VIA MEDICA

ORIGINAL PAPER / GYNECOLOGY

Ginekologia Polska
2024, vol. 95, no. 3, 181–189

ISSN 0017–0011, e-ISSN 2543–6767

DOI: 10.5603/gpl.94956

Ultrasonographic diagnosis of ovarian tumors through the deep convolutional neural network

Min Xi[1*], Runan Zheng[2*], Mingyue Wang[1], Xiu Shi[1], Chaomei Chen[1], Jun Qian[1], Xinxian Gu[3], Jinhua Zhou[1]

[1]The First Affiliated Hospital of Soochow University, Suzhou City, China
[2]Suzhou MicroClear Medical Ltd., Suzhou City, China
[3]Dushu Lake Hospital Affiliated to Soochow University, Suzhou City, China
**These authors contributed equally to this work*

ABSTRACT

Objectives: The objective of this study was to develop and validate an ovarian tumor ultrasonographic diagnostic model based on deep convolutional neural networks (DCNN) and compare its diagnostic performance with that of human experts.

Material and methods: We collected 486 ultrasound images of 192 women with malignant ovarian tumors and 617 ultrasound images of 213 women with benign ovarian tumors, all confirmed by pathological examination. The image dataset was split into a training set and a validation set according to a 7:3 ratio. We selected 5 DCNNs to develop our model: MobileNet, Xception, Inception, ResNet and DenseNet. We compared the performance of the five models through the area under the curve (AUC), sensitivity, specificity, and accuracy. We then randomly selected 200 images from the validation set as the test set. We asked three expert radiologists to diagnose the images to compare the performance of radiologists and the DCNN model.

Results: In the validation set, AUC of DenseNet was 0.997 while AUC was 0.988 of ResNet, 0.987 of Inception, 0.968 of Xception and 0.836 of MobileNet. In the test set, the accuracy was 0.975 with the DenseNet model vs 0.825 ($p < 0.0001$) with the radiologists, and sensitivity was 0.975 vs 0.700 ($p < 0.0001$), and specificity was 0.975 vs 0.908 ($p < 0.001$).

Conclusions: DensNet performed better than other DCNNs and expert radiologists in identifying malignant ovarian tumors from benign ovarian tumors based on ultrasound images, a finding that needs to be further explored in clinical trials.

Keywords: ultrasound; diagnosis; ovarian tumor; deep learning; radiologist

Ginekologia Polska 2024; 95, 3: 181–189

INTRODUCTION

Ovarian cancer is one of the deadliest gynecological malignancies. According to the Global Cancer Statistics 2020 [1], it is estimated to be 313 959 new cases and 207 252 deaths of ovarian cancer in 2020 in the world. The global 5-year survival is below 45% [2]. Ovarian cancer generally affects women over 50. The treatment is based on surgery and chemotherapy.

Ultrasound examination is the most appropriate first-line diagnostic technique for the preoperative evaluation of women with adnexal lesions. Ovarian cancer is usually diagnosed through ultrasound features such as the shape of the pelvic mass, the proportion of solid tissue, the presence of ascites, the number of papillary projections, and blood flow signals [3]. Whether a pelvic mass is benign or malignant, an expert radiologist discriminates through these features. Radiologists are limited in their abilities, and their judgment is subject to the influence of their working experience [4]. The accuracy of discriminating a pelvic mass through ultrasound by radiologists is approximately 82–92% [5]. Therefore, it is necessary to improve the precision of ultrasonographic diagnosis of ovarian tumors.

With the rapid development of artificial intelligence, the technique of computer-assisted image diagnosis in medicine has made substantial strides in the area of image--recognition [6, 7]. Recent advances in deep convolutional

Corresponding author:
Xinxian Gu, Dushu Lake Hospital Affiliated to Soochow University, Suzhou City, China, email: guxinxian@suda.edu.cn
and Jinhua Zhou, The First Affiliated Hospital of Soochow University, Suzhou City, China, e-mail: jsjhzh@126.com

Received: 1.04.2023 Accepted: 21.08.2023 Early publication date: 13.10.2023

neural networks (DCNN) have shown great promise for ultrasound diagnosis of diseases such as thyroid nodules and breast nodules [8, 9]. However, studies on ultrasonographic diagnosis of ovarian tumors through DCNN are few so far [10]. In contrast to typical machine learning algorithms, DCNN does not employ features that human experts identified as input. By taking raw image pixels and the corresponding class labels as inputs, DCNN automatically learns feature representations in a generalized manner [11].

One of the main challenges of DCNN models is vanishing gradients. A practical solution is to increase the connection between layers. This problem was overcome in some DCNN models such as ResNet [12], Highway Networks [13], and Stochastic depth [14]. Although these algorithms have different network structures, they all take advantage of short paths to link early and later layers. Therefore, we used the concept of DenseNet [15] to design our model architecture. A DenseNet network is an improved DCNN model that continues the idea mentioned above by directly connecting all layers to ensure maximum information flow between layers, using a shortcut connection to pass input from one block to another. Thus, DenseNet may offer great help for diagnosis of image-based examinations in clinical work.

Objectives

In this study, we aim to develop a DCNN-based ultrasound image analysis model and evaluate its performance for the automated diagnosis of ovarian tumors using real-world ultrasound images compared with human radiologists.

