An Annotation-free Whole-slide Training Approach to Pathological Classification of Lung Cancer Types by Deep Neural Network

Chi-Long Chen
Taipei Medical University, College of Medicine https://orcid.org/0000-0003-2875-1669
Chi-Chung Chen
aetherAI Co. Ltd.
Wei-Hsiang Yu
aetherAI, Co. Ltd
Szu-Hua Chen
aetherAI, Co. Ltd
Yu-Chan Chang
Academia Sinica https://orcid.org/0000-0003-0474-9935
Tai-I Hsu
Academia Sinica
Michael Hsiao
Academia Sinica https://orcid.org/0000-0001-8529-9213
Chao-Yuan Yeh (joeyeh@aetherai.com)
aetherAI, Co. Ltd
Cheng-Yu Chen (sandychen@tmu.edu.tw)
Taipei Medical University, College of Medicine

Article

Keywords: Deep neural network learning, Whole-slide Training, Lung cancer classification

Posted Date: August 1st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-48727/v1

Version of Record: A version of this preprint was published at Nature Communications on February 19th, 2021. See the published version at https://doi.org/10.1038/s41467-021-21467-y.

An Annotation-free Whole-slide Training Approach to Pathological Classification of Lung Cancer Types by Deep Neural Network

Chi-Long Chen[1,2,3*], Chi-Chung Chen[4*], Wei-Hsiang Yu[4], Szu-Hua Chen[4], Yu-Chan Chang[5], Tai-I Hsu[6], Michael Hsiao[6], Chao-Yuan Yeh[4**], Cheng-Yu Chen[7,8**]

1 Department of Pathology, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan

2 Department of Pathology, Taipei Medical University Hospital, Taipei, Taiwan

3 Research Center for Artificial Intelligence in Medicine, Taipei Medical University, Taipei, Taiwan

4 aetherAI, Co. Ltd., Taipei, Taiwan

5 Department of Biomedical Imaging and Radiological Sciences, National Yang-Ming University, Taipei, Taiwan

6 Genomics Research Center, Academia Sinica, Taipei, Taiwan

7 Department of Radiology, School of Medicine, College of Medicine, Taipei Medical University, Taipei Taiwan

8 Department of Radiology, Taipei Medical University Hospital, Taipei, Taiwan

Correspondences:

Cheng-Yu Chen, Department of Radiology, Taipei Medical University Hospital, 252, WuXing Street, Hsin-Yi District, Taipei 11031, Taiwan

Tel: 886-2-27372181 ext. 1131

e-mail: sandychen@tmu.edu.tw

Chao-Yuan Yeh, aetherAI, Co. Ltd., 3-2, YuanQu Street, Nan-Gang District, Taipei 115, Taiwan

Tel: 886-2-27856892

e-mail: joeyeh@aetherai.com

* These authors contribute equally in this work.

** Correspondence and requests for materials should be addressed to C-Y Y. (joeyeh@aetherai.com) or C-Y C. (email: sandychen@tmu.edu.tw).

Abstract

Deep learning for digital pathology is hindered by the extremely high spatial resolution of whole slide images (WSIs). Most studies adopt patch-based methods which, however, require well annotated data for training. These are typically done by laboriously free-hand contouring on the WSI by experts. To both alleviate annotation burdens of experts and enjoy benefits from scaling up amounts of data, we develop a whole-slide training method for entire WSIs to classify types of lung cancers using slide-level diagnoses. Our method leverages unified memory to offload the excessive amount of memory consumption to host memory to train a classifier by entire hundreds-of-million-pixels slides. Experiments were conducted on the lung cancer dataset which contains 9,662 digital slides with various main types. The results showed that the proposed method can achieve an AUC of 0.950 and 0.924 for adenocarcinoma and squamous cell carcinoma on a separate testing set respectively. Furthermore, critical regions highlighted by the class activation map (CAM) technique of our model reveals a high correspondence to cancerous areas annotated by pathologists.

Key Words: Deep neural network learning, Whole-slide Training, Lung cancer classification

Introduction

Lung cancer is among the most frequently diagnosed cancer and the leading cause of cancer-related mortality in recent decades worldwide including Taiwan[1]. Non-small cell lung cancer (NSCLC) accounts for about 85% of newly diagnosed lung cancer cases, with two major histological types: adenocarcinoma and squamous cell carcinoma accounting for nearly 50% and 30% of NSCLC, respectively[2]. Invasive adenocarcinoma of lung is a malignant epithelial tumor including five major patterns: lepidic, acinar, papillary, micropapillary, and solid. Squamous cell carcinoma is a malignant epithelial tumor with squamous differentiation and/or keratinization. Proper pathologic diagnosis can be challenging in many cases since morphological differences among lung cancer types are subtle. Examples of pathological features of adenocarcinoma and squamous cell carcinoma are shown in Figure 1.

In the past few years, deep neural networks, especially convolutional neural networks (CNNs), have taken over as the dominant method for image recognition since their performance surpassed most of traditional image analysis algorithms in the ImageNet Large Scale Visual Recognition (ILSVRC) in 2012[3-6]. In medical fields, deep learning algorithms have also been demonstrated to attain human-level performance on several tasks including tumor identification and segmentation through CT or MRI[7-8], cardiovascular risk assessment through images of eyes[9], and pneumonia detection through chest-X-ray[10]. However, analysis of digital whole slide image (WSI) is still challenging because of its extremely high spatial resolution compared to other medical images such as X-ray, CT, or MRIs.

Restricted by computing limitations, most histopathology studies used a two-stage patch-based workflow: a patch-level CNNs training using patches cropped from WSI, followed by a slide-level algorithm trained on features extracted by the patch-level model, to learn the final diagnosis. These patch-based methods have yielded successful results such as cancer identification[11-19], cancer types classification[14,20], cancer metastasis[13,16-18] and prognosis analysis[21-22].

Multiple instance learning (MIL) follows the same two-stage workflow as the traditional method while organizing the training procedure in a different way. The main idea of MIL on slide-level cancer classification is that if the patches with highest scores (the most possible K patches) on the slide were identified as carcinoma, the slide should be classified into cancer, and vice versa. Moreover, recent studies show that even state-of-the-art weak supervision methods still cannot attain the average performance of strong supervision methods in most computer vision fields such as object detection, semantic segmentation, and instance segmentation tasks.

In this work, we implemented the unified memory (UM) CNNs to train models with huge inputs. Additionally, we trained models parallel and distributed with supercomputing clusters. Experiment results show that our method can apply to WSI classification directly and outperform the MIL method. Our contributions are summarized as follows: 1). To the best of our knowledge, this is the first study evaluating methods that only use slide-level annotations, namely the MIL and our proposed method, on classifying lung specimens into adenocarcinoma, squamous cell carcinoma, or non-cancers. 2). Experiment results demonstrate a superior performance of our proposed method that achieved AUC scores of 0.950 and 0.924 of adenocarcinoma and squamous cell carcinoma respectively, which outperforms the MIL. 3). Critical regions highlighted by the class activation map (CAM)[23] technique of our model reveals a high correspondence to annotations provided by pathologists.

Results

All experiments were conducted with input of slides scaled to 4x magnification to train both the whole slide model and the MIL model. For whole-slide training methods, the size of inputs was 20,000 x 20,000 pixels on average. We experimented with both global average pooling (GAP) and global max pooling (GMP) layers as the functional aggregation layers for our proposed method. For MIL models, the size of instances was set to 224 x 224 pixels without overlapping to train the classifier and aggregated by the max-pooling operation to derive the final prediction. Unless specified, models were trained and evaluated by different subsets of slide data from [Anonymous Institute A], [Anonymous Institute B] and [Anonymous Institute C].

Comparison of Overall Model Performance

We describe the evaluation method and results among experiments. Models with different settings on the main type classifying, especially adenocarcinoma and squamous cell carcinoma, of lung specimens are measured by using the area under the receiver operating characteristic curve (AUC).

As shown in Figure 2, the whole-slide training method with GMP layer achieves significantly better AUC scores compared to the MIL on both adenocarcinoma (0.950 (0.939-0.960) vs 0.899 (0.883-0.914), p-value=1.19e-15) and squamous cell carcinoma (0.924 (0.904-0.943) vs 0.832 (0.804-0.858), p-value=3.91e-15) classification, respectively.

Interestingly, the whole-slide training method with GAP layers derives a nearly 0.5 AUC score as 0.546 (0.516-0.576), p-value=0.00330 for adenocarcinoma and 0.528 (0.486-0.571), p-value=0.188 for squamous cell carcinoma on this classification task, which indicates that the model captures only a limited amount of information from inputs and reveal a random-guessing-level performance. Although GAP layers are widely adopted by state-of-the-art CNN models [3-6] for natural image classification, GAP layers applied on high-resolution images are prone to lose subtle information presented by tiny features. Such inefficiency leads to a significant downgrade of model performance compared to the whole-slide training method with GMP layers.

In addition, learning curves of both MIL and our method are illustrated in Figure 3. It shows a significant difference between the two methods in the early training stage. The performance of our proposed method increases sharply during the first few epochs and converges gradually in the rest training time. This is a typical learning pattern in training

DNNs since models tend to seek and learn features that can easily be divided in high dimensional spaces at first, and hence contributes to a drastic improvement in accuracy. Along with the training procedure, most of the obvious features are used and model performance saturates gradually. Subtle features, instead, are extracted in later training stages since models try to seek other high dimensional planes to minimize losses, and hence refine the model. In contrast, the learning curves of MIL are relatively smooth at first few epochs. According to the MIL training procedure, it relies on its work-in-progress model to choose representative tiles from whole slide images before training its classifier. However, models in the early training stage reveal a random-guess behavior since it has not yet learned any information. Those wrongly selected tiles will inevitably misguide the model and thus slow down the convergence rate.

