How much can AI see in early pregnancy: A multi-center study of fetus head characterization in week 10-14 in ultrasound using deep learning

Yitao Jiang (joetao097@gmail.com)
Illuminate, LLC; Microport Prophecy

Qi Li
Shenzhen People's Hospital (The Second Clinical Medical College，Jinan University，The First Affiliated Hospital, Southern University of Science and Technolog

Yuli Zhou
Shenzhen People's Hospital (The Second Clinical Medical College，Jinan University，The First Affiliated Hospital, Southern University of Science and Technolog

Yujuan Zhang
Shenzhen People's Hospital (The Second Clinical Medical College，Jinan University，The First Affiliated Hospital, Southern University of Science and Technolog

Siyuan Shi
Illuminate, LLC; Microport Prophecy

Shaoli Yin
Shenzhen People's Hospital (The Second Clinical Medical College，Jinan University，The First Affiliated Hospital, Southern University of Science and Technolog

Xuye Liu
Shenzhen People's Hospital (The Second Clinical Medical College，Jinan University，The First Affiliated Hospital, Southern University of Science and Technolog

Qihui Peng
Shenzhen People's Hospital (The Second Clinical Medical College，Jinan University，The First Affiliated Hospital, Southern University of Science and Technolog

Shaoting Huang
pengqh16@126.com

Chen Cui
Illuminate, LLC; Microport Prophecy

Ruilian Zhe
Shenzhen People's Hospital (The Second Clinical Medical College，Jinan University，The First Affiliated Hospital, Southern University of Science and Technolog

Jinfeng Xu

henzhen People's Hospital, The Second Clinical Medical College Jinan University,The First Affiliated Hospital, Southern University of Science and Technology

Fajin Dong

Shenzhen People's Hospital (The Second Clinical Medical College Jinan University The First Affiliated Hospital, Southern University of Science and Technolog https://orcid.org/0000-0002-4558-4885

Article

Keywords:

Posted Date: February 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-754075/v1

Abstract

PURPOSE To investigate if artificial intelligence can identify fetus intracranial structures in pregnancy week 11-14; to provide an automated method of standard and non-standard sagittal view classification in obstetric ultrasound examination METHOD AND MATERIALS A data set of 1842 2D sagittal-view ultrasound images from 1842 females were collected to train and test a newly design scheme based on deep learning (DL) – Fetus Framework to identify nine fetus intracranial structures: thalami, midbrain, palate, 4th ventricle, cisterna magna, nuchal translucency (NT), nasal tip, nasal skin, and nasal bone. Results from Fetus Framework were further used for standard/non-standard (S-NS) plane classification, a key step for NT measurement and Down Syndrome assessment. S-NS classification were also tested with 156 images from a second medical center. Sensitivity, specificity and area under curve (AUC) were evaluated for comparison among Fetus Framework, three classic DL models and human experts with 1-, 3- and 5-year ultrasound training. Furthermore, a dataset of 316 standard images confirmed by the Fetus framework and another dataset of 316 standard images selected by physicians were utilized individually to train a random forest and perform the Fetal malformation classification task. Based on the hypothesis that random forest performs better on more standard dataset, mean AUC of 5-fold cross validation are compared. RESULTS Nine intracranial structures identified by Fetus Framework in validation are all consistent with that of senior radiologists. For S-NS sagittal view identification, Fetus Framework achieved AUC of 0.996 (95%CI: 0.987, 1.000) in internal test, at par with classic DL models. In external test, FF reaches an AUC of 0.974 (95%CI: 0.952, 0.995), while ResNet-50 arrives at AUC~0.883, 95% CI 0.828–0.939, Xception AUC~0.890, 95% CI 0.834–0.946, and DenseNet-121 AUC~0.894, 95% CI 0.839–0.949. For the internal test set, the sensitivity and specificity of the proposed framework is (0.905, 1), while the first-, third-, and fifth-year clinicians are (0.798, 0.986), (0.690, 0.958), and (0.619, 0.986), respectively. For the external test set, the sensitivity and specificity of FF is (0.989, 0.797) and first-, third-, and fifth-year clinicians are (0.663, 0.781), (0.609, 0.844), and (0.533, 0.875), respectively. In further validation of fetal malformation classification task, mean AUC of random forest in physician dataset is 0.768 (0.724 – 0.812) and in Fetus dataset is 0.806 (0.741 – 0.871), suggesting that Fetus framework identify standard images more accurately. CONCLUSION We proposed a new deep learning-based Fetus Framework for identifying key fetus intracranial structures. The framework was tested in data from two different medical centers. The results show consistency and improvement from classic models and human experts in standard and non-standard sagittal view classification during pregnancy week 11-13+6. CLINICAL RELEVANCE/APPLICATION With further refinement in larger population, the proposed model can improve the efficiency and accuracy in early pregnancy test using ultrasound examination.

