Real-time artificial intelligence for detection of Fetal Intracranial malformations in Ultrasonic images: A multicenter retrospective diagnostic study

Mei-Fang Lin[1], Xiaoqin He[2], Chun Hao[2], Miao He[2], Hongmei Guo[2], Lihe Zhang[2], Jianbo Xian[2], jv zheng[2], Qiuhong Xu[2], Jieling Feng[2], Yongzhong Yang[2], Nang Wang[2], and Hong-Ning Xie[1]

[1]Sun Yat-sen University First Affiliated Hospital
[2]Affiliation not available

May 6, 2020

Abstract

Objective: To develop an artificial intelligence (AI) model to detect congenital central nervous system (CNS) malformations in fetal cerebral-cranial ultrasound images, and to assess the efficacy of this algorithm in improving clinical doctors' diagnostic performance. Design: Retrospective, multicenter, diagnostic study Setting: Three Chinese hospitals Population: a cohort of 2397 fetuses with CNS malformations and 11316 normal fetuses. Methods: AI model was developed by training on 37450 images from 15264 fetuses and testing on 812 images from 449 fetuses. Three groups of doctors (trainee, competent, expert) were equipped with the AI system to test its enhancement of diagnosis performance. Main outcome measures: Diagnostic performance of AI model and that of doctors. Comparison of performance between AI model and doctors, and doctors with and without AI assistance. Results: The performance of AI model was comparable to that of expert in identifying 12 types of CNS malformations in terms of accuracy 79.8% (95% CI 77.0-82.6%) versus 78.9% (95% CI 75.2-85.2%), sensitivity 78.4% (75.3-81.3%) versus 77.5% (73.7-81.4%) , specificity of 94.4% (86.2-98.4%) versus 93.0% (84.1-100.0%), and AUC 0.864 (0.833-0.895) versus 0.853 (0.800-0.905). This AI model improved doctors' diagnostic performances, the trainee group received maximum improvement, whose diagnostic performance advanced to the level of expert group in terms of accuracy (80.2%, 95% CI 75.0-85.3%) and AUC (0.872, 95% CI 0.861-0.882). Conclusions: Our AI system achieved a high diagnostic performance comparable with that of experienced doctors and can support unexperienced doctors by improving their diagnostic accuracy to an expert-level.

Introduction

Congenital malformations are the leading cause of fetal loss and one of the top ten causes of mortality in children under five[1, 2]. It also accounted for 25-38 million disability-adjusted life-years worldwide[3], which causes heavy burden on individuals, families, health-care systems, and societies[4]. There are substantial inter-country differences worldwide in the reported prevalence of congenital malformations partly due to the unequal capacities of prenatal screening, leaving many cases undetected, especially in underdeveloped regions. For example, the reported prevalence of congenital cerebral anomalies in Europe increased by 2.4% per annum, but a six-fold difference was found in prevalence across different regions, with an association between prevalence and prenatal detection rate[5]. Therefore, early identification of congenital anomalies with efficiency is crucial in ensuring medical intervention, minimizing world healthcare disparity, and eventually leading to the optimization of healthcare resources. This goal calls for not only the detection equipment but also doctor expertise for prenatal diagnosis. Yet, training doctors is a timely and costly process, which causes enormous expense to provide prenatal surveillance for average citizens all over the world.

The implementation of artificial intelligence (AI) systems has shown its potential to revolutionize disease diagnosis by performing classification difficult for human experts [6-11]. The performance of most reported AI shows a promising trend[12-18], furthermore, it has significant advantages in terms of convenient open-source sharing, which have the potential to provide medical guidance to multiple hospitals simultaneously, especially for less developed and remote areas [19,20]. In the field of fetal congenital malformation diagnosis, AI development involved the differentiation of images of normal and abnormal fetuses was rare, only limited progress in AI-assisted fetal ultrasound identification of normal fetus structure were reported[14-18] , these studies laid a foundation for the development of AI system to identify abnormal structure in ultrasound images by training on fetuses with congenital malformation.

We have initially constructed an AI system involving abnormal fetal CNS ultrasound images to classify fetal CNS ultrasound images as either normal or abnormal and our system achieved a high performance[21]. Nonetheless, this system only classified images to provide binary outcomes, it is far from making diagnosis for specific CNS malformation. Here, we sought to further advance our system from binary classification to multi-classification, which is capable of detecting multiple types of CNS malformations. We also assessed the efficacy of this algorithm in improving clinical doctors' diagnostic performance. This is so far the first attempt to construct a deep learning AI system to aid both the experienced and unexperienced physicians in the prenatal ultrasound diagnosis on congenital anomalies.

