

# BMJ Open Automatic deep learning-based colorectal adenoma detection system and its similarities with pathologists

Zhigang Song,<sup>1</sup> Chunkai Yu,<sup>2</sup> Shuangmei Zou,<sup>3</sup> Wenmiao Wang,<sup>3</sup> Yong Huang,<sup>1</sup> Xiaohui Ding,<sup>1</sup> Jinhong Liu,<sup>1</sup> Liwei Shao,<sup>1</sup> Jing Yuan,<sup>1</sup> Xiangnan Gou,<sup>1</sup> Wei Jin,<sup>1</sup> Zhanbo Wang ,<sup>1</sup> Xin Chen,<sup>1</sup> Huang Chen,<sup>4</sup> Cancheng Liu,<sup>5</sup> Gang Xu,<sup>6</sup> Zhuo Sun,<sup>5</sup> Calvin Ku,<sup>5</sup> Yongqiang Zhang,<sup>1</sup> Xianghui Dong,<sup>1</sup> Shuhao Wang ,<sup>5,7</sup> Wei Xu,<sup>7</sup> Ning Lv,<sup>3</sup> Huaiyin Shi<sup>1</sup>

**To cite:** Song Z, Yu C, Zou S, *et al.* Automatic deep learning-based colorectal adenoma detection system and its similarities with pathologists. *BMJ Open* 2020;**10**:e036423. doi:10.1136/bmjopen-2019-036423

► Prepublication history for this paper is available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2019-036423>).

Received 19 December 2019  
Revised 04 July 2020  
Accepted 07 August 2020



© Author(s) (or their employer(s)) 2020. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

## Correspondence to

Dr. Shuhao Wang;  
ericwang@tsinghua.edu.cn,  
Prof. Huaiyin Shi;  
shihuiyin@sina.com and  
Prof. Shuangmei Zou;  
zousm@cicams.ac.cn

## ABSTRACT

**Objectives** The microscopic evaluation of slides has been gradually moving towards all digital in recent years, leading to the possibility for computer-aided diagnosis. It is worthwhile to know the similarities between deep learning models and pathologists before we put them into practical scenarios. The simple criteria of colorectal adenoma diagnosis make it to be a perfect testbed for this study.

**Design** The deep learning model was trained by 177 accurately labelled training slides (156 with adenoma). The detailed labelling was performed on a self-developed annotation system based on iPad. We built the model based on DeepLab v2 with ResNet-34. The model performance was tested on 194 test slides and compared with five pathologists. Furthermore, the generalisation ability of the learning model was tested by extra 168 slides (111 with adenoma) collected from two other hospitals.

**Results** The deep learning model achieved an area under the curve of 0.92 and obtained a slide-level accuracy of over 90% on slides from two other hospitals. The performance was on par with the performance of experienced pathologists, exceeding the average pathologist. By investigating the feature maps and cases misdiagnosed by the model, we found the concordance of thinking process in diagnosis between the deep learning model and pathologists.

**Conclusions** The deep learning model for colorectal adenoma diagnosis is quite similar to pathologists. It is on-par with pathologists' performance, makes similar mistakes and learns rational reasoning logics. Meanwhile, it obtains high accuracy on slides collected from different hospitals with significant staining configuration variations.

## INTRODUCTION

Computer-aided pathological diagnosis is becoming possible with the microscopic evaluation of slides has been gradually moving towards all digital in recent years. In the past 10 years, researchers have proposed various medical diagnosis systems using deep learning.<sup>1–7</sup> Deep learning has been widely studied in the field of object detection<sup>8–12</sup> and semantic segmentation.<sup>13 14</sup> Different

## Strengths and limitations of this study

- To study the similarities between deep learning models and pathologists before, we put them into practical scenarios, we used colorectal adenoma diagnosis as a testbed and established a semantic segmentation model for colorectal adenomas diagnosis using a deep convolutional neural networks.
- The deep learning model had achieved an area under the curve of 0.92 and obtained a slide-level accuracy of over 90% on the slides from two other hospitals.
- The performance of the deep learning model was on par with experienced pathologists.
- By investigating the feature maps and cases misdiagnosed by the model, we found the concordance of thinking process in diagnosis between the deep learning model and the pathologists.
- The current model was not at clinical grade due to the limited size of the training dataset. We need to include more types of adenomas in the training process and further improve the model performance.

from traditional machine learning methods, deep convolutional neural networks (CNNs) can learn directly from raw medical images, avoiding the feature engineering procedure and learn key features during the model training process automatically.<sup>15</sup>

The ability to interpret and elaborate histological features is crucial for artificial intelligence-powered medical diagnosis systems. Before applying deep learning under practical scenarios, we need to address the following non-trivial issues to understand the similarities between models and pathologists. The first and foremost question is whether the deep learning model can perform as good as pathologists. Second, as different hospitals operate under various staining configurations, the generalisation ability should be an important consideration when building the

systems. Third, we want to know when the deep learning model would make mistakes and whether they would be similar to pathologists. Lastly, the parameters of the model should be visualisable to enable interrogation of its reasoning logics.