Introduction

Obstetric examinations are critical during pregnancy. Non-radiative, economical and accessible, ultrasound is the most widely used tool to evaluate fetal development and anatomy throughout pregnancy for over half a century[1]. In many countries, it is clinically recommended to perform ultrasound-based screening during the first trimester (week 11 to 13+6), as it enables fetus viability check, accurate

gestation dating, ectopic pregnancy diagnosis and many other benefits[2]. In particular, the measurement of nuchal translucency (NT), a sonolucent area in the posterior fetal neck can be used to evaluate the risk for chromosome defects such as Down syndrome[3–6], affecting ~1 per 650-1000 newborns worldwide[7,8]. Moreover, an increased thickness of NT under ultrasonography in late first trimester is an established indicator of various structural abnormalities, congenital disease like cardiac defects and a number of other genetic syndromes[9–13]. The loss of nasal bone in fetus head scan is also associated with a few chromosome diseases including Down Syndrome[14–16]. Early detection of these preconditions provides an opportunity of intervention with maternal-fetal medicine at an early stage of pregnancy, and substantially reduces the level of emotional and physical burden if the choice of early termination is executed.

In spite of merits, challenges remain in the effective and accurate use of fetus head scan in the first trimester. As defined in the guidelines from International Society of Ultrasound in Obstetrics and Gynecology (ISUOG)[17], localization of mid-sagittal view of fetus head with high precision (aka standard plane or standard section) is required to secure a proper NT measurement. Education programs and certifications are routinely provided in a number of countries to help sonographers maintain proficiency[18–23]. However, quantification of NT thickness is still affected by inter-observer disagreement, resulting in variations of specificity for disease prediction[24–27]. In addition, the tall requirement for operators' skills also limits the impact of the technique from covering a wider population. Therefore, it is imperative for a new pathway to standardize the process of NT measurement acquisition and accurate fetus head visualization overall.

Meanwhile, deep learning is rapidly making inroads into medical imaging including ultrasound. Great interest is shown in using deep learning methods in obstetric ultrasound for precondition detection and exam optimization[28–30]. Previous efforts have focused on designing end-to-end models for parameter quantification and standard view detection during the second or third trimester with distinct imagery characteristics[28,30], or for cardiac function classification with time-serial video data[29,31]. End-to-end architecture is proven to deliver object recognition result with high accuracy, but potentially falls short for providing more nuanced information for users and hence might lack of generalizability to other similar problem setting. On the other hand, early pregnancy screening is still calling for more sophisticated deep learning-based methods to address its challenges such as less distinct features and more exam variations.

To address the above-mentioned challenges of high proficiency requirement for sonographers and inter-observer variability, we propose a new deep learning-based scheme: Fetus Framework (FF). By following a divide-and-conquer scheme with deep learning architecture, the framework identifies the existence of nine key structures in a fetus head scan: thalami, midbrain, palate, 4th ventricle, cisterna magna, nuchal translucency (NT), nasal tip, nasal skin, and nasal bone. Then a machine learning model is employed to perform the standard – non-standard (S-NS) section classification by incorporating the characterization results from previous step.

Methods And Materials

Deep learning-based framework for fetus head characterization and S-NS classification

In essence, a ‘divide-and-conquer’ principle is proposed in FF to detect nine key structures of fetus head. As in Figure 1, a CNN-based detector is deployed in the first step to detect midbrain, palate and thalami. Outputs are three boxed regions of interest (ROI) in the input image that most likely contains the three key structures, and their corresponding probabilities. Second, we designed a ROI proposal module (RPM) (Figure. 2) for nasal region, “IT&CM” region, and “NT” region based on the detection results from step one. Finally, three separate CNN-based detectors are deployed to find the boxed ROIs that best enclose respective structures, of which nasal bone, nasal skin, nasal tip by the first detector within the proposed nasal region, IT and CM by the second detector within “IT&CM” region, and NT within the by the third in “NT” region.

