scientific data

OPEN

DATA DESCRIPTOR

A large annotated cervical cytology images dataset for AI models to aid cervical cancer screening

Xuan Zhang[1], Jianxin Ji[1], Qi Zhang[1], Xiaohan Zheng[1], Kaiyuan Ge[1], Menglei Hua[1], Lei Cao[1] ✉ & Liuying Wang[2] ✉

Accurate detection of abnormal cervical cells in cervical cancer screening increases the chances of timely treatment. The vigorous development of deep learning methods has established a new ecosystem for cervical cancer screening, which has been proven to effectively improve efficiency and accuracy of cell detection in many studies. Although many contributing studies have been conducted, limited public datasets and time-consuming collection efforts may hinder the generalization performance of those advanced models and restrict further research. Through this work, we seek to provide a large dataset of cervical cytology images with exhaustive annotations of abnormal cervical cells. The dataset consists of 8,037 images derived from 129 scanned Thinprep cytologic test (TCT) slide images. Furthermore, we performed evaluation experiments to demonstrate the performance of representative models trained on our dataset in abnormal cells detection.

Background & Summary

Cervical cancer is the fourth most common cause of cancer incidence and mortality among women globally, while early diagnosis and treatment can help improve patient survival rates[1–3]. Cytology-based screening using pap smears or liquid-based preparation slides is a central requirement for the early diagnosis of cervical cancer[4,5]. For this, pathologists find out abnormal cervical cells under a microscope or in digital cytology images, and then issues a final report based on the results. However, this process is subject to high intra- and inter-pathologist variability, which can lead to high false-negative rates in routine diagnoses[6]. Additionally, since abnormal cervical cells usually account for only a small portion of all the sample cells, manual investigation leads to unnecessary waste of medical resources. Consequently, the computerized screening of abnormal cervical cells in digital cytology images is a relevant topic of ongoing scientific interest.

Since the development of deep learning[7], AI (Artificial intelligence)-powered cervical cancer screening systems have emerged, driving a wave of changes in cervical cancer diagnosis. These systems mainly based on modern object detection techniques of deep learning, and provide new automatic detection methods of abnormal cervical cells through domain-specific improvements. For instance, Li *et al.* adopted Faster R-CNN to detect and classify cervical exfoliated cells in the early diagnosis of cervical cancer[8]. Xiang *et al.* further cascaded a task-specific classifier on YOLOv3 to improve the classification performance and smooth cervical cell dataset distribution to weaken the influence of noisy labels[9]. Similarly, Ma *et al.* proposed a mask abnormal cell detection model based on Mask R-CNN, and used a fixed proposal module to generate fixed-sized feature maps[10]. The work by Liang *et al.* classified the proposals by comparing with the reference samples of each category thus circumvent the problem of the limited data in cervical abnormal cell detection[11]. Chen *et al.* developed a dynamic comparing module and an instance contrastive loss to imitate clinical diagnosis process of normal-abnormal cells comparing, effectively improving the efficiency of detecting abnormal cervical cells[12]. Another work by Liang *et al.* explored contextual relationships using RoI-relationship attention module (RRAM) and global RoI attention module (GRAM) to improve the performance of cervical abnormal cell detection[13]. These methods demonstrate the potential advantages of computer-assisted abnormal cervical cell screening, such as boosting sensitivity, and reducing the risk of misdiagnosis.

[1]Department of Biostatistics, School of Public Health, Harbin Medical University, Harbin, 150081, China. [2]Department of Health Management, Harbin Medical University, Harbin, 150081, China. ✉e-mail: caolei@hrbmu.edu.cn; wangliuying@hrbmu.edu.cn