Visualization

Though CNNs have led to impressive performance on the current classification task, it is more intriguing to understand how the models make decisions. Visualization is the most straightforward manner to investigate how models learn to solve the given task. Different visualization approaches are applied to different models since internal properties are not the same between MIL model and the whole-slide model.

For the MIL model, we derived a prediction map simply by feeding all the tiles cropped from the WSI into a patch-level classifier. These predictions indicated the probability of each local area containing cancerous features.

On the other hand, the whole-slide method is not capable of the above method because no patch-level classifier is available. Instead, we adopted class activation map (CAM)[23] to visualize critical regions.

As shown in Figure 4, both the MIL model and the whole-slide model could discover representative information, which was highlighted by heatmaps, after iteratively learning from slide-level diagnosis. Furthermore, our method revealed a more comprehensive ability to highlight all suspicious areas on the slide, especially small lesions. This could explain the reason why the whole-slide method achieves a better performance.

Throughput Comparison

To overcome the GPU memory limitation, our approach leverages Unified Memory (UM) to offload temporary data to host memory. However, the low access rate of system memory significantly cuts back the overall performance and lengthens the overall training

time. Therefore, we increased the UM efficiency by optimizing memory access and converting the model into mixed precision graph [26] during training time. Throughput improvements were measured by the number of slides being processed per second. To control the baseline among training methods and various speed improvement skills, all slides are resized to 10,000 pixels in both width and height dimensions beforehand. As shown in Figure 5, the throughput of the whole-slide training method can be speeded up 3.53x by incorporating both optimized memory access and using a mixed precision training.

Additionally, we deployed the training pipeline on [Anonymous Computing Center], a multi-GPU, a multi-node supercomputing environment. Given a hardware configuration of 4 computing nodes, 16 GPUs in total, the training process could achieve a 52.75x throughput compared to single-GPU, non-optimized one. As illustrated in Figure 5, the scaling efficiency was 93.32% which implies few performance overheads were paid when training models with our proposed method distributed.

Compared to the throughput of the MIL method which does not get involved with Unified Memory, it was 3.53x faster than our proposed method (Figure 5A) in model training. On the other hand, the inference throughput of our method, instead, outperforms that of MIL by 1.93x (Figure 5C). The huge discrepancy between the speeds of training and inference of our method is due to their different memory consumptions. In our case with Tesla V100 GPU and 10,000 x 10,000 slide input, all the temporary data produced during inference phase can fit in the GPU memory. Thus, model inference of our method yields a better throughput without the overhead incurred by Unified Memory access.

In general, given the setting of 8 GPUs in the training phase, it took around 2 weeks for our proposed method and a week for MIL to reach convergence on the 5,045-slide training dataset. As for the inference phase, it took nearly 4.5 hours by our proposed method and took about 9 hours to complete all 1,397 slides in test set by using a single GPU.

Discussion

Generating detailed annotations on WSIs is extremely laborious, cumbersome, and costly. For example, it takes an expert over half to an hour to annotate a single WSI into different parts of regions, for instance normal tissue region, carcinoma region and fibrosis region, and yet represents only partial regions in the WSI. Borders between tissue types are often ambiguous, leading inconsistent annotations between pathologists. The high variability of tissue morphology makes it difficult to cover all possible examples during annotation. These annotation shortcomings make deep learning models strongly biased by expert-defined annotations and difficult to learn comprehensively. To avoid requirements of enormous annotations and selection biases from experts, most of recent studies aim to seek weak supervision methods that can train deep learning models to explore relationships inside WSI to its clinical diagnosis directly.

Training cancer classifiers without detailed annotations alleviates the burden of annotation from experts, and allows deep neural network models to benefit from a huge amount of raw slide data with clinical diagnoses. Compared to previous works, methods on detecting cancers with strong supervision models by patch-wise annotations[11-19] still outperformed weak supervision models. However, research on weak supervision models for cancer detecting is gradually increasing since the model trained in a strong supervision manner is limited by how targets are annotated. On the other hand, even weak supervision is challenging but is easier to derive/update annotations according to different goals.

Although MIL can be trained with weak labels such as slide diagnosis, several drawbacks that may possibly affect the model performance: 1). Wrongly selected tiles in early training iterations may make models trapped in the local minimum due to random initialization of models. In some situations, models even stop improving after the first few epochs. 2). The MIL tries to utilize the k-most representative tiles while abandoning other relative information contained in others tiles. However, the size of k-most representative tiles is a hyperparameter and the true informative regions may vary from slides to slides. These k-most representative tiles may either overtake irrelevant tiles as positives or lack capacities to include atypical patterns that are crucial to diagnosis.

Instead of modifying algorithms and the training pipeline into a weak supervision form to bypass the out-of-memory issue, we leveraged the unified memory, UM, to train CNNs with huge image inputs directly without modifying any training pipeline. The UM enables GPUs to access the host memory directly, which expands the total capacity of temporally data from gigabytes level into terabytes level. Though the UM mechanism that can swap

memory between GPUs and the host, it will drastically slow down the forward and backward propagation of CNNs due to frequent exchange of data between GPUs and the host.

In this study, we demonstrated that CNNs can identify features of adenocarcinoma and squamous cell carcinoma of the lung by using merely slide-level labels. Instead of using MIL, a commonly used training procedure for weak supervision tasks, we leveraged the UM to enable whole-slide training. Our proposed method achieved an AUC of 0.950 and 0.924 for adenocarcinoma and squamous cell carcinoma respectively. It also showed a high correspondence between the class-activation-map (CAM) and cancerous areas identified by experts.

Importantly, improving both classification accuracy and lesion localization is plausible by acquiring more data. Comparing results of the cross-sites experiment against single site experiments, a significant improvement was made merely by augmenting the dataset. Compared to Coudray et al.[14], training patch-level models with 1,634 slides and tens of millions of patch-level annotations, and Campanella et al.[24], training MIL models with over tens of thousands of slides, our model had already achieved a competitive performance with only 7,003 weakly-labeled slides.

The lesion localization of our model by CAM revealed a great coverage in most cases. It is to be noted, however, that the semantics of CAM in our current method was slightly different from locating cancer cells. Areas highlighted by CAM were highly related to predictions. In other words, deep neural networks (DNNs) will use distinguishable features across the given dataset to classify an image into groups, which may often bring side-effects such as contextual bias. For instance, a classifier trained to learn cars and boats will highlight not only the boat itself but water since boats are always accompanied by water. In our case, the CAM for the squamous cell carcinoma highlighted not only cancerous regions but necrotic regions. (Figure 6). Despite the fact that necrosis can be caused by other reasons such as injury, infection or inflammation, squamous cell carcinoma has a tendency to consume nutrients on the border of tissues, leading to ischemia in inner regions and eventually, necrosis. Hence, it becomes a signature when classifying adenocarcinoma from squamous carcinoma since it is rare to get biopsy from purely injury, infection or inflammation patients and thus there is far less data in our current dataset.

Since deep learning models collect all possible clues to make decisions, such weak relation, or halo effect, is learned to be useful when classifying adenocarcinoma and squamous cell carcinoma. One way to resolve this issue is to add a small amount of detailed-annotated slides to specify cancer cells. These annotations provide hints for models to

separate cancerous representations from those weakly related representations. Such integration for leveraging both slide-level annotations and limited detail annotations can be developed to achieve a more comprehensive and precise model in the future.

Materials and Methods

Dataset

A total of 9,662 hematoxylin and eosin (H&E) stained specimens collected from 2,843 patients were retrieved from [Anonymous Institute A], [Anonymous Institute B] and [Anonymous Institute C]. For each case, it might have at least one H&E slides gathering from either biopsy or different parts of the resected lung tissue and diagnosed as the dominant type of cancer or non-cancer tissue.

Because of a lack of samples in certain types of rare lung cancer, we filtered out cases that have less than 500 samples, in specific, those cases diagnosed as small-cell carcinoma and large-cell carcinoma were excluded. The final dataset contains 7,003 slides, including 3,876 cases of adenocarcinoma, 1,088 cases of squamous cell carcinoma and 2,039 cases of non-cancer tissues. The diagnoses of these patients consisted of a wide range of cancerous and non-cancerous types, confirmed by at least two pathologists.

The dataset was randomly split into training, validation and testing sets, containing 5045, 561 and 1397 slides respectively using a stratified sampling. Detailed numbers of slides from each site are listed in Table 1. Unless specified, experiments are conducted using cross-site dataset configuration mentioned above. To compare cross-site generalization ability of the model, we conducted model training by using single-site dataset, in specific, training and validation dataset from [Anonymous Institute A] only; then, evaluating the model by testing sets from [Anonymous Institute B] and [Anonymous Institute C] to retrieve cross-site performance.

The access of data was compliant with policies and national legislation for research. All data regarding patient information were removed. The dataset only contained WSIs and de-identified diagnosis.

All slides were scanned by Hamamatsu NanoZoomer XR in 20x magnification (0.46 μm per pixel) with original resolution up to 102,399 pixels per dimension.