One primary feature of FF is that CNN-based detectors in all steps are non-specific, as the framework is designed be more generalizable in other hierarchical object detection applications. In this paper, a RetinaNet detector[32] with DenseNet-121[33] backbone is deployed to detect palate, midbrain, and thalami. The key advantage of RetinaNet is that its design of *focal loss* function. In general, detectors generate a number of candidates of boxed regions and chooses the one with highest probabilities as the final object location. The focal loss function in RetinaNet optimizes the search of boxed ROI region by prioritizing candidate ROIs with high probabilities in a large candidate pool, often >100K. DenseNet connects each layer to every other layer to maximize information flow in the network. Similarly, another three separate RetinaNet detectors with DenseNet-121 backbone are used to find the boxes that best enclose the structures.

Another novelty of FF is the ROI proposal module (RPM) for locating nasal, “IT&CM”, and “NT” regions by inheriting the relative location information of midbrain, palate and thalami from step one. The key notion of RPM is that it is generalizable of the location correlation for nasal, IT&CM, NT regions and midbrain, palate and thalami. In figure 2, RPM decodes the location correlation and generates three fine ROIs. These fine ROIs restricts the detector to focus on each region with specific relevant structures so nasal bone, nasal skin, nasal tip, IT, CM, and NT are further located respectively in next steps. Another major benefit is RPM is designed to be self-adaptive. For example, nasal structures are very finite (usually 2~3 pixels wide and tall), whereas NT boxed region is a narrow band with a long width. The proposed RPM locates ROIs and self-refines the size for best detection performance. Meanwhile, when the relative location is not found in step one, RPM would not find any fine ROI and the rest of FF is skipped. In this case, we assign 0 to the structure likelihood of nasal bone, nasal skin, nasal tip, IT, CM, and NT.

After all nine structures are detected, we trained a Logistic Regression (LR) classifier for standard and non-standard (S-NS) plane classification using the structure probabilities as input. The probability associated with the fine ROIs of each key structure indicates how likely the targeted structure is enclosed in the box. Intuitively, with a higher sum of weighted probabilities of all structures, a particular frame is

the standard plane with more certainty. Without loss of generosity, LR can also be replaced by many other machine learning models such as multi-layer perceptron (MLP), random forest and etc.

We used DenseNet-121, ResNet-50 and Xception trained end-to-end for S-NS classification as the control experiment.

Data curation and preparation

Images were acquired by multiple ultrasound specialists using GE Voluson E8 and Philips EPIQ 7 ultrasound machines. The internal dataset contains 2023 images from 2023 unique individuals who underwent NT measurement between 2017 and 2019. The external dataset contains 200 images from 200 unique individuals between 2018 and 2019. Each image corresponds to a unique individual and her single visit. All data is from single pregnancy of women at the age spectrum from 18-45 (Table 1). Average pregnancy age is 12 weeks and 5-6 days. 5 additional videos were also collected to record the process of NT exams during which the final standard plane is captured and saved manually by human expert. This retrospective study was performed in accordance with approved guidelines from participating hospitals. This study was approved by the Ethics Committee of the partnered medical center with consent forms approved by every patient.

323 images were excluded from the internal dataset and 8 images from the external dataset due to low quality or incorrect format. Recognizing the inherent variation in human assessment, two ultrasound experts, each of whom specializes in obstetric imaging with 10+ years of experiences, annotated the remaining images as SP or non-SP independently. Only images with same label as SP or non-SP from both experts are considered as the ground truth and we obtained 1133 SP and 447 non-SP images for the internal dataset, and 64 SP and 92 non-SP images for the external dataset. Bounding boxes for the nine structures in the SP images were labeled by one expert and reviewed by the other expert.

