LEARNING DEEP PATHOLOGICAL FEATURES FOR WSI-LEVEL CERVICAL CANCER GRADING

Ruixiang Geng[⋆†] *Qing Liu*[⋆†] *Shuo Feng*[⋆] *Yixiong Liang* [⋆‡]

[⋆]School of Computer Science, Central South University, Changsha 410083, P.R. China

ABSTRACT

Fully automated cervical cancer grading on the level of Whole Slide Images (WSI) is a challenge task. As WSIs are in gigapixel resolution, it is impossible to train a deep classification neural network with the entire WSIs as inputs. To bypass this problem, we propose a two-stage learning framework. In detail, we propose to first learn patch-level deep pathological features for smear patches via a patch-level feature learning module, which is trained via leveraging the cell instance detection task. Then, we propose to learn WSI-level pathological features from patch-level features for cervical cancer grading. We conduct extensive experiments on our private dataset and make comparisons with rule-based cervical cancer grading methods. Experimental results demonstrate that our proposed deep feature-based WSI-level cervical cancer grading method achieves state-of-the-art performance.

Index Terms— Cervical cancer screening/grading, deep pathological features, WSI analysis

1. INTRODUCTION

Cervical cancer is the fourth most common cause of cancer incidence and mortality among women [1], with approximately 604,000 new cases and 342,000 deaths worldwide in 2020 [2]. To increase the survival chance of patients, early diagnosis is essential and timely cytology-based cervical cancer screening is highly recommended. In clinical, cytopathologists manually find and analysis suspicious cells from WSIs of Pap smears to make diagnosis decision. However, manual identifying abnormal cells from hundreds of thousands cervical cells in gigapixel resolution WSIs is often tedious and labour-intensive. Thereby, developing automated analysis methods for WSIs is in extraordinary demand.

Motivated by the clinical practice that the presence of abnormal cells in different classes are always signs for cervical cancer screening/grading, many methods about cervical cell recognition in image patches cropped from WSIs have been proposed. Song *et al.* [3, 4, 5] follow a coarse-to-fine pipeline and first train a classifier to predict the labels of pixels in smear image patches, then employ post processes to further determine whether pixels with same predicted labels belong to the same cell component or instance. Instead, [6, 7, 8, 9, 10, 11] follow an end-to-end pipeline and update modern deep object detectors/segmentation approaches [12, 13, 14, 15] to directly detect/segment cell instances. Although those methods provide cell informations, massive efforts from cytopathologists are still required to gather predicted cell informations for further cervical cancer screening/grading, particularly when many normal cells are wrongly identified as abnormal cells.

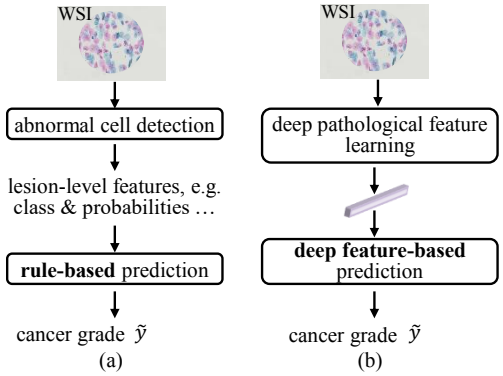


Fig. 1. Two paradigms of fully automated cervical cancer screening/grading on WSI-level. (a) Rule-based screening/grading with hand-crafted lesion-level features. (b) Our proposed deep pathological feature-based screening/grading.