MATERIAL AND METHODS

Dataset

We retrospectively collected ultrasound images of ovarian tumors from the First Affiliated Hospital of Soochow University between May 1st, 2017, to June 30th, 2020. Patients were included based on the following two eligibility criteria. The first requirement was that they were at least 18 years old. Secondly, all patients with benign or malignant ovarian tumors underwent a pathological examination. The pathological examination reports were provided by the pathological department of the First Affiliated Hospital of Soochow University.

If the patients fulfilled the inclusion criteria, ultrasound images within 120 days before the surgery were collected. The ultrasound imaging was manufactured by GE Healthcare system. Image quality control was performed by excluding images not containing tumor nidus based on the pathological review report, such as the uterus and opposite normal ovaries. The images were all in jpg format. As a final step, we established our image dataset of 1103 ultrasound images, including 486 images of malignant ovarian tumors from 192 patients and 617 images of benign ovarian tumors from 213 patients.

The construction of the DCNN models

The dataset was split into a training set and a validation set at random in a 7:3 ratio. The training set was utilized to learn the parameters of the ultrasound images, and the validation set was used to estimate the prediction error for hyperparameter tuning and model selection. The training set consisted of 340 images of malignant ovarian tumors and 432 images of benign ovarian tumors. The validation set consisted of 146 images of malignant ovarian tumors and 185 of benign ovarian tumors. Our training dataset was augmented with image data to increase training data and avoid overfitting artificially [16]. Image augmentation was not applied to the validation set. It is reported that after adopting the data enhancement method, the accuracy of the final recognition results can be improved by 3–4% [17]. We used the following methods to effectively enhance the ultrasound image data, including rotation ± 20°, horizontal translation 20°, vertical translation 20°, zoom 20%, and horizontal flip. The effect of augmentation of specific data is shown in Figure S1.

Afterwards, we selected five different DCNNs to develop our diagnostic models, including Inception [18], Mobilenet [19], Resnet, Xception [20], and DenseNet. We trained 50 rounds on the training set and evaluated the DCNN models using the training set. The output of the last layer was shown as the predicted probability of malignancy.

All experiments were conducted on a device with a Windows 10 system. The hardware capabilities included NVIDIA RTX 3080 GPU (10 GB memory), CPU AMD 5600X, and 32 GB RAM. In the experiment process, the size of all the images was set at 299 × 299 mm. We set the batch size to 16 due to GPU memory limitations. All programs were implemented by TensorFlow and Keras. The optimizer was Stochastic Gradient Descent, and the initial learning rate was 0.001. The momentum was 0.9, and the weight decay was 0.0001. We set the epoch at 50. Moreover, the warm-up was employed during the training process. A stable distribution could aid in maintaining the deep stability of the model, which could help to slow down the early overfitting of the mini-batch at the start of the model.

Comparison with radiologists

Futhermore, DenseNet showed the best performance among the five DCNN models and was used to compare whether the DCNN model has advantages over human radiologists in recognizing malignant ovarian tumors. Then, we randomly selected 200 images from the validation set as the test set. Three expert radiologists were invited to analyze the 200 images and determine whether they were malignant. The performance of human radiologists was

then compared with the DenseNet model on the test set. All radiologists had working experience more than six years and were required to complete the task within two hours independently.

Statistical analysis

The predictions of DCNN models and radiologists were compared with the pathological reports, considered the diagnostic gold standard. We applied the receiver operating characteristic (ROC) curve to compare the diagnostic abilities of different DCNN models in discriminating malignant ovarian tumors from benign ones. The ROC curve was drawn by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) by varying the predicted probability threshold, and the area under the curve (AUC) was calculated. We also calculated the accuracy, sensitivity, specificity, positive predicted values (PPV) and negative predictive values (NPV) to assess the diagnostic abilities of different DCNN models and radiologists. Sensitivity is the fraction of recognizing malignancies in the malignant data verified by pathological examination. Specificity is the fraction of recognizing benignities in benign data verified by pathological examination. Accuracy is the fraction of recognizing malignant/benign data in malignant/benign data verified by pathological examination. PPV is the fraction of malignancies verified by pathological examination in malignancies diagnosed by DCNN models or radiologists. NPV is the fraction of benignities verified by pathological examination in benignities diagnosed by DCNN models or radiologists. We calculated 95% confidence intervals (CIs) for sensitivity, specificity, accuracy, PPV, and NPV with the Clopper–Pearson method [21]. We also calculated kappa values and F1 scores. Kappa value measures the agreement between the prediction of one diagnostic method and the pathological reports. F1 score was calculated as the harmonic mean of sensitivity and PPV, which measures the accuracy of one diagnostic method against the pathological report.

We used the radiologists' average sensitivity, specificity, and accuracy when comparing the performance with the DenseNet model. A binomial test was applied to evaluate the difference in sensitivity, specificity, and accuracy between the DenseNet model and the radiologists. A p value less than 0.05 was considered statistically significant. The inter-radiologist agreement rate and Fleiss' kappa value [22] were also calculated. The figure plotting and statistical analyses were done with GraphPad Prism (version 8.0) and R software (version 4.0.3).

The flowchart depicting the process of our study is shown in Figure 1.