S-NS classification experiment design and clinical validation

The overall experiment consists of two major segments: 1) model training and validation and 2) internal and external model testing. In step 1), four RetinaNet-based detectors were firstly trained and validated for identifying nine structures with randomly chosen 812 standard plane images. Then Logistic Regression-based S-NS classifier was trained and validated with 560 images, including 249 standard planes 311 non-standard ones. The left 156 images in the internal dataset and 156 images from the second hospital were used for model testing in step 2). An automated preprocessing workflow was used to remove identifying information and eliminate unintended human labels. Each image was then padded to a square and resized to 320*320 pixels using Skimage in Python. The area under curve (AUC) is calculated by using the roc_curve function in the sklearn library. The confidence interval of AUC is evaluated using a fast implementation[33] of DeLong's algorithm[34]. Figure. 3 shows the data flow and major experiment steps.

Fetus Framework development and training

All four structure detectors were trained in Python using the Keras deep learning library. The Logistic Regression classifier was trained in Python using the scikit-learn machine learning library. We adopted the DenseNet-121 with a growth rate of 12 as the backbone network for the RetinaNet detector. DenseNet contains four dense blocks (DB) to extract feature maps at different resolutions. We constructed a multi-scale feature pyramid (P3, P4, P5) from the outputs of DB2, DB3, and DB4 to detect object at different scales. Given an input resolution of 3202, P3, P4, and P5 each has a resolution of 402, 202, and 102, respectively. The feature pyramid is further attached to a classification subnet for predicting the probability of object presence at candidate box for each object class, and a regression subnet for refining each candidate box. The first RetinaNet detector for palate, midbrain, and thalami was initialized with random weights, while the other three RetinaNet detectors were initialized with pretrained weights from the first RetinaNet detector. All four RetinaNet detectors were trained to minimize the focal loss ($\alpha=0.5,\gamma=2$). The models for the control experiments were initialized with ImageNet weights and were trained to minimize the binary cross entropy loss. All these models were trained using Adam optimizer with an initial learning rate of 0.0001, and batch size of 16 for at most 100 epochs. The learning rate was decayed by a factor of 0.1 if the training loss hasn't reduced for 5 epochs. The training process was forced to stop early if the validation loss hasn't reduced for 30 epochs, and we kept the weights with the lowest validation loss. Data augmentation techniques such as translation, scaling, and flipping on horizontal axis were applied during training.

Design of Region of interest Proposal Module (RPM)

As in Figure 2, we design a ROI proposal module (RPM) for locating nasal region, "IT&CM" region, and "NT" region by inheriting the relative location information of midbrain, palate and thalami from step one. We hypothesize that the relative locations of palate, midbrain, and thalami in a standard plane should follow a criterion (Figure. 4), where the center of midbrain is under thalami, and the center of palate is to the left or right of thalami based on fetus' orientation. We label a plane as potentially standard and move on to next step if it follows the criteria and label a plane as non-standard and stop testing if otherwise. Figure. 5 shows an example from the external test set that did not pass the criteria.

Once the plane passes the criteria, the RPM generates a focused ROIs for nasal, "IT&CM", and "NT" regions. We select the nasal region from above the palate and define its boundary according to the boxed region of palate and fetus' orientation. Similarly, the "IT&CM" region is chosen according to the boxed regions of midbrain and palate, the "NT" region according to the boxed region of midbrain. Figure. 6 is an example showing the focused ROIs after we applied the RPM to an input plane.

Evaluation of S-NS classification performance

To evaluate the robustness of the proposed framework, two sets of tests were performed in two independent datasets from the major medical center as internal test, and an affiliated medical center as external test. Both datasets are not previously seen by the trained framework. For each test, both comparison with state-of-the-art deep learning models and with human experts. Accuracy, sensitivity, specificity and area under curve (AUC) are reported for all experiments.

Additional test for fetal malformation classification task

In this test, a dataset of 316 standard images confirmed by the Fetus framework and another dataset of 316 standard images selected by physicians were utilized to prove the value of Fetus system in precise diagnosis. Both datasets were utilized individually to train a random forest and perform the Fetal malformation classification task. Crown rump length (CRL, continuous variable), thickness of the nuchal translucency (NT, continuous variable), visibility of the nasal bone (NB, categorical variables, invisible, visible on one or both sides) from the image, as well as the pregnant woman's self-reported age and abortion history, were used in random forest. The area under the curve (AUC) of 5-fold cross validation is used to evaluate the model's performance. We hypothesized that more standard dataset would achieve better model performance.

