Laboratory Investigation (2021) 101:513–524
https://doi.org/10.1038/s41374-021-00537-1

ARTICLE

Quantitative analysis of abnormalities in gynecologic cytopathology with deep learning

Jing Ke[1,2] · Yiqing Shen[3] · Yizhou Lu[4] · Junwei Deng[5] · Jason D. Wright[6] · Yan Zhang[7] · Qin Huang[8] · Dadong Wang[9] · Naifeng Jing[10] · Xiaoyao Liang[1,11] · Fusong Jiang[8]

Received: 16 November 2020 / Revised: 21 December 2020 / Accepted: 4 January 2021 / Published online: 1 February 2021

Abstract

Cervical cancer is one of the most frequent cancers in women worldwide, yet the early detection and treatment of lesions via regular cervical screening have led to a drastic reduction in the mortality rate. However, the routine examination of screening as a regular health checkup of women is characterized as time-consuming and labor-intensive, while there is lack of characteristic phenotypic profile and quantitative analysis. In this research, over the analysis of a privately collected and manually annotated dataset of 130 cytological whole-slide images, the authors proposed a deep-learning diagnostic system to localize, grade, and quantify squamous cell abnormalities. The system can distinguish abnormalities at the morphology level, namely atypical squamous cells of undetermined significance, low-grade squamous intraepithelial lesion, high-grade squamous intraepithelial lesion, and squamous cell carcinoma, as well as differential phenotypes of normal cells. The case study covered 51 positive and 79 negative digital gynecologic cytology slides collected from 2016 to 2018. Our automatic diagnostic system demonstrated its sensitivity of 100% at slide-level abnormality prediction, with the confirmation with three pathologists who performed slide-level diagnosis and training sample annotations. In the cellular-level classification, we yielded an accuracy of 94.5% in the binary classification between normality and abnormality, and the AUC was above 85% for each subtype of epithelial abnormality. Although the final confirmation from pathologists is often a must, empirically, computer-aided methods are capable of the effective extraction, interpretation, and quantification of morphological features, while also making it more objective and reproducible.

Introduction

Cervical cancer was the leading cause of cancer-related death in women in eastern, western, middle, and southern Africa, and the fourth most common cancer among women globally [1]. However, the cytological screening has led to a major decline in cervical cancer burden in resource-rich countries [1, 2]. The interpretation of cervical cytopathology on samples of cells or tissue fragments under the microscope has been the foundation of a cervical cancer

✉ Jing Ke
kejing@sjtu.edu.cn

1 Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

2 School of Computer Science and Engineering, University of New South Wales, Sydney, NSW, Australia

3 School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China

4 School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China

5 School of Information, University of Michigan, Ann Arbor, MI, USA

6 Department of Obstetrics and Gynecology, Columbia University, New York, NY, USA

7 Department of Pathology, Shanghai Tongshu Medical Laboratory Co.Ltd, Shanghai, China

8 Department of Endocrinology and Metabolism, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai Clinical Center for Diabetes, Shanghai, China

9 Quantitative Imaging, Data61 CSIRO, Sydney, NSW, Australia

10 Department of Micro-Nano Electronics, Shanghai Jiao Tong University, Shanghai, China

11 Biren Research, Shanghai, China, Shanghai, China

diagnosis for cytologists for over 90 years. A high-volume procedure in the laboratory, cytopathology, or cytology for short has shown its efficiency in lesion detection and decrease in invasive cervical cancer by Pap test. In more recent years, liquid-based cytology (LBC) has replaced conventional cytology attributed to its advantages in sample quality, reproducibility, sensitivity, and specificity [2, 3]. Today, the scanning of glass slides into digital imaging has further standardized the process and diagnoses of cytology [4]. The current gold standard is manual screening or review of glass slides [5].

Nevertheless, several limitations have been noticed with human interpretations of digital cytology. A disadvantage of using digital whole-slide images (WSIs) is that it might take a longer time to diagnose as compared with glass slides [4]. Containing gigabytes of pixels, WSIs suffer from the lack of diagnostic focusing capability, and are prone to miss some key information of tumors from numerous cells by the human eye. Second, possessing differential cytology expertise, cytologists may present diagnoses with inconsistencies. The diagnostic accuracy often varies greatly by country, as well as knowledge of pathologists, that manual pathology assessment may lead to high subjectivity. Take an example of a review on 40 studies with cervical cytology test samples from >140,000 women; for cervical intraepithelial neoplasia of moderate dysplasia (CIN 2 +) as the cutoff, the results for sensitivity ranged from 52 to 94% and specificity from 73 to 97% on LBC. Published few unbiased types of research suggest that the mean sensitivity of Pap test is 47% (range 30–80%) with a mean sensitivity of 95% (range 86–100%) [6]. Moreover, there is still a lack of approaches to routinely, automatically, and stably extract the rich morphologic information from WSIs. Consequently, clinical reports often highlight the most severe type of abnormality [4]. However, the quantification of differential abnormalities is considered to be more informative in further decisions and treatments [7].

In this decade, the diagnoses performed by pathologists have been and will continue to be assisted by deep learning interpretation. In this fashion, technical problems of intensive workload, low producibility, or high subjectivity might be well resolved. In particular, as LBC is stuck to the bare morphological evaluations, it is well-qualified to be detected with convolutional neural networks [3, 8]. However, in the field of cytopathology, compared with histopathology, the problems encompassed are more severe. Primarily, a drawback is the lack of datasets and associated annotations. Histopathology images and associated annotations are relatively publicly available like the cancer genome atlas [9–11] and many grand challenges [12]. To this end, a rich catalog of deep-learning-based approaches has been explored for classification, detection, and segmentation [9, 10, 13]. In contrast, large cohorts of cytology images are almost publicly unavailable with a very limited number of researches in cell classification or segmentation typically on small-image regions [14–16]. Furthermore, in contrast with histopathology that possesses a much higher level of local spatial correlation in the individual fields of view, the dysplasia is prone to be scattered in the WSIs. Nevertheless, in contrast with these obstacles, auto diagnosis is in extraordinary demand for gynecologic cytology as cervical screening is usually included in regular health checkups for women [17].

In this study, we bridged these gaps with a robust AI-assisted cytology diagnostic system. The overall target was to detect cellular abnormalities and make slide-level predictions for gynecologic cytology images. Our three major contributions are: (1) we collected and scanned a private set of cytology images and performed our study covering 130 slides. We annotated a variety of phenotypes, designed annotation strategies, and developed computer-aided approaches for the quantitative analysis. (2) We integrated the methodology of spatial correlation and evaluated cellular nuclear area to further improve the classification performance after the deep neural network. As a diagnostic assistant, the system eventually aggregated the quantitative results. By contrast, most of the existing works were often stuck on segmentation on individual cells in groups or classification of the prepared well-cropped single cells [18, 19]. Some previous machine-learning techniques used small regions as case studies [20, 21]. A very few works analyzed WSIs, yet lack the automatic and informative profile of phenotypes at slide level [5, 22]. (3) Overall, the prediction performance reached the sensitivity of 100% and the specificity of 91.1% in the diagnoses of positive and negative slides, along with an average accuracy of 94.5% in the abnormal and normal cell binary classification. By this research, we are capable of providing quantitative morphological recognitions to atypical cells as well as comparatively objective clinician decisions, which will further contribute to the clinical diagnosis for cervical cancer.