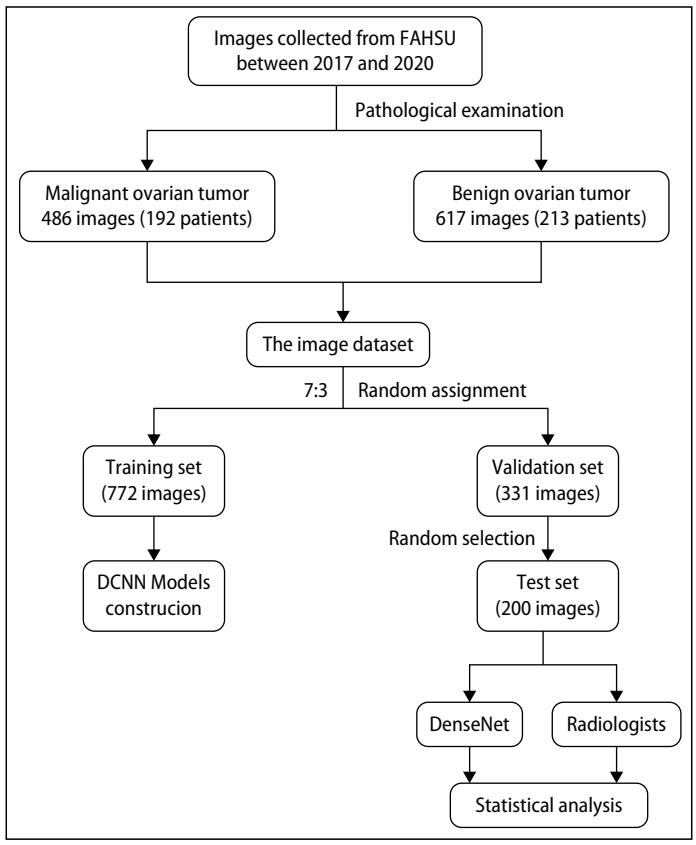

Figure 1. The flowchart of the study; FAHSU — the First Affiliated Hospital of Soochow University

Table 1. Baseline characteristics of the study. Data are n [%] or median (IQR)

	Training set (n = 282)		Validation set (n = 123)	
	Malignant group	Benign group	Malignant group	Benign group
Patients	137	145	55	68
Images	340	432	146	185
Age [years]	55 (49–64)	35 (30–45)	56 (49–66)	33 (28–44)
≤ 45 years	18 (13.1%)	109 (75.2%)	10 (18.2%)	52 (76.5%)
> 45 years	119 (86.9%)	36 (24.8%)	45 (81.8%)	16 (23.5%)
Histology				
Serous	95 (69.3%)	NA	46 (83.6%)	NA
Mucinous	9 (6.6%)	NA	3 (5.5%)	NA
Endometrioid	13 (9.5%)	NA	1 (1.8%)	NA
Clear cell	13 (9.5%)	NA	2 (3.6%)	NA
Others	7 (5.1%)	NA	3 (5.5%)	NA
FIGO				
Stage I	32 (23.4%)	NA	8 (14.5%)	NA
Stage II	16 (11.7%)	NA	9 (16.4%)	NA
Stage III	66 (48.2%)	NA	23 (41.8%)	NA
Stage IV	23 (16.8%)	NA	15 (27.3%)	NA

FIGO — International Federation of Gynecology and Obstetrics; NA — not applicable

Table 2. Performance of different different deep convolutional neural network (DCNN) models, assessed on the validation set. Data are n (95% CI)

	MobileNet	Xception	Inception	ResNet	DenseNet
Sensitivity	0.747 (0.668–0.815)	0.863 (0.796–0.914)	0.973 (0.931–0.991)	0.945 (0.895–0.976)	0.952 (0.904–0.981)
Specificity	0.795 (0.729–0.850)	0.941 (0.896–0.970)	0.849 (0.789–0.897)	0.957 (0.917–0.981)	0.973 (0.938–0.991)
Accuracy	0.773 (0.724–0.817)	0.906 (0.870–0.935)	0.903 (0.866–0.933)	0.952 (0.923–0.972)	0.964 (0.938–0.981)
Positive predictive value	0.741 (0.663–0.810)	0.920 (0.861–0.959)	0.835 (0.771–0.888)	0.945 (0.895–0.976)	0.965 (0.921–0.989)
Negative predictive value	0.799 (0.734–0.854)	0.897 (0.845–0.936)	0.975 (0.938–0.993)	0.957 (0.917–0.981)	0.963 (0.924–0.985)
Kappa	0.541	0.809	0.807	0.902	0.926
F1	0.744	0.890	0.899	0.945	0.959

CI — confidence interval; DCNN — deep convolutional neural networks

RESULTS

The baseline characteristics of the training set and the validation set are shown in Table 1. The median age of participants showed no apparent differences between the training set and the validation set, while the median age was higher in the malignant group than in the benign group [55 years (IQR 49–64) vs 35 years (30–45) in training set; 56 years (IQR 49–66) vs 33 years (28–44) in the validation set]. Since malignant ovarian tumors usually occur in older women, the proportion of participants over 45 was 86.9% in the malignant group, while the proportion was only 24.8% in the benign group in the training set. The age of onset was similar in the validation set. There were no significant differences between the training and validation sets regarding the histology of malignant ovarian tumors. Most of the participants were at stage III or IV, according to the International Federation of Gynecology and Obstetrics (FIGO).

