Deep learning-enabled pelvic ultrasound images for accurate diagnosis of ovarian cancer in China
a retrospective, multicentre, diagnostic study
88 authors, including; Chen, Lingxi

Published in:
The Lancet Digital Health

Published: 01/03/2022

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

License:
CC BY-NC-ND

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1016/S2589-7500(21)00278-8

Publication details:

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher's definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher's copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.
Deep learning-enabled pelvic ultrasound images for accurate diagnosis of ovarian cancer in China: a retrospective, multicentre, diagnostic study

Yue Gao*, Shaqing Zeng*, Xiaoyan Xu*, Huayi Li*, Zhuhong Yao*, Kun Song*, Xiao Li*, Lingxi Chen, Junying Tang, Hui Xing, Zhiyong Yu, Qinghua Zhang, Shue Zeng, Cunjian Yi, Hongning Xie, Xiaoming Xiong, Guangyao Cai, Zhi Wang, Yuan Wu, Jianhua Chi, Xiaofei Jiao, Yan Qin, Xiaogang Mao, Yu Chen, Xin Jin, Qingqing Mo, Pingbo Chen, Yi Huang, Yushuang Shi, Junmei Wang, Yimin Zhou, Shuping Ding, Shan Zhu, Xin Liu, Xiangyi Dong, Lin Cheng, Linlin Zhu, Huanhuan Cheng, Li Cha, Yanli Hao, Chunchun Jin, Lidan Zhang, Peng Zhou, Meng Sun, Qin Xu, Kehua Chen, Zeyan Gao, Xu Zhang, Yuanyan Ma, Yan Liu, Liling Xiao, Li Xu, Lin Peng, Zheyu Hao, Mi Yang, Yane Wang, Hongping Ou, Yongmei Jia, Lihua Tian, Wei Zhang, Ping Jin, Xun Tian, Lei Huang, Zhen Wang, Jiahao Liu, Tian Fang, Danmei Yan, Heng Cao, Jingjing Ma, Xiaoting Li, Xu Zheng, Hua Lou, Chunyan Song, Ruyuan Li, Siyuan Wang, Wenqian Li, Xulei Zheng, Jing Chen, Guannan Li, Ruqi Chen, Cheng Xu, Ruidi Yu, Ji Wang, Sen Xu, Beihua Kong, Xing Xie, Ding Ma, Qinglei Gao

Summary

Background Ultrasound is a critical non-invasive test for preoperative diagnosis of ovarian cancer. Deep learning is making advances in image-recognition tasks; therefore, we aimed to develop a deep convolutional neural network (DCNN) model that automates evaluation of ultrasound images and to facilitate a more accurate diagnosis of ovarian cancer than existing methods.

Methods In this retrospective, multicentre, diagnostic study, we collected pelvic ultrasound images from ten hospitals across China between September 2003, and May 2019. We included consecutive adult patients aged ≥18 years with adnexal lesions in ultrasonography and healthy controls and excluded duplicated cases and patients without adnexa or pathological diagnosis. For DCNN model development, patients were assigned to the training dataset (34 448 images of 3755 patients with ovarian cancer, 541 442 images of 101 777 controls). For model validation, patients were assigned to the internal validation dataset (3031 images of 266 patients with ovarian cancer, 5385 images of 602 with benign adnexal lesions), external validation dataset 1 (486 images of 67 with ovarian cancer, 933 images of 268 with benign adnexal lesions), and 2 (1253 images of 166 with ovarian cancer, 5257 images of 723 benign adnexal lesions). Using these datasets, we assessed the diagnostic value of DCNN, compared DCNN with 35 radiologists, and explored whether DCNN could augment the diagnostic accuracy of six radiologists. Pathological diagnosis was the reference standard.

Findings For DCNN to detect ovarian cancer, AUC was 0.911 (95% CI 0.886–0.936) in the internal dataset, 0.870 (95% CI 0.822–0.918) in external validation dataset 1, and 0.831 (95% CI 0.793–0.869) in external validation dataset 2. The DCNN model was more accurate than radiologists at detecting ovarian cancer in the internal dataset (88.8% vs 85.7%), external validation dataset 1 (86.9% vs 81.1%), accuracy and sensitivity of diagnosis increased more after DCNN-assisted diagnosis than assessment by radiologists alone (87.6% vs 80.3%, 72.1–84.5, p=0.0001; 82.7% vs 78.5–86.9%, p=0.0001) and reached 0.876 and were significantly augmented when they were DCNN-assisted (p=0.05).

Interpretation The performance of DCNN-enabled ultrasound exceeded the average diagnostic level of radiologists, matched the level of expert ultrasound image readers, and augmented radiologists’ accuracy. However, these observations warrant further investigations in prospective studies or randomised clinical trials.

Funding National Key Basic Research Program of China, National Sci-Tech Support Projects, and National Natural Science Foundation of China.

Copyright © 2022 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY-NC-ND 4.0 license.

