, IEEE Trans. Med. Imag. 35 (2015) 1013–1024.

[36] J. Melendez, B. van Ginneken, P. Maduskar, R.H. Philipsen, K. Reither, M. Breuninger, I.M. Adetifa, R. Maane, H. Ayles, C.I. Sánchez, A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest x-rays, IEEE Trans. Med. Imag. 34 (2014) 179–192.

[37] C. Mercan, S. Aksoy, E. Mercan, L.G. Shapiro, D.L. Weaver, J.G. Elmore, Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images, IEEE Trans. Med. Imag. 37 (2017) 316–325.

[38] K. Nagpal, D. Foote, Y. Liu, P.H.C. Chen, E. Wulczyn, F. Tan, N. Olson, J.L. Smith, A. Mohtashamian, J.H. Wren, et al., Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer, NPJ Digital Med. 2 (2019) 48.

[39] K. Nagpal, D. Foote, F. Tan, Y. Liu, P.H.C. Chen, D.F. Steiner, N. Manoj, N. Olson, J. L. Smith, A. Mohtashamian, et al., Development and validation of a deep learning algorithm for gleason grading of prostate cancer from biopsy specimens, JAMA Oncol. 6 (9) (2020).

[40] K. Nguyen, B. Sabata, A.K. Jain, Prostate cancer grading: gland segmentation and structural features, Pattern Recogn. Lett. 33 (2012) 951–961.

[41] G. Nir, D. Karimi, S.L. Goldenberg, L. Fazli, B.F. Skinnider, P. Tavassoli, D. Turbin, C.F. Villamil, G. Wang, D.J. Thompson, et al., Comparison of artificial intelligence techniques to evaluate performance of a classifier for automatic grading of prostate cancer from digitized histopathologic images, JAMA Network Open 2 (2019) e190442–e190442.

[42] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic Differentiation in Pytorch, 2017.

[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[44] G. Quellec, M. Lamard, M. Cozic, G. Coatrieux, G. Cazuguel, Multiple-instance learning for anomaly detection in digital mammography, IEEE Trans. Med. Imag. 35 (2016) 1604–1614.

[45] E. Reinhard, M. Adhikhmin, B. Gooch, P. Shirley, Color transfer between images, IEEE Comput. Graphics Appl. 21 (2001) 34–41.

[46] M. Saha, C. Chakraborty, D. Racoceanu, Efficient deep learning model for mitosis detection using breast histopathology images, Comput. Med. Imag. Graph. 64 (2018) 29–40.

[47] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual Explanations from Deep Networks via Gradient-Based Localization, ICCV, 2017, pp. 618–626.

[48] D. Shen, G. Wu, H.I. Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (2017) 221–248.

[49] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2019, CA A Cancer J. Clin. 69 (2019) 7–34.

[50] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014 arXiv preprint arXiv:1409.1556.

[51] K.K. Singh, Y.J. Lee, Hide-and-seek: forcing a network to be meticulous for weakly-supervised object and action localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 3544–3553.

[52] K.K. Singh, H. Yu, A. Sarmasi, G. Pradeep, Y.J. Lee, Hide-and-seek: A Data Augmentation Technique for Weakly-Supervised Localization and beyond, 2018 arXiv preprint arXiv:1811.02545.

[53] R. Tennakoon, G. Bortsova, S. Ørting, A.K. Gostar, M.M. Wille, Z. Saghir, R. Hoseinnezhad, M. de Bruijne, A. Bab-Hadiashar, Classification of volumetric images using multi-instance learning and extreme value theorem, in: IEEE Transactions on Medical Imaging, 2019.

[54] J.H. Thrall, X. Li, Q. Li, C. Cruz, S. Do, K. Dreyer, J. Brink, Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success, J. Am. Coll. Radiol. 15 (2018) 504–508.

[55] J.M. Tomczak, M. Ilse, M. Welling, Deep Learning with Permutation-Invariant Operator for Multi-Instance Histopathology Classification, 2017 arXiv preprint arXiv:1712.00310.

[56] O.J. del Toro, M. Atzori, S. Otálora, M. Andersson, K. Eurén, M. Hedlund, P. Rönnquist, H. Müller, Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade gleason score, in: Medical Imaging 2017: Digital Pathology, International Society for Optics and Photonics, 2017, 101400O.
[57] J.J. Tosoian, S. Loeb, J.I. Epstein, B. Turkbey, P. Choyke, E.M. Schaeffer, Active surveillance of prostate cancer: use, outcomes, imaging, and diagnostic tools, in: American Society of Clinical Oncology Educational Book/ASCO. American Society of Clinical Oncology. Meeting, NIH Public Access, 2016, p. e235.
[58] J.J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.C. Fillion-Robin, S. Pieper, H.J. Aerts, Computational radiomics system to decode the radiographic phenotype, Canc. Res. 77 (2017) e104–e107.
[59] S. Wang, Y. Zhu, L. Yu, H. Chen, H. Lin, X. Wan, X. Fan, P.A. Heng, Rmdl: recalibrated multi-instance deep learning for whole slide gastric image classification, Med. Image Anal. 58 (2019), 101549.
[60] X. Wang, F. Tang, H. Chen, L. Luo, Z. Tang, A.R. Ran, C.Y. Cheung, P.A. Heng, Ud-mil: uncertainty-driven deep multiple instance learning for oct image classification, IEEE J. Biomed. Health Infor. 24 (12) (2020).
[61] Y. Xu, Z. Jia, L.B. Wang, Y. Ai, F. Zhang, M. Lai, I. Eric, C. Chang, Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features, BMC Bioinf. 18 (2017) 281.
[62] Z. Yan, Y. Zhan, Z. Peng, S. Liao, Y. Shinagawa, S. Zhang, D.N. Metaxas, X.S. Zhou, Multi-instance deep learning: discover discriminative local anatomies for bodypart recognition, IEEE Trans. Med. Imag. 35 (2016) 1332–1343.
[63] X. Zhang, Y. Wei, J. Feng, Y. Yang, T.S. Huang, Adversarial complementary learning for weakly supervised object localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1325–1334.