It is estimated that more than 50% of western people may suffer from colorectal adenoma during their lifetime, among which 5%–16% develop to colorectal cancer (CRC).<sup>16–18</sup> The diagnosis and removal of these adenomas through colonoscopy lead to a reduction in the expected incidence of CRC, and individualised surveillance strategy for patients will be made according to the histological diagnosis of the resection specimen.<sup>19 20</sup> Analysing H&E-staining slides of colorectal adenoma is easier compared with CRC, making it to be a perfect testbed to understand deep learning models.

In this research, we established a semantic segmentation model for diagnosis of colorectal adenomas using a deep CNN and achieved an area under the curve (AUC) of 0.92, which was on par with the performance of experienced pathologists. The deep learning model achieved a slide-level accuracy of over 90% on the slides from two other hospitals. By investigating the cases misdiagnosed and feature maps of the model, we found a concordant thinking process in diagnosis between the deep learning model and the pathologists.

## METHODS

### Data construction

With the popularity of colonoscopy, the number of colorectal pathological slides occupied a large workload in pathology departments. All the histological colorectal slides used in this study were obtained as a part of surveillance colonoscopy. To effectively train the proof-of-concept deep CNN, we had collected a total of 411 slides from Chinese People's Liberation Army (PLA) General Hospital (PLAGH), of which 232 were diagnosed as colorectal adenomas, and 179 were normal mucosa or chronic inflammation which were categorised

as non-neoplasm. We selected 177 cases for the training set, 40 cases for validation and 194 cases as test samples. To further test the generalisation ability of the model, we had also collected 168 slides from two other hospitals, including China-Japan Friendship Hospital (CJFH) and Cancer Hospital, Chinese Academy of Medical Sciences (CH), composing the external test group. The detailed configuration of the datasets was shown in table 1. All slides were digitalised using KF-PRO-005 scanner (KFBio) with  $\times 40$  objective (eyepiece magnification fixed as  $\times 10$ ). Different from the traditional way of viewing slides on a microscope with fixed objectives, the whole slide images (WSIs) can be viewed at arbitrary levels via digital zooming.

The detailed labelling was prepared on 156 training and 20 validation slides containing adenomas using a self-developed annotation system based on iPad, by qualified pathologists. When adopting a rigorous definition that an adenomatous case contains one or more adenomatous glands, the diagnosis became remarkably subjective even between experienced pathologists. Therefore, a three-stage procedure was devised which included initial labelling, further verification, and the final expert check. Slides were first allocated to a pathologist, chosen randomly. When the labelling was finished, the slides along with annotations were then passed on to another randomly chosen pathologist for review. Finally, the senior pathologists spot-checked the slides that had passed the second reviewing stage. We were able to achieve a much better quality of training dataset using this elaborately designed labelling procedure.