The performance of different DCNN models on the validation set after 50 rounds of training is shown in Table 2, and the corresponding ROC curves are shown in Figure 2A. As the ROC curves show, the DenseNet model achieved the best performance in identifying benign malignant ovarian tumors in the validation set, with AUC of 0.997 (95% CI 0.995–1.000). AUC were 0.988 (0.980–0.997) of ResNet, 0,987 (0.978–0.996) of Inception, 0.968 (0.952–0.984) of Xception and 0.836 (0.792–0.880) of MobileNet. Moreover, the DenseNet model achieved the highest accuracy, sensitivity, specificity, PPV, and NPV on the validation

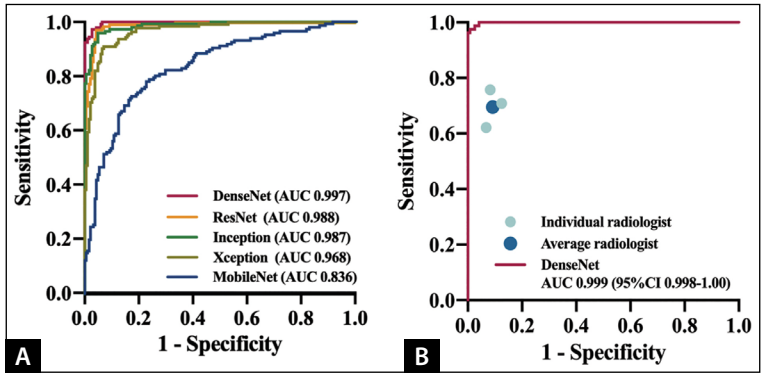

Figure 2. Performance of different deep convolutional neural network (DCNN) models and the radiologists in discriminating malignant ovarian tumors from benign ones; **A.** The receiver operating characteristic (ROC) curves for the performance of different DCNN models in the validation set; **B.** ROC for the performance of the DenseNet model versus 3 radiologists in the test set; AUC — area under the curve

Table 3. Performance of the DenseNet model versus radiologists, assessed on the test set

	Radiologist 1	Radiologist 2	Radiologist 3	Radiologist' mean	DenseNet	p value
Sensitivity	0.625 (0.510–0.731)	0.763 (0.654–0.851)	0.713 (0.600-0.808)	0.700 (0.587–0.797)	0.975 (0.913–0.997)	< 0.0001
Specificity	0.933 (0.873–0.971)	0.917 (0.852–0.959)	0.875 (0.802-0.928)	0.908 (0.842–0.953)	0.975 (0.929–0.995)	< 0.001
Accuracy	0.810 (0.749–0.862)	0.855 (0.798–0.901)	0.810 (0.749-0.862)	0.825 (0.765–0.875)	0.975 (0.943–0.992)	< 0.0001
Positive predictive value	0.862 (0.746–0.939)	0.859 (0.756–0.930)	0.792 (0.680-0.878)	0.836 (0.725–0.915)	0.963 (0.896–0.992)	
Negative predictive value	0.789 (0.712–0.853)	0.853 (0.780–0.909)	0.820 (0.743-0.883)	0.820 (0.744–0.881)	0.983 (0.941–0.998)	
Kappa	0.585	0.692	0.597	0.625	0.948	
F1	0.725	0.808	0.750	0.762	0.969	

set. For the DenseNet model, accuracy was 0.964 (0.938–0.981), sensitivity was 0.952 (0.904–0.981), specificity was 0.973 (0.938–0.991), PPV was 0.965 (0.921–0.989), and NPV was 0.963 (0.924–0.985). For the ResNet model, accuracy was 0.952 (0.923–0.972), sensitivity was 0.945 (0.895–0.976), specificity was 0.957 (0.917–0.981), PPV was 0.945 (0.895–0.976), and NPV was 0.957 (0.917–0.981). For the Inception model, accuracy was 0.903 (0.866–0.933), sensitivity was 0.973 (0.931–0.991), specificity was 0.849 (0.789-0.897), PPV was 0.835 (0.771–0.888), and NPV was 0.975 (0.938–0.993). For the Xception model, accuracy was 0.906 (0.870–0.935), sensitivity was 0.863 (0.796–0.914), specificity was 0.941 (0.896–0.970), PPV was 0.920 (0.861–0.959), and NPV was 0.897 (0.845–0.936). For the MobileNet model, accuracy was 0.773 (0.724–0.817), sensitivity was 0.747 (0.668–0.815), specificity was 0.795 (0.729–0.850), PPV was 0.741 (0.663–0.810), and NPV was 0.799 (0.734–0.854). Furthermore, the DenseNet model also had a higher kappa coefficient and F1 score than other DCNN models. From the above results, the DenseNet model has the best diagnostic capability compared to other DCNN models.