Materials and methods

Gynecologic cytology images

Collected from Shanxi Tumor Hospital, a total number of 130 specimens from 2016 to 2018 were made by LBC method and scanned into high-resolution images all with ethical approval. The signal-plane scanner was used to produce the high-resolution digital images at 40 × magnification. We also encompassed some out-of-focus cells in training a robust deep-learning model. All the 130 digital slides were identified as 79 normal, 24 atypical squamous cells of undetermined significance (ASC-US), 13 low-grade squamous

intraepithelial lesion (LSIL), 2 ASC-H, 7 high-grade squamous intraepithelial lesion (HSIL), and 5 squamous cell carcinoma (SCC) by the hospital. The slide-level diagnoses, the cellular contouring, as well as the image patch labeling work that was performed by two pathologists together. When there was inconsistency over an abnormality classification, to minimize inter- and intraobserver variability, we invited a senior pathologist to offer the classification label. The slide-level, patch-level, and image-level manual annotation is considered as ground truth in this AI-assist system.

Most successful approaches to training CNN models do not take the whole image as input for morphological feature extraction [11, 23]. Instead, image patches, usually the cropped ones with dimensions ranging from 32×32 up to 5000×5000 pixels, are adopted by the neural networks [24]. We tessellated all the valid areas of a whole-slide image, which is often presented as a circle shape, to non-overlapping patches of 256×256 pixels, as they are considered discriminative for a subtype to be well identified by CNNs at the $40\times$ magnified rate. The original digital slides ranged from about 55,000 to 65,000 pixels in width and height, and the overall number of patches in the experimental test was around seven million in this case study.

Annotation strategies

Today, successful applications of deep-learning techniques still heavily rely on the quantity and quality of data annotations, particularly in the medical imaging domain. With adequate clinically well-annotated datasets, minor differences that are hard to discriminate by human observers are sensitive to AI detectors. Predominantly, the high accuracy in deep-learning approaches is obtained via a strong supervised manner. For instance, given a WSI, experts are required to annotate every pixel in every patch, which is practically unfeasible. A simpler approach would be labeling a patch with a category while turning it to be a weakly supervised manner. Pixel-wise labeling for segmentation and image-level labeling for classification are two popular methods to provide essential knowledge. In our research, the advantages of the two methods were combined to train a system, characterized as labor-saving and highly precise, where a very small proportion of cells was pixel-wise annotated and the majority were image-level labeled to a certain category as shown in the data annotation stage in Fig. 1.

In the pixel-level annotation task, we performed per-pixel annotation to 800 image patches cropped from 20 whole-slide images. Instead of differentiating subtypes, we only contoured the nuclei and cytoplasm from cells and background, respectively. These image patches were trained to do the semantic segmentation to locate all the nuclei in WSIs. After the nucleus segmentation task was performed on individual image patches, they were pieced together to form WSIs again, where the nuclear areas were detected and masked out as a result of semantic segmentation. The patches with a nucleus as the center were then cropped from the original images at a fixed size of 256×256 to go through the following cell subtype classification stage. This selected patch size was considered suitable for nuclei grading by deep-learning framework [14, 25].

The image-level labeling targeted at the cell subtype classification. The pathologists selected and cropped 5000 representative cells of 12 categories with differential morphological features (shown in Fig. 2) from the original digital slides. The 12-category strategy was based on the distinguishable morphology features where normal cells encompass several categories. However, as the number of recognizable ASC-H cells was far away from using as one class of training samples, we excluded this phenotype in the classification task. To enhance the labeled dataset with more ambiguous patches, we use these labeled patches as the initialization of the annotated dataset pool by active learning [26]. In the unlabeled pool that was the collection of image patches whose center was identified as the nucleus in the 3-class semantic segmentation stage, we iteratively picked up a total number of 1000 image patches characterized with the highest uncertainty in training the classification model. These images were sent to pathologists for labeling, which progressively increased the labeled pool. To acquire more training samples without manual annotation, we used elastic transformation, flipping, and rotation to abnormal classes for data augmentation, which also achieved a balance in classes. Scaling was not adopted, as the nuclear area might be considered as a key feature in training a classification neural network [27]. The random sampling method that was often adopted in the routine histopathology annotation was also not used in our system. This sampling method might lead to a much higher class imbalance in cytology; moreover, many randomly cropped patches may contain incomplete cellular detail.

In addition, with the small proportion of patch-level contoured cells, pathologists are free to annotate each pixel in the succeeding annotation work, should there be more negative subtypes to be involved in the hard negative mining, or positives failed to be collected in this study, for instance, atypical glandular cells.

Positive and negative training sets

Abnormalities are cataloged to ASC-US, koilocytotic atypia, LSIL, HSIL, or SCC on the cellular level (Fig. 2A) in our training samples. Although the koilocyte information was not separated from LSIL to be labeled from the hospital, we also performed cytologic detection on it, as it has been considered to be closely related to human papillomavirus infections [28]. However, typically in a digital cytology slide, a considerable proportion of normal cells may present similar morphological