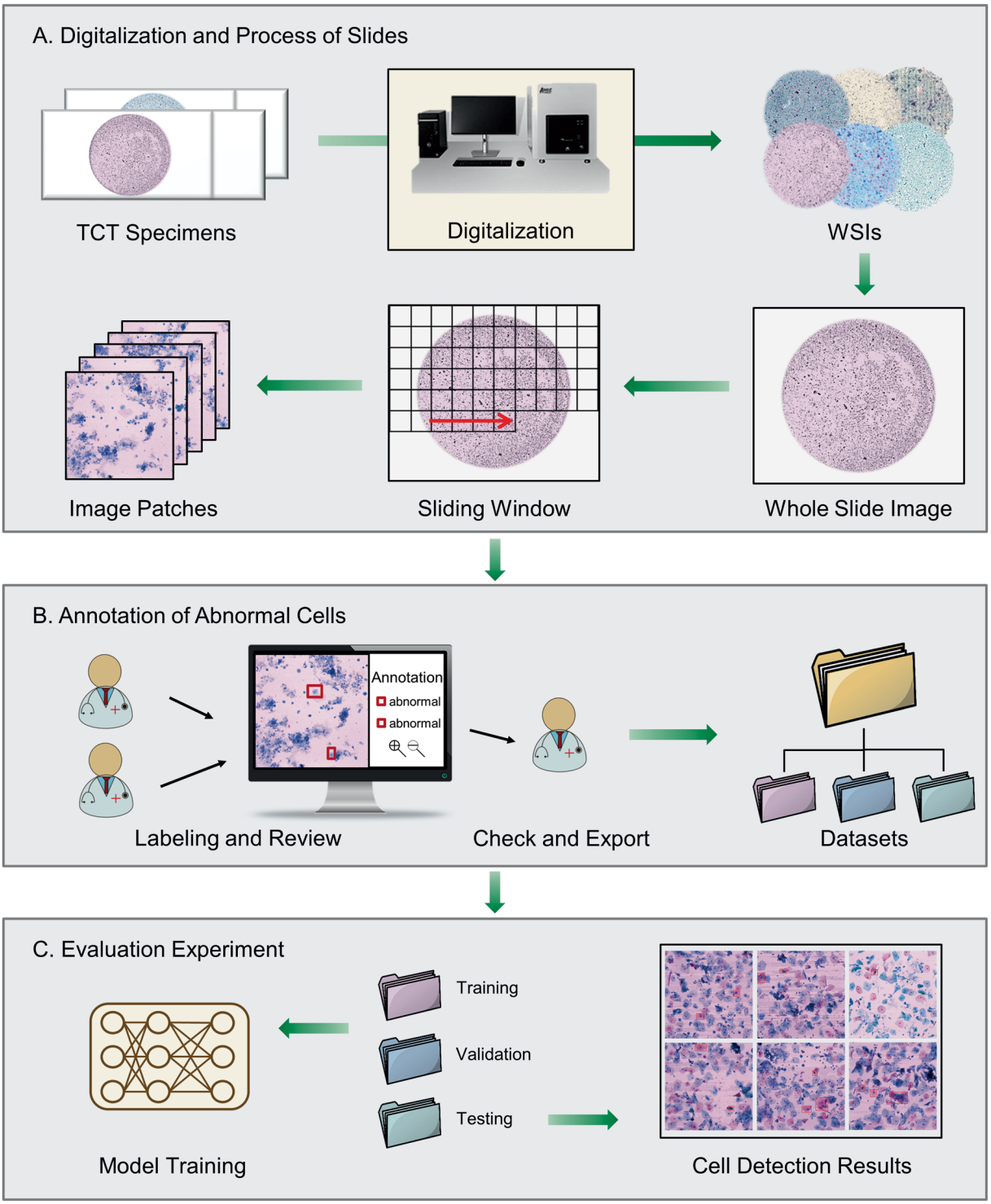

Fig. 1 Workflow of data generation and evaluation experiment. (**A**) The TCT specimens were digitalized into whole slide images (WSIs) and divided into 2048 × 2048 pixel patches. (**B**) All patches were labelled, reviewed by two pathologists, and finally checked by an experienced pathologists to finish annotation of abnormal cells. (**C**) Several representative detection models were adopted to validate our datasets. The experiment is performed by splitting the datasets, training the model and evaluating the prediction results.