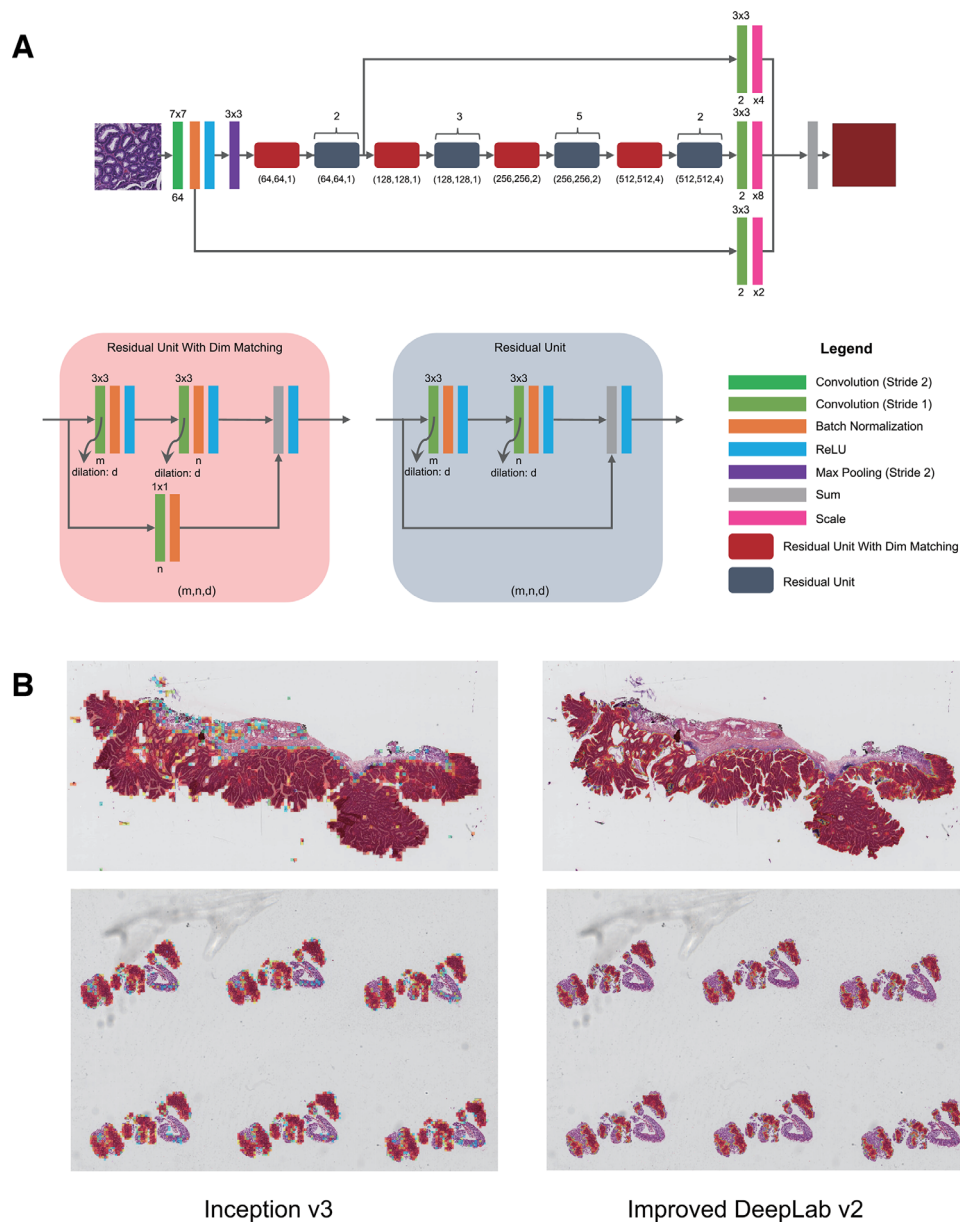
When preparing training and validation sets, the background areas of the slide were filtered out using the Otsu's method.<sup>21</sup> Then the slides were split into tiles with a stride of half of the tile size to form the training and validation data. For different field of views (FoVs), the tile number ranged from 203 212 to 2 265 945. Specifically, for the best performing model trained with  $\times 10$  FOV, we used a total of 113 090 adenomatous tiles and 90 122 normal ones for training.

**Table 1** Data distribution, where T, V, TV, H, L represent tubular, villous, tubulovillous, high grade, low grade, respectively

	Subtype	Grade	PLAGH (train)	PLAGH (validation)	PLAGH (test)	CJFH	CH
Adenoma	T	H	10	5	0	8	13
		L	151	5	56	43	58
	V	H	11	5	0	5	11
		L	28	5	2	3	46
	TV	H	10	0	0	5	11
		L	24	0	2	3	45
Non-neoplasm	–	–	21	20	138	13	44
Total	–	–	177	40	194	63	105

The counter is incremented by one when the slide contains a certain component.

CH, Cancer Hospital, Chinese Academy of Medical Sciences; CJFH, China-Japan Friendship Hospital; PLAGH, Chinese People's Liberation Army General Hospital.



**Figure 1** (A) Deep neural network structure; (B) predictions of both classification and segmentation models.

### Deep learning model

We built the model<sup>19</sup> based on DeepLab v2 with ResNet-34, which is illustrated in [figure 1A](#), with improvements. We introduced a skip layer fusion approach, in which we combined the upsampled lower layers with the higher layers to retain finer details containing semantic information. We also compared the performance of the improved DeepLab v2 against ResNet-50, DenseNet, Inception v3, U-Net, and DeepLab v3.

Since histological slides have no specific directions, we applied random rotating and mirroring to augment the training data. We used carefully designed data augmentation instead of stain normalisation during a training. Since histopathological slides have no specific orientation, we applied random rotations by 90°, 180° and 270° and random flips (horizontal and vertical) to the training patches. To boost the model stability for WSIs collected

from different hospitals, we also applied random scaling from  $\times 1.0$  to  $\times 1.5$ , Gaussian and motion blurs and colour jittering in brightness (0.0–0.2), saturation (0.0–0.25), contrast (0.0–0.2) and hue (0.0–0.04).

All models were trained and tested with TensorFlow on an Ubuntu server with 4 Nvidia GTX1080Ti graphics processing units (GPUs). The Adam optimiser with a fixed learning rate of 0.0001 was used to train the models. The batch size was set to 80 (20 on each GPU) and the training process was stopped after 25 epochs.

### Model test

A benefit from the fully CNN architecture is that the tile sizes during training and at inference need not be identical. In the inference stage, we cut the WSI into tiles with the size of 2000 $\times$ 2000 pixels. To further retain the environment information for the surrounding areas,

we adopted the overlap-patch approach<sup>22</sup> by feeding a 2200×2200 pixel tile into the model but only used the centred 2000×2000 pixel area for the final prediction.

We used the 15th largest pixel-level probability for slide-level prediction. The receiver operating characteristic (ROC) curve was derived by applying slide-level thresholding to the probability.

