Robust Whole Slide Image Analysis for Cervical Cancer Screening Using Deep Learning

Shenghua Cheng
Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology
Sibo Liu
Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology
Jingya Yu
Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology
Gong Rao
Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology
Yuwei Xiao
Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology
Wei Han
Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology
Wenjie Zhu
Tongji Medical College, Huazhong University of Science and Technology
Xiaohua Lv
Huazhong University of Science and Technology-Wuhan National Lab for Optoelectronics
Ning Li
Collaborative Innovation Center for Biomedical Engineering, Wuhan National Laboratory for Optoelectronics-Huazhong University of Science and Technology
Jing Cai
Tongji Medical College, Huazhong University of Science and Technology
Zehua Wang
Tongji Medical College, Huazhong University of Science and Technology https://orcid.org/0000-0001-6423-8219
Xi Feng
Tongji Medical College, Huazhong University of Science and Technology
Fei Yang
Tongji Medical College, Huazhong University of Science and Technology
Xiebo Geng
Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology
Jiabo Ma
Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology
Xu Li

Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology

Ziquan Wei

Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology

https://orcid.org/0000-0001-6553-4482

Xueying Zhang

Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology

Shaoqun Zeng

Huazhong University of Science & Technology, Britton Chance Center for Biomedical Photonics, Wuhan National Laboratory for Optoelectronics-Huazhong University of Science and Technology

Li Chen

Tongji Medical College, Huazhong University of Science and Technology

Junbo Hu

Tongji Medical College, Huazhong University of Science and Technology

Xiuli Liu (xlliu@mail.hust.edu.cn)

Wuhan National Laboratory for Optoelectronics, Huazhong University of Science & Technology

---

Article

Keywords: Cervical Cancer, Deep Learning, Aided Diagnosis, Whole Slide Image

Posted Date: April 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-377187/v1

Version of Record: A version of this preprint was published at Nature Communications on September 24th, 2021. See the published version at https://doi.org/10.1038/s41467-021-25296-x.

Abstract

Computer-assisted diagnosis is key for popularizing cervical cancer screening. However, current recognition algorithms are insufficient in accuracy and generalization for cervical lesion cells, especially when facing diversity data in clinical applications. Inspired by manual reading slide under microscopes, we develop a progressive lesion cell recognition method combing low and high resolutions WSIs to recommend lesion cells and a recurrent neural network-based WSI classification model to evaluate the lesion degree of WSIs. After validating our system on 3,545 patient-wise WSIs with 79,218 annotations from multiple hospitals and several imaging instruments, on multi-center independent test sets of 1,170 patient-wise WSIs, we achieve 93.5% *Specificity* and 95.1% *Sensitivity* for classifying slides, closely equivalent to the average level of three independent cytopathologists, and obtain 88.5% TPR (true positive rate) for recommending top 10 lesion cells on 447 positive slides. After deploying, our system recognizes one giga-pixel WSI in about 1.5 minutes using one Nvidia 1080Ti GPU.

Introduction

Cervical cancer is one of the most common cancers in women. In 2020, there were about 604,127 women diagnosed with cervical cancer worldwide, and 341,831 died of the disease[1]. Many studies show that periodic inspection can reduce the incidence and mortality of cervical cancer[2–5]. Traditional smear test requires doctors to read the slides under microscopes, which is very labor-intensive, and the manual screening is difficult to achieve uniform and objective interpretation. With the progresses of digital whole slide image scanning instruments[6] and computer image processing technologies[7], a lot of automated lesion cell recognition methods[8] are developed and bring hope for accurate and efficient computer-aided cervical cancer screening.

Traditional methods, mainly based on morphological and textural characteristics[9], usually consist of image segmentation, feature extraction and cell classification. The image segmentation is used to segment the nucleus or cytoplasm through image histogram threshold, optical density measuring and image gradient[10–13]. The feature extraction primarily focuses on the shape features and textural information of nuclei. Cervical cells are then classified by random forest, support vector machine and artificial neural network, etc[14–15]. The performance of such methods is highly dependent on the segmentation effect and feature engineering. Subject to the principles, traditional methods have limited accuracy and generalization for diverse cytology slides derived from staining and imaging. In order to solve this problem, some commercial systems such as BD FocalPoint Slide Profiler[16] and Hologic ThinPrep Imaging System[17] adopted a closed-loop strategy that integrates slide preparation, staining, imaging and recognizing to ensure the accuracy and stability of the systems. In fact, they circumvented the generalization problem without solving it, which limits the wide use of the products especially in impoverished areas.

With the development of deep learning[18], convolutional networks (CNNs) have been applied to the identification of cervical lesion cells. Some studies have shown CNNs improve the effect of nucleus segmentation [19–20], and others utilize image classification and object detection CNNs to directly identify lesion cells without traditional segmentation process[21–24]. Compared with traditional methods, CNN-based methods learn feature representations automatically and have better generalization potential, but require a large number of annotated datasets. The image volume and annotation number of existing public datasets, mainly including Herlev[25], ISBI14[26], ISBI15[27] and CERVIX93[28], are small and most of them are provided with image tiles instead of WSIs, and hindering the acquisition of robust deep learning recognition algorithms and WSI-level diagnosis analysis. In addition, inference speed of CNNs on giga-pixel WSIs is challenging. Consequently, it is still difficult to apply current CNN-based methods in clinical cervical cancer screening scenes.

To address the above challenges in terms of accuracy, generalization and speed, here we propose a clinical-level aided diagnosis system for cervical cancer screening based on deep learning and massive WSIs. In the cytopathologists' diagnosis process, usually they scan the slides under a low-power microscope to find suspicious cells, and then further confirm them under a high-power microscope. Inspired by the strategy and give consideration to the accuracy and speed, we design a progressive recognition method combing the low and high resolution WSIs. First, a CNN screens WSIs at low resolution (LR) to quickly locate the suspicious areas, and then these areas are further identified at high resolution (HR) by another CNN. Finally, the system recommends the 10 most suspicious lesion areas in each slide for further reviewing by cytopathologists. Besides recommending suspicious lesion areas, our system also evaluate the lesion degree of WSIs and give a probability through developing a recurrent neural network (RNN) based WSI classification model. The CNN image features of the top 10 areas are extracted and input to the RNN model to get the positive probability of WSIs. We integrate designed data augmentation, diverse data learning and hard sample mining to achieve high accuracy and good generalization of our system. We train and validate our system on patient-wise 3,545 WSIs and 79,218 annotations from 5 hospitals and 5 kinds of scanners. On multi-center independent test sets of 1170 patient-wise WSIs, we achieve 93.5% *Specificity* and 95.1% *Sensitivity* for classifying slides. For the most confusing 121 WSIs of them, we achieve 50.0% *Specificity* and 74.6% *Sensitivity*, closely equivalent to the average level of three independent cytopathologists. The recommended top 10 lesion cells on 447 positive slides have an average TPR of 88.5%. Compared with current Hologic ThinPrep Imaging System, our system has a higher TPR of recommended cells and is more robust to staining and imaging style. When deploying the system, multi-threading and TensorRt[29] are used to accelerate image processing and forward inference. Our system recognizes one giga-pixel WSI in about 1.5 minutes using one Nvidia 1080Ti GPU. This speed ensures a good user experience in clinical applications and provides the possibility of real-time augmented reality under microscopes.