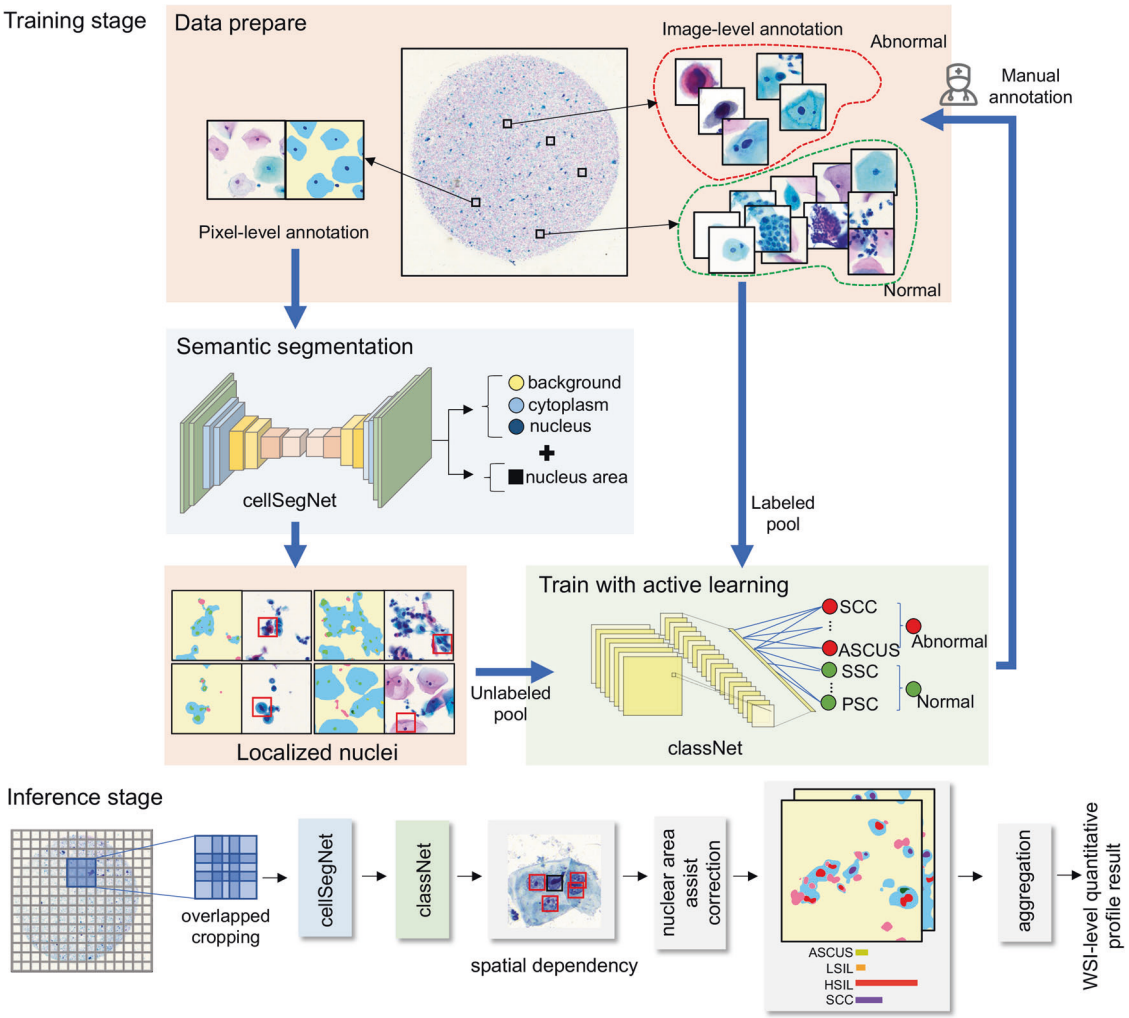

Fig. 1 The overall architecture of the proposed computer-aided cytology image diagnostic system. The AI-assisted system comprises the pixel-level and image-level annotation, the segmentation and classification neural networks, the spatial correlation model, the nuclear area profile, and the aggregation model.

features as abnormal cells, and they are widely spread over the cytology image. For example, a folded superficial squamous cell will be more likely to be identified as SCC. We hence used hard negative mining to increase normal samples, namely superficial squamous cells, intermediate squamous cells, parabasal squamous cells, endocervical cells, inflammatory cells, numerous epithelial cells (also include various cell aggregates without distinctive recognizable features to pathologists), and folded superficial squamous cells (Fig. 2B). The inclusion of these negative cells in the training samples significantly excluded a large number of false predictions and thus improved the overall specificity.

Diagnostic system

The overall architecture of the proposed computer-aided cytology diagnostic system consists of five functional components, namely the segmentation model, the classification model, the spatial correlation model, the nuclear area correction model, and the aggregation model. The overall framework is illustrated in Fig. 1.

Segmentation and classification

We employ a deep-learning framework consisting of two independent neural networks for the task of semantic segmentation. The first neural network, aiming to contour the boundary of cell and nucleus, is a hybrid architecture containing two conjunct paths for feature extraction and interpretation. The first half is the contraction path, also called the encoder, employing residual structure [29] to extract context information. Two types of residual blocks are integrated into the encoder, one RB halves the original high dimension of the code while doubles the number of

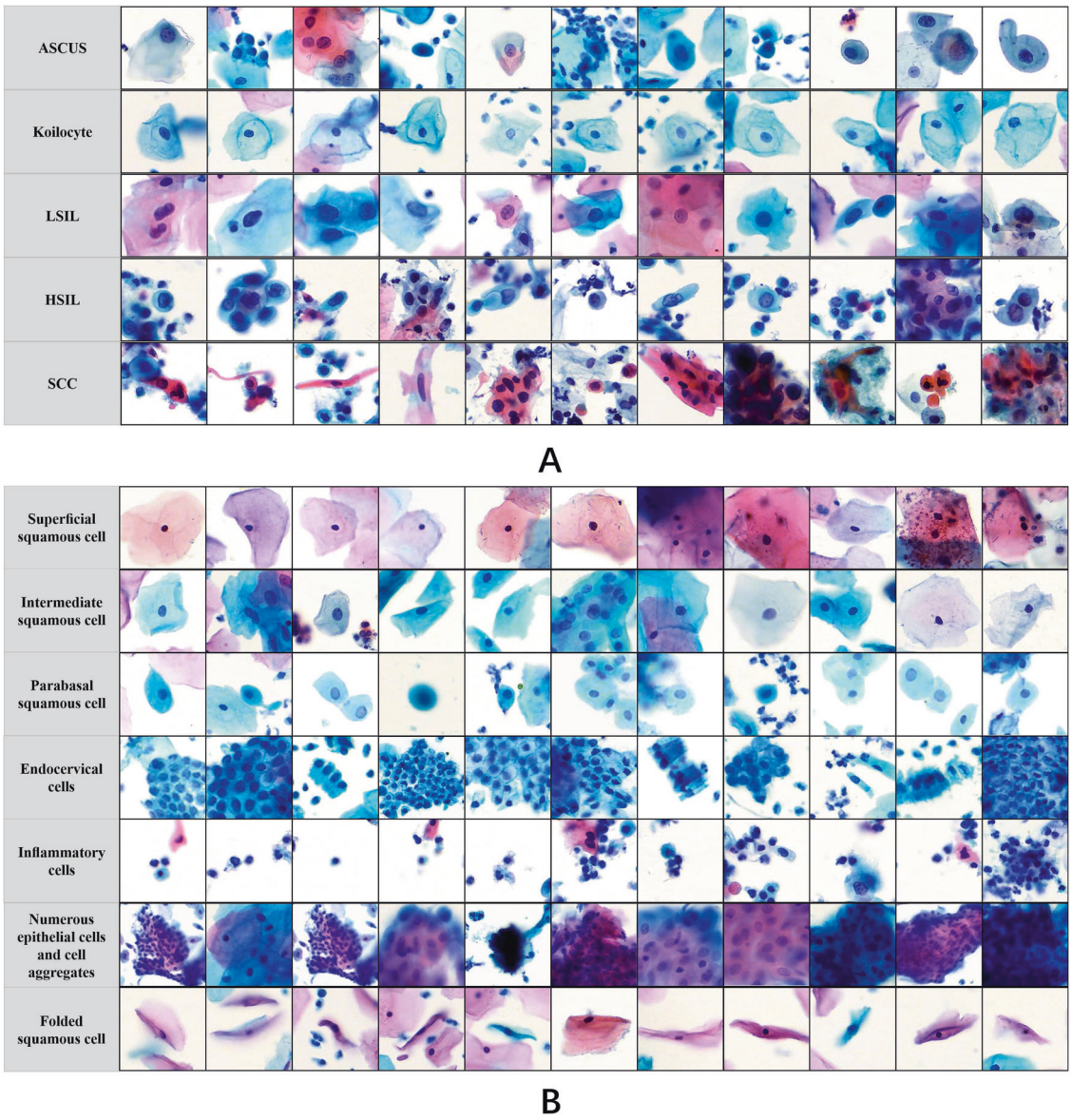

Fig. 2 A couple of example image patches cropped from the collected slides that were used as training input in the deep-learning system. **A** Positive samples. **B** Negative samples. The catalog of these classes targeted at a higher accuracy in the training process.