Introduction

Ovarian cancer occurs deep in the pelvis in ovarian tissue and produces vague and easily misattributed symptoms before it reaches an advanced stage. More than 75% of patients with ovarian cancer are initially diagnosed when it has advanced and they have a 5-year relative survival rate of 29% compared with 92% for patients with localised disease. No screening methods have been proven effective to date. Palpation of an adnexal mass during a pelvic examination and incidentally finding masses in imaging commonly initiate a diagnostic evaluation for ovarian cancer. Ultrasoundography is the most useful non-invasive diagnostic test for indeterminate adnexal masses because it is cheap, mostly harmless, readily accessible,
Research in context

Evidence before this study
We searched PubMed on Oct 10, 2021, for research articles containing terms “deep learning” OR “convolutional neural network” AND “ovarian cancer” AND “diagnosis” without date or language restrictions. We found six studies that used deep learning for diagnosis or predicting postsurgical outcomes in ovarian cancer. Only one study explored the combination of deep learning and ultrasound images for the diagnosis of ovarian cancer, which included less than 1000 cases and without external validations. Hence, the potential of deep-learning-enabled ultrasound for preoperative diagnosis of ovarian cancer was not clear. Literature had witnessed progress in machine-learning-assisted diagnosis, but explorations of integrating deep learning with ultrasound, the most useful non-invasive diagnostic test for ovarian cancer, were still rarely reported and largely insufficient, indicating the necessity of this study.

Added value of this study
In this study, we built and validated a deep convolutional neural network (DCNN) model using pelvic ultrasound images.
subtle signs. Low-level texture features and high-level deep features from images are extracted and cascaded to make an independent diagnosis by a convolutional neural network.2

In this study, we developed a DCNN model that automated detecting adnexal masses in ultrasound images and discriminating between malignant and benign masses and validated the model internally and externally. We also compared the DCNN model with 35 radiologists and explored whether it could augment the diagnostic accuracy of radiologists.

Methods

Study design and participants

In this retrospective, multicentre, diagnostic study, we developed, trained, and validated a DCNN model that aimed to differentiate between benign and malignant adnexal lesions using pelvic ultrasound images derived from ten hospitals across China, We collected ultrasound images from Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, and seven other hospitals for a training dataset (575,930 images of 105,532 cases) and an internal validation dataset (8416 images of 868 cases). We obtained images for external validation datasets from Jingzhou First People's Hospital (external validation dataset 1, 1419 images of 335 cases) and Xiangyang Central Hospital (external validation dataset 2, 6510 images of 889 cases). Recognising healthy ovaries was an indispensable step for the training process, so we included healthy controls and patients with adnexal lesions in the training dataset. Since the DCNN model was established to distinguish ovarian malignancy from benign lesions, the population of three validation datasets were exclusively patients with adnexal lesions. We compared the DCNN model with radiologists in the internal validation dataset (16 radiologists) and external validation dataset 1 (19 radiologists). Last, we explored whether the diagnostic accuracy of six radiologists could be augmented when they were DCNN-assisted.

We included consecutive adult patients (aged ≥18 years) who presented with adnexal lesions in ultrasound in ten hospitals between September 2003, and May 2019, and excluded duplicated cases, postoperative patients who were deprived of adnexa, and patients without histological diagnosis. We also included healthy adult (aged ≥18 years) women whose ultrasonography showed no adnexa-related abnormalities from Tongji Hospital and the other seven hospitals between September 2003, and March 2018. Data was collected retrospectively by trained investigators using the unified form and reviewed by two researchers (YG and SQZ).

All included patients with adnexal lesions were pathologically confirmed without knowledge of ultrasonography results after surgical removal by laparotomy or laparoscopy as the surgeon in the contributing hospitals considered appropriate. This method was in accordance with the WHO Classifications of Tumours of Female Reproductive Organs (2014),23 and the pathological diagnosis was the reference standard. The group with ovarian cancer included borderline tumours, primary ovarian cancer, metastatic cancer to the ovary, fallopian tube cancer, and peritoneal carcinoma. Benign adnexal lesions included lesions on ovaries and fallopian tubes, for example, benign adnexal cysts, benign cystic teratoma, cystadenoma, adnexal fibroma, hydrosalpinx, and adnexal abscess.

The use of ultrasound images, clinical information, and data collection protocols was approved by the ethical committee of Tongji Medical College, Huazhong University of Science and Technology. This study was done in accordance with the principles of the Declaration of Helsinki. As we used pre-existing medical data only, patient consent was waived.

Procedures

Ultrasound images were extracted from datasets of ten hospitals in JPEG format. Ultrasound was done in a transvaginal manner for non-virgins using miscellaneous machines, including Toshiba Aplio800, Siemens Acuson S3000, Philips EPIQ7, and GE Voluson S8. Image quality control was done to exclude images less than 224 pixels × 224 pixels. All images were acquired by examiners with more than 3 years of independent diagnostic experience. We collected an average of eight images per patient with adnexal lesions and four images per healthy woman. Each case was independently sampled and reviewed by three sonographers.