The performance of DenseNet versus the expert radiologists in the test set is shown in Figure 2B and Table 3. In the test set, the AUC value of the DenseNet model was 0.999 (95% CI 0.998–1.000). Among the radiologists, accuracy ranged from 0.810 (0.749–0.862) to 0.855 (0.798–0.901), sensitivity ranged from 0.625 (0.510–0.731) to 0.763 (0.654–0.851), specificity ranged from 0.875 (0.802–0.928) to 0.933 (0.873–0.971), PPV ranged from 0.792 (0.680–0.878) to 0.862 (0.746–0.939), and NPV ranged from 0.789 (0.712–0.853) to 0.853 (0.780–0.909). The interradiologist agreement rate was 0.735 (95% CI 0.668–0.795; Fleiss' kappa 0.667). Compared with the expert radiologists, the DenseNet model achieved higher performance in discriminating malignant ovarian tumors from benign ones. The accuracy was 0.975 (0.943–0.992) with the DenseNet model vs 0.825 (0.765–0.875; $p < 0.0001$) with the radiologists, and sensitivity was 0.975 (0.913–0.997) vs 0.700 (0.587–0.797; $p < 0.0001$), and specificity was 0.975 (0.929–0.995) vs 0.908 (0.842–0.953; $p < 0.001$). Furthermore, the DenseNet model also had higher PPV, NPV, kappa coefficient, and F1 score compared with the performance of the radiologists.

The ultrasound images misdiagnosed by DenseNet are shown in Figure S2. The confusion matrices reporting the number of true positive, false positive, false negative and true negative results achieved by Inception, MobileNet,

ResNet, Xception, DenseNet, and the radiologists are shown in Table S1 and Table S2.

DISCUSSION

In this study, an automatic DCNN model was developed and validated to discriminate malignant from benign tumors of the ovary on ultrasound images. According to the above results, DenseNet performed better than other DCNN models in the validation set with respect to AUC, accuracy, sensitivity, and specificity. Consequently, DenseNet was selected for the comparison with expert radiologists. The diagnostic capability of the DenseNet model significantly exceeded the average level of radiologists.

At present, studies on the application of deep learning in ovarian cancer are limited. The application fields include diagnosis, pathological classification and prognostic prediction. Meanwhile, magnetic resonance imaging and ultrasonography essentially take equal share of studies focusing on image recognition of ovarian tumor through deep learning. By February 2023, only 6 articles [23–28] on ultrasonographic diagnosis of ovarian tumor through deep learning were retrieved. A retrospective single-center study in South Korea [23] constructed a CNN-CAE model to make diagnoses through ultrasound images of ovarian tumors. The model consisted of two parts. The first part could automatically remove interfering information such as characters and rulers on ultrasound images through the CAE program, and the second part was the DenseNet model, which was used for image diagnosis. The accuracy of CNN-CAE model was 0.972 in distinguishing ovarian tumors from normal ovarian tissues, and the accuracy was 0.901 in recognizing malignant ovarian tumors. Another study from Taiwan, China [24] tested the performances of ten common DCNN models, and three of them with the highest accuracy (ResNet-18, ResNet-50 and Xception) were selected to construct an assembled diagnostic model. The average accuracy of the assembled model reached 0.922. However, none of the above deep learning models have been compared with the diagnostic performance of expert radiologists. Chen et al. [25] included a number of ultrasound images from 422 patients with ovarian tumors and trained two deep learning models based on ResNet, $DL_{decidion}$ and $DL_{feature}$. Then, the two models were compared with radiologists and the Ovarian-Adnexal Reporting and Data System (O-RADS). However, $DL_{decidion}$ and $DL_{feature}$ did not show superior diagnostic performances than radiologists and O-RADS. Radiologists from Shanghai represents the highest diagnostic level in China to a certain extent.

Another multicenter retrospective study [26] involving 106,400 patients showed that the AUC of the DenseNet-121 model reached 0.911 in the internal validation set, as well as 0.870 and 0.831 in the two external validation sets. With the assistance of the DCNN model, the average diagnostic accuracy of radiologists was improved from 0.783 to 0.876, revealing the great potential of DCNN model in the assistance of image diagnosis.

Since ultrasound examination is the most crucial assistant examination in the diagnosis of ovarian lesions, the accurate recognition of ovarian malignant tumors is dispensable. However, the discrimination of ovarian tumors is entirely up to radiologists, leading to subjective mistakes in accurate recognition and consistent interpretation of ovarian tumors by radiologists, as shown by the inter-radiologist agreement rate in the test set. Nevertheless, DenseNet is highly robust and can significantly avoid this defect since it learns the feature representations without subjectivity [29]. Thus, diagnostic consistency and reproducibility could be maintained by the DenseNet model effectively. On the one hand, fresh radiologists, without much experience, may be able to improve the accuracy of their diagnoses using the DenseNet model [27]. On the other hand, two radiologists are required to perform the ultrasonographic diagnosis during clinical work, one with less experience assessing the images to reach a primary diagnosis, and the other with more experience responsible for checking the primary diagnosis and offering the conclusion. The DenseNet model may relieve labor requirement, which may offer great help to remote areas in the lack of medical resources.

Furthermore, the DenseNet model has great application potential. Firstly, the DenseNet model works well for other diseases in addition to ovarian tumors, as mentioned above. Moreover, it could be applied not only in ultrasound examination but also in computerized tomography (CT), magnetic resonance imaging, retinal fundus photographs and other examinations requiring image generation [30, 31]. Finally, because the DenseNet model report is instantaneous, the diagnosis model may be integrated into the ultrasound workstations, creating a real-time diagnosis of dynamic images.