More recently, fully automated methods [16, 17] for cervical cancer screening/grading on WSI-level are emerging. They almost follow the diagnosis procedures of cytopathologists and first detect abnormal cells from WSIs, then extract lesion-level features and aggregate them according to a serials of rules based on the criteria of Bethesda diagnostic system [18] for cancer grading, as illustrated in Fig. 1(a). Specifically, in [16], abnormal cell classes and probabilities by abnormal cell detector are directly treated as lesion-level features, and with them, diagnosis and cancer grade are predicted according to predefined rules from Bethesda diagnostic system [18]. Instead, in [17], hand-crafted statistic features about abnormal cells are extracted as lesion-level features, and with them, XGBoost [19] is trained to learn a serials of rules for cancer screening/grading. However, hand-crafted le-

†Equal contribution.
‡Corresponding author: yxliang@csu.edu.cn

sion features generally lack of flexibility and are with limited representation ability, thereby leading to suboptimal performance and limited generalisation of rule-based cancer screening/grading models learnt from them. What's worse, the performance of these two methods is highly dependent on abnormal cell detectors and mis-identification on abnormal cells especially in early stage of cervical cancer would be a disaster to them.

Today, learning deep features for AI relative problems particularly classification problems has been an almost universal consensus due to their powerful representation ability and discriminability. Thereby, learning deep pathological features from WSIs for cancer screening/grading as illustrated in Fig. 1(b) would be a better option. However, WSIs are in gigapixel resolution. For example, the resolution in our datasets ranges from 90,112×16,384 to 137,984×135,168 and it is impossible to directly feed them into CNNs for training and inference with current computing power. Thus, we propose a two-stage CNN-based architecture. In the first stage, the gigapixel WSI is partitioned into small patches size of 4096×2816 and patch-level deep pathological features are learned from those patches. In the second stage, we aggregate patch-level deep pathological features into WSI-level features and learn to predict the cervical cancer. Fig. 2 illustrates our proposed method.

In summarise, the contributions of this work are two-fold: (1) We design a two-stage learning framework to learn WSI-level deep pathological features for WSI-level cervical cancer screening/grading, which is able to learn deep pathological features on WSI-level from gigapixel WSIs; (2) We conduct extensive experiments on our private dataset and results show that our method achieves state-of-the-art performance on cervical cancer grading. Especially on specificity, our method shows a significant improvement compared with the rule-based method, which shows that our method is able to significantly reduce the misdiagnosis of healthy samples.

2. METHOD

Our goal is to learn deep pathological features on WSI-level from gigapixel WSIs for cervical cancer screening/grading. To bypass the problem that current computing power is far away to allow CNNs directly handle gigapixel WSIs for deep pathological feature learning, we develop a two-stage learning framework.

Method Overview As illustrated in Fig. 2, our learning framework consists of two stages, i.e., patch-level feature learning stage and WSI-level deep feature learning stage. In the first stage, gigapixel WSI $\mathbf{X}$ is partitioned into K patches, denoted as $\mathbf{X} = \{\mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_K\}$, then they are fed into the patch-level feature learning module ϕ_{patch}, resulting K groups of patch-level feature maps $\{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_K\}$. In the second stage, our proposed WSI-level feature learning module ϕ_{WSI} first aggregates $\{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_K\}$ and maps them

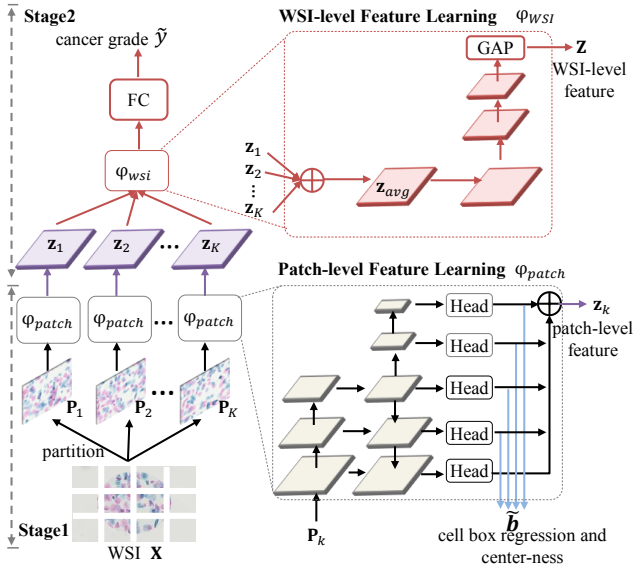


Fig. 2. Overview of our proposed method two-stage training framework for deep feature-based cervical cancer screening/grading on WSI-level. 'GAP' and 'FC' indicate Global Average Pooling and Full Connection layer respectively.