Materials and Methods

Ultrasound images datasets

This research was a retrospective multicenter diagnostic study. For AI model development and testing, abnormal pregnancies of 12 types of common CNS malformations and normal pregnancies were retrospectively collected from The first Affiliated Hospital of Sun Yat-sen University (March 2010 to September 2018), Dongguan Maternal and Child Health Hospital (January 2016 to December 2018), and the Women and Children's Hospital affiliated with Xiamen University (January 2016 to December 2018). These 12 types of malformations included: agenesis of corpus callosum (ACC), absence of cavum septi pellucidi (ASP), holoprosencephaly (HPE), Dandy-Walker malformation and variant (DWNv), Megacisterna magna (MCM), Blake's pouch cyst, hydrocephaly, ventriculomegaly, arachnoid cyst, choroid plexus cyst (CPC), midline cyst and subependymal cyst. All the prenatal ultrasonic diagnoses were confirmed by prenatal or postnatal MRI, follow-up examination or autopsy. Ultrasound examinations of the abnormal pregnancies over a period of four weeks were included as part of this study. The mean gestational age was 21+5 weeks and 25+4 weeks for normal and abnormal cases, respectively. Ultrasound examinations were performed using various machines from six different manufacturers (GE Voluson 730 Expert/E6/E8/E10, Aloka SSD-a10, Siemens Acuson S2000, Toshiba XARIO 200 TUS-X200, Samsung UGEO WS80A, Philips IU22). This retrospective study was approved by Institutional Review Board of The First Affiliated Hospital of Sun Yat-sen University. Informed consent from patients was waived because of the retrospective nature of the study.

Two-dimensional neurosonographic grayscale images were employed to develop and testing the AI system. If the images were 3D volume data or were with split-view, we would export it or divide it into qualified single two-dimensional grayscale images before use according to the methods introduced in our previously published study[21]. All the two-dimensional grayscale images should meet the following criteria of inclusion: 1) neurosonographic images of the standard axial planes, namely the transventricular (TV) plane, transthalamic (TT) plane or transcerebellar (TC) plane, acquired according to the guidelines of the International Society of Ultrasound in Obstetrics & Gynecology (ISUOG) [22,23]; 2) images with an integrated skull, properly magnified without measurement caliper overlays and without the obvious acoustical shadow. Consequently, after excluding unqualified images and redundant normal images in the test dataset at Xiamen hospital, the overall dataset contained 20,689 normal images and 17,573 abnormal images. The pixel sizes of images were 1920 × 1080, 1408 × 712, 1400 × 700, 1300 × 870, 960 × 720, 800 × 600, 768 × 576, 720 × 576 and 640 × 480. The detailed constitutions of the ultrasound image datasets for the development and testing of the AI system are shown in Table 1, and the workflow diagram is shown in Figure 1.

Image labeling and pretraining process

All images were labeled by a team of seven doctors with 3 to 23 years of experience using LabelImg software (v. 2.0) following two steps. First, five doctors with 3–8 years of experience identified lesions in the images independently and labeled them with minimum bounding rectangles. In addition, six normal structures were labeled if visible, including cavity of septum pellucidum, thalamus, lateral ventricles, Sylvian fissures, cerebellar and cisterna magna. Next, two senior independent ultrasound specialists with over 20 years of experience verified the labels for each image. After labeling, images from The First Affiliated Hospital of Sun Yat-sen University and Dongguan M&C Health Hospital were randomly assigned for training and evaluation with a ratio of 8:2. The assignment was made on a case level rather than an image level, ensuring that the testing dataset did not contain any images originating from the training cases. Details are shown in Figure 1. To make the algorithms robust, training datasets were augmented before training by randomly rotating images from 0° to 60°, and flipping the images horizontally and vertically to simulate various fetal positions. Additionally, the images were zoomed up and down across the whole image and were pseudo-color processed. After augmentation, all images were resized to 1600 × 900.

AI model development

Our AI model was developed based on the algorithm of YOLO (you only look once, V3) , a unified, real-time, efficient object detection algorithm, which was recently proposed in deep learning computer vision field[24-26]. Object detection algorithms were designed not only to recognize what objects are present but also to localize where they are, no matter how many objects are there. Thus, object detection is more complex and challenging compared with classification algorithms. It was initially used in face recognition in security field and self-driving. In the ultrasound imaging field, there might be unknown number of structures and lesions within one image that need to be recognized and precisely located. Also, we chose YOLO for its efficiency considering dynamic data analysis may be needed. We added a logic output network to YOLO in our current AI model, which would eliminate redundant labels on the same structure by comparing label scores. For example, for the same image, normal and abnormal labels could not simultaneously exit on the same side of the lateral ventricles. As a result, the model had only one input and two outputs. The input of the model was the ultrasound image of fetal brain. The first output was a bounding box with labels and scores (numbers range from 0 to 1). The second output was the final result which consisted of remaining bounding boxes with labels after label elimination in the logic output, as shown in Figure 2 and Figure 3. Note that, due to the logic output network, lesions detected by AI were not made only based on label scores which were continuous number from 0 to 1 but also on the higher score. Therefore, when we drew ROC, the data were treated as binary data (yes/no) like human making diagnosis, rather than continuous variable data.