Although these methods have demonstrated significant advantages, their development heavily relies on large amounts of annotated data. Consequently, substantial human and material resources were required for the collection of cervical cytology images at the outset of these studies. The main reason is the scarcity of publicly available image datasets containing annotated abnormal cervical cells. This situation not only prolongs the research

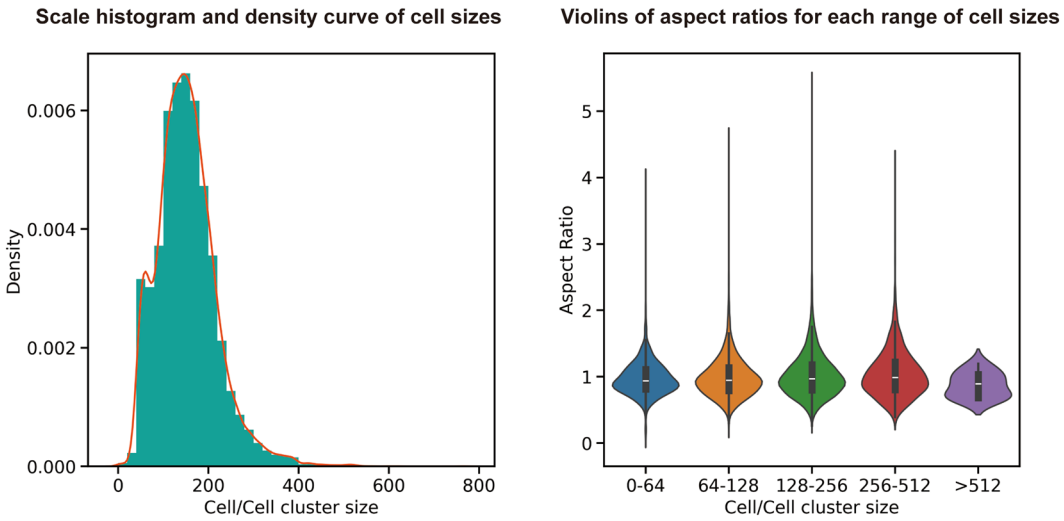

Fig. 2 Statistical description of annotated abnormal cells in dataset.

Method	$AP_{50:95}$	AP_{50}	AP_{75}	$AR_{50:95}$	F1-score
SSD	10.8	24.1	7.2	14.3	40.1
Retina Net	25.7	54.8	20.3	34.4	66.5
FCOS	27.6	61.6	21.2	37.7	68.9
Faster R-CNN	31.5	67.4	25.0	45.2	58.9
Cascade R-CNN	29.1	57.7	27.0	39.2	66.4
Sparse R-CNN	23.2	50.1	19.0	32.0	65.8
YOLOv3	9.1	28.3	2.8	17.6	46.3
YOLOv7	13.3	37.3	5.4	22.6	55.3
DETR	19.4	47.1	12.2	36.5	43.7

Table 1. Evaluation results of each model for abnormal cells detection.

cycle but also limits the generalization of models in clinical practice. Performance may be inconsistent when dealing with images derived from different centres, instruments, or staining techniques. In this work, we present and describe a large cervical cytology image dataset annotated with abnormal cells. Our goal is to provide a valuable public resource to optimize the development of abnormal cervical cell identification models. This dataset has the potential to enhance the generalization performance of existing models and open up new avenues for future research.

Methods

The following section describes the sample collection and preparation for the specimens included in the presented dataset. Furthermore, we elaborate on data annotation and the methods used for validating the presented dataset.

Ethics approval and consent to participate. This work obtained approval of ethics committee with number of 2022ZFYJ295-01. Relevant data for this work were acquired from Heilongjiang Maternal and Child Health Hospital (HMCHH) through a collaboration. The original data collection was approved by the institutional review board of HMCHH and adhered to the principles outlined in the Declaration of Helsinki. A waiver of Informed patient consent was granted by HMCHH as all samples were irreversibly anonymized by the institutional review board.

Specimen preparation and digitalization. The dataset contains samples from patients who underwent cervical cytology examination at Heilongjiang Maternal and Child Health Hospital between October 2018 and May 2019. According to the examination reports, a total of 129 TCT (Thinprep cytologic test) slides reported as abnormal levels were collected from the pathology department into this work (see Ethics approval and consent to participate). As shown in Fig. 1, each slide was digitized and divided into 333 non-overlapping patches (2048×2048 pixels) at 20x objective magnification using an Olympus BX53 optical microscope. Image patches with low information content, such as those covered by background or blurred patches, were removed. The remaining 8,037 patches were stored as cytology images in the dataset in.png file format.