However, our study has some limitations. Firstly, we did not set up external validation sets. Secondly, we excluded patients with borderline ovarian tumors because it may lead to confusion of features between samples. And the sample size of patients with borderline ovarian tumors is too small for DCNN to obtain enough effective features to ensure the accuracy of the DenseNet model. Lastly, the three radiologists were asked to make their judgments through only one single ultrasound image in our study. However, in the real world, radiologists usually make a comprehensive judgment by referring to more than one image. Not only that, but the blood flow signals also help them make diagnoses. Therefore, the diagnostic accuracy of human radiologists based on multi-modality data would likely be higher than the performance of DCNN.

CONCLUSIONS

To conclude, the Densnet model is valuable despite its limitations. In future, we plan to include more ultrasound images from external medical centers. We will also make efforts to refine our diagnostic model of ovarian tumors. And we hope our study will make a step to improve the accuracy of the diagnoses of ovarian tumors and to help the realization of AI-assisted ultrasonographic diagnoses in clinical work, which could bring benefit to both the patients and the radiologists.

Article information and declarations

Informed consent statement

Informed consent from patients with ovarian tumors was waived as the study design was based on a retrospective review of medical records and ultrasound images.

Ethics statement

This study was approved by the ethical committee of the First Affiliated Hospital of Soochow University in accordance with the principles of the Declaration of Helsinki [No:(2023)033, 2023/01/31].

Funding

This work was supported by the National Natural Science Foundation of China (No. 82172609, No. 82202898), the Jiangsu Province Sci-Tech Plan Special Fundation (No. BE2022729), the Foundation of Jiangsu Province Engineering Research Center of Precision Diagnostics and Therapeutics Development (No. SDGC2242).

Acknowledgements

All listed authors have made essential contributions to the work. Min Xi collected the data, performed the data analysis and drafted the original manuscript. Jun Qian and Chaomei Chen helped collect the data. Runan Zheng wrote DCNN program. Xinxian Gu and Mingyue Wang helped filtrate the ultrasound images and judge on the test set. Jinhua Zhou and Xiu Shi supervised the project design and reviewed the manuscript. All authors have read and approved the final manuscript for publication.

Conflict of interest

The authors declare no conflict of interest.

Supplementary material

Figure S1, S2, Table S1, S2.

REFERENCES

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021; 71(3): 209–249, doi: 10.3322/caac.21660, indexed in Pubmed: 33538338.
2. Webb PM, Jordan SJ. Epidemiology of epithelial ovarian cancer. Best Pract Res Clin Obstet Gynaecol. 2017; 41: 3–14, doi: 10.1016/j.bpobgyn.2016.08.006, indexed in Pubmed: 27743768.
3. Froyman W, Timmerman D. Methods of assessing ovarian masses: international ovarian tumor analysis approach. Obstet Gynecol Clin North Am. 2019; 46(4): 625–641, doi: 10.1016/j.ogc.2019.07.003, indexed in Pubmed: 31677746.
4. Van Holsbeke C, Daemen A, Yazbek J, et al. Ultrasound experience substantially impacts on diagnostic performance and confidence when adnexal masses are classified using pattern recognition. Gynecol Obstet Invest. 2010; 69(3): 160–168, doi: 10.1159/000265012, indexed in Pubmed: 20016188.
5. Timmerman D, Schwärzler P, Collins WP, et al. Subjective assessment of adnexal masses with the use of ultrasonography: an analysis of interobserver variability and experience. Ultrasound Obstet Gynecol. 1999; 13(1): 11–16, doi: 10.1046/j.1469-0705.1999.13010011.x, indexed in Pubmed: 10201081.
6. Bi WL, Hosny A, Schabath MB, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. CA Cancer J Clin. 2019; 69(2): 127–157, doi: 10.3322/caac.21552, indexed in Pubmed: 30720861.
7. Hosny A, Parmar C, Quackenbush J, et al. Artificial intelligence in radiology. Nat Rev Cancer. 2018; 18(8): 500–510, doi: 10.1038/s41568-018-0016-5, indexed in Pubmed: 29777175.
8. Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. Lancet Oncol. 2019; 20(2): 193–201, doi: 10.1016/S1470-2045(18)30762-9, indexed in Pubmed: 30583848.
9. Li J, Bu Y, Lu S, et al. Development of a deep learning-based model for diagnosing breast nodules with ultrasound. J Ultrasound Med. 2021; 40(3): 513–520, doi: 10.1002/jum.15427, indexed in Pubmed: 32770574.
10. Xu HL, Gong TT, Liu FH, et al. Artificial intelligence performance in image-based ovarian cancer identification: A systematic review and meta-analysis. EClinicalMedicine. 2022; 53: 101662, doi: 10.1016/j.eclinm.2022.101662, indexed in Pubmed: 36147628.
11. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521(7553): 436–444, doi: 10.1038/nature14539, indexed in Pubmed: 26017442.
12. He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA, 2016.
13. Rupesh KS, Klaus G,Jürgen S. Highway Networks. In Proceedings of the International Conference on Machine Learning. Lille, France, 2015.
14. Huang G, Sun Yu, Liu Z, et al. Deep networks with stochastic depth. Computer Vision – ECCV 2016. 2016: 646–661, doi: 10.1007/978-3-319-46493-0_39.
15. Huang G, Liu Z, Van ML, et al. Densely Connected Convolutional Networks. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA, 2017.
16. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017; 542(7639): 115–118, doi: 10.1038/nature21056, indexed in Pubmed: 28117445.
17. Forrest I, Matt M, Sergey K, et al. DenseNet: implementing efficient ConvNet descriptor pyramids. Computer Science. 2014.
18. Szegedy C, Vanhoucke V, Iofe S, et al. Rethinking the inception architecture for computer vision. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA, 2016.
19. Howard AG, Zhu M, Chen B, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv 2017.
20. Chollet F. Xception: Deep learning with depthwise separa-ble convolutions. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA, 2017.
21. CLOPPER CJ, PEARSON ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika. 1934; 26(4): 404–413, doi: 10.1093/biomet/26.4.404.
22. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971; 76(5): 378–382, doi: 10.1037/h0031619.
23. Jung Y, Kim T, Han MR, et al. Ovarian tumor diagnosis using deep convolutional neural networks and a denoising convolutional autoencoder. Sci Rep. 2022; 12(1): 17024, doi: 10.1038/s41598-022-20653-2, indexed in Pubmed: 36220853.
24. Hsu ST, Su YJ, Hung CH, et al. Automatic ovarian tumors recognition system based on ensemble convolutional neural network with ultrasound imaging. BMC Med Inform Decis Mak. 2022; 22(1): 298, doi: 10.1186/s12911-022-02047-6, indexed in Pubmed: 36397100.