### Evaluation metrics

We chose three evaluation metrics to describe the model performance

$$\text{accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FN} + \text{FP} + \text{TN}),$$

$$\text{sensitivity} = \text{TP} / (\text{TP} + \text{FN}),$$

$$\text{specificity} = \text{TN} / (\text{TN} + \text{FP}),$$

where TP, FP, TN, FN stood for true positive, false positive, true negative and false negative, respectively. The accuracy represented the ratio of the number of correctly predicted slides to the total number of slides. The sensitivity/specificity indicated the proportion of adenomatous/normal slides that were correctly identified. The statistics were made using self-developed Python scripts and plotted by Matlab.

### Model interpretability

Interpretability had been an issue to be considered in applying deep learning in medical practice. The deep CNNs were often described as black boxes, making it difficult to be applied for clinical use. This obstacle could be tackled by regarding the model either as a black box functional module or as a white box. From the black box perspective, we could study its input–output behaviour and compare it with expert pathologists. Meanwhile, we could also analyse the false predictions and compare them with mistakes made by pathologists. On the white box perspective, we could open the model and try to visualise what it has learnt. One of the most effective approaches of model visualisation is to output feature maps learnt by the CNN and infer what its reasoning logics look like. We visualised feature maps<sup>23</sup> to understand how the input samples went through the CNN. We normalised all the visualisation results to the range (0.0–1.0) according to the maximum and minimum values of all the feature maps derived from the corresponding CNN layer.

## RESULTS

### Performance of different deep learning models

In table 2, we gave the performance of six models trained and validated with 320×320 pixel patches under ×20 FoV. The improved DeepLab v2 outperformed both classification and segmentation models. Moreover, the segmentation model reveals more interpretable predictions, as shown in figure 1B. In the following, we chose the improved DeepLab v2 as the research object.

**Table 2** Performance of different deep learning models

Model	Accuracy, %
ResNet-50	89.8
DenseNet	87.7
Inception v3	90.3
U-Net	77.7
DeepLab v3	88.3
Improved DeepLab v2	90.4

### Comparison of models trained with different FoVs

We trained six models using ×10, ×20 and ×40 FoV tiles with sizes of 640×640 and 320×320 pixels, as illustrated in figure 2A. It can be easily seen that ×10 FoV captured glandular structure and gland–stromal relationships better than other smaller FoVs. In addition, with the help of larger tile size, the model trained with ×10 FoV and 640×640-pixel tile size outperformed others on the validation set, shown in figure 2B.

The computing speed was another important factor to be considered. It was worth investigating the prediction time under different FoVs. Since our deep learning model was fully convolutional, it was possible to make predictions for arbitrary sizes of input images. In our research, we fixed the tile size in the prediction stage to 2000×2000 pixels. The inference time of different FoVs was given in figure 2C, all numbers were normalised by the time taken at ×40. We could see that at ×10 we got both better accuracy and higher computing speed. The final diagnostic system developed with the trained deep learning model was demonstrated in the (online supplemental file 1).

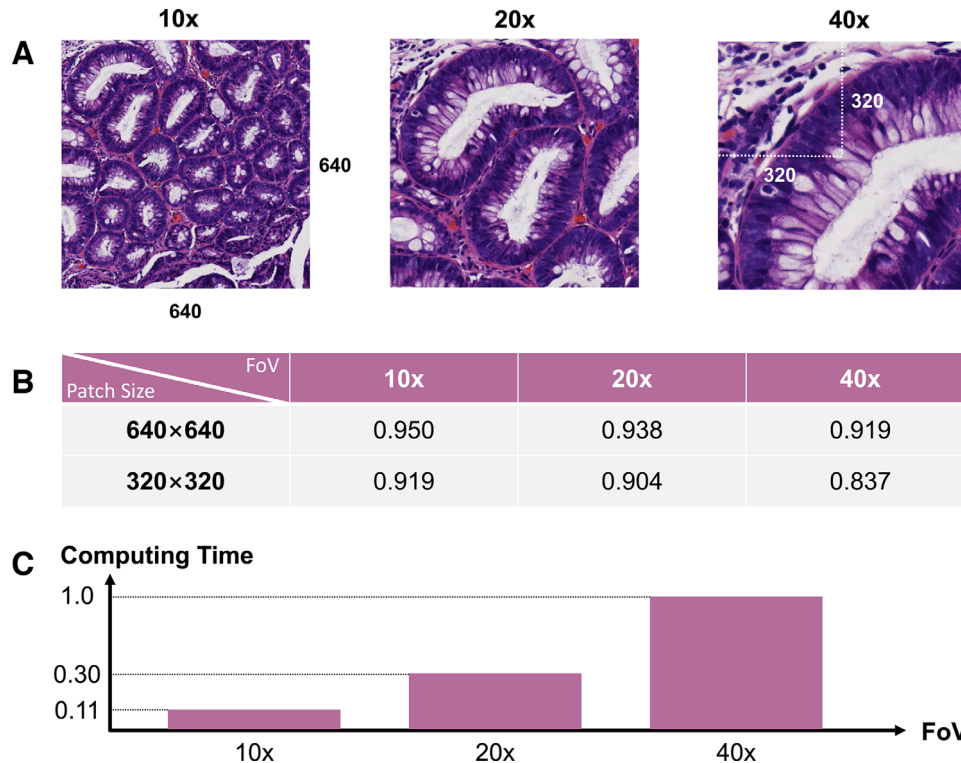
### Best model performance compared with pathologists