Radiologists were certified sonographers, trained in gynaecological scanning, and had at least 3 years of independent clinical experience in pelvic ultrasonography. Radiologists interpreted all images of a participant according to ultrasound features (eg, size of masses, boundaries, uniformity of echo, proportion of solid tissue, number of papillary projections, acoustic shadows, and ascites) on the basis of nomenclature and methodology proposed by the International Ovarian Tumor Analysis Group.24 The DCNN-assisted diagnosis was explored in six less experienced radiologists with 3–5 years of work experience. Six radiologists assessed images independently to reach the first conclusion (malignancy or benignity) at the patient level. Readouts of the DCNN model, including lesion labelling and probability of malignancy, were sent to radiologists to re-evaluate images. The second assessment of radiologists served as the final output. All readers were masked to pathological diagnosis and clinical information.

We trained a Dense Convolutional Network of 121 layers (DenseNet-121) on the basis of preprocessed pelvic ultrasound images in the training dataset (images 575,930). DenseNet-121 architecture enabled exploration of new features and improved information flow through the network by connecting each layer to every other
layer in a feed-forward manner. The advantage of DenseNet-121 was its ability to improve feature propagation, alleviate gradient vanishing, and reduce the number of parameters. The weights of DenseNet-121 were initialised from a model pretrained on ImageNet. The output of the last layer was changed to the two classes that matched the classification task of our study (ie, ovarian cancer vs non-ovarian cancer). Weighted cross-entropy loss was used as the objective function. The network was trained end-to-end with stochastic gradient descent using an initial learning rate of 0.001, momentum of 0.9, weight decay of 0.0001, and minibatches of 32. The learning rate was decreased by 0.1 after every 20 epochs. Data augmentations during the training process included random resize and crop, random horizontal flip, random rotation, random colour jittering, and normalisation. The predicted score of each image was calculated as the average prediction scores produced by the top five checkpoints of DenseNet-121.25,26 Eventually, the DCNN model generated a continuous prediction score between 0 and 1 for each image. For predictions of individual-level, weighted mean to the predicted probabilities of multiple images for each individual were combined into a single score. A tuning set containing 10% of images in the training dataset was randomly selected to optimise the network and set up hyperparameters. The Focal Loss method was applied to improve the accuracy of the DCNN model.27 For the test set, each image was resized to 256 pixels, cropped centrally to 224 pixels, and passed through the developed model. Logit values outputted from the model were transformed by softmax function into probabilities. For each individual, a malignancy score was calculated as the weighted mean of predicted probabilities of all ultrasound images from that person, specifically, for a given individual with n total images. The p value denotes probabilities of these images being classified as malignancy. The malignancy score was calculated as

$$\theta = -\left[ w_1 \times \log_{10}(1-p_1) + w_2 \times \log_{10}(1-p_2) + \ldots + w_n \times \log_{10}(1-p_n) \right]/n, \text{ where } w = p_i/(p_i + p_2 + \ldots + p_n).$$

The DCNN model was validated in the three validation datasets. The receiver operating characteristic (ROC) curve was plotted and area under curve, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score were used to assess model performance.

**Statistical analysis**

The predictions of the DCNN model and radiologists were compared with pathological diagnosis. ROC curves were constructed for the DCNN model in each validation dataset to calculate AUC. AUCs and 95% CIs, as well as other metrics of all datasets, were calculated using R-package (version 3.6), pROC (version 1.12.1), and GenBinomApps (version 1.0). 95% CIs of AUC, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score were used to assess model performance. The receiver operating characteristic (ROC) curve was plotted and area under curve, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score were used to assess model performance.

Figure 1: Study flowchart

We included healthy controls and patients with adnexal lesions in the training dataset for model training and optimisation. The goal of the DCNN model was to discriminate between malignant and benign ovarian masses so the population of three validation datasets were only patients with adnexal lesions. DCNN=deep convolutional neural network.
Role of the funding source
The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the manuscript.

Results
Between September 2003, and March 2018, 117746 individuals were assessed for eligibility in Tongji Hospital and seven other hospitals across China (figure 1). 11346 (10%) patients were excluded: 8436 (7%) duplicates, 1519 (1%) postoperative patients that lost adnexa, and 1391 (1%) patients without pathological diagnosis. 106400 eligible patients were assigned to the training dataset (105532 [99%]) or the internal validation dataset (868 [1%]). The training dataset comprised 575930 ultrasound images (34488 ovarian cancer and 541442 non-ovarian cancer) of 105532 cases (3755 ovarian cancer and 101777 non-ovarian cancer). The internal validation dataset comprised 8416 ultrasound images (3031 ovarian cancer and 5385 benign) of 868 cases (266 ovarian cancer and 602 benign) collected from the eight hospitals between September 2016, and March 2018. Between March 2018 and January 2019, we screened 390 consecutive patients with adnexal lesions using ultrasonography in Jingzhou First People’s Hospital. 64 patients were excluded (41 duplicates, 18 patients without pathological diagnosis, and five postoperative patients), which created external validation dataset 1 comprising 1419 images (486 ovarian cancer and 933 benign) of 335 cases (137 ovarian cancer and 268 benign). Between March 2012, and March 2018, 993 patients were screened from Xiangyang Central Hospital. External validation dataset 2 comprised 6510 images (1253 ovarian cancer and 5257 benign) of 889 cases (166 ovarian cancer and 723 benign) after excluding 104 patients (66 duplicated cases, 27 patients without pathological diagnosis, and 11 postoperative patients).