channels, and the other stays with the original high dimension of the code and the number of channels. The second path is the expanding path, also called the decoder, which retains the right-half structure of U-Net [30] to retrieve precise localization using transposed convolutions. This path is used to recover the data of code and also to retrieve high-resolution features sent by skip connections from each RB group to the output of the segmentation map. In this former network, we extract distinctive morphological features from the input image and then catalog each and every pixel into three categories, namely nucleus, cytoplasm, and background in a WSI, in which unknown tissues, blood cells, or mucus were also cataloged to the background. In this way, we get a clear boundary between the nucleus and its cytoplasm in a cell. Furthermore, via this segmentation approach, the parameter of the cellular nucleus can also be obtained and afterward used to improve the classification results from the deep learning. In the second neural network, we localize and analyze the image patches whose centers have already been identified as cellular nucleus in the first stage. These patches are cropped from original WSIs and subsequently processed with a pre trained ResNet-50 for the 12-category classification.

To further explore a higher interpretation capability with the limited training samples, we employ the transfer learning [31] from the pre trained ResNet-50 on the ImageNet and fine-tune our network. We do not train the network de novo. Regarding the differences at the high-level features

between the natural objects in the ImageNet and cytological cells, we retain the weights of the first seven convolutional blocks directly from the pre trained structure. Afterward, the first seven layers are frozen, and the trainable layers, including the nine convolutional blocks and fully connected layers, are initialized with random value and fine-tuned with our annotated multiclass sample cells.

Spatial correlation

In a patch-based fashion, the diagnosis of infections may be limited to the field of view. In some scenarios, contextual information can be decisive to the classification performance. In clinical practice, pathologists often incorporate context information together with morphological features to assist the recognition of cells and tissues. Correspondingly, data dependencies between adjacent patches can be integrated into the identification of individual patches in the histopathology image classification by deep-learning approaches, in particular for the phenotypes that cannot be well recognized within the limited patch size. Nevertheless, parallel approaches are rarely found to be applied in the cytopathology image. In this study, we take account of patch dependencies by a deformable conditional random field (DCRF) model [32]. This model learns the offsets and weights of the most representative patches in a spatial-adaptive manner and achieves an average of 5.0% improvement in classification, compared with its backbone residual learning structure.

Routinely, the patch size is fixed, e.g., 500×500. When a suspicious patch is located, spatial correlation will be incorporated for further analysis. We employ a DCRF with the following Gibbs distribution for higher classification performance, i.e.,

$$P(\mathbf{l} = l|\mathbf{p}) = \frac{1}{Z(\mathbf{p})}\exp(-\mathbb{E}(l, \mathbf{p})) \tag{1}$$

where $\mathbf{p}$ is a one-to-one mapping from the central coordinate in a WSI to a fixed-size patch, l is another mapping from the central coordinate to a specific label from the label set, $Z(\cdot)$ is a normalization constant to make Eq. (1) into a proper probability distribution, and $\mathbb{E}(l, \mathbf{p})$ is the energy function. In the DCRF model, the energy function is formulated with additional trainable offsets $\delta\mathbf{p}$ in a fully connected pairwise conditional random field model, i.e.,

$$\mathbb{E}(l, \mathbf{p}) = \sum_{\mathbf{p}\in\mathcal{G}} \psi_{\mathrm{u}}(l(\mathbf{p}+\delta\mathbf{p})) + \frac{1}{2}\sum_{\mathbf{p}\neq\mathbf{p}'\in\mathcal{G}} \psi_p(l(\mathbf{p}+\delta\mathbf{p}), l(\mathbf{p}'+\delta\mathbf{p}')) \tag{2}$$

where $\mathcal{G}$ denotes a collection of patches of interest. The unary potential ψ_u measures the cost of patch $\mathbf{p}+\delta\mathbf{p}$ taking the label $l(\mathbf{p}+\delta\mathbf{p})$, and the pairwise potential ψ_p measures the spatial correlation between the two patches, defined as

$$\psi_p(l(\mathbf{p}+\delta\mathbf{p}), l(\mathbf{p}'+\delta\mathbf{p}')) = w_{(\mathbf{p}+\delta\mathbf{p},\mathbf{p}'+\delta\mathbf{p}')} \cdot \mathbb{I}(\mathbf{p}+\delta\mathbf{p}, \mathbf{p}'+\delta\mathbf{p}') \cdot \exp\left(-\frac{\|\delta\mathbf{p}\|^2+\|\delta\mathbf{p}'\|^2}{2\sigma^2}\right) \cdot \left(1 - \frac{Y(\mathbf{p}+\delta\mathbf{p})\cdot Y(\mathbf{p}'+\delta\mathbf{p}')}{\|Y(\mathbf{p}+\delta\mathbf{p})\|\,\|Y(\mathbf{p}'+\delta\mathbf{p}')\|}\right). \tag{3}$$

In this definition, $\mathbb{I}(\cdot)$ stands for the indicator function, i.e., it is equal to 1 if and only if patches $\mathbf{p}+\delta\mathbf{p}$ and $\mathbf{p}'+\delta\mathbf{p}'$ have the same label and otherwise, it is equal to 0. The feature vector $Y(\cdot)$ is extracted by a CNN. The coefficient σ in the Gaussian kernel is tunable, and the trainable weight $w_{(\mathbf{p}+\delta\mathbf{p},\mathbf{p}'+\delta\mathbf{p}')}$ associated with patches $\mathbf{p}+\delta\mathbf{p}$ and $\mathbf{p}'+\delta\mathbf{p}'$ is updated by the back-propagation (BP) algorithm.

The CNN feature extractor block in DCRF is the pre trained ResNet-50 for the 12-category classification described in the previous section. Integrating the DCRF model to extract spatial information in the training and prediction stage, ResNet-50 achieves an observable classification performance improvement.

Quantitative nuclear area analysis

The state-of-the-art deep-learning approaches still tend to underperform in comprehensive cases, particularly when phenotypes often appear morphologically similar to each other. However, the nuclear area of individual cells can well discriminate ambiguous cases between two similar phenotypes, when deep-learning results fail to be universally suitable. For this reason, a quantitative profile of the pixels of the nucleus area is carried to the outcome of cellular classification and segmentation from the deep-learning framework. An analysis of over 2000 cells in each subtype, in both the training and test dataset, showed the approximate range of their corresponding nuclear area.

Aggregation

After the multiple-stage process of individual patches and cells, the results are aggregated. The quantitative profile, as well as the detected abnormal cells of individual slides, will be sent to pathologists for the final diagnoses and the succeeding medical treatment. It is important to note that, although AI-assisted methods are often considered as more objective and reproducible than manual evaluation, the gold standard is indeed created by experienced pathologists, which sometimes leads to a paradox in clinical evaluation [25]. Also, due to the limited prediction accuracy of AI, the varying knowledge of domain experts, and the relatively subjective classification of abnormalities, AI is used to support pathologists in making efficient and more accurate diagnoses, but not to replace them [33].