AI tests and comparison with human doctors

An external test set of 812 images from 449 patients was used to evaluate the performance of AI networks. The diagnostic accuracy, specificity, and sensitivity of AI in identifying CNS malformations were calculated, and the ROC curves were generated to evaluate the performance of the established AI algorithm. The performance of AI was then compared with that of doctors, who reviewed the same images in a separate testing. In this testing, images were shown one by one on the personal computer screen in a random order, and each image was along with 13 diagnosis choices (12 types of CNS abnormalities and normal). Ultrasonic doctors from different hospitals with varying degrees of expertise, who had experience >10 years (expert), 5-10 years (competent), and 1 year (trainee), reviewed one image with an optimal diagnosis and turned to the next image without returning to the previous one. The processing time for reading each image was recorded. All the doctors were blind to the diagnoses of images.

AI assistance strategy

Two months after the first reading, the doctors read the 812 ultrasonic images again (second reading) with a concurrent reading mode. This meant that, for each image, there would be two images (image without and with an AI diagnosis) shown on the screen side-by-side, and the doctors would read these two images and make a diagnosis. The diagnostic performance and time of the first and second readings were compared to evaluate the capability of AI in assisting diagnosis.

Statistical Analysis

The diagnostic performance of AI model and human doctors was assessed by multiple metrics, including accuracy, sensitivity, specificity and AUC. These parameters were defined as following:

Accuracy= the number of correctly labeled images divided by the total number of test images;

Type-specific sensitivity = the number of images correctly labeled with one type of abnormality divided by total number of images with that type of abnormalities;

Overall sensitivity=total number of images correctly labeled with each type of abnormality divided by total number of images with any type of abnormalities;

Type-specific specificity = the number of images correctly labeled without one type of abnormality divided by total number of images without that type of abnormalities;

Overall specificity=total number of images correctly labeled without corresponding type of abnormalities divided by total number of images without any types of abnormalities.

The mean accuracy, sensitivity, specificity, and AUC with 95% confidence intervals (CIs) were calculated. ROC curves were plotted by the sensitivity (true positive rate) versus the 1- specificity (false positive rate). The ROC curve shows the performance of a classification model at all classification thresholds. One sample t-tests were applied to compare the overall performance of AI to that of 13 doctors, as well as to that of doctors of three degrees respectively (AI vs. doctors, and AI vs. expert, competent or trainee). Paired t-tests were applied to comparing the performance of doctors without and with AI assistance. Analysis of variance was applied to compare the average improvement in performance level of doctor of three degrees and Bonferroni correction was applied for all multiple comparisons. All analyses were performed using statistical software (Stata, version 15.0; StataCorp LLC., College Station, TX), and a P value of less than 0.05 was considered significant for all analyses.

Results

AI performance

The AI system achieved an overall accuracy of 79.8% (95% CI 77.0-82.6%) in correctly identifying each type of CNS malformation, with a sensitivity of 78.4% (75.3-81.3%), specificity of 94.4% (86.2-98.4%) and an AUC of 0.864 (0.833-0.895). The performance of CPC identification was the best among all types of malformations detection, with a sensitivity of 92.0% (74.0-99.0%), specificity of 99.9% (99.3-100%)and AUC 0.959(0.905-1.000). Whereas, the performance of Blake's pouch cyst diagnosis was the lowest in terms of sensitivity of 42.9% (21.8- 66.0%), specificity of 99.6% (98.9-99.9%), and AUC 0.712 (0.604- 0.821). The diagnostic efficacy for the total and specific types of anomalies identification were shown in Table 2.

Comparison of performance between AI network and doctors

The AI outperformed the average efficacy of 13 doctors with respect to the overall types of malformations detection as shown in Table 3 and Figure 4a, the doctors' diagnostic accuracy [65.4% (95% CI 57.3-73.7%), p = 0.002], sensitivity [88.2% (82.3%-94.1%), p = 0.003], specificity 63.3% [(54.6-72.0%), p = 0.041] and AUC[0.758 (0.694, 0.821), p = 0.004] were all lower to that of AI system.

When compared AI performance with that of three groups of doctors respectively, we found the performance of AI model was similar to that of the expert doctors in terms of accuracy [78.9% (95%CI 75.2-82.5%), p = 0.528], sensitivity [77.5% (95%CI 73.7-81.4%), p = 0.521], and AUC [0.853 (95% CI 0.800-0.905), p = 0.681], while the performance of AI was higher than that of the competent {[accuracy: 69.6% (95% CI 75.2-85.2%), p = 0.016]; [sensitivity: 67.5% (95% CI 59.7-75.3%), p = 0.021]; [AUC: 0.793 (95% CI 0.777-0.809), p = 0.001]} and that of the trainees as well{[accuracy: 51.5%, 95% CI (39.4-63.6%), p = 0.001]; [sensitivity: 48.6% (

95% CI 36.0-61.2%), p = 0.003]; [AUC: 0.654(95% CI 0.538-0.770), p = 0.008)]}. However, specificity of AI did not differ to those of three categories of doctors. The comparison in performance between AI system and the various doctors is shown in Table 3 and Figure 4b.