Baseline characteristics of individuals from all datasets are provided in table 1. Because the training dataset included healthy controls, whereas participants in validation datasets were patients with adnexal lesions, the proportion of ovarian cancer in the training dataset was lower than that in validation datasets. Median age of participants was lowest in the training dataset, highest in external validation dataset 1, and the same in internal validation dataset, external validation dataset 2, and the same in internal validation dataset 1. EVD2=external validation dataset 2. FIGO=The International Federation of Gynaecology and Obstetrics.

<table>
<thead>
<tr>
<th>TD (n=105 532)</th>
<th>IVD (n=868)</th>
<th>EVD1 (n=335)</th>
<th>EVD2 (n=889)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ovarian cancer cases</td>
<td>3755 (3.6%)</td>
<td>266 (30.6%)</td>
<td>67 (20.0%)</td>
</tr>
<tr>
<td>Non-ovarian cancer cases</td>
<td>101777 (96.4%)</td>
<td>602 (64.9%)</td>
<td>268 (80.0%)</td>
</tr>
<tr>
<td>Age (years)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤45 years</td>
<td>32 (27-42)</td>
<td>38 (27-49)</td>
<td>43 (32-52)</td>
</tr>
<tr>
<td>&gt;45 years</td>
<td>36197 (34%)</td>
<td>273 (31%)</td>
<td>137 (41%)</td>
</tr>
<tr>
<td>Histology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serous</td>
<td>2550 (67.9%)</td>
<td>172 (64.6%)</td>
<td>43 (64.2%)</td>
</tr>
<tr>
<td>Mucinous</td>
<td>406 (10.8%)</td>
<td>42 (15.8%)</td>
<td>11 (16.3%)</td>
</tr>
<tr>
<td>Endometrioid</td>
<td>337 (9.0%)</td>
<td>21 (7.8%)</td>
<td>6 (9.0%)</td>
</tr>
<tr>
<td>Clear cell</td>
<td>242 (6.4%)</td>
<td>17 (6.5%)</td>
<td>5 (7.5%)</td>
</tr>
<tr>
<td>Others*</td>
<td>223 (5.9%)</td>
<td>14 (5.3%)</td>
<td>2 (3.0%)</td>
</tr>
<tr>
<td>FIGO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage I</td>
<td>791 (21.1%)</td>
<td>80 (30.1%)</td>
<td>20 (29.9%)</td>
</tr>
<tr>
<td>Stage II</td>
<td>468 (12.5%)</td>
<td>33 (12.4%)</td>
<td>9 (14.3%)</td>
</tr>
<tr>
<td>Stage III</td>
<td>1954 (52.0%)</td>
<td>101 (38.0%)</td>
<td>26 (38.5%)</td>
</tr>
<tr>
<td>Stage IV</td>
<td>542 (14.4%)</td>
<td>52 (19.5%)</td>
<td>12 (17.9%)</td>
</tr>
<tr>
<td>Grade</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>231 (5.2%)</td>
<td>22 (8.3%)</td>
<td>2 (3.0%)</td>
</tr>
<tr>
<td>Median</td>
<td>436 (11.6%)</td>
<td>39 (14.7%)</td>
<td>9 (13.4%)</td>
</tr>
<tr>
<td>High</td>
<td>2604 (69.3%)</td>
<td>161 (60.5%)</td>
<td>46 (68.7%)</td>
</tr>
<tr>
<td>Borderline</td>
<td>484 (12.9%)</td>
<td>44 (16.5%)</td>
<td>10 (14.9%)</td>
</tr>
</tbody>
</table>