to the final WSI-level feature $\mathbf{Z}$. Thereafter, a full connection layer (FC) is attached and used to predict the cervical cancer grade $\widetilde{y}$ with $\mathbf{Z}$. For clarify, we summarise the processes in first stage as follows:

$$\mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_K = partition(\mathbf{X}) \; ; \tag{1}$$

$$\mathbf{z}_1 = \phi_{patch}(\mathbf{P}_1; \mathbf{W}_{patch}), \cdots, \mathbf{z}_K = \phi_{patch}(\mathbf{P}_K; \mathbf{W}_{patch}) \; ; \tag{2}$$

and the processes in second stage as follows:

$$\mathbf{Z} = \phi_{WSI}(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_K; \mathbf{W}_{WSI}) \; ; \tag{3}$$

$$\widetilde{y} = FC(\mathbf{Z}; \mathbf{W}_{cls}) \; . \tag{4}$$

$\mathbf{W}_{patch}$ in Eq. 2, $\mathbf{W}_{WSI}$ in Eq. 3, and $\mathbf{W}_{cls}$ in Eq. 4 are parameters to be learned.

Patch-level Feature Learning Module. This module is designed to learn pathological features from image patches. Unlike WSIs whose ground truth labels are assigned with cancer grades by cytopathologists, it is impossible to assign smear patches with ground truth cervical cancer grade to guide the training. To motivate this module learn pathological features, we propose to leverage a relative auxiliary task, i.e, cell detection, to learn pathological features about abnormal cells since ground truth cell instances are available. For efficiency, here we choose one-stage object detector FCOS [20] to detect cell instances in each patch. FCOS [20] consists of a backbone equipped with a three-level feature pyramid, on which two convolution blocks are stitched to generate higher level feature maps, generating five scales of feature maps. With these feature maps, five cell instance detection

heads with sharing parameters are performed to predict the bounding boxes of cell instances, the center-ness and cell classes. We denote the output branches bounding boxes regression and center-ness prediction as $\widetilde{\mathbf{b}}$ while upscale the five groups of output maps by cell classification branch to same resolution to bottom ones, then employ an element-wise summation over them and obtain the patch-level deep pathological feature maps.

WSI-level Feature Learning Module. This module takes $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_K$ as input and first aggregates them via an element-wise average operator. We denote the output as $\mathbf{z}_{avg}$. Then we modify ResNet34 [21], a simple yet effective deep classification network, to learn WSI-level features from $\mathbf{z}_{avg}$. Specifically, we modify the kernel size of the first convolutional layer in ResNet34 [21] from $3 \times 7 \times 7 \times 64$ to $C \times 7 \times 7 \times 64$, where C is the channel number of $\mathbf{z}_{avg}$, i.e., the total class number of cell. In our experiments, C is set to ten. Feeding the modified ResNet34 [21] with $\mathbf{z}_{avg}$, we obtain the WSI-level features $\mathbf{Z}$. Finally, we attach a full connection (FC) layer on top of the modified ResNet34 [21] to predict the cervical cancer grade $\widetilde{y}$.

Learning. The training of our WSI-level cervical cancer grading method has two stages. In the first stage, we train FCOS [20], i.e the cell instance detector, with dataset consisting of patches in regular sizes with both normal and abnormal cell instance annotations. The training loss is same with FCOS [20] and only parameters $\mathbf{W}_{patch}$ in Eq. 2 and the detection heads are updated. The second stage involves the modified ResNet34 [21] for WSI-level feature learning and the full connection layer for cervical cancer grade classification. We jointly train them with a dataset consisting of gigapixel WSIs with ground truth cancer grade labels. In this stage, cross-entropy loss is adopted and only parameters $\mathbf{W}_{WSI}$ and $\mathbf{W}_{cls}$ are updated while $\mathbf{W}_{patch}$ are fixed.