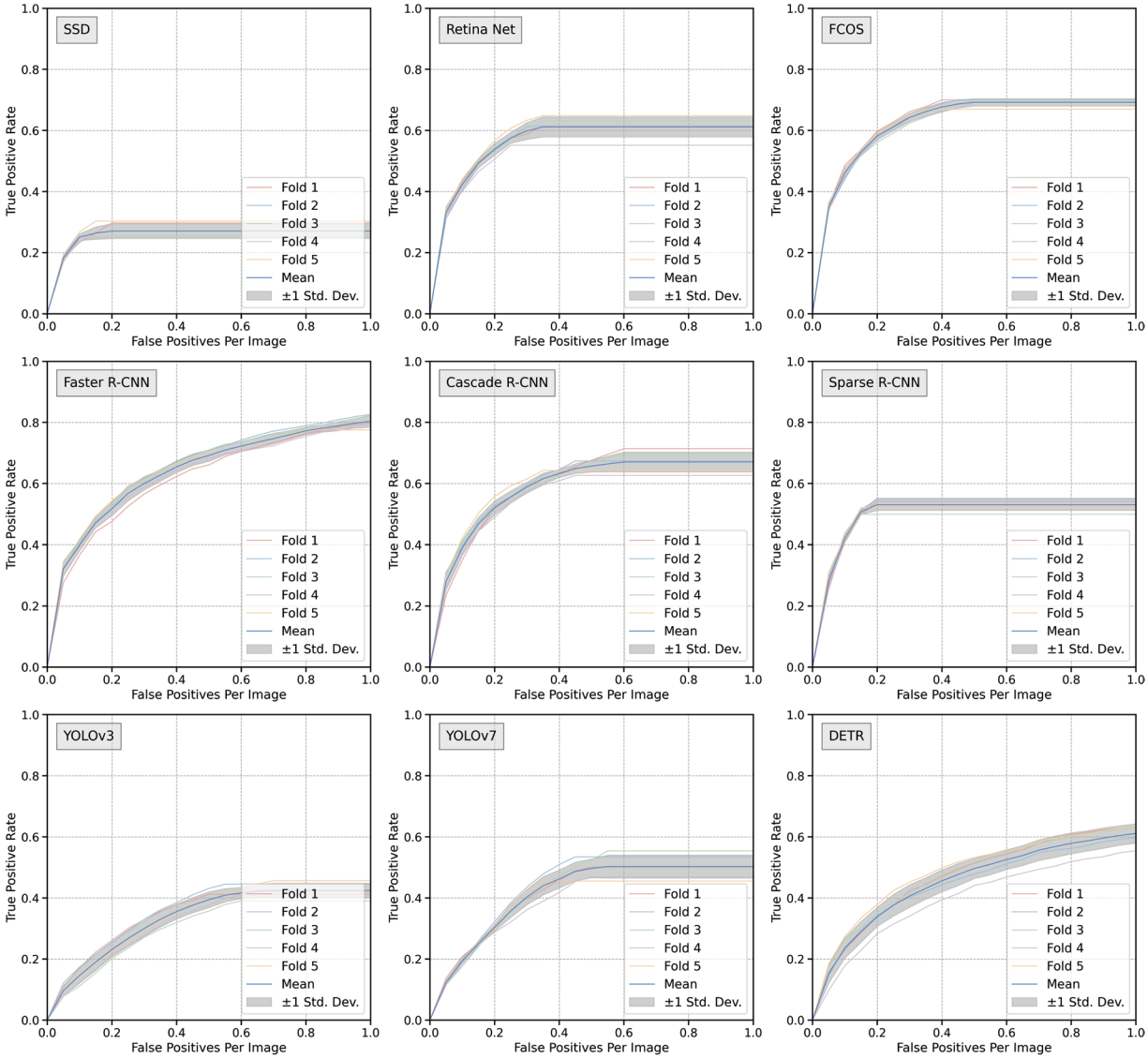

Fig. 3 FROC curves of detection models.

Annotation process. To obtain ground-truth annotations of abnormal cells, three pathologists were involved in the annotation preparation of this dataset, with the aim of making exhaustive annotations of abnormal cervical cells in each patch. We referred to three pathologists as A, B, and C; A had about 33 years of experience in reading cervical cytology images, while reader B and C had about 10 years of experience. The abnormality or normality of a cervical cell was defined according to the ACOGs guidance[14]. Based on the guidance, the readers draw bounding boxes around abnormal cells in each image patch using the Colabeler tool (http://www.jingling-biaozhu.com/) as shown in Fig. 1B.