Results

System architecture. Our progressive recognition system consists of the LR model, HR model and WSI classification model, as shown in Fig. 1. The LR model is designed to quickly locate suspicious lesion areas at low resolution. The HR model is to identify the lesion cells and recommend the top 10 lesion cells at high resolution. The WSI classification model uses a RNN to integrate the CNN image features of the top 10 lesion cells, outputting the positive confidence of WSIs. The LR model and HR model are both based on ResNet50[30]. For the LR model, we modify the fully connected layer of original ResNet50 and add a semantic segmentation branch for generating a rough location mask (Supplementary Fig. 1). Thus, the LR model can screen WSIs and locate the suspicious lesion areas. The semantic segmentation branch is constructed with residual blocks of dilated convolutions. The LR model accepts an image tile of 512 × 512 pixels (0.486 μm/pixel) as input and outputs a lesion probability and a location heatmap (Supplementary Fig. 2). Afterwards, for the areas with a probability higher than 0.5 predicted by the LR model, we perform some morphological operations on corresponding location heatmap to generate the location mask. A cropped image tile of 256 × 256 (0.243 μm/pixel) according to the location mask is input to the HR model and a new lesion probability is obtained. Finally, all identified lesion cells in WSIs are sorted by lesion probabilities, and the top 10 most typical lesion cells are recommended for cytopathologist reviewing. Further, the RNN model integrates the CNN image features of the recommended top 10 lesion cells to classify WSIs. For each lesion cell tile, 2048-dimensional features are extracted by the HR model. Then the total 10 × 2048 dimensional features are input to the RNN model, and positive probabilities of WSIs are output.

Multi-center WSI datasets. To assess the robustness and clinical applicability of our system, we collected 12 groups of datasets from 5 hospitals and 5 kinds of imaging instruments (see Dataset sources in Methods), which are referred as groups A-L (Fig. 2a). These 12 datasets include 1,467 (41.4%) positive WSIS and 2,073 (58.6%) negative WSIS with 79,218 annotated lesion cells by a consensus of three cytopathologists. Each WSI representatives one unique patient. The 12 datasets show completely different image styles of staining and imaging characteristics (Fig. 2b) and we quantified the difference in their numerical distributions (Fig. 2c). Groups A-D are used for training our system. Groups E-L are treated as completely independent test set to evaluate the generalization of our system. Groups A-D are randomly divided into training set, validation set and test set with a slide-wise ratio of 8: 1: 1 (Fig. 2d). WSIs of all groups are scanned under 20× or 40× magnification microscopes. We uniformly interpolate them to 0.243 μm/pixel in data preprocessing, since the different resolutions of various imaging instruments.

In order to verify the effect of our recognition system in practical applications, we invited three cytopathologists to evaluate the prediction results of 1,170 slides in groups E and F. Groups E and F are independent from the training data with new styles and thus are suitable for clinical-level experiments. We performed the below assessments: slide level accuracy, tile level accuracy and true positive rate of recommended top 10 lesion cells.

Assessment at the slide level. To assess the effectiveness of our system at the slide level, we compared the RNN classifier and cytopathologists in classifying WSIs on the independent groups E-F of 1,170

slides. Figure 3a shows the ROC (receiver operating characteristic) curve of our system for classifying positive and negative slides, achieving 93.5% *Specificity* and 95.1% *Sensitivity* with 0.979 AUC (the area under ROC). The most confusing 121 slides of the slides were classified by the RNN and the three cytopathologists. Each red dot in Fig. 3b refers to *1-Specificity* and *Sensitivity* of a cytopathologist's interpretation result. Our system achieves 0.5 *Specificity* and 0.746 *Sensitivity* with 0.647 AUC, which is comparable with the average level of cytopathologists. In addition, our system processes one giga-pixel WSI in about 1.5 minutes after deploying on single GPU card, much faster than manual slide reading time.

Analysis of false positive and false negative slides. From the frequency histogram of slide classification scores of groups E-F (Fig. 4), our system produces 0.8% false negative slides (the slide score threshold value is 0.5), all of which were confirmed as ASCUS slides (atypical squamous cells of undetermined significance). As we know that cervical cytology ASCUS slides and part hard negative slides are confusable, thus it is acceptable to misjudge a small amount of ASCUS slides. Meanwhile, our system produces 26.3% false positive slides, which is in line with the original intention of cervical cytology computer-aided diagnosis. These false positives will be further reviewed by cytopathologists. Further, our system achieves 49.3% *Specificity* while retaining 100% *Sensitivity*, which indicates 49.3% negative slides can be excluded.

Assessment at the tile level. To evaluate the difference between our system and the three cytopathologists at the tile level, we randomly selected 1,000 positive tiles and 3,000 negative tiles with a size of 256 × 256 (0.243 μm/pixel) from groups E-F as the test data. As shown in Fig. 3c, the ROC curve describes the performance of our system, and each red dot represents *1-Specificity* and *Sensitivity* of a cytopathologist's classification result. Our system achieves 95.3% *Specificity* and 92.8% *Sensitivity* with 0.979 AUC, better than the average level of cytopathologists.

Assessment on the recommended top 10 lesion cells. The three cytopathologists evaluated the recommended top 10 lesion cells on 447 positive slides in groups E-F. The average true positive rates of top 10 and top 20 recommended cells are 88.5% and 85.0% separately (Fig. 3d). Moreover, our system does not miss any positive slide when voting the evaluated results of three cytopathologists, i.e., positive slides at least have one true lesion cell in the recommended top 10 or top 20 cells. Figure 4 shows the recommended cells of slides with different classification scores. For high-risk slides, our system recommended typical lesion cells such as koilocytotic cells or hyperchromatic cells with large nucleus and irregular nuclear membrane. For medium-risk slides, some suspicious cells with slightly large or deep-stained nucleus were recommended. No typical lesion cells were recommended on low-risk slides. The results demonstrate our system can accurately recommend 10 lesion cells without missing positive slides, greatly reducing the workload of cytopathologists to screen WSIs.

Comparison of our system and Hologic ThinPrep Imaging System on recommending top lesion cells. To further verify the effectiveness of our aided screening system, we compared it with Hologic ThinPrep Imaging System (referred as TIS). The test data are cervical cytology samples from 58 positive patients

in Maternal and Child Hospital of Hubei Province equipped with TIS. First, 58 glass slides were prepared from the 58 samples, stained, imaged and identified by TIS in the hospital. 22 suspicious fields of view were recommended by TIS on each slide. Then, we used another instrument (Shenzhen Shengqiang Technology Ltd. with 0.180 μm/pixel under 40× magnification) to scan the 58 glass slides, and we used our system to recommend the top 20 suspicious cell regions (about 60×60mm^2, far less than TIS' fields of view) for each slide. We asked the three cytopathologists to evaluate the results recommended by TIS and our system at the same time. The statistical results in Fig. 5 show that the true positive rate of our system is higher than that of TIS. Notably, TIS can only work under the closed-loop strategy of preparation, staining, imaging and recognition, while our system is robust to staining and imaging of various sources.