Multiple Instance Learning

Most image binary classification tasks can be formed into a multiple instance learning (MIL) problem by dividing an image into multiple partial regions if the following criterion is met: a bag, or an image, is labeled as positive when the target shows up in at least one instance, or local region; and labeled as negative if target is absent from all instances. Hence, any positive bag can be represented by using several critical instances only whereas no suspicious instances should be left in negative bags.

This property makes training models with bag-level labels become possible by applying positive bag-level labels to critical instances in positive bags and negative bag-level labels to all instances in negative bags. During training, the MIL iteratively selects high-score instances from each bag when training a classifier.

To be more specific, the MIL method can be separated into two alternative steps: instance selection and classifier optimization, as illustrated in Figure 1. During instance selection, an instance selector, which computes probability of positive over instances of bags, is used to mine K-most positive instances from each bag. With selected instances, a classifier is trained to maximize the probability of instances selected from positive bags while minimizing the probability of instances selected from negative instances.

In the end of MIL, the classifier will be able to mine the most relative patterns of positive cases and the label of any given bag can be inferred by naive aggregation methods such as taking the maximum scores (max-pooling) or averaging scores of K-most instances (average-pooling).

Baseline MIL Method on Lung Cancer Type Classification

The lung cancer typing task with only slide-level labels can be considered as a MIL problem, since a slide is marked as either adenocarcinoma, squamous cell carcinoma, or non-cancer tissues.

As illustrated in Figure 7, we treated each slide as independent bags and cropped patches of 224x224 pixels inside each bag as instances to train a classifier and aggregated prediction of instances by the most widely adopted max-pooling method. For the instance classifier, a ResNet-50[3] with fixup initialization[25] is implemented. While the dominant lung cancer type classification mainly relies on inspecting tissue-level morphology rather than cell-level morphology as shown in Figure 1, we generate instances at a 4x magnification to provide sufficient observation scope for the classifier. The slides were augmented by flip, translation, and rotation followed by an intensity normalization before generating instances.

We removed instances belonging to the background by a color filter which drastically reduced total numbers of instances by 75% and speeded up the whole training process. To select representative instances of each bag, we set K=1 to pick up an instance that is most unlikely to be either adenocarcinoma or squamous cell carcinoma instance and mark instances of bags to its corresponding slide-level annotations.

Whole-slide Training Method

As a workaround algorithm for hardware memory constraints issue, the MIL alters typical CNNs training pipeline and thus produces several drawbacks. First, instances were nearly randomly selected in the early training phase because of random initialization of the classifier. These selected instances in the early stage strongly affect the trend of selection strategy of the classifier in the following training process. In some situations, wrongly selected instances, for instance accompany hyperplasia tissues but not cancer cells itself, may lead the classifier to fall into a local minimum and thus stop improving after a few epochs.

Second, the K-most representative instances may undertake informative regions or overtake irrelevant regions into account, which limits models to learn targets comprehensively. Since overtaking irrelevant instances as positives could be more severe to the training procedure, most implementations set K=1 to take the most relevant instance only, which turns out that models will lack capacities to include atypical patterns that are crucial to diagnosis.

Alternatively, we propose a whole-slide training method that incorporates the standard CNNs architecture with the unified memory (UM) mechanism to support inputs of hundreds of millions of pixels directly to train the models as usual (Figure 7). By using unified memory (UM), we are able to give GPUs direct access to host memory, which provides terabytes of memory instantly to accommodate most intermediate tensors during forwarding and back-propagation. While UM circumvents the memory constraint, the frequent data swapping for upcoming operations between the host memory and GPUs across rather slow hardware interfaces such as PCIe tremendously slows down the training throughput.

To mitigate the performance downgrade, several memory optimization techniques were applied. First, we prefetch data about to be used from host memory into GPU memory beforehand. By prefetching data, matrix operations can be executed continually without waiting. Arranged prefetch schedules were applied automatically through analyzing the computational graph of our model.

Second, we adopt the mixed precision training technique[26] to reduce memory consumption and the amount of data transfer. Most data are stored and computed in fp16 format instead of fp32, cutting back those costs to half. Overall, these memory optimization techniques speed up the whole-slide training method, even achieving higher throughput than MIL

Furthermore, we trained the model distributed with multiple GPUs. To compensate for the loss of randomness between batches caused by smoothness of averaging gradients

collected from different nodes, initial learning rate was multiplied by a square root of N nodes[27].

MIL Assumption in the Whole-slide Training

Training a classifier of natural images and a classifier of WSIs is a totally different scenario due to the difference in scale of image size. Typically, an image of 224 x 224 spatial resolution will be condensed into 2,048 feature maps of 7 x 7 spatial resolution after multiple stacks of convolutional and downsampling layers in the Resnet.

Since these layers conduct sliding window operations, each 7 x 7 feature map remains the same spatial arrangement as the input image. To be more precise, these layers can be deemed as a function that each pixel on the feature map encodes a certain size of region on the original input into a single 2,048 dimensional embedding vector.

The projection size of a pixel of feature maps corresponding to the original input can be referred to as a receptive field[28]. Information beyond a receptive field has no means to be encoded. According to operations of ResNet50, the receptive field of the final feature map is 483 x 483, which is larger than its common input size: 224 x 224. As a result, receptive fields of pixels on the final feature maps have already covered all information of the image.

Finally, the following global average pooling (GAP) layer is commonly applied to average feature maps of 7 x 7 x 2,048 into a 1 x 1 x 2048 vector. However, the receptive field of any given pixel on the final feature maps will no longer cover the whole image when enlarging the input into tens of thousands of pixels along its height and width. Such difference is critical in the cancer classification of WSIs. Since malignant regions may be relatively tiny compared to the whole tissues in the positive slides, only very few receptive fields cover critical areas.

With the majority voting aggregation, or the global averaging pooling (GAP), at the end of feature maps, critical signals were further diluted by signals coming from feature maps that are not relevant to patterns of cancers. It ultimately constraints the model to identify slides with small cancerous areas.

Inspired by the MIL, we replace the GAP layer by the global max pooling (GMP) layer, which only keeps the max value of each element of the 2048-long vectors, as shown in Figure 8. Large values appearing in the embedding vector implies meaningful features are extracted. By adopting GMP, those large values are kept and thus distinguishable signals behind them are preserved.

Experiment Setup

We conducted all experiments on [Anonymous Computing Center], a multiGPU, multi-node supercomputing environment. Each node is equipped with 8x Tesla V100 32GB-HBM2 GPUs. We used Tensorflow (version 1.15.0) for model building and training, and Horovod[29] (version 0.19.0) to enable multi-GPU parallel training. All the experiments were executed on 1 computing node with batch size 8, 1 sample per GPU.

Along with the training progress, the kernel weights were gradually updated with the process called stochastic gradient descent (SGD). We used the Adam optimizer[30] (with an initial learning rate of 0.00001 and decays to 0.000001 when validation loss does not improve in 16 epochs) to train the model and evaluate the performance per 704 training steps.

Statistics

We use the area under receiver operating characteristics (Area Under ROC, AUC) as evaluation metrics to measure the slide-level performance of different methods. The 95% confidence interval was obtained by using a bootstrap approach. The bootstrap procedure was as follows: stratified sampling with replacement n slides from the testing set, where n is the number of testing set size, and compute AUC. We repeated the sampling procedure to derive 10,000 bootstrap samples and report the 2.5 and 97.5 percentile values. When comparing the AUCs of two models, the p-value was also calculated by bootstrap approach with the same parameters above and one-sided hypothesis. To evaluate the significance level of the AUC of a model, we adopted a dummy model that always returns 0.5 as the null hypothesis. In this case, the p-value was calculated by bootstrap method with two-sided hypothesis.

For the throughput tests, we collected the elapsed time for a model to train on a batch, and repeated the same procedure for 30 times. The throughput value was obtained by the sample mean and the 95% confidence interval was acquired by calculating 1.96 times the sample standard deviation.

Acknowledgements

This work was supported by grants from Ministry of Sciences and Technology (grant number MOST108-3011-F-038-001), Taiwan. We thank Dr. Huai-Kuang Tsai and Dr. Trees-Juen Chuang for careful reading and giving advices of this manuscript.

Author contributions

C.Y. Chen and C.L. Chen initiated the study. C.C. Chen and W.H. Yu designed the experiments and wrote the code. C.C. Chen performed the experiments and analyzed the results. S.H. Chen and C.Y. Yeh reviewed the experiment results. Y.C. Chang, T.I. Hsu, and M. Hsiao critically reviewed and commended the manuscript. All authors contributed to the preparation of the manuscript.

Additional information

Data Availability

The slide data are not publicly available due to patient privacy constraints. The data that support the findings of this study are available on request from the corresponding author Chao-Yuan Yeh or Cheng-Yu Chen.

Code Availability

The code to generate the results in this study is available on request during the review stage.

Competing financial interests: The authors declare no competing financial interests.