The generation of the final annotation file follows three steps: the initial labelling step, the verification step, and the final check step. An image was firstly randomly assigned to reader B or C. Once the labelling was finished, the image and annotation were then passed on to another reader for review. Finally, the annotations were checked and exported by reader A.

All annotation files are stored in.xml file format and maintain the same number as the corresponding images.

Evaluation methods. To validate the dataset proposed in this study, we used several previously published representative detection models: the two-stage architecture Faster R-CNN[15], Cascade R-CNN[16], Sparse R-CNN[17], the one-stage architecture SSD[18], Retina Net[19], FCOS[20] and end-to-end architecture YOLOv3[21], YOLOv7[22], DETR[23] to perform the detection analysis of abnormal cervical cells. Initially, we randomly selected 20% of all patients as the testing subset. After that, the remaining patients were randomly divided into five folds to construct the training subset and validation subset for five-fold cross-validation. Note that all image patches corresponding to the same patient enter the same subset in this process.

In each fold, we trained for 30 epochs using training subset and retained the model that performed best on the validation subset during training process. Then, the testing subset was held out and used for model

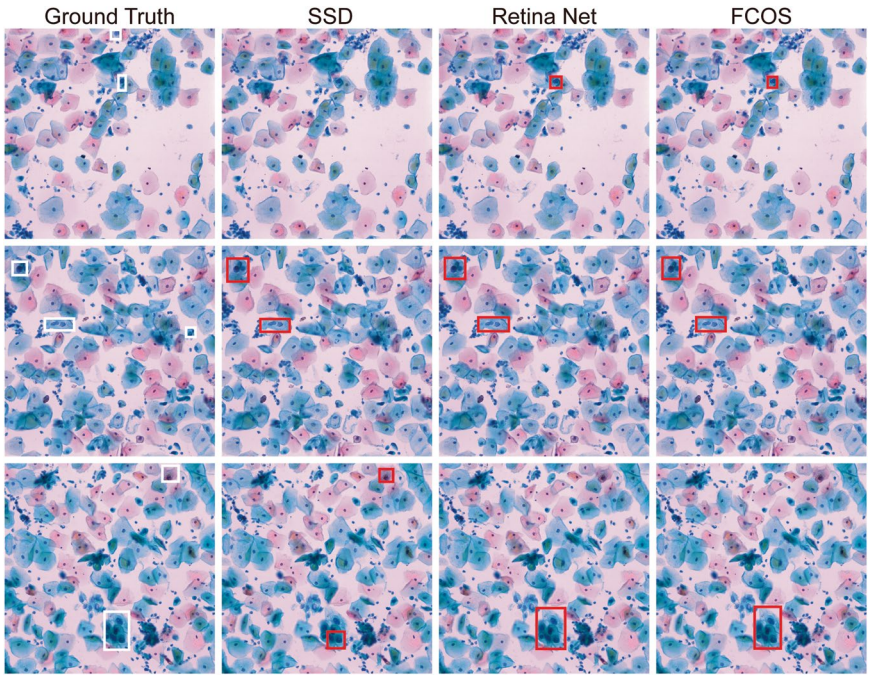

Fig. 4 Illustration of the predicted detection results using one-stage models after training.