Table 1 The quantity and scale of patches used in the performance evaluation.

	Training	Validation	Testing	Patch size
Classificatin	6803	1701	1701	256 × 256
Segementation	2001	650	650	776 × 776

Results

Implementation details

A large amount of seven million patches were cropped from the original digital slides. However, we used only a small proportion of high-quality patches to train the segmentation and classification networks. The number of nonoverlapping patches in the training set, validation set, and testing is shown in Table 1. The patch for the classification task was cropped at the size of 256 × 256, while for the segmentation task was 776 × 776 accordingly to comply with the annotation strategies described above.

The hyperparameters for the classification task were set as follows: the cross-entropy was used for the loss function and the Adam optimizer with a learning rate = 10^{-5} was adopted, and the number of training epochs was 40. The hyperparameters were set as follows: the cross-entropy was used for loss function, the Adam optimizer was valued by a learning rate = 10^{-4}, and the number of training epochs was 29. All the models were written with Python version 3.5. Experiments were conducted on the state-of-the-art NVIDIA Tesla V100 with 32-GB GPU memory.

Cellular-level segmentation and classification

We evaluated the system with three metrics, namely pixel accuracy, mean pixel accuracy, and mean IoU:

$$\text{pixel accuracy} = \frac{\sum_i^M n_{ii}}{\sum_i^M \sum_j^M n_{ij}} \tag{4}$$

$$\text{mean pixel accuracy} = \frac{1}{M+1} \sum_i^M \frac{n_{ii}}{\sum_j^M n_{ij}} \tag{5}$$

$$\text{mean IoU} = \frac{1}{M+1} \sum_i^M \frac{n_{ii}}{\sum_j^M n_{ij} + \sum_j^M n_{ji} - n_{ii}} \tag{6}$$

where $M = 3$ denotes the three categories of nucleus, cytoplasm, and background. n_{xy} represents the number of pixels classified to class y with ground truth x. To evaluate the performance of the hybrid ResNet and U-Net, we utilized the fivefold cross-validation to the pixel-wise annotated 800 patches. We demonstrated the semantic segmentation results in Table 2, as compared with two baselines where U-Net++ [34] is an updated version of the classical segmentation network U-Net with a differential downsampling.

To evaluate the cell classification performance, we performed the validations on the 6000 labeled image patches. We demonstrated the test results by a confusion matrix heat map of 12-category classification in Fig. 3A, and the receiver-operating characteristic curve of ASC-US, LSIL, HSIL, and SCC in Fig. 3B. To guarantee the specificity of the deep-learning diagnostic system, we also marginally increased the confidence of CIN + subtypes in comparison with normal phenotypes. The overall accuracy for normal and abnormal binary classification was 0.945 ± 0.006. For abnormal cell classification, we achieved the specificity of 0.929 ± 0.008 and the sensitivity of 0.923 ± 0.006, and the experiment results did not show significant deviations in the overall trend.

The final semantic segmentation output was composed of the results from multiple stages, where individually processed patches were pieced together to form a final wholly processed image. We show an example of an image patch of 2500 × 2500 pixels cropped from a WSI in Fig. 4.

Slide-level profile

We performed systematic diagnoses over the 130 digital WSIs by our proposed model and presented the quantitative profile result of abnormal cells in both positive slides (Fig. 5A) and negative slides (Fig. 5B). The labels were provided by three pathologists with at least two consistent decisions.

As the number of cells varies from slide to slide, it is the ratio of abnormal cells to normal cells. We made a quantitative analysis of both normal and abnormal cells in both positive and negative slides, with the results and ratio distribution demonstrated in Fig. 5C.

Discussion

All in all, in the diagnostic task to detect abnormality and to grade dysplasia, our proposed computer-aided system yielded the performance of 100% sensitivity at slide level, and the specificity varied from 96.2% to 98.1% in the identification of ASC-US prediction, 87.9% to 88.0% in CIN 1+ prediction, 93.2% to 95.7% in CIN 2+ prediction, and 99.2% in SCC prediction, based on the cytology images involved in the training and inference in this case study. There were moderate agreements between the pathologists. Although the grading result is often a trade-off between sensitivity and specificity in many applications, our system was designed to achieve higher sensitivity than the observers, while to retain a comparably lower specificity to

Table 2 The semantic segmentation results of nuclei, cytoplasm and background on the annotated image patches.

Methods	Pixel accuracy	Mean pixel accuracy	Mean IoU
U-Net	0.913 ± 0.009	0.906 ± 0.011	0.899 ± 0.005
Mask RCNN	0.922 ± 0.010	0.910 ± 0.010	0.898 ± 0.008
U-Net++	0.952 ± 0.003	0.941 ± 0.004	0.904 ± 0.005
Our proposed model without quantitative nuclear area analysis	0.970 ± 0.002	0.948 ± 0.007	0.914 ± 0.006
Our proposed model with quantitative nuclear area analysis	**0.974 ± 0.001**	**0.955 ± 0.007**	**0.913 ± 0.007**

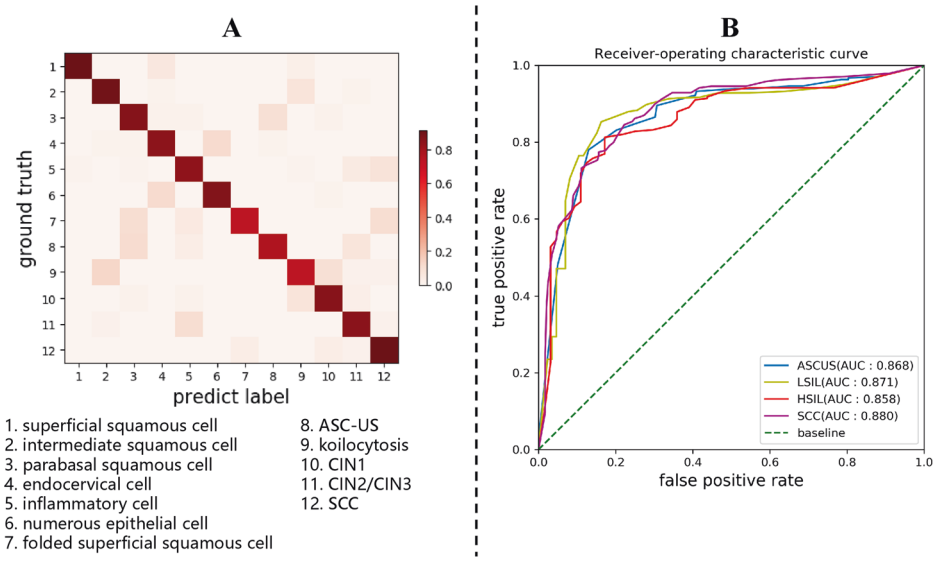

Fig. 3 Performance of the 12-category classification in cervical cells. A The confusion matrix heat map of 12-category classification in our cervical cell benchmark test. **B** The receiver-operating characteristic curve (ROC) of abnormalities, including ASC-US, LSIL, HSIL and SCC. The catalog of LSIL encompassed koilocyte detected in the WSIs, following the 2015 Bethesda System.

eliminate potential false negatives. It also successfully detected higher-level lesions as compared with the labels tagged from the hospital and confirmed by the three pathologists. However, suffering from verification bias, the sensitivity of cytology to detect definitive invasion squamous and glandular lesions is difficult to establish without histological confirmation [23]. Despite that, the advantage of cervical cytology in the detection of precursor lesions is clear, that it can be treated before the development of invasive carcinoma.