3. EXPERIMENTS

In this section, we conduct extensive experiments and make comparisons with stage-of-the-art methods to validate the effectiveness of our two-stage cervical cancer grading method.

Datasets. Our proposed cancer grading model are trained with two private datasets. The first one is the Cervical Cell DataSet (CCD) that includes 50000 smear patches of size 4096×2816 with ten-class cell instance annotations by pathologists, among which 30000, 10000 and 10000 for training, validation and testing, respectively. The second one is Cervical Smear Dataset (CSD), which contains 2625 WSIs and the diagnosis results by pathologists. Totally, according to Bethesda diagnostic system [18], there are six sub-categories of cervical cancer: NILM, ASC-US, LSIL, ASC-H, HSIL and SCC. NILM is the normal type while the rest are abnormal types and ranked from mild to severe. The details about CSD are listed in Table 1. 70% of the data is used for training and the remaining 30% is used for testing. Since the amount of SCC is too small, we merge it with HSIL into one category during training. Thus, two tasks are defined. One is 2-category classification task which requires the WSI-level cervical cancer grading model predict whether the input WSI belongs to normal or abnormal. The other is 5-category classification task which requires the grading model predict sub-categories.

Table 1. Cervical Smear Dataset (CSD)

Smear Category		num	total
NILM	NILM	1542	1542
ABN	ASC-US	740	1083
	LSIL	274	
	ASC-H	34	
	HSIL&SCC	35	

Table 2. Comparisons with the state-of-arts on 2-category classification task.

method	sens	spec	prec	sens.C	sens.H
Adaboost[22]	73.77	87.66	80.74	82.93	87.50
RandForest[23]	65.43	**91.78**	84.80	75.61	83.33
SVM[24]	79.94	90.04	84.92	89.02	91.67
Lin-0.90[16]	92.28	30.74	48.30	95.92	95.83
Lin-0.95[16]	95.68	27.71	48.14	97.96	96.30
Ours-0.90	94.44	90.91	**87.93**	96.34	95.83
Ours-0.95	**97.84**	85.50	82.55	**98.78**	**100**

Table 3. Comparisons with the state-of-arts on 5-category classification task.

method	accuracy
Adaboost[22]	66.75
Random Forest[23]	70.70
SVM [24]	68.54
Lin [16]	29.64
Ours	**79.74**

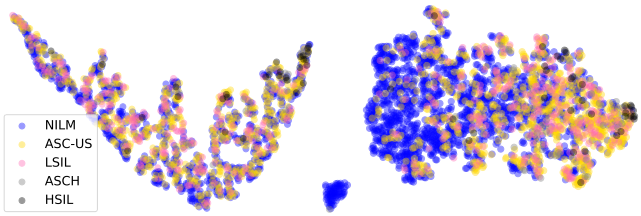
Evaluation Metrics. For 2-category classification task, we follow [16] and adopt sensitivity (sens), specificity (spec), sens.C (the sensitivity exclude ASC-US and ASC-H), sens.H (the sensitivity of high risk types including ASC-H, HSIL and SCC) and AUC as evaluation metrics. Additionally, we also report the precision (prec). For 5-category classification task, we adopt accuracy as evaluation metric.

Implementation Details. In the first training stage, the learning rate, batchsize, momentum and weight decay are set to 0.01, 16, 0.9 and 0.0001, and SGD is used as the optimiser. In second stage, learning rate, batchsize, momentum and weight decay are set to 0.002, 128, 0.9 and 0.0001. All models are trained on GeForce RTX 2080Ti with 12G GPU.