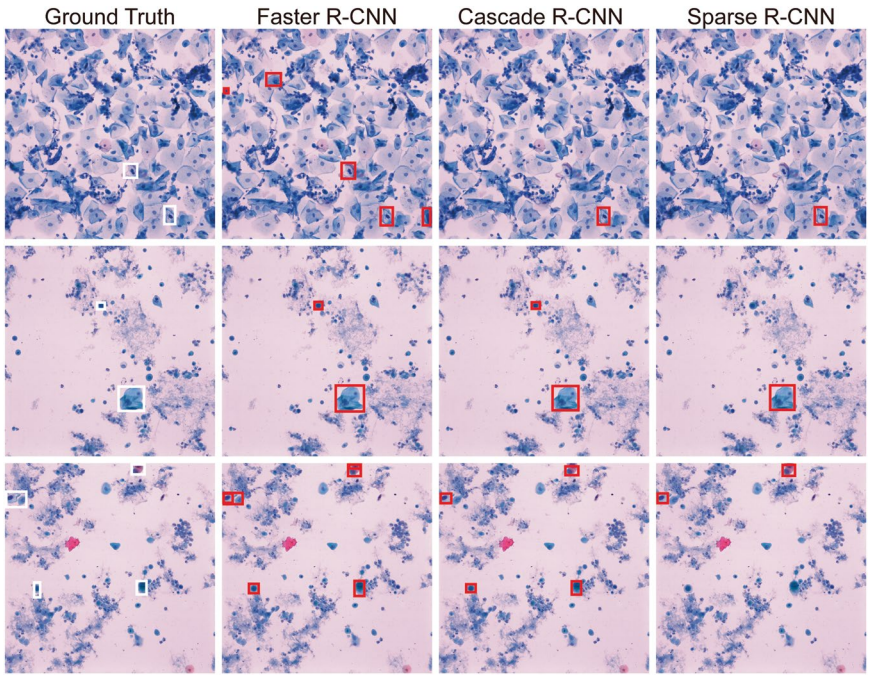

Fig. 5 Illustration of the predicted detection results using two-stage models after training.

evaluation when training was completed. The model parameters were updated via AdamW[24] optimizer with a batch size of 8, while the initial learning rate (lr0) was 2×10^{-5} with the weight decay of 1×10^{-4}. To avoid overfitting, the cosine annealing[25] algorithm was used to adjust the learning rate change.

Data augmentation was also adopted to overcome the overfitting. In our experiment, we performed random horizontal flip, vertical flip, rotation, brightness change, gray scale, and gaussian blur. These affine transformations helped the model to have a better understanding of the input image since it viewed the images in many transformed views. The input images were rotated randomly by 90, 180, 270 degrees. The range between [0.8, 1.5] was used for brightness change, and a sigma value between [0, 5] was used for the gaussian blur.

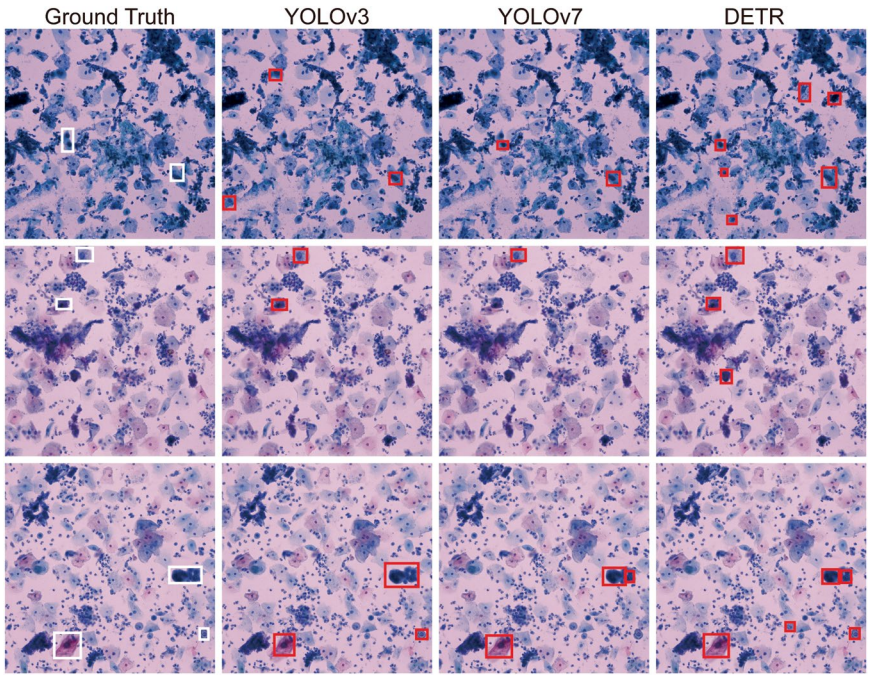

Fig. 6 Illustration of the predicted detection results using end-to-end models after training.

Data Records

The complete dataset, named HMCHH-TCT-CellDet, is provided on figshare for public non-restricted access[26]. It consists of two components: a folder called "JPEGImages" contains cytology images from the corresponding TCT slides in.png file format, and a folder called "Annotations", which contains annotation files of abnormal cells for each cytology image in.xml file format. The following section provides an overview of the presented dataset including the number of annotated instances and the size distribution of abnormal cells.