Importance of designed data enhancement, hard sample mining and diverse data learning. We conducted a set of ablation experiments to demonstrate the importance of designed data enhancement, hard sample mining and diverse data learning (see Methods). We used the three learning strategies to train a series of control high-resolution models step by step, and gave the classification accuracies on the test sets of groups A-F. Notably, the ratio of positive and negative tiles in the test set is 1:1. The ablation experimental model configs and results were provided in Fig. 6a. To evaluate model generalization, we treated groups E-F as the independent test data and showed the ROC curves of these control models on groups E-F in Fig. 6b. According to the results, with the designed data enhancement and hard sample mining, performance of the enhanced and mined models on groups E-F made great progress with AUC value increase of 0.138 and 0.072. The results indicate that our designed data enhancement and hard sample mining strategies are effective for improving model generalization and accuracy. Further, as more groups of training datasets were used, the AUC values of the mined, baseline and HR models increased gradually from 0.808 to 0.983. The results indicate that the diverse data learning of multiple groups with different styles is important for model generalization.

Generalized and rich feature representations of our models. We analyzed the alignment of features of high-resolution models between different groups of data by feature visualization. The dimension-reduced features of the original, enhanced, mined and baseline models by t-SNE[31] on groups A, B, E and F are shown in Fig. 7a. For the models, groups E-F are independent test data. From the original model to the baseline model, features of positive and negative tiles are gradually separated. Further, the features are gradually aligned between groups A-B and groups E-F. The results indicate that the designed data enhancement, hard sample mining and diverse data learning strategies improves the discrimination and alignment of features on unseen groups E-F.

We further analyzed the feature representations of the HR model on groups A-L in Fig. 7b. The tiles with high and low lesion probabilities from the 12 groups are clustered and well separated. Tiles corresponding to the far-right points are the typical lesion cells, including koilocytotic cells and hyperchromatic cells with large nucleus and irregular nuclear membrane. These lesion cells with different staining and imaging characteristics are clustered together and share similar features. Normal cells from different groups are clustered on the left points. At the junction regions are the suspicious cells with

about 0.5 lesion probabilities. The suspicious cells generally contain slightly large nucleus or deep-stained nucleus, but the degree is not enough. In addition, artifacts from staining and imaging may cause the suspicious cells. The results indicate that the learned features represent cervical lesion cells morphology well and the features are aligned between datasets with different staining and imaging characteristics. This is the key reason why our system has good generalization for unseen datasets of new styles.

Discussion

In this paper, we propose a clinical-level aided diagnosis system for cervical cancer screening based on deep learning and massive WSIs. We demonstrate the contributions of the low- and high- resolution combination strategy, the designed data augmentation, the diverse data learning and the hard sample mining for achieving high accuracy, good generalization and fast speed of our system. On multi-center test datasets of 1,170 WSIs with new styles, we achieve 93.5% *Specificity* and 95.1% *Sensitivity* for classifying WSIs, closely equivalent to cytopathologist interpretation level, and 88.5% TPR for recommending the top 10 lesion cells better than that of Hologic TIS. Further, about 49% negative slides can be excluded by our system while retaining 100% *Sensitivity*. The results prove that our system effectively improves the work efficiency of cytopathologists to screen WSIs. Therefore, for underdeveloped areas lacking cytopathologists, our system has important clinical and social significance for accelerating the popularization of cervical cancer screening.

The diversities of slide staining and imaging in different hospitals greatly limit the utility of current automated cervical cell recognition algorithms, thus model generalization is a key factor in the practicality. We design strategies of data augmentation and diverse data learning to obtain robust feature representations, ensuring well generalization and practical capabilities of our system. Our multi-center independent test datasets include differences in slide preparation (liquid-based preparation methods: membrane-based and sedimentation), dyeing schemes (fixing, clearing and dehydrating), imaging magnification (20× and 40×), imaging resolution (0.180–0.293 μm/pixel), imaging color characteristics (Fig. 2b&c), etc. The results on these diverse data prove the generalization of our system and lay a foundation for the practicability in diverse data scenarios.

Our system only distinguishes positive and negative classes instead of fine subclasses at both cell and slide levels, such as ASCUS, ASC-H (atypical squamous cells cannot exclude HSIL), LSIL (low-grade squamous intraepithelial lesion), HSIL (high-grade squamous intraepithelial lesion, SCC (squamous cell carcinoma), etc[32]. Though the subclasses are important in cervical screening, we don’t distinguish them based on the following considerations. Firstly, definition of the subclasses is based on cell morphology and the boundaries is often fuzzy especially at the cell level, which will produce a lot of noises in the actual manual annotations and inconsistency between different cytopathologists. Secondly, the key of the aided screening system is to improve the screening efficiency of cytopathologists by excluding low-risk slides and recommending a limited number of suspicious cells on high-risk slides for cytopathologist reviewing. The final precise diagnosis is up to cytopathology doctors, and this human machine

combination mode can reduce possible errors of artificial intelligence and ensure the accuracy of diagnosis.

In the future, we will focus on research about AI-enhanced portable microscopy and augmented reality microscopy to further expand our system. At present, professional but expensive scanners are still required, preventing the spread of cervical cancer screening in remote and underdeveloped areas. Thus, developing portable microscope-based cervical cancer computer aided diagnosis is necessary. In addition, developing real-time augmented reality microscope can provide friendly human-computer interaction for AI-assisted slide screening without changing the conventional working mode of cytopathologists.

Methods

Dataset sources. All 12 groups of glass slides are provided by Maternal and Child Hospital of Hubei Province (referred as H1), Tongji Hospital of Huazhong University of Science and Technology (referred as H2), Wuhan Union Hospital of Huazhong University of Science and Technology (referred as H3), Hubei Cancer Hospital (referred as H4) and KingMed Diagnostics Ltd. (referred as H5). The slide acquisition is performed in accordance with the guidelines of the Medical Ethics Committee of Tongji Medical College at Huazhong University of Science and Technology. These glass slides are scanned into WSIs by the instruments from 3DHisTech Ltd. with 0.243 mm/pixel under 20× magnification (referred as S1), Shenzhen Shengqiang Technology Ltd. with 0.180 mm/pixel under 40× magnification (referred as S2), Wuhan National Laboratory for Optoelectronics-Huazhong University of Science and Technology with 0.293 mm/pixel under 20× magnification (referred as S3), Huaiguang Intelligent Technology Ltd. with 0.238 mm/pixel under 20× magnification (referred as S4) and Konfoong Biotech Information Ltd. with 0.238 mm/pixel under 40× magnification (referred as S5). The details are shown in Table 1. Notably, the version 1 and version 2 of the scanner S1 are different generations of instruments, and they have different imaging color characteristics. The version 1, version 2 and version 3 of the hospital H2 are different in slide preparation and staining scheme.