Comparisons with the State-of-arts. We compare our method with Lin *et al.* [16] and three traditional classification methods. The results by [16] are obtained by our own reproduction and we try our best to follow the experimental pro-

Table 4. Experimental results of different feature fusion strategies

strategy	2-category task			5-category task
	sens	spec	prec	accuracy
element-wise average	**97.84**	85.50	82.55	**79.74**
maximum	86.77	49.13	54.55	68.45
element-wise average w sigmoid	95.38	86.36	83.11	76.94



(a) Rely on Manual Rules (b) Our method

Fig. 3. Visualization of the feature information extracted by different methods. (a) is the method in [16], (b) is our method.

cess and make the production be in accordance with [16]. For traditional classification methods, we make comparisons with SVM [24], Random Forest [23] and Adaboost [22]. We train the traditional classifiers with our WSI-level deep pathological features and compare them with ours. The experimental results are shown in Table 2 and Table 3.

Table 2 shows the results on 2-category task. We follow [16] and set 0.90 and 0.95 as two constraints of "minimum-required" sensitivity. When considering all the five abnormal classes, our method achieves the highest sensitivity 97.84%. For sens.C and sens.H, our method further improves to 98.78% and 100% respectively. We also find that our method is able to maintain high score in specificity and precision when we get the highest sensitivity. In Fig. 4, the AUC score 98.25% could further validate the power of our method. On 5-category task, our accuracy is also the highest which is 79.74%.

Ablation Study. To validate the effectiveness of element-wise average strategy to aggregate patch-level deep pathological features, we also provide the results of two other aggregation strategies: maximum and element-wise average followed by a sigmoid function. Table 4 reports the results and shows that element-wise average aggravation achieves best both in sensitivity on 2-category classification task and accuracy on the 5-category classification task.

A More In-depth Analysis. For a more in-depth analysis about our method and the rule-base method [16], we visualise their features via t-SNE[25] respectively. Fig. 3 shows the visualisation results. From Fig. 3(a), we can observe that the visualised features of ASC-US and NILM are very close and less of distinction, which may be difficult to distinguish them. This is also the possible reason that rule-base method [16] performs terrible on specificity, precision and accuracy

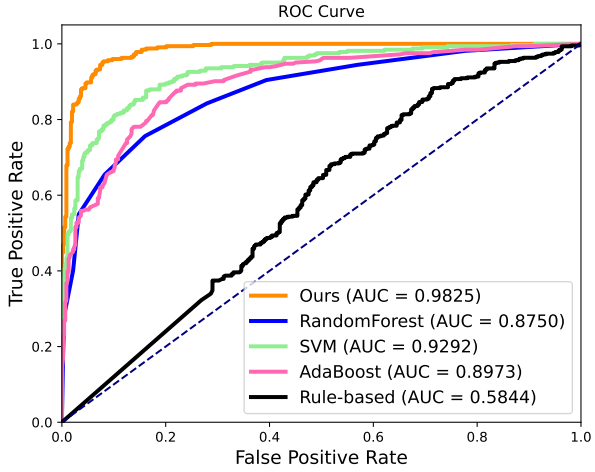


Fig. 4. The ROC curves of different methods.

on normal category NILM as reported in Table 2 and 3. As shown in Fig. 3 (b), our method has strong ability to distinguish between NILM and abnormal categories.

4. CONCLUSION

In this paper, we propose a two-stage learning framework to learn deep pathological features from WSIs for fully automated cervical cancer grading. In the first stage, we turn to the cell instance detection task to train the patch-level feature learning module. In the second stage, we jointly learn WSI-level features and cancer grade classifier. The experimental results prove that our method achieves the best results on both 2-category and 5-category cervical classification tasks.

5. ACKNOWLEDGE

This work is supported in part by the High Performance Computing Center of Central South University. Qing Liu is partially supported by the National Natural Science Foundation of China (NSFC No. 62006249), Changsha Municipal Natural Science Foundation (kq2014135) and Hunan Provincial Natural Science Foundation of China (2021JJ40788). Yixiong Liang is partially supported by the Hunan Provincial Science and Technology Innovation Leading Plan (No.2020GK2019).