Overall description. There are a total of 8,037 images of 2048 × 2048 pixels in the "JPEGImages" folder, and the same number of annotated files in the "Annotations" folder. The naming format of image files and annotation files is consistent, that is, "patient number_image number", which facilitates the split of datasets when training the model.

First, according to the count, there are a total of 15,761 annotation boxes of abnormal cells in the dataset, that is, an image contains an average of two abnormal cells. Additionally, we used area of annotation boxes (in pixels) to measure cell size while aspect ratio (i.e., height divided by width) to represent cell shape. As shown in Fig. 2, the scale histogram and density curve shown distribution of cell sizes while each violin shown the distribution of aspect ratios for a specific range of cell sizes. These statistics may be helpful in selecting the architecture of the detection model and setting hyperparameters such as size of anchor boxes.

Technical Validation

To evaluate detection performance of those representative models, we used the COCO-style[27] average precision (AP), average recall (AR) and the free-response receiver operating characteristic (FROC) analysis[28] in our evaluation experiments. We adopted the Darknet[21] as backbone network for YOLO-based models and VGG[29] for SSD to comply with original method, while all other models adopted ResNet50[30]. In testing process, the bounding boxes whose prediction probability is greater than the threshold 0.5 was retained as the final prediction results. The following section summarizes the performance results of our evaluation experiments.

Cell detection results. Table 1 and Fig. 3 displayed the evaluation results of different detection models. Note that all these results are mean values of the five trained models on the testing subset. Different model paradigms shown differences in detection performance, mainly due to differences in the model structures and proposals strategies. The overall performance of two-stage models is higher than other models, while the performance of dense anchor boxes is higher than that of learnable anchor boxes. Among the models adopted, Faster R-CNN achieved the best performance, as AP_{50} of 67.4%, AP_{75} of 25.0% and $AR_{50:95}$ of 45.2%, respectively. However, the F1-score of Faster R-CNN has declined compared to other dense anchor boxes strategy models, possibly due to the generation of more false positive boxes and the lack of correction modules like the Cascade R-CNN.

In Fig. 3, the FROC curves shown the performance of each detection models respectively, which represented the sensitivity achieved by the models as the number of false positive prediction boxes per image increases. When the number of false positive prediction boxes per image is 1, Faster R-CNN still achieved best sensitivity of 80.4% on average.

Detection visualization. The evaluation experiment also included visualized results to qualitatively demonstrate the performance of the models in detecting abnormal cells. The instances used for demonstration were taken from several randomly selected images from the testing subset, as shown in Figs. 4–6.

Usage Notes

The xml annotation files in the dataset can be processed using the parsing function provided in the code, thus a csv file containing all annotations can be generated. Once the dataset is divided, the images and annotations can be passed into models using the provided custom data loaders. The subsequent training and evaluation process can be reproduced by the provided code, while other customized models can also be plug-and-played.

Code availability

The code used in this study was written in Python3 and is available at GitHub (https://github.com/zx333445/TCT_data). The code is based on PyTorch (version 2.0.0) and provides complete training and evaluation processes for all representative models. In addition, the split csv files generated by the five-fold cross-validation were also uploaded to above code repository to facilitate researchers to conduct evaluation experiment.

Received: 28 August 2024; Accepted: 1 January 2025;
Published online: 07 January 2025