The slide-level ROC curve was given in figure 3, the AUC was 0.92. We had invited five pathologists to diagnosis the 194 test slides. As shown in figure 3, five pathologists gave significant different diagnosis results, showing the subjectiveness in the adenoma identification process. One could find the model performance was better than the average pathologist. In the following, the best model was chosen at the inverted triangle spotted in figure 3.

Some qualitative examples were shown in figure 4A. When we focused our attention on the regions with high probabilities (crimson), we could see the wedge-shaped adenomatous regions, which was consistent with a common observation from pathologists.

### Generalisation test

To further test the generalisation ability of the model, we fed slides from the generation testing group into our system and compared the predictions given by the model against the histological reports. Results were shown in table 3. Without any fine-tuning on the original model, it found 155 out of 168 slides (adenoma: 111; normal: 57) were correctly predicted, indicating the model still

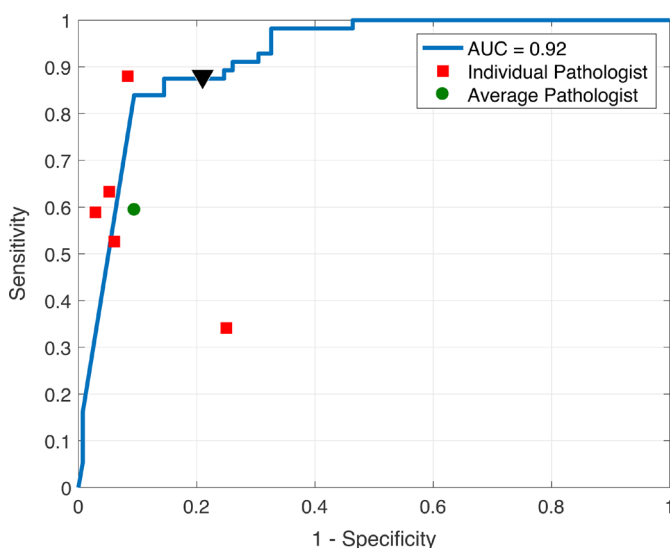


**Figure 2** (A) An example of tiles in  $\times 10$ ,  $\times 20$  and  $\times 40$  FoVs; (B) Tile-level classification accuracy on the validation set; (C) relative computing time of a WSI on different FoVs. FoVs, for different field of views.

maintained high accuracy under different staining configurations. [Figure 4B](#) shows three examples.

### System efficiency and scalability

Due to the large file size of histological slides, it was clear that building a system supporting multiple GPUs for the automatic diagnosis process was essential. The system completed the analysis of a slide with the size of 500 MB in 30s on a single GTX1080Ti GPU. As shown in [figure 4C](#), the system performance increased near linearly with the hardware configuration (ie, number of GPUs).