After the prediction from the deep neural networks, there were two noticeable misclassifications that required the succeeding aggregation model to improve. One ambiguous identification came from the identification between LSIL and koilocyte. Both phenotypes presented nuclear enlargement or hyperchromasia, while koilocyte is characterized by a rim of condensed cytoplasm, looking like a halo around the dysplastic nucleus [8]. We could not exclude the fact that, cytologically, a very clear borderline between these two phenotypes was absent at the annotation stage. However, as koilocytosis has been encompassed into LSIL in the 2015 Bethesda [7], the final LSIL diagnosis was not affected. Another obstacle for successful recognition occurred when heavy inflammation was presented in the image. Appearing in cluster or overlap, inflammatory cells might be misidentified as endocervical cells or ASC-US, while most of the time to be HSIL, as both of these two phenotypes are characterized with a large nucleus-to-cytoplasm ratio. However, sometimes, it was the same difficulty for pathologists to recognize well when the images were cropped into individual small patches. This is also a disadvantage of the current patch-based network. When a slide contains an extraordinary count of inflammatory cells, which overwhelms normal squamous epithelial cells in quantity and area, the overall cellular-level misclassification might observably arise and cause a false-positive prediction at the slide level.

The incorporation of handcrafted features, such as the nuclear area, had effectively improved the classification outcome from the neural network. The nuclear area of inflammatory cells was profiled as generally <150 pixels while HSIL >500 pixels, at the magnification rate of 40×, as shown in Fig. 6. We set 200 pixels as a criterion of the nuclear area to eliminate a majority of false inflammatory cells in the identification of HSIL. Likewise, as a small proportion of misclassification was between folded superficial squamous cells and SCC, we conservatively leave out those under 100 pixels in the nuclear area at the decision of

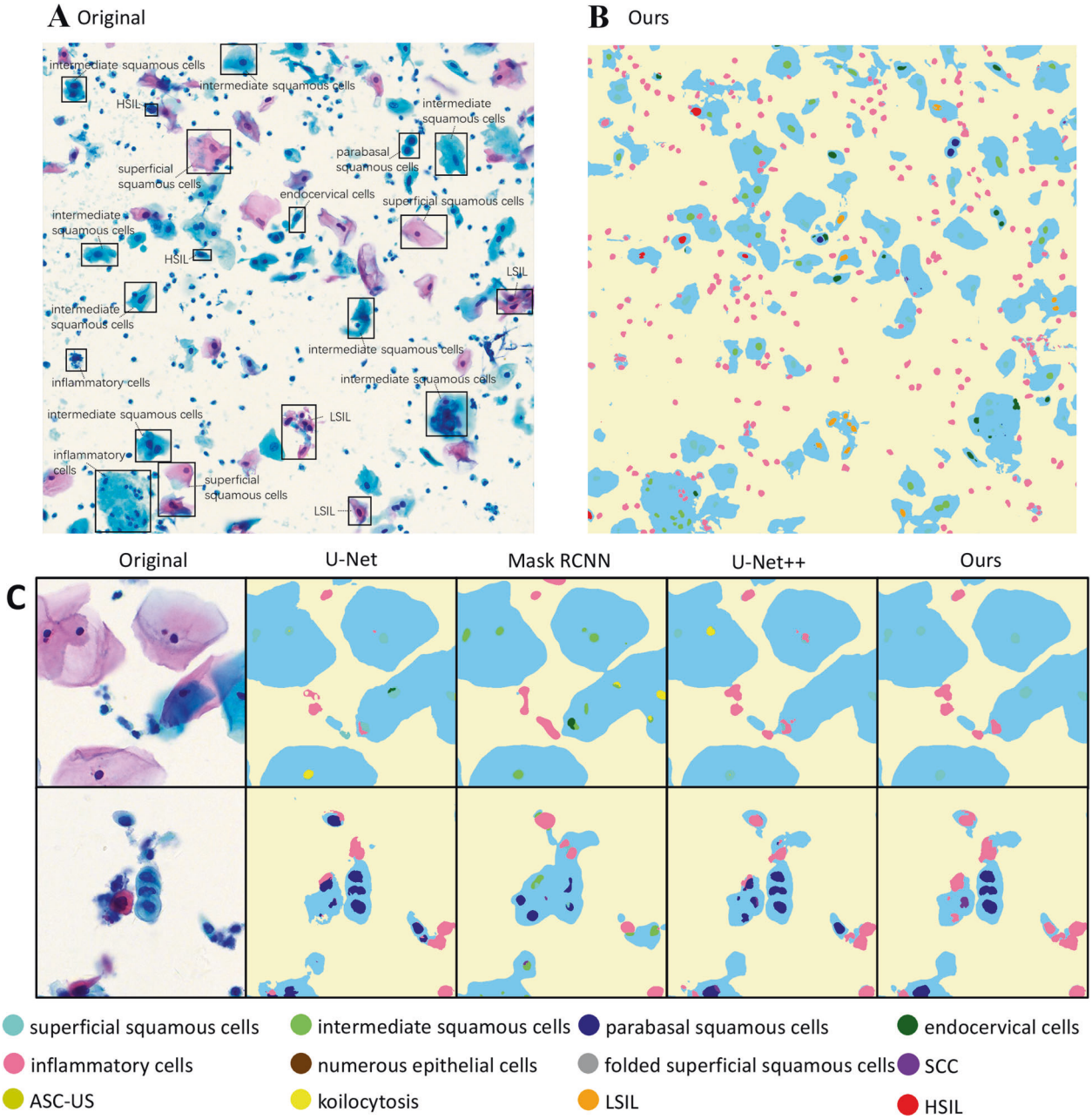

Fig. 4 An example of image patch of 2,500 × 2,500 pixels cropped from a whole-slide image. A The manual annotation for morphological interpretation. **B** The corresponding semantic segmentation result by the proposed two-stage deep-learning structure. This slide was labeled as HSIL from the hospital and also predicted as HSIL. **C** The segmentation comparison with state-of-the-art structures.

SCC. Overall, this system has the highest level of concordance and diagnostic confidence with pathologists at the slide-level prediction of SCC in the abnormality detection. It rarely showed overlap between normal squamous cells and high-grade dysplasia. Experimentally, it significantly improved the specificity when using CIN 2+ as a cutoff.

It is notable that the current deep-learning approaches in pathology still firmly follow the expert knowledge given to its annotated data. When training data come from one annotator and the test data from another, the knowledge and biases of the first annotator are sometimes systematic enough for a diagnostic system to learn them well and cause inaccurate results in the overall sensitivity and specificity of the system [9, 12]. Thus, it would be more objective to evaluate a deep learning system, when the annotation and confirmation were performed by the same pathologists and with high consistency [25].