Results

Comparison with classic deep learning methods

We compared the performance of FF to that of several additional state-of-the-art deep learning architectures trained on the same datasets (Figure. 7a, 7b). For the internal test set, the performance of FF (AUC~0.996, 95% CI 0.987–1) is close to that of ResNet-50 (AUC~0.997, 95% CI 0.993–1), Xception (AUC~0.999, 95% CI 0.998–1), and DenseNet-121 (AUC~0.995, 95% CI 0.988–1). For the external test set, FF achieves AUC~0.974, 95% CI 0.952–0.995 while ResNet-50 arrives at AUC~0.883, 95% CI 0.828–0.939, Xception AUC~0.890, 95% CI 0.834–0.946, and DenseNet-121 AUC~0.894, 95% CI 0.839–0.949.

Comparison with human expert

In the benchmark experiment between FF and human experts (Figure. 7c), each image in the test set was independently classified by FF, a first-, a third-, and a fifth-year ultrasound clinicians from the major medical center. For the internal test set, the sensitivity and specificity of the proposed framework is (0.905, 1), while the first-, third-, and fifth-year clinicians are (0.798, 0.986), (0.690, 0.958), and (0.619, 0.986), respectively. For the external test set, the sensitivity and specificity of FF is (0.989, 0.797) and first-, third-, and fifth-year clinicians are (0.663, 0.781), (0.609, 0.844), and (0.533, 0.875), respectively.

Direct application in videos of clinical scan

FF is also evaluated in video data collected directly from NT clinical exams. Figure 8 demonstrates an independent video clip of an NT scan process that is not seen by model in training and evaluation from the major medical center. The video clip had 93 frames and every frame was evaluated by FF. We plotted the predicted probability of each frame containing a standard plane over frames in Fig. 8. Also shown are examples of predicted standard planes (frame 13 and 14) and non-standard planes (frame 12 and 15). The predicted result is in agreement with the ground truth. In clinical setting, the Fetus Framework will remind the examiner once a standard plane is detected and then recommend those frames with highest predicted probabilities when the scan is completed.

Statistics of Region of interest Proposal Module (RPM)

Of the internal test data, 72 out of 72 (100.0%) standard planes passed the criteria, and 77 out of 84 (91.7%) non-standard planes passed the criteria. Of the external test data, 64 out of 64 (100.0%) standard planes passed the criteria, and 70 out of 92 (76.1%) non-standard planes passed the criteria.

To visualize how the structures are enclosed inside the RPM-proposed regions, we plotted the bottom left vertex, center, and top right vertex of the boxed region of a structure inside the RPM-proposed ROI (Figure. 9). We fed the ground-truth boxed regions of thalami, midbrain, and palate into the RPM, and obtained the statistics of the location of nasal bone, nasal skin, nasal tip, IT, CM, and NT inside the RPM-proposed ROIs as shown in Figure. 10 and summarized in Table. 2. The result indicated that the RPM-proposed nasal region can properly enclose nasal bone, nasal tip, and nasal skin, as well as the RPM-proposed “ITCM” region for IT and CM, and the RPM-proposed “NT” region for NT.

In addition, we fed the predicted boxed regions of thalami, midbrain, and palate into the RPM, and obtained the statistics of the location of nasal bone, nasal skin, nasal tip, IT, CM, and NT inside the RPM-proposed ROIs as shown in Figure. 11 and summarized in Table 3. The distribution of structures inside the RPM-proposed ROIs was similar to that of structures inside the ground-truth-based ROIs. Therefore, our model-based RPM was a reliable tool for locating the nasal, “IT&CM”, and “NT” regions.

Performances of random forest in Fetus and physician dataset

Performances of random forest in Fetus selected dataset and physician chosen dataset are compared by using AUC from 5-fold cross validation. (Figure. 12) Mean AUC of random forest in physician dataset is 0.768 (0.724 – 0.812) and in Fetus dataset is 0.806 (0.741 – 0.871). Based on the hypothesis that model achieves better performance in more standard dataset, Fetus framework identify standard images more accurately.