**Figure 3** Performance of the deep learning model and five pathologists. AUC, area under the curve.

### False analysis and model visualisation

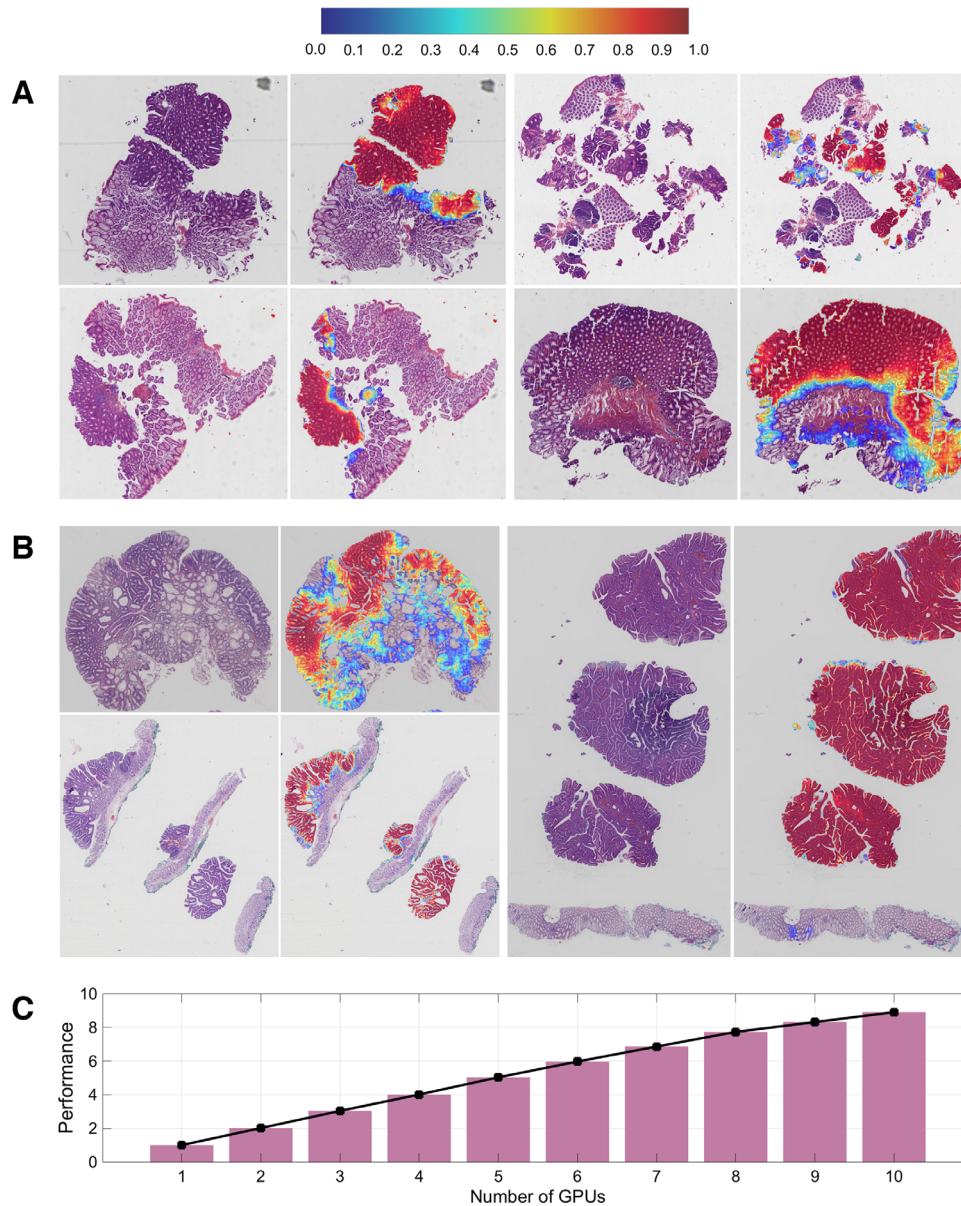
As shown in [figure 5A](#), in (I) and (II), the adenomatous grade was low, though the model successfully spots these glands, the probability was too low to reach an affirmative decision. The false positive predictions were closely related to tissue cauterisation and hyperplasia as shown in [figure 5A](#)-(III) and (IV), respectively.

We gave three representative examples in [figure 5B](#) to reveal the model inference process. Interestingly, as we looked carefully into the final probability map, we could infer where exactly the model put its attention on. The highlighted area matched with the area of adenomatous proliferation.

### DISCUSSION

We found the FoV had a substantial impact on the diagnostic accuracy for both machines and pathologists. Specifically, in addition to targeted lesional cells, the histological environment around the cells was also crucial for the diagnosis process. We discovered that the model showed better performance with the  $\times 10$  FoV than with  $\times 20$  or  $\times 40$  FoV. In the meantime, to further increase the FoV that can be perceived by the model, we enlarged the training tile size from the commonly adopted pixel size of  $320 \times 320$  to  $640 \times 640$ . The best deep learning model reached an AUC of 0.92, showing comparable performance to the pathologists, even better than the average pathologist.

This methodology could be applied to the detection of other syndromes from the histopathological



**Figure 4** (A) Predicted examples in the test set; (B) some predictions for slides from other hospitals; (C) system performance against the hardware configuration.

aspect. From the experience of pathologists, the FoV was specific to the disease type. For instance, for cancer detection,  $\times 20$  or  $\times 40$  FoV was necessary to make a