We implemented four popular architectures for the classification task, namely AlexNet, VGG-16, Google Inception-ResNet-V2, and ResNet-50. Resnet has fewer

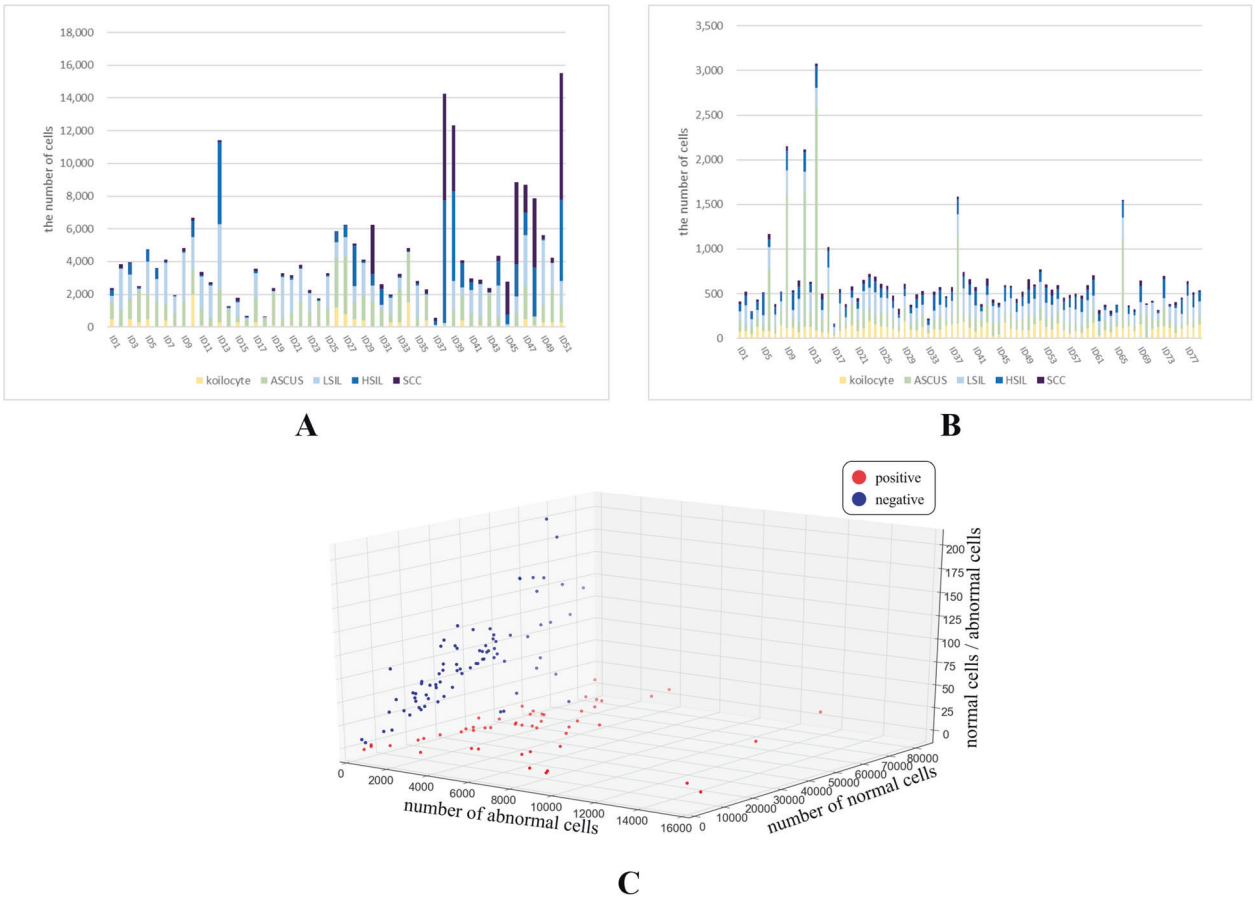

Fig. 5 The automatic quantitative profile of abnormal cells in whole slide images. A The identified positive slides by manual assessment. **B** The identified negative slides by manual assessment. **C** Slide-level normal and abnormal cells in both positive and negative slides.

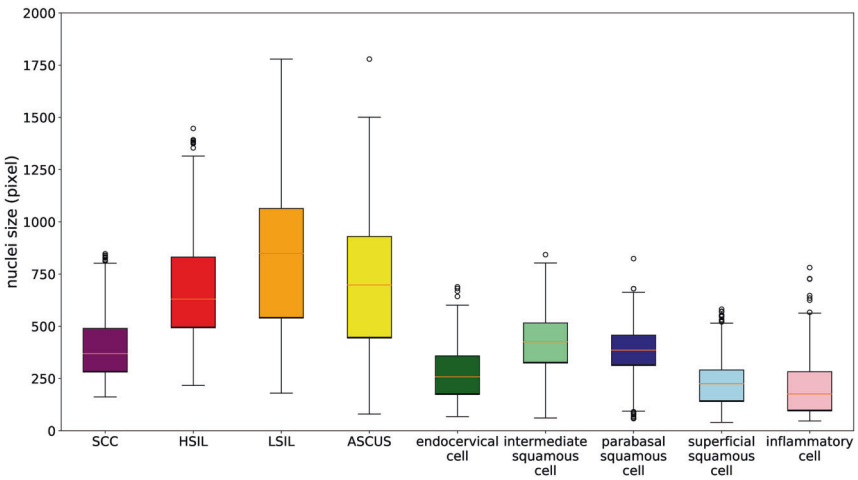

Fig. 6 Quantitative analysis on nuclear pixels of the annotated multiple-class cells at the magnification rate of 40×. This nuclear-area profile assists well at the recognition between some phenotypes, such as HSIL and inflammatory cells.

parameters compared with VGG and AlexNet, while being characterized as a deeper network. It demonstrated a better feature extraction capability at both pixel-level segmentation and patch-level classification. Inception is a multiscale feature extractor of multiple kernel size per layer, which showed outstanding performance on the ImageNet dataset. However, it did not demonstrate a better performance than other networks in our experimental dataset, and similar results had also been presented in the literature in pathology WSI classification [10, 11]. Consequently, leveraging the advantages in the training and prediction performance, we picked up ResNet-50 as the backbone for classification.

While crucially decreasing the annotation effort from pathologists, this coarse-to-fine two-stage semantic segmentation also outperformed the direct 12-catalog segmentation by a large margin empirically. The root cause is the extraordinary similarities in high-level features among the cell instances. Although a straightforward semantic segmentation to large-scale datasets of natural images, like COCO, did achieve outstanding performance, it failed to achieve high accuracy empirically in this cervical cell whole-slide image classification task.

Moreover, the inevitable obscuring tissues, such as blood, folded cytoplasmic borders, or thick areas of overlapping epithelial cells did not reduce the sensitivity in the diagnosis, compared with the diagnoses from cytologists. However, it might cause some false positives at the cellular level. Occasionally, when nuclei were presented to be overlapped or blurred, they were prone to be taken as an enlarged nucleus caused by dysplasia. In addition, false classifications on cells that were out of focus could not be completely excluded due to the signal plane focus system. To make the final decision like a human expert, we also included a clinical-grade decision model after the prediction of the quantitative abnormalities.