The developed AI algorithm could analyze 7–8 images per second(s) and took only 113s to complete the diagnosis of 812 ultrasound image. The time consuming was significantly less than the average time of the 13 doctors (113s vs. 11571s, p = 0.001). When compared with the subgroups, the time of the diagnosis process were also shorter than three groups of doctors respectively [113s vs. 8864s (expert), p=0.02; 12801s (competent), p=0.003; 12663s (trainee). p = 0.001].

AI improved the doctors' performance on CNS malformations identification

When facilitated with the AI diagnosis, the overall diagnostic efficacy of three subgroups of doctors got significantly improved (Table 3, Figure 5a, b, c) in terms of accuracy, sensitivity, and AUC. For the experts, the accuracy, sensitivity and AUC were improved from 78.9% to 84.7% (p = 0.002), from 77.5% to 83.4% (p = 0.003), and from 0.853 to 0.910 (p = 0.019), respectively. For the competent doctors, the improvements for accuracy was from 69.6% to 85.1% (p = 0.005), sensitivity was from 67.5% to 84.0% (p = 0.006), and AUC from 0.793 to 0.905(p = 0.002). For trainee doctors, the progress was shown in accuracy (51.5% vs. 80.2%, p = 0.001), sensitivity (48.6% vs.78.7%, p = 0.001), and AUC (0.654 vs. 0.872, p = 0.006), respectively. Whereas, no significant difference was noted in specificity with and without AI assistance. Among the three groups of doctors, the trainee group received maximum improvement with AI assistance, whose diagnostic performance advanced to the level of expert group in terms of accuracy[(80.2% (95% CI 75.0-85.3%) vs. 78.9 %(95% CI 75.2-85.2%), P = 0.593] and AUC [0.872 (95% CI 0.861-0.882) vs. 0.853(95% CI 0.809-0.905), p = 0.238]. (Table 4).

The average time for diagnosis required by 13 doctors reduced significantly (7040s vs. 11571s, $p < 0.001$) with AI assistance, compared to that without AI assistance. Compared the time in subgroup, the time required by trainee doctors (7383s vs. 12663s, p = 0.008) and competent doctors (7729s vs. 12801s, p = 0.018) also decreased. However, for experts, no significant time-saving was observed (5923s vs. 8864s, p = 0.114).

Discussion:

Main findings

We developed an AI model to detect 12 types of CNS malformations in fetal ultrasound images by training on 37450 images from 15264 fetuses and testing on 812 images from 449 fetuses, our AI system achieves performance on par with expert doctors demonstrating an artificial intelligence capable of detecting congenital malformations with a level of competence comparable experience doctors. Furthermore, with AI system assistance, the performance of all the groups doctors get improved, especially for the trainee doctors.

Strengths and limitations

There are some limitations to our study. First, although the brain is traditionally examined in the axial plane and the evaluation of this plane is widely used as a screening tool, to make a more comprehensive anatomy examining, coronal and sagittal planes are also required[22]. Our AI system was established only based on image of axial view and it was unable to provide a fully assessment of lesions, we will continue to train the current AI model with images of other planes to optimize its performance. Second, although transfer learning allows the development of an accurate model with a relatively small training dataset, our sample size might be relatively small considering for multiple kinds of anomalies identification, we will continue to optimize our system with larger amount of data[13]. Finally, our AI was trained and validated using datasets from southern China, and its efficacy for other populations is yet to be investigated.

The strength of our study is the multicenter design, AI system was training on data from two different hospitals and the high performance of the AI system was validated by the data from the third external hospital, the doctors took part in the test also came from different hospital all over the country, which contribute to the generalizability of AI system and ensure the objective assessment of AI performance.

Interpretation

To the best of our knowledge, this is the first attempt to develop AI system to detect specific CNS malformations. Previous studies showed that images of normal transventricular (TV) and transcerebellar (TC) planes could be recognized and biometric measured by CNN-based deep learning algorithms [14, 18]. For example, the AI system established by Yaqub et al. [14]can identify normal TV planes by detecting the fetal head and the visibility of the cavi septi pellucidi. Baumgartner et al.[18] reported a method for real-time detection and localization of 13 fetal standard planes, including the TV and TC planes. Nevertheless, rare studies involved cases with congenital malformations, and training to classify images as normal or abnormal, let alone to make a diagnosis for specific structural anomalies. Our previous study[21] used 15372 normal and 14047 abnormal fetal CNS ultrasound images to establish binary classification of an AI system, and the results showed that that AI system had a sensitivity of 96.9%, specificity of 95.9%, and AUC 0.989 (95% CI: 0.986–0.991) when identifying images as normal or abnormal. Thus, we verified the feasibility of CNN-based deep learning algorithms for binary classification. On the basis of that work, we established this multi-classification model to perform specific malformations diagnosis. This new AI system achieved a 0.798 (95% CI 0.770, 0.826) accuracy and an AUC of 0.86 (0.83–0.89) in identifying 12 types of CNS based on ultrasound images. The results demonstrated an artificial intelligence is capable of detecting specific congenital malformations.