Reference

1. Torre, L. A. F. *et al.* Global cancer statistics, 2012, *CA. Cancer J Clin* 65(2), 87–108 (2015).
2. Kim, H. S., Mitsudomi, T., Soo, R. A. & Cho, B. C. Personalized therapy on the horizon for squamous cell carcinoma of the lung. *Lung Cancer Amst Neth* 80(3) 249–255 (2013).
3. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *ArXiv151203385 Cs* (2015).
4. Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. Densely Connected Convolutional Networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI. 2261–2269 (2017).
5. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25*, Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q. Eds. Curran Associates, Inc., 1097–1105 (2012).
6. Simonyan K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv14091556 Cs* (2015).
7. Chen H. *et al.*, Low-dose CT via convolutional neural network. *Biomed Opt Express* 8(2) 679–694 (2017).
8. Lundervold A. S. & Lundervold, A. An overview of deep learning in medical imaging focusing on MRI. *Z Für Med Phys* 29(2) 102–127 (2019).
9. Poplin R. *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng* 2(3) 158–164 (2018).
10. Rajpurkar P. *et al.* CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *ArXiv171105225 Cs Stat* (2017).
11. Chuang W. Y. *et al.* Successful Identification of Nasopharyngeal Carcinoma in Nasopharyngeal Biopsies Using Deep Learning. *Cancers* 12(2) 507 (2020).
12. Burlutskiy, N. A Deep Learning Framework for Automatic Diagnosis in Lung Cancer. *CoRR* abs/1807.10466 (2018).
13. Liu Y. *et al.* Artificial Intelligence–Based Breast Cancer Nodal Metastasis Detection: Insights into the Black Box for Pathologists. *Arch Pathol Lab Med* 143(7) 859–868 (2019).
14. Coudray N. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat Med* 24(10) 1559–1567 (2018).

15. Gheisari, S., Catchpoole, D., Charlton, A. & Kennedy, P. Convolutional deep belief network with feature encoding for classification of neuroblastoma histological images. *J Pathol Inform* 9(1) 17 (2018).

16. Wang, D., Khosla, A., Gargeya, R., Irshad, H. & Beck, A. H. Deep Learning for Identifying Metastatic Breast Cancer. *ArXiv160605718 Cs Q-Bio* (2016).

17. Liu Y. *et al.* Detecting Cancer Metastases on Gigapixel Pathology Images. *ArXiv170302442 Cs* (2017).

18. Ehteshami Bejnordi B. *et al.* Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women with Breast Cancer. *JAMA* 318(22) 2199–2210 (2017.

19. Hou, L., Samaras, D., Kurc, T. M., Gao, Y., Davis, J. E. & Saltz, J. H. Patch-Based Convolutional Neural Network for Whole Slide Tissue Image Classification. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2424–2433 (2016).

20. Wei, J. W., Tafe, L. J., Linnik, Y. A., Vaickus, L. J., Tomita, N. & Hassanpour, S. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Sci Rep* 9(1) 3358 (2019).

21. Wang, S. *et al.* Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome. *Sci Rep* 8(1) 10393 (2018).

22 Mobadersany, P. *et al.* Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc Natl Acad Sci* 115(13) E2970–E2979 (2018).

23. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning Deep Features for Discriminative Localization. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA 2921–2929 2016).

24. Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* 25(8) 1301–1309 (2019).

25 Zhang, H., Dauphin, Y. N. & Ma, T. Fixup initialization residual learning without normalization. 16 (2019).

26. Micikevicius, P. *et al.* Mixed Precision Training. *ArXiv171003740 Cs* (2018).

27 Hoffer, E., Hubara, I., & Soudry, D. Train longer, generalize better: closing the generalization gap in large batch training of neural networks," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 1731–1741 (2017).

28 Araujo, A., Norris, W., & Sim, J. Computing Receptive Fields of Convolutional Neural Networks. *Distill* 4(11) e21 (2019).

29. Sergeev A. & Del Balso, M. Horovod: Fast and easy distributed deep learning in TensorFlow. *ArXiv180205799 Cs Stat* (2018).

30. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *ArXiv14126980 Cs*, (2017).

Figures

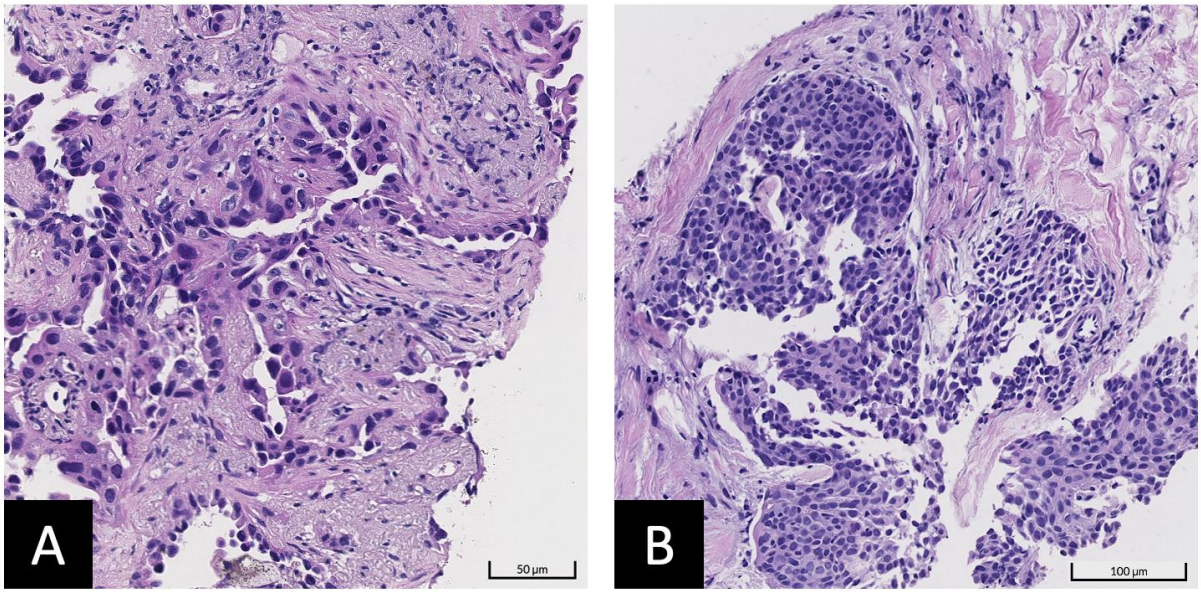

Figure 1 Examples of pathological images (A) adenocarcinoma and (B) squamous cell carcinoma.

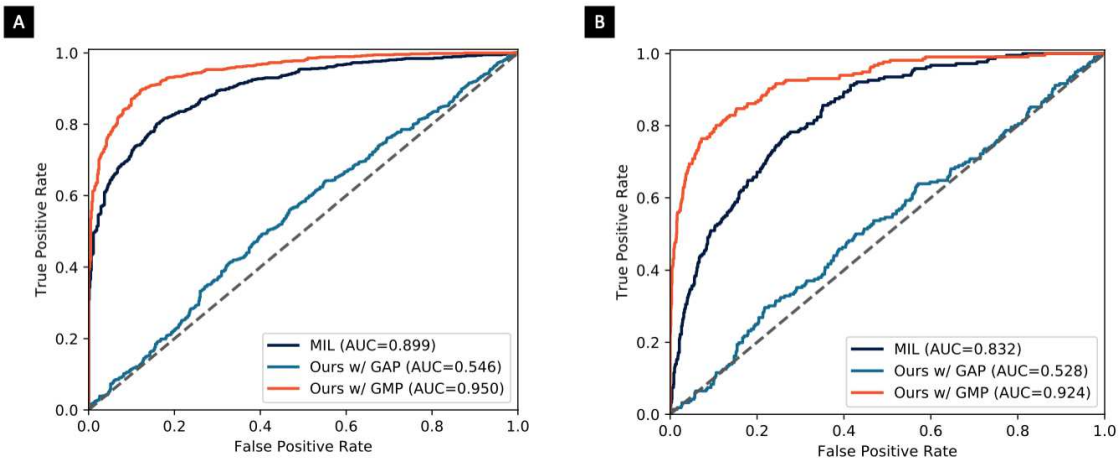

Figure 2 Receiver operating characteristic (ROC) curves of different training methods on classifying (A) adenocarcinoma and (B) squamous cell carcinoma on the testing dataset (n=1397). AUCs for classifying adenocarcinoma are 0.899 (0.883-0.914) by MIL, 0.546 (0.516-0.576) by our method with GAP and 0.950 (0.939-0.960) by our method with GMP. AUCs for classifying squamous cell carcinoma are 0.832 (0.804-0.858) by MIL, 0.528 (0.486-0.571) by our method with GAP and 0.924 (0.904-0.943) by our method with GMP.