Table 1. Dataset sources of the 12 groups of datasets.

Groups	Hospital	Scanner
A	H1	S1 - version 1
B	H1	S1 - version 2
C	H1	S3
D	H2 - version 1	S2
E	H1	S2
F	H2 - version 2	S2
G	H2 - version 1	S1 - version 2
H	H4	S4
I	H2 - version 3	S4
J	H3	S4
K	H5	S5
L	H1	S2

Data enhancement. Data enhancement is a common technique to expand the diversity of data distributions and improve the generalization ability of models. Based on the characteristics of cervical cell images, we designed a series of specific data enhancement, including transformations on hue, saturation, brightness, contrast, flipping and shifting, as well as adding noise such as blurring and sharpening. We determined the enhancement parameters to ensure that the recognition of cervical cells will not be affected.

Diverse data learning. Diverse data learning on different groups of data was employed to improve the robustness of feature representations. We first trained the LR model and HR model on groups A and B with a large number of annotations to obtain baseline models. The pre-training weights on ImageNet[33] were used as the initial weights. Then we incorporated groups C and D of different styles as extra training data. Based on the baseline models, we learned more robust models on mixed data of groups A-D.

Hard sample mining. Hard sample mining was adopted to improve the accuracy of our models. For training samples that fail to be successfully classified, we conducted a second round of learning on the hard samples. The proportion of hard samples in the entire training samples affected the tendency of models. The LR model of our recognition system was designed to find all possible lesion cells quickly and coarsely, while the HR model aimed to distinguish the true lesion cells from those candidates found by the LR model. Therefore, we used a small proportion of hard samples to train the LR model to ensure a high recall rate and a large proportion to train the HR model to ensure a high precision.

Training details. We employed the above three training strategies to enhance the generalization ability and recognition accuracy of the LR and HR models. According to the datasets of groups A-D in Figure 1d, we used the train set to optimize the LR and HR models, the validation set to adjust the hyper-parameters, and then tested the performance on the test set. Groups E-F were used as independent datasets for evaluating model generalization. The positive samples were cropped around the annotations of positive slides and the negative samples were randomly cropped from negative slides. The LR and HR models used Adam[34] as the optimizer with an initial learning rate of 10^{-3}. The strategy of learning rate decay was adopted during training.

We used a simple RNN with 1 hidden layer of 512 units for classifying WSIs. The RNN model was trained and validated on groups A-D and evaluated on independent groups E-F. We trained three kinds of RNNs with different inputs: the HR model features of the recommended top 10, top 20 and top 30 lesion cells. For each kind, we trained 2 RNNs, then integrated the total 6 RNNs as our WSI classifier. In the training process, we used data augmentation to improve the varieties of the input features of the RNNs, including enhancement and rearrangement of the top k lesion cell images. These strategies of model integration and data augmentation improved the robustness of the RNNs, since the limited number of slides (less than 10^4). Similarly, Adam[34] with an initial learning rate of 10^{-3} and the learning rate decay strategy were adopted during training the RNNs.

Declarations

Data availability

The statistical data supporting the conclusions of this paper were provided in the supplementary .zip files. The original WSI and annotation data are private and are not publicly available since the protection of patients' privacy in cooperative hospitals.

Code availability

The source codes of paper are available at https://github.com/ShenghuaCheng/Aided-Diagnosis-System-for-Cervical-Cancer-Screening.

Acknowledgements

We thank the members of Britton Chance Center for Biomedical Photonics for help in experiments and comments on the manuscript. This work is supported by the NSFC projects (grant 61721092) and the director fund of the WNLO.

Author contributions

X.L., J.H., L.C. and S.Z. conceived of the project. S.C., X.L., S.L. and J.Y. designed the aided diagnosis system. S.C., S.L. and J.Y. developed the algorithms. X.L., J.H., L.C., G.R., Y.X. and W.Z annotated the lesion cells, evaluated the experimental results, and completed the man-machine comparison

experiments. J.H., L.C., Z.W., J.C., X. F. and F. Y. provided the cervix cytopathology glass slides. X.L., G.R., Y.X, X.L. and N.L. scanned the slides into WSIs. W.H. performed the deployment and performance optimization of the system. S.C., S.L., J.Y., X.G., J.M., X.L., Z.W. and X.Z. performed the image analysis and processing. S.C., X.L., S.L. and J.Y. wrote the manuscript. All the authors revised the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary files were submitted together with the manuscript.

References

1. Ferlay, J. et al. Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. Available from: https://gco.iarc.fr/today (2020).
2. Peto, J., Gilham, C., Fletcher, O. & Matthew, F. E. The cervical cancer epidemic that screening has prevented in the UK. *The Lancet* 364, 249-256 (2004).
3. Sasieni, P., Adams, J. & Cuzick, J. Benefit of cervical screening at different ages: evidence from the UK audit of screening histories. *British Journal of Cancer* 89, 88-93 (2003).
4. Levi, F., Lucchini, F., Negri, E., Franceschi, S. & Vecchia, C. L. Cervical cancer mortality in young women in Europe: patterns and trends. *European Journal of Cancer* 36, 2266-2271 (2000).
5. Parkin, D. M., Nguyen-Dinh X. & Day N. E. The impact of screening on the incidence of cervical cancer in England and Wales. *British Journal of Obstetrics and Gynaecology* 92, 150-157 (1985).
6. Wright, A. M. et al. Digital slide imaging in cervicovaginal cytology: a pilot study. *Pathol. Lab. Med.* 137, 618-624 (2013).
7. Litjens, G. et al. A survey on deep learning in medical image analysis. *Medical Image Analysis* 42(9), 60-88 (2017).
8. Conceição, T., Braga, C., Rosado, L. & Vasconcelos, M. J. M. A review of computational methods for cervical cells segmentation and abnormality classification. *J. Mol. Sci.* 20, 5114 (2019).
9. Cahn, R.L., Poulsen, R. S. & Toussaint, G. Segmentation of cervical cell images. *Histochem. Cytochem.* 25, 681-688 (1977).
10. Borst, H., Abmayr, W. & Gais, P. A thresholding method for automatic cell image segmentation. *Histochem. Cytochem.* 27, 180-187 (1979).
11. Chang, C. W. et al. Automatic segmentation of abnormal cell nuclei from microscopic image analysis for cervical cancer screening. In *Proceedings of the IEEE 3rd International Conference on Nano/Molecular Medicine and Engineering* 77-80 (IEEE, 2009).
12. Kim, K. B., Song, D. H. & Woo, Y. W. Nucleus segmentation and recognition of uterine cervical pap-smears. *In International Workshop on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing;*

Lecture Notes in Computer Science 153-160 (Springer, 2007).