In the clinical testing, our AI system assisted doctors of all expertise levels in improving their detection performance of fetal CNS malformations. This was especially prominent for the trainee doctors, whose performance was improved to a level comparable with that of expert doctors after AI assistance. This might be attributed to the lesion localization function of the AI model, which can help doctors to recognized the lesions then to make diagnosis. This advantage would be especially useful in clinical practice. As we know, the prenatal diagnosis for CNS anomalies is one of the most difficult and challenging task and needs a special technique, namely neurosonography, a targeted ultrasound examination of the fetal brain performed by an expert [27]. However, such expertise requires years of experience and cannot be equivalent in all centers, especially in undeveloped countries and remote areas[28]. Hence, with our AI assistance, the detection rate of fetal CNS anomalies is expected to be improved even in clinical unit lacking of expert. Additionally, the ultrasound images used for training and validation in current AI system were collected by a variety of ultrasound equipments from different companies, which will indicate it can be used universally.

Conclusions

In a short summary, we developed an AI system to help diagnose congenital CNS malformations based on ultrasound images of fetal craniocerebral standard transverse planes. Our AI model achieved a high diagnostic performance compared with that of experienced doctors and can support unexperienced doctors by improving their diagnostic accuracy to an expert-level.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (grant numbers 81571687 and 61771007), the Science and Technology Development Plan of Guangdong Province (grant numbers 2017A020214013 and 2017B020226004) and the Health & Medical Collaborative Innovation Project of Guangzhou City (grant number 201803010021)

Author contributions

Meifang Lin, Nan Wang and Hongning Xie designed the research; Xiaoqin He, Miao He, Lihe Zhang, Jv Zheng, Jieling Feng, Yongzhong Yang, Hongmin Cai, Jianbo Xian, Hongmei Guo, and Qiuhong Xu acquired data and/or executed the research; Meifang Lin and Chun Hao analysed and/or interpreted the data; Meifang Lin and Nan Wang prepared the manuscript.

Competing interests

No conflicts of interest to declare

Details of ethics approval

Research ethics committee approval (2019421) was obtained from Institutional Review Board of The First Affiliated Hospital of Sun Yat-sen University on 25/10/2019

References

1. WHO. Section on congenital anomalies. Available from:*http://wwwwhoint/mediacentre/factsheets/fs370/en/* 2012.

2. Darmstadt GL, Howson CP, Walraven G, Armstrong RW, Blencowe HK, Christianson AL, et al. Prevention of Congenital Disorders and Care of Affected Children: A Consensus Statement. JAMA pediatrics. 2016; 170:790-3.

3. Sitkin NA, Ozgediz D, Donkor P, Farmer DL. Congenital anomalies in low- and middle-income countries: the unborn child of global surgery. World journal of surgery. 2015; 39:36-40.

4. Hospital Costs for Birth Defects Reach Tens of Billions. Jama. 2017; 317:799.

5. Morris JK, Wellesley DG, Barisic I, Addor MC, Bergman JEH, Braz P, et al. Epidemiology of congenital cerebral anomalies in Europe: a multicentre, population-based EUROCAT study. Archives of disease in childhood. 2019; 104:1181-7.

6. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017; 542:115-8.

7. Becker AS, Mueller M, Stoffel E, Marcon M, Ghafoor S, Boss A. Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. The British journal of radiology. 2018; 91:20170576.

8. Chi J WE, Babyn P, Wang J, Groot G, Eramian M. Thyroid Nodule Classification in Ultrasound Images by Fine-Tuning Deep Convolutional Neural Network. J Digit Imaging. 2017; 30:477-86.

9. Yap MH PG, Marti J, Ganau S, Sentis M, Zwiggelaar R, Davison AK, Marti R. Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks. IEEE J Biomed Health Inform. 2018; 22:1218-26.

10. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. Jama. 2017; 318:2199-210.

11. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. Jama. 2016; 316:2402-10.

12. Luo H, Xu G, Li C, He L, Luo L, Wang Z, et al. Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study. The Lancet Oncology. 2019; 20:1645-54.

13. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell. 2018; 172:1122-31 e9.

14. Yaqub M, Kelly B, Papageorghiou AT, Noble JA. A Deep Learning Solution for Automatic Fetal Neurosonographic Diagnostic Plane Verification Using Clinical Standard Constraints. Ultrasound in medicine & biology. 2017; 43:2925-33.

15. Yu Z TE, Ni D, Qin J, Chen S, Li S, Lei B, Wang T. A Deep Convolutional Neural Network-Based Framework for Automatic Fetal Facial Standard Plane Recognition. IEEE J Biomed Health Inform. 2018;

22:874-5.

16. L. W, JZ. C, S. L, B. L, T. W, FUIQA ND. Fetal Ultrasound Image Quality Assessment With Deep Convolutional Networks. IEEE Trans Cybern. 2017; 47:1336-49.

17. Chen H WL, Dou Q, Qin J, Li S, Cheng JZ, Ni D, Heng PA. Ultrasound Standard Plane Detection Using a Composite Neural Network Framework. IEEE Trans Cybern. 2017; 47:1576-86.

18. Baumgartner CF, Kamnitsas K, Matthew J, Fletcher TP, Smith S, Koch LM, et al. SonoNet: Real-Time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound. IEEE transactions on medical imaging. 2017; 36:2204-15.