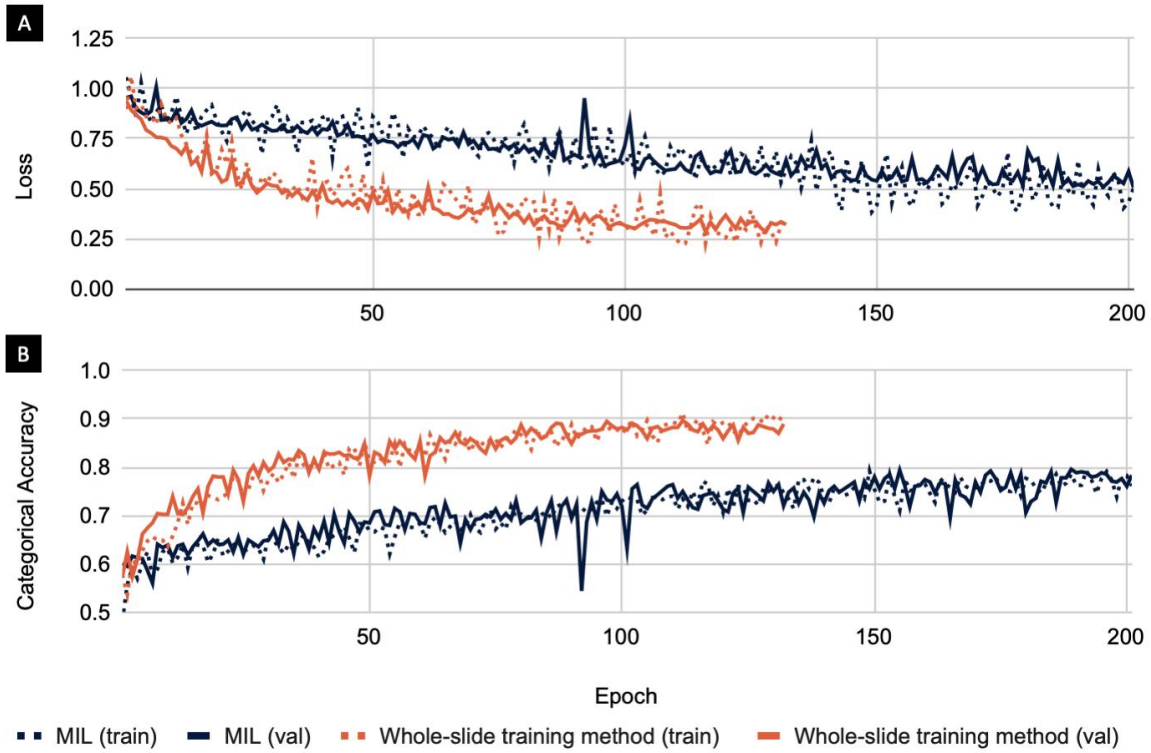

Figure 3 Learning curves of MIL and whole-slide training method. The dotted lines represent (A) the loss and (B) accuracy on the training set. Otherwise, the solid lines represent those on the validation set.

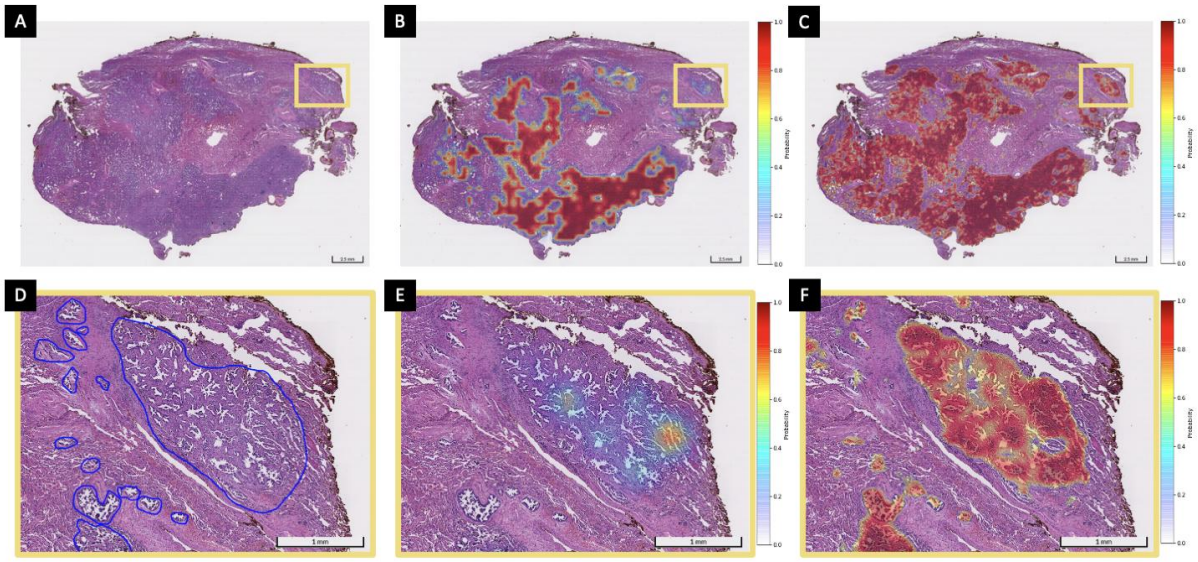

Figure 4 Visualization of heatmaps generated by MIL and whole-slide training method. (A) Whole-slide view on a slide containing adenocarcinoma lesions. (B) MIL heatmap on whole-slide view. (C) Heatmap generated by whole-image method. (D) Zoom-in view with reference human annotation. (E) Zoom-in view of MIL heatmap. (F) Zoom-in view of heatmap generated by whole-slide training method.

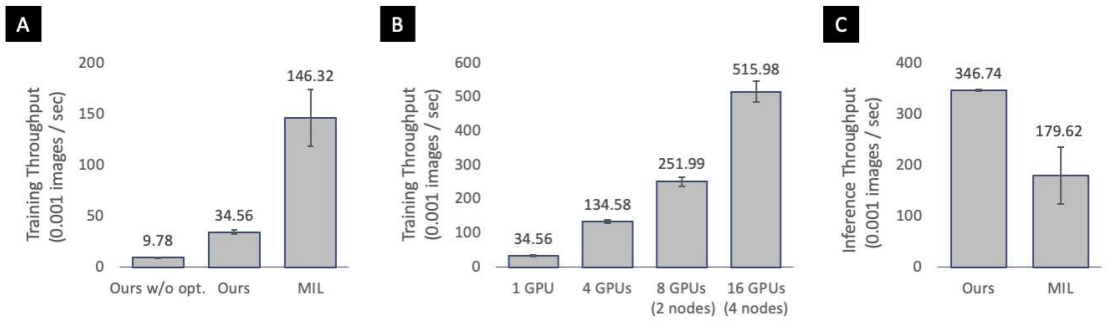

Figure 5 Throughput results of (A) whole-slide training method (ours) against non-optimized version and MIL, (B) distributed whole-slide training on multi-node, multi-GPU, and (C) inference on whole-slide training method and MIL.

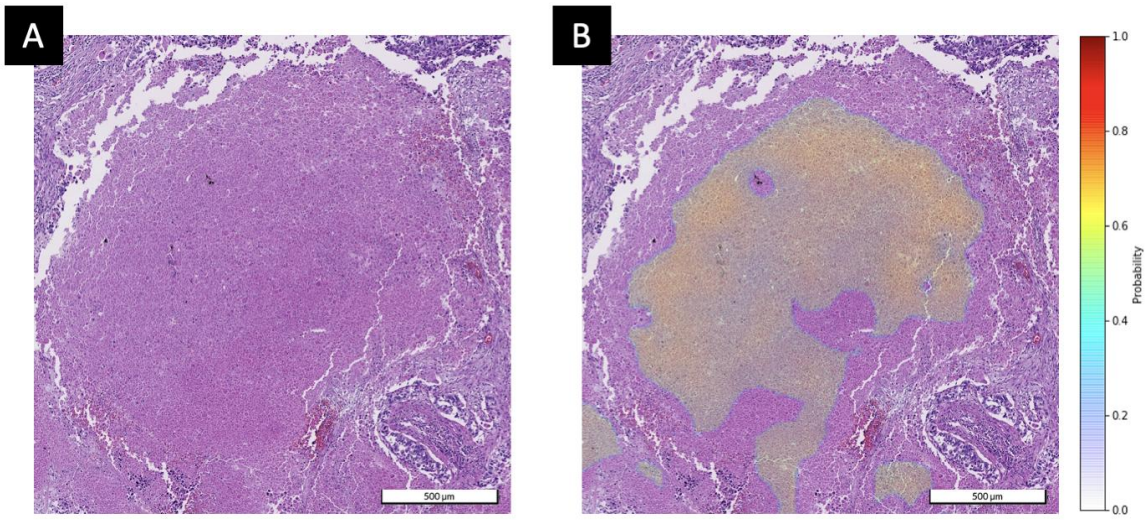

Figure 6 Tissue necrosis highlighted by CAM.