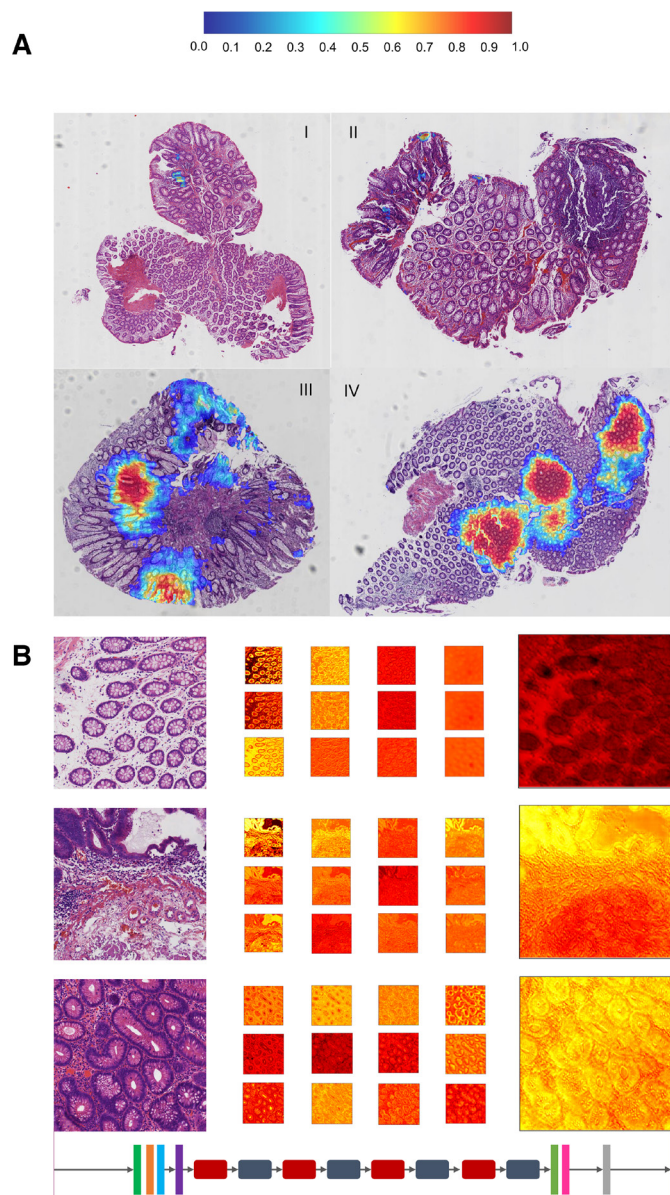
confirmative diagnosis. Despite all this, increasing the tile size should always be effective with abundant GPU resources.

**Table 3** Model performance on three test datasets, where T, V, TV, H, L represent tubular, villous, tubulovillous, high-grade, low-grade adenomas, respectively

Dataset	Adenoma, %	T, %	V, %	TH, %	TL, %	VH, %	VL, %	TVH, %	TVL, %
PLAGH	89.3/79.0	89.3	100.0	–	89.3	–	100.0	–	100.0
CJFH	90.0/92.3	89.8	100.0	100.0	88.3	100.0	100.0	–	100.0
CH	93.4/93.2	96.6	95.7	92.3	96.6	90.9	95.7	–	97.78

The second column gives sensitivity/specificity and the last columns list the sensitivity.

CH, Cancer Hospital, Chinese Academy of Medical Sciences; CJFH, China-Japan Friendship Hospital; PLAGH, Chinese People's Liberation Army General Hospital.



**Figure 5** (A) Falsely predicted examples in the test set; (B) feature maps extracted by the deep CNN. CNN, convolutional neural networks.

The model was required to have consistent and robust performance, that is, generalisation ability, in order to deal with the different staining configurations of histological slides across different hospitals. We had collected 168 additional slides from two other hospitals and achieved a slide-level accuracy of over 90%.

The false negative predictions were the cases which we need to be more cautious about. As shown in figure 5A, for false negative cases, when the adenomatous glands were small and hard to differentiate from regenerative changes at the base of crypts, the model tended to miss them. This behaviour was similar to that of junior pathologists who very often underdiagnose these ambiguous areas. False positive cases were related closely to tissue cauterisation and hyperplasia. Coincidentally, these tissue structures often confused the junior pathologists, some

may overlook an inflammatory background and mistaken regenerative atypia as adenomas. It was necessary to introduce quality assurance step to filter out low-quality slides, such as shredded and folded sections and marked cauterisation, before feeding them into the segmentation model.

To make a decision on whether there were adenomas, the pathologists mainly focused on gland and cell morphologies. From model visualisation results shown in figure 5B, we could observe that the lower CNN layers extracted the edge and colour information from the raw image. As the network went deeper, some of the feature maps gradually revealed glands and cells, especially gland shapes, nucleus and cell forms. For cases with abnormal gland shape and cell morphology, the model made the final decision and determined they were adenomatous glands. Otherwise, the tile was considered normal. This reasoning path was very similar to that of an experienced pathologist.

To apply the methodology to more disease for various organs, it is necessary to recruit a large number of WSIs in the training phase covering diverse tumour subtypes. These WSI should be labelled with accurate pixel-level annotations by experienced pathologists. The augmented data should be generated from domain-specific features of histopathology to further improve the robustness and generability under complicated scenarios.

## CONCLUSION

It was necessary to know whether deep learning models were similar to pathologists. To answer this question, we established a semantic segmentation model for colorectal adenomas diagnosis using deep CNNs and achieved an AUC of 0.92, which is on par with the performance of experienced pathologists. By carefully studying the influence of FoV on the model performance, we found that a larger FoV brings better diagnostic accuracy, which was consistent with pathologists' experience.

The model generalisation ability was proved by the multicenter test by slides collected from two other hospitals. We had discovered that the model made similar mistakes on the samples as junior pathologists would do. Meanwhile, model visualisation showed the reasoning path of the deep CNN was very similar to that of experts. By increasing the number of the training samples to include more types of adenomas in the training process, we could further improve the model performance.