19. Lin H, Li R, Liu Z, Chen J, Yang Y, Chen H, et al. Diagnostic Efficacy and Therapeutic Decision-making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial. EClinicalMedicine. 2019; 9:52-9.

20. Azad N, Amos S, Milne K, Power B. Telemedicine in a rural memory disorder clinic-remote management of patients with dementia. Canadian geriatrics journal : CGJ. 2012; 15:96-100.

21. Xie H, Wang N, He M, Zhang L, Cai H, Xian J, et al. Using deep learning algorithms to classify fetal brain ultrasound images as normal or abnormal. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2020;7.

22. International Society of Ultrasound in O, Gynecology Education C. Sonographic examination of the fetal central nervous system: guidelines for performing the 'basic examination' and the 'fetal neurosonogram'. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2007; 29:109-16.

23. Salomon LJ, Alfirevic Z, Berghella V, Bilardo C, Hernandez-Andrade E, Johnsen SL, et al. Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2011;37:116-26.

24. Redmon J DS, Girshick R , et al. You Only Look Once: Unified, Real-Time Object Detection. arXiv. 2015.

25. Redmon J FA. YOLO9000: Better, Faster, Stronger. arXiv. 2016.

26. Redmon J FA. YOLOv3: An Incremental Improvement. arXiv. 2018.

27. Paladini D, Quarantelli M, Sglavo G, Pastore G, Cavallaro A, D'Armiento MR, et al. Accuracy of neurosonography and MRI in clinical management of fetuses referred with central nervous system abnormalities. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2014; 44:188-96.

28. Edwards L, Hui L. First and second trimester screening for fetal structural anomalies. Seminars in fetal & neonatal medicine. 2018; 23:102-11.

Table 1 Details of ultrasound images datasets (images/pregnancies) included in development and test of AI system

	1^{st} Hospital of SYSU	Dongguan M&C Health Hospital	W&C Hospital of Xiamen University
Normal	17610/11370	3008/1904	71/42
ACC	3610/222	139/14	135/54
ASP	1076/64	22/4	50/25
DWMv	1063/115	54/13	30/18
HPE	762/103	228/36	95/53
MCM	922/331	377/75	88/64

Hydrocephaly	2793/184	655/83	139/70
Ventriculomegaly	972/153	362/67	57/45
Blake's pouch Cyst	798/55	14/3	21/10
Arachnoid Cyst	1363/83	125/18	31/13
CPC	614/154	39/10	25/14
Midline Cyst	361/88	52/15	29/23
Subependymal Cyst	375/91	56/9	40/18
Total	32319/13013	5131/2251	812/449

SYSU, Sun Yat-sen University; M&C, Maternal and Child; W&C, Women and Children's; ACC, Absence of corpus callosum; ASP, absence of cavum septi pellucidi; DWNv, Dandy-Walker malformation or variant; HPE, holoprosencephaly; MCM, Megacisterna magna; CPC, choroid plexus cyst.

Table 2 The performance of AI of overall and each type of anomalies identification

	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
ACC	93.6(91.9 - 95.3)	64.4(55.8 - 72.5)	99.4(98.5 - 99.8)	0.819 (0.779 - 0.860)
ASP	98.7(97.8 - 99.4)	78.0(64.0 - 88.5)	100(99.5 - 100)	0.890 (0.832 - 0.948)
DWMv	98.0(97.1 - 99.0)	86.7(69.3 - 96.2)	98.5(97.3 - 99.2)	0.926 (0.864 - 0.988)
HPE	97.5(96.5 - 98.6)	91.6(84.1 - 96.3)	98.3(97.1 - 99.1)	0.950 (0.921 - 0.981)
MCM	98.0(97.1 - 99.0)	84.1(74.8 - 91.0)	99.7(99.0 - 100)	0.919 (0.881 - 0.958)
Hydrocephaly	95.4(94.0 - 96.9)	87.1(80.3 - 92.1)	97.2(95.6 - 98.3)	0.921 (0.893 - 0.950)
Ventriculomegaly	95.0(93.4 - 96.5)	87.7(76.3 - 94.9)	95.5(93.8 - 96.9)	0.917 (0.873 - 0.960)
Blake's pouch Cyst	98.2(97.2 - 99.1)	42.9(21.8 - 66.0)	99.6(98.9 - 99.9)	0.712 (0.604 - 0.821)
Arachnoid Cyst	97.4(96.3 - 98.5)	51.6(33.1 - 69.8)	99.2(99.2 - 99.3)	0.754 (0.665 - 0.844)
CPC	99.6(99.2 - 100)	92.0(74.0 - 99.0)	99.9(99.3 - 100)	0.959 (0.905 - 1.000)
Midline Cyst	97.5(96.5 - 98.6)	56.7(37.4 - 74.5)	99.1(98.2 - 99.6)	0.779 (0.689 - 0.869)
Subependymal Cyst	98.9(98.2 - 99.6)	80.0(64.4 - 90.9)	99.8(99.3 - 100)	0.899 (0.837 - 0.962)
Overall	79.8(77.0 - 82.6)	78.4(75.3 - 81.3)	94.4(86.2 - 98.4)	0.864 (0.833 - 0.895)

ACC, Absence of corpus callosum; ASP, absence of cavum septi pellucidi; DWNv, Dandy-Walker malformation or variant; HPE, holoprosencephaly; MCM, Megacisterna magna; CPC, choroid plexus cyst.