25. Chen H, Yang BW, Qian Le, et al. Deep learning prediction of ovarian malignancy at US compared with O-RADS and expert assessment. Radiology. 2022; 304(1): 106–113, doi: 10.1148/radiol.211367, indexed in Pubmed: 35412367.
26. Gao Y, Zeng S, Xu X, et al. Deep learning-enabled pelvic ultrasound images for accurate diagnosis of ovarian cancer in China: a retrospective, multicentre, diagnostic study. Lancet Digit Health. 2022; 4(3): e179–e187, doi: 10.1016/S2589-7500(21)00278-8, indexed in Pubmed: 35216752.
27. Christiansen F, Epstein EL, Smedberg E, et al. Ultrasound image analysis using deep neural networks for discriminating between benign and malignant ovarian tumors: comparison with expert subjective assessment. Ultrasound Obstet Gynecol. 2021; 57(1): 155–163, doi: 10.1002/uog.23530, indexed in Pubmed: 33142359.
28. Wang H, Liu C, Zhao Z, et al. Application of deep convolutional neural networks for discriminating benign, borderline, and malignant serous ovarian tumors from ultrasound images. Front Oncol. 2021; 11: 770683, doi: 10.3389/fonc.2021.770683, indexed in Pubmed: 34988015.
29. Gu J, Wang Z, Kuen J, et al. Recent advances in convolutional neural networks. Pattern Recognition. 2018; 77: 354–377, doi: 10.1016/j.patcog.2017.10.013.
30. Akazawa M, Hashimoto K. Artificial intelligence in gynecologic cancers: Current status and future challenges — A systematic review. Artif Intell Med. 2021; 120: 102164, doi: 10.1016/j.artmed.2021.102164, indexed in Pubmed: 34629152.
31. Guo C, Yu M, Li J. Prediction of different eye diseases based on fundus photography via deep transfer learning. J Clin Med. 2021; 10(23), doi: 10.3390/jcm10235481, indexed in Pubmed: 34884192.