13. Chen, Y. et al. Semi-automatic segmentation and classification of pap smear cells. *IEEE J. Biomed. Health Inf.* 18, 94-108 (2014).
14. Mariarputham, E. J. et al. Nominated texture based cervical cancer classification. *Math. Methods Med.* 2015, 1-10 (2015).
15. Huang, P. C. et al. Quantitative assessment of pap smear cells by PC-based cytopathologic image analysis system and support vector machine. *In International Conference on Medical Biometrics; Lecture Notes in Computer Science* 192-199 (Springer 2007).
16. Renshaw, A. & Elsheikh, T. M. A validation study of the Focalpoint GS imaging system for gynecologic cytology screening. *Cancer Cytopathology* 121(12), 737-738 (2013).
17. Quddus, M., Neves, T., Reilly, M., Steinhoff, M. & Sung, C. Does the ThinPrep Imaging System increase the detection of high-risk HPV-positive ASC-US and AGUS? The Women and Infants Hospital experience with over 200,000 cervical cytology cases. *Cytojournal* 6(1), 15 (2009).
18. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* 521(7553), 436-444 (2015).
19. Zhang, L. et al. Combining fully convolutional networks and graph-based approach for automated segmentation of cervical cell nuclei. In *Proceedings of the IEEE 14th International Symposium on Biomedical Imaging* (IEEE 2017).
20. Song, Y. et al. A deep learning based framework for accurate segmentation of cervical cytoplasm and nuclei. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 2903-2906 (IEEE 2014).
21. Lin, H., et al. Dual-path network with synergistic grouping loss and evidence driven risk stratification for whole slide cervical image analysis. *Medical Image Analysis* 69, 101955 (2021).
22. Nirmal-Jith, U. et al. DeepCerv: deep neural network for segmentation free robust cervical cell classification. MICCAI, 2018.
23. Wieslander, H. et al. Deep convolutional neural networks for detecting cellular changes due to malignancy. IEEE International Conference on Computer Vision Workshop. IEEE (2018).
24. Lin, H., Hu, Y., Chen, S., Yao, J. & Zhang, L. Fine-grained classification of cervical cells using morphological and appearance based convolutional neural networks. *IEEE Access* 7, 71541-71549 (2019).
25. Jantzen, J., Norup, J., Dounias, G. & Beth, B. Pap-smear benchmark data for pattern classification. In *Inspir. Smart Inf. Syst.* 1-9 (NiSIS 2005).
26. Lu, Z., Carneiro, G. & Bradley, A. P. An improved joint optimization of multiple level set functions for the segmentation of overlapping cervical cells. *IEEE Trans. Image Process* 24, 1261-1272 (2015).
27. Lu, Z. et al. Evaluation of three algorithms for the segmentation of overlapping cervical cells. *IEEE J. Biomed. Health Inf.* 21, 441-450 (2017).
28. Phoulady, H. A. & Mouton, P. R. A new cervical cytology dataset for nucleus detection and image classification (Cervix93) and methods for cervical nucleus detection. Preprint at:

https://arxiv.org/abs/1811.09651 (2018).
29. Available from: https://developer.nvidia.com/zh-cn/tensorrt.
30. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 770-778 (IEEE, 2016).
31. Van der Maaten, L., & Hinton, G. Visualizing data using t-SNE. *Mach. Learn. Res.* 9, 2579-2605 (2008).
32. Nayar, R. & Wilbur, D.C. The Bethesda System for Reporting Cervical Cytology: Definitions, Criteria, and Explanatory Notes. (Springer, 2015).
33. Deng, J. et al. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 248-255 (IEEE, 2009).
34. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. Preprint at: https://arxiv.org/abs/1412.6980 (2015).