Discussion

FF is a deep learning framework with a set of cascading computational modules that achieves classifies standard or non-standard sagittal planes with high accuracy in obstetrical US examination between week 10 and week 13+6. It uses expert-labeled nine structures to train multiple CNN architecture at the same time in a novel divide-and-conquer framework for hierarchical object detection. The probability of each structure obtained from the trained object detectors is then deployed to train a machine learning-based classifier to generate the final classification outcome. FF demonstrates better performance than that of state-of-the-art end-to-end deep learning models and differently experienced human experts on both internal dataset and external dataset from an independent hospital without additional model tuning. With only one GPU (GeForce RTX 2080) FF is able to perform the classification in magnitude of milliseconds per frame, possessing a great potential to aid sonographers in capturing standard sagittal planes in real time.

We attribute the effectiveness of FF to the divide-and-conquer approach that decomposes the S-NS classification task into a hierarchy of object detection steps. Each step focuses on detecting one structure or region of structures which is a necessary indicator for the final standard plane classification. This explicit merit-based mechanism enables FF to perform better than state-of-the-art end-to-end models with only a dataset of 1700 images. Meanwhile, FF's predicted and labeled structures provide lucid model interpretation to examiners, making it convenient for quality assurance and adjustment if human expert decides to overturn the result from AI.

In essence, FF is a hierarchy of object detectors. It is designed that deep learning modules within FF, including object detectors and classifier, can be updated by state-of-the-art models, thus improving FF's performance. In addition, FF can be generalized as a classifier building on a navigator (CBN) if object locations are measurably relative in the task. For instance, future extension of the proposed framework can be potentially applied to other tasks of obstetrical US examination, such as measuring head circumference or biparietal diameter, and organ detection in the first trimester, as well as measuring cardiac activity or abdominal circumference in later trimesters. On the other hand, CBN can be potentially useful to other US exams such as screening for thyroid cancer, where CBN locates a thyroid nodule by searching for trachea, carotid artery, and thyroid, then the classifier assesses the condition of the nodule. In short, CBN aims to standardize the process of identifying representative structures within a US scan, and the classifier makes predictions accordingly.

FF can be further optimized by training with more data in the absence of nasal bone which is also considered acceptable in the standard sagittal view. NT measurement can be also automated by using deep CNN-based image segmentation techniques after FF as a completion of NT exam. A wide spectrum of NT metrics in both preconditioned or normal cases can not only help further train the framework, but also establish a baseline of automated NT calibration for chromosome defect prediction in clinical setting. The proposed framework has proven robust in initial experiments, but more work can be done to understand model adaptability with a diversity of clinical situations.

To our knowledge, this work is the first to propose a merit-based framework without a specific CNN model requirement to automate key structure measurement in obstetric ultrasound exams. By working with key partners of this work, we expect to release full dataset after de-identification as well as our full code and data-processing workflow. With these efforts, we hope to encourage better model development and performance comparison in AI and ultrasound communities.

Declarations

Competing interests: The authors declare no competing interests.

References

1. Woo, J. A short history of the development of ultrasound in obstetrics and gynecology. *Hist. Ultrasound Obstet. Gynecol.* **3**, 1–25 (2002).

2. Whitworth, M., Bricker, L. & Mullan, C. Ultrasound for fetal assessment in early pregnancy. *Cochrane Database Syst. Rev.* **2015**, (2015).

3. Ghaffari, S. R. *et al.* First-trimester screening for chromosomal abnormalities by integrated application of nuchal translucency, nasal bone, tricuspid regurgitation and ductus venosus flow combined with maternal serum free β-hCG and PAPP-A: A 5-year prospective study. *Ultrasound Obstet. Gynecol.* **39**, 528–534 (2012).

4. Avgidou, K., Papageorghiou, A., Bindra, R., Spencer, K. & Nicolaides, K. H. Prospective first-trimester screening for trisomy 21 in 30,564 pregnancies. *Am. J. Obstet. Gynecol.* **192**, 1761–1767 (2005).

5. Nicolaides, K. H. Screening for chromosomal defects. *Ultrasound Obstet. Gynecol.* **21**, 313–321 (2003).

6. Fetal Chromosomal. 704–707 (1992).

7. Bittles, A. H., Bower, C., Hussain, R. & Glasson, E. J. The four ages of Down syndrome. **17**, 221–225 (2006).

8. Frid, C., Drott, P., Lundell, B., Rasmussen, F. & Annere, G. Mortality in Down ' s syndrome in relation to congenital malformations. **43**, 234–241 (1999).

9. Berger, V. K. *et al.* The utility of nuchal translucency ultrasound in identifying rare chromosomal abnormalities not detectable by cell-free DNA screening. 185–190 (2020) doi:10.1002/pd.5583.