AUC was 0·911 (95% CI 0·886–0·936) in the internal dataset for the DCNN model to discriminate ovarian cancer from its benign mimickers (figure 2A). AUC was 0·870 (95% CI 0·822–0·918) in external dataset 1 and 0·831 (95% CI 0·793–0·869) in external dataset 2 for the model to differentiate ovarian malignancy from benignity (figure 2B, 2C, table 2). Moreover, internal dataset had an accuracy of 88·8% (95% CI 86·5–90·8), a sensitivity of 78·9% (95% CI 73·5–83·7), and a specificity of 93·2% (95% CI 90·9–95·1) for the DCNN model. In external datasets 1 and 2, an accuracy of 86·9% (95% CI 82·8–90·3%) and 85·3% (95% CI 82·8–87·5%), a sensitivity of 40·3% (95% CI 28·5–53·0%) and 57·8% (95% CI 49·9–65·4%), and a specificity of 98·5% (95% CI 96·2–99·6%) and 91·6% (95% CI 89·3–93·5%) were observed for the DCNN model to discern ovarian malignancy from benign pathology (table 2). The calibration curve showed that the DCNN model could discern between borderline and benign tumours for their malignancy potential. We thus gathered borderline tumours (n=79) and benign controls (n=1593) in three validation datasets to assess whether the DCNN model could discern between borderline and benign tumours.
benign lesions. AUC was 0·821 (95% CI 0·765–0·876) with an accuracy of 82·4% (95% CI 80·5–84·2) for DCNN to distinguish borderline tumours from benign controls, with a Brier score of 0·04 (appendix p 7).

To characterise the diagnostic value of the DCNN model, we compared it with radiologist’s values using all images in the internal validation dataset and external dataset 1 (16 radiologists vs 19 radiologists; table 2). Radiologists classified adnexal lesions by simultaneously interpreting all case images to better mirror clinical practice. The DCNN model had higher accuracy, lower sensitivity, and better specificity than the average diagnostic performance of radiologists in the internal validation dataset (88·8% vs 85·7%, p<0·0001; 78·9% vs 83·1%, p<0·0001; 93·2% vs 86·8%, p<0·0001). For external validation dataset 1, the DCNN model had better accuracy and specificity than radiologist’s accuracy and specificity (86·9% vs 81·1%, p<0·0001; 98·5% vs 87·5%, p<0·0001; appendix p 3; table 2). The diagnostic value of the DCNN model significantly exceeded the average diagnostic level of ultrasound image readers (figure 2). The diagnostic performance of individual radiologists is available in the appendix (pp 2–3). We included 438 consecutive patients with adnexal lesions in ultrasonography between February 2019 and May 2019, in Tongji Hospital to explore whether embedding the DCNN model in clinic could facilitate more accurate assessments of ultrasound images (932 images of 80 patients with ovarian cancer, 1742 images of 358 patients with benign lesions). Baseline characteristics of 438 patients are available in the appendix (p 4).

Table 3 shows the comparison of capability of discerning ovarian cancer from its benign mimickers between radiologists alone and radiologists with DCNN assistance. The diagnostic assessment accuracy for six radiologists without the DCNN model ranged between 70·3% (95% CI 65·8–74·6) and 85·4% (95% CI 81·7–88·6). Their average accuracy when DCNN-assisted was 87·6% (85·0–90·2); their diagnostic accuracy was significantly augmented when DCNN-assisted (p<0·0001). Moreover, average sensitivity and specificity of radiologists also improved by varying degrees before and after aid of the DCNN model (82·7% vs 70·4%, p<0·0001, 88·7% vs 80·1%, p<0·0001). The readers’ rating scale with or without assistance of DCNN model is shown in the appendix. The performance of DCNN alone is shown in table 3.

A latent window of opportunity of 4·3 years is reported and believed to exist for early detection of ovarian cancer.13 Of patients with ovarian cancer from Tongji Hospital, we found that 12 had images produced 0·9–3·7 years before it was diagnosed when visual interpretations of these images did not indicate potential ovarian malignancy. 10 (83%) of 12 individuals were identified as patients with ovarian cancer by the DCNN model using these images (appendix p 5). Two images derived from one of the ten cases are displayed in the appendix (p 8). In addition to hinting at the importance of closely monitoring indeterminate adnexal masses, this observation suggested that the DCNN model could potentially locate radiographic abnormalities that are undiscernible by the human eye, thus offering opportunities to detect ovarian cancer early. However, it should be interpreted with caution owing to
the small number of patients involved and the paucity of pathological evidence indicating that ovarian cancer existed when the image was produced.

**Discussion**

We built and validated a DCNN model that automated interpreting ultrasound images for discerning ovarian cancer from its benign tumours. The comparison between the DCNN model and 35 radiologists revealed that the diagnostic accuracy of the DCNN model significantly exceeded the average level of ultrasound image readers. More importantly, the diagnostic accuracy of six radiologists was significantly improved when assisted with the DCNN model.

The first strength of the DCNN model is the automated capability. The tenet of applying artificial intelligence is to streamline clinical workflow and improve diagnostic accuracy. Assessing ultrasound images is an iterative and labour-consuming process. Deep-learning algorithms can process large quantities of image data efficiently, are not vulnerable to fatigue, and have high throughput and stability in diagnosis. Moreover, image interpretations involve recognising nuanced patterns for which DCNN that mimics human neural networks is best suited.\(^\text{14,15}\)

The DCNN model could be embedded in ultrasound workstations to provide radiologists with real-time guidance. Sonographers can also devote more time to obscure images that can cause high risks of misdiagnosis and omissions. Furthermore, the model can be used in remote regions in China, subject to further validation, to help alleviate the uneven distribution of medical and human resources.