References

1. Bray, F. *et al*. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **68**, 394–424, https://doi.org/10.3322/caac.21492 (2018).
2. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer Statistics, 2021. *CA: a cancer journal for clinicians* **71**, 7–33, https://doi.org/10.3322/caac.21654 (2021).
3. Boulet, G. A., Horvath, C. A., Berghmans, S. & Bogers, J. Human papillomavirus in cervical cancer screening: important role as biomarker. *Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **17**, 810–817, https://doi.org/10.1158/1055-9965.Epi-07-2865 (2008).
4. Hashmi, A. A. *et al*. Comparison of Liquid-Based Cytology and Conventional Papanicolaou Smear for Cervical Cancer Screening: An Experience From Pakistan. *Cureus* **12**, e12293, https://doi.org/10.7759/cureus.12293 (2020).
5. Curry, S. J. *et al*. Screening for Cervical Cancer: US Preventive Services Task Force Recommendation Statement. *Jama* **320**, 674–686, https://doi.org/10.1001/jama.2018.10897 (2018).
6. Branca, M. & Longatto-Filho, A. Recommendations on Quality Control and Quality Assurance in Cervical Cytology. *Acta cytologica* **59**, 361–369, https://doi.org/10.1159/000441515 (2015).
7. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444, https://doi.org/10.1038/nature14539 (2015).
8. Li, X. & Li, Q. Detection and classification of cervical exfoliated cells based on faster R-CNN. *2019 IEEE 11th international conference on advanced infocomm technology (ICAIT)*, 52–57 (2019).
9. Xiang, Y. *et al*. A novel automation-assisted cervical cancer reading method based on convolutional neural network. *Biocybernetics Biomedical Engineering* **40**, 611–623 (2020).
10. Ma, B., Zhang, J., Cao, F. & He, Y. MACD R-CNN: an abnormal cell nucleus detection method. *IEEE Access* **8**, 166658–166669 (2020).
11. Liang, Y. *et al*. Comparison detector for cervical cell/clumps detection in the limited data scenario. *Neurocomputing* **437**, 195–205, https://doi.org/10.1016/j.neucom.2021.01.006 (2021).
12. Chen, T. *et al*. A task decomposing and cell comparing method for cervical lesion cell detection. *IEEE Transactions on Medical Imaging* **41**, 2432–2442 (2022).
13. Liang, Y. *et al*. Exploring contextual relationships for cervical abnormal cell detection. *IEEE Journal of Biomedical Health Informatics*, (2023).
14. Randel, A. ACOG Releases Guideline on Cervical Cancer Screening. *American Family Physician* **88**, 776–777 (2013).
15. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, 91–99 (2015).
16. Cai, Z. & Vasconcelos, N. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *IEEE transactions on pattern analysis and machine intelligence* **43**, 1483–1498, https://doi.org/10.1109/tpami.2019.2956516 (2021).
17. Sun, P. *et al*. Sparse r-cnn: End-to-end object detection with learnable proposals. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14454–14463 (2021).
18. Berg, A. C. *et al*. SSD: Single Shot MultiBox Detector. *Computer Vision – ECCV 2016*, https://doi.org/10.1007/978-3-319-46448-0_2 (2015).
19. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision*, 2980–2988 (2017).
20. Tian, Z., Shen, C., Chen, H. & He, T. Fcos: Fully convolutional one-stage object detection. *Proceedings of the IEEE/CVF international conference on computer vision*, 9627–9636 (2019).
21. Farhadi, A. & Redmon, J. Yolov3: An incremental improvement. *Computer vision and pattern recognition* **1804**, 1–6 (2018).
22. Wang, C.-Y., Bochkovskiy, A. & Liao, H.-Y. M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7464–7475 (2023).
23. Carion, N. *et al*. End-to-end object detection with transformers. *European conference on computer vision*, 213–229 (2020).
24. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *International Conference on Learning Representations* (2019).
25. Loshchilov, I. & Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. *International Conference on Learning Representations* (2017).
26. Zhang, X. *et al*. A large annotated cervical cytology images dataset for AI models to aid cervical cancer screening. *figshare* https://doi.org/10.6084/m9.figshare.27901206 (2024).
27. Lin, T.-Y. *et al*. Microsoft COCO: Common Objects in Context. *Computer Vision – ECCV 2014*, 740–755, https://doi.org/10.1007/978-3-319-10602-1_48 (2014).
28. Chakraborty, D. P. & Winter, L. H. Free-response methodology: alternate analysis and a new observer-performance experiment. *Radiology* **174**, 873–881, https://doi.org/10.1148/radiology.174.3.2305073 (1990).
29. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR 2015)*, 1–14 (2015).
30. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *IEEE* (2016).

Acknowledgements
The study was supported by National Natural Science Foundation of China (NSFC) (82304250, 82273734).

Author contributions
L.C. and L.W. conceived of the project, X.Z., J.J., K.G and Q.Z. contributed to the preprocess of the data, J.J., M.H. and X.Z. participated in workstation environment deployment, X.Z., J.J. and X.Z. designed the evaluation experiments and constructed the models, X.Z. and L.C. wrote the manuscript. All the authors revised the paper.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to L.C. or L.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.