Figures

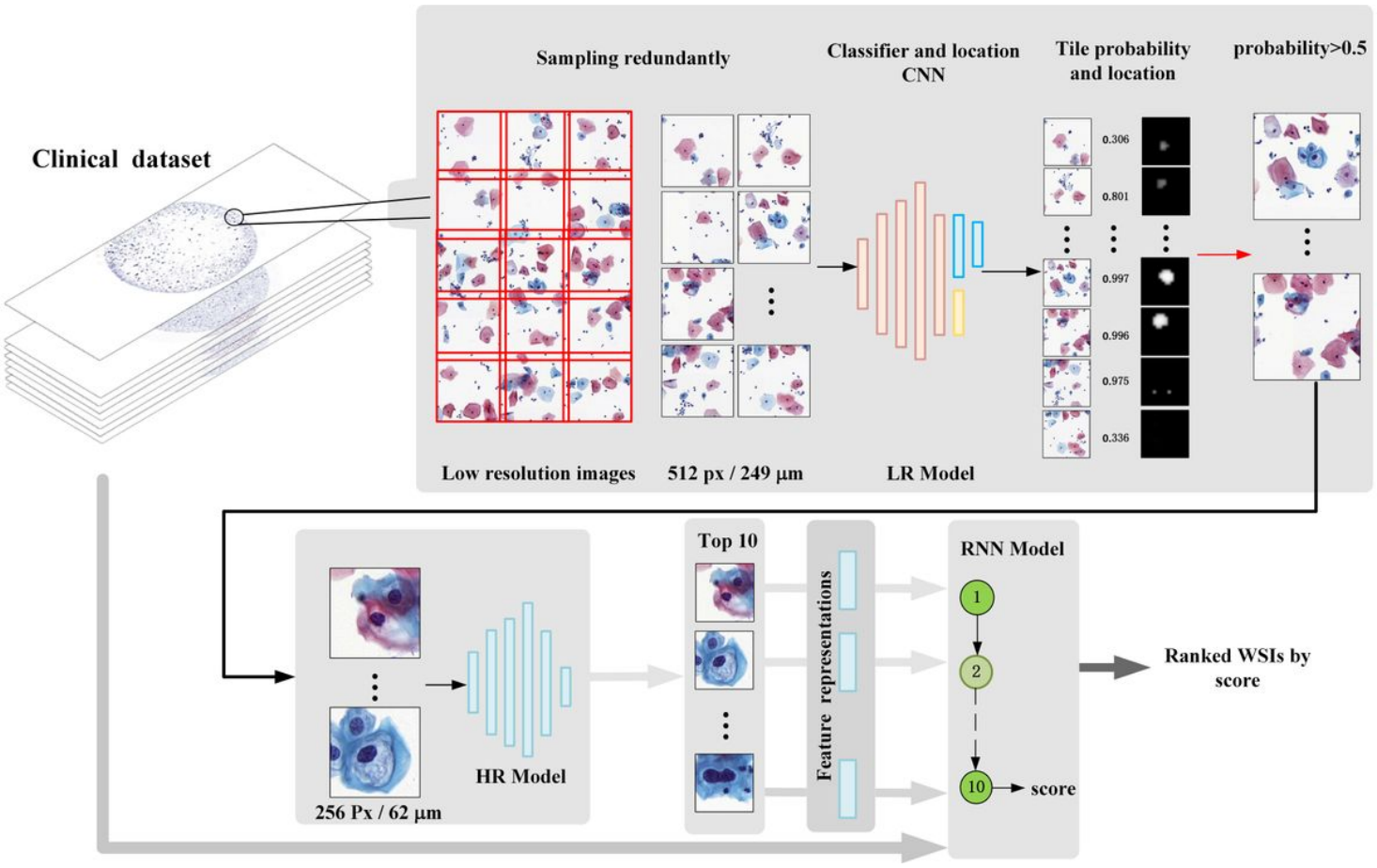


Figure 1

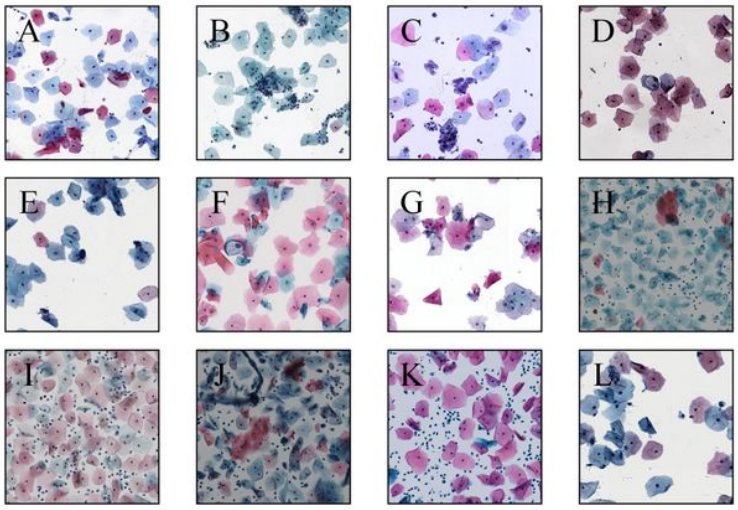
The proposed cervical cancer aided diagnosis system. Our system consists of WSI redundant division, LR model, HR model and RNN model. The LR model takes a divided image tile of 512 × 512 pixels (0.486

um/pixel) as input and outputs a lesion probability and a location heatmap to identify and locate the suspicious lesion areas on WSIs. The HR model takes an image tile of 256 × 256 (0.243 um/pixel) cropped according to the location heatmap as input and outputs a new lesion probability. The RNN model integrates the HR model image features of the top 10 lesion cells and outputs positive probabilities of WSIs.

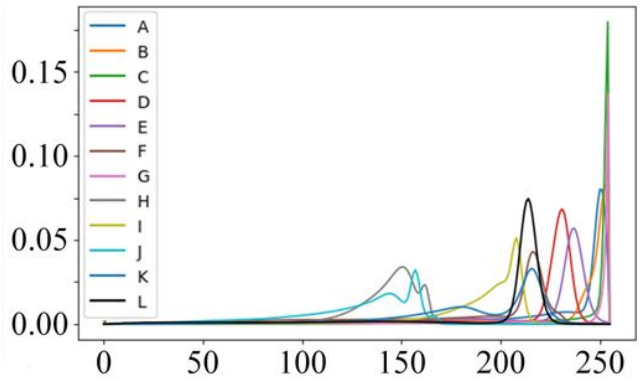
a

Groups	Positive slides	Negative slides	Labels
A	105	106	31,978
B	160	307	11,211
C	240	94	7,127
D	424	879	11,954
E	226	159	11,579
F	285	500	5,369
G-L	27	33	0
Total	1,467	2,078	79,218

b



c



d

	Positive slides			Negative slides		
	train	val	test	train	val	test
A	83/25,208	11/2,991	11/3,779	84	11	11
B	130/8,728	15/1,100	15/1,383	247	30	30
C	192/4,863	24/936	24/1,328	76	9	9
D	340/691	42/1,452	42/3,601	703	88	88
Total	745/45,700	92/6,479	92/10,091	1,110	138	138

Figure 2

Overview of multi-center WSI datasets. a, The collected 12 groups of WSI datasets from five hospitals and five kinds of scanners. There are total 3,545 WSIs with 79,218 annotated lesion cells by a consensus of three cytopathologists. b, Cervical image instances of the 12 groups, showing diverse image styles of staining and imaging characteristics. c, The numerical distributions of the 12 groups in the value channel of HSV space (hue-saturation-value), further confirming the difference of image styles. d, The division of training set, validation set and test set with a slide-wise ratio of 8: 1: 1 on groups A-D, which are used for training our system.