Table 3 Detailed performance comparisons between AI and ultrasonic doctors alone, doctors with and without AI assistance.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC

AI model		79.8 (77.0 - 82.6)	78.4 (75.3 - 81.3)	94.4 (86.2 - 98.4)	0.860 (0.830 - 0.890)
Doctors Only	Expert	78.9 (75.2 - 85.2)	77.5 (73.7 - 81.4)	93.0 (84.1 - 100.0)	0.853 (0.800 - 0.905)
	Competent	69.6* (75.2 - 85.2)	67.5* (59.7 - 75.3)	91.2 (84.7 - 97.7)	0.793* (0.777 - 0.809)
	Training	51.5* (39.4 - 63.6)	48.6* (36.0 - 61.2)	82.0 (65.8 - 98.2)	0.654* (0.538 - 0.770)
Doctors with AI Assistants	Expert	84.7# (82.4 - 86.9)	83.4# (80.8 - 85.9)	98.3 (96.1 - 100)	0.910# (0.897 - 0.923)
	Competent	85.1# (82.9 - 87.4)	84.0# (81.6 - 86.4)	96.9 (92.6 - 100)	0.905# (0.884 - 0.925)
	Training	80.2# (75.0 - 85.3)	78.7# (72.6 - 84.8)	95.8 (91.7 - 99.9)	0.872# (0.861 - 0.882)

* a statistically significant difference between AI and doctors alone. #: a statistically significant difference between doctors with and without AI assistance.

Table 4 The improvement of diagnostic performance with AI assistance

	Accuracy difference (%)	Sensitivity difference (%)	Specificity difference (%)	AUC difference
Trainee	28.7(19.5 - 37.8) [a,b]	30.1(21.2 - 39.1) [a,b]	13.8(-1.4 - 29.0)	0.218(0.011 - 0.330) [b]
Competent	15.6(8.9 - 22.3) [a]	16.6(9.2 - 24.0) [a]	5.7(-0.1 - 11.5)	0.113(0.077 - 0.148)
Expert	5.8(3.9 - 7.7) [b]	5.9(3.7 - 8.0) [b]	5.3(-2.1 - 12.7)	0.058(0.018 - 0.097) [b]
	p<0.001	p<0.001	p=0.270	p=0.007

[abc]: represent the results of bonferroni comparison,[a] significant difference between trainee and competent, p<0.05;[b] significant difference between trainee and expert, p<0.05

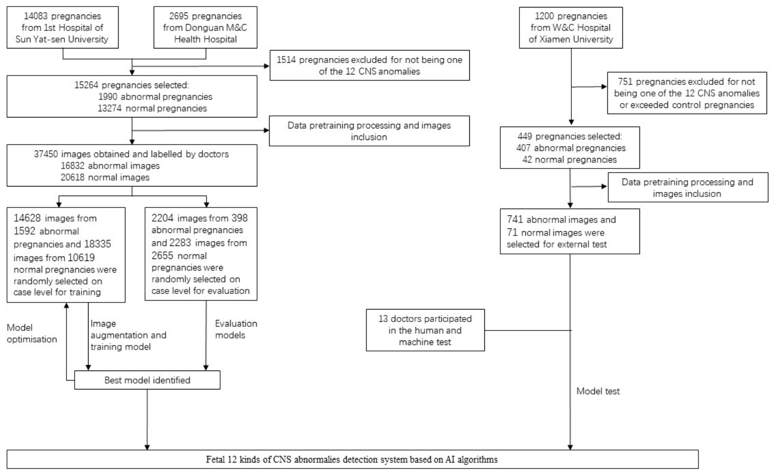
Figure 1 Flowchart for the development and test of the algorithms. M&C, Maternal and Child; W&C, Women and Children's; CNS, central nervous system; AI, artificial intelligence.

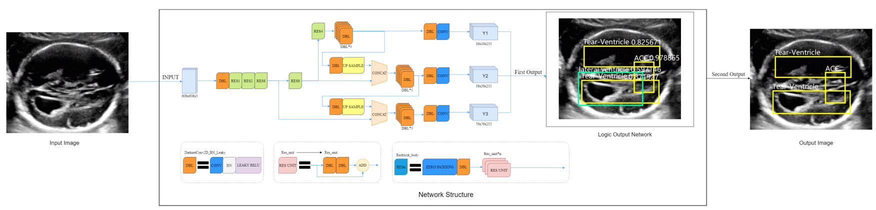
Figure 2 Flow chart illustrating the entire process of the network. As shown in the figure, our process contains one input and two outputs. In the first output, two labels were detected on the same side of ventricle by the model, which were lateral ventricle (green box, the label score was 0.597146) and tear-ventricle (lower yellow box, the label score was 0.871927). After label elimination in the logic output network according to the scores, only one label with the higher score remained in output image (tear-ventricle, lower yellow box).