The second strength of the DCNN model is its robustness. Compared with the available DCNN-enabled models based on ultrasound for diagnosis of ovarian cancer that have insufficient images for training and validation.

<table>
<thead>
<tr>
<th>IVD</th>
<th>DCNN</th>
<th>EVD1</th>
<th>DCNN</th>
<th>EVD2</th>
<th>DCNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radiologists</td>
<td>DCNN</td>
<td>Radiologists</td>
<td>DCNN</td>
<td>Radiologists</td>
<td>DCNN</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>85.7% (83.4–88.0)</td>
<td>88.8% (86.5–90.8)</td>
<td>81.1% (77.8–84.5)</td>
<td>86.9% (82.8–90.3)</td>
<td>85.3% (82.8–87.5)</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>83.1% (81.4–84.8)</td>
<td>78.9% (73.5–83.7)</td>
<td>55.5% (47.3–63.7)</td>
<td>40.3% (28.5–53.0)</td>
<td>57.8% (49.9–65.4)</td>
</tr>
<tr>
<td>Specificity</td>
<td>86.8% (83.6–90.1)</td>
<td>93.2% (90.9–95.1)</td>
<td>87.5% (82.3–92.7)</td>
<td>98.5% (96.2–99.6)</td>
<td>91.6% (89.3–93.5)</td>
</tr>
<tr>
<td>Positive predictive value</td>
<td>74.6% (70.1–79.2)</td>
<td>83.7% (78.5–88.0)</td>
<td>58.1% (51.8–64.5)</td>
<td>87.1% (70.2–96.4)</td>
<td>61.1% (53.1–68.8)</td>
</tr>
<tr>
<td>Negative predictive value</td>
<td>92.1% (91.3–92.8)</td>
<td>90.9% (88.4–93.1)</td>
<td>89.0% (87.5–90.5)</td>
<td>86.8% (82.5–90.4)</td>
<td>90.4% (88.1–92.5)</td>
</tr>
</tbody>
</table>

**Kappa** | 0.678 | 0.733 | 0.472 | 0.486 | 0.504 |

**F1** | 0.784 | 0.812 | 0.539 | 0.551 | 0.594 |

Data are n (95% CI), unless otherwise specified. DCNN=deep convolutional neural network. IVD=internal validation dataset. EVD1=external validation dataset 1. EVD2=external validation dataset 2.