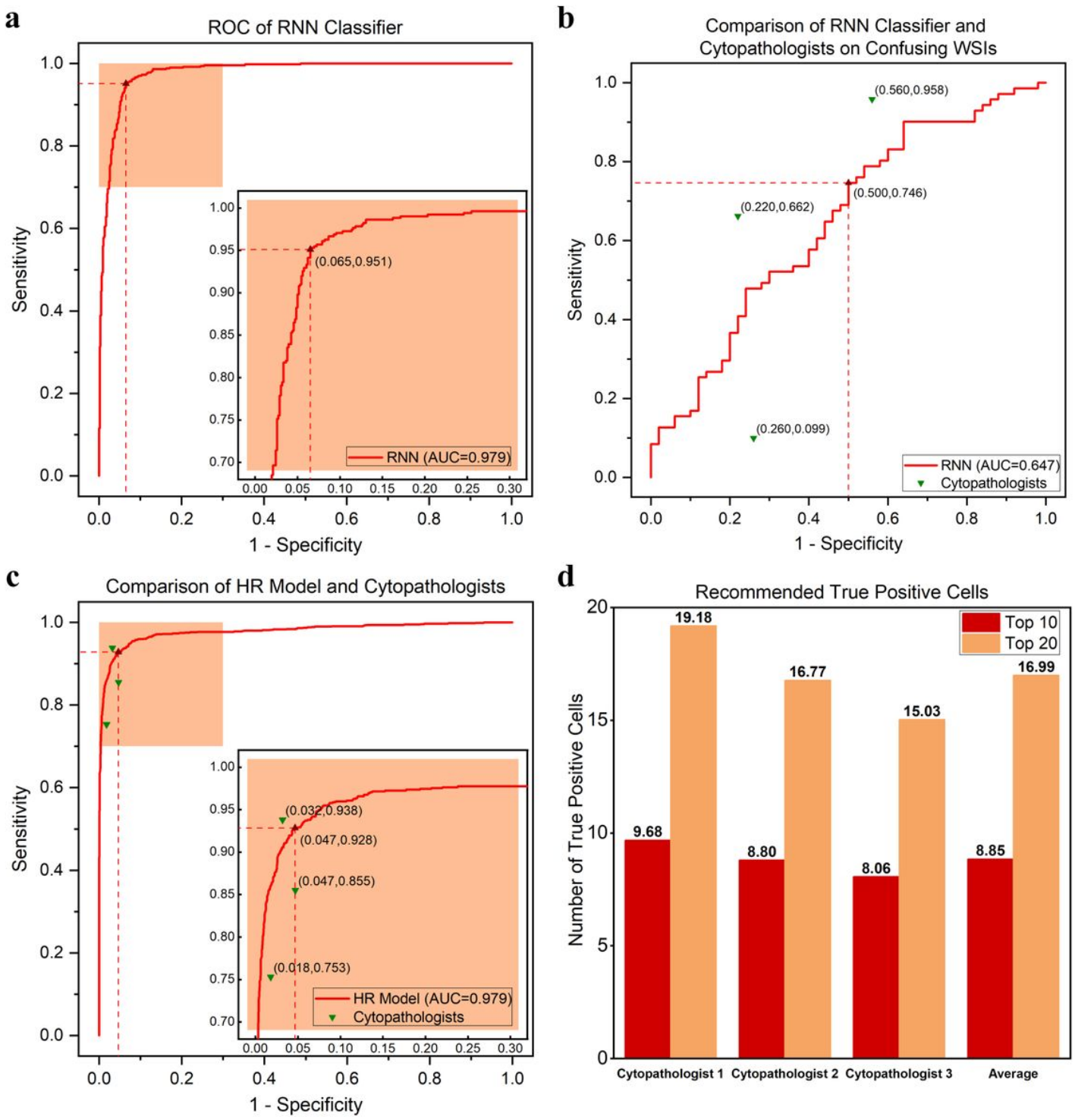


Figure 3

Clinical-level experimental results. a, The ROC of the RNN model for classifying the 1,170 slides of groups E-F. b, Comparison of the RNN model and three cytopathologists at the slide level. The most confusing 121 slides of the 1,170 slides were classified by the RNN model and the three cytopathologists. Each green triangle refers to 1-Specificity and Sensitivity of a cytopathologist's result. c, Comparison of the HR model and the three cytopathologists at the tile level. Randomly selected 1,000 positive test tiles and 3,000 negative test tiles with a size of 256 × 256 (0.243 um/pixel) from groups E-F were simultaneously

classified by the HR model and the three cytopathologists. d, The average true positive numbers of the recommended top 10 and top 20 lesion cells on 447 positive slides in groups E-F, evaluated by the three cytopathologists separately.

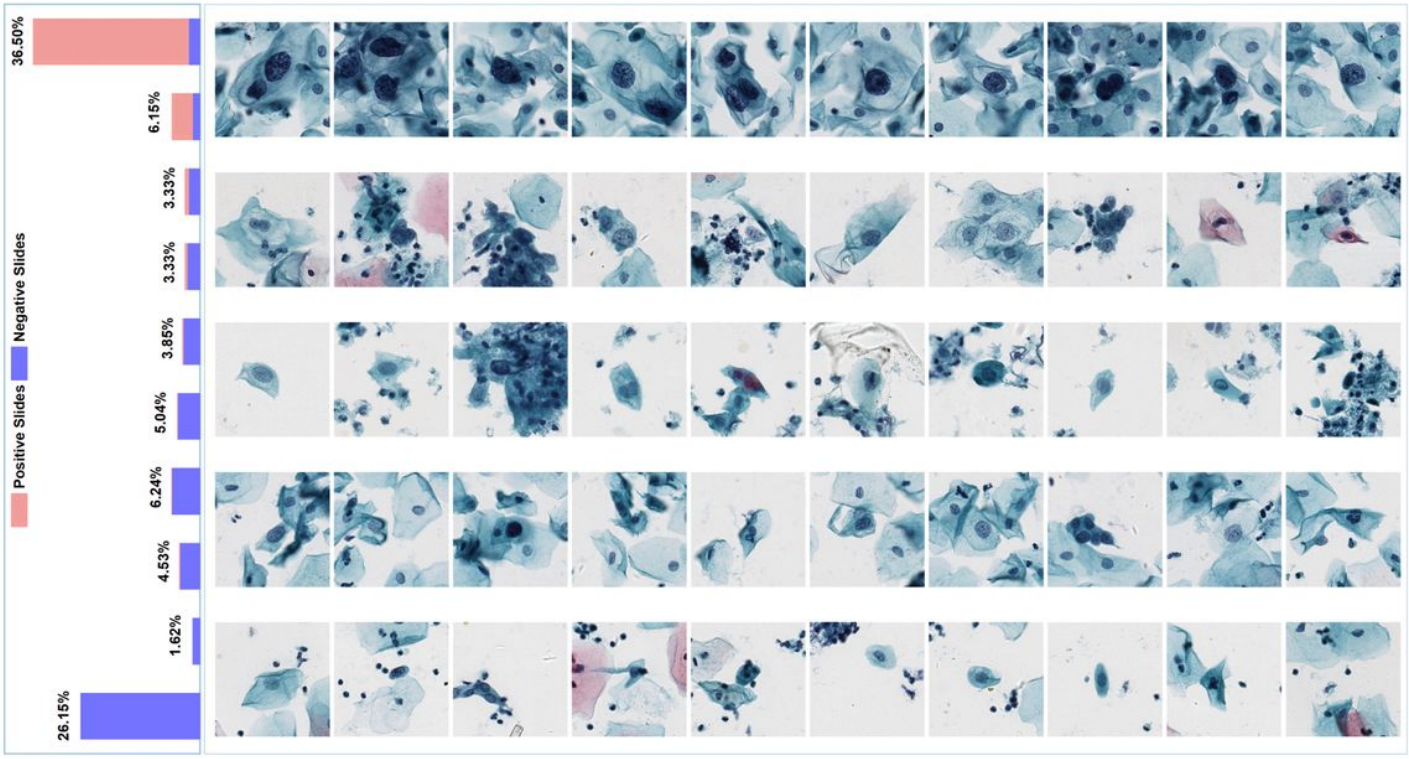


Figure 4

The recommended top 10 lesion cells of WSIs with different classification scores. The left subgraph is the frequency histogram of slide scores from 0 to 1 with an interval of 0.1 in groups E-F (n = 1,170). The right subgraph is the recommended top 10 lesion cells of slides with different scores.