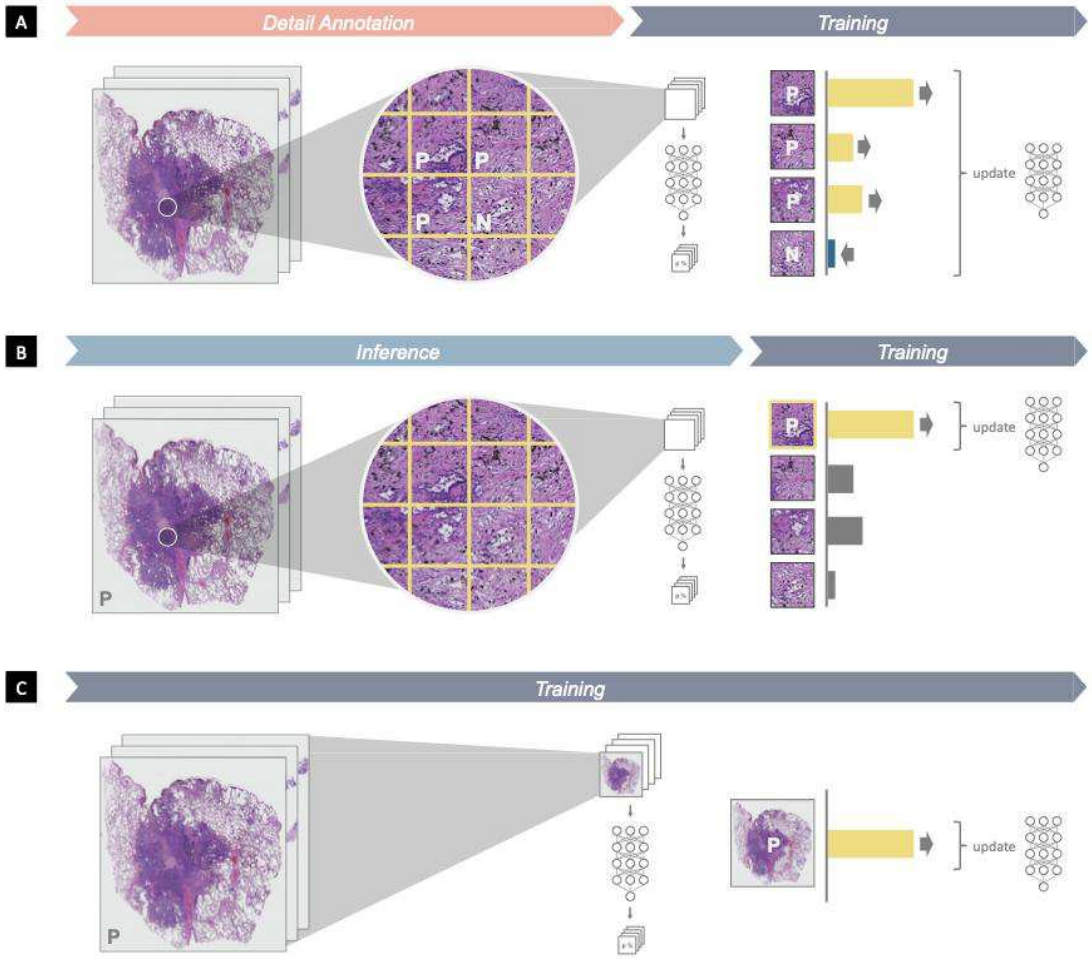

Figure 7 Workflows of (A) patch-based method (B) MIL method and (C) our proposed whole-slide training method.

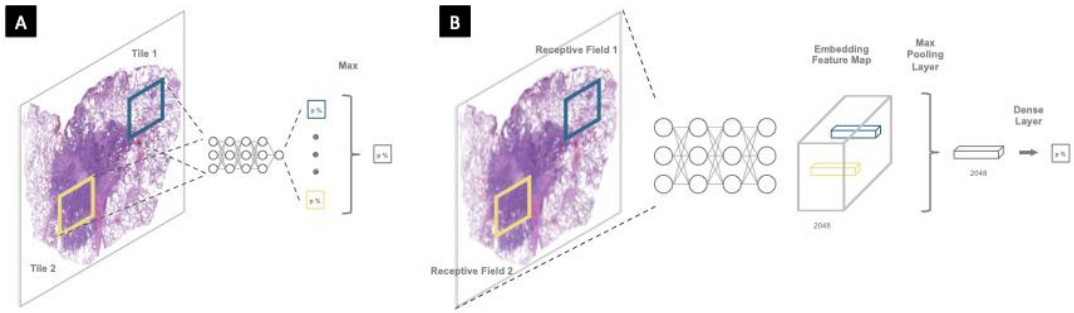

Figure 8 (A) Illustration of the mappings from input tiles to output predictions in MIL training. (B) Illustration of the mapping from receptive fields on input images to embedding feature maps using GMP in whole-slide training method.

Tables

Table 1 The number of slides of different lung cancer types for cross-site data collected from [Anonymous Institute A], [Anonymous Institute B] and [Anonymous Institute C].

		Non-cancer	Adenocarcinoma	Squamous cell carcinoma	Total
Training Set	[Anonymous Institute A]	953	401	98	5045
	[Anonymous Institute B]	509	1123	318	
	[Anonymous Institute C]	7	1271	365	
Validation Set	[Anonymous Institute A]	103	46	11	561
	[Anonymous Institute B]	58	123	38	
	[Anonymous Institute C]	2	138	42	
Testing Set	[Anonymous Institute A]	264	111	27	1397
	[Anonymous Institute B]	141	311	88	
	[Anonymous Institute C]	2	352	101	
					7003

Figures

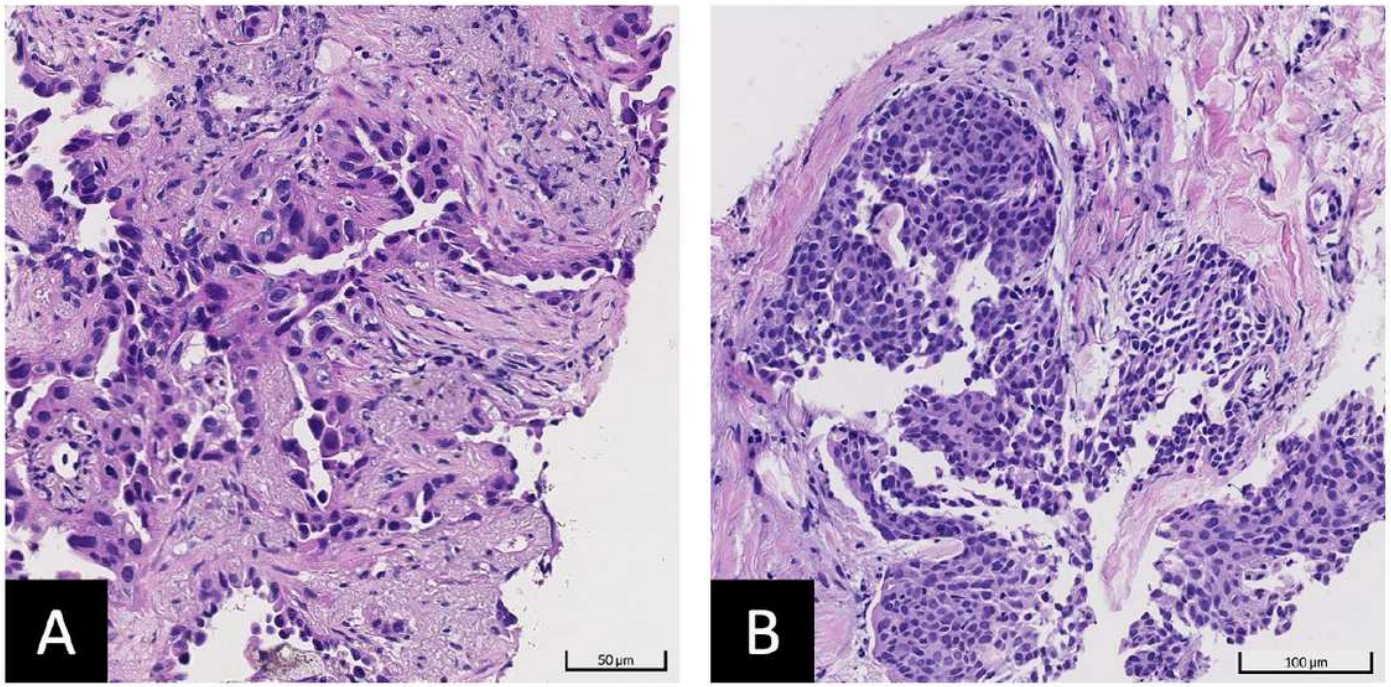

Figure 1

Examples of pathological images (A) adenocarcinoma and (B) squamous cell carcinoma.

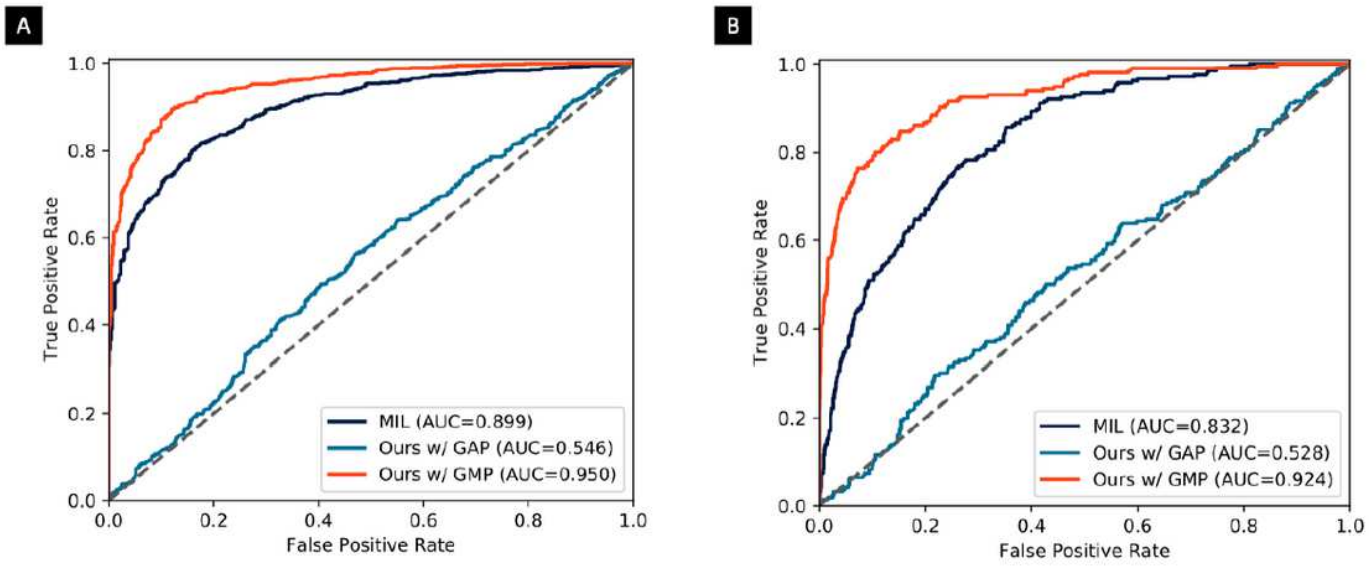

Figure 2

Receiver operating characteristic (ROC) curves of different training methods on classifying (A) adenocarcinoma and (B) squamous cell carcinoma on the testing dataset (n=1397). AUCs for classifying

adenocarcinoma are 0.899 (0.883-0.914) by MIL, 0.546 (0.516-0.576) by our method with GAP and 0.950 (0.939-0.960) by our method with GMP. AUCs for classifying squamous cell carcinoma are 0.832 (0.804-0.858) by MIL, 0.528 (0.486-0.571) by our method with GAP and 0.924 (0.904-0.943) by our method with GMP.