**Table 2:** Performance of DCNN model and average diagnostic level of radiologists in validation datasets

<table>
<thead>
<tr>
<th>Radiologists with DCNN model</th>
<th>84.0% (80.2–87.3)</th>
<th>86.3% (76.7–92.9)</th>
<th>83.5% (79.3–87.2)</th>
<th>53.9% (44.9–62.8)</th>
<th>96.5% (93.7–98.2)</th>
<th>0.566</th>
<th>0.664</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radiologist A with DCNN</td>
<td>86.5% (83.0–89.6)</td>
<td>85.0% (75.3–92.0)</td>
<td>86.9% (82.9–90.2)</td>
<td>59.1% (49.6–68.2)</td>
<td>96.3% (93.6–98.1)</td>
<td>0.614</td>
<td>0.697</td>
</tr>
<tr>
<td>Radiologist C with DCNN</td>
<td>90.2% (87.0–92.8)</td>
<td>75.0% (64.1–84.0)</td>
<td>93.6% (90.5–95.9)</td>
<td>72.3% (61.4–81.6)</td>
<td>94.4% (91.4–96.5)</td>
<td>0.676</td>
<td>0.736</td>
</tr>
<tr>
<td>Radiologist D with DCNN</td>
<td>87.9% (84.5–90.8)</td>
<td>83.8% (73.8–91.1)</td>
<td>88.8% (85.1–91.9)</td>
<td>62.6% (52.7–71.8)</td>
<td>96.1% (93.4–97.9)</td>
<td>0.642</td>
<td>0.717</td>
</tr>
<tr>
<td>Radiologist E with DCNN</td>
<td>90.6% (87.5–93.2)</td>
<td>82.5% (72.4–90.1)</td>
<td>92.5% (89.2–95.0)</td>
<td>71.0% (60.6–79.9)</td>
<td>95.9% (93.3–97.8)</td>
<td>0.705</td>
<td>0.763</td>
</tr>
<tr>
<td>Radiologist F with DCNN</td>
<td>86.3% (82.7–89.4)</td>
<td>83.8% (73.8–91.1)</td>
<td>86.9% (82.9–90.2)</td>
<td>58.8% (49.2–67.9)</td>
<td>96.0% (93.2–97.9)</td>
<td>0.606</td>
<td>0.691</td>
</tr>
<tr>
<td>Radiologists’ mean with DCNN</td>
<td>87.6% (85.0–90.2)</td>
<td>82.7% (78.5–86.9)</td>
<td>88.7% (84.7–92.7)</td>
<td>63.0% (55.3–70.6)</td>
<td>95.9% (95.1–96.6)</td>
<td>0.635</td>
<td>0.711</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Radiologists without DCNN model</th>
<th>79.2% (75.1–82.9)</th>
<th>55.0% (43.5–66.2)</th>
<th>84.6% (80.5–88.2)</th>
<th>44.4% (34.5–54.8)</th>
<th>89.4% (85.6–92.5)</th>
<th>0.363</th>
<th>0.492</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radiologist A</td>
<td>82.4% (78.5–85.9)</td>
<td>68.8% (57.4–78.7)</td>
<td>85.5% (81.4–89.0)</td>
<td>51.4% (41.5–61.2)</td>
<td>92.5% (89.1–95.1)</td>
<td>0.479</td>
<td>0.588</td>
</tr>
<tr>
<td>Radiologist C</td>
<td>70.3% (65.7–74.6)</td>
<td>72.5% (61.4–81.9)</td>
<td>69.8% (56.8–74.6)</td>
<td>34.9% (27.7–42.7)</td>
<td>91.9% (88.0–94.9)</td>
<td>0.299</td>
<td>0.427</td>
</tr>
<tr>
<td>Radiologist D</td>
<td>80.4% (76.3–84.0)</td>
<td>73.8% (62.7–83.0)</td>
<td>81.8% (77.5–85.7)</td>
<td>47.6% (38.5–56.7)</td>
<td>93.3% (90.0–95.8)</td>
<td>0.458</td>
<td>0.578</td>
</tr>
<tr>
<td>Radiologist E</td>
<td>72.2% (67.7–76.3)</td>
<td>87.5% (78.2–92.8)</td>
<td>68.7% (63.6–73.5)</td>
<td>38.5% (31.4–46.0)</td>
<td>96.1% (92.9–98.1)</td>
<td>0.376</td>
<td>0.534</td>
</tr>
<tr>
<td>Radiologist F</td>
<td>85.4% (81.7–88.6)</td>
<td>65.0% (53.5–75.3)</td>
<td>89.9% (86.4–92.9)</td>
<td>59.1% (48.1–69.5)</td>
<td>92.0% (88.7–94.6)</td>
<td>0.529</td>
<td>0.619</td>
</tr>
<tr>
<td>Radiologists’ mean without DCNN model</td>
<td>78.3% (73.1–84.5)</td>
<td>70.4% (59.1–81.7)</td>
<td>80.1% (70.9–89.3)</td>
<td>46.0% (36.8–55.2)</td>
<td>92.5% (90.2–94.8)</td>
<td>0.417</td>
<td>0.547</td>
</tr>
</tbody>
</table>

DCNN=deep convolutional neural network. PPV=positive predictive value. NPV=negative predictive value.

**Table 3:** DCNN-assisted evaluation of pelvic ultrasound images
external validations,29 our model is built using 575,930 images of 105,532 cases collected from ten hospitals across China and validated in independent, large-scale, and heterogeneous datasets. Three validation datasets containing 16,345 aggregated images of 2,092 cases are derived from three hospitals, two of which are unknown to the DCNN model. The participants included in this study have a myriad of histopathological types of malignant and benign adnexal lesions, and images are generated by various professionals using different devices. The heterogeneity of geography, cases, and images ensure the reproducibility of the DCNN model. Third, the DCNN model is a potential tool to assist radiologists. Surgeons often misinterpret adnexal masses, although they understand ultrasound assessments and clinical factors like age, menopausal status, and cancer antigen-125. Artificial intelligence could provide radiologists with quantitative features of images other than qualitative ones, enabling more comprehensive evaluations of images. Deep learning could learn image features representative of characteristics and locate considerably subtle signs, which are beyond the capability of visual observation.35 Auxiliary use of the DCNN model could improve diagnostic accuracy in less experienced radiologists, indicating that we could leverage the DCNN model in developed areas where experts are overloaded to meet medical demands and in remote regions with suboptimal medical resources.

The accuracy of radiologists is lower in this study than in subjective assessments of ultrasound for diagnosis of ovarian cancer in other studies.7 As some images were produced as early as 2003 using less updated machines, the accuracy could be impaired. Second, some images were collected from municipal hospitals in the less developed areas of China, suggesting the appropriateness of leveraging the DCNN model in regions where prime medical resources are scarce. The declined accuracy could also be attributed to the different diagnostic workflow between studies. Readers in this study made classifications on the basis of static images without referring to clinical factors. By contrast, radiologists in the real diagnostic scenario could continuously and dynamically observe the lesion and three-dimensional changes and take multiple clinical parameters into account. Compared with ultrasonography in superficial organs, such as the thyroid and breast, the accuracy of pelvic ultrasound for diagnosis of ovarian cancer is vulnerable to body shape, subcutaneous fat thickness, ascites, intestinal cavity, and depth of ovarian masses. Besides, the DCNN model is developed on static images and could not visualise lesions from multiple perspectives in a real-time scenario, which might deteriorate model performance. To address this problem, we did extensive data augmentation during model development and increased image variations, shown by the comparable and generalisable performance of DCNN in external validations. Images in validation datasets derived from two smaller hospitals had a greater effect on the performance of DCNN than the training dataset, image quality, and acquisition equipment in the testing datasets. Therefore, the performance of DCNN in testing datasets is slightly worse than that in the training dataset.