6. REFERENCES

[1] World Health Organization, "Cervical cancer," . 1

[2] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, et al., "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021. 1

[3] Youyi Song, Ling Zhang, Siping Chen, et al., "Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning," *TBME*, vol. 62, no. 10, pp. 2421–2433, 2015. 1

[4] Youyi Song, Ee-Leng Tan, Xudong Jiang, et al., "Accurate cervical cell segmentation from overlapping clumps in pap smear images," *TMI*, vol. 36, no. 1, pp. 288–300, 2016. 1

[5] Youyi Song, Lei Zhu, Jing Qin, et al., "Segmentation of overlapping cytoplasm in cervical smear images via adaptive shape priors extracted from contour fragments," *TMI*, vol. 38, no. 12, pp. 2849–2862, 2019. 1

[6] Yanning Zhou, Hao Chen, Jiaqi Xu, et al., "IRNet: Instance relation network for overlapping cervical cell segmentation," in *MICCAI*. Springer, 2019, pp. 640–648. 1

[7] Changzheng Zhang, Dong Liu, Lanjun Wang, et al., "DCCL: A benchmark for cervical cytology analysis," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2019, pp. 63–72. 1

[8] Yao Xiang, Wanxin Sun, Changli Pan, et al., "A novel automation-assisted cervical cancer reading method based on convolutional neural network," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 2, pp. 611–623, 2020. 1

[9] Lei Cao, Jinying Yang, et al., "A novel attention-guided convolutional network for the detection of abnormal cervical cells in cervical cancer screening," *MedIA*, vol. 73, pp. 102197, 2021. 1

[10] Yixiong Liang, Changli Pan, Wanxin Sun, et al., "Global context-aware cervical cell detection with soft scale anchor matching," *Computer Methods and Programs in Biomedicine*, vol. 204, pp. 106061, 2021. 1

[11] Yixiong Liang, Zhihong Tang, Meng Yan, et al., "Comparison detector for cervical cell/clumps detection in the limited data scenario," *Neurocomputing*, vol. 437, pp. 195–205, 2021. 1

[12] Shaoqing Ren, Kaiming He, Ross Girshick, et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," *NIPS*, vol. 28, pp. 91–99, 2015. 1

[13] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, et al., "Feature pyramid networks for object detection," in *CVPR*, 2017, pp. 2117–2125. 1

[14] TY Lin, P Goyal, R Girshick, et al., "Focal loss for dense object detection.," *TPAMI*, vol. 42, no. 2, pp. 318–327, 2018. 1

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, et al., "Mask R-CNN," in *ICCV*, 2017, pp. 2961–2969. 1

[16] Huangjing Lin, Hao Chen, Xi Wang, et al., "Dual-path network with synergistic grouping loss and evidence driven risk stratification for whole slide cervical image analysis," *MedIA*, vol. 69, pp. 101955, 2021. 1, 2, 3, 3, 3

[17] Xiaohui Zhu, Xiaoming Li, Kokhaur Ong, et al., "Hybrid ai-assistive diagnostic model permits rapid tbs classification of cervical liquid-based thin-layer cell smears," *Nature Communications*, vol. 12, no. 1, pp. 1–12, 2021. 1

[18] Ritu Nayar and David C Wilbur, *The Bethesda system for reporting cervical cytology: Definitions, criteria, and explanatory notes*, Springer, 2015. 1, 3

[19] Tianqi Chen and Carlos Guestrin, "XGBoost: A scalable tree boosting system," in *SIGKDD*, 2016, pp. 785–794. 1

[20] Zhi Tian, Chunhua Shen, Hao Chen, et al., "FCOS: Fully convolutional one-stage object detection," in *ICCV*, 2019, pp. 9627–9636. 2

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al., "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778. 2

[22] Yoav Freund and Robert E Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997. 2, 3

[23] Andy Liaw, Matthew Wiener, et al., "Classification and regression by randomForest," *R news*, vol. 2, no. 3, pp. 18–22, 2002. 2, 3

[24] Johan AK Suykens and Joos Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999. 2, 3

[25] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *JMLR*, vol. 9, no. 11, 2008. 3