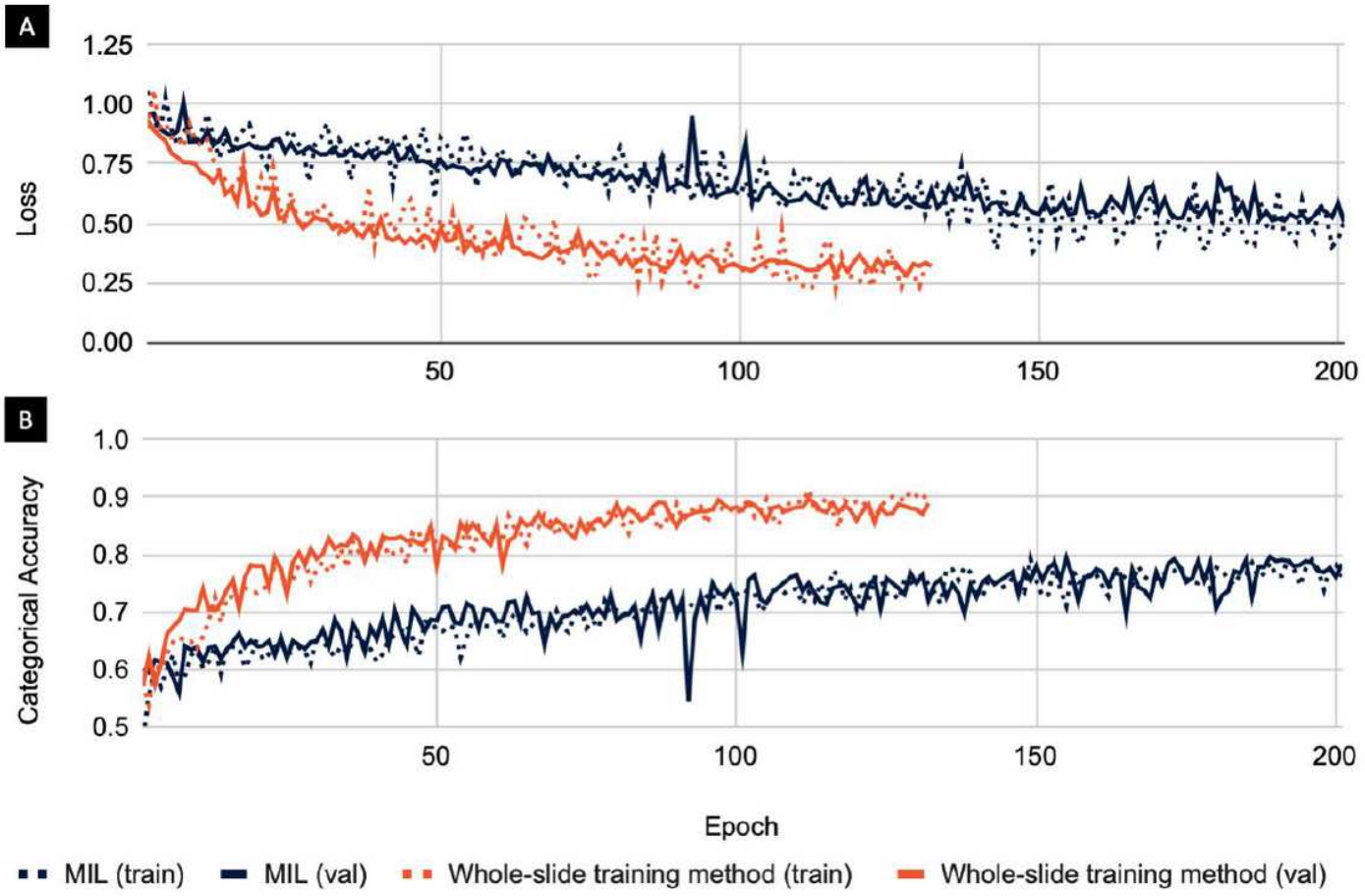

Figure 3

Learning curves of MIL and whole-slide training method. The dotted lines represent (A) the loss and (B) accuracy on the training set. Otherwise, the solid lines represent those on the validation set.

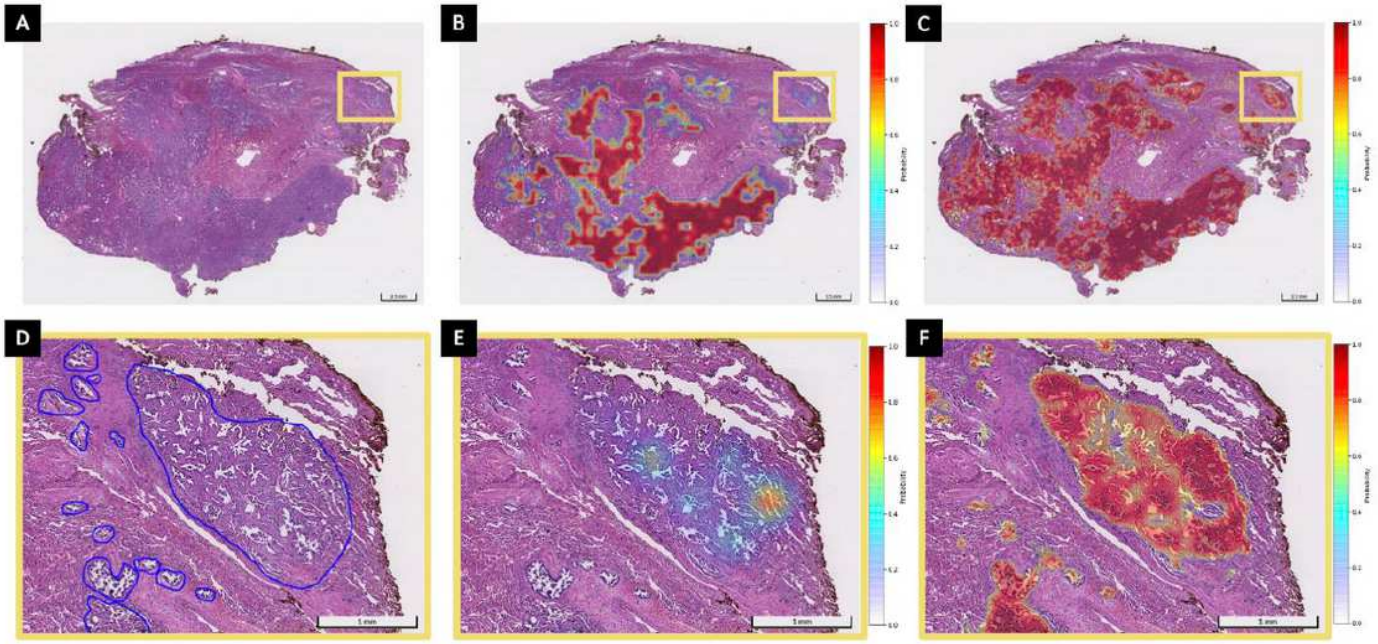

Figure 4

Visualization of heatmaps generated by MIL and whole-slide training method. (A) Whole-slide view on a slide containing adenocarcinoma lesions. (B) MIL heatmap on whole-slide view. (C) Heatmap generated by whole-image method. (D) Zoom-in view with reference human annotation. (E) Zoom-in view of MIL heatmap. (F) Zoom-in view of heatmap generated by whole-slide training method.

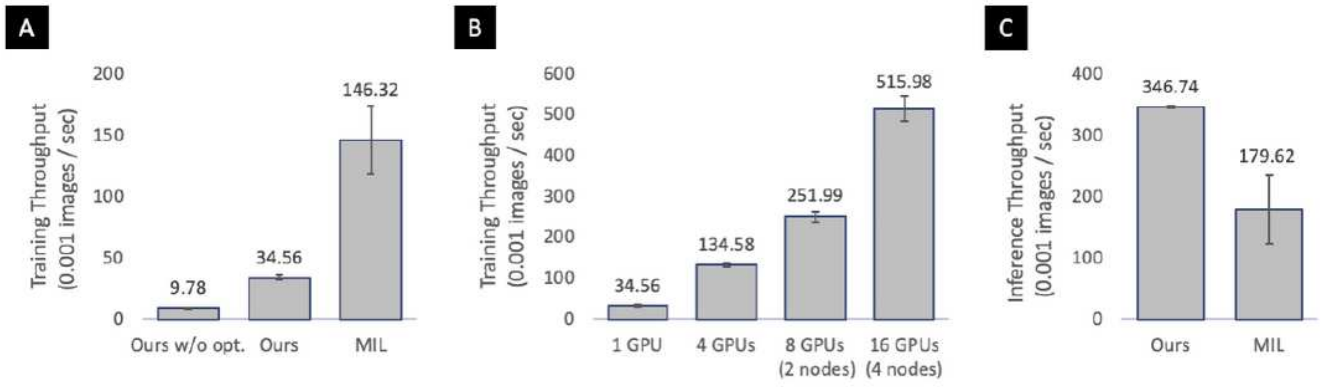

Figure 5

Throughput results of (A) whole-slide training method (ours) against non-optimized version and MIL, (B) distributed whole-slide training on multi-node, multi-GPU, and (C) inference on whole-slide training method and MIL.