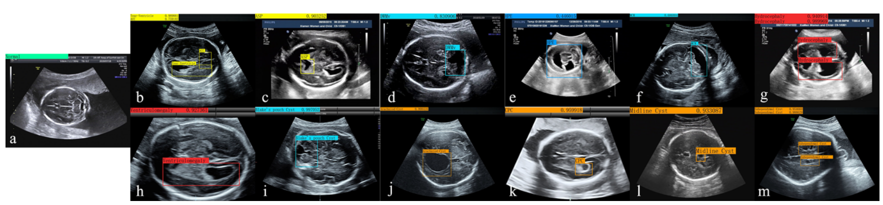
Figure 3 The composite image shows the AI output correctly labeled with corresponding type of specific malformations in each image, as well as normal image. ACC, Absence of corpus callosum; ASP, absence of cavum septi pellucidi; DWNv, Dandy-Walker malformation or variant; HPE, holoprosencephaly; MCM, Megacisterna magna; CPC, choroid plexus cyst.

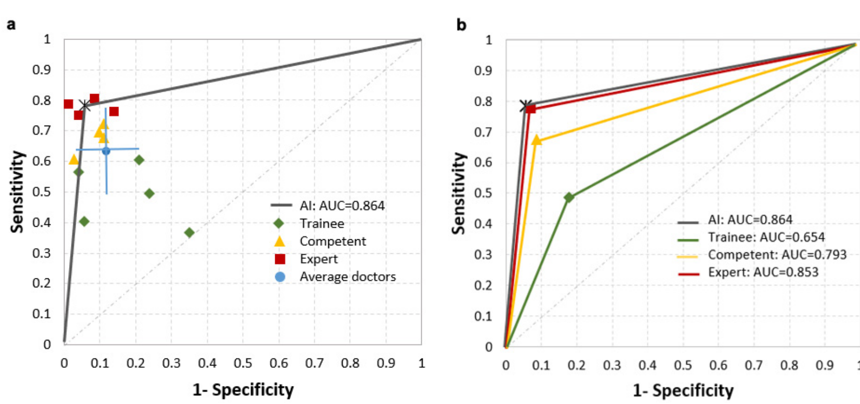
Figure 4 The performance of the AI system and Ultrasonic doctors in CNS malformations identification a. AI system outperforms the average of the ultrasonic doctors at CNS malformations identification. Each point represented the sensitivity and specificity of a single ultrasonic doctors, the blue points are the average of the doctors, with error bars denoting one standard deviation. The AI system achieves superior performance to a doctor if the sensitivity–specificity point of the lies below the blue curve, which most do. b, The performance of AI model versus that of experts, competent and trainee doctors.

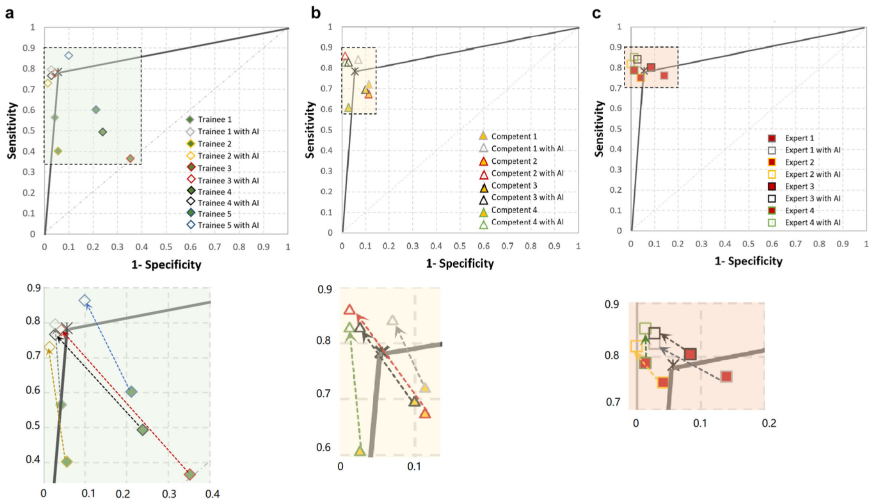
Figure 5 The improvement of overall performance of three degrees of doctors in CNS malformations identification with AI assistance (a. trainee, b. competent, c. expert).










a
b
c
Sensitivity
1- Specificity
Trainee 1
Trainee 1 with AI
Trainee 2
Trainee 2 with AI
Trainee 3
Trainee 3 with AI
Trainee 4
Trainee 4 with AI
Trainee 5
Trainee 5 with AI
Competent 1
Competent 1 with AI
Competent 2
Competent 2 with AI
Competent 3
Competent 3 with AI
Competent 4
Competent 4 with AI
Expert 1
Expert 1 with AI
Expert 2
Expert 2 with AI
Expert 3
Expert 3 with AI
Expert 4
Expert 4 with AI