The system was designed to flexibly grade the severity of lesions, just like making diagnostic decisions by different pathologists. To the best of our knowledge, the novel and noncostly labeling strategies, the handcraft feature incorporation into the deep-learning model, along with the aggregation methodology, have not been explicitly proposed in the previous deep-learning-based methods for quantitative diagnosis of cervical cytology. Its successful location of atypical or suspicious image patches can significantly reduce the tedious work of pathologists to find the comparably modest abnormal cells among numerous normal cells in a slide. Its performance demonstrates the potential of wide application in clinical practice.

Author contributions JK conceived and designed the system, performed the analysis, interpreted the results, and wrote the paper. YS constructed the mathematical models. YS and YL worked on the figures and graphs. YL and JD implemented the system and profiled the experiment results. JDW, YZ, QH, and FJ made the pathological annotation and diagnosed the slides. NJ and DW suggested the overall architecture. XL and JK collected the data.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval The data used in this study were collected from Shanxi Tumor Hospital, China. All experiments were conducted in accordance with the Ethical Guidelines for Shanxi Tumor Hospital. An ethics commitment from Shanxi Tumor Hospital granted this dataset to the corresponding author for this research on July 29, 2019, and the ethical approval ID was 201903.

References

1. Arbyn M, Weiderpass E, Bruni L, de Sanjosé S, Saraiya M, Ferlay J, et al. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. Lancet Glob Health. 2020;8: e191–203.
2. Nandini NM, Nandish SM, Pallavi P, Akshatha SK, Chandrashekhar AP, Anjali S, et al. Manual liquid based cytology in primary screening for cervical cancer—a cost effective preposition for scarce resource settings. Asian Pac J Cancer Prev. 2012;13:3645–51.
3. Gibb RK, Martens MG. The impact of liquid-based cytology in decreasing the incidence of cervical cancer. Rev Obstet Gynecol. 2011;4:S2.
4. Wilbur DC. Digital cytology: current state of the art and prospects for the future. Acta Cytol. 2011;55:227–38.
5. Khalbuss WE, Pantanowitz L, Parwani AV. Digital imaging in cytopathology. Patholog Res Int. 2011;2011:1–10.
6. Cibas ES, Ducatman BS. Cytology E-Book: diagnostic principles and clinical correlates. 4th ed. Elsevier Health Sci, Canda. 2013.
7. Nayar R, Wilbur DC. The Bethesda system for reporting cervical cytology: definitions, criteria, and explanatory notes. 3rd ed. Springer, Cham. 2015.
8. Carozzi F, Negri G, Sani C Molecular Cytology Applications on Gynecological Cytology. Springer, Cham. 2018;127–49.
9. Campanella G, Hanna MG, Geneslaw L, Miraflor AP, Silva VWK, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019;25:1301–9.
10. Kather JN, Krisam J, Charoentong P, Luedde T, Herpel E, Weis CA, et al. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. PLoS Med. 2019;16:e1002730.
11. Kather JN, Pearson AT, Halama N, Jager D, Krause J, Loosen SH, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med. 2019;25:1054–6.
12. Bejnordi BE, Veta M, Van Diest PJ, Van Ginneken B, Karssemeijer N, Litjens GJS, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA. 2017;318:2199–210.
13. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyo D, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature Med. 2018;24:1559–67.
14. Zhang L, Lu L, Nogues I, Summers RM, Liu S, Yao J. DeepPap: deep convolutional networks for cervical cell classification. IEEE J Biomed Health Inform. 2017;21:1633–43.
15. Song Y, Zhang L, Chen S, Ni D, Lei B, Wang T. Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning. IEEE Trans Biomed Eng. 2015;62:2421–33.
16. Byju NB, Sujathan VK, Malm P, Kumar RR. A fast and reliable approach to cell nuclei segmentation in pap stained cervical smears. CSI Trans on ICT. 2013;1:309–15.
17. Volerman A, Cifu A. Cervical cancer screening. JAMA. 2014;312:2279–80.
18. Song Y, Tan EL, Jiang X, Cheng JZ, Ni D, Chen S, et al. Accurate cervical cell segmentation from overlapping clumps in pap smear images. IEEE Trans Med Imaging. 2016;36:288–300.
19. Kuko M, Pourhomayoun M. Single and clustered cervical cell classification with ensemble and deep learning methods. Inf Syst Front. 2020;22:1039–51.

20. Tareef A, Song Y, Cai W, Huang H, Chang H, Wang Y, et al. Automatic segmentation of overlapping cervical smear cells based on local distinctive features and guided shape deformation. Neurocomputing. 2017;221:94–107.
21. Zhou Y, Chen H, Xu J, Dou Q, Heng PA. Irnet: instance relation network for overlapping cervical cell segmentation. MICCAI. 2019;LNCS(11764):640–8.
22. Bao H, Sun X, Zhang Y, Pang B, Li H, Zhou L, et al. The artificial intelligence-assisted cytology diagnostic system in large-scale cervical cancer screening: a population-based cohort study of 0.7 million women. Cancer Med. 2020;9:6896–906.
23. Cruzroa A, Gilmore H, Basavanhally A, Feldman M, Ganesan S, Shih N, et al. High-throughput adaptive sampling for whole-slide histopathology image analysis (HASHI) via convolutional neural networks: application to invasive breast cancer detection. PLoS ONE. 2018;13:e0196828.
24. Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH. Patch-based convolutional neural network for whole slide tissue image classification. CVPR. 2016;2424–33.
25. Aeffner F, Wilson K, Martin NT, Black JC, Hendriks CLL, Bolon B, et al. The gold standard paradox in digital image analysis: manual versus automated scoring as ground truth. Arch Pathol Lab Med. 2017;141:1267–75.
26. Rączkowski Ł, Możejko M, Zambonelli J, Szczurek E. ARA: accurate, reliable and active histopathological image classification framework with Bayesian deep learning. Sci. Reports. 2019;9: 1–12.
27. Abels E, Pantanowitz L, Aeffner F, Zarella MD, van der Laak J, Bui MM, et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. J Pathol. 2019;249:286–94.
28. Koliopoulos G, Nyaga VN, Santesso N, Bryant A, Martin-Hirsch PP, Mustafa RA, et al. Cytology versus HPV testing for cervical cancer screening in the general population. Cochrane Database Syst Rev. 2017;8:1–105.
29. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. CVPR. 2016;770–8.
30. Ronneberger OP, Fischer Brox T. U-net: convolutional networks for biomedical image segmentation. MICCAI. 2015;LNCS(9351): 234–41.
31. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? Adv Neural Inf Process Syst. 2014;2:3320–8.
32. Shen Y, Ke J. A deformable CRF model for histopathology whole-slide image classification. MICCAI. 2020;LNCS(12265): 500–8.
33. Tosun AB, Pullara F, Becich MJ, Taylor DL, Chennubhotla SC, Fine JL. HistoMapr^{TM}: An explainable AI (xAI) platform for computational pathology solutions. Patent App. 2020;LNCS (12090):204–7.
34. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. Unet++: a nested u-net architecture for medical image segmentation. Deep Learn Med Image Anal Multimodal Learn Clin Decis Support. 2018;11045:3–11.