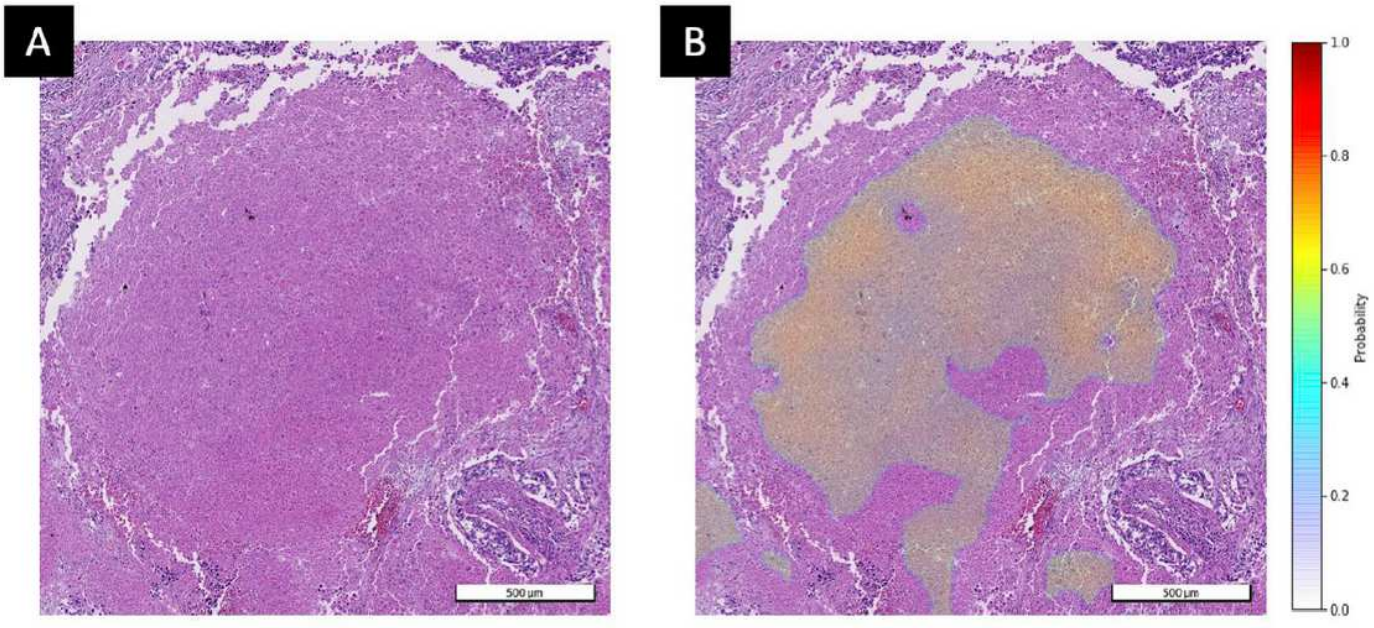

Figure 6

Tissue necrosis highlighted by CAM.

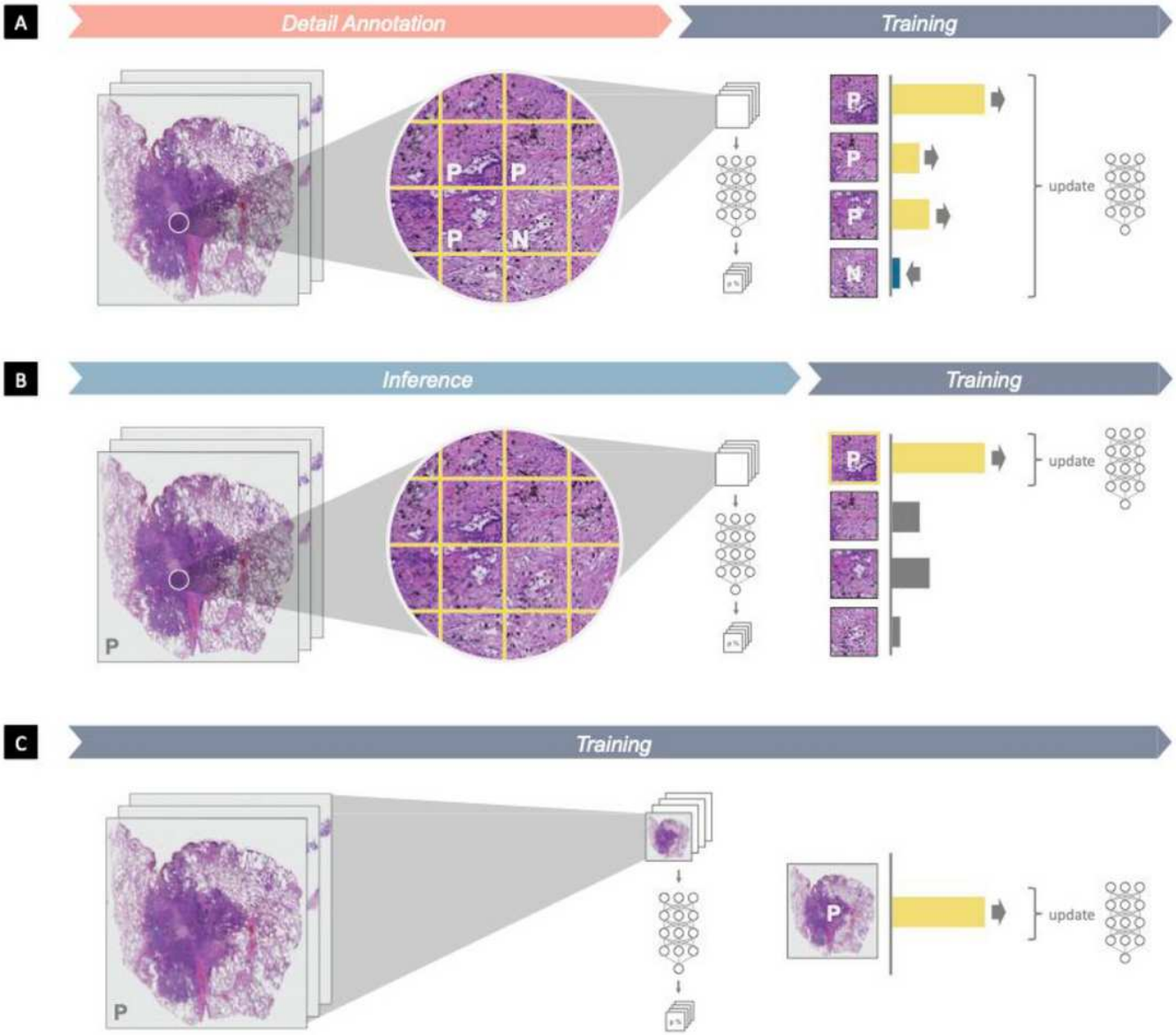

Figure 7

Workflows of (A) patch-based method (B) MIL method and (C) our proposed whole-slide training method.

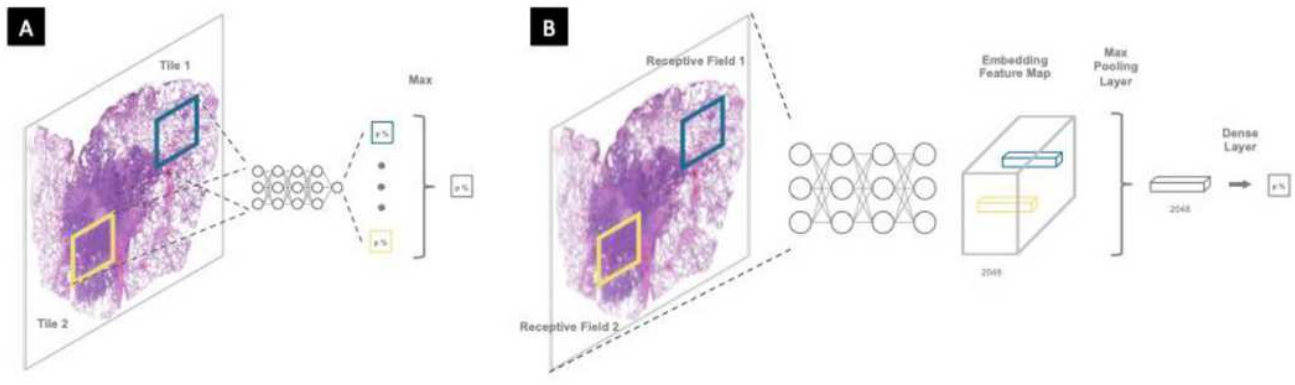

Figure 8

(A) Illustration of the mappings from input tiles to output predictions in MIL training. (B) Illustration of the mapping from receptive fields on input images to embedding feature maps using GMP in whole-slide training method.