10. Karim, J. N., Roberts, N. W., Salomon, L. J. & Papageorghiou, A. T. Systematic review of first-trimester ultrasound screening for detection of fetal structural anomalies and factors that affect screening performance Search strategy. 429–441 (2017) doi:10.1002/uog.17246.

11. Zalel, Y. & Zemet, R. The added value of detailed early anomaly scan in fetuses with increased nuchal translucency. (2016) doi:10.1002/pd.4997.

12. Sebire, N. J., Murphy, K. W., Carvalho, J. S. & Hall, C. M. Increased first-trimester fetal nuchal translucency thickness in association with chondroectodermal dysplasia (Ellis – Van Creveld syndrome). 412–414 (2005) doi:10.1002/uog.1849.

13. Huang, W. H. & Porto, M. Abnormal First-Trimester Fetal Nuchal Translucency and Cornelia de Lange Syndrome. **99**, 956–958 (2002).

14. Orlandi, F. *et al.* Measurement of nasal bone length at 11-14 weeks of pregnancy and its potential role in Down syndrome risk assessment. *Ultrasound Obstet. Gynecol.* **22**, 36–39 (2003).

15. Bunduki, V. *et al.* Fetal nasal bone length: Reference range and clinical application in ultrasound screening for trisomy 21. *Ultrasound Obstet. Gynecol.* **21**, 156–160 (2003).

16. Cicero, S., Avgidou, K., Rembouskos, G., Kagan, K. O. & Nicolaides, K. H. Nasal bone in first-trimester screening for trisomy 21. *Am. J. Obstet. Gynecol.* **195**, 109–114 (2006).

17. Committee, C. S. isuog. 102–113 (2013) doi:10.1002/uog.12342.

18. Marinescu, P. S. *et al.* 203: Maintaining optimal performance: Characteristics of providers requiring remediation within the Nuchal Translucency Quality Review program. *Am. J. Obstet. Gynecol.* **222**, S140 (2020).

19. Bjerring, L., Rice, B. & Okun, N. Nuchal Translucency Quality Assurance (NTQA) in Ontario. *J. Med. Imaging Radiat. Sci.* **50**, S11 (2019).

20. Palomaki, G. E. *et al.* Quality assessment of routine nuchal translucency measurements: A North American laboratory perspective. *Genet. Med.* **10**, 131–138 (2008).

21. Sahota, D. S. *et al.* Quality assurance of nuchal translucency for prenatal fetal Down syndrome screening. *J. Matern. Neonatal Med.* **25**, 1039–1043 (2012).

22. Nisbet, D., Robertson, A., Mannil, B., Pincham, V. & Mclennan, A. Quality management of nuchal translucency ultrasound measurement in Australia. *Aust. New Zeal. J. Obstet. Gynaecol.* **59**, 54–58 (2019).

23. Koster, M. P. H. *et al.* Quality of nuchal translucency measurements in the Netherlands: A quantitative analysis. *Ultrasound Obstet. Gynecol.* **34**, 136–141 (2009).

24. Evans, M. I., Evans, S. M., Bennett, T. A. & Wapner, R. J. The price of abandoning diagnostic testing for cell-free fetal DNA screening. *Prenat. Diagn.* **38**, 243–245 (2018).

25. Evans, M. I., Van Decruyes, H. & Nicolaides, K. H. Nuchal translucency measurements for first-trimester screening: The 'price' of inaccuracy. *Fetal Diagn. Ther.* **22**, 401–404 (2007).

26. Kagan, K. O., Wright, D., Etchegaray, A., Zhou, Y. & Nicolaides, K. H. Effect of deviation of nuchal translucency measurements on the performance of screening for trisomy 21. *Ultrasound Obstet. Gynecol.* **33**, 657–664 (2009).

27. Fillman, T., Matteson, J., Sciortino, S. & Saha, S. 1058: Disparities in nuchal translucency uptake in California. *Am. J. Obstet. Gynecol.* **220**, S678–S679 (2019).

28. Yaqub, M., Kelly, B., Papageorghiou, A. T. & Noble, J. A. A Deep Learning Solution for Automatic Fetal Neurosonographic Diagnostic Plane Verification Using Clinical Standard Constraints. *Ultrasound Med. Biol.* **43**, 2925–2933 (2017).