The past two decades have witnessed many effective models that combine ultrasound assessments and clinical factors to reach a degree of suspicion that the ovarian mass might be malignant, such as Logistic Regression Models 1 and 2, Simple Rules, ADNEX, and Risk of Malignancy Index.33 The DCNN model could supplement these algorithms because it is adept in extracting ultrasound features and making classifications independent of clinical information.34 At present, the DCNN model aims to distinguish between ovarian cancer and benign adnexal lesions. However, ovarian malignancy comprises distinct histological subtypes with unique genomic characteristics. Besides, BRCA status is an important signature for managing ovarian cancer and a major risk factor of developing ovarian cancer, but genetic tests are unaffordable for many people worldwide. Exploring the capabilities of the DCNN model to classify subtypes of ovarian cancer and characterise BRCA status are among important directions for future research.

The DCNN model has some limitations. First, this study is retrospective; therefore, validations in prospective settings can provide more powerful conclusions.35 Another consequence of the retrospective nature is the unbalanced number of images between patients with and without ovarian cancer. Second, ultrasound images are exclusively collected in China. The model’s capabilities to discriminate between malignant and benign adnexal lesions in other ethnicities merit investigation. Third, although we strive to evaluate the DCNN model realistically, empirical validations and real-world applications are inconsistent, especially for DCNN-assisted assessments.36 Moreover, metastatic cancer to the ovary is distinct from primary ovarian cancer regarding clinical presentation, management, and treatment. However, we could not explore the model’s potential to separate primary ovarian cancer from secondary disease because of insufficient secondary ovarian cancers in the validation datasets. Last, BRCA mutation, a crucial signature for ovarian cancer treatment, could not be investigated owing to the scarcity of genetic analysis.

The DCNN model based on ultrasonography enables expert-level distinctions between ovarian cancer and its benign mimickers. The efficacy of radiologists to interpret pelvic ultrasound images for ovarian cancer diagnosis is augmented when readers are assisted by the DCNN model.

Contributors
YG and QLG contributed to the study design. LXC and SQZ performed the data analysis and were responsible for data collection. YG and HYL contributed to the data interpretation and writing of the manuscript. QLG and DM were the senior supervisors of the project. SZY, KS, XI, JYT, HX, ZYY, QHZ, SEZ, CjY, HNX, XMX, GYC, ZW, YQ, XGM, YC, XI, QQM, PBC, YH, YSS, SPD, SZ, BHK, and XX provided or interpreted images. JHC, XJF, ZJL, TF, DMY, HC, JMJ, XTL, XZ, HL, CVS, RYL, SYW, WQL, XIZ, JC, GNJ, BGQ, CX, RDY, JW, and SX...
curated pathological examinations. XYX, SEZ, JW, JMW, YMZ, XL, XDY, LC, LLZ, HHC, LC, YLH, CCJ, LDZ, FZ, MS, QS, KHC, ZYG, XZ, YYM, XL, LX, LP, ZYH, MY, YD, HPO, YMJ, LH, TZ, PJ, XT, and LH assessed the randomly selected images from internal validation dataset and external validation dataset. 1. QLG and YG had access to and verified the data supporting this study. All authors reviewed and approved the submitted manuscript.

**Declaration of interests**

QLG will apply for a patent for the DCNN model used in this study pending to Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology. All other authors declare no competing interests.

**Data sharing**

Requests for patient-related data not included in the article will not be considered. Data will only be available on execution of appropriate Material Transfer Agreement. We used code that is freely available at https://github.com/pytorch/vision/tree/master/references/classification

**Acknowledgments**

This study is supported by the National Science and Technology Major Project-2018ZX10304204-002, the National Natural Science Foundation of China (82072889, 81772787, 81630060, 81702574, and 81702572), the National Key Basic Research Program of China (2015CB553903), the National Key Research & Development Program of China (2016YFC0902001), the Technical Innovation Special Project of Hubei Province (2018ACA138), and the Fundamental Research Funds for the Central Universities (2019kyXMZ024). We especially thank Wenzhi Lv, Leigang Sun, and Youcai Ba from the Julei technology for their support in data extraction and processing. We are grateful for the husband of YG, Lei Cai, for his help in editing the tables and support during the revision. We acknowledge Dr G Kaushik Subramanian and Dr Amit Bhat, Indegene, Bangalore, India, for their assistance in medical writing and revising.

**References**