## Author affiliations

<sup>1</sup>Department of Pathology, Chinese PLA General Hospital, Beijing, China

<sup>2</sup>Department of Pathology, Capital Medical University Affiliated Beijing Shijitan Hospital, Beijing, China

<sup>3</sup>Department of Pathology, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

<sup>4</sup>Department of Pathology, China-Japan Friendship Hospital, Beijing, China

<sup>5</sup>Thorough Images, Beijing, China

<sup>6</sup>School of Life Sciences, Tsinghua University, Beijing, China

<sup>7</sup>Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China

**Acknowledgments** The authors would like to thank Xiang Gao, Lang Wang, Yuefeng Wang, Siqi Zheng, Cunguang Wang, Fangjun Ding at Thorough Images for data processing and helpful discussions.

**Contributors** ZS, SZ, SW, WX, NL and HS proposed the research, CY, YH, XD, JL, LS, JY, XG, WJ, ZW, XC, YZ and XD performed the slide annotation, WW and HC led the multicentre test, CL, GX, ZS and CK wrote the deep learning code and performed the experiment, ZS and SW wrote the manuscript, WX, NL and HS reviewed the manuscript.

**Funding** This work is supported by CAMS Innovation Fund for Medical Sciences (CIFMS) (grant number 2018-I2M-AI-008); the National Natural Science Foundation of China (NSFC) [grant number 61532001; and the Tsinghua Initiative Research Programme Grant (grant number 20151080475).

**Competing interests** SW is the cofounder, chief technology officer, and equity holder of Thorough Images. CL, ZS, CK are algorithm researchers of Thorough Images. All remaining authors have declared no conflicts of interest.

**Patient and public involvement** Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

**Patient consent for publication** Not required.

**Ethics approval** The study was approved by the institutional review board of each participating hospital (Medical Ethics Committee, Chinese PLA General Hospital, S2018-163-01; Research Ethics Committee, China-Japan Friendship Hospital Clinical, 2018-106 K75; Ethics Committee of National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences, NCC1789).

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available on request to the corresponding authors.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

#### ORCID iDs

Zhanbo Wang <http://orcid.org/0000-0002-4316-887X>

Shuhao Wang <http://orcid.org/0000-0002-5467-3548>

#### REFERENCES

- Litjens G, Kooi T, Bejnordi BE, *et al*. A survey on deep learning in medical image analysis. *Med Image Anal* 2017;42:60–88.
- Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.
- Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus Photographs. *JAMA* 2016;316:2402–10.
- Kermany DS, Goldbaum M, Cai W, *et al*. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 2018;172:1122–31.
- De Fauw J, Ledsam JR, Romera-Paredes B, *et al*. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018;24:1342–50.
- Wang P, Xiao X, Glissen Brown JR, *et al*. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nat Biomed Eng* 2018;2:741–8.
- Coudray N, Ocampo PS, Sakellaropoulos T, *et al*. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med* 2018;24:1559–67.
- Fast R-CNN Girshick R. *Proceedings of the IEEE International Conference on computer vision*, 2015: 1440–8.
- Ren S, He K, Girshick R, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell* 2017;39:1137–49.
- Liu W, Anguelov D, Erhan D. *SSD: single shot multibox detector*. European Conference on Computer Vision, 2016: 21–37.
- Dai J, Li Y, He K. *R-FCN: object detection via region-based fully convolutional networks*. *Advances in neural information processing systems*, 2016: 379–87.
- Redmon J, Divvala S, Girshick R. *You only look once: unified, real-time object detection*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779–88.
- Long J, Shelhamer E, Darrell T. *Fully convolutional networks for semantic segmentation*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3431–40.
- Chen LC, Papandreou G, Kokkinos I, *et al*. Semantic image segmentation with deep convolutional nets and fully connected CRFs. *arXiv* 2014;7062:1412.
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436–44.
- Benson VS, Patnick J, Davies AK, *et al*. Colorectal cancer screening: a comparison of 35 initiatives in 17 countries. *Int J Cancer* 2008;122:1357–67.
- Kinzler KW, Vogelstein B. Lessons from hereditary colorectal cancer. *Cell* 1996;87:159–70.
- Chino A, Yamamoto N, Kato Y, *et al*. The frequency of early colorectal cancer derived from sessile serrated adenoma/polyps among 1858 serrated polyps from a single institution. *Int J Colorectal Dis* 2016;31:343–9.
- Lieberman DA, Rex DK, Winawer SJ, *et al*. Guidelines for colonoscopy surveillance after screening and polypectomy: a consensus update by the US Multi-Society Task force on colorectal cancer. *Gastroenterology* 2012;143:844–57.
- Lin JS, Piper MA, Perdue LA, *et al*. Screening for colorectal cancer: updated evidence report and systematic review for the US preventive services Task force. *JAMA* 2016;315:2576–94.
- Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern* 1979;9:62–6.
- Ronneberger O, Fischer P, Brox T. *U-Net: Convolutional networks for biomedical image segmentation*. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015: 234–41.
- Zeiler MD, Fergus R. *Visualizing and understanding convolutional networks*. European Conference on Computer Vision, 2014: 818–33.