SUPPLEMENTARY MATERIALS

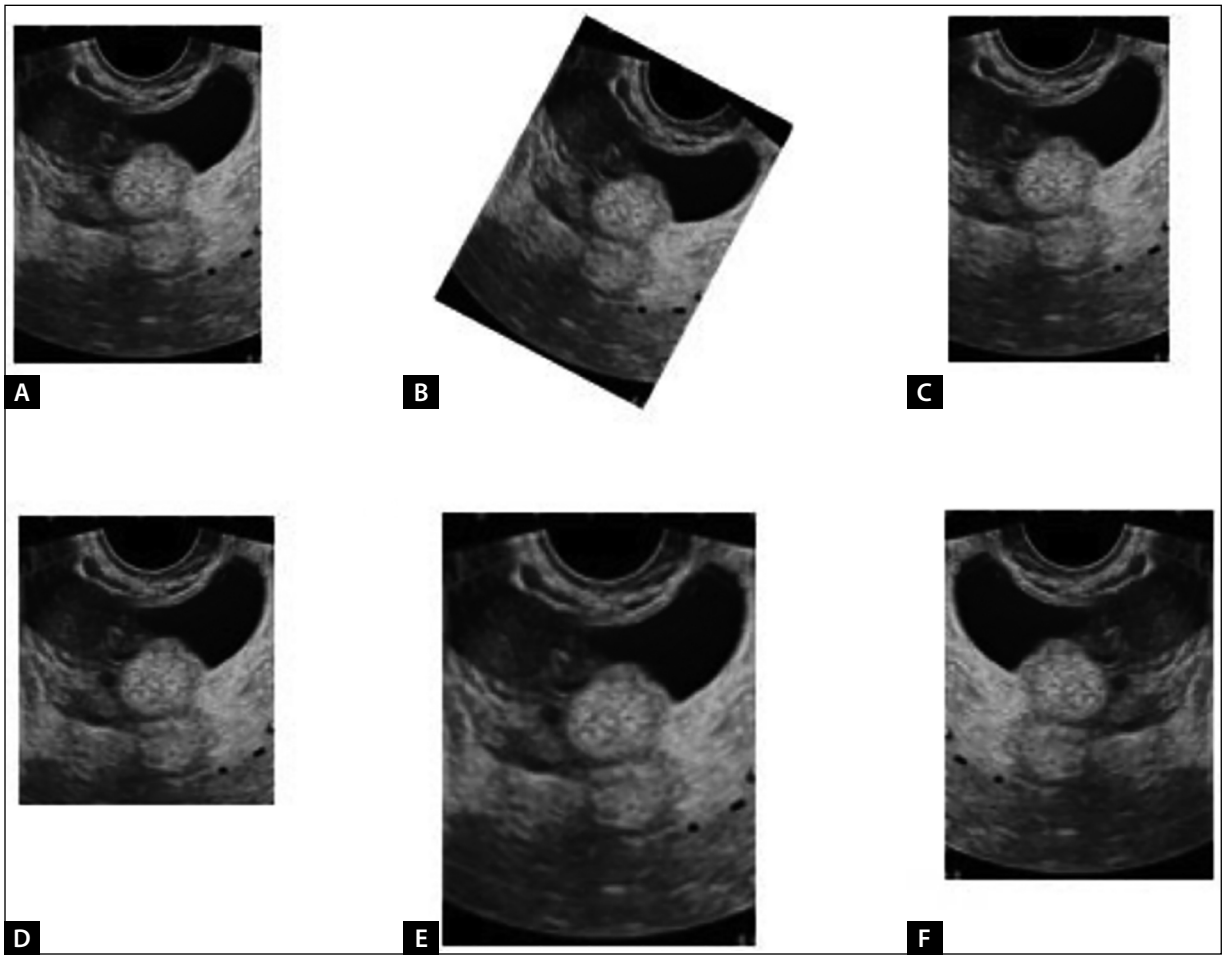

Figure S1. Data augmentation effect; **A.** Original image; **B.** Rotate; **C.** Horizontal translation; **D.** Vertical translation; **E.** Zoom; **F.** Horizontal flip

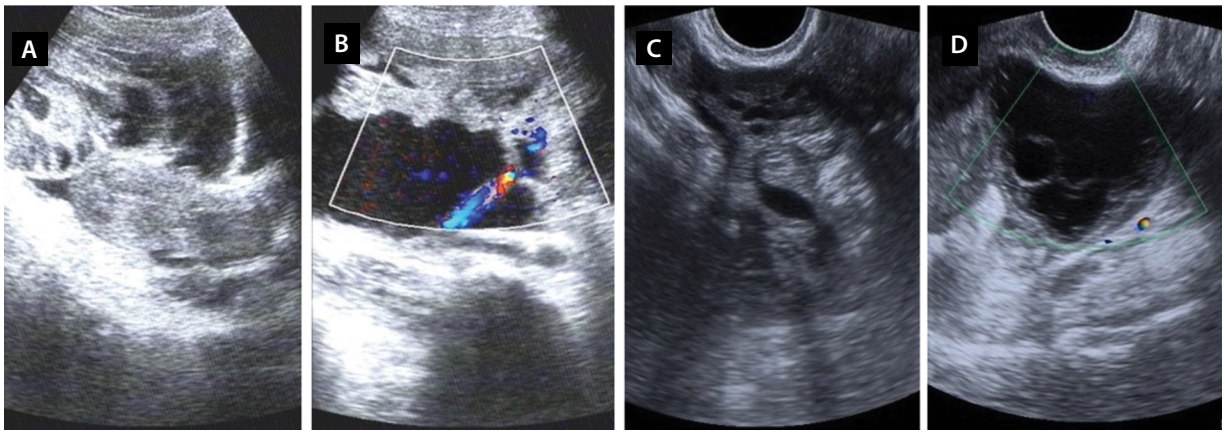

Figure S2. Images misdiagnosed by the DenseNet Model; **A, B.** Malignant images classified as be-nign; **C, D.** Benign images classified as malignant

Table S1. Confusion matrices of different different deep convolutional neural network (DCNN) models on the validation set

	MobileNet		Xception		Inception	
	Truth		Truth		Truth	
Prediction	Malignancy	Benign	Malignancy	Benign	Malignancy	Benign
Malignancy	109	38	126	11	142	28
Benign	37	147	20	174	4	157
	ResNet		DenseNet			
	Truth		Truth			
Prediction	Malignancy	Benign	Malignancy	Benign		
Malignancy	138	8	139	5		
Benign	8	177	7	180		

Table S2. Confusion matrices of radiologists and DenseNet on the test set

	Radiologist 1		Radiologist 2		Radiologist 3	
	Truth		Truth		Truth	
Prediction	Malignancy	Benign	Malignancy	Benign	Malignancy	Benign
Malignancy	50	8	61	10	57	15
Benign	30	112	19	110	23	105
	DenseNet					
	Truth					
Prediction	Malignancy	Benign				
Malignancy	78	3				
Benign	2	117				