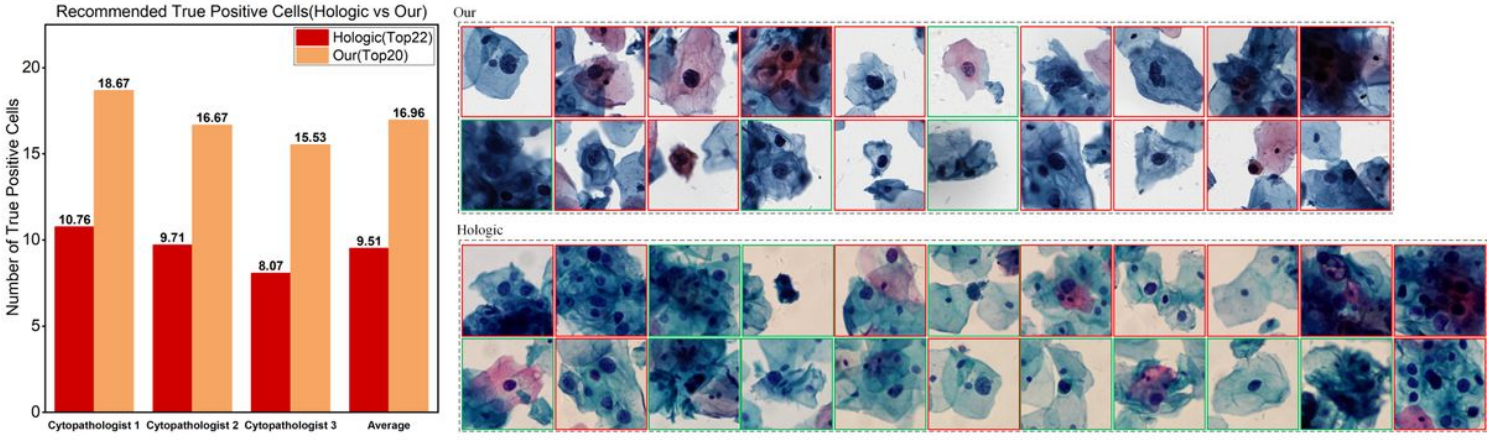


Figure 5

Comparison of our system and Hologic TIS on recommending top lesion cells. The histogram on the left shows the true positive cell number of our system's top 20 cell regions (about 60×60um2) and TIS' top 22

fields of view (far greater than 60×60um2) on 58 positive slides. The subgraph on the right is recommended cells of one positive slide. Notably, the most suspicious cells were cropped from TIS' top 22 fields of view according to cytopathologists' evaluation results. The cells with red border line were evaluated as true positive cell by cytopathologists.

a

Models	Train Strategies			Train Groups	Accuracy of Test Groups					
	Data Enhance-ment	Hard Samples Mining	Diverse Data Learning		A	B	C	D	E	F
Original	×	×	×	A	0.842	0.809	0.745	0.611	0.571	0.564
Enhanced	√	×	×	A	0.894	0.903	0.738	0.710	0.686	0.606
Mined	√	√	×	A	0.913	0.916	0.773	0.788	0.664	0.784
Baseline	√	√	×	AB	0.958	0.930	0.865	0.703	0.778	0.838
HR model	√	√	√	ABCD	0.988	0.941	0.958	0.916	0.965	0.905

b

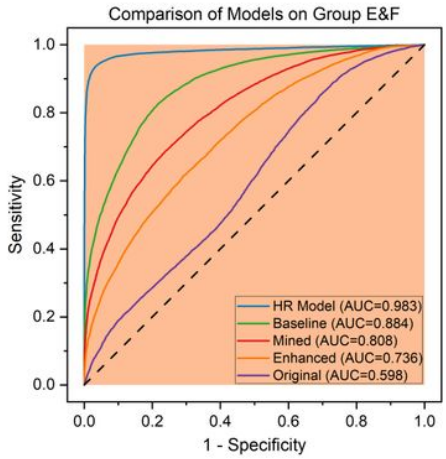


Figure 6

Importance of designed data enhancement, hard sample mining and diverse data learning. a, The ablation experimental results about the designed three learning strategies of data enhancement, hard sample mining and diverse data learning on the HR model. We used these learning strategies to train a series of control models step by step, and gave the classification accuracies on the test sets of groups A-F. b, AUC-ROC comparison of the Original, Enhanced, Mined, baseline and HR models on Groups E-F.

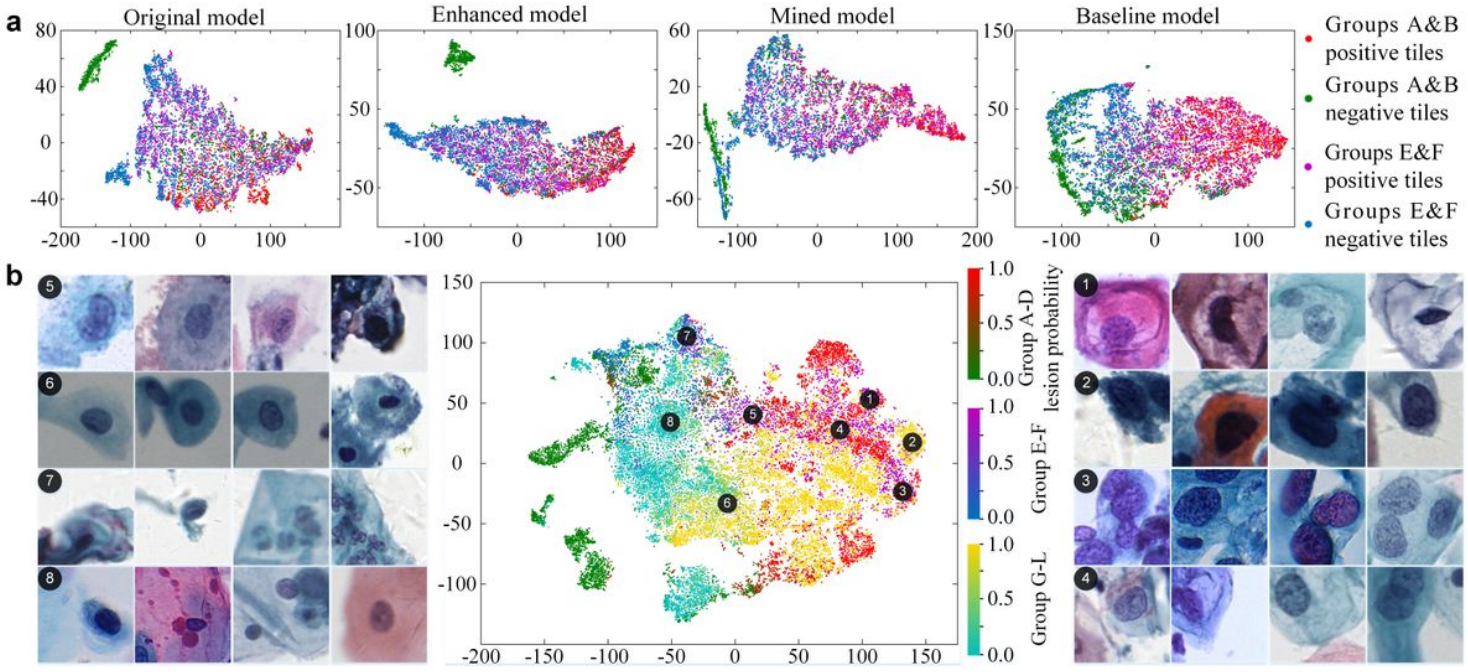


Figure 7

Generalized and rich feature representations of our models. a, Gradually aligned features between different groups of datasets. The last 2048-dimensional features of the original, enhanced, mined and baseline models were reduced to two dimensions using t-SNE on the test sets of groups A, B, E and F and were plotted in four subgraphs respectively. 1,000 positive test tiles and 1,000 negative test tiles in each group were randomly selected for the visualization. b, The distribution of dimension-reduced features of the HR model on the 12 groups test sets and the interpretation of the feature representations by examining corresponding cervical image tiles. 2,000 test tiles in each group were randomly selected for the visualization.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.docx
- Supplementary.docx
- Supplementary.docx