29. Gao, Y. & Alison Noble, J. Detection and characterization of the fetal heartbeat in free-hand ultrasound sweeps with weakly-supervised two-streams convolutional networks. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **10434 LNCS**, 305–313 (2017).

30. Ravishankar, H., Prabhu, S. M., Vaidya, V. & Singhal, N. Hybrid approach for automatic segmentation of fetal abdomen from ultrasound images using deep learning. *Proc. - Int. Symp. Biomed. Imaging* **2016-June**, 779–782 (2016).

31. Huang, W., Bridge, C. P., Noble, J. A. & Zisserman, A. Temporal HeartNet: Towards human-level automatic analysis of fetal cardiac screening video. *arXiv* 341–349 (2017) doi:10.1007/978-3-319-66185-8.

32. Lin, T. Y., Goyal, P., Girshick, R., He, K. & Dollar, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 318–327 (2020).

33. Delong, E. R. & Carolina, N. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves : A Nonparametric Approach Author (s): Elizabeth R . DeLong , David M . DeLong and Daniel L . Clarke-Pearson Published by : International Biometric Society Stable . *Biometrics* **44**, 837–845 (2016).

34. Sun, X. & Xu, W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process. Lett.* **21**, 1389–1393 (2014).

Tables 1-3

Tables 1-3 are available in the Supplementary Files section.

Figures

Figure 1

The Fetus Framework. a). The input ROI. b). The input ROI with detected palate, midbrain, and thalami boxes with scores by CNN detector 1. c). The refined ROIs proposed by the RPM. The red, orange, and green regions are nasal, IT&CM, and NT regions, respectively. d). The nasal ROI is scaled up by 1.5 times. e). The IT&CM region. f). The NT region is scaled up by 3 times. g). The nasal region with detected nasal bone, skin, and tip boxes with scores. h). The IT&CM region with detected IT and CM boxes and scores. i). The NT region with detected NT box and score. j). The LR classifier predicts the probability of the input as standard by using the 9 structure scores gathered from b), g), h), and i).

Figure 2

See image above for figure legend

Figure 3

Experiment work flow. The entire experiment can be divided into three major components: 1) data preparation from collection, exclusion, annotation and preprocessing. 2) FF training and validation with data from one major center. 3) Internal and external model testing with multi-center datasets.

Figure 4

See image above for figure legend

Figure 5

A non-standard plane that did not pass the RPM because the relative locations of palate, thalami, and midbrain does not follow the RPM's criteria. a. The input plane. b. The detected palate, thalami, and midbrain of the input plane.

Figure 6

a). The dashed green box is the NT region (scaled vertically by 3 times) proposed by RPM. The solid green box is the ground truth box for NT. The blue dot, orange dot, and red dot are the bottom left vertex, center, and top right vertex of the ground truth box for NT. b). Three dots mark the location of NT within the NT region. c). A standardized version of b).

Figure 7

AUC generated by Fetus Framework, Xception, ResNet and DenseNet in a) internal test with 156 cases from the major medical center and b) external test with 156 cases from the affiliated medical center. c).

Test results for both internal (Test set A) and external (Test set B). SP: standard plane; Non-SP: nonstandard plane. Non SP and SP differentiation becomes less obvious among observers but more consistent among models with the highest precision from the proposed framework.

Figure 8

Example of Fetus Framework applications in video cases. This video was a clip collected by a senior ultrasound specialist from a real exam at the Department of Ultrasound of our major medical center. Video contains over 120 frames and FF detects 7 sagittal standard planes with accepted presentation of the nine structures. The intensity change at nasal and NT regions are subtle with milliseconds, posing difficulties for examiners to capture the best moment for measurement. FF is able to capture and save the frame automatically without extra effort from doctors.

Figure 9

The distribution of nasal bone, nasal skin, nasal tip, IT, CM, and NT inside the ground truth-based RPM-proposed regions for both internal and external data.

Figure 10

The distribution of nasal bone, nasal skin, nasal tip, IT, CM, and NT inside the model-based RPM-proposed regions for both internal and external data.

Figure 11

Fig 12. Receiver operating characteristic curves of random forest on physician selected dataset and fetus selected dataset.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.png
- Table2.png

- Table3.png